Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αυτόματη Περίληψη Κειμένου: Μηχανική Μάθηση και Σημασιολογικές Τεχνικές

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Παναγιώτη Ε. Κουρή

Αθήνα, Δεκέμβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αυτόματη Περίληψη Κειμένου: Μηχανική Μάθηση και Σημασιολογικές Τεχνικές

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Παναγιώτη Ε. Κουρή

Συμβουλευτική Επιτροπή: Γεώργιος Στάμου
Στέφανος Κόλλιας
Ηρακλής Βαρλάμης

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή τη 12η Δεκεμβρίου 2022

.....	.....	.....
Γεώργιος Στάμου	Στέφανος Κόλλιας	Ηρακλής Βαρλάμης
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Αναπληρωτής Καθηγητής
		Χαροκοπείου Πανεπιστημίου
.....	.....	.....
Αθανάσιος Βουλόδημος	Παναγιώτης Τσανάκας	Ελένη Ευθυμίου
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Ερευνήτρια Α΄
		Ερευνητικού Κέντρου “Αθηνά”

.....
Γεώργιος Καρυδάκης
Αναπληρωτής Καθηγητής
Πανεπιστημίου Αιγαίου

Αθήνα, Δεκέμβριος 2022

.....

Παναγιώτης Ε. Κουρής
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η συνεχώς αυξανόμενη πληροφορία κειμένου έχει οδηγήσει στην ανάπτυξη έντονης ερευνητικής δραστηριότητας στο πεδίο της αυτόματης περίληψης κειμένου, το οποίο αποτελεί έναν σημαντικό ερευνητικό τομέα της επεξεργασίας φυσικής γλώσσας. Η έρευνα, που διεξάγεται σήμερα στο πλαίσιο της αυτόματης περίληψης κειμένου, επικεντρώνεται κυρίως σε ανάπτυξη προσεγγίσεων μηχανικής μάθησης, χωρίς, τις περισσότερες φορές, να εξετάζεται ο συνδυασμός μοντέλων μηχανικής μάθησης με άλλες τεχνικές που βασίζονται σε επεξεργασία φυσικής γλώσσας, οι οποίες θα μπορούσαν να συνεισφέρουν στην περαιτέρω βελτίωση του πεδίου αυτού. Με αφορμή το ερευνητικό αυτό κενό, η παρούσα διδακτορική διατριβή, με αντικείμενο την αυτόματη περίληψη κειμένου ενός εγγράφου με τη μέθοδο της παραγωγής κειμένου, εξετάζει αρχιτεκτονικές βαθιάς μάθησης και παρουσιάζει νέες μεθόδους που συνδυάζουν μηχανική μάθηση και σημασιολογικές τεχνικές, με σκοπό τη βελτίωση της αυτόματης περίληψης κειμένου.

Η συνεισφορά της διατριβής περιλαμβάνει: (i) τη διερεύνηση αρχιτεκτονικών βαθιάς μάθησης για την αυτόματη περίληψη κειμένου, (ii) την πρόταση νέας μεθοδολογίας σημασιολογικών μετασχηματισμών του περιεχομένου και μηχανικής μάθησης για την αντιμετώπιση του προβλήματος της διαχείρισης νέου κειμένου, το οποίο δεν έχει επαρκή παρουσία στο σύνολο εκπαίδευσης ενός μοντέλου μηχανικής μάθησης, (iii) την εισαγωγή ενός νέου πλαισίου σημασιολογικής αναπαράστασης του περιεχομένου και βαθιάς μάθησης, προς την κατεύθυνση της παραγωγής περιλήψεων με σημασιολογική συνάφεια περιεχομένου και (iv) την παρουσίαση ενός συνόλου μετρικών για παροχή ποιοτικής αξιολόγησης του περιεχομένου των παραγόμενων περιλήψεων.

Στο πρώτο μέρος η ερευνητική προσπάθεια εστιάζει στη διερεύνηση αρχιτεκτονικών βαθιάς μάθησης και προτείνει την αξιοποίηση κατάλληλων μοντέλων νευρωνικών δικτύων για την αυτόματη περίληψη κειμένου. Στην κατεύθυνση αυτή, διερευνώνται διαφορετικές αρχιτεκτονικές μοντέλων μηχανικής μάθησης, όπως είναι τα δίκτυα νευρωνικών δικτύων τύπου κωδικοποιητή-αποκωδικοποιητή, η ενισχυτική μάθηση, οι αρχιτεκτονικές μετασχηματιστών ή τα προεκπαιδευμένα μοντέλα γλωσσικής αναπαράστασης.

Στη συνέχεια, η έρευνα εστιάζει στην ανάπτυξη ενός πλαισίου σημασιολογικών μετασχηματισμών του περιεχομένου, το οποίο δίνει λύσεις στο πρόβλημα των νέων υποψήφιων για περίληψη κειμένων, τα οποία περιλαμβάνουν περιεχόμενο που ενδεχομένως δεν έχει επαρκή παρουσία στο σύνολο εκπαίδευσης ενός μοντέλου μηχανικής μάθησης. Το προτεινόμενο πλαίσιο περιλαμβάνει τρία βασικά στάδια: την προ-επεξεργασία, τις προβλέψεις μηχανικής μάθησης και τη μετα-επεξεργασία. Το στάδιο της προ-επεξεργασίας βασίζεται σε μια καλά καθορισμένη μεθοδολογία γενίκευσης περιεχομένου, η οποία αξιοποιεί πόρους γνώσης, ταξινομίες εννοιών, σημασιολογική αποσαφήνιση έννοιας λέξεων και αναγνώριση ονοματικών οντοτήτων για τον

μετασχηματισμό του περιεχομένου σε μια γενικευμένη μορφή. Η εφαρμογή της μεθοδολογίας γενίκευσης του περιεχομένου βελτιώνει την ακρίβεια των προβλέψεων μηχανικής μάθησης, με την παραγωγή περιλήψεων σε μια γενικευμένη μορφή. Το στάδιο της μετα-επεξεργασίας βασίζεται σε ευρεστικές μεθόδους, που αξιοποιούν αντίστοιχους πόρους γνώσης με εκείνους που χρησιμοποιούνται στη φάση της προ-επεξεργασίας, για τον μετασχηματισμό των γενικευμένων περιλήψεων στην τελική τους μορφή.

Στο τρίτο μέρος, η ερευνητική προσπάθεια επικεντρώνεται στην αξιοποίηση της σημασιολογικής αναπαράστασης του περιεχομένου σε μορφή γραφήματος, σε συνδυασμό με προβλέψεις βαθιάς μάθησης για τη βελτίωση της αυτόματης περίληψης κειμένου, προς την κατεύθυνση της παραγωγής περιλήψεων με σημασιολογική συνάφεια περιεχομένου. Η κύρια συνεισφορά της προτεινόμενης μεθοδολογίας περιλαμβάνει τη μοντελοποίηση του προβλήματος ως ένα πρόβλημα μάθησης από γράφημα σε περίληψη με μεθόδους βαθιάς μάθησης, την παρουσίαση σημασιολογικών αναπαραστάσεων κειμένου σε μορφή γραφήματος και τη διερεύνηση της επίδοσης διαφορετικών αρχιτεκτονικών βαθιάς μάθησης σε συνδυασμό με διάφορα σχήματα δεδομένων. Η προτεινόμενη προσέγγιση βασίζεται σε ένα καλά καθορισμένο πλαίσιο για την ανάκτηση των σημασιολογικών γραφημάτων για κάθε περίοδο ενός αρχικού κειμένου, την κατασκευή του σημασιολογικού γραφήματος του περιεχομένου ενός κειμένου, τον μετασχηματισμό ενός σημασιολογικού γραφήματος σε κατάλληλη μορφή για είσοδο σε κάποιο μοντέλο μηχανικής μάθησης και τις προβλέψεις μηχανικής μάθησης. Η προσέγγιση αυτή οργανώνει τη μη δομημένη πληροφορία και αναπαριστά σημασιολογικά το περιεχόμενο, σε μια προσπάθεια βελτίωσης των προβλέψεων μηχανικής μάθησης και την παροχή περιλήψεων με σημασιολογική συνάφεια περιεχομένου.

Προς την κατεύθυνση της παροχής μιας ποιοτικής αξιολόγησης για την αυτόματη περίληψη κειμένου, η παρούσα διατριβή προτείνει ένα νέο σύνολο μετρικών, οι οποίες προσδιορίζουν τη συνέπεια απόδοσης πληροφορίας των παραγόμενων περιλήψεων σε σχέση με το αρχικό κείμενο. Οι εν λόγω μετρικές παρέχουν μια σταθμισμένη τιμή αξιολόγησης, σύμφωνα με την έκταση του αρχικού κειμένου και της περίληψης συστήματος, προσδιορίζοντας τη σημασιολογική επικάλυψη μεταξύ της πληροφορίας που περιλαμβάνει η παραγόμενη περίληψη σε σχέση με το αρχικό κείμενο. Το νέο σύνολο μετρικών μπορεί να συνεισφέρει στην αξιολόγηση και βελτίωση των συστημάτων αυτόματης περίληψης κειμένου.

Οι προσεγγίσεις που παρουσιάζονται, καθορίστηκαν θεωρητικά, υλοποιήθηκαν και διερευνήθηκαν πειραματικά. Στο πλαίσιο της πειραματικής διαδικασίας εξετάστηκαν σημαντικές πτυχές της προτεινόμενης μεθοδολογίας, καθώς, επίσης, διερευνήθηκε και συγκρίθηκε η επίδοση της εκάστοτε προσέγγισης με άλλες συναφείς εργασίες. Ο προσδιορισμός των βέλτιστων επιλογών, που οδηγούν στη βελτιστοποίηση της επίδοσης των προτεινόμενων λύσεων, τα θετικά αποτελέσματα, καθώς και τα συμπεράσματα, που προέκυψαν, αναδεικνύουν τα οφέλη της παρούσας ερευνητικής προσπάθειας, η οποία μπορεί να οδηγήσει στην περαιτέρω έρευνα για την ανάπτυξη ευφυών συστημάτων στον τομέα της αυτόματης περίληψης κειμένου.

Λέξεις Κλειδιά

Αυτόματη Περίληψη Κειμένου, Μηχανική Μάθηση, Νευρωνικά Δίκτυα, Βαθιά Μάθηση, Αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή, Ενισχυτική Μάθηση, Δίκτυα Μετασχηματιστών,

Προ-εκπαιδευμένα Μοντέλα, Επεξεργασία Φυσικής Γλώσσας, Σημασιολογικές Τεχνικές, Ταξινομία Εννοιών, Αποσαφήνιση Έννοιας Λέξεων, Αναγνώριση Ονοματικών Οντοτήτων, Σημασιολογικό Γράφημα, Αναπαράσταση Αφηρημένης Εννοιας, Γράφημα σε Περίληψη, Μετρικές Αξιολόγησης, Συνέπεια Απόδοσης Πληροφορίας

Abstract

The constantly growing amount of textual information has led to the development of automatic text summarization, which constitutes an important research area in natural language processing. The current research that is conducted in this field is mainly focused on developing machine learning approaches, without, in most cases, considering the combination of machine learning models with other techniques based on natural language processing, which could contribute to further improvement in this field. In view of this research gap, the present dissertation focuses on the field of abstractive text summarization of single documents, examining deep learning architectures and presenting new methodologies that combine machine learning and semantic-based techniques in order to improve automatic text summarization.

The contribution of the dissertation includes; (*i*) the investigation of deep learning architectures for automatic text summarization, (*ii*) a novel methodology that is based on semantic content transformations and machine learning to address the problem of managing new content, without sufficient presence in the training set of a machine learning model, (*iii*) a new framework that combines methodology of semantic content representation and deep learning, towards the production of summaries with semantic content relevance and (*iv*) a set of metrics that provides a qualitative assessment of the estimated summaries.

The first part covers the investigation of a range of deep learning architectures for estimating a sequence of words that composes the summary of an original text. These architectures include encoder-decoder recurrent neural networks, reinforcement learning, transformer-based architectures, and pre-trained neural language models.

The second part presents a novel framework that is based on semantic content transformations along with machine learning predictions. The proposed framework is capable of dealing with the problem of out-of-vocabulary or rare words, improving the performance of the deep learning models. The framework is composed of three components; a pre-processing task, a machine learning methodology and a post-processing task. The pre-processing task is based on a well-defined theoretical model of semantic-based content generalization, which utilizes ontological knowledge resources, word-sense-disambiguation and named-entity recognition to transform ordinary text into a generalized form. A range of deep learning models is trained on a generalized version of text-summary pairs, learning to predict summaries in a generalized form. The post-processing task utilizes knowledge resources, word embeddings, word-sense disambiguation and heuristic algorithms based on text similarity methods to transform the generalized version of a predicted summary into a final, human-readable form.

The third part includes a novel approach that combines semantic graph representations

along with deep learning predictions to generate abstractive summaries of single documents, in an effort to utilize a semantic representation of the unstructured textual content in a machine-readable, structured, and concise manner. The main contribution of this approach includes; the graph-to-summary formulation of the problem of abstractive text summarization using deep learning techniques, the examination of a range of deep learning models, and the investigation of semantic graph-based representation schemes. The overall framework is based on a well-defined methodology for performing semantic graph parsing, graph construction, graph transformations for machine learning models, and deep learning predictions. This approach organizes unstructured textual information through a semantic representation of the content, in an effort to improve machine learning predictions and provide semantically relevant summaries.

Another important contribution is an introduction of a set of measures for assessing the factual consistency of the generated summaries in an effort to provide a qualitative evaluation. These measures provide a weighted evaluation value, according to the length of an original text and the system summary, by determining the semantic overlap between the information contained in the generated summary and the original text. The new set of metrics can contribute to the evaluation and improvement of automatic text summarization systems.

The approaches presented were theoretically defined, implemented and experimentally investigated. Considering the research conducted, the novel methodology, the positive results and the useful conclusions may contribute to further improvement of intelligent systems in the field of automatic text summarization.

Keywords

Automatic Text Summarization, Abstractive Text Summarization, Machine Learning, Neural Networks, Deep Learning, Encoder-Decoder Architecture, Reinforcement Learning, Transformer Networks, Pre-trained Models, Natural Language Processing, Semantic Techniques, Taxonomy of Concepts, Word Sense Disambiguation, Named-Entity Recognition, Semantic Graph, Abstract Meaning Representation, Graph-to-Summary, Evaluation Metrics, Factual Consistency

Ευχαριστίες

Η παρούσα διδακτορική διατριβή εκπονήθηκε, από τον Νοέμβριο του 2016 έως τον Δεκέμβριο του 2022, στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης, του τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ) υπό την επίβλεψη του κύριου Ανδρέα-Γεωργίου Σταφυλοπάτη, Καθηγητή του ΕΜΠ. Θα ήθελα να ευχαριστήσω θερμά τον κύριο Σταφυλοπάτη για την εμπιστοσύνη, την ενθάρρυνση, την υποστήριξη, το ενδιαφέρον τόσο σε ακαδημαϊκό όσο και σε προσωπικό επίπεδο, καθώς και τη βοήθεια που μου προσέφερε σε όλη τη διάρκεια των Μεταπτυχιακών-Διδακτορικών μου σπουδών. Ο κύριος Σταφυλοπάτης, είχε τυπικά την επίβλεψη της διδακτορικής μου διατριβής έως τον Οκτώβριο του 2022. Ουσιαστικά, όμως, με στήριξε μέχρι την ολοκλήρωση των σπουδών μου.

Οφείλω επίσης ευχαριστίες στον κύριο Γεώργιο Στάμου, Καθηγητή ΕΜΠ, αρχικά, για τη συμβολή του ως μέλους της τριμελούς συμβουλευτικής επιτροπής και, στη συνέχεια, από τον Οκτώβριο του 2022, για την ανάληψη της επίβλεψης της διδακτορικής μου διατριβής, λόγω υπηρεσιακού κωλύματος του κύριου Σταφυλοπάτη. Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον κύριο Στάμου τόσο για τη συμβολή του στην ολοκλήρωση της διδακτορικής μου διατριβής όσο και για τη συνεργασία που είχαμε στο άρθρο σημασιολογικού εμπλουτισμού, που οι εύστοχες παρατηρήσεις του με βοήθησαν να βελτιώσω την ποιότητα της δουλειάς μου στα μετέπειτα ερευνητικά μου βήματα.

Θα ήθελα να ευχαριστήσω τον κύριο Ηρακλή Βαρλάμη, Αναπληρωτή Καθηγητή του Χαροκοπείου Πανεπιστημίου, για τη συμμετοχή του στη συμβουλευτική επιτροπή εκπόνησης της διδακτορικής μου διατριβής. Ιδιαίτερα, οφείλω ευχαριστίες στον κύριο Βαρλάμη για τη συμβολή του στα πρώτα ερευνητικά μου βήματα, τόσο κατά την εκπόνηση της διπλωματικής μου εργασίας στο πλαίσιο των μεταπτυχιακών μου σπουδών στο Χαροκόπειο Πανεπιστήμιο, όσο και για τη βοήθεια που μου προσέφερε κατά τη συνέχιση της ερευνητικής μου προσπάθειας και τη συνεργασία μας στη συγγραφή και δημοσίευση ερευνητικών άρθρων στον τομέα των συστημάτων συστάσεων, που αποτέλεσαν την αφετηρία των διδακτορικών μου σπουδών.

Θα ήθελα, επίσης, να ευχαριστήσω τον κύριο Στέφανο Κόλλια, Καθηγητή του ΕΜΠ, αρχικά για τη συμμετοχή του στην επιτροπή εξέτασης ενδιάμεσης κρίσης και, στη συνέχεια, από τον Οκτώβριο του 2022, για τη συμμετοχή του στην συμβουλευτική επιτροπή εκπόνησης της διδακτορικής μου διατριβής.

Θα ήθελα να ευχαριστήσω τον κύριο Γιώργο Αλεξανδρίδη, ΕΔΙΠ του ΕΜΠ, για τη συνεργασία που είχαμε και την υποστήριξη που μου προσέφερε κατά τη διάρκεια των ερευνητικών μου βημάτων. Οι εύστοχες παρατηρήσεις του, η αυστηρή επιστημονική του ματιά, οι παραινέσεις και η ενθάρρυνση του με βοήθησαν να αντεπεξέλθω στη δύσκολη αυτή πορεία προς την ολοκλήρωση της διδακτορικής

μου διατριβής.

Θα ήθελα, επίσης, να ευχαριστήσω την κυρία Ελένη Ευθυμίου, Ερευνήτρια του Ε.Κ. «Αθηνά», τον κύριο Παναγιώτη Τσανάκα, Καθηγητή ΕΜΠ, τον κύριο Αθανάσιο Βουλόδημο, Επίκουρο Καθηγητή του ΕΜΠ, και τον κύριο Γιώργο Καρυδάκη, Αναπληρωτή Καθηγητή του Πανεπιστημίου Αιγαίου για την τιμή που μου έκαναν να συμμετάσχουν στην επιτροπή αξιολόγησης της διδακτορικής μου διατριβής.

Θα ήθελα να ευχαριστήσω τον κύριο Γιάννη Μαΐστρο, Επίκουρο Καθηγητή του ΕΜΠ, και την κυρία Στέλλα Μαρκαντωνάτου, Ερευνήτρια του Ι.Ε.Λ., Ε.Κ. «Αθηνά», τόσο για τη διδασκαλία του μεταπτυχιακού μαθήματος «Παράσταση και Επεξεργασία Γλωσσικής Γνώσης», το οποίο με έφερε σε επαφή με θέματα επεξεργασίας φυσικής γλώσσας και διαμόρφωσε την κατεύθυνση της διδακτορικής μου διατριβής, όσο και για τη μετέπειτα συνεργασία μας στο πλαίσιο ερευνητικής δουλειάς σε συναφές αντικείμενο με το προαναφερόμενο μάθημα. Ευχαριστώ, επίσης, την κυρία Μαρκαντωνάτου που ως επιστημονική υπεύθυνη με εμπιστεύτηκε σε θέση ερευνητικού προγράμματος του Ε.Κ. «Αθηνά», στο οποίο συμμετείχα ως Υ.Δ. του ΕΜΠ. Θα ήθελα να ευχαριστήσω και τους ερευνητές του εν λόγω ερευνητικού προγράμματος, κυρία Μαριέττα Σιόντη και κύριο Βαλάντη Κορφύτη, Διδάκτορες του Καποδιστριακού Πανεπιστημίου, για την άριστη συνεργασία που είχαμε.

Θα ήθελα να ευχαριστήσω όλα τα μέλη του εργαστηρίου, κυρίους Γιώργο Σιόλα, Ε.ΔΙ.Π. του ΕΜΠ, Άρη Λαναρίδη, Γιώργο Στρατογιάννη, κυρία Ελένη Βάθη, κύριο Θάνο Τάγαρη (Διδάκτορες του ΕΜΠ), κυρίους Τάσο Παπαγιάννη, Γιώργο Ιωάννου και Θάνο Τασάκο (Υ.Δ. του ΕΜΠ) για την ευχάριστη ατμόσφαιρα που επικρατούσε στο εργαστήριο και την άριστη συνεργασία που είχαμε.

Μπορεί να μην είχα φτάσει έως εδώ χωρίς τη συμβολή του κύριου Χρήστου Χήρα, Εκπαιδευτικού μέσης εκπαίδευσης, ο οποίος στα μαθητικά μου χρόνια με ενέπνευσε πραγματικά προς την πορεία της γνώσης.

Θα ήθελα να ευχαριστήσω τα αδέρφια και τους γονείς μου, που ο καθένας είχε τη συμβολή του στον δύσκολο αυτό δρόμο από τις προπτυχιακές έως τις διδακτορικές μου σπουδές. Θα ήθελα να ευχαριστήσω τη σύζυγό μου, Λένα, για την υπομονή, την ανοχή και τη βοήθεια που μου προσέφερε όλα αυτά τα χρόνια για να μπορέσω να αντεπεξέλθω στον δύσκολο δρόμο προς την ολοκλήρωση των διδακτορικών μου σπουδών. Τέλος, θα ήθελα να ευχαριστήσω τους δύο μου γιους, Άγγελο και Σοφοκλή, που με τον δικό τους τρόπο, μου επέτρεψαν να συνεχίσω τις σπουδές μου και, συγχρόνως, θα ήθελα να τους υποσχεθώ ότι θα προσπαθήσω να αναπληρώσω τον χρόνο που μας στέρησε η αφοσίωσή μου στις διδακτορικές μου σπουδές.

Αφιερώνεται στους γιους μου
Άγγελο και Σοφοκλή

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά

Η παρούσα διατριβή εστιάζει στο πρόβλημα της αυτόματης *περίληψης κειμένου (ΠΚ) (text summarization - TS)* και, πιο συγκεκριμένα, επικεντρώνεται στο πεδίο της αυτόματης *ΠΚ* ενός εγγράφου με τη μέθοδο της παραγωγής κειμένου, όπως αναφέρεται με λεπτομέρεια στη συνέχεια. Στην Ενότητα 1.2 που ακολουθεί περιγράφεται το πρόβλημα της αυτόματης *ΠΚ*, οι βασικές κατευθύνσεις έρευνας και οι προκλήσεις για περαιτέρω μελέτη. Η συνεισφορά της διατριβής συνοψίζεται στην Ενότητα 1.3 και η δομή της διατριβής παρουσιάζεται στην Ενότητα 1.4.

1.2 Πρόβλημα και προκλήσεις

Η συνεχώς αυξανόμενη πληροφορία κειμένου που διατίθεται, κυρίως μέσω του διαδικτύου, έχει καταστήσει την πρόσβαση σε αυτή μια δύσκολη εργασία και, κατά συνέπεια, είναι επιτακτική η ανάγκη ανάπτυξης αυτοματοποιημένων μεθόδων με σκοπό την επεξεργασία και την εύκολη κατανόηση δεδομένων κειμένου. Ένας από τους κύριους τρόπους επίτευξης αυτού του στόχου είναι η ανάπτυξη μεθοδολογίας μείωσης της διάστασης του αρχικού κειμένου, μετατρέποντάς το σε ένα νέο κείμενο συνοπτικής μορφής, το οποίο, ιδανικά, ανταποκρίνεται στο περιεχομένου του αρχικού κειμένου. Σε αυτή την κατεύθυνση συμβάλλει η ανάπτυξη του πεδίου της αυτόματης *ΠΚ* που αποτελεί ενεργό ερευνητικό πεδίο για περισσότερο από μισό αιώνα [1]. Ο στόχος του πεδίου αυτού είναι η ανάπτυξη συστημάτων τα οποία παράγουν ενημερωτικές και αναγνώσιμες από τον άνθρωπο περιλήψεις, διατηρώντας την ουσία και το περιεχόμενο του αρχικού κειμένου. Από την εμφάνιση των πρώτων εργασιών στον τομέα της αυτόματης *ΠΚ* [2, 3], έχουν αναπτυχθεί διάφορες προσεγγίσεις, οι οποίες, ανάλογα με την πηγή πληροφορίας και το είδος του αρχικού κειμένου, διακρίνονται κυρίως σε *ΠΚ ενός εγγράφου (single-document TS*, π.χ., άρθρα, ειδήσεις, ιστορίες, βιβλία, επιστημονικές εργασίες, πρόγνωση καιρού κλπ.), *ΠΚ πολλαπλών εγγράφων (multi-documents TS*, π.χ., κριτικές χρηστών, ειδήσεις από διάφορες πηγές, μηνύματα ηλεκτρονικού ταχυδρομείου κλπ.) και *ΠΚ που βασίζεται σε ερωτήσεις (query-based TS*, δηλ., *ΠΚ* που εστιάζει σε συγκεκριμένες πληροφορίες του κειμένου) [4, 5].

Επιπροσθέτως, οι τεχνικές αυτόματης *ΠΚ*, ανάλογα με την προσέγγιση σύνθεσης της

περίληψης, ταξινομούνται περαιτέρω σε δύο ομάδες: (*i*) *ΠΚ με τη μέθοδο της εξαγωγής κειμένου (extractive TS)* και (*ii*) *ΠΚ με τη μέθοδο της παραγωγής κειμένου (abstractive TS)* [6, 7, 5]. Η πρώτη μέθοδος βασίζεται στην εξαγωγή ενός υποσύνολου προτάσεων από το αρχικό κείμενο οι οποίες συνθέτουν την περίληψη. Η μέθοδος αυτή επιδιώκει η περίληψη που προκύπτει να περιλαμβάνει σημαντικές πληροφορίες από το αρχικό κείμενο και, παράλληλα, στοχεύει στην ελαχιστοποίηση του πλεονασμού της παρεχόμενης πληροφορίας. Η δεύτερη ομάδα προσεγγίσεων αυτόματης *ΠΚ* εστιάζει στη σύνθεση μιας αφηρημένης αναπαράστασης του αρχικού κειμένου, χρησιμοποιώντας τεχνικές παραγωγής φυσικής γλώσσας για τη δημιουργία των περιλήψεων. Με άλλα λόγια, ένα σύστημα παραγωγής κειμένου συνθέτει νέο κείμενο, το οποίο μπορεί να αποτελείται από λέξεις, φράσεις ή προτάσεις που ενδεχομένως δεν εμφανίζονται στο αρχικό κείμενο, και ταυτόχρονα ενσωματώνει σημαντικές πτυχές πληροφορίας του αρχικού κειμένου. Η μεθοδολογία που βασίζεται στην παραγωγή κειμένου έχει σκοπό τη δημιουργία περιλήψεων υψηλής ποιότητας όσον αφορά τη συνοχή, την αναγνωσιμότητα και τη μείωση του πλεονασμού. Κατά συνέπεια, η ανάπτυξη μίας τέτοιας προσέγγισης είναι ένα δύσκολο εγχείρημα, καθώς το είδος αυτό, της αυτόματης *ΠΚ* με την μέθοδο της παραγωγής κειμένου, εστιάζει στη σύνθεση περιλήψεων που μοιάζουν ή προσεγγίζουν τις περιλήψεις που γράφονται από τον άνθρωπο.

Είναι γνωστό ότι οι προσεγγίσεις της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου παρουσιάζουν μειωμένη απόδοση σε σύγκριση με τα μοντέλα εξαγωγής κειμένου [1, 8]. Παρά τις αδυναμίες τους, τα συστήματα παραγωγής κειμένου βελτιώνονται συνεχώς. Το κύριο πλεονέκτημά τους είναι η αντιμετώπιση των προβλημάτων συνοχής, πλεονασμού και αναφοράς, τα οποία είναι δύσκολο να αντιμετωπιστούν με τεχνικές εξαγωγής κειμένου. Επιπλέον, οι προσεγγίσεις παραγωγής κειμένου οδηγούν σε συνοπτικές περιλήψεις, μειώνοντας το μέγεθος των αρχικών προτάσεων (δηλ., επιτυγχάνουν συμπίεση ή συγχώνευση προτάσεων του αρχικού κειμένου) και ταυτόχρονα δημιουργούν συνεκτικές, γραμματικά σωστές και αναγνώσιμες από τον άνθρωπο περιλήψεις.

Έχουν αναπτυχθεί διάφορες προσεγγίσεις για την αυτόματη *ΠΚ* με τη μέθοδο της παραγωγής κειμένου, οι οποίες διακρίνονται κυρίως σε προσεγγίσεις που βασίζονται σε μοντέλα μηχανικής μάθησης, σε σημασιολογικές τεχνικές, σε πόρους γνώσης ή σε μεθοδολογία δομής περιεχομένου [9, 5]. Με δεδομένο ότι το μεγαλύτερο μέρος της έρευνας, που διεξάγεται σήμερα σε αυτόν το τομέα, εστιάζει περισσότερο στην ανάπτυξη μοντέλων μηχανικής μάθησης και λιγότερο σε άλλες τεχνικές για την αυτόματη *ΠΚ*, υπάρχει ερευνητικός χώρος για τη διερεύνηση του συνδυασμού των προαναφερόμενων κατευθύνσεων έρευνας, με σκοπό την περαιτέρω βελτίωση της αυτόματης *ΠΚ*.

Ένα ζήτημα που επηρεάζει αρνητικά τα συστήματα παραγωγής κειμένου, τα οποία βασίζονται σε μεθοδολογία μηχανικής μάθησης, είναι η περίληψη νέων κειμένων τα οποία ενδεχομένως περιλαμβάνουν *λέξεις εκτός λεξιλογίου (ΛΕΛ) (out-of-vocabulary - OOV)* ή σπάνιες λέξεις με περιορισμένη παρουσία στο σύνολο εκπαίδευσης ενός μοντέλου μηχανικής μάθησης. Αυτό το πρόβλημα έχει ισχυρή αρνητική επίδραση στα συστήματα μηχανικής μάθησης, τα οποία απαιτούν σύνολα εκπαίδευσης με επαρκή αριθμό παραδειγμάτων χρήσης για αποτελεσματικές προβλέψεις. Επίσης, τα συστήματα βαθιάς μάθησης, τα οποία σε σύγκριση με άλλες τεχνικές επιτυγχάνουν τις καλύτερες επιδόσεις στην αυτόματη *ΠΚ* με τη μέθοδο της παραγωγής κειμένου [9], σχεδόν αποτυγχάνουν να κάνουν ακριβείς προβλέψεις όταν το υποψήφιο προς περίληψη κείμενο περιλαμβάνει σπάνιες ή άγνωστες λέξεις (δηλ., περιεχόμενο με λίγες εμφανίσεις ή περιεχόμενο

που δεν περιλαμβάνεται στο σύνολο εκπαίδευσης του μοντέλου μηχανικής μάθησης). Για την αντιμετώπιση αυτού του προβλήματος, η εργασία εισάγει ένα νέο πλαίσιο που αποσκοπεί στην αποδοτική διαχείριση αυτής της κατηγορίας των δεδομένων, ενισχύοντας τις επιδόσεις των μοντέλων νευρωνικών δικτύων παραγωγής κειμένου.

Ένα δεύτερο ζήτημα, εξίσου σημαντικό, που επηρεάζει τις προβλέψεις μηχανικής μάθησης είναι η μορφή ενός υποψήφιου για περίληψη κειμένου, το οποίο, τις περισσότερες φορές, είναι σε μορφή μη δομημένης πληροφορίας. Με δεδομένο ότι η μορφή των δεδομένων εισόδου ενός μοντέλου μηχανικής μάθησης επηρεάζει την ποιότητα των προβλέψεων ενός τέτοιου συστήματος, η αναπαράσταση του αρχικού κειμένου έχει άμεση σχέση με την ποιότητα των παραγόμενων περιλήψεων. Επομένως, η μετατροπή του κειμένου σε μια δομημένη, αναγνώσιμη από τις μηχανές, και συνοπτική μορφή, η οποία περιλαμβάνει και στοιχεία σημασιολογίας, θα μπορούσε να επιφέρει βελτίωση στην αυτόματη *ΠΚ*. Το ζήτημα αυτό εξετάζεται στο ερευνητικό μέρος της παρούσας εργασίας, η οποία προτείνει νέα μεθοδολογία, που αξιοποιεί τη σημασιολογική αναπαράσταση του αρχικού κειμένου σε μορφή γραφήματος για την παραγωγή περιλήψεων.

Ένα ακόμη σημαντικό ζήτημα για τη βελτίωση της αυτόματης *ΠΚ* είναι η ανάπτυξη μεθοδολογίας για την αξιολόγηση των συστημάτων που αναπτύσσονται, με σκοπό την περαιτέρω βελτίωσή τους. Η αξιολόγηση των παραγόμενων περιλήψεων μπορεί να γίνει είτε από άνθρωπο είτε είτε με χρήση αυτοματοποιημένων μεθόδων [7, 10, 5, 11]. Η αξιολόγηση από άνθρωπο αποτελεί μια αξιόπιστη μέθοδο ποιοτικής αξιολόγησης των παραγόμενων περιλήψεων. Ωστόσο, αυτή η μορφή αξιολόγησης πολλές φορές εφαρμόζεται δύσκολα και μπορεί να προκαλέσει καθυστερήσεις στην ερευνητική διαδικασία. Η αξιολόγηση από άνθρωπο αποτελεί μια απαιτητική διαδικασία, καθώς, για να θεωρηθεί έγκυρη, χρειάζεται να αξιολογηθεί ένα πλήθος παραδειγμάτων χρήσης από ένα πλήθος αξιολογητών, οι οποίοι ιδανικά πρέπει να έχουν ως μητρική τους γλώσσα τη γλώσσα των παραδειγμάτων χρήσης που εξετάζουν, να έχουν λάβει κάποιο είδος εκπαίδευσης σχετικά με τα κριτήρια της αξιολόγησης και ο βαθμός συμφωνίας των αποτελεσμάτων αξιολόγησης των αξιολογητών να βρίσκεται σε ικανοποιητικά επίπεδα. Συνεπώς, η αξιολόγηση από άνθρωπο είναι μια εργασία που απαιτεί ανθρώπινο δυναμικό, χρόνο και πόρους. Το δεύτερο είδος αξιολόγησης που βασίζεται σε αυτοματοποιημένες μεθόδους μπορεί να εφαρμοστεί άμεσα και να οδηγήσει σε, επίσης, άμεση βελτίωση των συστημάτων κατά την ερευνητική διαδικασία. Ωστόσο, σε αντίθεση με την αξιολόγηση από άνθρωπο, οι μετρικές που έχουν αναπτυχθεί στο πλαίσιο της αυτοματοποιημένης αξιολόγησης παρέχουν κυρίως μια ποσοτική αξιολόγηση (π.χ., προσδιορισμός του βαθμού επικάλυψης των λέξεων μιας περίληψης σε σχέση με τις λέξεις του αντίστοιχου αρχικού κειμένου) και αποτυγχάνουν σε περιπτώσεις παράφρασης του περιεχομένου ή δεν λαμβάνουν υπόψη τη σημασιολογία. Επομένως, για μια περισσότερο ποιοτική αξιολόγηση (παρά ποσοτική), χρειάζεται να αναπτυχθούν προσεγγίσεις αυτοματοποιημένης αξιολόγησης προς την κατεύθυνση της ποιοτικής αξιολόγησης που λαμβάνει υπόψη της χαρακτηριστικά σημασιολογίας. Σημειώνουμε ότι η αυτόματη αξιολόγηση των περιλήψεων δύσκολα θα μπορέσει να αντικαταστήσει την αξιολόγηση από άνθρωπο, ωστόσο, λόγω του βασικού πλεονεκτήματος που παρουσιάζει (δηλ., της άμεσης εφαρμογής της) χρειάζεται περαιτέρω διερεύνηση. Προς αυτή την κατεύθυνση, της ποιοτικής αξιολόγησης, η παρούσα εργασία επιχειρεί να παρουσιάσει ένα σύνολο μετρικών για αυτόματη αξιολόγηση των παραγόμενων περιλήψεων, το οποίο προσδιορίζει την συνέπεια απόδοσης πληροφορίας μίας περίληψης σε σχέση με το αρχικό κείμενο.

Στο πλαίσιο της αυτόματης *ΠΚ* ενός εγγράφου με τη μέθοδο της παραγωγής κειμένου, η

διατριβή αυτή εστιάζει στα παραπάνω προβλήματα, των οποίων η αντιμετώπιση αποτελεί πρόκληση για το πεδίο της αυτόματης *ΠΚ*. Προς την κατεύθυνση αντιμετώπισης των προαναφερόμενων προβλημάτων, εξετάζονται αρχιτεκτονικές βαθιάς μάθησης και ταυτόχρονα προτείνεται νέα μεθοδολογία που βασίζεται σε επεξεργασία φυσικής γλώσσας και σημασιολογικές τεχνικές, προκειμένου να βελτιωθεί περαιτέρω τόσο η απόδοση των συστημάτων μηχανικής μάθησης όσο και η ποιότητα των παραγόμενων περιλήψεων. Πιο συγκεκριμένα, οι σημασιολογικές τεχνικές που χρησιμοποιούνται βασίζονται σε πόρους γνώσης, ταξινομίες εννοιών, προσεγγίσεις σημασιολογικής *αποσαφήνισης έννοιας λέξεων (ΑΕΛ) (word sense disambiguation - WSD)*, *αναγνώριση ονοματικών οντοτήτων (ΑΟΟ) (named entity recognition - NER)*, μεθοδολογία σημασιολογικής γενίκευσης του περιεχομένου και τεχνικές σημασιολογικής αναπαράστασης σε μορφή γραφήματος. Οι τεχνικές αυτές αξιοποιούνται προκειμένου να γίνει εφικτός ο κατάλληλος μετασχηματισμός των δεδομένων ή η κατάλληλη σημασιολογική αναπαράσταση του περιεχομένου, με σκοπό τη βελτίωση των προβλέψεων μηχανικής μάθησης και την παραγωγή ενημερωτικών και αναγνώσιμων από τον άνθρωπο περιλήψεων που αποτυπώνουν το περιεχόμενο του αρχικού κειμένου.

1.3 Συνεισφορά της διατριβής

Η παρούσα διατριβή περιλαμβάνει τη μελέτη και την παρουσίαση αρχιτεκτονικών βαθιάς μάθησης για την αυτόματη *ΠΚ* (Κεφάλαιο 3), καθώς και την εισαγωγή νέας μεθοδολογίας, που συνδυάζει και αξιοποιεί βασικές πτυχές από τρεις διαφορετικές κατευθύνσεις έρευνας του πεδίου της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου (Κεφάλαια 4, 5, 6 και 7). Πιο συγκεκριμένα, συνδυάζονται χαρακτηριστικά και πτυχές μεθοδολογιών που βασίζονται στη δομή, στη σημασιολογία και στη μηχανική μάθηση [9], τις οποίες συναντάμε κυρίως ως ξεχωριστές κατευθύνσεις έρευνας στη σχετική βιβλιογραφία (Ενότητες 3.2, 4.2 και 6.2, οι οποίες αναφέρουν τις σχετικές εργασίες). Στο πλαίσιο αυτό, η παρούσα εργασία προσπαθεί να αξιοποιήσει, να συνδυάσει και να ενοποιήσει στοιχεία και βασικές πτυχές των τριών αυτών πεδίων έρευνας, προκειμένου να προτείνει νέες μεθόδους που συνδυάζουν μηχανική μάθηση και σημασιολογικές τεχνικές για την αυτόματη *ΠΚ*.

Η συνεισφορά της διατριβής εντοπίζεται σε τέσσερα κύρια μέρη: (*i*) τη διερεύνηση αρχιτεκτονικών βαθιάς μάθησης για την αυτόματη *ΠΚ*, (*ii*) την πρόταση νέας μεθοδολογίας που συνδυάζει σημασιολογικούς μετασχηματισμούς του περιεχομένου και μηχανική μάθηση για τη βελτίωση της αυτόματης *ΠΚ*, (*iii*) την εισαγωγή ενός νέου πλαισίου που συνδυάζει μεθοδολογία σημασιολογικής αναπαράστασης του περιεχομένου και βαθιά μάθησης, προς την κατεύθυνση της παραγωγής περιλήψεων με σημασιολογική συνάφεια περιεχομένου, και (*iv*) την παρουσίαση ενός συνόλου μετρικών με σκοπό την παροχή ποιοτικής αξιολόγησης του περιεχομένου των παραγόμενων περιλήψεων.

Σύμφωνα με τα παραπάνω, η παρούσα εργασία, αρχικά, εστιάζει στη διερεύνηση αρχιτεκτονικών βαθιάς μάθησης και προτείνει την αξιοποίηση κατάλληλων μοντέλων νευρωνικών δικτύων για την αυτόματη *ΠΚ*. Στην κατεύθυνση αυτή, διερευνώνται διαφορετικές αρχιτεκτονικές μοντέλων μηχανικής μάθησης, όπως είναι τα δίκτυα νευρωνικών δικτύων τύπου κωδικοποιητή-αποκωδικοποιητή, η ενισχυτική μάθηση, οι αρχιτεκτονικές μετασχηματιστών ή τα

προεκπαιδευμένα μοντέλα γλωσσικής αναπαράστασης. Η ερευνητική διαδικασία αποκαλύπτει τις αδυναμίες ή τα οφέλη των διαφορετικών προσεγγίσεων μηχανικής μάθησης και προσδιορίζει τις επιλογές εκείνες που οδηγούν στη βελτιστοποίηση των προβλέψεων μηχανικής μάθησης για την εκτίμηση περιλήψεων.

Στη συνέχεια, η έρευνα εστιάζει στη βελτίωση των προβλέψεων μηχανικής μάθησης μέσω κατάλληλων σημασιολογικών μετασχηματισμών των δεδομένων. Οι σημασιολογικοί μετασχηματισμοί των δεδομένων εντάσσονται σε ένα νέο πλαίσιο, το οποίο εισάγεται στο Κεφάλαιο 4 και δίνει λύσεις στο πρόβλημα εμφάνισης νέων υποψηφίων για περίληψη κειμένων, τα οποία περιλαμβάνουν περιεχόμενο που ενδεχομένως δεν έχει επαρκή παρουσία στο σύνολο εκπαίδευσης ενός μοντέλου μηχανικής μάθησης. Με άλλα λόγια, τα νέα αυτά κείμενα ενδεχομένως να περιλαμβάνουν ΛΕΛ ή σπάνιες λέξεις, όπως αναφέρθηκε στην προηγούμενη ενότητα. Το προτεινόμενο πλαίσιο περιλαμβάνει τρία βασικά στάδια: την προ-επεξεργασία, τις προβλέψεις μηχανικής μάθησης και τη μετα-επεξεργασία. Το στάδιο της προ-επεξεργασίας βασίζεται σε μια καλά καθορισμένη μεθοδολογία γενίκευσης περιεχομένου, η οποία αξιοποιεί πόρους γνώσης, ταξινομίες εννοιών, σημασιολογική αποσαφήνιση έννοιας λέξεων και αναγνώριση ονοματικών οντοτήτων για τον μετασχηματισμό του περιεχομένου σε μια γενικευμένη μορφή, που δίνεται ως είσοδος σε ένα μοντέλο μηχανικής μάθησης. Η εφαρμογή της μεθοδολογίας γενίκευσης του περιεχομένου βελτιώνει την ακρίβεια των προβλέψεων μηχανικής μάθησης, με την παραγωγή περιλήψεων σε μια γενικευμένη μορφή. Το στάδιο της μετα-επεξεργασίας βασίζεται σε ευρεστικές μεθόδους, που αξιοποιούν αντίστοιχους πόρους γνώσης με εκείνους που χρησιμοποιούνται στη φάση της προ-επεξεργασίας, για τον μετασχηματισμό των γενικευμένων περιλήψεων στην τελική τους μορφή. Η εφαρμογή του πλαισίου των σημασιολογικών μετασχηματισμών του περιεχομένου οδηγεί σε βελτίωση των εκτιμώμενων περιλήψεων σε σύγκριση με την περίπτωση των προβλέψεων μηχανικής μάθησης, που εφαρμόζεται χωρίς την εφαρμογή του προτεινόμενου πλαισίου.

Στο τρίτο μέρος, η ερευνητική προσπάθεια επικεντρώνεται στην αξιοποίηση της σημασιολογικής αναπαράστασης του περιεχομένου σε μορφή γραφήματος, σε συνδυασμό με προβλέψεις βαθιάς μάθησης για τη βελτίωση της αυτόματης *ΠΚ*, προς την κατεύθυνση της παραγωγής περιλήψεων με σημασιολογική συνάφεια περιεχομένου. Η κύρια συνεισφορά της προτεινόμενης μεθοδολογίας περιλαμβάνει τη μοντελοποίηση του προβλήματος ως ένα πρόβλημα μάθησης από γράφημα σε περίληψη με μεθόδους βαθιάς μάθησης, την παρουσίαση σημασιολογικών αναπαραστάσεων κειμένου σε μορφή γραφήματος και τη διερεύνηση της επίδοσης διαφορετικών αρχιτεκτονικών βαθιάς μάθησης σε συνδυασμό με διάφορα σχήματα δεδομένων. Η προτεινόμενη προσέγγιση βασίζεται σε ένα καλά καθορισμένο πλαίσιο για την ανάκτηση των σημασιολογικών γραφημάτων για κάθε περίοδο ενός αρχικού κειμένου, την κατασκευή του σημασιολογικού γραφήματος του περιεχομένου ενός κειμένου, τον μετασχηματισμό ενός σημασιολογικού γραφήματος σε κατάλληλη μορφή για είσοδο σε κάποιο μοντέλο μηχανικής μάθησης και τις προβλέψεις μηχανικής μάθησης. Με δεδομένο ότι η αναπαράσταση του αρχικού κειμένου έχει άμεση σχέση με την ποιότητα των παραγόμενων περιλήψεων, η προσέγγιση αυτή οργανώνει τη μη δομημένη πληροφορία και αναπαριστά σημασιολογικά το περιεχόμενο, σε μια προσπάθεια βελτίωσης των προβλέψεων μηχανικής μάθησης και την παροχή περιλήψεων με σημασιολογική συνάφεια περιεχομένου.

Προς την κατεύθυνση της παροχής μιας ποιοτικής αξιολόγησης για την αυτόματη *ΠΚ*, η παρούσα εργασία προτείνει ένα νέο σύνολο μετρικών, οι οποίες προσδιορίζουν τη συνέπεια απόδοσης πληροφορίας των παραγόμενων περιλήψεων σε σχέση με το αρχικό κείμενο. Οι μετρικές

αυτές, οι οποίες παρουσιάζονται στην Ενότητα 7.8.5, συμπληρωματικά με άλλες μετρικές που χρησιμοποιούνται στο πειραματικό μέρος αυτής της εργασίας (Κεφάλαια 3, 5 και 7), μπορούν να συνεισφέρουν στην αξιολόγηση και βελτίωση των συστημάτων αυτόματης *ΠΚ*.

Η έρευνα που διεξάγεται στο πλαίσιο αυτής της διατριβής έχει ως αφετηρία τη μελέτη της σχετικής βιβλιογραφίας με σκοπό την περαιτέρω διερεύνηση και την παρουσίαση νέας μεθοδολογίας για τη βελτίωση της αυτόματης *ΠΚ* ενός εγγράφου με τη μέθοδο της παραγωγής κειμένου. Οι προσεγγίσεις που παρουσιάζονται, καθορίζονται θεωρητικά, υλοποιούνται και διερευνώνται πειραματικά. Στο πλαίσιο της πειραματικής διαδικασίας εξετάζονται σημαντικές πτυχές της προτεινόμενης μεθοδολογίας, καθώς επίσης, διερευνάται και συγκρίνεται η επίδοση της εκάστοτε προσέγγισης με άλλες συναφείς εργασίες. Ο προσδιορισμός των βέλτιστων επιλογών, που οδηγούν στη βελτιστοποίηση της επίδοσης των προτεινόμενων λύσεων, καθώς και τα συμπεράσματα, που προκύπτουν, αναδεικνύουν τα οφέλη της παρούσας ερευνητικής προσπάθειας, η οποία μπορεί να οδηγήσει στην περαιτέρω έρευνα και ανάπτυξη ευφυών συστημάτων στον τομέα της αυτόματης *ΠΚ*.

1.4 Δομή της διατριβής

Το Κεφάλαιο 2, που ακολουθεί, παρουσιάζει το βασικό θεωρητικό υπόβαθρο που χρησιμοποιείται στο ερευνητικό μέρος αυτής της εργασίας, το οποίο περιλαμβάνει δύο κύριες ενότητες που αφορούν τη μηχανική μάθηση και την επεξεργασία φυσικής γλώσσας. Στο κεφάλαιο αυτό επιχειρείται μια εισαγωγή σε βασικές γνώσεις που αποτελούν την αφετηρία για τη διενέργεια της ερευνητικής εργασίας, η οποία παρουσιάζεται με λεπτομέρεια στο υπόλοιπο της διατριβής.

Το Κεφάλαιο 3 εξετάζει αρχιτεκτονικές νευρωνικών δικτύων βαθιάς μάθησης για την αυτόματη περίληψη κειμένου. Οι αρχιτεκτονικές αυτές αναλύονται, αξιολογούνται και συγκρίνονται με τη χρήση τριών δημοφιλών συνόλων δεδομένων. Διερευνώνται οι βασικές πτυχές, καθώς και οι αδυναμίες ή τα πλεονεκτήματα των μοντέλων μηχανικής μάθησης. Ειδικότερα, στο κεφάλαιο αυτό παρουσιάζονται η σχετική βιβλιογραφία, οι αρχιτεκτονικές βαθιάς μάθησης, η πειραματική διαδικασία, τα πειραματικά αποτελέσματα, η περιγραφή και η ερμηνεία των αποτελεσμάτων, καθώς και η εξαγωγή χρήσιμων συμπερασμάτων.

Το Κεφάλαιο 4 παρουσιάζει ένα νέο πλαίσιο για την αυτόματη *ΠΚ* που συνδυάζει μεθοδολογία μηχανικής μάθησης και σημασιολογικών μετασχηματισμών του περιεχομένου. Στο κεφάλαιο αυτό περιλαμβάνονται οι σχετικές εργασίες, η αρχιτεκτονική του προτεινόμενου πλαισίου και οι τρεις διακριτές φάσεις της προτεινόμενης προσέγγισης. Οι φάσεις αυτές είναι η προ-επεξεργασία για τη γενίκευση του περιεχομένου, οι προβλέψεις μηχανικής μάθησης και η μετα-επεξεργασία για τη διαμόρφωση της εκτιμώμενης περίληψης.

Το Κεφάλαιο 5 παρουσιάζει το πειραματικό μέρος για την αξιολόγηση του πλαισίου που συνδυάζει μεθοδολογία μηχανικής μάθησης και σημασιολογικών μετασχηματισμών του περιεχομένου για την αυτόματη *ΠΚ*. Το κεφάλαιο αυτό περιλαμβάνει τα χρησιμοποιούμενα σύνολα δεδομένων, τις μετρικές αξιολόγησης, την πειραματική διαδικασία, τη βελτιστοποίηση των παραμέτρων, τις προσεγγίσεις σύγκρισης, τα πειραματικά αποτελέσματα, την περιγραφή και ερμηνεία των αποτελεσμάτων, τα συμπεράσματα που προκύπτουν, καθώς και τις προοπτικές επέκτασης.

Το κεφάλαιο 6 εισάγει ένα νέο πλαίσιο που αξιοποιεί σημασιολογική αναπαράσταση του

περιεχομένου σε μορφή γραφήματος και μεθοδολογία βαθιάς μάθησης για την αυτόματη *ΠΚ*. Το κεφάλαιο αυτό περιλαμβάνει μια καταγραφή των σχετικών εργασιών, περιγράφει τα σημασιολογικά γραφήματα, εισάγει την αρχιτεκτονική της προτεινόμενης προσέγγισης και αναλύει τις επιμέρους βαθμίδες της μεθοδολογίας. Οι βαθμίδες αυτές περιλαμβάνουν την ανάκτηση, την κατασκευή και τον μετασχηματισμό των σημασιολογικών γραφημάτων, καθώς και τις προβλέψεις μηχανικής μάθησης για την εκτίμηση των περιλήψεων.

Το κεφάλαιο 7 παρουσιάζει το πειραματικό μέρος για την αξιολόγηση του πλαισίου που αξιοποιεί σημασιολογική αναπαράσταση περιεχομένου σε μορφή γραφήματος και μεθοδολογία βαθιάς μάθησης για την αυτόματη *ΠΚ*. Το κεφάλαιο αυτό περιλαμβάνει τα χρησιμοποιούμενα σύνολα δεδομένων, τις μετρικές αξιολόγησης, την πειραματική διαδικασία, τη βελτιστοποίηση των παραμέτρων και τις συναφείς προσεγγίσεις σύγκρισης. Ακολουθούν τα πειραματικά αποτελέσματα, η μελέτη περίπτωσης, σύμφωνα με παραδείγματα χρήσης για την παραγωγή περιλήψεων, και η ερμηνεία των αποτελεσμάτων. Τέλος, το κεφάλαιο αναφέρει τα συμπεράσματα και τις κατευθύνσεις για μελλοντική εργασία.

Η διατριβή ολοκληρώνεται με το Κεφάλαιο 8, το οποίο παρουσιάζει μια σύνοψη και τα γενικά συμπεράσματα που προκύπτουν από το σύνολο της ερευνητικής εργασίας, καθώς και τις προοπτικές που ανοίγονται για περαιτέρω έρευνα στο πεδίο της αυτόματης *ΠΚ*.

Κεφάλαιο 2

Εισαγωγή στη μηχανική μάθηση και στην επεξεργασία φυσικής γλώσσας

2.1 Γενικά

Στο κεφάλαιο αυτό γίνεται μια σύντομη εισαγωγή τόσο στη μηχανική μάθηση όσο και στην επεξεργασία φυσικής γλώσσας. Τα δύο αυτά πεδία αποτελούν τις κύριες περιοχές γνώσης στις οποίες βασίζεται η παρούσα διατριβή. Τα αναφερόμενα σε αυτό το κεφάλαιο, ως βασικό θεωρητικό υπόβαθρο, αποτελούν μια απλή εισαγωγή και συγχρόνως την αφετηρία για το σύνολο της ερευνητικής εργασίας που παρουσιάζεται στη συνέχεια.

Πιο συγκεκριμένα, στην Ενότητα 2.2 γίνεται μια εισαγωγή στη μηχανική μάθηση. Στο πλαίσιο αυτό, αναφέρονται τα είδη της μηχανικής μάθησης (Ενότητα 2.2.1) και περιγράφονται οι γενικές αρχές των τεχνητών νευρωνικών δικτύων (Ενότητα 2.2.2). Ακολουθούν οι αρχές εκπαίδευσης των τεχνικών νευρωνικών δικτύων (Ενότητα 2.2.3) και περιγράφονται τα αναδρομικά νευρωνικά δίκτυα (Ενότητα 2.5). Στην Ενότητα 2.3 παρουσιάζεται το πεδίο της επεξεργασίας φυσικής γλώσσας με μια εισαγωγή στα επίπεδα ανάλυσης της φυσικής γλώσσας (Ενότητα 2.3.1), στη διαδικασία προ-επεξεργασίας δεδομένων κειμένου (Ενότητα 2.3.2), στους πόρους γλωσσικής γνώσης και ειδικότερα στο ηλεκτρονικό λεξικό *WordNet* (Ενότητα 2.3.3), στην αποσαφήνιση της έννοιας των λέξεων (Ενότητα 2.3.4) και στην αναγνώριση ονοματικών οντοτήτων (Ενότητα 2.3.5).

2.2 Μηχανική μάθηση

Η *μηχανική μάθηση (machine learning - ML)* αποτελεί κλάδο της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη συστημάτων τα οποία έχουν τη δυνατότητα να αποκτούν εμπειρία και να βελτιώνουν την απόδοσή τους κατά την εκτέλεση κάποιας διαδικασίας, μέσω της αξιοποίησης προηγούμενης γνώσης. Σε αυτά τα συστήματα αναμένεται ότι η βελτίωση της επίδοσης θα επιτυγχάνεται αυτόματα, μέσα από την εμπειρία που αποκτά το σύστημα, χωρίς να απαιτείται

κάποιος επιπρόσθετος προγραμματισμός. Ένας τυπικός ορισμός που χρησιμοποιείται ευρέως για τη μηχανική μάθηση δόθηκε από τον Mitchell (1997) [12], ο οποίος είναι ο ακόλουθος:

"Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κατηγορία εργασιών T και σε σχέση με ένα μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες της T, όπως αποτιμάται με το μέτρο απόδοσης P, βελτιώνεται μέσω της εμπειρίας E."

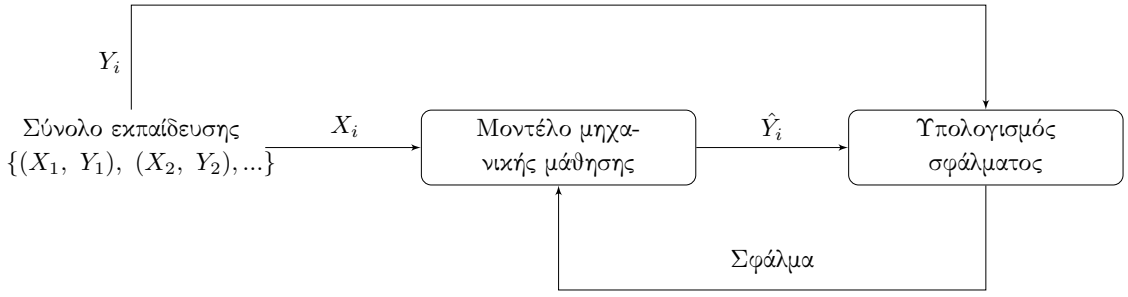
Ο ορισμός αυτός εισάγει τους όρους εμπειρία (E), εργασίες (T) και μέτρο απόδοσης (P) οι οποίοι θέτουν το πλαίσιο σχεδίασης και μελέτης ενός συστήματος μηχανικής μάθησης. Ακολουθούν τα είδη μηχανικής μάθησης και μια εισαγωγή στα νευρωνικά δίκτυα.

2.2.1 Είδη μηχανικής μάθησης

Ακολουθεί η περιγραφή των τριών βασικών ειδών μηχανικής μάθησης που είναι η επιβλεπόμενη μάθηση, η μη-επιβλεπόμενη μάθηση και η ενισχυτική μάθηση. Να αναφερθεί ότι το ερευνητικό μέρος αυτής της εργασίας βασίζεται σε μοντέλα επιβλεπόμενης μάθησης και ενισχυτικής μάθησης.

Επιβλεπόμενη μάθηση

Οι προσεγγίσεις *επιβλεπόμενης μάθησης (supervised learning)* [13] μαθαίνουν σε δεδομένη είσοδο να προβλέπουν την επιθυμητή έξοδο. Κατά τη διαδικασία μάθησης, οι προσεγγίσεις αυτές δέχονται ένα σύνολο από παραδείγματα χρήσης της μορφής $D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_i, Y_i), ..., (X_n, Y_n)\}$, που περιλαμβάνουν τόσο τα δεδομένα εισόδου (X_i) όσο και την επιθυμητή έξοδο Y_i σε αυτά (δηλ., ζεύγη της μορφής (X_i, Y_i)), με σκοπό την εκπαίδευση ενός μοντέλου μηχανικής μάθησης για την πρόβλεψη της εξόδου σε νέα παραδείγματα χρήσης, τα οποία δεν περιλαμβάνουν την τιμή εξόδου. Πιο συγκεκριμένα, τα συστήματα αυτά έχουν στόχο την προσαρμογή των παραμέτρων (ή την εκπαίδευση) μιας συνάρτησης $f : X \to Y$, η οποία, με χρήση δεδομένων εκπαίδευσης, μαθαίνει την αντιστοιχία μεταξύ των τιμών της μεταβλητής εσόδου X (π.χ., το διάνυσμα των χαρακτηριστικών εισόδου για την ταξινόμηση εγγράφων) και εξόδου Y (π.χ., οι κλάσεις ταξινόμησης εγγράφων) και εφαρμόζει αυτή την αντιστοιχία για την εκτίμηση της εξόδου σε νέα δεδομένα εισόδου. Οι μεταβλητές X και Y αποτελούν διανύσματα της μιας ή περισσότερων διαστάσεων, τα οποία αναπαριστούν την είσοδο και την έξοδο των δεδομένων (π.χ., σε προβλήματα ταξινόμησης, συνήθως, η μεταβλητή X αποτελεί διάνυσμα πολλών διαστάσεων που αντιστοιχούν στα χαρακτηριστικά του προβλήματος και η μεταβλητή Y διάνυσμα μιας διάστασης που αντιστοιχεί στην κλάση ή την ετικέτα των εγγράφων).



Σχήμα 2.1: Η διαδικασία επιβλεπόμενης μάθησης.

Με δεδομένο ότι στις περισσότερες εφαρμογές τα ζεύγη εισόδου - εξόδου (X, Y) δεν ακολουθούν μια απόλυτα αιτιοκρατική (ντετερμινιστική) σχέση της μορφής $Y = f(X)$, ως μια προσέγγιση, υποθέτουμε ότι τα δεδομένα ακολουθούν το μοντέλο $Y = f(X) + \epsilon$, όπου ϵ είναι το τυχαίο σφάλμα το οποίο είναι ανεξάρτητο της μεταβλητής X και έχει μέση τιμή $E(\epsilon) = 0$. Το τυχαίο σφάλμα ϵ μπορεί να αντιστοιχεί και σε μεταβλητές που δεν μπορούν να ενταχθούν στο μοντέλο, καθώς είτε είναι άγνωστες είτε είναι δύσκολη η διερεύνησή τους.

Με χρήση των παραδειγμάτων εκπαίδευσης, ένα μοντέλο μηχανικής μάθησης βασίζεται στη σταδιακή προσαρμογή της συνάρτησης f σύμφωνα με την υπολογιζόμενη διαφορά μεταξύ της εκτιμώμενης τιμής $\hat{Y}_i = \hat{f}(X_i)$ και της πραγματικής τιμής της εξόδου Y_i των παραδειγμάτων χρήσης. Η διαδικασία αυτή, η οποία απεικονίζεται στο Σχήμα 2.1, αποτελεί τη διαδικασία επιβλεπόμενης μάθησης που έχει σκοπό να προσεγγίσει όσο περισσότερο γίνεται την πραγματική τιμή της εξόδου με δεδομένη είσοδο, μέσω της προσαρμογής του μοντέλου μηχανικής μάθησης. Κοινά παραδείγματα των αλγόριθμων επιβλεπόμενης μάθησης αποτελούν τα προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression).

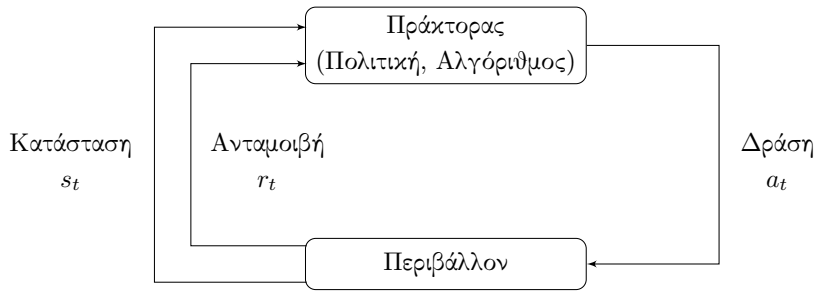
Μη-επιβλεπόμενη μάθηση

Οι προσεγγίσεις μη-επιβλεπόμενης μάθησης (unsupervised learning) [14] βασίζονται σε αλγόριθμους και στατιστικά μοντέλα που έχουν σκοπό τον προσδιορισμό της δομής των δεδομένων εισόδου. Πιο συγκεκριμένα, ας υποθέσουμε ότι ένα σύστημα μη επιβλεπόμενης μάθησης λαμβάνει ως είσοδο μια ακολουθία δεδομένων (x_1, x_2, ..., x_i, ..., x_n), το σύστημα αυτό πρέπει να είναι σε θέση να ανακαλύψει μοτίβα ή δομές στα δεδομένα εισόδου, χωρίς να έχει κάποια ανατροφοδότηση από το περιβάλλον ή χωρίς να γνωρίζει την επιθυμητή έξοδο. Παραδείγματα μη-επιβλεπόμενης μάθησης είναι η ανάλυση συσχετίσεων (association analysis) και η συσταδοποίηση ή ομαδοποίηση (clustering) των δεδομένων. Σε αυτή την εργασία δεν θα ασχοληθούμε περαιτέρω με τη μη-επιβλεπόμενη μάθηση, καθώς το ερευνητικό μέρος βασίζεται σε μοντέλα είτε επιβλεπόμενης είτε ενισχυτικής μάθησης. Επομένως, στη συνέχεια εστιάζουμε κυρίως στην επιβλεπόμενη και στην ενισχυτική μάθηση.

Ενισχυτική μάθηση

Ένα τυπικό μοντέλο ενισχυτικής μάθησης (ΕΜ) (reinforcement learning - RL) [15], το οποίο απεικονίζεται στο Σχήμα 2.2, αποτελείται από έναν πράκτορα, ο οποίος αλληλεπιδρά με το περιβάλλον, με σκοπό να βελτιώσει την απόδοσή του μέσω διαδικασίας μάθησης. Σε κάθε βήμα αλληλεπίδρασης t, ο πράκτορας λαμβάνει ως είσοδο κάποια ένδειξη για την τρέχουσα κατάσταση (state) s_t του περιβάλλοντος και επιλέγει μια δράση (action) a_t, για να προσαρμόσει και να παράγει την έξοδό του. Στη συνέχεια, η επιβολή της δράσης αλλάζει την κατάσταση του περιβάλλοντος και η αλλαγή αυτή κοινοποιείται στον πράκτορα μέσω μιας ανταμοιβής (reward) r_t. Η συμπεριφορά του πράκτορα μεταβάλλεται με σκοπό να επιλέγει δράσεις που τείνουν να αυξήσουν την ανταμοιβή σε αυτόν. Με αυτόν τον τρόπο και μέσω της επανάληψης, το μοντέλο μαθαίνει από τις δράσεις και τα σφάλματα να εκτελεί κάποια διαδικασία.

Στην ΕΜ χρησιμοποιείται μια πολιτική (policy), η οποία καθορίζει τον τρόπο συμπεριφοράς του πράκτορα σε μια δεδομένη στιγμή. Η πολιτική καθορίζει μια αντιστοιχία από τις καταστάσεις



Σχήμα 2.2: Βασικό μοντέλο ενισχυτικής μάθησης.

του περιβάλλοντος σε ενέργειες που πρέπει να γίνουν. Η πολιτική μπορεί να είναι από μια απλή συνάρτηση ή ένας πίνακας αναζήτησης έως μια σύνθετη υπολογιστική προσέγγιση ή κάποια στοχαστική διαδικασία. Η πολιτική αποτελεί βασικό στοιχείο για ένα σύστημα ενισχυτικής μάθησης καθώς καθορίζει τη συμπεριφορά του συστήματος και τον τρόπο που αυτό μαθαίνει.

Σημαντικό ρόλο στην ΕΜ παίζει και η ανταμοιβή, η οποία υπολογίζεται από μια συνάρτηση ανταμοιβής σύμφωνα με την έξοδο του πράκτορα σε δεδομένη είσοδο. Η τιμή της ανταμοιβής χρησιμοποιείται για την προσαρμογή του πράκτορα, καθώς ορίζει τι είναι καλό και τι κακό για τον πράκτορα, σύμφωνα με το πρόβλημα που αντιμετωπίζει. Ο πράκτορας έχει στόχο την μεγιστοποίηση της ανταμοιβής που λαμβάνει, χωρίς ωστόσο να μπορεί να επηρεάσει τον τρόπο υπολογισμού της ανταμοιβής, ο οποίος είναι ανεξάρτητος από τον πράκτορα. Ουσιαστικά η πολιτική που χρησιμοποιείται προσαρμόζει τη συμπεριφορά του πράκτορα με σκοπό τη μεγιστοποίηση της ανταμοιβής, η οποία υποδηλώνει την επιτυχή εκτέλεση της διαδικασίας. Σε αυτή την εργασία χρησιμοποιείται ΕΜ για την αυτόματη περίληψη κειμένου. Το συγκεκριμένο μοντέλο ενισχυτικής μάθησης, οι παραλλαγές και ο τρόπος εφαρμογής του περιγράφονται με λεπτομέρεια στο ερευνητικό μέρος της διατριβής, που ακολουθεί στα επόμενα κεφάλαια.

2.2.2 Τεχνητά νευρωνικά δίκτυα

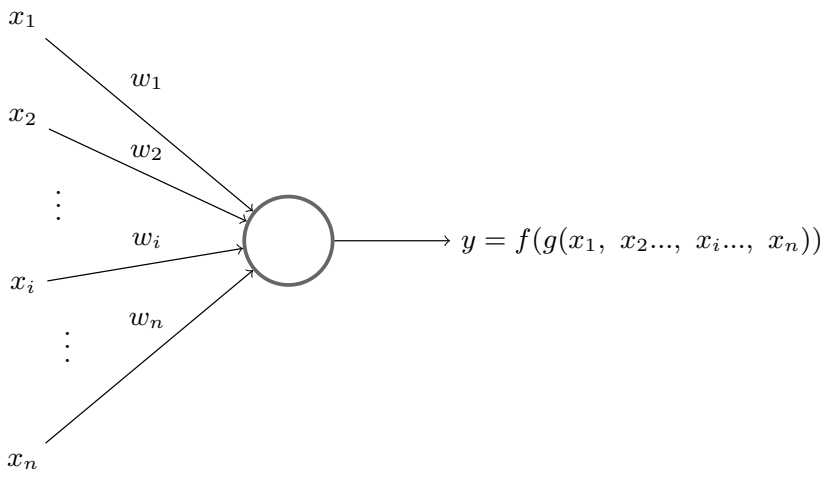
Τα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) (artificial neural networks - ANN) βασίζονται σε υπολογιστικά μοντέλα των οποίων η λειτουργικότητα είναι εμπνευσμένη από αντίστοιχα βιολογικά μοντέλα που αφορούν τη δομή και συμπεριφορά του ανθρώπινου εγκεφάλου. Γενικά, ένα νευρωνικό δίκτυο αποτελείται από ένα σύνολο από διασυνδεδεμένα στοιχεία επεξεργασίας που ονομάζονται νευρώνες, μονάδες ή κόμβοι. Η υπολογιστική ικανότητα του δικτύου βασίζεται κυρίως στις τιμές των βαρών που αποκτούν οι συνδέσεις (συνάψεις) μεταξύ των μονάδων. Οι τιμές των βαρών διαμορφώνονται μέσα από μια διαδικασία προσαρμογής (ή μάθησης) με τη χρήση ενός συνόλου εκπαίδευσης [16].

Τεχνητός νευρώνας

Το Σχήμα 2.3 απεικονίζει έναν τεχνητό νευρώνα [16] που αποτελείται από έναν κόμβο με n εισόδους και μια έξοδο. Κάθε κανάλι ή ακμή εισόδου i μεταφέρει μια τιμή x_i η οποία αντιστοιχεί στην πληροφορία εισόδου και φέρει ένα βάρος w_i. Η έξοδος ενός νευρώνα υπολογίζεται σε δύο φάσεις:

i Αρχικά πραγματοποιείται ένας μετασχηματισμός των εισόδων του νευρώνα μέσω μιας συνάρτησης g και

ii στη συνέχεια, εφαρμόζεται μια συνάρτηση ενεργοποίησης f (activation function) στο αποτέλεσμα της g για τον υπολογισμό της εξόδου y.



Σχήμα 2.3: Ένας τεχνητός νευρώνας με n εισόδους και μία έξοδο y, ο οποίος, για τον υπολογισμό της εξόδου, εκτελεί έναν μετασχηματισμό των εισόδων μέσω μιας συνάρτησης g και εφαρμόζει στο αποτέλεσμα της g μια συνάρτηση ενεργοποίησης f.

Η Εξίσωση 2.1 δίνει την έξοδο y ενός νευρώνα τύπου perceptron μετά από έναν γραμμικό μετασχηματισμό των εισόδων, μέσω της συνάρτησης g, και την εφαρμογή μιας συνάρτησης ενεργοποίησης f.

$$y = f(g(x_1,\ x_2...,\ x_n)) = f(w_1 \cdot x_1 + w_2 \cdot x_2 + ... + w_n \cdot x_n + b) \tag{2.1}$$

όπου τα βάρη $w_i \in R$ και η σταθερά πόλωσης (bias) $b \in R$ αποτελούν τις παραμέτρους του νευρώνα και προσαρμόζονται κατά την διαδικασία εκπαίδευσης.

Μερικές από τις περισσότερο χρησιμοποιούμενες μη γραμμικές συναρτήσεις ενεργοποίησης [17] είναι η σιγμοειδής συνάρτηση (sigmoid function) της Εξίσωσης 2.2,

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

η συνάρτηση υπερβολικής εφαπτομένης (hyperbolic tangent) της Εξίσωσης 2.3,

$$f(x) = tanh(x) = \frac{\mathrm{e^x} - \mathrm{e^{-x}}}{e^x + e^{-x}} \tag{2.3}$$

η συνάρτηση ανορθωμένης γραμμικής μονάδας ReLU (rectified linear unit) της Εξίσωσης 2.4,

$$f(x) = ReLU(x) = max(0,\ x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2.4}$$

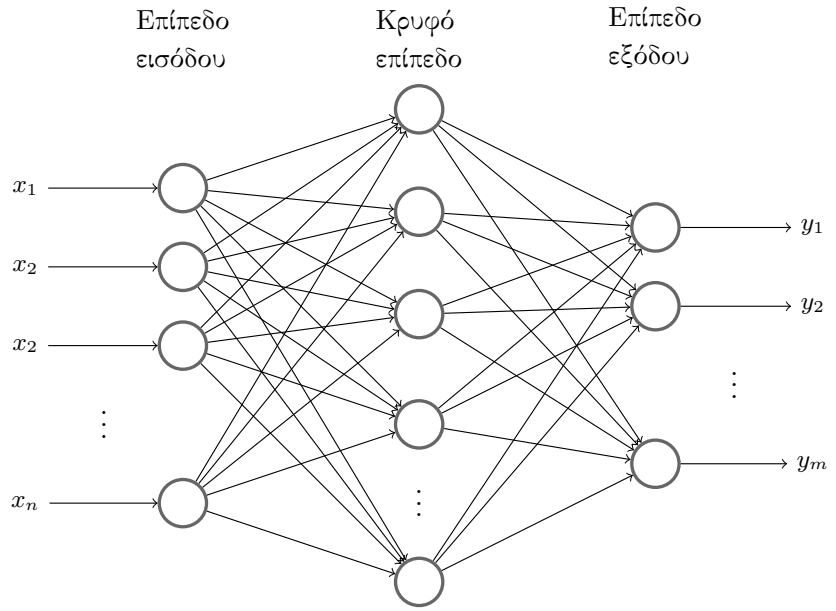
και η συνάρτηση εκθετικής γραμμικής μονάδας ELU (exponential linear unit), με $a > 0$, η οποία περιγράφεται από την Εξίσωση 2.5.

$$f(x) = ELU(x) = \begin{cases} x, & x > 0 \\ a(e^x - 1), & x \leq 0 \end{cases} \tag{2.5}$$

Επίπεδα νευρώνων

Γενικά, μια αρχιτεκτονική νευρωνικών δικτύων αποτελείται από ένα σύνολο εισόδων I, ένα σύνολο υπολογιστικών μονάδων (νευρώνες) N, ένα σύνολο εξόδων O και ένα σύνολο από κατευθυνόμενες ακμές σύνδεσης των μονάδων E [16]. Συνεπώς, η πλειάδα $(I,\ N,\ O,\ E)$ περιγράφει την αρχιτεκτονική ενός νευρωνικού δικτύου. Κάθε υπολογιστική μονάδα του συνόλου N συλλέγει πληροφορία από n εισόδους και, στη συνέχεια, με τη διαδοχική εφαρμογή μιας συνάρτησης μετασχηματισμού των εισόδων $g : R^n \to R$ και μιας συνάρτησης ενεργοποίησης $f : R \to R$ υπολογίζεται η έξοδος της μονάδας. Κάθε κατευθυνόμενη ακμή του συνόλου E περιγράφεται από μια πλειάδα $(u,\ v,\ w)$, όπου $u \in I \cup N$ και $v \in N \cup O$ είναι οι μονάδες που διασυνδέονται από τη συγκεκριμένη ακμή και $w \in R$ το βάρος που φέρει η ακμή. Συνηθίζεται οι είσοδοι I ενός ΤΝΔ να ονομάζονται μονάδες εισόδου χωρίς ωστόσο να εκτελείται κάποια υπολογιστική διαδικασία σε αυτές. Ενώ ως έξοδοι O, τις περισσότερες φορές, θεωρούνται οι υπολογιστικές μονάδες οι οποίες υπολογίζουν το αποτέλεσμα της εξόδου.

Στην περίπτωση που οι υπολογιστικές μονάδες N ενός νευρωνικού δικτύου οργανώνονται σε επίπεδα τότε έχουμε L υποσύνολα υπολογιστικών μονάδων $N_1, N_2, ...N_L$, τα οποία διασυνδέονται με τέτοιο τρόπο μεταξύ τους, ώστε να έχουμε συνδέσεις από τις υπολογιστικές μονάδες του ενός επιπέδου σε εκείνες του επόμενου. Σε πολυεπίπεδες αρχιτεκτονικές ΤΝΔ, έχουμε το επίπεδο εισόδου που αντιστοιχεί στις εισόδους του δικτύου, το επίπεδο εξόδου που αντιστοιχεί στις υπολογιστικές μονάδες εξόδου του δικτύου και ένα ή περισσότερα ενδιάμεσα επίπεδα, τα οποία ονομάζονται κρυφά επίπεδα. Κάθε επίπεδο ενός ΤΝΔ μπορεί να είναι πλήρως ή μερικώς διασυνδεδεμένο με το επόμενο επίπεδο. Για παράδειγμα, το Σχήμα 2.4 απεικονίζει ένα νευρωνικό δίκτυο με τρία επίπεδα: το επίπεδο εσόδου, το κρυφό επίπεδο και το επίπεδο εξόδου.



Σχήμα 2.4: Ένα πλήρως διασυνδεδεμένο πολυεπίπεδο νευρωνικό δίκτυο με n εισόδους και m εξόδους, το οποίο αποτελείται από ένα επίπεδο εισόδου με n κόμβους, ένα κρυφό επίπεδο με k κόμβους και ένα επίπεδο εξόδου με m κόμβους.

Αν θεωρήσουμε μια πολυεπίπεδη αρχιτεκτονική ΤΝΔ τύπου perceptron, τότε η έξοδος του

νευρώνα j του επιπέδου $L+1$ δίνεται από την Εξίσωση 2.6, όπου f είναι μια συνάρτηση ενεργοποίησης, w_i το βάρος της i ακμής εισόδου στον νευρώνα j, y_i^L μια έξοδος από το προηγούμενο επίπεδο L που οδηγείται μέσω της ακμής i στον νευρώνα j και b_j η παράμετρος πόλωσης του νευρώνα j.

$$y_j^{L+1} = f(\sum_i w_i \cdot y_i^L + b_j) \tag{2.6}$$

2.2.3 Εκπαίδευση νευρωνικών δικτύων

Έστω ότι έχουμε ένα ΤΝΔ πρόσθιας τροφοδότησης με n εισόδους και m εξόδους και ένα σύνολο εκπαίδευσης $D_{train} = \{(X_1, Y_1), ..., (X_p, Y_p)\}$, το οποίο περιλαμβάνει ζεύγη διανυσμάτων εισόδου (X_i) και εξόδου (Y_i) διάστασης n και m, αντίστοιχα. Κατά το στάδιο εκπαίδευσης, αν δοθεί ένα διάνυσμα εισόδου X_i στο ΤΝΔ, τότε αυτό παράγει ένα διάνυσμα εξόδου $\hat{Y_i}$ το οποίο, γενικά, διαφέρει από την επιθυμητή έξοδο Y_i. Αν θεωρήσουμε μια δεδομένη συνάρτηση σφάλματος L, η οποία υπολογίζει την απόκλιση μεταξύ μιας εκτιμώμενης εξόδου $\hat{Y_i}$ και της επιθυμητής εξόδου Y_i, τότε ο στόχος της εκπαίδευσης του ΤΝΔ είναι η ελαχιστοποίηση της τιμής της συνάρτησης σφάλματος μέσω της κατάλληλης επιλογής των παραμέτρων $\theta = \{\theta_1, \theta_2, ..., \theta_z\}$ (δηλ., προσαρμογή των βαρών και των παραμέτρων πόλωσης) του ΤΝΔ για ένα δεδομένο σύνολο εκπαίδευσης.

Μετά την ελαχιστοποίηση της συνάρτησης σφάλματος για ένα δεδομένο σύνολο εκπαίδευσης, το ΤΝΔ είναι σε θέση να δέχεται νέα παραδείγματα εισόδου, για τα οποία υπολογίζει μια εκτιμώμενη έξοδο, η οποία ιδανικά προσεγγίζει σε μεγάλο βαθμό την επιθυμητή έξοδο, σύμφωνα πάντα με την συνάφεια που έχει το νέο παράδειγμα με άλλα παραδείγματα χρήσης που χρησιμοποιήθηκαν στη διαδικασία εκπαίδευσης.

Συνάρτηση σφάλματος

Για την αξιολόγηση της εξόδου ενός μοντέλου μηχανικής μάθησης χρησιμοποιείται ένα κριτήριο για τον προσδιορισμό των σφαλμάτων με σκοπό τη διόρθωση ή προσαρμογή των τιμών των παραμέτρων για την επίτευξη όλο και μικρότερων σφαλμάτων κατά τη διαδικασία μάθησης. Η πιο απλή συνάρτηση σφάλματος (loss function ή error function) είναι η συνάρτηση μέσου τετραγωνικού σφάλματος (mean squared error) της Εξίσωσης 2.7.

$$L_{MSE} = \frac{1}{N}\sum_{c=1}^{N}(y_c - \hat{y_c})^2 \tag{2.7}$$

όπου y_c και $\hat{y_c}$ είναι η τιμή της ορθής εξόδου και της εκτιμώμενης εξόδου, αντίστοιχα, για την συνιστώσα c του διανύσματος εξόδου Y ενός μοντέλου μάθησης.

Η συνάρτηση μέσου τετραγωνικού σφάλματος είναι χρήσιμη σε γραμμικά προβλήματα ή προβλήματα που οι έξοδοι παίρνουν συνεχείς τιμές. Σε προβλήματα ταξινόμησης κλάσεων που η έξοδοι παίρνουν διακριτές ή κατηγορικές τιμές, η συνάρτηση η οποία είναι περισσότερο κατάλληλη και χρησιμοποιείται ευρέως είναι η συνάρτηση σφάλματος διασταυρούμενης εντροπίας (cross

entropy loss function) της Εξίσωσης 2.8.

$$L_{CE} = -\sum_{c=1}^{N} P(y_c|X) \cdot log(\hat{P}(y_c|X)) \tag{2.8}$$

όπου $P(y_c|X)$ είναι η πιθανότητα παρουσίας μια ορθής κλάσης y_c σε δεδομένη είσοδο X (π.χ., σε γλωσσικό μοντέλο $P(y_c|X) = 1$ στην περίπτωση παρουσίας της λέξης y_c στην έξοδο) και $\hat{P}(y_c|X)$ είναι η εκτιμώμενη πιθανότητα παρουσίας της κλάσης y_c από το μοντέλο μάθησης σε δεδομένη είσοδο X. Η συνάρτηση σφάλματος διασταυρούμενης εντροπίας είναι γνωστή και ως συνάρτηση αρνητικής λογαριθμικής πιθανοφάνειας (negative log likelihood), καθώς σχετίζεται άμεσα με το μοντέλο εκτίμησης της λογαριθμικής πιθανοφάνειας (η μεγιστοποίηση της λογαριθμικής πιθανοφάνειας αντιστοιχεί σε ελαχιστοποίηση της αρνητικής λογαριθμικής πιθανοφάνειας).

Σε ένα ΤΝΔ ο υπολογισμός της κατανομής των πιθανοτήτων για τις διάφορες κλάσεις του προβλήματος γίνεται με την προσθήκη ενός στρώματος κανονικοποιημένης εκθετικής συνάρτησης (γνωστής ως softmax), το οποίο μετατρέπει τις τιμές εξόδου του ΤΝΔ σε τιμές πιθανοτήτων για κάθε κλάση, με αποτέλεσμα το σύνολο των πιθανοτήτων αυτών να αθροίζεται στη μονάδα. Η Εξίσωση 2.9 παρουσιάζει την κανονικοποιημένη εκθετική συνάρτηση η οποία υπολογίζει την πιθανότητα της κλάσης i.

$$softmax(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{N} e^{y_j}} \tag{2.9}$$

όπου y_i μια είσοδος σε έναν κόμβο του επιπέδου κανονικοποίησης. Το άθροισμα του παρανομαστή υπολογίζεται για τις N εισόδους $y_j, j \in \{1, 2, ..., N\}$ (ή N εξόδους του τελευταίου στρώματος πριν το softmax) στους N κόμβους του στρώματος κανονικοποίησης. Το άθροισμα των εκτιμώμενων πιθανοτήτων των N εξόδων του ΤΝΔ για δεδομένη είσοδο X αθροίζεται στην μονάδα σύμφωνα με την Εξίσωση 2.10.

$$\sum_{j=1}^{N} \hat{P}(y_j|X) = 1 \tag{2.10}$$

όπου $\hat{P}(y_i|X) = softmax(y_i)$. Στο ερευνητικό μέρος αυτής της εργασίας που αφορά την αυτόματη περίληψη κειμένου, η κανονικοποιημένη εκθετική συνάρτηση μπορεί να εφαρμοστεί σε ολόκληρο το λεξιλόγιο για την εκτίμηση της πιθανότητας που έχει κάθε λέξη να παρουσιαστεί στην εκτιμώμενη περίληψη.

Κατάβαση κλίσης

Η τεχνική που χρησιμοποιείται για τον υπολογισμό της μεταβολής και της ενημέρωσης των τιμών των παραμέτρων ενός ΤΝΔ είναι η μεθοδολογία κατάβασης κλίσης (gradient descent) και οι βελτιώσεις ή οι επεκτάσεις της [18], οι οποίες χρησιμοποιούνται για την ελαχιστοποίηση της συνάρτησης σφάλματος ενός ΤΝΔ. Σύμφωνα με αυτή την μεθοδολογία, θεωρούμε ότι το σφάλμα που καταγράφεται σε κάθε επανάληψη μάθησης του δικτύου, μέσω της συνάρτησης σφάλματος L, εξαρτάται από τις τιμές των παραμέτρων του δικτύου $L(\theta) = L(\theta_1, \theta_2, ..., \theta_z)$ (δηλ., τα βάρη w και οι παράμετροι πόλωσης b). Ο υπολογισμός των μερικών παραγώγων της συνάρτησης σφάλματος ως προς τις παραμέτρους του δικτύου, σύμφωνα με την Εξίσωση 2.11, επιτρέπει την μεταβολή της τιμής των παραμέτρων αντίθετα από την υπολογιζόμενη τιμή κλίσης (όπως προκύπτει από την κάθε

μερική παράγωγο) με αποτέλεσμα το σφάλμα να οδηγείται σε τοπικό ή ολικό ελάχιστο (δηλ., η μεταβολή των βαρών οδηγεί σε κάθοδο ή κατάβαση στην καμπύλη της συνάρτησης σφάλματος).

$$\nabla L(\theta) = (\frac{\partial L(\theta)}{\partial \theta_1}, \frac{\partial L(\theta)}{\partial \theta_2}, ..., \frac{\partial L(\theta)}{\partial \theta_z}) \tag{2.11}$$

Γενικά, υπάρχουν τρεις τρόποι εφαρμογής της μεθόδου καθόδου κλίσης, οι οποίοι διαφέρουν στον αριθμό των παραδειγμάτων χρήσης που χρησιμοποιούνται για τον υπολογισμό των κλίσεων και την ενημέρωση των παραμέτρων της συνάρτησης σφάλματος. Οι εναλλακτικές αυτές ακολουθούν.

i Κατάβαση κλίσης στο σύνολο των δεδομένων (batch gradient descent): Η ενημέρωση των παραμέτρων του ΤΝΔ γίνεται με βάση τη συνάρτηση σφάλματος του συνόλου των παραδειγμάτων εκπαίδευσης σύμφωνα με την Εξίσωση 2.12 (δηλ., έχουμε ενημέρωση των βαρών του δικτύου στο τέλος κάθε εποχής εκπαίδευσης).

$$\theta_i = \theta_i - \eta \cdot \frac{\partial L(\theta)}{\partial \theta_i} \tag{2.12}$$

όπου θ_i η παράμετρος i του δικτύου και η ο ρυθμός μάθησης (learning rate). Η μέθοδος αυτή δεν ενδείκνυται για μεγάλα σύνολα δεδομένων καθώς είναι αργή και απαιτεί αρκετή μνήμη. Ωστόσο συγκλίνει πάντα σε τοπικό ή ολικό ελάχιστο της συνάρτησης σφάλματος.

ii Στοχαστική κατάβαση κλίσης (stochastic gradient descent - SGD): Η ενημέρωση των παραμέτρων γίνεται για κάθε παράδειγμα χρήσης (X_j, Y_j), σύμφωνα με την Εξίσωση 2.13. Λόγω των συχνών ενημερώσεων των βαρών του ΤΝΔ, η συνάρτηση σφάλματος παρουσιάζει μεγάλη διακύμανση και αυτό οδηγεί σε ταχύτερη σύγκλιση από την προαναφερόμενη μέθοδο, αλλά και στον απεγκλωβισμό από τοπικά ελάχιστα. Αν ο ρυθμός μάθησης μειώνεται σταδιακά κατά την διαδικασία εκπαίδευσης, η σύγκλιση γίνεται, όπως στην περίπτωση της μαζικής καθόδου κλίσης, με την συνάρτηση σφάλματος να συγκλίνει σε κάποιο τοπικό ή ολικό ελάχιστο.

$$\theta_i = \theta_i - \eta \cdot \frac{\partial L(\theta; X_j, Y_j)}{\partial \theta_i} \tag{2.13}$$

iii Κατάβαση κλίσης σε δέσμη παραδειγμάτων (mini-batch gradient descent): Η ενημέρωση των βαρών γίνεται μετά από κάθε δέσμη παραδειγμάτων (δηλ., η εξίσωση σφάλματος βασίζεται σε ένα υποσύνολο παραδειγμάτων του συνόλου εκπαίδευσης) που αποτελείται από n_b παραδείγματα εκπαίδευσης σύμφωνα με την Εξίσωση 2.14. Αυτή είναι μια μέση λύση σε σύγκριση με τις δύο προαναφερόμενες, η οποία παρουσιάζει μείωση της διακύμανσης της συνάρτησης σφάλματος και οδηγεί σε πιο ομαλή σύγκλιση. Το μέγεθος της δέσμης παραδειγμάτων (batch size) ($|X_j : X_{j+n_b}| = |Y_j : Y_{j+n_b}| = n_b$) διαφέρει ανάλογα με την εφαρμογή χρήσης (π.χ., στο ερευνητικό μέρος αυτής της εργασίας χρησιμοποιούμε μέγεθος δέσμης $n_b = 16$ έως 64, ανάλογα με την περίπτωση).

$$\theta_i = \theta_i - \eta \cdot \frac{\partial L(\theta; X_j : X_{j+n_b}, Y_j : Y_{j+n_b})}{\partial \theta_i} \tag{2.14}$$

Αλγόριθμοι βελτιστοποίησης

Οι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούνται στην εκπαίδευση των ΤΝΔ ουσιαστικά επεκτείνουν την τεχνική της κατάβασης κλίσης που παρουσιάστηκε στην προηγούμενη παράγραφο. Ακολουθούν κάποιοι από τους περισσότερο χρησιμοποιούμενος αλγόριθμους βελτιστοποίησης.

Βελτιστοποίηση ορμής (Momentum) [19]: Ο αλγόριθμος αυτός επεκτείνει την στοχαστική κατάβαση κλίσης με την εισαγωγή ενός όρου ορμής γ. Η ενημέρωση των παραμέτρων θ γίνεται σύμφωνα με τις Εξισώσεις 2.15.

$$\begin{aligned} u_t &= \gamma \cdot u_{t-1} + \eta \cdot \nabla L(\theta) \\ \theta &= \theta - u_t \end{aligned} \tag{2.15}$$

Ουσιαστικά, στον χρόνο εκπαίδευσης t, προστίθεται ένα μέρος $\gamma < 1$ του διανύσματος u_{t-1} του προηγούμενου χρόνου εκπαίδευσης $t-1$ για να ενημερωθεί το τρέχον διάνυσμα u_t, το οποίο με τη σειρά του χρησιμοποιείται για την ενημέρωση των παραμέτρων θ. Η τεχνική αυτή επιταχύνει την κατάβαση (δηλ., μείωση της τιμής) της συνάρτησης σφάλματος (λαμβάνοντας υπόψη την προηγούμενη κατάσταση u_{t-1}), συγκλίνοντας σε τοπικό ή ολικό ελάχιστο, ταχύτερα από την στοχαστική κατάβαση κλίσης. Ουσιαστικά, κατά την κατάβαση κλίσης συσσωρεύεται ορμή, η οποία βοηθά να γίνεται η κατάβαση όλο και με ταχύτερο ρυθμό. Επίσης, η βελτιστοποίηση ορμής αντιμετωπίζει το πρόβλημα που εμφανίζεται στην στοχαστική κατάβαση κλίσης, με την οποία η συνάρτηση σφάλματος μπορεί να κινείται γύρω από ένα τοπικό ή ολικό ελάχιστο με πολύ αργή σύγκλιση. Ο όρος ορμής συνήθως παίρνει τιμές $\gamma \simeq 0.9$.

Βελτιστοποίηση διάδοσης ρίζας μέσων τετραγώνων (Root mean square propagation - RMSProp): Η μέθοδος RMSProp παρουσιάστηκε σε διάλεξη του καθηγητή Geoffrey Hinton του πανεπιστημίου του Τορόντο, χωρίς να δημοσιευτεί σε κάποιο ερευνητικό άρθρο, ωστόσο η μέθοδος αυτή έγινε πολύ δημοφιλής και χρησιμοποιείται ευρέως στην εκπαίδευση μοντέλων ΤΝΔ [20, 18]. Πρόκειται για μέθοδο προσαρμογής του ρυθμού μάθησης η οποία ενημερώνει τις παραμέτρους ενός μοντέλου μάθησης σύμφωνα με τις Εξισώσεις 2.16.

$$\begin{aligned} E[g^2]_t &= \gamma \cdot E[g^2]_{t-1} + (1-\gamma) \cdot g_t^2 \\ \theta_{t+1} &= \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} \cdot g_t \end{aligned} \tag{2.16}$$

όπου $g = \nabla L(\theta)$ είναι το διάνυσμα κλίσεων (μερικών παραγώγων) ως προς τις παραμέτρους της συνάρτησης σφάλματος, $E[g^2]_t$ είναι η μέση τιμή των τετραγώνων των κλίσεων g στον χρόνο εκπαίδευσης t, γ είναι όπως ο αντίστοιχος όρος που χρησιμοποιείται στην βελτιστοποίηση ορμής που περιγράφεται παραπάνω, ϵ είναι ένας όρος εξομάλυνσης για την αποφυγή διαίρεσης με το 0 (π.χ., $\epsilon = 10^{-8}$). Συνιστάται από τους δημιουργούς της μεθόδου [20], η παράμετρος γ να τίθεται ίση με $\gamma = 0.9$ ενώ η τιμή του ρυθμού μάθησης συνήθως τίθεται ίση με $\eta = 10^{-3}$.

Αλγόριθμος εκτίμησης προσαρμοστικών ροπών (Adaptive Moment Estimation - Adam) [21]: Η μέθοδος βελτιστοποίησης Adam είναι επίσης μια μέθοδος προσαρμογής του ρυθμού μάθησης για τις παραμέτρους του δικτύου. Η μέθοδος αυτή διατηρεί τη μέση τιμή κλίσεων και των τετραγώνων των κλίσεων των προηγούμενων βημάτων εκπαίδευσης, παρόμοια με την προαναφερόμενη μέθοδο (RMSProp), υπολογίζοντας τις ροπές πρώτης (first moment) και δεύτερης τάξης (second moment) m_t και u_t, αντίστοιχα, των κλίσεων των παραμέτρων θ ενός μοντέλου μάθησης, σύμφωνα με τις Εξισώσεις 2.17.

$$\begin{aligned} m_t &= \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \\ u_t &= \beta_2 \cdot u_{t-1} + (1-\beta_2) \cdot g_t^2 \end{aligned} \tag{2.17}$$

Επειδή τα διανύσματα m_t και u_t αρχικοποιούνται σε μηδενικές τιμές, αυτό προκαλεί πρόβλημα πόλωσης προς το μηδέν, ιδιαίτερα στα πρώτα βήματα εκτέλεσης και για αυτό το λόγο υπολογίζονται

οι εκτιμήσεις των δύο αυτών διανυσμάτων, σύμφωνα με τις Εξισώσεις 2.18.

$$\begin{aligned} \hat{m_t} &= \frac{m_t}{1-\beta_1^t} \\ \hat{u_t} &= \frac{u_t}{1-\beta_2^t} \end{aligned} \tag{2.18}$$

Στη συνέχεια, ενημερώνονται οι τιμές των παραμέτρων θ στο βήμα εκπαίδευσης $t+1$ σύμφωνα με την Εξίσωση 2.19.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{u_t}} + \epsilon} \cdot \hat{m_t} \tag{2.19}$$

Προτείνεται από τους ερευνητές που παρουσίασαν τον αλγόριθμο Adam, οι προεπιλεγμένες τιμές των παραμέτρων να είναι $\beta_1 = 0.9$, $\beta_2 = 0.999$ και $\epsilon = 10^-8$ [21].

Γενικά, υπάρχουν και άλλοι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούν τη μέθοδο της προσαρμογής του ρυθμού μάθησης όπως οι αλγόριθμοι Adagrad, Adadelta, AdaMax και Nadam [18]. Να σημειωθεί ότι οι αλγόριθμοι RMSProp και Adam που παρουσιάστηκαν παραπάνω λύνουν το πρόβλημα που εμφανίζεται στον Adagrad, ο οποίος οδηγεί σε πολύ μικρές τιμές τον ρυθμό μάθησης, διακόπτοντας ουσιαστικά τη μάθηση. Την αντιμετώπιση του ίδιου προβλήματος επιδιώκει και ο Adadelta που έχει κοινά στοιχεία τόσο με τον RMSProp όσο και με τον Adam. Ο Adam επεκτείνει τον RMSProp και, με τη σειρά τους, οι αλγόριθμοι AdaMax και Nadam αποτελούν επεκτάσεις του Adam. Τέλος, να σημειωθεί ότι οι αλγόριθμοι προσαρμογής του ρυθμού μάθησης, και ιδιαίτερα οι RMSProp και Adam, είναι αυτοί που χρησιμοποιούνται ευρέως σε προβλήματα βαθιάς μάθησης.

Αλγόριθμος οπισθοδιάδοσης σφάλματος

Ο αλγόριθμος οπισθοδιάδοσης σφάλματος (backpropagation) [22] χρησιμοποιείται για τη διόρθωση των τιμών των παραμέτρων θ ενός ΤΝΔ με σκοπό την ελαχιστοποίηση της συνάρτησης σφάλματος. Έστω ότι έχουμε ένα ΤΝΔ πολυεπίπεδης αρχιτεκτονικής, πρόσθιας τροφοδότησης, ένα σύνολο εκπαίδευσης $\{(X_1, Y_1), ..., (X_p, Y_p)\}$, το οποίο περιλαμβάνει ζεύγη διανυσμάτων εισόδου (X_z) και εξόδου (Y_z) και μια συνάρτηση σφάλματος $L(\theta)$. Τότε τα βήματα του αλγόριθμου είναι τα εξής:

(i) Εμπρόσθια τροφοδότηση: Ένα διάνυσμα εισόδου X_i δίνεται στο δίκτυο το οποίο παράγει στην έξοδό του το διάνυσμα $\hat{Y_i}$. Στο βήμα αυτό υπολογίζονται οι έξοδοι και οι παράγωγοι των συναρτήσεων ενεργοποίησης σε κάθε κόμβο του ΤΝΔ. Οι υπολογιζόμενες τιμές διατηρούνται, καθώς απαιτούνται στα επόμενα βήματα για τον υπολογισμό της οπισθοδιάδοσης του σφάλματος.

(ii) Οπισθοδιάδοση στο επίπεδο των κόμβων εξόδου: Έστω ότι κάθε κόμβος i του τελευταίου κρυφού στρώματος (πριν το στρώμα εξόδου) έχει μια έξοδο h_i και συνδέεται με τους κόμβους εξόδου j μέσω ακμών που φέρουν βάρη w_{ij}. Για κάθε έξοδο y_j υπολογίζεται η μερική παράγωγος της συνάρτησης σφάλματος ως προς κάποιο βάρος w_{ij}, που συνδέει ένα κόμβο του κρυφού στρώματος με ένα κόμβο του στρώματος εξόδου σύμφωνα με τον κανόνα της αλυσίδας (Εξίσωση 2.20).

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial f_j} \cdot \frac{\partial f_j}{\partial g_j} \cdot \frac{\partial g_j}{\partial w_{ij}} = \delta_j \cdot h_i \tag{2.20}$$

όπου f_j η συνάρτηση ενεργοποίησης του κόμβου j, g_j η συνάρτηση γραμμικού μετασχηματισμού των εισόδων του κόμβου j (δηλ., $g_j = \sum_i w_{ij} \cdot h_i + b_j$), δ_j το σφάλμα οπισθοδιάδοσης και h_i η έξοδος του κόμβου i (του τελευταίου κρυφού επιπέδου) που συνδέεται με τον κόμβο εξόδου j.

iii Οπισθοδιάδοση στο κρυφό επίπεδο (ή στα κρυφά επίπεδα): Αν θεωρήσουμε ότι κάθε κόμβος i του τελευταίου κρυφού επιπέδου (πριν το επίπεδο εξόδου) συνδέεται με έναν αριθμό κόμβων $j = 1, 2, ..., m$ του επιπέδου εξόδου μέσω ακμών που φέρουν βάρη w_{ij}, τότε υπολογίζεται το σφάλμα οπισθοδιάδοσης δ_i του κόμβου i λαμβάνοντας υπόψη κάθε δυνατή διαδρομή οπισθοδιάδοσης των σφαλμάτων δ_j των κόμβων εξόδου (που υπολογίστηκαν στο προηγούμενο βήμα) προς τον κόμβο αυτό. Η μερική παράγωγος της συνάρτησης σφάλματος L ως προς ένα βάρος w_{qi}, που συνδέει έναν κόμβο q του προηγούμενου επιπέδου με τον κόμβο i του τρέχοντος κρυφού επιπέδου, δίνεται από την Εξίσωση 2.21 (με εφαρμογή του κανόνα της αλυσίδας και λαμβάνοντας υπόψη τα σφάλματα οπισθοδιάδοσης δ_j του προηγούμενου βήματος).

$$\frac{\partial L}{\partial w_{qi}} = \sum_{j=1}^{m} \frac{\partial L}{\partial f_j} \cdot \frac{\partial f_j}{\partial g_j} \cdot \frac{\partial g_j}{\partial f_i} \cdot \frac{\partial f_i}{\partial g_i} \cdot \frac{\partial g_i}{\partial w_{iq}} = \sum_{j=1}^{m} \delta_j \cdot w_{ij} \cdot \frac{\partial f_i}{\partial g_i} \cdot \frac{\partial g_i}{\partial w_{iq}} = \delta_i \cdot h_q \tag{2.21}$$

Όπου f_i, g_i (ή f_j, g_j) οι συναρτήσεις ενεργοποίησης και γραμμικών μετασχηματισμών των κόμβων i ή j, αντίστοιχα, και h_q η έξοδος του κόμβου q του προηγούμενου επιπέδου (πριν το τρέχον κρυφό επίπεδο) που συνδέεται με τον κόμβο εξόδου i.

Η διαδικασία αυτή επαναλαμβάνεται με οπισθοδιάδοση του σφάλματος για όλα τα κρυφά επίπεδα μέχρι το στρώμα εισόδου.

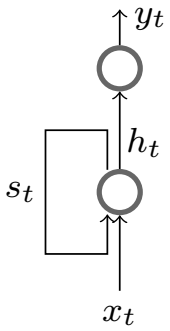
vi Ενημέρωση βαρών: Μετά τον υπολογισμό όλων των μερικών παραγώγων της συνάρτησης σφάλματος για κάθε παράμετρο του δικτύου ακολουθεί η ενημέρωση των βαρών σύμφωνα με τον αλγόριθμο βελτιστοποίησης που χρησιμοποιείται. Στην περίπτωση της κατάβασης κλίσης, σε κάθε βάρος μεταξύ δυο κόμβων i, j προστίθεται η διόρθωση του βάρους Δw_{ij}, που αντιστοιχεί στην αντίθετη τιμή της υπολογιζόμενης κλίσης πολλαπλασιαζόμενη με τον ρυθμό μάθησης η, σύμφωνα με την Εξίσωση 2.22.

$$\Delta w_{ij} = -\eta \cdot \frac{\partial L}{\partial w_{ij}} \tag{2.22}$$

Γενικά, από τα παραπάνω καταβάλλεται προσπάθεια να γίνει κατανοητό ότι ο αλγόριθμος οπισθοδιάδοσης χρησιμοποιείται για τη διάδοση του σφάλματος από τους κόμβους εξόδου προς τους αρχικούς κόμβους, με υπολογισμό των μερικών παραγώγων της συνάρτησης σφάλματος ως προς τις παραμέτρους του δικτύου και, τελικά, την ενημέρωση των παραμέτρων με σκοπό την μείωση του σφάλματος.

2.2.4 Αναδρομικά νευρωνικά δίκτυα

Ένα αναδρομικό νευρωνικό δίκτυο (recurrent neural network - RNN) [23, 24] αποτελεί έναν τύπο ΤΝΔ που έχει τουλάχιστον έναν βρόχο ανατροφοδότησης όπως το δίκτυο του Σχήματος 2.5. Τα δίκτυα αυτά παρουσιάζουν ένα είδος μνήμης, καθώς η έξοδος δεν εξαρτάται μόνο από την



Σχήμα 2.5: Ένα αναδρομικό νευρωνικό δίκτυο με μια μονάδα κρυφού στρώματος που περιλαμβάνει βρόχο ανατροφοδότησης και μια μονάδα εξόδου.

τρέχουσα είσοδο αλλά και από τις προηγούμενες εισόδους, καθώς διαμορφώνεται μια εσωτερική κατάσταση στο δίκτυο που διαδίδεται μέσω των βρόχων ανατροφοδότησης. Στην παρούσα εργασία χρησιμοποιούνται αναδρομικά νευρωνικά δίκτυα, καθώς παρουσιάζουν πλεονεκτήματα σε προβλήματα γλωσσικής μοντελοποίησης, που απαιτούν την πρόβλεψη της επόμενης λέξης, με δεδομένες όλες τις προηγούμενες λέξεις ενός κειμένου. Στο επόμενο κεφάλαιο θα παρουσιάσουμε με λεπτομέρεια την αρχιτεκτονική των αναδρομικών νευρωνικών δικτύων που χρησιμοποιούμε στο ερευνητικό μέρος αυτής της εργασίας.

2.3 Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας (ΕΦΓ) (natural language processing - NLP) αποτελεί ένα εξαιρετικά ενεργό πεδίο έρευνας που εστιάζει τόσο στην ανάπτυξη θεωριών όσο και σε υλοποίηση τεχνολογίας για την ανάλυση, επεξεργασία και κατανόηση της φυσικής γλώσσας [25, 26, 27]. Με τον όρο φυσική γλώσσα εννοούμε τη γλώσσα που χρησιμοποιείται για την καθημερινή επικοινωνία των ανθρώπων. Διευκρινίζεται ότι η παρούσα εργασία εστιάζει στη φυσική γλώσσα σε μορφή κειμένου. Η ΕΦΓ, ως κλάδος της τεχνητής νοημοσύνης και πεδίο έρευνας της υπολογιστικής γλωσσολογίας (η οποία συνδυάζει στοιχεία από τις επιστημονικές περιοχές της γλωσσολογίας και της πληροφορικής), έχει στόχο την επεξεργασία της γλώσσας από τις μηχανές σε τέτοιο βαθμό που να προσεγγίζει την επεξεργασία της γλώσσας από τον άνθρωπο [25]. Με άλλα λόγια, η ΕΦΓ κινείται προς την κατεύθυνση της κατανόησης φυσικής γλώσσας (ΚΦΓ) (natural language understanding - NLU) χωρίς, ωστόσο, να έχει επιτευχθεί πλήρως αυτός ο στόχος και για αυτόν τον λόγο το πεδίο αυτό αναφέρεται ως ΕΦΓ και όχι ως ΚΦΓ. Ένα σύστημα ΚΦΓ θα ήταν σε θέση να εκτελέσει με επιτυχία εργασίες όπως είναι οι ακόλουθες: αναδιατύπωση και παράφραση κειμένου, μετάφραση κειμένου από μια γλώσσα σε μια άλλη, σύνοψη του περιεχομένου ενός κειμένου, απάντηση σε ερωτήσεις σχετικές με το περιεχόμενο ενός κειμένου, παραγωγή φυσικής γλώσσας και εξαγωγή συμπερασμάτων από ένα κείμενο.

Σε εφαρμογές της ΕΦΓ, τα τελευταία χρόνια έχουν γίνει σοβαρές προσπάθειες προς την κατεύθυνση της ΚΦΓ με ανάπτυξη τόσο θεωρίας όσο και υπολογιστικών προσεγγίσεων που διαρκώς εξελίσσονται [28, 29]. Το ερευνητικό πεδίο της ΕΦΓ που εστιάζει η παρούσα εργασία είναι η αυτόματη περίληψη κειμένου, η οποία βασίζεται σε μεθοδολογία παραγωγής φυσικής γλώσσας

για τη σύνθεση μιας περίληψης.

2.3.1 Επίπεδα ανάλυσης φυσικής γλώσσας

Με σκοπό τη διερεύνηση, επεξεργασία και κατανόηση της φυσικής γλώσσας διακρίνονται τα ακόλουθα επίπεδα ανάλυσης της φυσικής γλώσσας [30, 31, 25, 32].

Φωνολογικό επίπεδο. Το επίπεδο αυτό αναφέρεται στον τρόπο που οι ήχοι οργανώνονται, χρησιμοποιούνται ή προφέρονται στις φυσικές γλώσσες. Αυτό το επίπεδο δεν αφορά το γραπτό κείμενο αλλά την κατανόηση της προφορικής γλώσσας (π.χ., ανάπτυξη συστημάτων αναγνώρισης φωνής). Αυτό το επίπεδο δεν θα μας απασχολήσει περαιτέρω καθώς η παρούσα εργασία εστιάζει σε δεδομένα κειμένου.

Μορφολογικό επίπεδο. Αφορά την ανάλυση των μερών σύνθεσης ή της δομής των λέξεων. Η ανάλυση αυτή περιλαμβάνει τη μελέτη των συστατικών στοιχείων των λέξεων, όπως είναι το πρόθεμα, θέμα, επίθεμα ή ρίζα των λέξεων.

Συντακτικό επίπεδο. Αναφέρεται στην ανάλυση της δομής των προτάσεων διερευνώντας πτυχές της ακολουθίας των λέξεων, της έννοιας που προκύπτει από τη σύνταξη και των συσχετίσεων μεταξύ των λέξεων. Το επίπεδο αυτό αφορά τη δομή και τη γραμματική των προτάσεων του κειμένου.

Σημασιολογικό επίπεδο. Αφορά την ανάλυση της έννοιας τον λέξεων και των προτάσεων, καθώς η φυσική γλώσσα δεν έχει μόνο μορφολογία και σύνταξη αλλά και σημασία. Το σημασιολογικό επίπεδο εξετάζει την έννοια των λέξεων σύμφωνα με τα συμφραζόμενα ή το σημασιολογικό πλαίσιο μέσα στο οποίο αυτές βρίσκονται. Ένα σημαντικό πρόβλημα που αντιμετωπίζει η σημασιολογική ανάλυση αποτελεί η λεξιλογική αμφισημία. Δηλαδή, το πρόβλημα των πολλαπλών εννοιών μιας λέξης ή φράσης ενός κειμένου και για την αποσαφήνιση απαιτείται ο προσδιορισμός της συγκεκριμένης έννοιας που αντιπροσωπεύει η εκάστοτε λέξη ενός κειμένου. Στην Ενότητα 2.3.4, η οποία ακολουθεί, αναφέρονται περισσότερες λεπτομέρειες για τη διαδικασία αποσαφήνισης της έννοιας των λέξεων.

Πραγματολογικό επίπεδο. Η ανάλυση σε αυτό το επίπεδο αφορά την κατανόηση της φυσικής γλώσσας, λαμβάνοντας υπόψη τόσο το περιεχόμενο του προς ανάλυση λόγου ή κειμένου όσο και το γενικότερο νοηματικό πλαίσιο των συμφραζομένων. Στην κατεύθυνση αυτή αξιοποιείται, ενδεχομένως, γνώση από το ευρύτερο περιβάλλον ή γενικότερα από τον κόσμο μας. Ο στόχος είναι να διερευνηθεί πώς η γνώση αυτή, η πληροφορία της οποίας δεν αναφέρεται ρητά μέσα στο κείμενο, μπορεί να χρησιμοποιηθεί για την κατανόηση κειμένου (π.χ., τμήματος κειμένου, πρότασης ή φράσης). Προς την κατεύθυνση αυτή, αναπτύσσονται εφαρμογές ΕΦΓ που αξιοποιούν πόρους γνώσης σε συνδυασμό με μεθόδους εξαγωγής συμπερασμάτων.

2.3.2 Προ-επεξεργασία δεδομένων κειμένου

Με δεδομένο ότι το κείμενο συνήθως αποτελεί μη δομημένη μορφή πληροφορίας, τις περισσότερες φορές χρειάζεται να γίνουν κατάλληλοι μετασχηματισμοί για την μετατροπή του σε αναγνώσιμη από τις μηχανές μορφή. Για τον σκοπό αυτό πραγματοποιούνται κάποια στάδια προ-επεξεργασίας του κειμένου, στα οποία τυπικά περιλαμβάνονται οι διαδικασίες που περιγράφονται

παρακάτω.

Διαχωρισμός κειμένου σε λεκτικές μονάδες (tokenization) [33]. Χρησιμοποιείται για την μετατροπή του κειμένου σε μια ακολουθία λεκτικών μονάδων. Οι λεκτικές μονάδες μπορεί να είναι λέξεις, σημεία στίξης, αριθμοί, σύμβολα, ονοματικές οντότητες ή φράσεις. Διαχωρισμός μπορεί να γίνει ακόμη και σε προτάσεις (sentence segmentation) ή παραγράφους, προσδιορίζοντας την ακολουθία των προτάσεων ή παραγράφων που περιέχει το κείμενο.

Αφαίρεση κοινών λέξεων (stop words) [34]. Αυτές είναι λέξεις που εμφανίζονται στο κείμενο με πολύ υψηλή συχνότητα και μπορούν να αγνοηθούν κατά την ΕΦΓ στις περιπτώσεις που θεωρείται ότι δεν περιέχουν κάποια χρήσιμη πληροφορία. Η αφαίρεση των κοινών λέξεων (π.χ., I, the, of, my, it, to, from για την αγγλική γλώσσα) από το κείμενο μειώνει το μέγεθος της εισόδου των αλγόριθμων και μπορεί να αυξήσει τις επιδόσεις, σύμφωνα με την εκάστοτε εφαρμογή ΕΦΓ (π.χ., εξόρυξη πληροφορίας από κείμενα).

Κανονικοποίηση κειμένου. Αφορά την αναγωγή όλων των μορφολογικών τύπων μιας λέξης σε μια ενιαία μορφή (π.χ., οι λέξεις U.S.A και USA θεωρούνται ισοδύναμες). Η κανονικοποίηση του κειμένου μπορεί να αφορά των ορισμό σχέσεων ισοδυναμίας μεταξύ όρων του κειμένου, την μετατροπή των κεφαλαίων γραμμάτων σε πεζά, την εφαρμογή αποκατάληξης ή λημματοποίησης των λέξεων, διαδικασίες που αναφέρονται παρακάτω.

Αποκοπή κατάληξης ή αποκατάληξη (stemming) [35, 36]. Πρόκειται για διαδικασία κανονικοποίησης του κειμένου, η οποία αφαιρεί την κατάληξη των λέξεων (π.χ., οι λέξεις runs, studying, study, better μετά την διαδικασία αποκατάληξης γίνονται run, studi, studi, better). Αυτό είναι ιδιαίτερα χρήσιμο σε εφαρμογές ΕΦΓ (π.χ., εφαρμογές ανάκτησης πληροφορίας, μηχανές αναζήτησης κλπ.).

Λημματοποίηση (lemmatization) [37]. Δηλαδή αναγωγή των λέξεων στη βασική τους μορφή (π.χ. το λήμμα των λέξων was, worse, better είναι be, bad, good, αντίστοιχα). Η διαδικασία λημματοποίησης μετατρέπει τις λέξεις σύμφωνα με το αρχικό τους λήμμα. Δηλαδή, κατά την λημματοποίηση ελέγχεται η γλωσσολογική προέλευση των λέξεων σύμφωνα με ένα λεξικό ώστε να εξαχθεί το ζητούμενο λήμμα. Η διαδικασία λημματοποίησης είναι περισσότερο απαιτητική ή εξελιγμένη από την διαδικασία αποκοπής των καταλήξεων, η οποία θεωρείται περισσότερο επιφανειακή, καθώς ασχολείται μόνο με τη λέξη χωρίς να αντλεί κάποιες γλωσσολογικές πληροφορίες για αυτή.

Επισημείωση μέρους του λόγου (part-of-speech tagging - POS tagging) [38, 39]. Η διαδικασία αυτή αποδίδει μια ετικέτα σε κάθε λέξη σύμφωνα με το μέρος του λόγου στο οποίο ανήκει (π.χ., ουσιαστικό, ρήμα, επίθετο κλπ.). Ως μέρη του λόγου ορίζονται οι κλάσεις στις οποίες ομαδοποιούνται οι λέξεις. Η διαδικασία αυτή βασίζεται σε μορφολογική και συντακτική ανάλυση για την ταξινόμηση των λέξεων στα αντίστοιχα μέρη του λόγου που ανήκουν.

2.3.3 Πόροι γλωσσικής γνώσης

Για τη διερεύνηση, επεξεργασία και κατανόηση της φυσικής γλώσσας, είναι σημαντικό να υπάρχουν διαθέσιμοι πόροι γλωσσικής γνώσης που επιτρέπουν την αξιοποίησή τους σε εφαρμογές ΕΦΓ. Σήμερα, ανάλογα με τις ανάγκες, έχουν αναπτυχθεί πόροι γνώσης που αξιοποιούνται στο πεδίο της ΕΦΓ, όπως επισημειωμένα σώματα κειμένων, ηλεκτρονικά λεξικά, θησαυροί λέξεων,

ταξινομίες εννοιών, οντολογίες πεδίων γνώσης, γραφήματα αναπαράστασης γνώσης και βάσεις γνώσης [40, 41, 42]. Στην συνέχεια, παρουσιάζεται το ηλεκτρονικό λεξικό *WordNet*, το οποίο χρησιμοποιείται στο ερευνητικό μέρος της παρούσας διατριβής.

Το ηλεκτρονικό λεξικό WordNet

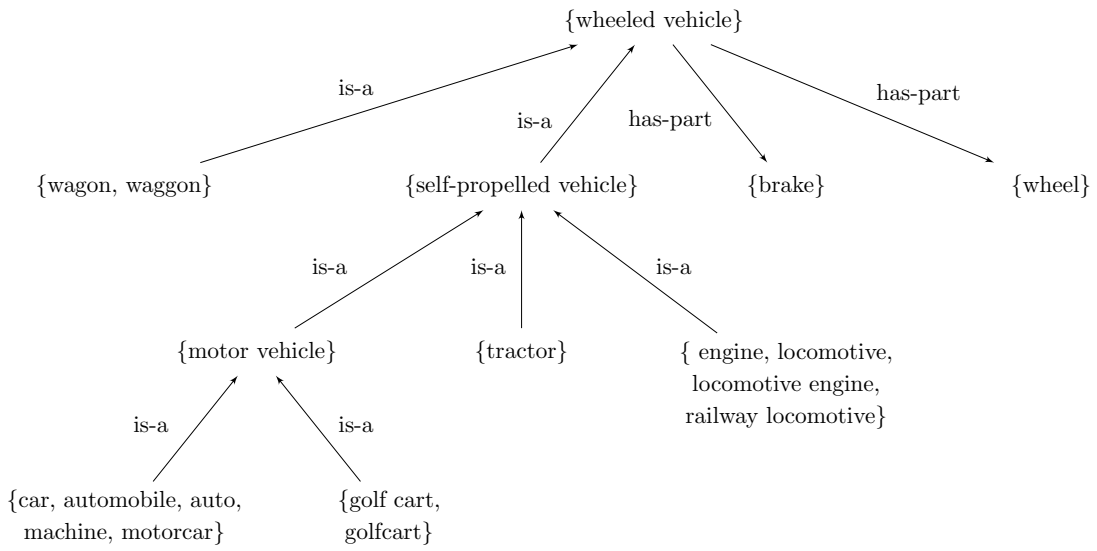
Το *WordNet* [43, 44] αποτελεί ένα ηλεκτρονικό λεξικό της αγγλικής γλώσσας όπου κάθε έννοια οργανώνεται σε ένα σύνολο συνώνυμων λέξεων (synset). Κάθε σύνολο συνώνυμων λέξεων (ή κάθε έννοια) έχει έναν ορισμό (gloss) που περιγράφει την εκάστοτε έννοια. Επίσης, για τις έννοιες δίνονται και παραδείγματα χρήσης. Το *WordNet* δημιουργήθηκε στο πανεπιστήμιο Πρίνσετον [45] και διατηρείται σε ηλεκτρονική μορφή [46]. Η τελευταία του έκδοση (*WordNet* 3.0) περιλαμβάνει περίπου 155.000 λέξεις οργανωμένες σε περίπου 117.000 σύνολα συνώνυμων.

Οι έννοιες στο *WordNet* αναπαρίστανται με μια τριάδα της μορφής *[λέξη].[αύξων αριθμός έννοιας].[μέρος του λόγου]*. Για παράδειγμα, τα αναγνωριστικά εννοιών *play.01.n* και *play.03.n* υποδηλώνουν ουσιαστικά της λέξης *play* με σημασίες *"θεατρικό έργο"* και *"σχέδιο δράσης για ομαδικά αθλήματα"*, αντίστοιχα. Ενώ τα αναγνωριστικά εννοιών *play.03.v* και *play.04.v* αντιστοιχούν σε ρήματα της λέξης *play* με τις σημασίες, *"παίζω μουσικό όργανο"* και *"παίζω έναν ρόλο"*, αντίστοιχα.

Οι έννοιες συνδέονται η μια με την άλλη μέσω σημασιολογικών σχέσεων. Αυτές οι σχέσεις συνδέουν μόνο τις έννοιες λέξεων του ίδιου μέρους του λόγου. Οι σχέσεις μεταξύ εννοιών που αντιστοιχούν σε λέξεις ουσιαστικών περιλαμβάνουν σχέσεις υπερωνυμίας, υπωνυμίας (δηλ., όταν το Α είναι είδος του Β, τότε το Α είναι υπώνυμο του Β και το Β είναι υπερώνυμο του Α), μερωνυμίας και ολωνυμίας (δηλ., όταν το Α είναι μέρος ή τμήμα του Β, τότε το Α είναι μερώνυμο του Β και το Β ολώνυμο του Α). Για παράδειγμα, το *"όχημα"* είναι υπερώνυμο (*hypernym*) του *"αυτοκινήτου"* ή το *"δίκυκλο"* αποτελεί υπώνυμο (*hyponym*) του *"οχήματος"*. Επίσης, η *"μύτη"* είναι μερώνυμο (*meronym*) του *"προσώπου"* ή το *"πρόσωπο"* είναι ολώνυμο (*holonym*) της *"μύτης"*.

Παρόμοιες σχέσεις μεταξύ εννοιών υπάρχουν και στην περίπτωση των ρημάτων, τα οποία οργανώνονται επίσης σε σύνολα συνώνυμων. Στην περίπτωση των ρημάτων έχουμε σχέσεις υπερώνυμων και σχέσεις τροπώνυμων εννοιών. Δηλαδή, αν μια έννοια ρήματος Α αποτελεί ένα υπερώνυμο της έννοιας ρήματος Β τότε και η Β αποτελεί τροπώνυμο της Α. Για παράδειγμα, το ρήμα *"διευθύνω"* αποτελεί ένα υπερώνυμο του ρήματος *"οδηγώ"* και το ρήμα *"οδηγώ"* είναι τροπώνυμο του ρήματος *"διευθύνω"*. Οι σχέσεις τροπώνυμων εννοιών για τα ρήματα είναι ανάλογες με τις σχέσεις υπώνυμων εννοιών για τα ουσιαστικά. Να σημειωθεί ότι στη συνέχεια αυτής της εργασίας, όταν αναφέρουμε σχέσεις υπώνυμων εννοιών και πρόκειται για ρήματα, αυτές οι σχέσεις είναι σχέσεις τροπωνυμίας.

Συνεπώς, η δομή του ηλεκτρονικού λεξικού *WordNet* σχηματίζει μια ταξινομία εννοιών που περιλαμβάνει τις παραπάνω ιεραρχικές σχέσεις μεταξύ των εννοιών. Το Σχήμα 2.6 παρουσιάζει ένα παράδειγμα της ιεραρχίας του *WordNet* με σχέσεις υπερωνυμίας και μερωνυμίας μεταξύ συνόλων συνώνυμων. Υπάρχουν και άλλες σχέσεις μεταξύ των εννοιών στο *WordNet*, ωστόσο αυτές που περιγράφονται παραπάνω, καταλαμβάνουν το μεγαλύτερο μέρος της ταξινομίας εννοιών και είναι αυτές που χρησιμοποιούνται στην παρούσα εργασία.



Σχήμα 2.6: Ένα παράδειγμα των σχέσεων μεταξύ υπερώνυμων (is-a) και μερώνυμων (has-part) εννοιών του *WordNet* που οργανώνονται σε σύνολα συνώνυμων.

2.3.4 Αποσαφήνιση έννοιας λέξεων

Είναι γνωστό ότι η φυσική γλώσσα (η γλώσσα που χρησιμοποιείται για την επικοινωνία μεταξύ των ανθρώπων) περιλαμβάνει πολλές λέξεις που μπορούν να ερμηνευτούν με πολλούς τρόπους, ανάλογα με το σημασιολογικό πλαίσιο στο οποίο εμφανίζονται. Για τον άνθρωπο ο προσδιορισμός της συγκεκριμένης σημασίας μίας λέξης αποτελεί μια απλή διαδικασία καθώς προκύπτει εύκολα, λαμβάνοντας υπόψη τα συμφραζόμενα ή το γενικότερο σημασιολογικό πλαίσιο μέσα στο οποίο αναφέρεται η συγκεκριμένη λέξη. Ωστόσο, για τις μηχανές, η διαδικασία αυτή αποτελεί μια πρόκληση καθώς δεν είναι τόσο απλή. Οι μηχανές χρειάζεται να επεξεργαστούν τη μη-δομημένη πληροφορία κειμένου, η οποία πρέπει να αναλυθεί και σύμφωνα με κάποια ακολουθούμενη μεθοδολογία να προσδιοριστεί η έννοια των λέξεων.

Η υπολογιστική διαδικασία για τον προσδιορισμό της σημασίας των λέξεων ενός κειμένου ονομάζεται *αποσαφήνιση της έννοιας των λέξεων (ΑΕΛ) (word sense disambiguation - WSD)*. Πιο συγκεκριμένα, η ΑΕΛ είναι ο υπολογιστικός προσδιορισμός της έννοιας μιας λέξης η οποία ανήκει σε ένα συγκεκριμένο σημασιολογικό πλαίσιο [28]. Για να ορίσουμε το πρόβλημα, αν έχουμε ένα κείμενο T το οποίο αναπαρίσταται ως μια ακολουθία λέξεων $(w_1,\ w_2, ...,\ w_i,\ ..., w_m)$ ή λεκτικών μονάδων και ένα σύνολο από έννοιες $S = \{s_1,\ s_2, ...,\ s_j, ...,\ s_n\}$, μια διαδικασία ΑΕΛ προσδιορίζει μια αντιστοίχιση A από λέξεις σε έννοιες, έτσι ώστε $A_{w_i} \subseteq S_{w_i}$, όπου A_{w_i} η αντιστοίχιση της λέξης w_i με το υποσύνολο εννοιών S_{w_i}, το οποίο περιλαμβάνει κατάλληλες έννοιες για την λέξη w_i, σύμφωνα με το σημασιολογικό πλαίσιο που ανήκει η λέξη w_i. Η διαδικασία αυτή μπορεί να αντιστοιχίσει περισσότερες από μια έννοιες σε μια λέξη. Ωστόσο, το ζητούμενο του προβλήματος είναι για κάθε λέξη w_i να επιλεγεί η έννοια $s_j \in S$ που ταιριάζει περισσότερο με αυτή τη λέξη και, συνήθως, οι προσεγγίσεις ΑΕΛ επιλέγουν μια έννοια για κάθε λέξη, δηλαδή $|A_{w_i}| = 1$.

Προσεγγίσεις αποσαφήνισης έννοιας λέξεων

Οι προσεγγίσεις ΑΕΛ που έχουν αναπτυχθεί διακρίνονται κυρίως σε τρεις ομάδες [47, 48], οι οποίες αναφέρονται παρακάτω:

i. ΑΕΛ που βασίζεται σε γνώση: Σε αυτές τις προσεγγίσεις χρησιμοποιούνται εξωτερικές πηγές γνώσης που παρέχουν δεδομένα τα οποία είναι απαραίτητα για τη συσχέτιση των εννοιών με τις λέξεις. Οι πόροι γνώσης μπορεί να είναι:

- Θησαυροί λέξεων που παρέχουν πληροφορίες για τη σημασιολογική συσχέτιση μεταξύ των λέξεων.
- Ηλεκτρονικά λεξικά που αποτελούν τους πιο δημοφιλείς πόρους γνώσης για την ΑΕΛ. Ένα από τα περισσότερο χρησιμοποιούμενα ηλεκτρονικά λεξικά είναι το *WordNet* (Ενότητα 2.3.3) το οποίο χρησιμοποιείται στο ερευνητικό μέρος αυτής της εργασίας.
- Οντολογίες που παρέχουν γνώση για συγκεκριμένους τομείς ενδιαφέροντος. Συνήθως παρέχουν μια ταξινομία η οποία περιλαμβάνει τις σχέσεις μεταξύ εννοιών.
- Σώματα κειμένων (corpora) μη δομημένης πληροφορίας στα οποία μπορεί να έχει γίνει επισημείωση εννοιών (sense-annotated corpora) ή όχι (raw corpora).

Ακολουθεί η περιγραφή του αλγόριθμου LESK και της εκτεταμένης έκδοσής του που χρησιμοποιείται στο ερευνητικό μέρος αυτής της εργασίας.

Ο αλγόριθμος LESK: Ένας από τους πρώτους αλγόριθμους αξιοποίησης ηλεκτρονικού λεξικού είναι ο αλγόριθμος LESK [49]. Ο αλγόριθμος αυτός βασίζεται στην επικάλυψη του ορισμών της υποψήφιας προς αποσαφήνιση λέξης (λέξη-στόχος) με τους ορισμούς των άλλων λέξεων, που περιλαμβάνει η εκάστοτε φράση που περιέχει την λέξη-στόχο. Σύμφωνα με τον αλγόριθμο αυτό, αρχικά επιλέγεται μια σύντομη φράση η οποία περιέχει τη λέξη-στόχο. Στη συνέχεια, συλλέγονται με χρήση ενός ηλεκτρονικού λεξικού οι διάφοροι ορισμοί για την λέξη-στόχο και οι ορισμοί των άλλων λέξεων, που περιέχονται στη φράση της λέξης-στόχου. Στο επόμενο βήμα, όλοι οι ορισμοί της λέξης-στόχου συγκρίνονται με τους ορισμούς των άλλων λέξεων της προαναφερόμενης φράσης. Η έννοια, για την οποία προκύπτει η μέγιστη επικάλυψη περιεχομένου μεταξύ των προαναφερόμενων ορισμών, αποτελεί την ζητούμενη έννοια για τη λέξη-στόχος.

Εκτεταμένος αλγόριθμος LESK: Ένας σημαντικός περιορισμός του αλγόριθμου LESK είναι ότι οι ορισμοί των εννοιών ενός ηλεκτρονικού λεξικού είναι συνήθως πολύ σύντομοι (π.χ., οι ορισμοί των όρων στο *WordNet* έχουν μέσο μήκος 7 λέξεις) με αποτέλεσμα να μην λειτουργεί επαρκώς ο αλγόριθμος. Για αυτό τον λόγο ο αλγόριθμος αυτός έχει επεκταθεί ενσωματώνοντας μια μετρική, η οποία αξιοποιεί την ιεραρχική δομή και τις συσχετίσεις μεταξύ των λέξεων που παρέχονται από ένα ηλεκτρονικό λεξικό (π.χ., το *WordNet*), για να προσδιοριστεί πληρέστερα η επικάλυψη μεταξύ των ορισμών μιας έννοιας και των ορισμών των λέξεων που περιέχονται στη φράση που περιλαμβάνει τη λέξη-στόχος. Πιο συγκεκριμένα, η ιδέα για τον εκτεταμένο αλγόριθμο LESK είναι η επέκταση των συγκρίσεων επικάλυψης, οι οποίες δεν λαμβάνουν υπόψη μόνο την επικάλυψη μεταξύ των προαναφερόμενων ορισμών, αλλά επεκτείνονται σε επιπλέον συγκρίσεις. Οι εκτεταμένες αυτές συγκρίσεις επεκτείνονται στον προσδιορισμό της επικάλυψης μεταξύ ορισμών των υπερώνυμων, υπώνυμων, μερώνυμων, ολώνυμων και τροπώνυμων εννοιών των

λέξεων εισόδου (δηλ., μεταξύ της λέξης-στόχου και των άλλων λέξεων της φράσης που περιλαμβάνει τη λέξη-στόχο). Περισσότερες λεπτομέρειες για τον εκτεταμένο αλγόριθμο LESK αναφέρονται στην εργασία [50]. Με δεδομένο ότι το προτεινόμενο πλαίσιο της παρούσας εργασίας που αφορά τους σημασιολογικούς μετασχηματισμούς του περιεχομένου (Κεφάλαιο 4) βασίζεται σε πόρους γνώσης, ο εκτεταμένος αλγόριθμος LESK είναι αυτός που επιλέγεται στο ερευνητικό μέρος αυτής της εργασίας για εφαρμογή ΑΕΛ, καθώς ταιριάζει απόλυτα με τη συνολική προσέγγιση.

ii. ΑΕΛ που βασίζεται σε επιβλεπόμενη μάθηση: Πρόκειται για προσεγγίσεις που εκτελούν ταξινόμηση των λέξεων σε κλάσεις εννοιών αξιοποιώντας τεχνικές μηχανικής μάθησης. Το σύνολο εκπαίδευσης που χρησιμοποιείται για την εκπαίδευση ενός τέτοιου ταξινομητή αποτελείται συνήθως από ένα σύνολο παραδειγμάτων χρήσης, που περιλαμβάνουν κείμενο στο οποίο έχει γίνει επισημείωση της έννοιας κάθε λέξης. Ο ταξινομητής εκπαιδεύεται για να αναθέτει την κατάλληλη έννοια σε κάθε λέξη του κειμένου. Οι προσεγγίσεις που βασίζονται σε επιβλεπόμενη μάθηση δίνουν καλύτερα αποτελέσματα από άλλες. Ωστόσο, το βασικό μειονέκτημα χρήσης τέτοιων μεθόδων είναι ότι χρειάζεται να υπάρχει ένα επαρκές σώμα κειμένου, το οποίο είναι επισημειωμένο με τις έννοιες των λέξεων για την εκπαίδευση των μοντέλων μηχανικής μάθησης. Ειδικά στην περίπτωση των προσεγγίσεων βαθιάς μάθησης [51, 52], οι οποίες αποτελούν τις πλέον σύγχρονες προσεγγίσεις, το επισημειωμένο σώμα κειμένου απαιτείται να περιλαμβάνει μεγάλο πλήθος παραδειγμάτων χρήσης ως σύνολο δεδομένων, ώστε να εκπαιδευτεί επαρκώς το νευρωνικό δίκτυο.

iii. ΑΕΛ που βασίζεται σε μη-επιβλεπόμενη μάθηση: Είναι προσεγγίσεις συσταδοποίησης εννοιών που, γενικά, δεν αναθέτουν κάποια συγκεκριμένη έννοια στις λέξεις, αλλά διακρίνουν ή ομαδοποιούν τις λέξεις με κοινή έννοια [47, 48]. Επίσης, ένα χαρακτηριστικό των μεθόδων αυτών είναι ότι δεν εξαρτώνται από εξωτερικές πηγές γνώσης, όπως ηλεκτρονικά λεξικά ή επισημειωμένα σώματα κειμένου, αλλά χρησιμοποιούν μη-επισημειωμένα σώματα κειμένου ως σύνολα δεδομένων. Οι προσεγγίσεις μη-επιβλεπόμενης μάθησης για την ΑΕΛ δεν είναι κατάλληλες και δεν χρησιμοποιούνται στο ερευνητικό μέρος αυτής της εργασίας, ωστόσο αναφέρονται εδώ για λόγους πληρότητας.

2.3.5 Αναγνώριση ονοματικών οντοτήτων

Η *αναγνώριση ονοματικών οντοτήτων (ΑΟΟ) (named entity recognition - NER)* [53, 54] αφορά το πρόβλημα εντοπισμού και κατηγοριοποίησης ονομάτων ή οντοτήτων ενός κειμένου (π.χ., πρόσωπο, τοποθεσία, οργανισμός, αριθμός, ημερομηνία, ώρα κ.λπ.). Οι ονοματικές οντότητες αφορούν λέξεις ή φράσεις που αντιστοιχούν σε συγκεκριμένες κατηγορίες, με τις λέξεις ή τις φράσεις κάθε κατηγορίας να έχουν κοινές ιδιότητες. Οι ονοματικές οντότητες ενός κειμένου παρέχουν χρήσιμη πληροφορία και για αυτό το λόγο η ΑΟΟ αποτελεί ένα σημαντικό και αρκετά ενεργό πεδίο της ΕΦΓ. Μερικές από τις περιοχές της ΕΦΓ που έχει εφαρμογή η ΑΟΟ είναι η εξαγωγή πληροφορίας, η απάντηση ερωτήσεων, η μηχανική μετάφραση, η αυτόματη περίληψη κειμένου, η συσταδοποίηση κειμένου και η ανάκτηση πληροφορίας.

Για να περιγράψουμε το πρόβλημα, με δεδομένη μια ακολουθία λέξεων $S = (w_1,\ w_2,\ ...\ w_i,\ ...,\ w_n)$ και ένα σύνολο ονοματικών οντοτήτων $E = \{e_1,\ e_2,\ ...e_j, ...\ e_m\}$, η εφαρμογή μιας διαδικασίας ΑΟΟ επιστρέφει μια λίστα από πλειάδες της μορφής $(ind_{start},\ ind_{end},\ e_j)$, όπου $ind_{start} \in [1,\ n]$ και $ind_{end} \in [ind_{start},\ n]$ αντιπροσωπεύουν τους

δείκτες στην αρχική και τελική λέξη της S που προσδιορίζουν τη λέξη ή τη φράση της S στην οποία ανατίθεται η ονοματική οντότητα $e_j \in E$.

Οι προσεγγίσεις ΑΟΟ χωρίζονται κυρίως σε συστήματα που βασίζονται σε κανόνες και σε συστήματα μηχανικής μάθησης [55]. Οι προσεγγίσεις που βασίζονται σε κανόνες χρησιμοποιούν βάσεις γνώσης, καθώς και κανόνες προτύπων για τον εντοπισμό και την ταξινόμηση ονοματικών οντοτήτων. Οι προσεγγίσεις μηχανικής μάθησης διακρίνονται σε συστήματα επιβλεπόμενης ή μη επιβλεπόμενης μάθησης. Τα συστήματα επιβλεπόμενης μάθησης, στη διαδικασία εκπαίδευσης, χρησιμοποιούν δεδομένα στα οποία είναι επισημειωμένη η κατηγορία των ονοματικών οντοτήτων. Μετά τη διαδικασία της εκπαίδευσης, τα συστήματα επιβλεπόμενης μάθησης αναμένεται να εντοπίζουν τις ονοματικές οντότητες σε νέα παραδείγματα κειμένου και να τις κατηγοριοποιούν στην κατηγορία που ανήκουν. Ενώ τα συστήματα μη επιβλεπόμενης μάθησης είναι συνήθως προσεγγίσεις συσταδοποίησης, οι οποίες βασίζονται σε στατιστικές κατανομές και αξιοποιούν μετρήσεις ομοιότητας κειμένου για τον εντοπισμό και την ομαδοποίηση ονοματικών οντοτήτων, χωρίς να απαιτείται επισημειωμένο σύνολο δεδομένων. Υπάρχουν και προσεγγίσεις που συνδυάζουν μεθοδολογία που βασίζεται σε κανόνες και σε τεχνικές μηχανικής μάθησης για την ΑΟΟ. Τα τελευταία χρόνια, οι προσεγγίσεις βαθιάς μάθησης έχουν κυριαρχήσει στον χώρο αυτό και φαίνεται να παρουσιάζουν αξιόλογες επιδόσεις [56, 57].

Κεφάλαιο 3

Μοντέλα βαθιάς μάθησης για την αυτόματη περίληψη κειμένου

3.1 Γενικά

Με δεδομένο ένα κείμενο εισόδου, η εκτίμηση μιας περίληψης μπορεί να επιτευχθεί με χρήση νευρωνικών δικτύων βαθιάς μάθησης [23], τα οποία αποτελούν την κυρίαρχη προσέγγιση στο πεδίο της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου (*abstractive text summarization*). Πιο συγκεκριμένα, ένα μοντέλο μηχανικής μάθησης εκτιμά μια ακολουθία λέκτικών μονάδων $\hat{Y} = (\hat{y}_1,\ \hat{y}_2, ..., \hat{y}_m)$ (την εκτιμώμενη περίληψη), με δεδομένη μια ακολουθία λεκτικών μονάδων εισόδου $X = (x_1,\ x_2, ..., x_n)$ (το αρχικό κείμενο).

Στο πλαίσιο αυτής της εργασίας, ακολουθεί μια ανασκόπηση της σχετικής βιβλιογραφίας και, στη συνέχεια, εξετάζονται πέντε μοντέλα βαθιάς μάθησης για την αυτόματη *ΠΚ*, τα οποία είναι τα εξής: (i) μοντέλο κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής, (ii) μοντέλο με μηχανισμό αντιγραφής λέξεων εκτός λεξιλογίου, (iii) μοντέλο ενισχυτικής μάθησης, (iv) μοντέλο μετασχηματιστών και (v) μοντέλο μετασχηματιστών με προ-εκπαιδευμένο κωδικοποιητή. Τα πέντε αυτά μοντέλα νευρωνικών δικτύων υλοποιούνται και αξιολογείται η επίδοσή τους με χρήση τριών δημοφιλών συνόλων δεδομένων, όπως αναφέρεται με λεπτομέρεια στις παρακάτω ενότητες.

3.2 Σχετικές εργασίες

Οι προσεγγίσεις για την αυτόματη *ΠΚ* που βασίζονται σε αρχιτεκτονικές νευρωνικών δικτύων είναι ικανές να παράγουν περιλήψεις με τη χρήση ενός κατάλληλου μοντέλου αρχιτεκτονικής νευρωνικών δικτύων, χωρίς άλλη πολύπλοκη επεξεργασία φυσικής γλώσσας. Τέτοιες τεχνικές βασίζονται συχνά σε μοντέλα αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή [58], τα οποία είναι μοντέλα νευρωνικών δικτύων βαθιάς μάθησης, όπου ο κωδικοποιητής λαμβάνει μια ακολουθία λέξεων και ο αποκωδικοποιητής αποδίδει μια ακολουθία λέξεων που αποτελεί την περίληψη. Τα εν λόγω δίκτυα εκπαιδεύονται από άκρη σε άκρη (*end-to-end*) με χρήση ενός συνόλου δεδομένων που περιλαμβάνει ένα μεγάλο πλήθος παραδειγμάτων χρήσης από ζεύγη κειμένου-περίληψης. Τα

δίκτυα αυτά μαθαίνουν να εκτιμούν τις περιλήψεις των κειμένων που δίνονται ως είσοδος. Τα μοντέλα βαθιάς μάθησης βασίζονται κυρίως σε αναδρομικά νευρωνικά δίκτυα και, πρωτίστως, σε μονάδες μακράς και βραχείας μνήμης (*long short term memory - LSTM*), σε αναδρομικά δίκτυα με πύλη (*gated recurrent units - GRU*) και σε αρχιτεκτονικές ενισχυτικής μάθησης. Τελευταία κερδίζουν έδαφος οι αρχιτεκτονικές που βασίζονται σε μετασχηματιστές (*tranformers*) οι οποίοι δεν χρησιμοποιούν αναδρομή. Τα προαναφερόμενα δίκτυα επιτυγχάνουν τις καλύτερες επιδόσεις στην αυτόματη *ΠΚ* με τη μέθοδο της παραγωγής κειμένου σύμφωνα με τις σχετικές εργασίες [59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69]

Το βασικό μοντέλο κωδικοποιητή-αποκωδικοποιητή μπορεί να βελτιωθεί περαιτέρω μέσω της εισαγωγής ενός *μηχανισμού προσοχής* [70], καθιστώντας έτσι τα μοντέλα αυτά μια τυπική αρχιτεκτονική για εφαρμογές ακολουθία-σε-ακολουθία (*sequence-to-sequence*) (δηλ., εκτίμηση μιας ακολουθίας εξόδου σε δεδομένη ακολουθία εισόδου) [71] και ειδικά για την περίπτωση της αυτόματης *ΠΚ*, που το δίκτυο δέχεται μια ακολουθία λέξεων εισόδου και αποδίδει μια ακολουθία λέξεων στην έξοδο ως περίληψη [60, 61, 72, 64]. Ο μηχανισμός προσοχής επικεντρώνεται σε σημαντικές λέξεις του κειμένου εισόδου. Δεδομένης της ακολουθίας εισόδου του κωδικοποιητή, ο αποκωδικοποιητής τροφοδοτείται με ένα διάνυσμα που περιλαμβάνει τη σημασία της κάθε λέξης (δηλ., μια κατανομή πιθανοτήτων ή βαρών για τις λέξεις του κειμένου εισόδου) με αποτέλεσμα το μοντέλο να εστιάζει στις περισσότερο σημαντικές λέξεις.

Επιπλέον, τα μοντέλα κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής μπορούν να επεκταθούν για την αντιμετώπιση του προβλήματος των λέξεων εκτός λεξιλογίου (*ΛΕΛ*) μέσω της προσθήκης ενός δικτύου αντιγραφής άγνωστων λέξεων *(pointer-generator network)* [60, 61]. Η εργασία [61] προτείνει έναν μηχανισμό αποφυγής της επανάληψης των ίδιων λέξεων στην εκτιμώμενη περίληψη. Στην ίδια κατεύθυνση, στην εργασία [73] παρουσιάζεται ένα μοντέλο κωδικοποίησης για την επίλυση του προβλήματος της επανάληψης. Επιπλέον, η εργασία [62] προτείνει ένα μοντέλο εξαγωγής φράσεων από το κείμενο για τη δημιουργία περιλήψεων, το οποίο βασίζεται σε μονάδες *LSTM* και συνελικτικά νευρωνικά δίκτυα (*convolutional neural networks - CNN*). Στην παρούσα εργασία εξετάζεται τόσο ένα βασικό μοντέλο κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής όσο και ένα μοντέλο αντιγραφής *ΛΕΛ*, το οποίο περιλαμβάνει και μηχανισμό αποφυγής του προβλήματος της επανάληψης των λέξεων στην έξοδο. Τα μοντέλα αυτά αναλύονται με λεπτομέρεια στη συνέχεια του κεφαλαίου.

Η εξέλιξη στην έρευνα στο πεδίο των *ΤΝΔ* οδήγησε στη δημιουργία ενός μοντέλου αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή που βασίζεται σε δίκτυο μετασχηματιστών (*transformers*), το οποίο χρησιμοποιεί μηχανισμό ενδο-προσοχής για την επιλογή και διάκριση σημαντικών πτυχών ή πληροφοριών για την δημιουργία των περιλήψεων [67]. Τα δίκτυα μετασχηματιστών δεν περιλαμβάνουν αναδρομή ή συνέλιξη, αλλά βασίζονται κυρίως στον μηχανισμό προσοχής. Να σημειωθεί ότι τα μοντέλα μετασχηματιστών και οι παραλλαγές τους που βασίζονται σε προ-εκπαιδευμένα μοντέλα μετασχηματιστών αποτελούν σήμερα τις κυρίαρχες προσεγγίσεις στην αυτόματη *ΠΚ* [65, 74, 66, 67, 68, 75, 69]. Οι μετασχηματιστές αποτελούν απλά μοντέλα που ενσωματώνουν κατανεμημένο μηχανισμό προσοχής προκειμένου να μειώσουν το υπολογιστικό φορτίο, παραλληλίζοντας τους υπολογισμούς στη διαδικασία εκπαίδευσης. Στην παρούσα εργασία εξετάζονται: ένα μοντέλο μετασχηματιστών και ένα μοντέλων μετασχηματιστών προ-εκπαιδευμένου κωδικοποιητή τα οποία περιγράφονται με λεπτομέρεια στη συνέχεια.

Η κύρια αδυναμία των παραπάνω αρχιτεκτονικών, κωδικοποιητή-αποκωδικοποιητή, είναι ότι

ελαχιστοποιούν μια συνάρτηση σφάλματος μέγιστης πιθανότητας, ενώ η αξιολόγηση της αυτόματης *ΠΚ* βασίζεται σε κάποια διαφορετική μετρική (π.χ., *Rouge*). Για την αντιμετώπιση αυτού του προβλήματος, κάποιες πρόσφατες προσεγγίσεις χρησιμοποιούν μοντέλα ενισχυτικής μάθησης (*EM*) (Ενότητα 2.2.1) [76, 77], για να μεγιστοποιήσουν ένα μέγεθος ανταμοιβής που υπολογίζεται σύμφωνα με μια χρησιμοποιούμενη και συγκεκριμένη μετρική αξιολόγησης. Συγκεκριμένα, στην εργασία [78] προτείνεται ένα μοντέλο *EM* που βελτιστοποιεί μια συγκεκριμένη μετρική αξιολόγησης (π.χ., σε όρους μετρικής $Rouge_L$) και ένα συνδυαστικό μοντέλο, το οποίο συνδυάζει τη βελτιστοποίηση μιας συνάρτησης σφάλματος και μιας συγκεκριμένης μετρικής αξιολόγησης. Στην εργασία [79] παρουσιάζεται ένα μοντέλο ΕΜ για *ΠΚ* σε επίπεδο εγγράφου που χρησιμοποιεί πολλούς συνεργαζόμενους πράκτορες για την κωδικοποίηση διαφορετικών τμημάτων του κειμένου (π.χ., προτάσεις ή παραγράφους). Η αποτελεσματική επικοινωνία μεταξύ των πρακτόρων βελτιώνει τις επιδόσεις. Στην εργασία [80] παρουσιάζεται μια προσέγγιση πολλαπλών ανταμοιβών με σκοπό τη δημιουργία περιλήψεων που περιέχουν σημαντικές πληροφορίες. Στο ερευνητικό μέρος της παρούσας εργασίας εξετάζεται ένα μοντέλο ΕΜ το οποίο περιγράφεται με λεπτομέρεια παρακάτω.

3.2.1 Διανυσματική αναπαράσταση λέξεων

Το πεδίο της μηχανικής μάθησης περιλαμβάνει την ανάπτυξη μοντέλων για τη διανυσματική αναπαράσταση των λεκτικών μονάδων ενός κειμένου που αποτελεί την τυπική μορφή αναπαράστασης. Οι τεχνικές διανυσματικής αναπαράστασης λέξεων, που βασίζονται σε επεξεργασία φυσικής γλώσσας με χρήση νευρωνικών δικτύων, αντιστοιχούν κάθε λέξη σε ένα διάνυσμα πραγματικών αριθμών D διαστάσεων. Αυτά τα διανύσματα είναι γνωστά ως ενσωματώσεις λέξεων (*word embeddings*) [81, 82, 83, 84]. Ένα από τα κύρια πλεονεκτήματα των διανυσμάτων ενσωμάτωσης λέξεων είναι ότι οι όροι διατηρούν τη σημασιολογική σχέση μεταξύ τους, με την έννοια ότι συναφείς λέξεις αναπαρίστανται σε κοντινές αποστάσεις στον διανυσματικό χώρο. Οι προσεγγίσεις διανυσματικής αναπαράστασης των λέξεων που βασίζονται στη μηχανική μάθηση διακρίνονται σε δύο κύριες κατηγορίες, στις μεθόδους που η διανυσματική αναπαράσταση των λεκτικών μονάδων είναι ανεξάρτητη των συμφραζομένων [81, 85] και σε εκείνες που εξαρτάται από τα συμφραζόμενα [86].

Ένα κύριο πλεονέκτημα που παρουσιάζουν οι ενσωματώσεις λέξεων που είναι ανεξάρτητες από τα συμφραζόμενα [81, 85] είναι ότι τα διανύσματα διατηρούν τη σημασιολογική σχέση μεταξύ των λέξεων, όπως έχει αναφερθεί ήδη παραπάνω. Ωστόσο, ένα βασικό μειονέκτημα αυτής της αναπαράστασης είναι ότι κάθε λέξη αναπαρίσταται από ένα μοναδικό διάνυσμα το οποίο δεν λαμβάνει υπόψη του τις διαφορετικές σημασίες της ίδιας λέξης στο κείμενο (δηλ., το ίδιο διάνυσμα ανατίθεται σε μια λέξη που μπορεί να έχει περισσότερες από μια έννοιες, ανεξάρτητα από την εκάστοτε έννοια). Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί από προσεγγίσεις που παράγουν διανυσματική αναπαράσταση λέξεων εξαρτώμενη από τα συμφραζόμενα.

Έχουν αναπτυχθεί τεχνικές διανυσματική αναπαράστασης λέξεων που εξαρτώνται από τα συμφραζόμενα, οι οποίες αναθέτουν σε κάθε λέξη ένα διάνυσμα, λαμβάνοντας υπόψη τα συμφραζόμενα της λέξης αυτής [86]. Επομένως, οι ενσωματώσεις λέξεων που βασίζονται στα συμφραζόμενα καταγράφουν συντακτικές και σημασιολογικές ιδιότητες των λέξεων (ή των επιμέρους τμημάτων που συνθέτουν μια λέξη), σύμφωνα με την ακολουθία των λέξεων του κειμένου που περιλαμβάνει τη λέξη-στόχο. Η αναπαράσταση αυτή βασίζεται κυρίως σε προ-εκπαιδευμένα

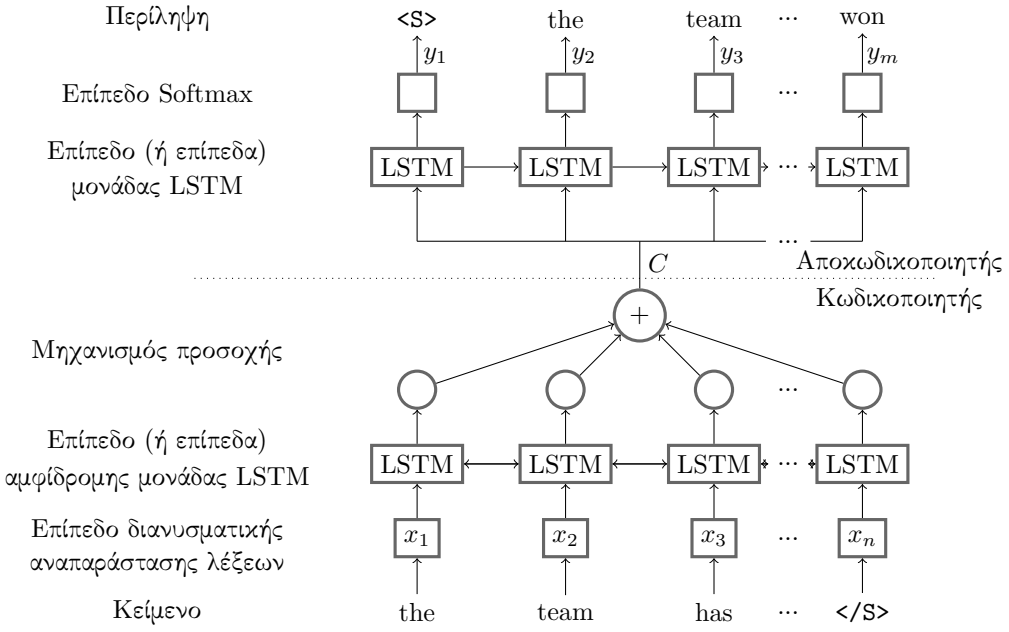
μοντέλα, τα οποία έχουν εκπαιδευτεί σε σώματα κειμένων μεγάλης κλίμακας. Να αναφερθεί ότι σε εφαρμογές του τομέα της επεξεργασίας φυσικής γλώσσας (π.χ., ανάλυση συναισθήματος, απάντηση ερωτήσεων, μηχανική μετάφραση, *ΠΚ* κλπ.) αξιοποιούνται ενσωματώσεις λέξεων που βασίζονται στα συμφραζόμενα, οι οποίες επιτυγχάνουν επιδόσεις αιχμής [87].

Τα μοντέλα μηχανικής μάθησης, που θα εξετάσουμε στη συνέχεια, χρησιμοποιούν ενσωματώσεις λέξεων και των δύο προαναφερόμενων κατηγοριών οι οποίες αναπαριστούν την ακολουθία των λέξεων εισόδου. Τα συγκεκριμένα μοντέλα ενσωμάτωσης λέξεων, που χρησιμοποιούνται, και η διαδικασία εκπαίδευσης των μοντέλων αυτών για την παραγωγή των διανυσματικών αναπαραστάσεων των λέξεων περιγράφονται στην Ενότητα 3.4.3. Οι αρχιτεκτονικές βαθιάς μάθησης που εξετάζουμε για την αυτόματη *ΠΚ*, οι οποίες αξιοποιούν τις ενσωματώσεις λέξεων, παρουσιάζονται με λεπτομέρεια στη συνέχεια.

3.3 Αρχιτεκτονικές βαθιάς μάθησης

3.3.1 Μοντέλο κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής

Το Σχήμα 3.1 απεικονίζει ένα νευρωνικό δίκτυο βαθιάς μάθησης, το οποίο είναι ένα μοντέλο αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής (ΚΑΜΠ) [61]. Ο κωδικοποιητής δέχεται τις διανυσματικές αναπαραστάσεις των λέξεων του αρχικού κειμένου και ο αποκωδικοποιητής μαθαίνει να προβλέπει την αντίστοιχη περίληψη. Η αρχιτεκτονική και τα στοιχεία του δικτύου περιγράφονται με λεπτομέρεια παρακάτω.



Σχήμα 3.1: Μοντέλο βαθιάς μάθησης αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής.

Επίπεδο διανυσματικής αναπαράστασης λέξεων: Το επίπεδο διανυσματικής

αναπαράστασης λέξεων λαμβάνει μια λέξη από το αρχικό κείμενο, την μετατρέπει στο αντίστοιχο διάνυσμα και στη συνέχεια την προωθεί στο επόμενο επίπεδο του κωδικοποιητή. Η αρχικοποίηση των διανυσμάτων αναπαράστασης των λέξεων γίνεται με δύο τρόπους: (i) είτε αρχικοποιούνται τυχαία σε έναν διανυσματικό χώρο και, στη συνέχεια, κατά τη διάρκεια εκπαίδευσης του συνολικού δικτύου προσαρμόζουν τα βάρη τους είτε (ii) χρησιμοποιούνται προ-εκπαιδευμένα διανύσματα αναπαράστασης λέξεων (π.χ., έχει προηγηθεί εκπαίδευση *word2vec* διανυσμάτων που αναπαριστούν το δεδομένο λεξιλόγιο). Στη δεύτερη περίπτωση, τα προ-εκπαιδευμένα διανύσματα των λέξεων μπορεί να παραμείνουν σταθερά σε όλη τη διάρκεια εκπαίδευσης του μοντέλου ή να συνεχίσει η εκπαίδευσή τους και η προσαρμογή τους κατά τη διάρκεια της εκπαίδευσης του συνολικού δικτύου.

Επίπεδο αμφίδρομης μονάδας *LSTM* (κωδικοποιητή): Το δεύτερο επίπεδο του κωδικοποιητή αποτελείται από αμφίδρομες μονάδες *LSTM* [88, 23], στις οποίες παρέχεται η διανυσματική αναπαράσταση μιας ακολουθίας λέξεων του αρχικού κειμένου $X = (x_1,\ x_2, ...,\ x_n)$ (δηλ., παρέχεται το διάνυσμα μιας λέξης x_t σε κάθε βήμα t). Η ακολουθία των διανυσμάτων παρέχεται ταυτόχρονα σε κανονική και ανάστροφη διαδοχή των λέξεων, καθώς έχουμε αμφίδρομες μονάδες *LSTM*, δημιουργώντας μια κρυφή κατάσταση $H_t = bi\_lstm(x_t,\ H_{t-1})$ (Εξίσωση 3.2) στην έξοδό τους για κάθε λέξη εισόδου. Αυτή η κρυφή κατάσταση είναι στην πραγματικότητα η συνένωση των διανυσμάτων κρυφής κατάστασης και των δύο κατευθύνσεων του αμφίδρομου δικτύου *LSTM*. Το επίπεδο αυτό μπορεί να επαναληφθεί περισσότερες από μια φορές δημιουργώντας πολλαπλά επίπεδα αμφίδρομων μονάδων *LSTM* στον κωδικοποιητή, τα οποία διαμορφώνουν ένα βαθύ δίκτυο. Η εξήγηση της συνάρτησης bi_lstm η οποία βασίζεται στη λειτουργικότητα μονάδων *LSTM* μονής κατεύθυνσης ακολουθεί.

Σε κάθε υπολογιστικό βήμα, μια μονάδα *LSTM* [89, 90, 91], υπολογίζει την κρυφή κατάσταση h_t όπως περιγράφεται από τις Εξισώσεις 3.1.

$$
\begin{aligned}
f_t &= \sigma(W_f x_t \ + \ U_f h_{t-1} \ + \ b_f) \\
i_t &= \sigma(W_i x_t \ + \ U_i h_{t-1} \ + \ b_i) \\
o_t &= \sigma(W_o x_t \ + \ U_o h_{t-1} \ + \ b_o) \\
s_t &= f_t \odot s_{t-1} \ + \ i_t \odot tanh(W_s x_t \ + \ U_s h_{t-1} \ + \ b_s) \\
h_t &= o_t \odot tanh(s_t)
\end{aligned}
\tag{3.1}
$$

όπου x_t είναι το διάνυσμα εισόδου (π.χ., *word2vec* διανυσματική αναπαράσταση) της μονάδας *LSTM*, f_t η κατάσταση της πύλης λήθης, i_t η κατάσταση της πύλης εισόδου, o_t η κατάσταση της πύλης εξόδου, s_t η κατάσταση του κελιού μνήμης και h_t η κρυφή κατάσταση στο χρονικό βήμα t. Οι παράμετροι W, U και b αντιπροσωπεύουν τα βάρη εισόδου, τα βάρη αναδρομής και τις πολώσεις, αντίστοιχα. Τέλος, η συνάρτηση σ είναι η σιγμοειδής συνάρτηση και το σύμβολο $\odot$ αντιστοιχεί στην πράξη εσωτερικού γινομένου μεταξύ διανυσμάτων.

Η κρυφή κατάσταση H_t μιας αμφίδρομη μονάδας *LSTM* στο βήμα t προκύπτει από τη συνένωση των διανυσμάτων των κρυφών καταστάσεων δύο μονάδων *LSTM* μονής κατεύθυνσης που δέχονται την είσοδο σε δύο κατευθύνσεις σύμφωνα με τις Εξισώσεις 3.2. Η μια κατεύθυνση αφορά την κανονική τροφοδότηση της ακολουθίας εισόδου σε μια μονάδα *LSTM* και η δεύτερη κατεύθυνση αφορά την ανάστροφη τροφοδότηση της ακολουθίας εισόδου σε δεύτερη μονάδα *LSTM*.

$$\begin{aligned} \overrightarrow{h_t} &= lstm(x_i,\ \overrightarrow{h_{t-1}}),\ x_i \in (x_1,\ x_2,\ ...,\ x_{n-1},\ x_n) \\ \overleftarrow{h_t} &= lstm(x_j,\ \overleftarrow{h_{t-1}}),\ x_j \in (x_n,\ x_{n-1}\ ,...,\ x_2,\ \ x_1) \\ H_t &= [\overrightarrow{h_t} + \overleftarrow{h_t}] \end{aligned} \tag{3.2}$$

Επίπεδο μηχανισμού προσοχής: Το τελευταίο επίπεδο του κωδικοποιητή αποτελείται από έναν μηχανισμό προσοχής (*attention mechanism*) [70, 61] που εστιάζει σε λέξεις του κειμένου εισόδου, ενισχύοντας την ακρίβεια των προβλέψεων εξόδου. Αυτός ο μηχανισμός υπολογίζει το διάνυσμα περιβάλλοντος c_t. Πιο συγκεκριμένα, το διάνυσμα περιβάλλοντος c_t υπολογίζεται ως σταθμισμένο άθροισμα των κρυφών καταστάσεων H_i του κωδικοποιητή, σύμφωνα με τις Εξισώσεις 3.3.

$$\begin{aligned} c_t &= \sum_{i=1}^{|X|} a_{ti} \cdot H_i \\ a_{ti} &= softmax(e_{ti}) \\ e_{ti} &= tanh(W_h \cdot H_i + W_s \cdot s_{t-1} + b) \end{aligned} \tag{3.3}$$

όπου a_{ti} είναι το βάρος κάθε βήματος t για την κρυφή κατάσταση του κωδικοποιητή H_i (δήλ., a_{ti} δείχνει την σημασία της κρυφής κατάστασης H_i), e_{ti} είναι η παράμετρος που δείχνει πόσο καλά ταιριάζει η έξοδος του βήματος t με την είσοδο γύρω από την λέξη i και s_{t-1} είναι η προηγούμενη κατάσταση του κωδικοποιητή. W_h, W_s είναι βάρη του δικτύου και b αντιπροσωπεύει την παράμετρο πόλωσης, οι παράμετροι αυτοί προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης.

Επίπεδο μονάδας *LSTM* (αποκωδικοποιητή): Ο αποκωδικοποιητής αποτελείται από μονάδες *LSTM* μονής κατεύθυνσης που μπορεί να σχηματίζουν ένα ή περισσότερα στρώματα. Σκοπός του αποκωδικοποιητή είναι η πρόβλεψη της επόμενης λέξης y_t της περίληψης, υπολογίζοντας την κρυφή κατάσταση $h_t = lstm(c_t,\ h_{t-1})$ σε κάθε βήμα t, με δεδομένο το διάνυσμα περιβάλλοντος c_t όπως αυτό υπολογίστηκε από το στρώμα προσοχής (Εξίσωση 3.3) και την προηγούμενη κρυφή κατάσταση h_{t-1} του αποκωδικοποιητή. Κατά τη διάρκεια της εκπαίδευσης, τα διανύσματα των λέξεων της ακολουθίας στόχου (της περίληψης αναφοράς) $Y = (y_1,\ y_2, ...,\ y_m)$ τίθενται επίσης στη διάθεση του αποκωδικοποιητή (δηλ., παρέχεται ένα διάνυσμα λέξης y_t σε κάθε βήμα t) και ο αποκωδικοποιητής μαθαίνει να προβλέπει τη επόμενη λέξη, διαμορφώνοντας την τελική περίληψη.

Επίπεδο κανονικοποιημένης εκθετικής συνάρτησης (*Softmax*): Το τελευταίο επίπεδο του αποκωδικοποιητή είναι το στρώμα κατανομής πιθανοτήτων (*softmax*) στο λεξιλόγιο του συνόλου δεδομένων (ή στο λεξιλόγιο εξόδου), το οποίο, ουσιαστικά, δημιουργεί την κατανομή πιθανοτήτων της επόμενης λέξης σε κάθε βήμα λειτουργίας του δικτύου. Συγκεκριμένα, σε κάθε βήμα t, το επίπεδο αυτό υπολογίζει την πιθανότητα κάθε υποψήφιας λέξης y_i του λεξιλογίου Y για την εκτίμηση της συμμετοχής της στην εκτιμώμενη περίληψη, σύμφωνα με την Εξίσωση 3.4.

$$p_t(y_i|X,\ y_{t-1}) = \frac{e^{h_i^T w_i + b_i}}{\sum_{j=1}^{k} e^{h_t^T w_j + b_j}} \tag{3.4}$$

όπου X, y_{t-1} και h_t είναι η ακολουθία λέξεων εισόδου (δηλ., το υποψήφιο για περίληψη κείμενο), η προηγούμενη εκτιμώμενη λέξη και η κρυφή κατάσταση του κωδικοποιητή, αντίστοιχα. Οι παράμετροι w και b αντιστοιχούν στα βάρη και την πόλωση που προσαρμόζονται κατά τη διαδικασία μάθησης. Το άθροισμα των πιθανοτήτων στο σύνολο των υποψήφιων λέξεων είναι ίσο με ένα (Εξίσωση 3.5).

$$\sum_{i=1}^{k} p_t(y_i|X,\ y_{t-1}) = 1 \tag{3.5}$$

Το μοντέλο βαθιάς μάθησης, που περιγράφεται παραπάνω, εκπαιδεύεται από άκρη-σε-άκρη μέσω διαδικασίας επιβλεπόμενης μάθησης με χρήση ενός συνόλου εκπαίδευσης, το οποίο αποτελείται από ζεύγη κειμένου-περίληψης. Κατά τη διάρκεια της εκπαίδευσης του δικτύου, χρησιμοποιείται στοχαστική κατάβαση κλίσης με σκοπό την ελαχιστοποίηση της συνάρτησης αρνητικής λογαριθμικής πιθανοφάνειας (συνάρτηση σφάλματος διασταυρούμενης εντροπίας), όπως περιγράφεται από την Εξίσωση 3.6

$$Loss = -\sum_{t=1}^{T} logP(y_t|X) \tag{3.6}$$

όπου $P(y_t|X)$ είναι η πιθανότητα παρουσίας μιας λέξης στόχου y_t στο βήμα t (δηλ., στο βήμα t από ένα σύνολο βημάτων T που αντιστοιχούν σε T λέξεις της παραγόμενης περίληψης), δεδομένης μιας ακολουθίας λέξεων X που αποτελούν το κείμενο εισόδου.

Επιπλέον, για να αποφευχθεί το φαινόμενο της υπερ-προσαρμογής (*overfitting*), χρησιμοποιείται, επίσης, η τεχνική απόρριψης συνδέσεων μεταξύ κόμβων του δικτύου (*dropout*) [92, 93, 94]. Η τεχνική αυτή απορρίπτει τυχαία συνδέσεις κόμβων από το νευρωνικό δίκτυο κατά τη διάρκεια της εκπαίδευσης.

Αλγόριθμος αναζήτησης δέσμης

Στο στάδιο της εκτίμησης μιας περίληψης για ένα νέο παράδειγμα εισόδου X, η εκτιμώμενη περίληψη αξιολογείται περαιτέρω μέσω του αλγόριθμου αναζήτησης δέσμης (*beam search*), που έχει σκοπό τον προσδιορισμό της βέλτιστης ακολουθίας λέξεων [95, 96], σύμφωνα με την υπολογιζόμενη κατανομή πιθανοτήτων στο λεξιλόγιο εξόδου. Πιο συγκεκριμένα, σύμφωνα με την αναζήτηση δέσμης, σε κάθε βήμα λειτουργίας του αποκωδικοποιητή διατηρούνται w υποψήφιες λέξεις με την υψηλότερη εκτιμώμενη πιθανότητα παρουσίας στην περίληψη, όπου w είναι η παράμετρος πλάτους δέσμης του αλγόριθμου. Ο αλγόριθμος αναζήτησης δέσμης εκτελεί αναζήτηση κατά πλάτος (*breadth-first search - BFS*), όπου μόνο οι w επικρατέστεροι κόμβοι (οι οποίοι αντιστοιχούν στις λέξεις με τη μεγαλύτερη πιθανότητα παρουσίας στην περίληψη) στο βάθος αναζήτησης t διατηρούνται για τη διερεύνηση των μονοπατιών. Το μέγιστο βάθος αναζήτησης είναι ίσο με T, δηλαδή ίσο με το μήκος της ακολουθίας των λέξεων που αποτελούν την περίληψη. Στον χώρο αναζήτησης που δημιουργείται, ο *αλγόριθμος αναζήτησης δέσμης* προσδιορίζει το επικρατέστερο μονοπάτι με τη μεγαλύτερη αθροιστικά πιθανότητα των κόμβων του που αντιστοιχεί στην επικρατέστερη ακολουθία λέξεων, η οποία αποτελεί την βέλτιστη εκτιμώμενη περίληψη εξόδου. Σημειώνεται, ότι ο αλγόριθμος αυτός μπορεί να επεκταθεί ώστε να επιστρέφει τις N καλύτερες περιλήψεις.

3.3.2 Μοντέλο αντιγραφής λέξεων εκτός λεξιλογίου

Το δίκτυο με μηχανισμό αντιγραφής άγνωστων λέξεων (*poineter-generator network*) [61] ή διαφορετικά το μοντέλο αντιγραφής λέξεων εκτός λεξιλογίου (*ΑΛΕΛ*), επεκτείνει το προαναφερόμενο βασικό μοντέλο, κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής της Ενότητας 3.3.1, επιτρέποντας την αντιγραφή λέξεων εκτός λεξιλογίου από το αρχικό κείμενο στην εκτιμώμενη περίληψη. Αυτό το δίκτυο προωθεί στην έξοδο λέξεις που προέρχονται είτε από ένα σταθερό λεξιλόγιο (δηλ., το λεξιλόγιο εξόδου του συνόλου εκπαίδευσης) είτε από τις λέξεις του αρχικού κειμένου. Επομένως, το λεξιλόγιο εξόδου επεκτείνεται ώστε να περιλαμβάνει και λέξεις εκτός λεξιλογίου που ενδεχομένως περιέχονται στο αρχικό κείμενο.

Πιο συγκεκριμένα, το δίκτυο αντιγραφής άγνωστων λέξεων αποτελεί συνδυασμό του βασικού μοντέλου που παρουσιάστηκε παραπάνω (Ενότητα 3.3.1) και ενός δικτύου αρχιτεκτονικής γεννήτριας δεικτών (*pointer generator network*) [97]. Το δίκτυο γεννήτριας δεικτών έχει τη δυνατότητα να υπολογίζει την υπό συνθήκη πιθανότητα μιας ακολουθίας εξόδου, η οποία αποτελείται από διακριτά στοιχεία (π.χ., λέξεις), που περιλαμβάνονται σε θέσεις της ακολουθίας εισόδου. Τα δίκτυα γεννήτριας δεικτών μπορούν να εφαρμοστούν σε προβλήματα με μεταβλητό μήκος λεξιλογίου εξόδου. Το μοντέλο αντιγραφής άγνωστων λέξεων επεκτείνει τη βασική αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή ενσωματώνοντας ένα δίκτυο γεννήτριας δεικτών που υπολογίζει την πιθανότητα γεννήτριας $P_{g,t} \in [0, 1]$ στο βήμα t της λειτουργίας του αναδρομικού νευρωνικού δικτύου. Η πιθανότητα γεννήτριας δίνεται από την Εξίσωση 3.7.

$$P_{g,t} = \sigma(w_c \cdot c_t + w_s \cdot s_t + w_x \cdot x_t + b_p) \tag{3.7}$$

όπου με σ συμβολίζουμε τη σιγμοειδή συνάρτηση, c_t είναι το διάνυσμα περιβάλλοντος όπως υπολογίζεται από τον μηχανισμό προσοχής (Εξίσωση 3.3), s_t το διάνυσμα κατάστασης του αποκωδικοποιητή και x_t το διάνυσμα εισόδου. Οι παράμετροι που προσαρμόζονται κατά τη διαδικασία εκπαίδευσης w_h, w_s, w_x αποτελούν τα βάρη και b_g την παράμετρο πόλωσης του δικτύου. Η πιθανότητα $P_t(w_i)$ παρουσίας της λέξης w_i στο βήμα t στην εκτιμώμενη περίληψη υπολογίζεται σύμφωνα με την Εξίσωση 3.8, η οποία δίνει την κατανομή πιθανοτήτων του εκτεταμένου λεξιλογίου (δηλ., το λεξιλόγιο που προκύπτει από τη συνένωση του σταθερού λεξιλογίου του συνόλου εκπαίδευσης και των άγνωστων λέξεων του κειμένου εισόδου).

$$P_t(w_i) = P_{g,t} \cdot P_v(w_i) + (1 - P_{g,t}) \cdot a_{it} \tag{3.8}$$

όπου $P_v(w_i)$ είναι η πιθανότητα παρουσίας μιας λέξης w_i στην ακολουθία εξόδου. Να διευκρινίσουμε, ότι P_v είναι η υπολογιζόμενη κατανομή πιθανοτήτων των λέξεων του σταθερού λεξιλογίου του συνόλου εκπαίδευσης, όπως υπολογίζεται από το στρώμα κανονικοποιημένης εκθετικής συνάρτησης (Εξίσωση 2.9). Η παράμετρος a_{it} αντιπροσωπεύει το βάρος της λέξης i στο βήμα t, όπως υπολογίζεται στο επίπεδο του μηχανισμού προσοχής. Αν η λέξη w_i είναι λέξη εκτός λεξιλογίου (*ΛΕΛ*) τότε ισχύει $P_V(w_i) = 0$. Παρόμοια, αν μια λέξη δεν εμφανίζεται στο κείμενο εισόδου, τότε $a_{it} = 0$. Με αυτόν τον τρόπο υπολογίζεται η τελική κατανομή πιθανοτήτων στο εκτεταμένο λεξιλόγιο το οποίο περιλαμβάνει και τις λέξεις εκτός λεξιλογίου.

Επιπλέον, το μοντέλο αντιγραφής άγνωστων λέξεων ενσωματώνει έναν μηχανισμό αποφυγής της επανάληψης των λέξεων στην έξοδο (*coverage mechanism*), προσαρμόζοντας τη λύση που προτείνεται στην εργασία [98] για το πρόβλημα της μηχανικής μετάφρασης. Για τον σκοπό αυτό υπολογίζεται ένα διάνυσμα κάλυψης cov_t σε κάθε βήμα t του αναδρομικού νευρωνικού δικτύου, το

οποίο είναι ίσο με το άθροισμα των βαρών του μηχανισμού προσοχής σύμφωνα με την Εξίσωση 3.9 για όλα τα προηγούμενα βήματα t'.

$$cov_t = \sum_{t'=0}^{t-1} a_{t'} \tag{3.9}$$

Το διάνυσμα cov_t είναι μια κατανομή των λέξεων του κειμένου εισόδου που δείχνει τον βαθμό κάλυψης αυτών των λέξεων από τον μηχανισμό προσοχής. Σημειώνεται ότι $cov_0 = 0$, καθώς κατά το πρώτο βήμα καμία λέξη του αρχικού κειμένου δεν έχει εξεταστεί. Το διάνυσμα κάλυψης δίνεται ως μια επιπλέον είσοδος στον μηχανισμό προσοχής με την τροποποίηση της Εξίσωσης 3.3 του διανύσματος $e_{t,i}$ που παρουσιάστηκε παραπάνω σύμφωνα με την Εξίσωση 3.10.

$$e_{t,i} = tanh(w_h \cdot H_i + w_s \cdot s_{t-1} + w_{cov} \cdot cov_t + b) \tag{3.10}$$

Όπου w_{cov} είναι παράμετρος που προσαρμόζεται κατά τη διαδικασία εκπαίδευσης. Αυτή η τροποποίηση αποτρέπει τον μηχανισμό προσοχής στο να δίνει προσοχή στις ίδιες λέξεις του κειμένου εισόδου, αποφεύγοντας την επανάληψη λέξεων στην έξοδο.

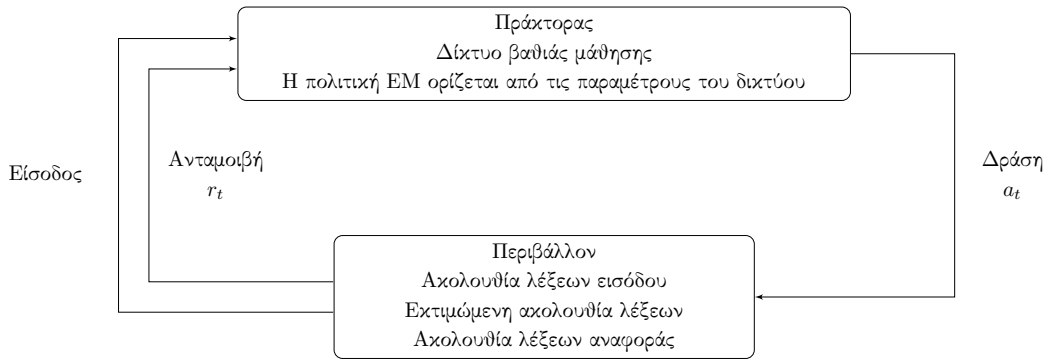
Η συνάρτηση σφάλματος που χρησιμοποιείται για την εκπαίδευση αυτού του μοντέλου, όπως και παραπάνω, είναι η συνάρτηση αρνητικής λογαριθμικής πιθανοφάνειας (Εξίσωση 3.6) με τη διαφορά ότι οι πιθανότητες των λέξεων εξόδου υπολογίζονται σύμφωνα με την Εξίσωση 3.8.

3.3.3 Μοντέλο ενισχυτικής μάθησης

Τα δυο προαναφερόμενα μοντέλα βαθιάς μάθησης (Ενότητες 3.3.1 και 3.3.2) επιδιώκουν την ελαχιστοποίηση μιας συνάρτησης σφάλματος, η οποία όμως, όπως θα δούμε και στη συνέχεια, δεν αποτελεί μέτρο αξιολόγησης για την αυτόματη *ΠΚ*. Σε αντίθεση με τα προαναφερόμενα μοντέλα, ένα μοντέλο ενισχυτικής μάθησης (*EM*) μπορεί να βασίζεται στη μεγιστοποίηση μιας μετρικής αξιολόγησης η οποία ταιριάζει με το συγκεκριμένο πρόβλημα (δηλ., μεγιστοποίηση μιας συγκεκριμένης μετρικής αξιολόγησης για την αυτόματη *ΠΚ*).

Πιο συγκεκριμένα, το μοντέλο *EM*, ακολουθώντας την προσέγγιση του [78], ενσωματώνει το δίκτυο κωδικοποιητή-αποκωδικοποιητή που παρουσιάστηκε παραπάνω, το οποίο, στο πλαίσιο της ΕΜ, λειτουργεί ως πράκτορας που επιδρά με το περιβάλλον του. Το Σχήμα 3.2 παρουσιάζει τα βασικά στοιχεία ενός μοντέλου *EM* για την αυτόματη περίληψη κειμένου. Οι παράμετροι του δικτύου θ ορίζουν μια πολιτική, η οποία καθορίζει τις δράσεις για την εκτίμηση μιας ακολουθίας λέξεων που αποτελεί την περίληψη σε δεδομένη είσοδο. Ουσιαστικά, μια δράση a_t πραγματοποιείται με στόχο την εκτίμηση μιας νέας λέξης σε κάθε χρονικό βήμα t της ακολουθίας των λέξεων της εξόδου. Μετά από κάθε δράση, ο πράκτορας (το δίκτυο βαθιάς μάθησης) λαμβάνει μια ανταμοιβή (*reward*) r_t με σκοπό την ενημέρωση των παραμέτρων του θ. Η ανταμοιβή υπολογίζεται με τη σύγκριση μεταξύ της εκτιμώμενης ακολουθίας λέξεων και της ακολουθίας λέξεων στόχου (περίληψη αναφοράς), σύμφωνα με τη χρησιμοποιούμενη μετρική αξιολόγησης.

Για την εκπαίδευση του δικτύου χρησιμοποιείται αλγόριθμος *EM* που εφαρμόζει πολιτική αυτοκριτικής για τη μάθηση της εκτίμησης μιας ακολουθίας λέξεων εξόδου, για δεδομένη ακολουθία λέξεων εισόδου (*self-critical sequence training*) [99]. Σύμφωνα με τον αλγόριθμο αυτό, το μοντέλο μηχανικής μάθησης παράγει δυο ακολουθίες λέξεων σε κάθε επανάληψη εκπαίδευσης.



Σχήμα 3.2: Μοντέλο ενισχυτικής μάθησης για την αυτόματη περίληψη κειμένου.

Η μια ακολουθία λέξεων είναι η $Y_s = (y_{s,1}, y_{s,2}, ..., y_{s,t}, ..., y_{s,T})$ που ανακτάται με δειγματοληψία από την κατανομή πιθανοτήτων $p(y_{s,t}|y_1, ..., y_{t-1}, X)$ του μοντέλου, σύμφωνα με τις πιθανότητες που δίνονται από το στρώμα *softmax* σε δεδομένη είσοδο X για κάθε βήμα του δικτύου t (από ένα σύνολο T βημάτων που αντιστοιχούν στο μήκος της ακολουθίας εξόδου). Σημειώνεται ότι $y_{s,t}$ αντιπροσωπεύει μια λέξη δειγματοληψίας στο χρονικό βήμα t. Αντίστοιχα, $\hat{Y} = (\hat{y_1}, \hat{y_2}, ..., \hat{y_t}, ..., \hat{y_T})$ είναι η εκτιμώμενη ακολουθία εξόδου, η οποία ανακτάται μεγιστοποιώντας την κατανομή πιθανοτήτων του λεξιλογίου εξόδου σε κάθε χρονικό βήμα t (δηλ., εκτίμηση της ακολουθίας εξόδου όπως γίνεται στη φάση πρόβλεψης της εξόδου σε νέα παραδείγματα εισόδου με χρήση του αλγόριθμου αναζήτησης δέσμης που περιγράφεται παραπάνω).

Στη φάση εκπαίδευσης, το μοντέλο αναδρομικού νευρωνικού δικτύου, το οποίο χρησιμοποιείται ως πράκτορας, εκτιμά την ακολουθία δειγματοληψίας Y_s των λέξεων εξόδου γνωρίζοντας σε κάθε βήμα t την προηγούμενη λέξη της ακολουθίας στόχου. Αντίθετα, κατά την εκτίμηση της ακολουθίας εξόδου $\hat{Y}$, το μοντέλο λαμβάνει υπόψη του τις προηγούμενες εκτιμώμενες λέξεις και όχι τις προηγούμενες λέξεις της ακολουθίας στόχου. Συνεπώς, η ακολουθία στόχος είναι γνωστή στο μοντέλο κατά τη διάρκεια της εκπαίδευσης, ενώ είναι άγνωστη στη φάση πρόβλεψης με είσοδο νέων παραδειγμάτων. Αυτό έχει ως αποτέλεσμα, η ακολουθία δειγματοληψίας να προσεγγίζει με μεγαλύτερη ακρίβεια την ακολουθία αναφοράς σε σχέση με την εκτιμώμενη ακολουθία. Αυτό αποτελεί πρόβλημα μεροληψίας κατά την εκπαίδευση (γνωστό ως *exposure bias*).

Ακολουθώντας την προσέγγιση [99], το μοντέλο εκπαιδεύεται μεγιστοποιώντας την ανταμοιβή που λαμβάνεται για την ακολουθία δειγματοληψίας ή ελαχιστοποιώντας τη συνάρτηση σφάλματος της Εξίσωσης 3.11.

$$L(\theta) = -E_{Y_s \sim \theta}[r(Y_s)] \simeq -r(Y_s) \tag{3.11}$$

Όπου $-E_{Y_s \sim P_\theta}[(r(Y_s)]$ είναι η αναμενόμενη τιμή ανταμοιβής της ακολουθίας δειγματοληψίας Y_s σύμφωνα με την κατανομή πιθανοτήτων P_θ του λεξιλογίου εξόδου που ορίζεται από τις παραμέτρους θ του μοντέλου.

Οι μερικές παράγωγοι των παραμέτρων υπολογίζονται σύμφωνα με την εξίσωση 3.12.

$$\nabla_\theta L(\theta) = -E_{Y_s \sim \theta}[r(Y_s) \nabla_\theta log(P_\theta(Y_s)]) \simeq -r(Y_s) \nabla_\theta log(P_\theta(Y_s)) \tag{3.12}$$

Για να μειωθεί η διακύμανση, αφαιρείται από την ανταμοιβή της ακολουθίας δειγματοληψίας η ανταμοιβή της εκτιμώμενης ακολουθίας σύμφωνα με την Εξίσωση 3.13.

$$\nabla_\theta L(\theta) = -(r(Y_s) - r(\hat{Y})) \nabla_\theta log(P_\theta(Y_s)) \tag{3.13}$$

Σύμφωνα με τα παραπάνω, το μοντέλο *EM* εκπαιδεύεται ελαχιστοποιώντας τη συνάρτηση της Εξίσωσης 3.14.

$$L_{RL} = -(r(Y_s) - r(\hat{Y})) \sum_{i=1}^{T} log(p(y_{s,t}|y_{s,1}, y_{s,2}, ..., y_{s,t-1}; \theta, X)) \tag{3.14}$$

Είναι φανερό ότι η ελαχιστοποίηση της συνάρτησης L_{RL} ισοδυναμεί με την μεγιστοποίηση της υπό συνθήκη πιθανότητας της ακολουθίας δειγματοληψίας, όταν η ανταμοιβή της ακολουθίας δειγματοληψίας $r(Y_s)$ είναι μεγαλύτερη από την ανταμοιβή της εκτιμώμενης ακολουθίας $r(\hat{Y})$, με αποτέλεσμα να αυξάνεται η ανταμοιβή που λαμβάνεται από το μοντέλο.

Επιπλέον, το μοντέλο ενσωματώνει μηχανισμό αντιγραφής λέξεων εκτός λεξιλογίου και μηχανισμό αποφυγής της επανάληψης λέξεων όπως περιγράφεται παραπάνω στην Ενότητα 3.3.2.

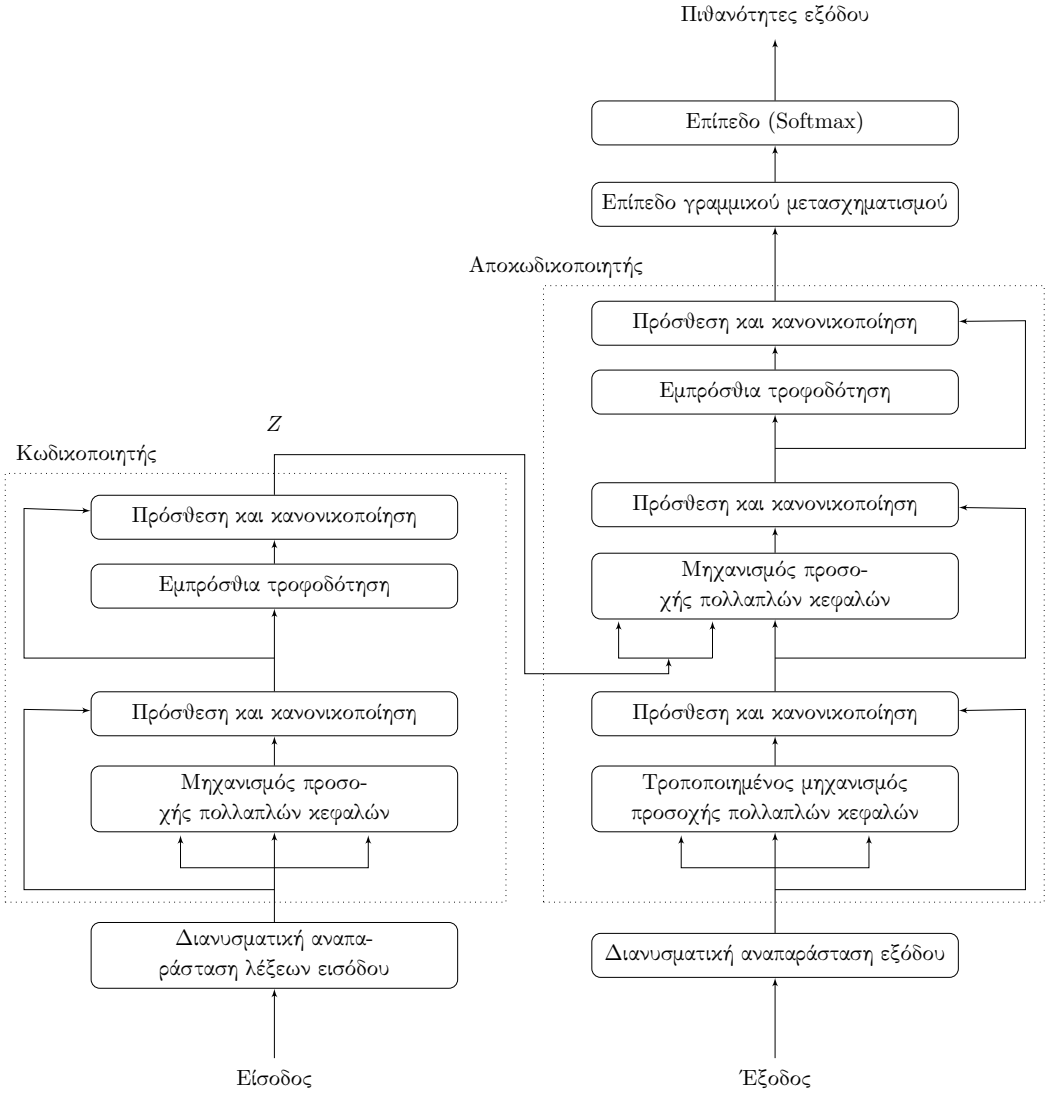
3.3.4 Μοντέλα που βασίζονται σε αρχιτεκτονική μετασχηματιστών

Δεδομένου ότι τα μοντέλα που βασίζονται σε αρχιτεκτονική τύπου μετασχηματιστών (*transformers*) [65, 74, 67, 66, 75, 68] αποτελούν την επικρατέστερη προσέγγιση σε προβλήματα κατανόησης φυσικής γλώσσας [69], ακολουθούμε τη μεθοδολογία του [66] χρησιμοποιώντας δύο προσεγγίσεις δικτύων μετασχηματιστών οι οποίες είναι οι εξής: (i) μοντέλο μετασχηματιστών ($M\Sigma$) και (ii) μοντέλο μετασχηματιστών με προ-εκπαιδευμένο κωδικοποιητή ($M\Sigma\Pi K$). Τα δύο αυτά μοντέλα παρουσιάζονται με λεπτομέρεια παρακάτω.

Μοντέλο μετασχηματιστών

Το μοντέλο μετασχηματιστών ($M\Sigma$) [65] αποτελεί ένα δίκτυο αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή που βασίζεται κυρίως σε μηχανισμό προσοχής, ο οποίος επιτρέπει τον παραλληλισμό των εργασιών για τον υπολογισμό των εξαρτήσεων μεταξύ εισόδου και εξόδου, αποφεύγοντας τη χρήση αναδρομικών ή συνελικτικών δικτύων. Ο κωδικοποιητής δέχεται ως είσοδο μια ακολουθία διανυσμάτων $(x_1, x_2, ..., x_n)$ που αντιστοιχούν στην ακολουθία των λεκτικών μονάδων εισόδου και υπολογίζει την ακολουθία διανυσμάτων $z = (z_1, z_2, ..., z_n)$. Στη συνέχεια, με δεδομένη την ακολουθία z, ο αποκωδικοποιητής σε κάθε χρονικό βήμα t παράγει μια έξοδο y_t, χρησιμοποιώντας ως επιπλέον είσοδο την ακολουθία των προηγούμενων εξόδων $(y_1, y_2, ..., y_{t-1})$. Τελικά, μετά από m χρονικά βήματα, ο αποκωδικοποιητής παράγει μια ακολουθία εξόδου $(y_1, y_2, ..., y_m)$. Το Σχήμα 3.3 απεικονίζει την αρχιτεκτονική ενός δικτύου $M\Sigma$ το οποίο περιγράφεται με λεπτομέρεια παρακάτω.

Στρώμα διανυσματικής αναπαράστασης λέξεων: Αποτελεί το στρώμα που μετατρέπει τις λέξεις εισόδου σε μια διανυσματική αναπαράσταση, και, στη συνέχεια, αυτές οι διανυσματικές αναπαραστάσεις προωθούνται στον κωδικοποιητή. Τα διανύσματα που αντιπροσωπεύουν τις λέξεις, διάστασης d_{model}, μπορεί είτε να είναι προ-εκπαιδευμένα και να παραμένουν σταθερά κατά τη διάρκεια της εκπαίδευσης είτε να εκπαιδεύονται και να προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης (με τυχαία αρχικοποίηση ή χρήση προ-εκπαιδευμένων διανυσμάτων που προσαρμόζονται περαιτέρω).



Σχήμα 3.3: Αρχιτεκτονική μετασχηματιστών.

Κωδικοποίηση θέσης: Με δεδομένο ότι το μοντέλο $M\Sigma$ δεν περιλαμβάνει αναδρομή, χρειάζεται να δοθεί πληροφορία για τη σχετική ή απόλυτη θέση κάθε λεκτικής μονάδας στην ακολουθία των λεκτικών μονάδων εισόδου. Για τον σκοπό αυτό, για κάθε λεκτική μονάδα εισόδου, χρησιμοποιείται κωδικοποίηση θέσης που αντιστοιχεί σε μια διανυσματική αναπαράσταση θέσης (διάστασης d_{model}) που προστίθεται στις αντίστοιχες διανυσματικές αναπαραστάσεις των λέξεων, με σκοπό να προκύψει η διανυσματική αναπαράσταση εισόδου για κάθε λέξη. Στην παρούσα προσέγγιση χρησιμοποιείται η ημιτονοειδής συνάρτηση της Εξίσωσης 3.15 για την κωδικοποίηση θέσης $PE(p, i)$, όπου p είναι η θέση στην ακολουθία των λεκτικών μονάδων και i είναι μια διάσταση ενός διανύσματος εισόδου από ένα σύνολο d_{model} διαστάσεων. Δηλαδή, υπολογίζεται μια τιμή ημιτόνου για κάθε διάσταση $i \in (1, 2, 3, ..., d_{model})$, δημιουργώντας ένα διάνυσμα κωδικοποίησης θέσης για κάθε θέση p συνολικής διάστασης d_{model}.

$$PE(p, i) = sin(\frac{p}{10000^{\frac{2i}{d_{model}}}}) \tag{3.15}$$

Κωδικοποιητής: Ο κωδικοποιητής αποτελείται από μια στοίβα πανομοιότυπων στρωμάτων (στο Σχήμα 3.3 απεικονίζεται μόνο ένα τέτοιο στρώμα), όπου κάθε στρώμα περιλαμβάνει δύο υπό-επίπεδα: το πρώτο επίπεδο υλοποιεί έναν μηχανισμό αυτο-προσοχής (ή ένδο-προσοχής) πολλαπλών κεφαλών (*multi-head self-attention mechanism*) και το δεύτερο υπό-επίπεδο αποτελεί ένα, ως προς τη θέση (*position wise*), πλήρως διασυνδεδεμένο δίκτυο εμπρόσθιας τροφοδότησης (*fully connected feed-forward network*). Γύρω από αυτά τα δύο υπο-επίπεδα χρησιμοποιούνται υπολειμματικές συνδέσεις (*Residual connections*) [100] και ακολουθεί κανονικοποίηση στρώματος (*layer normalization*) [101]. Συνεπώς, η έξοδος κάθε υπο-επιπέδου ισοδυναμεί με $f_{LayerNorm}(x + f_{Sublayer}(x))$, όπου x είναι η είσοδος ενός υπο-επιπέδου και $f_{Sublayer}(x)$ είναι η συνάρτηση που υλοποιεί το αντίστοιχο υπο-επίπεδο (π.χ., υπο-επίπεδο μηχανισμού προσοχής ή υπο-επίπεδο δικτύου εμπρόσθιας τροφοδότησης). Για να υλοποιηθούν οι υπολειμματικές συνδέσεις, όλα τα επίπεδα του κωδικοποιητή παράγουν διανύσματα διάστασης d_{model}. Ίδια είναι η διάσταση και της εισόδου στον αποκωδικοποιητή, η οποία διατηρείται από το επίπεδο διανυσματικής αναπαράστασης των λεκτικών μονάδων εισόδου.

Να σημειώσουμε ότι, σχετικά με τις υπολειμματικές συνδέσεις, θεωρείται ότι ένα στρώμα που περιβάλλεται από τέτοιες συνδέσεις μαθαίνει μια συνάρτηση $\hat{f}_r(x)$, η οποία δέχεται μια είσοδο x και η έξοδος του στρώματος $\hat{f}_r(x) + x$, προσεγγίζει μια επιθυμητή συνάρτηση $f(x)$. Συνεπώς, η συνάρτηση $\hat{f}_r(x)$ του εκάστοτε στρώματος υπολειμματικής σύνδεσης μαθαίνει την αντιστοιχία $\hat{f}_r(x) := f(x) - x$ [100].

Αποκωδικοποιητής: Ο αποκωδικοποιητής, επίσης, αποτελείται από μια στοίβα πανομοιότυπων επιπέδων (το Σχήμα 3.3 απεικονίζει ένα τέτοιο επίπεδο το οποίο μπορεί να επαναληφθεί περισσότερες φορές). Τα επίπεδα του αποκωδικοποιητή, με τη σειρά τους, αποτελούνται από τρία υπο-επίπεδα το κάθε ένα, παρόμοια με τον κωδικοποιητή, όπως φαίνεται στο Σχήμα 3.3 (δηλ., χρησιμοποιείται μηχανισμός αυτο-προσοχής, υπολειμματικές συνδέσεις και κανονικοποίηση επιπέδου για κάθε υπο-επίπεδο). Στο πρώτο υπο-επίπεδο έχει τροποποιηθεί ο μηχανισμός αυτό-προσοχής για να αποτρέψει την προσοχή σε επόμενες θέσεις (δηλ., σε κάθε χρονικό βήμα λειτουργίας του αποκωδικοποιητή να γίνεται η εκτίμηση της επόμενης λέξης με τη γνώση μόνο των προηγούμενων λέξεων της ακολουθίας εξόδου). Αυτό γίνεται με κάλυψη της πρώτης θέσης της ακολουθίας εξόδου που δίνεται στον αποκωδικοποιητή με ένα σύμβολο αρχής (το οποίο έχει συγκεκριμένη διανυσματική αναπαράσταση) και είναι ίδιο σε όλες τις ακολουθίες

εξόδου. Συνεπώς η ακολουθία εξόδου, που δίνεται στον αποκωδικοποιητή έχει μετακινηθεί μια θέση δεξιά ώστε σε κάθε χρονικό βήμα να μη γνωρίζει ο αποκωδικοποιητής την λέξη-στόχο της θέσης αυτής, αλλά να αποφασίζει σύμφωνα με τις προηγούμενες λέξεις.

Μηχανισμός προσοχής πολλαπλών κεφαλών: Ο μηχανισμός προσοχής μπορεί να περιγραφεί ως μια αντιστοίχιση ενός ερωτήματος (*query*) Q και ενός συνόλου από ζεύγη κλειδιού-τιμής (*key, value*) K, V σε μια έξοδο. Όπου Q, K και V είναι διανύσματα. Η έξοδος υπολογίζεται ως ένα σταθμισμένο άθροισμα των τιμών, όπου τα βάρη κάθε τιμής υπολογίζονται από μια συνάρτηση συμβατότητας του ερωτήματος με το αντίστοιχο κλειδί. Σύμφωνα με την εργασία [65], η συνάρτηση προσοχής υπολογίζεται από την Εξίσωση 3.16.

$$f_{Att}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.16}$$

Η προσοχή πολλαπλών κεφαλών χρησιμοποιεί μια διαδικασία παραλληλισμού των ερωτημάτων, κλειδιών και τιμών, με αρχική διάσταση d_{model} το καθένα, σε h διανύσματα διάστασης d_q, d_k και d_v, αντίστοιχα. Όπου $d_q = d_k = d_v = \frac{d_{model}}{h}$ (π.χ., αν $d_{model} = 512$, και $h = 8$, τότε $d_q = d_k = d_v = 64$). Οι h αυτές προβολές των αρχικών διανυσμάτων συμμετέχουν σε διαδικασίες που εκτελούνται παράλληλα, κάτι που αποτελεί βασικό πλεονέκτημα για ένα μοντέλο $MΣ$. Μετά τον παραλληλισμό των υπολογισμών, η κάθε κεφαλή ($head_i$) του μηχανισμού προσοχής υπολογίζεται από την Εξίσωση 3.17.

$$head_i = f_{Att}(QW_{Q,i}, KW_{K,i}, VW_{V,i}) \tag{3.17}$$

όπου $W_{Q,i} \in R^{d_{model} \times d_Q}$, $W_{K,i} \in R^{d_{model} \times d_K}$ και $W_{V,i} \in R^{d_{model} \times d_V}$ είναι οι πίνακες βαρών των ερωτημάτων, κλειδιών και τιμών, αντίστοιχα.

Η έξοδος του μηχανισμού προσοχής πολλαπλών κεφαλών είναι ίση με τη συνένωση των επιμέρους κεφαλών ($head_i$) σύμφωνα με την Εξίσωση 3.18.

$$f_{MultiHeadAtt} = Concat(head_1, head_2, ..., head_h) \tag{3.18}$$

Στα επίπεδα αυτό-προσοχής (*self-attention*) του κωδικοποιητή, τα ερωτήματα Q, τα κλειδιά K και οι τιμές V προέρχονται από το ίδιο μέρος, την έξοδο του προηγούμενου στρώματος του κωδικοποιητή. Κάθε θέση της ακολουθίας εισόδου (δηλ., ακολουθία λέξεων του αρχικού κειμένου) στον κωδικοποιητή μπορεί να παρακολουθεί όλες τις θέσεις του προηγούμενου επιπέδου του κωδικοποιητή. Αντίστοιχα, τα στρώματα αυτό-προσοχής του αποκωδικοποιητή επιτρέπουν σε κάθε θέση της ακολουθίας εισόδου (δηλ., ακολουθία λέξεων της περίληψης) του αποκωδικοποιητή να παρακολουθεί όλες τις θέσεις του αποκωδικοποιητή έως την τρέχουσα θέση. Το επίπεδο προσοχής μεταξύ του κωδικοποιητή και αποκωδικοποιητή δέχεται τα ερωτήματα από το προηγούμενο επίπεδο του αποκωδικοποιητή, ενώ τα κλειδιά και τις τιμές από την έξοδο του κωδικοποιητή. Αυτό επιτρέπει κάθε θέση του αποκωδικοποιητή να παρακολουθεί όλες τις θέσεις της ακολουθίας εισόδου στον κωδικοποιητή.

Εμπρόσθιας τροφοδότησης δίκτυο ως προς τη θέση: Κάθε επίπεδο του κωδικοποιητή ή του αποκωδικοποιητή περιλαμβάνει ένα πλήρως διασυνδεδεμένο δίκτυο εμπρόσθιας τροφοδότησης (ως προς τη θέση), το οποίο εφαρμόζεται σε κάθε θέση χωριστά και πανομοιότυπα. Η Εξίσωση 3.19 περιγράφει τη συνάρτηση του δικτύου εμπρόσθιας τροφοδότησης f_{ffn}, όπου έχουμε δύο γραμμικούς μετασχηματισμούς με συνάρτηση ενεργοποίησης γραμμικού ανορθωτή

($ReLU$) μεταξύ των δύο μετασχηματισμών.

$$f_{ffn} = ReLU(x \cdot W_1 + b_1)W2 + b2 = max(0,\ x \cdot W_1 + b_1)W2 + b2 \tag{3.19}$$

Η διάσταση της εισόδου και εξόδου σε αυτό το δίκτυο είναι ίση με d_{model} ενώ η εσωτερική διάσταση είναι ίση με d_{ffn} (π.χ., $d_{model} = 512$ και $d_{ffn} = 2048$).

Υπολογισμός κατανομής πιθανοτήτων εξόδου: Η έξοδος του αποκωδικοποιητή οδηγείται σε ένα στρώμα συνάρτησης γραμμικού μετασχηματισμού και, στη συνέχεια, σε ένα στρώμα εκθετικής κανονικοποιημένης συνάρτησης (*softmax*) για τον υπολογισμό της κατανομής πιθανοτήτων στο λεξιλόγιο εξόδου, σε κάθε χρονικό βήμα του αποκωδικοποιητή.

Μοντέλο μετασχηματιστών προ-εκπαιδευμένου κωδικοποιητή

Το δεύτερο μοντέλο που υλοποιεί αρχιτεκτονική μετασχηματιστών χρησιμοποιεί έναν προ-εκπαιδευμένο κωδικοποιητή που βασίζεται στο μοντέλο *αναπαραστάσεων αμφίδρομου κωδικοποιητή από μετασχηματιστές* (*Bidirectional Encoder Representations from Transformers*) γνωστό ως *BERT* [102]. Ο αποκωδικοποιητής του μοντέλου είναι αρχιτεκτονικής μετασχηματιστών, σύμφωνα με την εργασία [66] όπως έχει ήδη περιγραφεί. Σε αντίθεση με το μοντέλο *ΜΣ* που παρουσιάστηκε παραπάνω, το οποίο εκπαιδεύεται από την αρχή, το μοντέλο μετασχηματιστών προ-εκπαιδευμένου κωδικοποιητή (*ΜΣΠΚ*) ακολουθεί μια διαδικασία μάθησης που βασίζεται στα χρησιμοποιούμενα σύνολα δεδομένων με σκοπό την περαιτέρω προσαρμογή των βαρών του κωδικοποιητή και επίσης την προσαρμογή των βαρών του αποκωδικοποιητή.

Προ-εκπαιδευμένα γλωσσικά μοντέλα: Τα προ-εκπαιδευμένα γλωσσικά μοντέλα έχουν επικρατήσει σε διαδικασίες ΕΦΓ. Πιο συγκεκριμένα, το μοντέλο *BERT*, το οποίο χρησιμοποιείται στην παρούσα εργασία, αποτελεί ένα μοντέλο γλωσσικής αναπαράστασης το οποίο έχει εκπαιδευτεί σε διαδικασία πρόβλεψης λέξεων και πρόβλεψης της επόμενης πρότασης με χρήση των σωμάτων κειμένου *BookCorpus* (800 εκατομμυρίων λέξεων) [103] και της αγγλικής *Wikipedia* (2,5 δισεκατομμυρίων λέξεων). Σε αυτή την αρχιτεκτονική, για την αναπαράσταση κάθε λέξης εισόδου, χρησιμοποιούνται τριών ειδών διανυσματικές αναπαραστάσεις: (*i*) διανυσματική αναπαράσταση λέξεων που ενσωματώνει την έννοια κάθε λέξης (*word embeddings*), (*ii*) διανυσματική αναπαράσταση διαχωρισμού προτάσεων (*segmentation embeddings*) που χρησιμοποιείται για την διάκριση μεταξύ δύο προτάσεων και (*iii*) διανυσματική αναπαράσταση θέσης (*position embeddings*) που δείχνει τη θέση της κάθε λέξης στην πρόταση. Αυτά τα τρία διανύσματα προστίθενται και προκύπτει ένα διάνυσμα που δίνεται ως είσοδος στο μοντέλο αμφίδρομου μετασχηματιστή, το οποίο περιλαμβάνει πολλαπλά επίπεδα. Η αρχιτεκτονική των μετασχηματιστών ακολουθεί την προσέγγιση του μοντέλου μετασχηματιστών με μηχανισμό προσοχής πολλαπλών κεφαλών [65] που παρουσιάστηκε παραπάνω. Τον μοντέλο *BERT* παράγει στην έξοδο ένα διάνυσμα t_i για κάθε λέξη i με πλούσια πληροφορία σχετικά με τα συμφραζόμενα. Τα προ-εκπαιδευμένα γλωσσικά μοντέλα χρησιμοποιούνται σε εργασίες κατανόησης φυσικής γλώσσας.

Κωδικοποιητής *BERT* για την αυτόματη περίληψη κειμένου: Για τη διαδικασία περαιτέρω προσαρμογής των βαρών (γνωστή ως *fine-tuning*) του κωδικοποιητή *BERT*, έχουν γίνει κάποιες τροποποιήσεις σε σχέση με το αρχικό μοντέλο *BERT* με σκοπό την προσαρμογή του στο πρόβλημα της αυτόματης περίληψης κειμένου. Πιο συγκεκριμένα, στην αρχή κάθε

πρότασης προστίθεται η λέξη $[CLS]$ που δείχνει την έναρξη της πρότασης. Για τη διανυσματική αναπαράσταση διαχωρισμού των προτάσεων χρησιμοποιούμε δύο διανυσματικές αναπαραστάσεις: την E_a για τις προτάσεις που βρίσκονται σε άρτια θέση i και E_b για τις προτάσεις που βρίσκονται σε περιττή θέση i, όπου i είναι ο δείκτης θέσης κάθε πρότασης στην ακολουθία των προτάσεων του κειμένου $(s_1, s_2, ..., s_i..., s_k)$ (π.χ., για τις προτάσεις (s_1, s_2, s_3, s_4) έχουμε τις αναπαραστάσεις (E_a, E_b, E_a, E_b). Για τις διανυσματικές αναπαραστάσεις που δείχνουν τη θέση κάθε λέξης, χρησιμοποιούνται ημιτονοειδής αναπαραστάσεις θέσης (*sinusoid positional embeddings*) όπως και στην περίπτωση του μοντέλου $MΣ$ που παρουσιάστηκε παραπάνω.

3.4 Πειραματικό μέρος

Στην ενότητα αυτή παρουσιάζεται η πειραματική διαδικασία για την αξιολόγηση των μοντέλων μηχανικής μάθησης που παρουσιάστηκαν παραπάνω, με σκοπό τη βελτιστοποίηση των παραμέτρων και τη διερεύνηση των επιδόσεών τους. Η διαδικασία αξιολόγησης ακολουθεί την τυπική μεθοδολογία που χρησιμοποιείται σε σχετικές εργασίες [104, 60, 59, 61, 63]. Στη συνέχεια, η Ενότητα 3.4.1 παρουσιάζει τα χρησιμοποιούμενα σύνολα δεδομένων, η Ενότητα 3.4.2 περιγράφει τις μετρικές αξιολόγησης, η Ενότητα 3.4.3 παρουσιάζει την πειραματική διαδικασία και τις επιλογές βελτιστοποίησης των παραμέτρων, η Ενότητα 3.4.4 παραθέτει τα πειραματικά αποτελέσματα και η Ενότητα 3.4.5 περιλαμβάνει την περιγραφή και ερμηνεία των αποτελεσμάτων.

3.4.1 Σύνολα δεδομένων

Η αξιολόγηση βασίζεται στη χρήση τριών συνόλων δεδομένων τα οποία χρησιμοποιούνται ευρέως στο πεδίο της αυτόματης *ΠΚ*. Τα σύνολα δεδομένων, τα οποία περιγράφονται αναλυτικά παρακάτω, είναι τα εξής: *Gigaword* [105], *DUC 2004* [106] και *CNN/DailyMail* [107].

Σύνολο δεδομένων *Gigaword*: Η έκδοση του συνόλου δεδομένων *Gigaword* που χρησιμοποιείται προέρχεται από την εργασία [104] και είναι αυτή η έκδοση που έχει υιοθετηθεί ευρέως σε εργασίες της σχετικής βιβλιογραφίας [8]. Το *Gigaword* περιέχει περίπου $3,8$ εκατομμύρια ζεύγη άρθρου-περίληψης, τα οποία περιλαμβάνουν συνολικά 123 εκατομμύρια λέξεις, με το λεξιλόγιο να αποτελείται από 119.000 διακριτές λέξεις. Το μέσο μήκος ενός άρθρου είναι $31,4$ λέξεις και το μέσο μήκος μιας περίληψης είναι $8,3$ λέξεις. Από το αρχικό σύνολο δεδομένων έχουν επιλεγεί τυχαία 2.000 ζεύγη άρθρου-περίληψης, που αποτελούν το σύνολο ελέγχου (test set, για την αξιολόγηση του συστήματος), και άλλα 2.000 ζεύγη, τα οποία έχουν επιλεγεί με παρόμοιο τρόπο για τη δημιουργία του συνόλου επικύρωσης (validation set, για τη βελτιστοποίηση των παραμέτρων ή την παρακολούθηση της διαδικασίας εκπαίδευσης). Ο τρόπος επιλογής που ακολουθήσαμε τόσο για το συνόλου ελέγχου όσο για το σύνολο επικύρωσης αποτελεί μια συνηθισμένη πρακτική στις σχετικές εργασίες [104, 60, 61]. Μετά την αφαίρεση των προαναφερόμενων παραδειγμάτων χρήσης από το αρχικό σύνολο δεδομένων, τα ζεύγη κειμένου-περίληψης που απομένουν αποτελούν το σύνολο εκπαίδευσης (training set).

Σύνολο δεδομένων *DUC 2004*: Το δεύτερο σύνολο δεδομένων, γνωστό ως *DUC 2004* [106], περιέχει 500 άρθρα ειδησεογραφικού περιεχομένου μαζί με τέσσερις περιλήψεις αναφοράς για κάθε άρθρο. Οι περιλήψεις είναι γραμμένες από αναγνώστες των άρθρων. Το σύνολο

δεδομένων έχει υποβληθεί σε προ-επεξεργασία, διατηρώντας μόνο την πρώτη περίοδο των άρθρων και μειώνοντας τις περιλήψεις σε μέγιστο μήκος 75 χαρακτήρων η κάθε μία, όπως αναφέρεται στις σχετικές εργασίες [104, 59, 63]. Με δεδομένο ότι το σύνολο δεδομένων DUC περιέχει πολύ λίγα παραδείγματα χρήσης για την εκπαίδευση ενός μοντέλου βαθιάς μάθησης, χρησιμοποιείται αποκλειστικά για σκοπούς αξιολόγησης (δηλ., χρησιμοποιείται ως σύνολο ελέγχου) [63, 59].

Σύνολο δεδομένων *CNN/DailyMail*: Το τελευταίο σύνολο δεδομένων, *CNN/DailyMail* [107], χρησιμοποιείται επίσης ευρέως για την αξιολόγηση συστημάτων αυτόματης *ΠΚ* και αποτελείται από άρθρα ή ιστορίες και τις περιλήψεις τους, οι οποίες έχουν έκταση πολλών προτάσεων. Στα πειράματα χρησιμοποιείται η μη ανώνυμη έκδοση του συνόλου δεδομένων και ακολουθούνται τα βήματα προ-επεξεργασίας που προτείνονται από τους δημιουργούς του συνόλου δεδομένων [60]. Μετά το στάδιο της προ-επεξεργασίας, το σύνολο δεδομένων περιέχει 287.227 ζεύγη κείμενου-περίληψης ως σύνολο εκπαίδευσης, 11.490 παραδείγματα χρήσης ως σύνολο ελέγχου και 13.368 δείγματα ως σύνολο επικύρωσης. Τα άρθρα περιορίστηκαν σε μήκος 400 λέξεων και οι περιλήψεις σε μήκος 100 λέξεων, ακολουθώντας την τυπική διαδικασία χρήσης αυτού του συνόλου δεδομένων για την αξιολόγηση μοντέλων βαθιάς μάθησης [61, 108, 72]. Όταν το σύνολο δεδομένων πήρε την τελική του μορφή, το μέσο μήκος κειμένου και περίληψης προέκυψαν $386,42$ λέξεις και $61,08$ λέξεις, αντίστοιχα.

Τέλος, πρέπει να σημειωθεί, ότι η παρούσα εργασία διατηρεί τη μορφή των συνόλων δεδομένων όπως χρησιμοποιήθηκαν σε σχετικές προσεγγίσεις. Επομένως, είναι δυνατή μια άμεση σύγκριση μεταξύ των πειραματικών αποτελεσμάτων που προέκυψαν με εκείνα των άλλων προσεγγίσεων.

3.4.2 Μετρικές αξιολόγησης: Το σύνολο μετρικών Rouge

Για την αξιολόγηση της αυτόματης *ΠΚ* χρησιμοποιείται ευρέως το σύνολο μετρικών *Rouge* [109]. Πιο συγκεκριμένα, στην παρούσα εργασία χρησιμοποιούμε τις μετρικές $Rouge_1$ (βαθμός επικάλυψης λέξεων), $Rouge_2$ (επικάλυψη δύο διαδοχικών λέξεων) και $Rouge_L$ (η μεγαλύτερη κοινή ακολουθία).

Αναλυτικότερα, η μετρική $Rouge_N$ (π.χ., $Rouge_1$ ή $Rouge_2$) εξετάζει την επικάλυψη N λέξεων μεταξύ μιας εκτιμώμενης περίληψης και μιας περίληψης αναφοράς. Η μετρική αυτή περιλαμβάνει τις παραλλαγές εκτίμησης της ανάκλησης ($Rouge_N(recall)$), της ακρίβειας ($Rouge_N(precision)$) και του αρμονικού μέσου ($Rouge_N(f1)$), οι οποίες παρουσιάζονται στις Εξισώσεις 3.20.

$$
\begin{aligned}
Rouge_N(recall) &= \frac{|Ngrams_{Sr} \cap Ngrams_{Se}|}{|Ngrams_{Sr}|} \\
Rouge_N(precision) &= \frac{|Ngrams_{Sr} \cap Ngrams_{Se}|}{|Ngrams_{Se}|} \\
Rouge_N(f1) &= \frac{2 \cdot Rouge_N(precision) \cdot Rouge_N(recall)}{Rouge_N(precision) + Rouge_N(recall)}
\end{aligned}
\tag{3.20}
$$

όπου $Ngrams_{Sr}$ είναι το σύνολο των N διαδοχικών λέξεων στην περίληψη αναφοράς και $Ngrams_{Se}$ είναι το σύνολο των N διαδοχικών λέξεων στην εκτιμώμενη περίληψη.

Η μετρική $Rouge_L$, επίσης, συναντάται σε τρεις παραλλαγές σύμφωνα με τις Εξισώσεις 3.21.

$$\begin{aligned} Rouge_L(recall) &= \frac{LCS(S_r, S_e)}{|S_r|} \\ Rouge_L(precision) &= \frac{LCS(S_r, S_e)}{|S_e|} \\ Rouge_L(f1) &= \frac{2 \cdot Rouge_L(precision) \cdot Rouge_L(recall)}{Rouge_L(precision) + Rouge_L(recall)} \end{aligned} \tag{3.21}$$

όπου η μετρική $Rouge_L(recall)$ υπολογίζει το ποσοστό της μεγαλύτερης κοινής ακολουθίας λέξεων $LCS(S_r, S_e)$ μεταξύ της περίληψης αναφοράς S_r και της εκτιμώμενης περίληψης S_e ως προς το πλήθος των λέξεων της περίληψης αναφοράς $|S_r|$. Η μετρική $Rouge_L(precision)$ υπολογίζει το ποσοστό της μεγαλύτερης κοινής ακολουθίας λέξεων $LCS(S_r, S_e)$ ως προς το πλήθος των λέξεων της εκτιμώμενης περίληψης $|S_e|$.

Ειδικότερα, στις μετρήσεις για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail* χρησιμοποιείται ο αρμονικός μέσος (f_1) των τριών προαναφερόμενων μετρικών του πακέτου *Rouge*. Στην περίπτωση του συνόλου δεδομένων *DUC 2004*, χρησιμοποιείται η *ανάκληση* (*recall*) των μετρικών *Rouge*. Αυτή είναι η κοινή πρακτική κατά την αξιολόγηση της αυτόματης *ΠΚ* με χρήση των προαναφερόμενων συνόλων δεδομένων, σύμφωνα με τις σχετικές εργασίες [60, 59, 63].

3.4.3 Πειραματική διαδικασία και βελτιστοποίηση παραμέτρων

Διανυσματική αναπαράσταση λέξεων

Για την διανυσματική αναπαράσταση λέξεων ανεξάρτητη των συμφραζομένων χρησιμοποιείται το μοντέλο *Word2Vec* [82] αρχιτεκτονικής *CBOW (continuous bag of words)*. Το μοντέλο αυτό εκπαιδεύεται με χρήση των δεδομένων εκπαίδευσης των χρησιμοποιούμενων συνόλων δεδομένων *Gigaword* και *CNN/DailyMail* (δηλ., προκύπτουν δύο εκπαιδευμένα μοντέλα διανυσματικής αναπαράστασης λέξεων, ένα για κάθε σύνολο δεδομένων). Κατά τη διάρκεια εκπαίδευσης του μοντέλου *Word2Vec*, το μέγεθος του παραθύρου ορίζεται σε 5 και η εκπαίδευση διαρκεί 10 εποχές. Ο ρυθμός μάθησης μεταβάλλεται σταδιακά από $0,025$ έως $0,001$ κατά τη διαδικασία εκπαίδευσης. Τα διανύσματα των λέξεων έχουν 300 διαστάσεις. Στην περίπτωση του μοντέλου αρχιτεκτονικής προ-εκπαιδευμένου κωδικοποιητή χρησιμοποιούνται ενσωματώσεις λέξεων εξαρτώμενες από τα συμφραζόμενα τύπου *BERT*, όπως έχει περιγραφεί στην Ενότητα 3.3.4.

Εκπαίδευση μοντέλων μηχανικής μάθησης

Εκπαίδευση μοντέλου κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής: Οι παράμετροι του μοντέλου *ΚΑΜΠ* (Ενότητα 3.3.1) βελτιστοποιούνται για το σύνολο επικύρωσης του συνόλου δεδομένων *Gigaword*. Η βελτιστοποίηση των παραμέτρων έγινε μέσα από πειραματική διαδικασία (με αναζήτηση πλέγματος) και οι επιλογές των παραμέτρων ακολουθούν. Η αμφίδρομη μονάδα *LSTM* του κωδικοποιητή αποτελείται από δύο στρώματα ίδιας διάστασης το καθένα (200 διαστάσεων), ενώ ο αποκωδικοποιητής περιλαμβάνει ένα στρώμα μονάδας *LSTM* μονής κατεύθυνσης, διάστασης, επίσης, 200. Το μέγεθος δέσμης των παραδειγμάτων

εκπαίδευσης (batch size) έχει οριστεί σε 64 για το σύνολο δεδομένων *Gigaword* και σε 32 για το σύνολο δεδομένων *CNN/DailyMail*. Τα δεδομένα εκπαίδευσης για κάθε εποχή παρέχονται στο δίκτυο βαθιάς μάθησης με τυχαία σειρά. Ο ρυθμός μάθησης ορίζεται αρχικά σε $0,002$ και σταδιακά μειώνεται κατά 25% μετά από κάθε εποχή εκπαίδευσης. Επιπλέον, ως μέθοδος βελτιστοποίησης χρησιμοποιείται ο αλγόριθμος *εκτίμησης προσαρμοστικών ροπών* (*adaptive moment estimation - Adam*) [21] και η μέθοδος *κατάβασης κλίσης με ψαλιδισμό* (*gradient norm clipping*) [110], με σκοπό την ελαχιστοποίηση της συνάρτησης αρνητικής λογαριθμικής πιθανοφάνειας [111], η οποία χρησιμοποιείται ως *συνάρτηση σφάλματος* (*loss function*) στην εκπαίδευση του δικτύου. Χρησιμοποιείται, επίσης, η τεχνική απόρριψης συνδέσεων δικτύου (*dropout*) με πιθανότητα απόρριψης $p = 20\%$. Στην περίπτωση του συνόλου δεδομένων *CNN/DailyMail*, το λεξιλόγιο περιορίζεται σε 150.000 λέξεις (χρησιμοποιώντας τις πιο συχνές λέξεις του συνόλου εκπαίδευσης) ενώ για το σύνολο δεδομένων *Gigaword* δεν τίθεται κάποιος περιορισμός στο πλήθος των λέξεων. Η εκπαίδευση του μοντέλου πραγματοποιήθηκε με χρήση μονάδας γραφικής επεξεργασίας (*graphics processing unit - GPU*) για την επιτάχυνση της διαδικασίας εκπαίδευσης μέσω του παραλληλισμού των υπολογισμών. Η σύγκλιση των μοντέλων επιτεύχθηκε μετά από περίπου 15 εποχές εκπαίδευσης.

Εκπαίδευση του μοντέλου αντιγραφής λέξεων εκτός λεξιλογίου: Μία διαφορά μεταξύ του μοντέλου *ΑΛΕΛ* (Ενότητα 3.3.2) και του βασικού μοντέλου *ΚΑΜΠ* (Ενότητα 3.3.1) είναι, ότι ο κωδικοποιητής του μοντέλου *ΑΛΕΛ* αποτελείται από δύο στρώματα αμφίδρομων μονάδων *LSTM* διάστασης 256 το καθένα, ενώ ο αποκωδικοποιητής χρησιμοποιεί ένα στρώμα μονάδας *LSTM* μονής κατεύθυνσης διάστασης 512. Τα υπόλοιπα στοιχεία του δικτύου παραμένουν ίδια με εκείνα του μοντέλου *ΚΑΜΠ*, καθώς το δίκτυο αυτό αποτελεί μια επέκταση της αρχιτεκτονικής *ΚΑΜΠ*.

Εκπαίδευση μοντέλου ενισχυτικής μάθησης: Ο πράκτορας του μοντέλου *ΕΜ* (Ενότητα 3.3.3) αποτελεί μια επέκταση του μοντέλου *ΑΛΕΛ* με τα περισσότερα στοιχεία του να διατηρούνται ίδια με εκείνα του μοντέλου *ΑΛΕΛ*, εκτός από κάποιες μικρές διαφορές. Ο ρυθμός μάθησης καθορίζεται σε 10^{-4} και το μέγεθος δέσμης παραδειγμάτων εκπαίδευσης τίθεται σε 32 και 16 για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*, αντίστοιχα. Οι διαστάσεις των στρωμάτων *LSTM* είναι ίδιες με εκείνες του μοντέλου *ΑΛΕΛ* που περιγράφεται παραπάνω. Για τον υπολογισμό της ανταμοιβής χρησιμοποιείται η μετρική $Rouge_L$. Σχετικά με τις υπόλοιπες παραμέτρους, ισχύουν οι ίδιες υποθέσεις με τις παραπάνω αρχιτεκτονικές.

Εκπαίδευση μοντέλου μετασχηματιστών: Σύμφωνα με την αρχιτεκτονική του μοντέλου *ΜΣ* (Ενότητα 3.3.4), ο κωδικοποιητής και ο αποκωδικοποιητής αποτελούνται από μια στοίβα 6 πανομοιότυπων στρωμάτων ο καθένας. Η διάσταση του μοντέλου είναι $d_{model} = 512$ και η διάσταση του εσωτερικού στρώματος του δικτύου πρόσθιας τροφοδότησης έχει οριστεί σε $d_{ffn} = 2048$. Υποθέτουμε $h = 8$ κεφαλές (δηλαδή, 8 παράλληλα επίπεδα του μηχανισμού προσοχής που μειώνουν τη διάσταση του μοντέλου για κάθε στρώμα προσοχής σε $512/8 = 64$). Ως μέθοδος βελτιστοποίησης χρησιμοποιείται η *Adam* με παραμέτρους $\beta_1 = 0.9$ και $\beta_2 = 0.99$. Ο ρυθμός μάθησης lr προσαρμόζεται κατά τη διάρκεια της εκπαίδευσης σύμφωνα με την Εξίσωση 3.22 (δηλ., αύξηση του ρυθμού μάθησης για τα πρώτα $warmupSteps$ βήματα εκπαίδευσης και στη συνέχεια μείωση του), όπου $warmupSteps = 5000$ και $a = 0,05$.

$$lr = a \cdot \min\{step^{-0.5},\ step \cdot warmupSteps^{-1.5}\} \tag{3.22}$$

Η μέθοδος απόρριψης συνδέσεων dropout εφαρμόζεται με πιθανότητα απόρριψης $p = 10\%$ και το

μέγεθος δέσμης των παραδειγμάτων εκπαίδευσης ορίζεται σε 64 και 16 για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*, αντίστοιχα.

Εκπαίδευση μοντέλου μετασχηματιστών προ-εκπαιδευμένου κωδικοποιητή: Η διαφορά μεταξύ των μοντέλων *ΜΣ* και *ΜΣΠΚ* (Ενότητα 3.3.4) είναι ότι η τελευταία αρχιτεκτονική χρησιμοποιεί έναν προ-εκπαιδευμένο κωδικοποιητή τύπου *BERT* και έναν αποκωδικοποιητή τύπου μετασχηματιστών 6 στρωμάτων. Τα αρχικά βάρη του κωδικοποιητή έχουν οριστεί σύμφωνα με το προ-εκπαιδευμένο μοντέλο *BERT*, ενώ τα βάρη του αποκωδικοποιητή αρχικοποιούνται τυχαία. Επιπλέον, για σταθερότητα στη σύγκλιση του μοντέλου, χρησιμοποιούνται δύο διαδικασίες βελτιστοποίησης τύπου *Adam* (με παραμέτρους $\beta_1 = 0.9$ και $\beta_2 = 0.99$), ξεχωριστά για τον κωδικοποιητή και τον αποκωδικοποιητή. Κάθε διαδικασία βελτιστοποίησης θέτει διαφορετικό ρυθμό μάθησης σύμφωνα με την Εξίσωση 3.22, όπου $a_{enc} = 10^{-3}$, $a_{dec} = 0.1$, $warmupSteps_{enc} = 10.000$ και $warmupSteps_{dec} = 5.000$. Οι διαφορετικοί ρυθμοί μάθησης στοχεύουν σε μια ομαλή προσαρμογή των βαρών του δικτύου, λαμβάνοντας υπόψη ότι ο προ-εκπαιδευμένος κωδικοποιητής χρειάζεται μικρότερο ρυθμό μάθησης και ομαλότερη μείωση του ρυθμού μάθησης σε σύγκριση με τον αποκωδικοποιητή, ο οποίος εκπαιδεύεται από την αρχή. Η μέθοδος αυτή βοηθάει στην αποφυγή του φαινομένου της υπερ-προσαρμογής του κωδικοποιητή και της υπο-προσαρμογής (uderfitting) του αποκωδικοποιητή (ή το αντίστροφο), όπως αναφέρεται στην εργασία [66].

Παραγωγή εκτιμώμενων περιλήψεων

Για τη μεγιστοποίηση της κατανομής πιθανοτήτων στο λεξιλόγιο εξόδου της εκτιμώμενης περίληψης, χρησιμοποιούμε τον αλγόριθμο αναζήτησης δέσμης (beam search) ο οποίος έχει παρουσιαστεί στην Ενότητα 3.3.1. Ο αλγόριθμος αυτός χρησιμοποιείται σε όλα τα μοντέλα μηχανικής μάθησης στο στάδιο εκτίμησης μια περίληψης ενός νέου κειμένου εισόδου. Το πλάτος δέσμης του αλγόριθμου ορίζεται σε 4.

3.4.4 Πειραματικά αποτελέσματα

Τα αποτελέσματα των μετρήσεων αναφέρονται στον Πίνακα 3.1, ο οποίος περιλαμβάνει τις τιμές των μετρικών *Rouge* για τα τρία σύνολα δεδομένων. Οι μετρήσεις αφορούν τα πέντε μοντέλα βαθιάς μάθησης που περιγράφηκαν παραπάνω και άλλων αντίστοιχων προσεγγίσεων της σχετικής βιβλιογραφίας. Στις περιπτώσεις των σχετικών προσεγγίσεων, οι οποίες δεν χρησιμοποιούν όλα τα σύνολα δεδομένων, υπάρχουν παύλες στον πίνακα των αποτελεσμάτων για τα μη χρησιμοποιούμενα σύνολα δεδομένων.

3.4.5 Περιγραφή και ερμηνεία αποτελεσμάτων

Σύμφωνα με τα αποτελέσματα των μετρήσεων (Πίνακας 3.1), παρατηρούμε ότι το βασικό μοντέλο αρχιτεκτονικής *ΚΑΜΠ* υπολείπεται σε επιδόσεις σε σχέση με τα υπόλοιπα μοντέλα, τα οποία είναι περισσότερο προηγμένα σύμφωνα με όσα αναφέρθηκαν παραπάνω. Με την εισαγωγή του μηχανισμού αντιγραφής *ΛΕΛ* και την αντιμετώπιση του προβλήματος της επανάληψης των λέξεων στην έξοδο, παρατηρούμε ότι βελτιώθηκε αρκετά η επίδοση του μοντέλου *ΑΛΕΛ* σε σύγκριση με

Πίνακας 3.1: Οι τιμές επίδοσης *Rouge* για τα δίκτυα βαθιάς μάθησης κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής (*ΚΑΜΠ*), αντιγραφής λέξεων εκτός λεξιλογίου (*ΑΛΕΛ*), ενισχυτικής μάθησης (*ΕΜ*), μετασχηματιστών (*ΜΣ*), μετασχηματιστών με προ-εκπαιδευμένο κωδικοποιητή (*ΜΣΠΚ*), καθώς και για άλλες προσεγγίσεις της σχετικής βιβλιογραφίας (δοκιμή-t: $p_{value} < 0.01$).

	Gigaword			*DUC 2004*			*CNN/DailyMail*		
Μοντέλο	$Rouge_1$	$Rouge_2$	$Rouge_L$	$Rouge_1$	$Rouge_2$	$Rouge_L$	$Rouge_1$	$Rouge_2$	$Rouge_L$
ΚΑΜΠ	36.12	15.36	33.96	26.72	8.16	24.36	36.88	15.53	29.14
ΑΛΕΛ	39.33	18.31	37.10	27.45	09.72	25.60	41.40	19.26	29.56
ΕΜ	40.57	19.72	38.36	27.82	09.79	25.75	41.67	19.34	29.42
ΜΣ	40.18	18.96	37.26	27.43	09.75	25.65	41.08	18.12	33.37
ΜΣΠΚ	40.95	19.78	38.61	27.91	09.85	26.05	41.79	19.38	35.62
RAS-Elman [59] (*ΚΑΜΠ*)	33.78	15.97	31.15	28.97	8.26	24.06	-	-	-
Words-lvt2k-1sent [60] (*ΚΑΜΠ*)	34.97	17.17	32.70	28.35	9.46	24.59	35.46	13.30	32.65
point.-gener.+cov. [61] (*ΑΛΕΛ*)	-	-	-	-	-	-	39.53	17.28	36.38
RL+intra-attention [78] (*ΕΜ*)	-	-	-	-	-	-	41.16	15.75	39.08
SAGCopy-Indegree-1 [68] (*ΜΣ*)	38.84	20.39	36.27	-	-	-	-	-	-
SAGCopy-Outdegree [68] (*ΜΣ*)	-	-	-	-	-	-	42.53	19.92	39.44
BertSumAbs [66] (*ΜΣ*)	-	-	-	-	-	-	41.72	19.39	38.76

τη βασική αρχιτεκτονική. Η επίδοση βελτιώνεται περαιτέρω με χρήση του μοντέλου *ΜΣ*, ενώ οι προσεγγίσεις *ΕΜ* και *ΜΣΠΚ* παρουσιάζουν τις καλύτερες επιδόσεις σε όρους μετρικών *Rouge*.

Σε σύγκριση με τις άλλες σχετικές προσεγγίσεις (Πίνακας 3.1) παρατηρούμε ότι η αρχιτεκτονική *ΚΑΜΠ* ξεπερνά ή έχει παρόμοιες επιδόσεις με τα αντίστοιχα μοντέλα (*RAS-Elman* και *Words-lvt2k-1sent*) που υλοποιούν παρόμοια αρχιτεκτονική. Το μοντέλο *ΑΛΕΛ*, για το σύνολο δεδομένων *CNN/DailyMail*, ξεπερνά σε επιδόσεις το αντίστοιχο σχετικό μοντέλο (*pointer-generator+coverage*) για τις μετρικές $Rouge_1$ και $Rouge_2$, ενώ υπολείπεται σε επιδόσεις στη μετρική $Rouge_L$. Το ίδιο με το μοντέλο *ΑΛΕΛ* ισχύει και για το μοντέλο *ΕΜ* συγκρινόμενο με το αντίστοιχο μοντέλο της βιβλιογραφίας. Τα μοντέλα *ΜΣ* και *ΜΣΠΚ* παρουσιάζουν τις καλύτερες επιδόσεις και ξεπερνάνε για κάποιες μετρικές τις αντίστοιχες προσεγγίσεις της βιβλιογραφίας.

Να σημειώσουμε εδώ, ότι ο σκοπός της μελέτης αυτής, η οποία περιλαμβάνει, διερεύνηση της βιβλιογραφίας, υλοποίηση, βελτιστοποίηση και έλεγχο της επίδοσης διαφορετικών μοντέλων μηχανικής μάθησης, δεν είναι η δημιουργία ενός μοντέλου που ξεπερνά σε επιδόσεις τις σχετικές προσεγγίσεις αλλά η μελέτη των προσεγγίσεων αυτών και η ανάδειξη των πλεονεκτημάτων ή μειονεκτημάτων που αυτές παρουσιάζουν συγκρινόμενες μεταξύ τους. Η σύγκριση με τις σχετικές εργασίες γίνεται για να επιβεβαιώσουμε ότι τα μοντέλα μηχανικής μάθησης που εξετάζουμε παρουσιάζουν επιδόσεις αντίστοιχες με αυτές των πλέον σύγχρονων προσεγγίσεων που συναντάμε στη σχετική βιβλιογραφία.

Συμπληρωματικά με αυτά που αναφέρθηκαν παραπάνω, να σημειώσουμε ότι το μοντέλο *ΜΣ*, το οποίο αποτελεί μια διαφορετική προσέγγιση από τα άλλα μοντέλα, καθώς δε χρησιμοποιεί αναδρομή, φαίνεται ότι υπερέχει σε σχέση με τις αρχιτεκτονικές *ΚΑΜΠ* και *ΑΛΕΛ* που βασίζονται σε αναδρομικά νευρωνικά δίκτυα. Επίσης, το μοντέλο *ΜΣΠΚ* παρουσιάζει πλεονέκτημα σε σχέση με το μοντέλο *ΜΣ* λόγω του προ-εκπαιδευμένου κωδικοποιητή και της περαιτέρω προσαρμογής των βαρών μέσω της διαδικασίας εκπαίδευσης. Αντίστοιχες επιδόσεις με το μοντέλο *ΜΣΠΚ* παρουσιάζει και το μοντέλο *ΕΜ*, το οποίο βασίζεται στη μεγιστοποίηση της μετρικής αξιολόγησης

$Rouge_L$.

3.5 Συμπεράσματα και προοπτικές επέκτασης

Εξετάστηκαν πέντε αρχιτεκτονικές νευρωνικών δικτύων για την αυτόματη περίληψη κειμένου. Η διερεύνηση αυτή αφορά την υλοποίηση πέντε διαφορετικών αρχιτεκτονικών βαθιάς μάθησης, τη βελτιστοποίηση των παραμέτρων και τον έλεγχο της επίδοσης, καθώς και τη σύγκριση της επίδοσης με άλλες συναφείς προσεγγίσεις της σχετικής βιβλιογραφίας. Η πειραματική διαδικασία, που διεξήχθη, με χρήση τριών δημοφιλών συνόλων δεδομένων, έδειξε ότι οι αρχιτεκτονικές που βασίζονται σε μετασχηματιστές παρουσιάζουν πλεονεκτήματα έναντι των υπολοίπων, καθώς πρόκειται για απλά μοντέλα που δε χρησιμοποιούν αναδρομή ή συνέλιξη, αλλά βασίζονται κυρίως σε έναν κατανεμημένο μηχανισμό ενδο-προσοχής που έχει τη δυνατότητα παραλληλισμού των υπολογισμών. Αντίστοιχες επιδόσεις με τα μοντέλα που βασίζονται σε *ΜΣ* παρουσιάζει και το μοντέλο *ΕΜ*, το οποίο χρησιμοποιεί ως πράκτορα ένα αναδρομικό νευρωνικό δίκτυο τύπου κωδικοποιητή-αποκωδικοποιητή, αντίστοιχο με εκείνο του μοντέλου *ΑΛΕΛ*, που μέσω της αλληλεπίδρασης με το περιβάλλον επιδιώκει τη βελτιστοποίηση μιας ανταμοιβής, ο υπολογισμός της οποίας βασίζεται σε συγκεκριμένη μετρική αξιολόγησης της αυτόματης *ΠΚ*.

Με δεδομένο ότι το μοντέλο *ΕΜ* παρουσιάζει βελτιωμένη επίδοση σε σύγκριση με την αρχιτεκτονική *ΑΛΕΛ*, η οποία χρησιμοποιείται ως πράκτορας στο δίκτυο *ΕΜ*, ως μελλοντική εργασίας θα μπορούσε να συνδυαστεί η προσέγγιση *ΕΜ* με άλλα είδη πρακτόρων, όπως είναι ένα δίκτυο *ΜΣ*, για να διερευνηθεί αν αυτή η προοπτική θα μπορούσε να επιφέρει επιπλέον βελτίωση. Τέλος, οι προσεγγίσεις μηχανικής μάθησης θα μπορούσαν να συνδυαστούν με σημασιολογικές τεχνικές και γενικά με μεθοδολογία επεξεργασίας φυσικής γλώσσας, με σκοπό την περαιτέρω βελτίωση. Σε αυτή την κατεύθυνση, στη συνέχεια αυτής της διατριβής, εξετάζεται ο συνδυασμός προβλέψεων μηχανικής μάθησης με νέα μεθοδολογία που βασίζεται σε τεχνικές σημασιολογικού μετασχηματισμού και σημασιολογικής αναπαράστασης του περιεχομένου για την περαιτέρω βελτίωση της αυτόματης *ΠΚ*.

Κεφάλαιο 4

Αυτόματη περίληψη κειμένου με χρήση μηχανικής μάθησης και σημασιολογικών μετασχηματισμών περιεχομένου

4.1 Γενικά

Στο κεφάλαιο αυτό παρουσιάζεται ένα νέο πλαίσιο το οποίο συνδυάζει μεθοδολογία μηχανικής μάθησης και σημασιολογικών μετασχηματισμών περιεχομένου για την αυτόματη *ΠΚ* ενός εγγράφου με τη μέθοδο της παραγωγής κειμένου (*single document abstractive TS*). Οι σημασιολογικές τεχνικές που προτείνονται βασίζονται, κυρίως, σε πόρους γνώσης, όπως ταξινομίες εννοιών, προσεγγίσεις αποσαφήνισης έννοιας λέξεων (*Word Sense Disambiguation - WSD*), αναγνώριση ονοματικών οντοτήτων (*named-entity recognition - NER*) και μεθοδολογία σημασιολογικής γενίκευσης του περιεχομένου, προκειμένου να γίνει εφικτός ο μετασχηματισμός των δεδομένων, με σκοπό την αποδοτικότερη εκπαίδευση μοντέλων μηχανικής μάθησης που οδηγεί στη βελτίωση της ακρίβειας των εκτιμήσεων. Πιο συγκεκριμένα, συνδυάζονται χαρακτηριστικά και πτυχές μεθοδολογιών που βασίζονται στη δομή, στη σημασιολογία και στη μηχανική μάθηση [9], τις οποίες συναντάμε κυρίως ως ξεχωριστές προσεγγίσεις στη σχετική βιβλιογραφία (όπως αναφέρεται στην παρουσίαση των σχετικών εργασιών στην Ενότητα 4.2), ειδικά για τις προσεγγίσεις βαθιάς μάθησης. Στο πλαίσιο αυτό, η παρούσα εργασία προτείνει μια νέα προσέγγιση, η οποία αντιμετωπίζει το πρόβλημα των σπάνιων λέξεων ή των λέξεων εκτός λεξιλογίου (*ΛΕΛ*) και ταυτόχρονα επιτυγχάνει τη βελτίωση της επίδοσης μοντέλων βαθιάς μάθησης, μέσω ενός πλαισίου σημασιολογικών μετασχηματισμών του περιεχομένου, το οποίο παρουσιάζεται, αναλύεται και αξιολογείται.

Σε αυτή την κατεύθυνση, η προτεινόμενη μεθοδολογία περιλαμβάνει τρεις διακριτές φάσεις για την εκτίμηση της τελικής περίληψης, οι οποίες είναι: (*i*) προ-επεξεργασία για τη σημασιολογική γενίκευση του περιεχομένου, (*ii*) προβλέψεις μηχανικής μάθησης και (*iii*) μετά-επεξεργασία για την διαμόρφωση της εκτιμώμενης περίληψης.

Το πρώτο στάδιο επιδιώκει τη σημασιολογική γενίκευση του περιεχομένου. Για τον σκοπό αυτό αξιοποιούνται πόροι γνώσης όπως οντολογίες (οι οποίες μπορεί να περιγράφουν ένα συναφές με τα δεδομένα πεδίο γνώσης) ή ταξινομίες εννοιών, μεθοδολογία σημασιολογικής αποσαφήνισης λέξεων και αναγνώριση ονοματικών οντοτήτων (προκειμένου να ανακτηθούν προκαθορισμένες ονοματικές οντότητες από το αρχικό κείμενο). Στη συνέχεια, το γενικευμένο κείμενο παρέχεται σε ένα μοντέλο μηχανικής μάθησης, το οποίο εκπαιδεύεται με σκοπό την εκτίμηση μιας ακολουθίας λέξεων που αποτελεί την περίληψη ενός αρχικού κειμένου. Η εκτιμώμενη περίληψη ενός κειμένου γενικευμένης μορφής αποτελεί και εκείνη μια περίληψη γενικευμένης μορφής. Μετά την εφαρμογή της τρίτης φάσης, της μετα-επεξεργασίας, η περίληψη αποκτά την τελική της μορφή. Το στάδιο της μετα-επεξεργασίας βασίζεται σε πόρους γνώσης αντίστοιχους με αυτούς που χρησιμοποιούνται στη φάση της προ-επεξεργασίας και, επιπλέον, αξιοποιούνται ευρεστικές μέθοδοι που βασίζονται σε διαδικασίες μετρήσεων ομοιότητας κειμένου, με σκοπό την αντιστοίχιση των γενικευμένων εννοιών της γενικευμένης περίληψης με συγκεκριμένες, για τη διαμόρφωση της τελικής περίληψης.

Στο στάδιο των προβλέψεων μηχανικής μάθησης, συνδυάζεται το προτεινόμενο πλαίσιο σημασιολογικών μετασχηματισμών με τις πέντε αρχιτεκτονικές νευρωνικών δικτύων βαθιάς μάθησης που παρουσιάστηκαν με λεπτομέρεια στο Κεφάλαιο 3, οι οποίες είναι οι εξής: (*i*) μοντέλο κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής, (*ii*) δίκτυο με μηχανισμό αντιγραφής λέξεων εκτός λεξιλογίου, (*iii*) μοντέλο ενισχυτικής μάθησης, (*iv*) δίκτυο μετασχηματιστών και (*v*) μοντέλο μετασχηματιστών προεκπαιδευμένου κωδικοποιητή.

Για την αξιολόγηση της προτεινόμενης προσέγγισης πραγματοποιήθηκε μια εκτεταμένη πειραματική διαδικασία χρησιμοποιώντας τρία ευρέως χρησιμοποιούμενα σύνολα δεδομένων (*Gigaword* [105], *DUC 2004* [106] και *CNN/DailyMail* [107]). Η αξιολόγηση οδήγησε σε θετικά αποτελέσματα, καθώς η παρούσα προσέγγιση βελτιώνει την επίδοση των μοντέλων μηχανικής μάθησης και αντιμετωπίζει αποτελεσματικά το πρόβλημα των σπάνιων λέξεων ή των λέξεων εκτός λεξιλογίου. Η πειραματική διαδικασία για την αξιολόγηση της προτεινόμενης μεθοδολογίας παρουσιάζεται στο Κεφάλαιο 5.

Το υπόλοιπο αυτού του κεφαλαίου οργανώνεται ως εξής: Η Ενότητα 4.2 παρουσιάζει τη σχετική βιβλιογραφία. Η Ενότητα 4.4 παρουσιάζει την αρχιτεκτονική της προτεινόμενης προσέγγισης, η οποία αναλύεται με λεπτομέρεια στις Ενότητες 4.5, 4.6 και 4.7, που παρουσιάζουν τις φάσεις της προ-επεξεργασίας, των προβλέψεων μηχανικής μάθησης και της μετα-επεξεργασίας, αντίστοιχα.

4.2 Σχετικές εργασίες

Οι πρώτες προσεγγίσεις που εμφανίστηκαν στο πεδίο της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου (*abstractive TS*) περιλαμβάνουν τεχνικές συμπίεσης προτάσεων [112, 113] για τη μείωση του μεγέθους των αρχικών προτάσεων του υποψήφιου προς περίληψη κειμένου, καθώς και μεθοδολογίες συγχώνευσης κειμένου [114, 115, 116, 117], για τον συνδυασμό παρόμοιων φράσεων ή προτάσεων του αρχικού κειμένου που οδηγεί σε μια συνοπτική εκδοχή του. Οι περισσότερες προσεγγίσεις, ωστόσο, βασίζονται είτε στη δομή είτε στη σημασιολογία [118], ενώ, τα τελευταία χρόνια, έχει προκύψει μια τρίτη κατηγορία που αφορά προσεγγίσεις παραγωγής κειμένου που βασίζονται σε προβλέψεις νευρωνικών δικτύων [9]. Πιο συγκεκριμένα, οι προσεγγίσεις που βασίζονται στη δομή εκμεταλλεύονται προκαθορισμένες δομές, όπως δέντρα,

οντολογίες, γραφήματα, κανόνες και πρότυπα για να δημιουργήσουν μια περίληψη. Οι προσεγγίσεις που βασίζονται σε σημασιολογικές τεχνικές, από την άλλη πλευρά, χρησιμοποιούν συστήματα παραγωγής φυσικής γλώσσας, βασίζονται σε σημασιολογικά γραφήματα, χρησιμοποιούν σχέσεις κατηγορήματος – ορίσματος και οντότητες πληροφοριών για τη δημιουργία της περίληψης, αξιοποιώντας τη σημασιολογική αναπαράσταση του αρχικού κειμένου. Τέλος, οι προσεγγίσεις νευρωνικών δικτύων βαθιάς μάθησης κυριαρχούν τα τελευταία χρόνια στην αυτόματη *ΠΚ* με τη μέθοδο της παραγωγής κειμένου, καθώς επιτυγχάνουν της καλύτερες επιδόσεις [64, 9], συγκρινόμενες με άλλες μεθοδολογίες της σχετικής βιβλιογραφίας.

Οι μέθοδοι που εστιάζουν στη δομή βασίζονται συνήθως σε ιεραρχικές οντολογίες ή ταξινομίες εννοιών ως πηγές γνώσης που χρησιμοποιούνται για τη σημασιολογική αναπαράσταση ενός εγγράφου και βοηθούν στην επίλυση ζητημάτων αμφισημίας [119]. Σε αυτές τις προσεγγίσεις, χρησιμοποιούνται βάσεις γνώσης για την παραγωγή του κειμένου των περιλήψεων [119], όπως το ηλεκτρονικό λεξικό *WordNet* [43, 44], η *DBPedia* [120] ή οντολογίες συγκεκριμένου πεδίου γνώσης. Έχουν αναπτυχθεί αρκετά τέτοια συστήματα που εστιάζουν σε οντολογικές αναπαραστάσεις με σκοπό την εξαγωγή εννοιών ή φράσεων-κλειδιά από ένα κείμενο για τη δημιουργία μιας περίληψης [121, 122, 123, 124]. Στην παρούσα εργασία, χρησιμοποιούνται προκαθορισμένες ταξινομίες εννοιών για τον σημασιολογικό μετασχηματισμό των δεδομένων και, πιο συγκεκριμένα, για τη γενίκευση του περιεχομένου (στη φάση της προ-επεξεργασίας) και την αντιστοίχιση των γενικευμένων εννοιών σε συγκεκριμένες (στη φάση της μετά-επεξεργασίας), όπως περιγράφεται με λεπτομέρεια στις επόμενες ενότητες.

Οι προσεγγίσεις που βασίζονται στη σημασιολογία χρησιμοποιούν μια σημασιολογική αναπαράσταση του κειμένου, η οποία αποτυπώνει σχέσεις μεταξύ οντοτήτων, προκειμένου να προσδιοριστούν σημαντικές προτάσεις, φράσεις ή εκφράσεις που συνδυάζονται κατάλληλα για να συνθέσουν την περίληψη. Συγκεκριμένα, οι σημασιολογικές μέθοδοι που βασίζονται σε γραφήματα μετατρέπουν το κείμενο σε μια αναπαράσταση γραφήματος, η οποία αποτυπώνει σημασιολογικές και συντακτικές σχέσεις [125, 126] των οντοτήτων του κειμένου. Συνήθως, η περίληψη δημιουργείται με την εύρεση σημαντικών οντοτήτων και εννοιών, είτε χρησιμοποιώντας τις σχέσεις ενός γραφήματος αναπαράστασης κειμένου, είτε μειώνοντας το μέγεθος ενός τέτοιου γραφήματος, εξαλείφοντας τον πλεονασμό ή απορρίπτοντας μη σημαντικές οντότητες [127]. Επιπλέον, οι μέθοδοι που βασίζονται σε στοιχεία πληροφορίας (*information items*) χρησιμοποιούν σχέσεις υποκειμένου, ρήματος και αντικειμένου για τη δημιουργία προτάσεων. Σε αυτή την περίπτωση, η περίληψη διαμορφώνεται με την κατάταξη των προτάσεων σύμφωνα με τα στοιχεία πληροφορίας που περιέχουν [128]. Επίσης, οι λύσεις που βασίζονται σε σχέσεις κατηγορήματος- ορίσματος (*predicate-argument relations*) ανακτούν δομές, όπως τριάδες της μορφής οντότητα-σχέση-οντότητα, οντότητα-σχέση-τιμή, οντότητα-ιδιότητα-τιμή ή και σχέσεις υποκειμένου-ρήματος-αντικειμένου, οι οποίες συνδυάζονται σημασιολογικά για να δημιουργήσουν την περίληψη [129, 130]. Με δεδομένο ότι στη σχετική βιβλιογραφία δεν εξετάζεται ο συνδυασμός των προσεγγίσεων αυτής της κατηγορίας με προσεγγίσεις μηχανικής μάθησης [9, 5], η παρούσα εργασία επιδιώκει να διερευνήσει αυτό το κενό, χρησιμοποιώντας μεθοδολογίες σημασιολογικών τεχνικών από κοινού με αρχιτεκτονικές νευρωνικών δικτύων προκειμένου να βελτιώσει την αποτελεσματικότητα των μοντέλων μηχανικής μάθησης. Η προτεινόμενη προσέγγιση χρησιμοποιεί ταξινομίες εννοιών και μεθοδολογίες σημασιολογικής αποσαφήνισης έννοιας λέξεων για τον μετασχηματισμό του αρχικού κειμένου σε μια γενικευμένη μορφή, που διευκολύνει τις προβλέψεις μηχανικής μάθησης.

Σε αντίθεση με τις μεθόδους που περιγράφονται παραπάνω, οι οποίες απαιτούν δύσκολες επιμέρους εργασίες, όπως εξαγωγή πληροφορίας, επιλογή περιεχομένου και παραγωγή φυσικής γλώσσας [64], οι προσεγγίσεις που βασίζονται σε αρχιτεκτονικές νευρωνικών δικτύων είναι ικανές να παράγουν περιλήψεις με τη χρήση ενός κατάλληλου μοντέλου μηχανικής μάθησης, χωρίς άλλη επεξεργασία φυσικής γλώσσας. Τέτοιες τεχνικές βασίζονται συχνά σε μοντέλα νευρωνικών δικτύων βαθιάς μάθησης, κυρίως, αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή [58], όπου ο κωδικοποιητής λαμβάνει μια ακολουθία λέξεων και ο αποκωδικοποιητής αποδίδει μια εκτιμώμενη ακολουθία λέξεων που αποτελεί την περίληψη. Τα δίκτυα αυτά, τα οποία περιγράφονται με λεπτομέρεια στο Κεφάλαιο 3, εκπαιδεύονται από άκρη-σε-άκρη με χρήση ενός συνόλου δεδομένων, το οποίο περιλαμβάνει μεγάλο πλήθος από ζεύγη κειμένου-περίληψης. Συνεπώς, το μοντέλο μηχανικής μάθησης εκπαιδεύεται να εκτιμά μια περίληψη ενός αρχικού κειμένου. Στην Ενότητα 3.2 που προηγήθηκε έχει περιγραφεί με μεγαλύτερη λεπτομέρεια η σχετική ερευνητική δουλειά, που έχει γίνει στον τομέα της αυτόματης *ΠΚ* με χρήση μοντέλων βαθιάς μάθησης.

Τα τελευταία χρόνια, οι προσεγγίσεις που βασίζονται σε αρχιτεκτονικές νευρωνικών δικτύων για την αυτόματη *ΠΚ* εξελίσσονται διαρκώς με σκοπό τη βελτίωσή τους, χωρίς, ωστόσο, να συνδυάζουν ή να επωφελούνται από άλλες τεχνικές επεξεργασίας φυσικής γλώσσας, όπως αυτές που αναφέρθηκαν παραπάνω. Για να συμπληρώσει αυτό το κενό, η παρούσα διατριβή προτείνει ένα νέο πλαίσιο, το οποίο συνδυάζει χαρακτηριστικά μεθοδολογίας που βασίζεται στη δομή, στη σημασιολογία και τη μηχανική μάθηση, για την αντιμετώπιση του προβλήματος των λέξεων εκτός λεξιλογίου ή των σπάνιων λέξεων και, γενικά, για τη βελτίωση της επίδοσης των προβλέψεων μηχανικής μάθησης. Σε αυτή την κατεύθυνση, η τρέχουσα εργασία προτείνει μεθοδολογία σημασιολογικών μετασχηματισμών του περιεχομένου, η οποία χρησιμοποιείται για τη βελτίωση των προβλέψεων μηχανικής μάθησης και τη διαμόρφωση των εκτιμώμενων περιλήψεων στην τελική τους μορφή, μέσω της σημασιολογικής αντιστοίχισης των γενικευμένων εννοιών με συγκεκριμένες. Επιπλέον, εξετάζεται η ευελιξία, η προσαρμοστικότητα και η ικανότητα γενίκευσης της μεθοδολογίας σε συνδυασμό με διαφορετικές προσεγγίσεις μηχανικής μάθησης. Τα συγκεκριμένα μοντέλα μηχανικής μάθησης που εξετάζονται έχουν παρουσιαστεί στο Κεφάλαιο 3.

4.3 Το παράδειγμα της ταξινόμησης εγγράφων μη ισορροπημένων δεδομένων

Σε προηγούμενη ερευνητική μας εργασία [131], η οποία, θα μπορούσαμε να πούμε, ότι αποτελεί την αφετηρία του σημασιολογικού πλαισίου που παρουσιάζεται σε αυτό το κεφάλαιο, επιδιώκουμε τον σημασιολογικό εμπλουτισμό εγγράφων μέσω μεθοδολογίας που βασίζεται στην ταξινόμηση των σημασιολογικών τους περιγραφών. Οι σημασιολογικές περιγραφές, ουσιαστικά, αποτελούν διανυσματικές αναπαραστάσεις των εγγράφων. Οι διαστάσεις των διανυσμάτων αυτών αντιστοιχούν σε χαρακτηριστικά (*features*) που αντιπροσωπεύουν οντότητες, είτε μιας οντολογίας (αναπαράστασης γνώσης ενός συγκεκριμένου πεδίου), είτε μιας ταξινομίας εννοιών. Στο πλαίσιο αυτό μπορούν να χρησιμοποιηθούν μια ή περισσότερες οντολογίες γνώσης ή ταξινομίες εννοιών, με την προϋπόθεση ότι έχουν ιεραρχική δομή και ενσωματώνουν σχέσεις μεταξύ οντοτήτων ή εννοιών τύπου μερωνυμίας-ολωνυμίας, υπωνυμίας-υπερωνυμίας ή γενικότερου-ειδικότερου. Στην

κατεύθυνση αυτή, τα χαρακτηριστικά εξάγονται από τα έγγραφα σύμφωνα με τη χρησιμοποιούμενη ταξινομία και τα έγγραφα ταξινομούνται σε κλάσεις που αντιστοιχούν σε οντότητες της χρησιμοποιούμενης ταξινομίας. Ο σκοπός της προσέγγισης αυτής είναι η ταξινόμηση των εγγράφων σε κλάσεις οι οποίες στη συνέχεια προστίθενται ως ετικέτες στα έγγραφα. Ο σημασιολογικός εμπλουτισμός προκύπτει από τη διαδικασία ταξινόμησης και την προσθήκη επιπρόσθετων ετικετών στα έγγραφα. Συνεπώς, η εργασία αυτή διερευνά ένα πρόβλημα ταξινόμησης εγγράφων.

Η σύνδεση του περιγραφόμενου παραδείγματος [131] με την παρούσα εργασία αφορά τη διαδικασία εξισορρόπησης των κλάσεων (*class balancing*) των παραδειγμάτων χρήσης που οδηγεί σε βελτίωση των επιδόσεων της ταξινόμησης. Με δεδομένο ότι χρησιμοποιούνται μοντέλα μηχανικής μάθησης που εκπαιδεύονται με τη μέθοδο της επιβλεπόμενης μάθησης, παρατηρήθηκε ότι οι κλάσεις με χαμηλή συχνότητα εμφανίσεων στο σύνολο εκπαίδευσης οδηγούν σε μειωμένες επιδόσεις ταξινόμησης. Από την άλλη πλευρά, η εξισορρόπηση των κλάσεων οδηγεί σε βελτίωση των επιδόσεων ταξινόμησης. Για την ισορρόπηση του συνόλου δεδομένων παρουσιάζεται ένας αλγόριθμος που αντικαθιστά τις κλάσεις μειωμένης συχνότητας εμφάνισης στο σύνολο εκπαίδευσης με κλάσεις υψηλότερης συχνότητας, οι οποίες όμως έχουν ευρύτερη σημασιολογική έννοια ή αποτελούν ολώνυμα ή υπερώνυμα των αρχικών κλάσεων. Πιο συγκεκριμένα, με δεδομένη μια ταξινομία κλάσεων, αν ένα παράδειγμα χρήσης ανήκει στην κλάση c_1 η οποία συνεπάγεται σημασιολογικά την κλάση c_2 (δηλ., $c_1 \models c_2$), τότε η c_1 μπορεί να αντικατασταθεί από την c_2 (π.χ. το *"αυτοκίνητο"* συνεπάγεται σημασιολογικά το *"όχημα"* με αποτέλεσμα το *"αυτοκίνητο"* να μπορεί να αντικατασταθεί από το *"όχημα"*).

Ο μετασχηματισμός αυτός των δεδομένων, ο οποίος αποτελεί ένα είδος σημασιολογικής γενίκευσης, επιφέρει βελτίωση στις επιδόσεις των μοντέλων ταξινόμησης, όπως διαπιστώθηκε μέσα από την πειραματική διαδικασία. Συνεπώς, η προσέγγιση αυτή αποτελεί μια αφετηρία για την σχεδίαση ενός πλαισίου σημασιολογικής γενίκευσης του περιεχομένου για την αυτόματη *ΠΚ*, όπως θα δούμε στη συνέχεια. Να αναφερθεί, επίσης, ότι η εξισορρόπηση των κλάσεων στην εργασία σημασιολογικού εμπλουτισμού [131], παρά τη βελτίωση των επιδόσεων της ταξινόμησης, οδήγησε σε απώλεια πληροφορίας, καθώς οι κλάσεις με χαμηλή συχνότητα εμφάνισης στο σύνολο εκπαίδευσης αντικαθίστανται με άλλες κλάσεις υψηλότερης συχνότητας, με αποτέλεσμα την οριστική απώλεια των πρώτων. Αυτό αποτέλεσε ένα πρόβλημα στην εργασία σημασιολογικού εμπλουτισμού και επιδιώκεται η αντιμετώπισή του στην παρούσα εργασία που αφορά την αυτόματη *ΠΚ*.Το πρόβλημα αυτό αντιμετωπίζεται κατά τη φάση της μετα-επεξεργασίας, η οποία αντιστοιχεί τις γενικευμένες έννοιες της εκτιμώμενης περίληψης με συγκεκριμένες, διαμορφώνοντας την τελική περίληψη. Συνεπώς, η παρούσα εργασία, με αφετηρία την εργασία σημασιολογικού εμπλουτισμού, που περιγράφηκε παραπάνω, εκμεταλλεύεται το πλεονέκτημα της γενίκευσης του περιεχομένου για πιο ακριβείς προβλέψεις μηχανικής μάθησης και ταυτόχρονα επιδιώκει την αποφυγή της απώλειας πληροφορίας, που δημιουργεί η σημασιολογική γενίκευση.

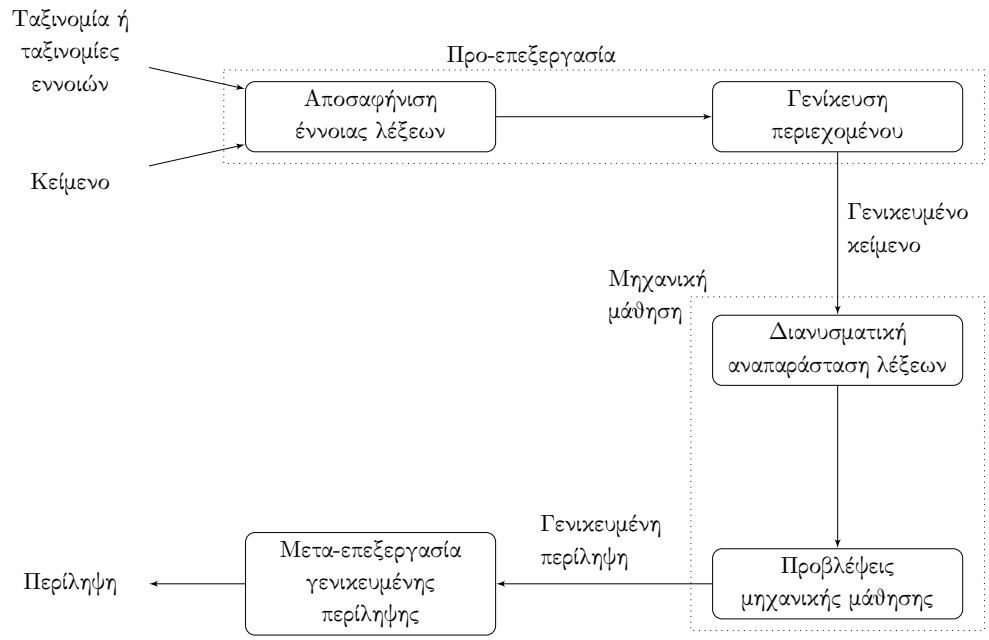
4.4 Εισαγωγή στην αρχιτεκτονική της προτεινόμενης προσέγγισης

Το Σχήμα 4.1 απεικονίζει το διάγραμμα ροής της προτεινόμενης μεθοδολογίας. Σύμφωνα με το διάγραμμα αυτό, η είσοδος στο σύστημα αποτελείται από ένα έγγραφο κειμένου, μαζί με μία

ταξινομία εννοιών T (ή ένα σύνολο από ταξινομίες εννοιών $T = \{T_1, T_2, ...\}$), ενώ η έξοδος είναι η εκτιμώμενη περίληψη του αρχικού κειμένου. Τα κύρια μέρη της αρχιτεκτονικής του συστήματος, όπως φαίνεται και στο σχήμα 4.1, είναι πέντε, τα οποία ξεκινάνε με τον μηχανισμό αποσαφήνισης της έννοιας των λέξεων ($ΑΕΛ$) που έχει σκοπό την αντιμετώπιση του προβλήματος της αμφισημίας. Αυτό είναι ένα σημαντικό βήμα για την εύστοχη εφαρμογή του σταδίου που ακολουθεί, το οποίο αποτελεί τη γενίκευση του περιεχομένου και αφορά την αντικατάσταση των λέξεων εκτός λεξιλογίου ($ΛΕΛ$) με συναφείς όρους εντός λεξιλογίου ή την αντικατάσταση των σπάνιων λέξεων ή των λέξεων χαμηλής συχνότητας εμφάνισης με όρους υψηλότερης συχνότητας. Και τα δύο προαναφερόμενα βήματα αποτελούν τη φάση της προ-επεξεργασίας, η οποία παρουσιάζεται με λεπτομέρεια στην Ενότητα 4.5.

Οι λεκτικές μονάδες που περιλαμβάνει το γενικευμένο κείμενο μετατρέπονται, στη συνέχεια, σε διανυσματικές αναπαραστάσεις ενός συνεχούς διανυσματικού χώρου. Μια ακολουθία τέτοιων διανυσμάτων, η οποία αντιστοιχεί σε μια ακολουθία λεκτικών μονάδων του αρχικού κειμένου, παρέχεται, στη συνέχεια, σε ένα μοντέλο μηχανικής μάθησης. Το μοντέλο αυτό, έχοντας εκπαιδευτεί σε ένα σύνολο από ζεύγη κειμένου - περίληψης, εκτιμά μια περίληψη σε γενικευμένη μορφή για κάθε κείμενο που δίνεται στην είσοδο. Τόσο η διανυσματική αναπαράσταση των λέξεων του κειμένου, όσο και οι εκτιμήσεις μηχανικής μάθησης εντάσσονται στη φάση της μηχανικής μάθησης και αναλύονται περαιτέρω στην Ενότητα 4.6.

Τέλος, η εκτιμώμενη περίληψη, η οποία είναι γενικευμένης μορφής, ακολουθεί το στάδιο της μετα-επεξεργασίας προκειμένου να διαμορφωθεί η τελική περίληψη σε αναγνώσιμη από τον άνθρωπο μορφή. Η τρίτη αυτή φάση περιλαμβάνει μια σειρά από διαδικασίες, όπως είναι ο εντοπισμός των γενικευμένων εννοιών της γενικευμένης περίληψης και ο προσδιορισμός της αντιστοίχισης των γενικευμένων εννοιών με συγκεκριμένες του αρχικού εγγράφου. Αυτή η τελική φάση παρουσιάζεται με λεπτομέρεια στην Ενότητα 4.7.



Σχήμα 4.1: Διάγραμμα ροής της προτεινόμενης προσέγγισης για την αυτόματη περίληψη κειμένου με χρήση μηχανικής μάθησης και σημασιολογικών μετασχηματισμών περιεχομένου

4.5 Φάση πρώτη: Προ-επεξεργασία

Όπως είδαμε στην προηγούμενη ενότητα που παρουσιάζει την αρχιτεκτονική της προτεινόμενης προσέγγισης, η φάση της προ-επεξεργασίας αποτελείται από δύο βαθμίδες. Τη βαθμίδα της αποσαφήνισης της έννοιας των λέξεων (*ΑΕΛ*), η οποία περιγράφεται στην Ενότητα 4.5.1, παρακάτω, και τη βαθμίδα της σημασιολογικής γενίκευσης του περιεχομένου, η οποία παρουσιάζεται στην Ενότητα 4.5.2. Στο πλαίσιο της γενίκευσης περιεχομένου, εξετάζονται μια σειρά από στρατηγικές γενίκευσης περιεχομένου, οι οποίες αναλύονται με λεπτομέρεια στην Ενότητα 4.5.3.

4.5.1 Αποσαφήνιση έννοιας λέξεων

Η (*ΑΕΛ*) αφορά τη διαδικασία προσδιορισμού της έννοιας μιας λέξης σε ένα συγκεκριμένο εννοιολογικό πλαίσιο, αξιολογώντας έτσι τις διαφορετικές σημασίες που μπορεί να έχει μια λέξη για τον προσδιορισμό της κατάλληλης έννοιας [47, 132]. Ο σκοπός της *ΑΕΛ* στην παρούσα εργασία είναι τριπλός:

(*i*) Η μεθοδολογία *ΑΕΛ* χρησιμοποιείται στη διαδικασία γενίκευσης περιεχομένου για τον προσδιορισμό των εννοιών των λέξεων και την στοχευμένη αντικατάστασή τους με γενικότερες έννοιες στη φάση της προ-επεξεργασίας (Ενότητα 4.5.2).

(*ii*) Τα αναγνωριστικά *ΑΕΛ* που χρησιμεύουν για τη διάκριση των διαφορετικών εννοιών των λέξεων αξιοποιούνται στη φάση της μετα-επεξεργασίας (Ενότητα 4.7) για την αντιστοίχιση των γενικευμένων εννοιών με συγκεκριμένες.

(*iii*) Η μεθοδολογία *ΑΕΛ* χρησιμοποιείται για τη μετατροπή ενός κειμένου σε μια μορφή πλήρως αποσαφηνισμένου κειμένου (*ΠΑΚ*), σύμφωνα με την οποία, οι λέξεις του κειμένου αντικαθίστανται με αναγνωριστικά *ΑΕΛ* που προσδιορίζουν τη συγκεκριμένη έννοια των λέξεων. Στη συνέχεια, ένα σύνολο από ζεύγη κειμένου - περίληψης σε μορφή *ΠΑΚ*, χρησιμοποιείται στη φάση της εκπαίδευσης ενός μοντέλου μηχανικής μάθησης, το οποίο, με τη σειρά του, μαθαίνει να εκτιμά μια περίληψη σε μορφή *ΠΑΚ*, με δεδομένο ένα αρχικό κείμενο, επίσης, σε μορφή *ΠΑΚ*. Η τρίτη αυτή χρήση της *ΑΕΛ* αξιοποιείται με σκοπό να διερευνηθεί αν μπορεί η μορφή αυτή του κειμένου να συμβάλει στη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης (Ενότητα 4.6). Αναμένουμε βελτίωση της ακρίβειας των προβλέψεων, καθώς περιορίζεται ο χώρος αναζήτησης για την εκτίμηση μια ακολουθίας λέξεων που αποτελεί την εκτιμώμενη περίληψη, με την προϋπόθεση, βέβαια, ότι η μετατροπή του συνόλου εκπαίδευσης σε μορφή *ΠΑΚ* δεν θα μειώσει δραματικά τη συχνότητα εμφάνισης της κάθε λέξης στο σύνολο εκπαίδευσης και θα επιτραπεί στο μοντέλο μηχανικής μάθησης να εκπαιδευτεί επαρκώς.

Οι τρεις αυτές προοπτικές της *ΑΕΛ* θα διερευνηθούν σε βάθος στο πειραματικό μέρος της εργασίας που παρουσιάζεται στο επόμενο κεφάλαιο.

Η αναπαράσταση των εννοιών που χρησιμοποιείται στο πλαίσιο αυτής της εργασίας ακολουθεί τον τρόπο αναπαράστασης εννοιών του ηλεκτρονικού λεξικού *WordNet* [43, 44]. Δηλαδή, χρησιμοποιείται ένα αναγνωριστικό *ΑΕΛ* για κάθε έννοια, η μορφή του οποίου γίνεται φανερή στα ακόλουθα παραδείγματα εννοιών *book.n.01, book.n.02, watch.v.02* και *watch.v.06*. Συνεπώς,

τα αναγνωριστικά *ΑΕΛ* αποτελούνται από τρία μέρη που χωρίζονται μεταξύ τους με τελείες. Τα μέρη αυτά με τη σειρά, που αναγράφονται, αναπαριστούν: (*i*) μία λέξη (π.χ., *book* ή *watch*), (*ii*) ένα μέρος του λόγου (δηλ., *n* για ουσιαστικά και *v* για ρήματα) και (*iii*) έναν αριθμό που χαρακτηρίζει την έννοια μιας λέξης (π.χ., *01, 02, 06* κλπ. για τις διαφορετικές έννοιες της ίδιας λέξης). Για παράδειγμα, η λέξη *bank* έχει περισσότερες από μία έννοιες, όπως *bank.n.01, bank.n.02* ή *bank.v.01*, όπου οι δύο πρώτες αναφέρονται σε ουσιαστικά με διαφορετική σημασία (καθώς οι αριθμοί που προσδιορίζουν την έννοιά της λέξης είναι διαφορετικοί), ενώ η τρίτη αντιπροσωπεύει ένα ρήμα. Να διευκρινιστεί ότι για να προκύψει μία έκδοση *ΠΑΚ*, οι λέξεις του κειμένου αντικαθίστανται από τα αντίστοιχα αναγνωριστικά *ΑΕΛ* (π.χ., η λέξη *table* αναπαρίστανται με *table.n.01*), σύμφωνα με την εκάστοτε έννοιά τους.

Με δεδομένο ότι αυτή η παρούσα προσέγγιση βασίζεται σε πόρους γνώσης, ταιριάζουν περισσότερο τεχνικές *ΑΕΛ* που, επίσης, βασίζονται σε πόρους γνώσης [133, 134], καθώς εύκολα μπορούν να ενσωματωθούν στο προτεινόμενο πλαίσιο. Εναλλακτικά, θα μπορούσαν να χρησιμοποιηθούν και άλλες προσεγγίσεις, όπως είναι τα μοντέλα *ΑΕΛ* επιβλεπόμενης μάθησης [135], με την προϋπόθεση ότι υπάρχουν επαρκείς πόροι παραδειγμάτων χρήσης για την εκπαίδευση τέτοιων συστημάτων.

Στο πειραματικό μέρος αυτής της εργασίας (Κεφάλαιο 5), διερευνάται η επίδραση του μηχανισμού *ΑΕΛ* στην ακρίβεια προβλέψεων των μοντέλων μηχανικής μάθησης και στην εκτίμηση της τελικής περίληψης.

4.5.2 Σημασιολογική γενίκευση περιεχομένου

Η σημασιολογική γενίκευση περιεχομένου (ή κειμένου) αποτελεί βασική προοπτική της παρούσας προσέγγισης με σκοπό τη βελτίωση της απόδοσης των προβλέψεων μηχανικής μάθησης. Ο κύριος στόχος της γενίκευσης περιεχομένου είναι η αντικατάσταση άγνωστων ή σπάνιων λέξεων με σημασιολογικά συναφείς όρους. Συγκεκριμένα, μια λέξη που δεν υπάρχει στο λεξιλόγιο του συνόλου εκπαίδευσης (λέξη εκτός λεξιλογίου - *ΛΕΛ*) ενός μοντέλου μηχανικής μάθησης μπορεί να γενικευτεί σε μια λέξη με γενικότερη έννοια, η οποία έχει επαρκή παρουσία στο σύνολο εκπαίδευσης. Με παρόμοιο τρόπο, οι σπάνιες λέξεις, με μικρή συχνότητα στο σύνολο εκπαίδευσης, μπορούν να γενικευτούν σε όρους υψηλότερης συχνότητας.

Στο σημείο αυτό θα παρουσιάσουμε το θεωρητικό μοντέλο που περιγράφει τις αρχές και τις ιδιότητες της σημασιολογικής γενίκευσης περιεχομένου, η οποία στη συνέχεια θα διερευνηθεί και πειραματικά. Καταρχήν, η διαδικασία γενίκευσης του περιεχομένου απαιτεί μια ταξινομία εννοιών σε ιεραρχική δομή (όπως, για παράδειγμα, η ταξινομία εννοιών του Σχήματος 4.2), η οποία περιγράφεται στον Ορισμό 4.1. Μια ταξινομία εννοιών βασίζεται σε πόρους γνώσης (π.χ., *WordNet*) και αποτελείται από λέξεις ή έννοιες και τη σημασιολογική σχέση τους σε μια ιεραρχική δομή. Πιο συγκεκριμένα, ο τύπος ταξινομίας που μελετάται σε αυτή την εργασία περιλαμβάνει όρους οι οποίοι αντιστοιχούν σε έννοιες λέξεων.

Ορισμός 4.1 (Ταξινομία εννοιών)**.** *Μια ταξινομία εννοιών αποτελείται από μια ιεραρχική δομή σημασιολογικά αποσαφηνισμένων εννοιών που σχετίζονται με έναν τύπο σχέσης γενικότερης - ειδικότερης ή ευρύτερης - πιο συγκεκριμένης σημασίας.*

Οι σχέσεις μεταξύ των εννοιών περιγράφονται με τη χρήση των όρων *υπώνυμο* (Ορισμός 4.2)

και *υπερώνυμο* (Ορισμός 4.3). Ένα υπώνυμο μιας λέξης αναφέρεται σε ένα πιο συγκεκριμένο σημασιολογικό πεδίο της δεδομένης λέξης (π.χ., το *"άλογο"* είναι ένα υπώνυμο του *"ζώου"*). Από την άλλη πλευρά, ένα υπερώνυμο μιας λέξης αναφέρεται σε ένα ευρύτερο σημασιολογικό πεδίο από τη δεδομένη λέξη (π.χ., το *"όχημα"* είναι ένα υπερώνυμο του *"δίκυκλου"*). Σε μια ιεραρχική ταξινομία εννοιών μορφής *δένδρου*, η *ρίζα* (*root*) της ταξινομίας αντιστοιχεί στη λέξη με το ευρύτερο ή γενικότερο σημασιολογικό πεδίο, ενώ τα *φύλλα* περιέχουν τις πιο συγκεκριμένες έννοιες.

Ορισμός 4.2 (Υπώνυμο). *Δεδομένης μιας ταξινομίας εννοιών, μια έννοια c_b αποτελεί ένα υπώνυμο μιας άλλης έννοιας c_a αν και μόνο αν η έννοια c_b σημασιολογικά συνεπάγεται την έννοια c_a ($c_b \models c_a$).*

Ορισμός 4.3 (Υπερώνυμο). *Δεδομένης μιας ταξινομίας εννοιών, μια έννοια c_a αποτελεί ένα υπερώνυμο μιας άλλης έννοιας c_b αν και μόνο αν η έννοια c_b σημασιολογικά συνεπάγεται την έννοια c_a ($c_b \models c_a$)*

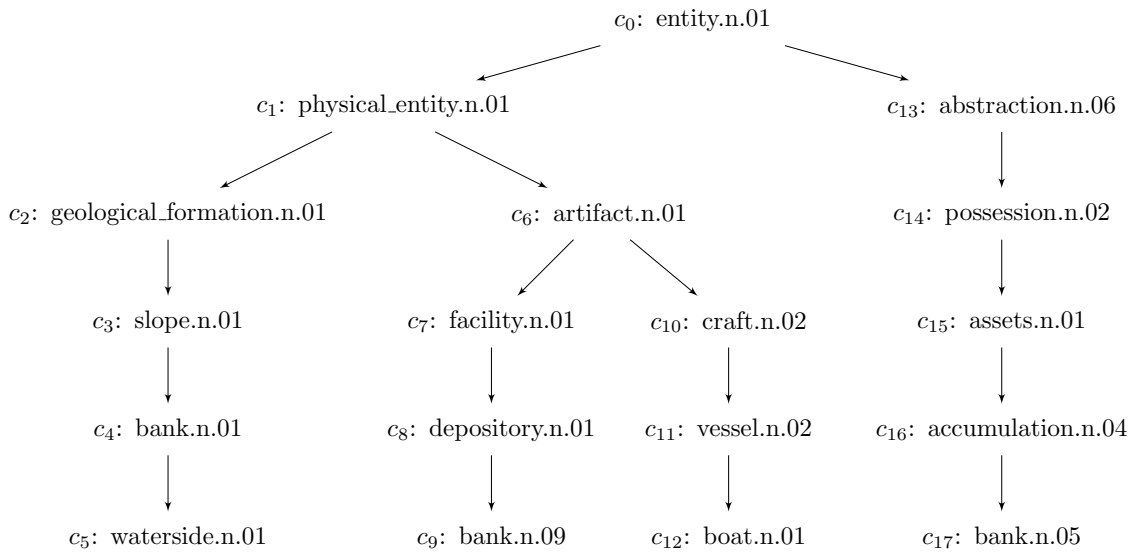
Επιπλέον, το μονοπάτι ταξινομίας υπώνυμων και υπερώνυμων εννοιών περιγράφεται στους Ορισμούς 4.4 και 4.5, αντίστοιχα. Δεδομένου ότι η ταξινομία των εννοιών θεωρείται ότι έχει ιεραρχική δομή (π.χ., μορφή δένδρου), ένα μονοπάτι υπώνυμων εννοιών αντιπροσωπεύει τη διατεταγμένη ακολουθία όρων από μια συγκεκριμένη έννοια (π.χ., κόμβο ταξινομίας) μέχρι την πιο συγκεκριμένη, η πιο ειδική έννοια αυτού του όρου (π.χ., έως ένα φύλλο της ταξινομίας). Αντίστοιχα, ένα μονοπάτι υπερώνυμων εννοιών περιλαμβάνει τους όρους από μια συγκεκριμένη έννοια μέχρι την πιο γενική ή ευρύτερη έννοια αυτού του όρου (π.χ., έως τη ρίζα της ταξινομίας).

Ορισμός 4.4 (Μονοπάτι ταξινομίας υπώνυμων εννοιών). *Δεδομένης μιας ταξινομίας εννοιών, ένα μονοπάτι υπώνυμων εννοιών από έναν όρο c_a είναι μια διατεταγμένη ακολουθία εννοιών $\{c_a,\ c_{a+1},\ ...,\ c_{i-1},\ c_i,\ c_{i+1},\ ...,\ c_{n-1},\ c_n\}$, όπου η έννοια c_{i+1} σημασιολογικά συνεπάγεται την έννοια c_i και η έννοια c_n αποτελεί ένα φύλλο της ταξινομίας.*

Ορισμός 4.5 (Μονοπάτι ταξινομίας υπερώνυμων εννοιών). *Δεδομένης μιας ταξινομίας εννοιών, ένα μονοπάτι υπερώνυμων εννοιών από έναν όρο c_a είναι μια διατεταγμένη ακολουθία εννοιών $\{c_a,\ c_{a-1},\ ...,\ c_{i+1},\ c_i,\ c_{i-1},\ ...,\ c_{r+1},\ c_r\}$, όπου η έννοια c_{i+1} σημασιολογικά συνεπάγεται την έννοια c_i και ο όρος c_r αποτελεί τον όρο ρίζα της ταξινομίας.*

Δεδομένου ενός κειμένου από το οποίο έχουν εξαχθεί οι έννοιες, σύμφωνα με μια χρησιμοποιούμενη ταξινομία, το μονοπάτι ταξινομίας υπερώνυμων εννοιών χρησιμοποιείται για τη γενίκευση των εξαγόμενων όρων από το εν λόγω κείμενο με σκοπό τη γενίκευση του περιεχομένου. Αντίστοιχα, ένα μονοπάτι ταξινομίας υπώνυμων εννοιών χρησιμοποιείται, για να αντικατασταθούν οι όροι που μπορεί να έχουν υποστεί γενίκευση με όρους πιο ειδικής ή συγκεκριμένης έννοιας. Τόσο το μονοπάτι υπερώνυμων όσο και το μονοπάτι υπώνυμων εννοιών μπορεί να χρησιμοποιηθεί για τη μετάβαση από μια συγκεκριμένη έννοια σε μια γενικότερη και από μια γενικότερη έννοια σε μια πιο συγκεκριμένη, αντίστοιχα (δηλ., γενίκευση ή συγκεκριμενοποίηση των όρων ενός κειμένου). Το Παράδειγμα 4.1 που ακολουθεί παρουσιάζει κάποια βασικά χαρακτηριστικά μιας ταξινομίας εννοιών.

Παράδειγμα 4.1 (Ταξινομία εννοιών).
Το Σχήμα 4.2 απεικονίζει ένα παράδειγμα ταξινομίας 18 εννοιών, αποτυπώνοντας τις σχέσεις υπερωνυμίας ή υπωνυμίας μεταξύ τους. Κάθε κόμβος περιέχει ένα αναγνωριστικό ΑΕΛ (π.χ., facility.v.01). Η λέξη bank εμφανίζεται με τρεις διαφορετικές έννοιες, bank.n.01 (επικλινής γη),



Σχήμα 4.2: Ένα παράδειγμα ταξινομίας εννοιών.

bank.n.05 (προμήθεια ή απόθεμα για μελλοντική χρήση) και bank.n.09 (κτίριο στο οποίο λειτουργεί ένα τραπεζικό κατάστημα). Το μονοπάτι ταξινομίας υπώνυμων εννοιών του όρου c_3 (slope.n.01) είναι $\{c_3, c_4, c_5\}$, ενώ το μονοπάτι υπερώνυμων εννοιών της ίδιας έννοιας είναι $\{c_3, c_2, c_1, c_0\}$. Η πιο γενική έννοια, με το ευρύτερο σημασιολογικό πεδίο, αυτής της ταξινομίας είναι η entity.v.01, η οποία αποτελεί τη ρίζα της ταξινομίας. Αυτή η ταξινομία περιέχει μόνο ουσιαστικά ενώ εναλλακτικά, μια άλλη ταξινομία θα μπορούσε να περιέχει και άλλα μέρη του λόγου, όπως ρήματα.

Οι Ορισμοί 4.6 και 4.7 που ακολουθούν ορίζουν πότε μια έννοια θεωρείται γενικεύσιμη και πότε ένα κείμενο ή τμήμα κειμένου είναι γενικεύσιμο, αντίστοιχα. Ένας όρος c_i με μονοπάτι ταξινομίας υπερώνυμων εννοιών Ph_{c_i} είναι γενικεύσιμος όταν η έννοια c_i συνεπάγεται σημασιολογικά μια έννοια $c_j \in Ph_{c_i}$ με μικρότερο βάθος στην ταξινομία (και πιο γενική έννοια) από τον όρο c_i (δηλ., η λέξη c_i στο κείμενο θα μπορούσε να αντικατασταθεί από τη λέξη c_j, δημιουργώντας μια γενικευμένη έκδοση του κειμένου). Επίσης, ένα κείμενο ή τμήμα κειμένου μπορεί να γενικευτεί όταν μία ή περισσότερες από τις έννοιές του είναι γενικεύσιμες. Διευκρινίζεται ότι ως βάθος ταξινομίας θεωρείται το πλήθος των διατεταγμένων εννοιών από τη ρίζα της ταξινομίας έως κάποια συγκεκριμένη έννοια. Η έννοια με το μέγιστο βάθος ταξινομίας ενός μονοπατιού υπώνυμων εννοιών βρίσκεται σε φύλλο της ταξινομίας.

Ορισμός 4.6 (Γενικεύσιμη έννοια)**.** *Μια έννοια c_i είναι γενικεύσιμη όταν το μονοπάτι ταξινομίας υπερώνυμων εννοιών της Ph_{c_i} περιέχει τουλάχιστον μία έννοια $c_j \in Ph_{c_i}$ τέτοια ώστε η έννοια c_i σημασιολογικά να συνεπάγεται την έννοια c_j.*

Ορισμός 4.7 (Γενικεύσιμο κείμενο)**.** *Ένα κείμενο ή τμήμα κειμένου μπορεί να γενικευτεί ή θεωρείται γενικεύσιμο όταν περιέχει τουλάχιστον μία έννοια η οποία είναι γενικεύσιμη.*

Μια ταξινομία μπορεί να περιέχει έννοιες πολύ γενικής σημασίας, όπως *entity* ή *object*. Για να αποφευχθεί η γενίκευση σε έννοιες με πολύ ευρύ σημασιολογικό πεδίο, πρέπει να καθοριστεί ένα επίπεδο, το οποίο θέτει όρια στον βαθμό γενίκευσης. Ο Ορισμός 4.8 καθορίζει ότι μια έννοια μπορεί να γενικευθεί έως ένα ελάχιστο βάθος της ταξινομίας εννοιών, σύμφωνα με το δεδομένο επίπεδο γενίκευσης.

Ορισμός 4.8 (Επίπεδο γενίκευσης). *Το επίπεδο γενίκευσης ενός κειμένου είναι ίσο με το ελάχιστο βάθος της ταξινομίας εννοιών που μπορεί να γενικευτεί μια έννοια.*

4.5.3 Στρατηγικές γενίκευσης κειμένου

Σε συνέχεια της θεωρητικής περιγραφής που προηγήθηκε, σε αυτή την ενότητα θα παρουσιάσουμε έξι διαφορετικές στρατηγικές γενίκευσης περιεχομένου. Η εφαρμογή των στρατηγικών αυτών αποτελεί τη φάση της προ-επεξεργασίας για την προετοιμασία των δεδομένων, τα οποία θα χρησιμοποιηθούν με σκοπό τη βελτίωση της απόδοσης ενός μοντέλου μηχανικής μάθησης (Ενότητα 4.6). Συγκεκριμένα, οι μέθοδοι που περιγράφονται εδώ χρησιμοποιούνται για να διερευνηθεί, αν η εκάστοτε στρατηγική μπορεί να επιφέρει βελτίωση στην αυτόματη *ΠΚ*, σε σύγκριση με τις προσεγγίσεις που βασίζονται μόνο στη μηχανική μάθηση. Οι εξεταζόμενες στρατηγικές ακολουθούν τρεις διαφορετικές κατευθύνσεις:

(*i*) Γενίκευση εννοιών με χρήση των υπερώνυμων εννοιών τους, αξιοποιώντας μια σημασιολογική ταξινομία εννοιών.

(*ii*) Γενίκευση εννοιών σε γνωστές ονοματικές οντότητες με χρήση αναγνώρισης ονοματικών οντοτήτων.

(*iii*) Ο συνδυασμός των δυο προαναφερόμενων προσεγγίσεων γενίκευσης εννοιών.

Οι στρατηγικές γενίκευσης εφαρμόζονται είτε στο αρχικό κείμενο (δηλ., κείμενο που δεν έχει δεχθεί προηγούμενη επεξεργασία) είτε σε κείμενο μορφής *ΠΑΚ* (δηλ., σε κείμενο του οποίου έχει γίνει σημασιολογική αποσαφήνιση του συνόλου των λέξεων που περιλαμβάνονται σε αυτό).

Οι τεχνικές που εξετάζονται λαμβάνουν υπόψη τη συχνότητα εμφάνισης κάθε όρου (θ_f) του αρχικού κειμένου. Αυτή η επιλογή βασίζεται στο γεγονός ότι τα συστήματα επιβλεπόμενης μάθησης απαιτούν έναν επαρκή αριθμό παραδειγμάτων χρήσης για την αποτελεσματική εκπαίδευσή τους. Σε αυτή την κατεύθυνση, η γενίκευση άγνωστων ή σπάνιων λέξεων και η αντικατάστασή τους από πιο συχνούς όρους μπορεί να οδηγήσει σε πιο ακριβείς προβλέψεις, βελτιώνοντας έτσι τις επιδόσεις ενός συστήματος μηχανικής μάθησης. Η επίδραση της τιμής της παραμέτρου θ_f στην ακρίβεια προβλέψεων του συστήματος μελετάται στο πειραματικό μέρος αυτής της εργασίας (Ενότητα 5), όπου εκεί προσδιορίζεται η βέλτιστη τιμή της για τα χρησιμοποιούμενα σύνολα δεδομένων.

Μια ακόμη υπερ-παράμετρος που επηρεάζει τη λειτουργικότητα των εξεταζόμενων στρατηγικών είναι το επίπεδο γενίκευσης θ_d (Ορισμός 4.8), το οποίο χρησιμοποιείται για να περιοριστεί η γενίκευση των όρων σε έννοιες με πολύ ευρύ σημασιολογικό πεδίο. Η υπερ-γενίκευση θεωρούμε ότι θα πρέπει να περιοριστεί καθώς δεν αναμένεται να έχει θετική επίδραση στην ακρίβεια των προβλέψεων μηχανικής μάθησης. Αυτό συμβαίνει, καθώς ένα μοντέλο μηχανικής μάθησης δεν θα είναι σε θέση να κάνει διάκριση μεταξύ των διαφορετικών σημασιών των πολύ γενικών και ιδιαίτερα συχνών εννοιών, που εμφανίζονται στο σύνολο εκπαίδευσης, με αποτέλεσμα την αποτυχία εκτίμησης μιας κατάλληλης περίληψης. Στο πειραματικό μέρος αυτής της εργασίας εξετάζεται η επίδραση της παραμέτρου θ_d στην ακρίβεια των προβλέψεων, προσδιορίζοντας μια βέλτιστη τιμή αυτής της παραμέτρου για τα χρησιμοποιούμενα σύνολα δεδομένων.

Ακολουθεί η παρουσίαση των στρατηγικών γενίκευση περιεχομένου.

Γενίκευση βασισμένη στο βάθος ταξινομίας (ΓBT)

Αυτή η στρατηγική επιχειρεί γενίκευση περιεχομένου λαμβάνοντας υπόψη: (i) το δεδομένο κατώφλι θ_f, σύμφωνα με το οποίο οι έννοιες με συχνότητα εμφάνισης στο σύνολο εκπαίδευσης ίση ή μικρότερη από την τιμή θ_f είναι υποψήφιες για γενίκευση και (ii) το δεδομένο επίπεδο γενίκευσης θ_d, σύμφωνα με το οποίο οι έννοιες με βάθος ταξινομίας μεγαλύτερο από την τιμή θ_d μπορούν να γενικευτούν. Το Παράδειγμα 4.2 παρουσιάζει τη στρατηγική ΓBT, όπου οι έννοιες μιας πρότασης γενικεύονται στα υπερώνυμά τους ανάλογα με τη συχνότητα εμφάνισής τους στο σύνολο εκπαίδευσης.

Παράδειγμα 4.2 (ΓBT)**.**
Δίνονται:

(i) *η πρόταση "he is sitting on the bank of the river watching a boat",*

(ii) *η αποσαφηνισμένη έκδοσή της σε μορφή ΠΑΚ: "he is sitting on the bank.n.01 of the river.n.01 watching a boat.n.01",*

(iii) *η ταξινομία εννοιών T του Σχήματος 4.2,*

(iv) *ένα σύνολο εννοιών και η συχνότητα εμφάνισής τους στο σύνολο εκπαίδευσης*

$$F = \{(\text{“}bank.n.01\text{”},\ 58), (\text{“}slope.n.01\text{”},\ 120),\ (\text{“}boat.n.01\text{”},\ 45),\\ (\text{“}vessel.n.02\text{”},\ 98),\ (\text{“}craft.n.02\text{”},\ 160),\ (\text{“}river.n.01\text{”},\ 220)\}$$

(v) *οι τιμές των παραμέτρων $\theta_d = 3$ και $\theta_f = 100$*

Σύμφωνα με τα παραπάνω, οι υποψήφιες έννοιες για γενίκευση είναι οι "bank.n.01" και "boat.n.01", καθώς η συχνότητά τους στο σύνολο εκπαίδευσης είναι μικρότερη από θ_f και το βάθος ταξινομίας τους στην T είναι μεγαλύτερο από θ_d.

Το μονοπάτι υπερώνυμων εννοιών για κάθε μία από αυτές τις δύο έννοιες εξάγεται από την ταξινομία T όπως ακολούθως:

$$P_{bank.n.01} = \{\text{“}bank.n.01\text{”},\ \text{“}slope.n.01\text{”},\ \text{“}geological\_formation.n.01\text{”},\\ \text{“}physical\_entity.n.01\text{”},\ \text{“}entity.n.01\text{”}\}$$
$$P_{boat.n.01} = \{\text{“}boat.n.01\text{”},\ \text{“}vessel.n.02\text{”},\ \text{“}craft.n.02\text{”},\ \text{“}artifact.n.01\text{”},\\ \text{“}physical\_entity.n.01\text{”},\ \text{“}entity.n.01\text{”}\}$$

Σύμφωνα με τα παραπάνω, η πρόταση γενικεύεται σε: "he is sitting on the slope.n.01 of the river watching a craft.n.01", ικανοποιώντας τα δεδομένα όρια. Σημειώνεται επίσης ότι οι γενικευμένες έννοιες αντικαθίστανται από αναγνωριστικά αποσαφήνισης εννοιών (π.χ., slope.n.01) για ευκολότερη αναγνώριση τους στη φάση της μετα-επεξεργασίας, η οποία θα περιγραφεί στη συνέχεια.

Ο Αλγόριθμος 4.1 περιγράφει τη διαδικασία και τα βήματα της στρατηγικής ΓBT. Ο αλγόριθμος δέχεται ως είσοδο ένα κείμενο ($text$), την αποσαφηνισμένη έκδοση του κειμένου σε μορφή $ΠΑΚ$ ($wsdText$), ένα σύνολο με τις συχνότητες εμφάνισης των εννοιών στο σύνολο δεδομένων F, το

Αλγόριθμος 4.1 Γενίκευση βασισμένη στο βάθος ταξινομίας (*ΓΒΤ*)

Require: $text$, $wsdText$, F, T, θ_d, θ_f

1: $genText \leftarrow text$
2: **for all** $token \in text$ **do**
3: $f_{token} \leftarrow$ Συχνότητα της λέξης $token$ από το F
4: **if** $f_{token} \leq \theta_f$ **then**
5: $c \leftarrow$ Αναγνωριστικό αποσαφήνισης της λέξης $token$ από το $wsdText$
6: $P_c \leftarrow$ Μονοπάτι υπερώνυμων της έννοιας c από την T
7: $d_c \leftarrow$ Βάθος ταξινομίας της έννοιας c
8: $f_c \leftarrow$ Συχνότητα της έννοιας c από το F
9: **while** $f_c < \theta_f$ **και** $d_c > \theta_d$ **do**
10: $c \leftarrow$ Υπερώνυμο της έννοιας c από το P_{token}
11: $d_c \leftarrow$ Βάθος ταξινομίας της έννοιας c
12: $f_c \leftarrow$ Συχνότητα της έννοιας c από το F
13: **end while**
14: **if** (Η λέξη της έννοιας c) $\neq token$ **then**
15: $genText \leftarrow$ Γενίκευση της λέξης $token$ του $genText$ στην έννοια c
16: $F \leftarrow (F \setminus \{(token,\ f_{token})\}) \cup \{(token,\ f_{token} - 1)\}$
17: $F \leftarrow (F \setminus \{(c,\ f_c)\}) \cup \{(c,\ f_c + 1)\}$
18: **end if**
19: **end if**
20: **end for**
21: **return** $genText$

όριο βάθους ταξινομίας θ_d και το κατώφλι της συχνότητας των εννοιών θ_f, οι οποίες πρόκειται να γενικευτούν. Στον βρόχο επανάληψης (for, γραμμές 2-20) εξετάζονται όλες οι λέξεις του αρχικού κειμένου και αυτές που έχουν συχνότητα μικρότερη ή ίση με θ_f είναι υποψήφιες για γενίκευση (γραμμή 4). Σε αυτό το σημείο, αν αληθεύει η έκφραση της δομής επιλογής, το αναγνωριστικό *ΑΕΛ* της λέξεις $token$ εκχωρείται στη μεταβλητή c (γραμμή 5). Επίσης, το μονοπάτι υπερώνυμων εννοιών, το βάθος ταξινομίας και η συχνότητα της έννοιας c εκχωρούνται στις μεταβλητές P_c, d_c και f_c, αντίστοιχα (γραμμές 6-8). Όσο η συχνότητα και το βάθος ταξινομίας της τρέχουσας επιλεγμένης έννοιας c δεν πληρούν τις απαιτήσεις κατωφλίου θ_f και θ_d (γραμμή 9), τότε εκχωρείται στην μεταβλητή c το επόμενο υπερώνυμο της τρέχουσας έννοιας από το μονοπάτι υπερώνυμων εννοιών P_c (γραμμή 10). Αντίστοιχα, ενημερώνονται οι τιμές του βάθους ταξινομίας d_c και της συχνότητας f_c της νέας έννοιας c (γραμμές 11-12). Εάν ο προηγούμενος βρόχος είχε ως αποτέλεσμα την ανάκτηση μιας νέας γενικευμένης έννοιας c (γραμμή 14), τότε η λέξη $token$ αντικαθίσταται από την έννοια c (γραμμή 15) και οι συχνότητες εμφάνισης της λέξης $token$ και της έννοιας c ενημερώνονται (γραμμές 16-17). Μόλις εκτελεστεί αυτή η διαδικασία για όλες τις λέξεις του αρχικού κειμένου ($text$), ο αλγόριθμος τερματίζει και επιστρέφει τη γενικευμένη έκδοση του αρχικού κειμένου στη μεταβλητή $genText$ (γραμμή 21).

Η υπολογιστική πολυπλοκότητα του Αλγόριθμου 4.1 είναι $O(k \cdot n)$, καθώς στη χειρότερη περίπτωση θα εξεταστούν όλες οι λέξεις εισόδου πλήθους n (δηλ., $f_{token} \leq \theta_f\ \forall\ token \in text$).

Για κάθε λέξη εισόδου, ο βρόχος επανάληψης (while) των γραμμών 9-13 εξετάζει όλες τις έννοιες του μεγαλύτερου μονοπατιού υπερώνυμων εννοιών της ταξινομίας ($k = |P_{token}|$). Στην πράξη ισχύει ότι $k << n$ (ειδικά σε μεγάλα κείμενα). Επομένως, η υπολογιστική πολυπλοκότητα του Αλγόριθμου 4.1 θεωρείται $O(n)$. Είναι μια γραμμική πολυπλοκότητα που επιτρέπει στον αλγόριθμο την αποδοτική λειτουργία με μικρούς χρόνους εκτέλεσης.

Γενίκευση βασισμένη στο βάθος ταξινομίας πλήρως αποσαφηνισμένου κειμένου (*ΓΒΤ-ΠΑΚ*)

Αυτή η στρατηγική αποτελεί μια τροποποίηση της *ΓΒΤ* που παρουσιάστηκε παραπάνω. Η διαφορά είναι ότι στο κείμενο εισόδου αρχικά εφαρμόζεται *ΑΕΛ* και στη συνέχεια η έκδοση του κειμένου μορφής *ΠΑΚ* δίνεται ως είσοδος στον Αλγόριθμο 4.1. Αντίστοιχα, και σε αυτή τη στρατηγική γενίκευσης, οι σπάνιες έννοιες, οι οποίες δεν πληρούν το όριο θ_f, γενικεύονται. Η στρατηγική *ΓΒΤ-ΠΑΚ* μπορεί να γίνει περαιτέρω κατανοητή στο Παράδειγμα 4.3, το οποίο σε μεγάλο μέρος βασίζεται στα δεδομένα που παρουσιάστηκαν στο προαναφερόμενο Παράδειγμα 4.2.

Παράδειγμα 4.3 (*ΓΒΤ-ΠΑΚ*)**.**
Με δεδομένο το κείμενο εισόδου και τις άλλες παραμέτρους του Παραδείγματος 4.2 που προηγήθηκε, το αρχικό κείμενο μετατρέπεται σε μορφή ΠΑΚ ως εξής: "he is sitting on the bank.n.01 of the river.n.01 watching a boat.n.01" που δίνεται ως είσοδος στον Αλγόριθμο 4.1 και τελικά γενικεύεται σε "he is sitting on the slope.n.01 of the river.n.01 watching a craft.n.01".
Το κείμενο εξόδου, επίσης, διατηρεί τα αναγνωριστικά αποσαφήνισης των εννοιών εκείνων που δεν γενικεύτηκαν, όπως είναι η έννοια river.n.01, καθώς είναι μορφής ΠΑΚ.

Γενίκευση βασισμένη σε ονοματικές οντότητες (*ΓΟΟ*)

Αυτή η στρατηγική εκτελεί γενίκευση περιεχομένου σύμφωνα με (i) τις προκαθορισμένες ονοματικές οντότητες (π.χ., τοποθεσία, άτομο, οργανισμός) και (ii) το όριο θ_f, σύμφωνα με το οποίο οι όροι με συχνότητα εμφάνισης στο σύνολο εκπαίδευσης ίση η μικρότερη από την τιμή θ_f είναι υποψήφιοι για γενίκευση. Η στρατηγική *ΓΟΟ* παρουσιάζεται με περισσότερη λεπτομέρεια στο Παράδειγμα 4.4, όπου οι έννοιες μιας πρότασης γενικεύονται σε ονοματικές οντοτήτων, λαμβάνοντας υπόψη τη συχνότητά τους στο σύνολο εκπαίδευσης.

Παράδειγμα 4.4 (*ΓΟΟ*)**.**
Δίνονται:

(i) *το κείμενο "Elizabeth works at an antique shop in New York City",*

(ii) *ένα σύνολο από ονοματικές οντότητες*
$E = \{LOCATION,\ PERSON,\ ORGANIZATION\}$,

(iii) *ένα σύνολο εννοιών με τις συχνότητες εμφάνισής τους στο σύνολο εκπαίδευσης*
$F = \{(\text{“}Elizabeth\text{”},\ 58),\ (\text{“}antique\ shop\text{”},\ 22),\ (\text{“}New\ York\ City\text{”},\ 140)\}$,

(iv) $\theta_f = 100$

Μετά από διαδικασία αναγνώρισης ονοματικών οντοτήτων, προκύπτουν οι εξής ονοματικές οντότητες: "Elizabeth" (PERSON), "antique shop" (ORGANIZATION) και "New York City" (LOCATION).
Οι υποψήφιες έννοιες για γενίκευση είναι "Elizabeth" και " antique shop"", καθώς η συχνότητά τους στο σύνολο εκπαίδευσης είναι μικρότερη από θ_f.
Συνεπώς, το αρχικό κείμενο γενικεύεται σε "PERSON works at an ORGANIZATION in New York City".

Αλγόριθμος 4.2 Γενίκευση κειμένου βασιζόμενη σε ονοματικές οντότητες (*ΓΟΟ*)

Require: $text$, E, F, θ_f
1: $genText \leftarrow text$
2: $tokenNamedEntities \leftarrow$ Ονοματικές οντότητες του $text$ απο E
3: **for all** $(token,\ namedEntity) \in tokenNamedEntities$ **do**
4: $\quad f_{token} \leftarrow$ Συχνότητα του $token$ από F
5: $\quad$ **if** $f_{token} < \theta_f$ **then**
6: $\quad\quad genText \leftarrow$ Γενίκευσε $token$ του $genText$ σε $namedEntity$
7: $\quad$ **end if**
8: **end for**
9: **return** $genText$

Ο Αλγόριθμος 4.2 περιγράφει τα βήματα της στρατηγικής *ΓΟΟ* με λεπτομέρεια. Ο αλγόριθμος αυτός δέχεται ως είσοδο το υποψήφιο κείμενο για γενίκευση $text$, ένα σύνολο ονοματικών οντοτήτων E, ένα σύνολο εννοιών F και το όριο συχνότητας εμφάνισης των εννοιών θ_f. Αρχικά, το κείμενο εισόδου εκχωρείται στην μεταβλητή $genText$ (γραμμή 1). Στη συνέχεια, εφαρμόζεται αναγνώριση ονοματικών οντοτήτων στο αρχικό κείμενο ($text$), με αποτέλεσμα τον προσδιορισμό πλειάδων του τύπου (αναγνωριστικό λέξης, ονοματική οντότητα), οι οποίες εκχωρούνται στη μεταβλητή $tokenNamedEntity$ (γραμμή 2). Στη συνέχεια, εξετάζονται όλες οι πλειάδες $tokenNamedEntity$ (γραμμές 3-8) και η συχνότητα f_{token} για κάθε λέξη (ή αναγνωριστικό λέξης) ανακτάται (γραμμή 4). Αν η εν λόγω συχνότητα είναι ίση ή μικρότερη από θ_f (γραμμή 5), τότε ο εξεταζόμενος όρος $token$ γενικεύεται στην αντίστοιχη ονοματική οντότητα και ενημερώνεται η γενικευμένη έκδοση του κειμένου ($genText$, γραμμή 6). Ο αλγόριθμος τερματίζεται όταν εξεταστούν όλα τα ζεύγη $tokenNamedEntity$, επιστρέφοντας τη γενικευμένη έκδοση του αρχικού κειμένου ($genText$).

Η υπολογιστική πολυπλοκότητα του Αλγόριθμου 4.2 είναι γραμμική ($O(n)$), καθώς στη χειρότερη περίπτωση, η στρατηγική *ΓΟΟ* εξετάζει όλες τις πλειάδες εισόδου $(token, namedEntity)$ πλήθους n.

Γενίκευση βασισμένη σε ονοματικές οντότητες πλήρως αποσαφηνισμένου κειμένου (*ΓΟΟ-ΠΑΚ*)

Αυτή η στρατηγική αποτελεί μια τροποποίηση της *ΓΟΟ*, με τον ίδιο τρόπο που η *ΓΒΤ-ΠΑΚ* είναι μια τροποποίηση της *ΓΒΤ* όπως περιγράψαμε παραπάνω. Στην στρατηγική *ΓΟΟ-ΠΑΚ*, αντί να παρέχεται το αρχικό κείμενο ως είσοδος στον Αλγόριθμο 4.2, παρέχεται η έκδοση του κειμένου

μορφής *ΠΑΚ*, όπου οι λέξεις του έχουν αντικατασταθεί με τα αντίστοιχα αναγνωριστικά *ΑΕΛ*. Η στρατηγική αυτή παρουσιάζεται στο Παράδειγμα 4.5, το οποίο βασίζεται στο προαναφερόμενο Παράδειγμα 4.4.

Παράδειγμα 4.5 (*ΓΟΟ-ΠΑΚ*)**.**
Σύμφωνα με το κείμενο εισόδου και τις άλλες παραμέτρους του Παραδείγματος 4.4, το αποσαφηνισμένο κείμενο είναι: "Elizabeth works at an antique.n.02 shop.n.01 in New_York_City.n.01", το οποίο γενικεύεται σε "PERSON works at an ORGANIZATION in New_York_City.n.01".
Οι ονοματικές οντότητες αντιστοιχούν σε λέξεις ή φράσεις του αρχικού κειμένου οι οποίες έχουν αντικατασταθεί με τη χρήση μηχανισμού αναγνώρισης ονοματικών οντοτήτων. Σημειώνεται ότι το γενικευμένο κείμενο μπορεί να διατηρεί τα αναγνωριστικά ΑΕΛ των μη γενικευμένων εννοιών (π.χ., New_York_City.n.01 στην παραπάνω πρόταση) καθώς πρόκειται για ΠΑΚ.

Συνδυασμός των στρατηγικών *ΓΟΟ* και *ΓΒΤ* (*ΓΟΟ-ΓΒΤ*)

Αυτή η στρατηγική εκτελεί γενίκευση περιεχομένου σε δύο βήματα. Αρχικά εφαρμόζεται η στρατηγική *ΓΟΟ* και στη συνέχεια η στρατηγική *ΓΒΤ*. Το Παράδειγμα 4.6 παρουσιάζει την εφαρμογή της στρατηγικής *ΓΟΟ-ΓΒΤ*.

Παράδειγμα 4.6 (*ΓΟΟ-ΓΒΤ*)**.**
Θεωρούμε το κείμενο "Elizabeth, who works at an antique shop in New York City, is sitting on the bank of the river watching a boat", καθώς και τις παραμέτρους που δίνονται στα προαναφερόμενα παραδείγματα 4.2 και 4.4, με την διαφορά ότι στο σύνολο F προστίθεται η λέξη "who", η οποία έχει 520 *εμφανίσεις στο σύνολο εκπαίδευσης.*
Οι ονοματικές οντότητες του κειμένου, όπως αναγνωρίστηκαν από τον μηχανισμό αναγνώρισης ονοματικών οντοτήτων, είναι οι ακόλουθες: "Elizabeth" (PERSON), "who" (PERSON), "antique shop" (ORGANIZATION) και "New York City" (LOCATION).
Αφού η στρατηγική ΓΟΟ εφαρμόζεται πρώτη, οι υποψήφιες έννοιες για γενίκευση είναι οι "Elizabeth" και "antique shop", καθώς οι συχνότητές τους στο σύνολο εκπαίδευσης είναι μικρότερες από το δεδομένο όριο θ_f. Επομένως, το αρχικό κείμενο το οποίο ήταν σε μορφή ΠΑΚ γίνεται: "PERSON, who works at an ORGANIZATION in New York City, is sitting on the bank.n.01 of the river.n.01 watching the boat.n.01".
Στη συνέχεια, εφαρμόζεται η στρατηγική ΓΒΤ, η οποία γενικεύει τις έννοιες bank.n.01 και boat.n.01, καθώς οι συχνότητες τους στο σύνολο εκπαίδευσης είναι μικρότερες από θ_f. Τέλος, μετά την εφαρμογή και των δύο στρατηγικών γενίκευσης, το αρχικό κείμενο γενικεύεται σε: "PERSON, who works at an ORGANIZATION in New York City, is sitting on the slope.n.01 of the river watching the craft.n.01".

Συνδυασμός στρατηγικών *ΓΟΟ* και *ΓΒΤ-ΠΑΚ*(*ΓΟΟ-ΓΒΤ-ΠΑΚ*)

Η τελευταία στρατηγική γενίκευσης που εξετάζεται συνδυάζει τις μεθοδολογίες *ΓΟΟ* και *ΓΒΤ-ΠΑΚ*. Το Παράδειγμα 4.7 παρουσιάζει τη σχετική μεθοδολογία.

Παράδειγμα 4.7 (*ΓΟΟ-ΓΒΤ-ΠΑΚ*)**.**
Δίνονται:

(i) Το κείμενο "Elizabeth, who works at an antique shop in New York City, is sitting on the bank of the river watching a boat",

(ii) το αρχικό κείμενο σε μορφή ΠΑΚ "Elizabeth, who.n.01 works at an antique.n.02 shop.n.01 in New_York_city.n.01, is sitting on the bank.n.01 of the river.n.01 watching a boat.n.01" και

(iii) οι παράμετροι του Παραδείγματος 4.6.

Το αρχικό κείμενο γενικεύεται σε "PERSON, who.n.01 works at an ORGANIZATION in New_York_city.n.01, is sitting on the slope.n.01 of the river.n.01 watching the craft.n.01", σύμφωνα με την στρατηγική γενίκευσης ΓΟΟ-ΓΒΤ-ΠΑΚ.
Και εδώ, επειδή έχουμε γενίκευση ΠΑΚ, το γενικευμένο κείμενο διατηρεί τα αναγνωριστικά ΑΕΛ και των λέξεων που δε γενικεύτηκαν.

4.6 Φάση δεύτερη: Μηχανική μάθηση

Η φάση της μηχανικής μάθησης, στο πλαίσιο της προτεινόμενης προσέγγισης, αποτελείται από δύο διαδικασίες. Αρχικά, ανακτώνται οι ενσωματώσεις λέξεων (*word embeddings*) από το γενικευμένο κείμενο, το οποίο προκύπτει από την προηγούμενη φάση της προ-επεξεργασίας, με την σχετική μεθοδολογία να παρουσιάζεται παρακάτω στην Ενότητα 4.6.1. Οι διανυσματικές αναπαραστάσεις των λέξεων χρησιμοποιούνται τόσο στην διαδικασία μάθησης ενός νευρωνικού δικτύου όσο και στη φάση της μετα-επεξεργασίας, για τις μετρήσεις ομοιότητας κειμένου, όπως θα δούμε παρακάτω. Στη συνέχεια, ακολουθεί η διαδικασία εκπαίδευσης και εκτίμησης νέων περιλήψεων σε δεδομένη είσοδο. Σύμφωνα με τη διαδικασία μάθησης, ένα σύνολο από ακολουθίες λέξεων κειμένου και περίληψης παρέχονται σε ένα μοντέλο βαθιάς μάθησης που εκπαιδεύεται να εκτιμά την περίληψη ενός αρχικού κειμένου. Μετά το στάδιο της εκπαίδευσης, ένα μοντέλο μηχανικής μάθησης εκτιμά μια ακολουθία λεκτικών μονάδων $\hat{Y} = (\hat{y}_1,\ \hat{y}_2, ..., \hat{y}_m)$ (την εκτιμώμενη περίληψη), με δεδομένη μια ακολουθία λεκτικών μονάδων εισόδου $X = (x_1,\ x_2, ..., x_n)$ (το αρχικό κείμενο).

4.6.1 Διανυσματική αναπαράσταση λέξεων

Η προτεινόμενη προσέγγιση χρησιμοποιεί ενσωματώσεις λέξεων, οι οποίες έχουν περιγραφεί στην Ενότητα 3.2.1, στη διαδικασία εκπαίδευσης των μοντέλων μηχανικής μάθησης (Ενότητα 3.3) και στη φάση της μετα-επεξεργασίας (Ενότητα 4.7). Τα ζεύγη κειμένου-περίληψης του συνόλου δεδομένων, το οποίο έχει τροποποιηθεί σύμφωνα με μία συγκεκριμένη στρατηγική γενίκευσης, χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου της αντίστοιχης διανυσματικής αναπαράστασης των λέξεων. Στην εργασία αυτή μπορεί να χρησιμοποιηθεί κάθε προσέγγιση ενσωμάτωσης λέξεων είτε ανεξάρτητη από τα συμφραζόμενα, (όπως *word2vec, GloVe, fastText* [83, 136]) είτε εξαρτώμενη από τα συμφραζόμενα (όπως *BERT, ELMO* [86]), αρκεί τα διανύσματα αυτά να διατηρούν τις σημασιολογικές σχέσεις μεταξύ των λέξεων στον διανυσματικό χώρο.

Στο πειραματικό μέρος αυτής της εργασίας (Κεφάλαιο 5), έχει χρησιμοποιηθεί η μεθοδολογία *Word2Vec* και, πιο συγκεκριμένα, το μοντέλο συνεχούς συνόλου λέξεων (*continuous bag-of-word*

- *CBOW*) [82, 137] για την περίπτωση των ενσωματώσεων λέξεων ανεξάρτητων των συμφραζομένων. Στην περίπτωση των ενσωματώσεως λέξεων, οι οποίες εξαρτώνται από τα συμφραζόμενα, χρησιμοποιείται το μοντέλο *BERT* [102]. Ακολουθεί η περιγραφή των μοντέλων μηχανικής μάθησης, τα οποία αξιοποιούν τις ενσωματώσεις λέξεων.

4.6.2 Μοντέλα νευρωνικών δικτύων βαθιάς μάθησης

Στο πλαίσιο αυτής της εργασίας, εξετάζονται τα ακόλουθα πέντε μοντέλα βαθιάς μάθησης: (*i*) μοντέλο κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής (*ΚΑΜΠ*), (*ii*) μοντέλο με μηχανισμό αντιγραφής λέξεων εκτός λεξιλογίου (*ΑΛΕΛ*), (*iii*) μοντέλο ενισχυτικής μάθηση (*ΕΜ*), (*iv*) μοντέλο μετασχηματιστών (*ΜΣ*), (*v*) μοντέλο μετασχηματιστών με προ-εκπαιδευμένο κωδικοποιητή (*ΜΣΠΚ*).

Τα πέντε αυτά μοντέλα νευρωνικών δικτύων έχουν παρουσιαστεί με λεπτομέρεια στο Κεφάλαιο 3 και χρησιμοποιούνται στην πειραματική διαδικασία της προτεινόμενης προσέγγισης.

4.7 Φάση τρίτη: Μετα-επεξεργασία

Όπως έχει αναφερθεί παραπάνω (Ενότητες 4.4 και 4.6), η μεθοδολογία μηχανικής μάθησης εφαρμόζεται σε μια γενικευμένη έκδοση του αρχικού κειμένου και, κατά συνέπεια, οι παραγόμενες περιλήψεις είναι, επίσης, σε γενικευμένη μορφή. Αυτό σημαίνει ότι περιέχουν σημασιολογικά γενικευμένες έννοιες όπως υπερώνυμα ή ονοματικές οντότητες των αρχικών όρων. Επομένως, είναι απαραίτητη μια ακόμη διαδικασία, η φάση της μετα-επεξεργασίας των εκτιμώμενων (από τα μοντέλα μηχανικής μάθησης) περιλήψεων, προκειμένου να αντικατασταθούν οι γενικευμένες έννοιες της εκτιμώμενης περίληψης με συγκεκριμένες που θα δώσουν το τελικό νόημα στην περίληψη. Συνεπώς, η διαδικασία της μετα-επεξεργασίας διαμορφώνει την τελική περίληψη σε αναγνώσιμη από τον άνθρωπο μορφή. Ο Αλγόριθμος 4.3, που ακολουθεί, έχει ως αντικείμενο αυτήν την εργασία, καθώς μετατρέπει μια γενικευμένη περίληψη στην τελική της μορφή, αντικαθιστώντας γενικευμένες έννοιες της εκτιμώμενης (από τη διαδικασία μηχανικής μάθησης) περίληψης με τις αντίστοιχες συγκεκριμένες έννοιες, σύμφωνα με το αρχικό κείμενο.

4.7.1 Αλγόριθμος μετα-επεξεργασίας

Η φάση της μετα-επεξεργασίας βασίζεται στον Αλγόριθμο 4.3 ο οποίος περιγράφεται με λεπτομέρεια παρακάτω.

Ο Αλγόριθμος 4.3 δέχεται ως είσοδο την εκτιμώμενη από το σύστημα μηχανικής μάθησης περίληψη ($predSum$), το αντίστοιχο αρχικό κείμενο ($text$) και μια ταξινομία εννοιών T. Αρχικά, το σύνολο των υποψήφιων αντικαταστάσεων των γενικευμένων εννοιών cr και το σύνολο των γενικευμένων εννοιών gc αρχικοποιούνται σε κενά σύνολα (γραμμές 1-2). Στη συνέχεια, κάθε λέξη της περίληψης ($token_s$) διατρέχεται (γραμμές 3-13) και, αν κάποια λέξη είναι σε γενικευμένη μορφή (γραμμές 4-12), τότε προστίθεται στο σύνολο gc και το παράθυρο του κειμένου γύρω από τη λέξη της περίληψης ($token_s$) εκχωρείται στη μεταβλητή w_s (γραμμές 5-6). Ο βρόχος που ακολουθεί (γραμμές 7-11) διατρέχει κάθε λέξη ($token_t$) του αρχικού κειμένου ($text$), όπου το

Αλγόριθμος 4.3 Μετα-επεξεργασία: μετατροπή της γενικευμένης περίληψης ($predSum$) στην τελική της μορφή.

Require: $predSum$, $text$, T

1: $cr \leftarrow \{\}$ ▷ Υποψήφιες αντικαταστάσεις γενικευμένων εννοιών
2: $gc \leftarrow \{\}$ ▷ Γενικευμένες έννοιες
3: **for all** $token_s \in predSum$ **do**
4: **if** $token_s$ αντιστοιχεί σε γενικευμένη έννοια **then**
5: $gc \leftarrow gc \cup \{token_s\}$
6: $w_s \leftarrow$ Παράθυρο κειμένου γύρω από $token_s$
7: **for all** $token_t \in text$ **do**
8: $w_t \leftarrow$ Παράθυρο κειμένου γύρω από $token_t$
9: $sim \leftarrow$ Ομοιότητα μεταξύ w_s και w_t (Αλγόριθμος 4.4 ή 4.5)
10: $cr \leftarrow cr \cup \{(token_s,\ token_t,\ sim)\}$
11: **end for**
12: **end if**
13: **end for**
14: $summary \leftarrow$ Αντιστοίχιση των γενικευμένων εννοιών της $predSum$ με συγκεκριμένες έννοιες (βέλτιστη αντιστοίχιση εννοιών της Ενότητας 4.7.3 ή άπληστη αντιστοίχιση εννοιών της Ενότητας 4.7.3)
15: **return** $summary$

παράθυρο κειμένου γύρω από κάθε λέξη $token_t$ εκχωρείται στη μεταβλητή w_t (γραμμή 8). Η ομοιότητα κειμένου sim μεταξύ των παραθύρων w_s και w_t εκτιμάται σύμφωνα με τον Αλγόριθμο 4.4 ή 4.5 (γραμμή 9) και η πλειάδα $(token_s, token_t, sim)$ προστίθεται στο σύνολο cr. Στη γραμμή 14, ο Αλγόριθμος 4.3 εκτελεί μια αντιστοιχία των γενικευμένων εννοιών της εκτιμώμενης από το σύστημα μηχανικής μάθησης περίληψης $perdSum$ με συγκεκριμένες έννοιες, χρησιμοποιώντας τη βέλτιστη αντιστοίχιση εννοιών (Ενότητα 4.7.3) ή την άπληστη προσέγγιση (Ενότητα 4.7.3). Τέλος, ο αλγόριθμος επιστρέφει την περίληψη στην τελική της μορφή.

Προσδιορισμός γενικευμένων εννοιών: Στη γραμμή 4 του Αλγόριθμου 4.3, ελέγχεται κάθε λέξη της εκτιμώμενης περίληψης αν αντιστοιχεί σε γενικευμένη έννοια. Γενικευμένοι θεωρούνται οι όροι που φέρουν ονοματική οντότητα (από το σύνολο των προκαθορισμένων ονοματικών οντοτήτων) ή αναγνωριστικό *ΑΕΛ*. Για παράδειγμα, όταν η γενίκευση βασίζεται στην στρατηγική *ΓΟΟ*, οι γενικευμένες λέξεις αντιστοιχούν σε ονοματικές οντότητες (της μορφής *PERSON, LOC, ORG* κ.λπ.), ενώ στην περίπτωση που η γενίκευση βασίζεται στην στρατηγική *ΓΒΤ*, οι γενικευμένες έννοιες είναι της μορφής *tree.n.01, bank.n.02* κ.λπ.). Επιπλέον, σε περιπτώσεις συνδυασμού των προαναφερόμενων προσεγγίσεων γενίκευσης κειμένου όπως η *ΓΟΟ-ΓΒΤ*, οι γενικευμένες έννοιες αντιστοιχούν είτε σε ονοματικές οντότητες είτε σε αναγνωριστικά *ΑΕΛ*.

4.7.2 Ομοιότητα κειμένου

Μετρήσεις ομοιότητας μεταξύ αποσπασμάτων κειμένου, της γενικευμένης περίληψης και του αρχικού κειμένου (γραμμή 9, Αλγόριθμος 4.4), χρησιμοποιούνται, για να προσδιορίσουν τις συγκεκριμένες έννοιες ενός αρχικού κειμένου που ταιριάζουν περισσότερο με αντίστοιχες γενικευμένες έννοιες μιας γενικευμένης περίληψης. Οι έννοιες αυτές μπορούν να αντικαταστήσουν μια γενικευμένη έννοια της περίληψης, διαμορφώνοντας την τελική περίληψη στη φάση της μετα-επεξεργασίας. Για τις μετρήσεις αυτές, προτείνονται οι Αλγόριθμοι 4.4 και 4.5 για τις στρατηγικές γενίκευσης που χρησιμοποιούν *ΓΒΤ* (δηλ., *ΓΒΤ*και *ΓΒΤ-ΠΑΚ*) και *ΓΟΟ* (δηλ., *ΓΟΟ* και *ΓΟΟ-ΠΑΚ*), αντίστοιχα.

Αλγόριθμος 4.4 Προσδιορισμός της ομοιότητας μεταξύ αποσπασμάτων κειμένου που βασίζονται στη στρατηγική *ΓΒΤ*

Require: $token_s,\ token_t,\ token_g,\ w_s,\ w_t,\ T,\ a_1,\ a_2,\ a_3$
1: $hyponyms \leftarrow$ Υπώνυμα της λέξης $token_s$
2: $hypernyms \leftarrow$ Υπερώνυμα της λέξης $token_s$
3: $sim \leftarrow similarity\left((token_s, w_s),\ (token_t, w_t)\right)$
4: **if** $token_s \equiv\ token_g$ **then**
5: $\quad sim = a_1\ \cdot\ sim$
6: **else if** $token_t \in hyponyms$ **then**
7: $\quad sim = a_2\ \cdot\ sim$
8: **else if** $token_g \in hypernyms$ **then**
9: $\quad sim = a_3\ \cdot\ sim$
10: **end if**
11: **return** sim

Ομοιότητα κειμένου που βασίζεται σε *ΓΒΤ*

Ο Αλγόριθμος 4.4 εκτιμά την ομοιότητα μεταξύ των εννοιών για μοντέλα που βασίζονται στην *ΓΒΤ* (*ΓΒΤ*ή *ΓΒΤ-ΠΑΚ*). Ο αλγόριθμος δέχεται ως είσοδο μια λέξη ($token_s$) της εκτιμώμενης γενικευμένης περίληψης, μία λέξη του γενικευμένου υποψήφιου για περίληψη κειμένου ($token_g$), ένα παράθυρο κειμένου από τη γενικευμένη περίληψη (w_s), ένα παράθυρο κειμένου από το αρχικό κείμενο (w_t), μια ταξινομία εννοιών T και τις τρεις παραμέτρους a_1, a_2, a_3, οι οποίες καθορίζουν τη βαρύτητα της ομοιότητας μεταξύ w_s και w_t. Οι βέλτιστες τιμές των παραμέτρων a_1, a_2 και a_3 καθορίζονται πειραματικά. Αρχικά, ανακτώνται από την ταξινομία εννοιών τα μονοπάτια υπώνυμων και υπερώνυμων εννοιών της λέξης $token_s$ και εκχωρούνται σε αντίστοιχες μεταβλητές (γραμμές 1, 2). Στη συνέχεια, η ομοιότητα μεταξύ $token_s$ και $token_g$ υπολογίζεται και αποθηκεύεται στη μεταβλητή sim (γραμμή 3), για παράθυρα κειμένου μεγεθών w_s και w_t, αντίστοιχα, σύμφωνα με την επιλεγμένη συνάρτηση ομοιότητας (π.χ., ομοιότητα συνημιτόνου ή ομοιότητα με βάση την απόσταση μετακίνησης λέξεων η οποία είναι γνωστή ως *word mover's distance - WMD*). Σε περίπτωση που η λέξη $token_s$ ταυτίζεται με τη λέξη $token_g$ (γραμμή 4), δηλαδή όταν μια λέξη της γενικευμένης περίληψης *genSum* ταυτίζεται με μία λέξη του γενικευμένου κειμένου, η αρχικά μετρούμενη ομοιότητα (sim) ενισχύεται περαιτέρω σύμφωνα με την τιμή του συντελεστή

a_1. Διαφορετικά, εάν το $token_g$ βρίσκεται στο μονοπάτι υπώνυμων ή υπερώνυμων εννοιών της λέξης $token_s$, η ομοιότητα sim ενισχύεται σύμφωνα με τις τιμές των συντελεστών a_2 ή a_3, αντίστοιχα (γραμμές 6-10). Τέλος, ο αλγόριθμος επιστρέφει την ομοιότητα sim μεταξύ των δύο αποσπασμάτων κειμένου (γραμμή 11).

Ομοιότητα κειμένου που βασίζεται σε *ΓΟΟ*

Ο Αλγόριθμος 4.5 εκτιμά την ομοιότητα κειμένου για τη στρατηγική *ΓΟΟ* (*ΓΟΟ* ή *ΓΟΟ-ΠΑΚ*). Ο αλγόριθμος δέχεται ως είσοδο μία λέξη της γενικευμένης περίληψης ($token_s$) και ένα παράθυρο κειμένου γύρω από αυτή τη λέξη (w_s), μια λέξη του αρχικού κειμένου ($token_t$) και ένα παράθυρο κειμένου γύρω από αυτή τη λέξη (w_t), μια λέξη του γενικευμένου κειμένου ($token_g$), μια ταξινομία εννοιών T και την υπερ-παράμετρο b. Στην αρχή, υπολογίζεται η ομοιότητα μεταξύ των παραθύρων κειμένου w_s, w_t (γραμμή 1), σύμφωνα με τη συνάρτηση ομοιότητας που χρησιμοποιείται (η οποία μπορεί να είναι οποιαδήποτε από αυτές που συζητήθηκαν στην προηγούμενη παράγραφο). Εάν η λέξη $token_s$ ταυτίζεται με τη λέξη $token_g$, τότε η ομοιότητα κειμένου που υπολογίστηκε αρχικά πολλαπλασιάζεται με τον συντελεστή b (γραμμή 3), ο οποίος ενισχύει την τιμή της ομοιότητας σε αυτή την περίπτωση. Τέλος, ο αλγόριθμος επιστρέφει την ομοιότητα (sim) μεταξύ των δύο αποσπασμάτων κειμένου (γραμμή 5).

Αλγόριθμος 4.5 Προσδιορισμός της ομοιότητας μεταξύ αποσπασμάτων κειμένου που βασίζονται στη στρατηγική *ΓΟΟ*

Require: $token_s,\ token_t,\ token_g,\ w_s,\ w_t,\ T,\ b$
1: $sim \leftarrow similarity\,((token_s, w_s),\ (token_t, w_t))$
2: **if** $token_s \equiv\ token_g$ **then**
3: $\quad sim = b\ \cdot\ sim$
4: **end if**
5: **return** sim

Ομοιότητα κειμένου για τις στρατηγικές *ΓΟΟ-ΠΑΚ*, *ΓΟΟ-ΓΒΤ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*

Σε αυτές τις περιπτώσεις γενίκευσης, η τελική περίληψη παράγεται με την εφαρμογή του Αλγόριθμου 4.3 σε δύο διαδοχικά βήματα. Στο πρώτο βήμα, η γενικευμένη περίληψη μετατρέπεται σε μια ενδιάμεση μορφή με χρήση του Αλγόριθμου 4.4 για τον προσδιορισμό της ομοιότητας κειμένου *ΓΒΤ*. Στη συνέχεια, η ενδιάμεση περίληψη, που έχει προκύψει από το προηγούμενο βήμα, παρέχεται ως είσοδος στον Αλγόριθμο 4.3, ο οποίος εκτελείται για δεύτερη φορά. Σε αυτή την εκτέλεση του αλγόριθμου μετα-επεξεργασίας, για τον προσδιορισμό της ομοιότητας κειμένου *ΓΟΟ* χρησιμοποιείται ο Αλγόριθμος 4.5. Η παραγόμενη περίληψη που προκύπτει από το δεύτερο βήμα αποτελεί την τελική περίληψη.

4.7.3 Αντιστοίχιση εννοιών

Ο Αλγόριθμος 4.3 που εκτελεί την εργασία της μετα-επεξεργασίας απαιτεί μια διαδικασία αντικατάστασης των γενικευμένων εννοιών της εκτιμώμενης (από ένα μοντέλο μηχανικής μάθησης) περίληψης με συγκεκριμένες (γραμμή 14), προκειμένου να παραχθεί η τελική περίληψη σε αναγνώσιμη από τον άνθρωπο μορφή. Το πρόβλημα της αντιστοίχισης εννοιών μπορεί να διατυπωθεί ως ένα πρόβλημα διμερούς αντιστοίχισης μεταξύ των γενικευμένων και των συγκεκριμένων εννοιών που ανακτώνται από το αρχικό κείμενο. Η αναπαράσταση και απεικόνιση του προβλήματος μπορεί να γίνει με χρήση ενός διμερούς γράφου, που το πρώτο σύνολο κόμβων αποτελεί τις γενικευμένες έννοιες και το δεύτερο σύνολο κόμβων τις συγκεκριμένες έννοιες. Για την αντιστοίχιση αυτή ακολουθούν δύο μεθοδολογίες: (i) η βέλτιστη αντιστοίχιση μεταξύ των εννοιών και (ii) η άπληστη αντιστοίχιση μεταξύ των εννοιών.

Τόσο η βέλτιστη όσο και η άπληστη μεθοδολογία έχουν παρουσιαστεί στις εργασίες μας [138, 139], που αφορούν τον χώρο των συστημάτων συστάσεων (*recommender systems*), και έχουν σκοπό την αντιστοίχιση αντικειμένων με τις κατηγορίες τους για τη δημιουργία ενός πακέτου συστάσεων. Συνεπώς, κατά ανάλογο τρόπο για την αντιστοίχιση μεταξύ εννοιών, έγινε μεταφορά και προσαρμογή των μεθοδολογιών αυτών από τον χώρο των συστημάτων συστάσεων στον χώρο της αυτόματης *ΠΚ* [140]. Ακολουθεί η ανάλυση και η περιγραφή τόσο της βέλτιστης όσο και της άπληστης μεθοδολογίας αντιστοίχισης εννοιών.

Βέλτιστη αντιστοίχιση εννοιών

Το πρόβλημα της διμερούς αντιστοίχισης μεταξύ γενικευμένων και συγκεκριμένων εννοιών μπορεί να μοντελοποιηθεί ως πρόβλημα ροής ελάχιστου κόστους [141, 142, 143, 138, 139]. Πιο συγκεκριμένα, το γράφημα ροής ελάχιστου κόστους G (όπως, για παράδειγμα αυτό του Σχήματος 4.3) αποτελείται από έναν αρχικό κόμβο Src, ένα σύνολο από κόμβους που αντιστοιχούν στις γενικευμένες έννοιες N_G (δηλ., έννοιες της γενικευμένης περίληψης που χρειάζεται να αντικατασταθούν από συγκεκριμένες), ένα σύνολο κόμβων που αντιστοιχούν στις υποψήφιες συγκεκριμένες έννοιες N_C (δηλ., έννοιες που περιλαμβάνονται στο αρχικό έγγραφο) και ένας τερματικός κόμβος Trm. Επίσης, οι κόμβοι συνδέονται με ακμές που φέρουν τις πληροφορίες της ελάχιστης (min_f) και μέγιστης (max_f) ροής, καθώς και του κόστους ροής $cost_{i,j}$ μεταξύ δύο κόμβων (i, j). Το κόστος ροής υπολογίζεται σύμφωνα με την Εξίσωση 4.1 η οποία ακολουθεί.

$$cost_{ij} = \begin{cases} M - similarity(i,\ j), & i \in N_G \text{ and } j \in N_C \\ 0, & i = Src \text{ or } j = Trm \end{cases} \tag{4.1}$$

όπου M υποδηλώνει τη μέγιστη τιμή που μπορεί να πάρει η μετρική ομοιότητας και $similarity(i,\ j)$ είναι η ομοιότητα μεταξύ ενός παραθύρου κειμένου γύρω από τη γενικευμένη έννοια του κόμβου $i \in N_G$ (δηλ., ενός παραθύρου κειμένου γύρω από κάποια λέξη μιας γενικευμένης περίληψης) και ενός παραθύρου κειμένου γύρω από μια υποψήφια έννοια του κόμβου $j \in N_C$ (δηλ., ενός παραθύρου κειμένου γύρω από μία λέξη του αρχικού κειμένου).

Εναλλακτικά, αν χρησιμοποιηθεί μετρική που υπολογίζει την απόσταση μεταξύ αποσπασμάτων κειμένου $(distance(i,\ j))$, τότε ο όρος $M - similarity(i,\ j)$ της εξίσωσης 4.1 μπορεί να

αντικατασταθεί με την μετρούμενη απόσταση (δηλ., $distance(i,\ j) = M - similarity(i,\ j)$) για τον υπολογισμό του κόστους ροής.

Οι περιορισμοί min_f, max_f μεταξύ των κόμβων i, j εκφράζονται με τη μορφή πλειάδας, σύμφωνα με την Εξίσωση 4.2. Η ροή των ακμών μεταξύ του αρχικού κόμβου Src και των κόμβων N_G είναι ίση με 1, καθώς όλες οι γενικευμένες έννοιες πρέπει να ταιριάξουν με μία από τις υποψήφιες συγκεκριμένες έννοιες. Επιπλέον, η ροή μεταξύ των ακμών των κόμβων N_G και N_C και, με τη σειρά της, η ροή μεταξύ των ακμών των κόμβων N_C και του τερματικού κόμβου Trm είναι ίση με 0 ή 1, καθώς ο περιορισμός αυτός υποδηλώνει ότι μια γενικευμένη έννοια μπορεί να ταιριάξει με μία μόνο υποψήφια συγκεκριμένη έννοια. Τέλος, η ροή μεταξύ των κόμβων Src και Trm είναι ίση με τον αριθμό των γενικευμένων εννοιών $|N_G|$, καθώς όλες οι γενικευμένες έννοιες πρέπει να ταιριάξουν με τις υποψήφιες των κόμβων N_C.

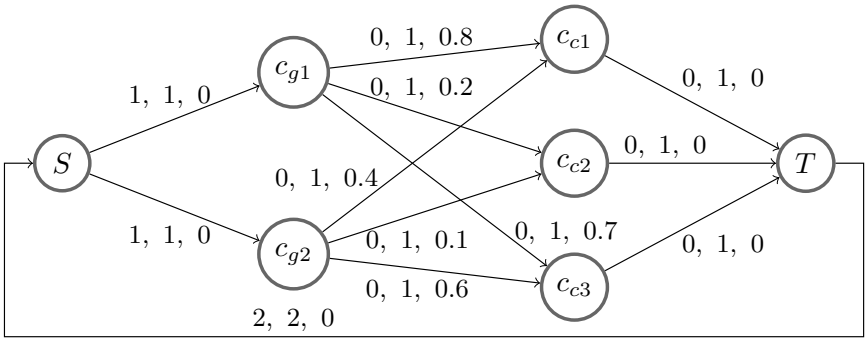
$$(min_{f_{ij}},\ max_{f_{ij}}) = \begin{cases} (1,\ 1), & i = Src,\ j \in N_G \\ (0,\ 1), & i \in N_G,\ j \in N_C \\ (0,\ 1), & i \in N_C,\ j = Trm \\ (|N_G|,\ |N_G|), & i = Src,\ j = Trm \end{cases} \tag{4.2}$$

Εναλλακτικά, το πρόβλημα ροής ελάχιστου κόστους μπορεί να αναδιατυπωθεί και να μοντελοποιηθεί και ως πρόβλημα ακέραιου γραμμικού προγραμματισμού [144], όπως εκφράζεται από τις Εξισώσεις 4.3, όπου $V = N_C \cup N_G \cup \{Src,\ Trc\}$ είναι το σύνολο των προαναφερόμενων κόμβων του προβλήματος ροής ελάχιστου κόστους.

$$\begin{aligned} \text{minimize:} \quad & \sum_{i,j \in V, i \neq j} cost_{ij} \cdot f_{ij} \\ \text{subject to:} \quad & min_{f_{ij}} \leq f_{ij} \leq max_{f_{ij}}, \forall i, j \in V,\ i \neq j \\ & \sum_{j \in V, i \neq j} f_{ij} = 0,\ \forall i \in V \\ & f_{ij} \in \mathbb{Z}_{\geq 0} \end{aligned} \tag{4.3}$$

Το πρόβλημα της βέλτιστης αντιστοίχισης εννοιών περιγράφεται περαιτέρω στον Αλγόριθμο 4.6 και στο Παράδειγμα 4.8 που ακολουθούν.

Ο Αλγόριθμος 4.6 περιγράφει τη διαδικασία εκτέλεσης της βέλτιστης αντιστοίχισης εννοιών μεταξύ των γενικευμένων εννοιών c_g της εκτιμώμενης περίληψης και των υποψήφιων εννοιών c_c του αρχικού κειμένου. Ο αλγόριθμος δέχεται ως είσοδο μια γενικευμένη περίληψη $genSum$ και ένα σύνολο υποψηφίων αντικαταστάσεων cr, το οποίο περιέχει πλειάδες που περιλαμβάνουν μια γενικευμένη έννοια, μία υποψήφια συγκεκριμένη έννοια και την τιμή της μεταξύ τους ομοιότητας. Αρχικά, δημιουργείται το γράφημα ροής (γραμμή 1) και επιλύεται το πρόβλημα προσδιορίζοντας τη ροή ελάχιστου κόστους (γραμμή 2, π.χ. με γραμμικό προγραμματισμό). Στη συνέχεια, οι ακμές με ροή $f_{ij} > 0$ αντιστοιχούν σε έννοιες που ταιριάζουν μεταξύ τους (γραμμή 3) και οι πλειάδες αυτών των εννοιών εκχωρούνται στο σύνολο των αντικαταστάσεων ($replacements$) (γραμμή 4). Στη συνέχεια, στη μεταβλητή $summary$ εκχωρείται η γενικευμένη περίληψη (γραμμή 5) και για κάθε πλειάδα του συνόλου αντικαταστάσεων, οι γενικευμένες έννοιες της περίληψης ($summary$)



Σχήμα 4.3: Ένα απλό παράδειγμα γραφήματος ροής ελάχιστου κόστους για βέλτιστη αντιστοίχιση μεταξύ γενικευμένων εννοιών c_{gi} και υποψηφίων συγκεκριμένων εννοιών c_{cj}, με τις ακμές μεταξύ των κόμβων να φέρουν την πληροφορία min_f, max_f και $cost$.

Αλγόριθμος 4.6 Βέλτιστη αντιστοίχιση εννοιών

Require: $genSum$, cr

1: Δημιουργείται το γράφημα ροής σύμφωνα με το σύνολο cr των υποψήφιων αντικαταστάσεων μεταξύ εννοιών
2: Επιλύεται το πρόβλημα ροής ελάχιστου κόστους
3: Οι ακμές με ροή $f_{ij} > 0$ υποδεικνύουν τις έννοιες που ταιριάζουν μεταξύ τους
4: $replacements \leftarrow$ Το σύνολο με τις πλειάδες $(c_{gi},\ c_{cj})$ των εννοιών που ταιριάζουν μεταξύ τους
5: $summary \leftarrow genSum$
6: **for all** $(c_{gi},\ c_{cj}) \in replacements$ **do**
7: $\quad summary \leftarrow$ Αντικατάσταση της έννοιας c_{gi} με την έννοια c_{cj}
8: **end for**
9: **return** $summary$

αντικαθίστανται από τις συγκεκριμένες έννοιες (γραμμές 6 - 8). Τέλος, ο αλγόριθμος επιστρέφει την τελική περίληψη $summary$ (γραμμή 9).

Παράδειγμα 4.8 (Βέλτιστη αντιστοίχιση εννοιών)**.**
Ας υποθέσουμε ότι έχουμε μια εκτιμώμενη περίληψη που περιλαμβάνει γενικευμένες έννοιες οι οποίες πρέπει να αντικατασταθούν με συγκεκριμένες του αρχικού κειμένου. Σε αυτή τη φάση της προτεινόμενης προσέγγισης, στη φάση της μετα-επεξεργασίας, σύμφωνα με τον Αλγόριθμο 4.3, θεωρούμε ότι η μέτρηση της ομοιότητας μεταξύ των εννοιών διαμόρφωσε το σύνολο των υποψήφιων αντικαταστάσεων cr ως εξής:

$$cr = \{(c_{g1},\ c_{c1},\ 0.2),\ (c_{g1},\ c_{c2},\ 0.8),\ (c_{g1},\ c_{c3},\ 0.3),$$
$$(c_{g2},\ c_{c1},\ 0.6),\ (c_{g2},\ c_{c2},\ 0.9),\ (c_{g2},\ c_{c3},\ 0.4),\}$$

όπου προέκυψαν δύο γενικευμένες έννοιες $c_{gi} \in N_G,\ i = \{1,\ 2\}$ και τρεις υποψήφιες έννοιες $c_{cj} \in N_C,\ j = \{1,\ 2,\ 3\}$. Το Σχήμα 4.3 απεικονίζει το αντίστοιχο γράφημα ροής, όπου οι ακμές μεταξύ των κόμβων φέρουν τους περιορισμούς της ελάχιστης και μέγιστης ροής καθώς και το κόστος ροής $(min_f, max_f, cost)$. Η ομοιότητα μεταξύ των εννοιών θεωρούμε ότι λαμβάνει τιμές στο διάστημα $[0, 1]$ (δηλ., $M = 1$) και στη συνέχεια μετατρέπεται σε κόστος, σύμφωνα με την

Εξίσωση 4.1.

Από τα δεδομένα του προβλήματος διαπιστώνουμε ότι η ροή μεταξύ των κόμβων $c_{g1} - c_{c2}$ και $c_{g2} - c_{c1}$ αποτελεί τη ροή ελάχιστου κόστους η οποία αντιστοιχεί και στο βέλτιστο ταίριασμα των αντίστοιχων εννοιών. Το ελάχιστο κόστος εκτιμάται σε $0,6$ (δηλ., $0,2+0,4$). Επομένως, οι έννοιες c_{g1}, c_{g2} της γενικευμένης περίληψης θα αντικατασταθούν από τις έννοιες c_{c2}, c_{c1}, αντίστοιχα.

Άπληστη αντιστοίχιση εννοιών

Με δεδομένο ότι η βέλτιστη προσέγγιση αντιστοίχισης εννοιών που συζητήθηκε παραπάνω παρουσιάζει υψηλή υπολογιστική πολυπλοκότητα για την αποδοτική επίλυση του προβλήματος γραμμικού προγραμματισμού [141, 145]. Εναλλακτικά, προτείνεται ένας άπληστος αλγόριθμος, ο οποίος έχει χαμηλότερη υπολογιστική πολυπλοκότητα και είναι εύκολα υλοποιήσιμος. Ο Αλγόριθμος 4.7, ο οποίος ταιριάζει τις γενικευμένες έννοιες της εκτιμώμενης περίληψης με τις υποψήφιες έννοιες του αρχικού κειμένου, δέχεται ως είσοδο μια γενικευμένη περίληψη $genSum$, μια λίστα υποψήφιων αντικαταστάσεων cr και μια λίστα γενικευμένων εννοιών gc. Αρχικά, οι πλειάδες του συνόλου cr ταξινομούνται κατά φθίνουσα σειρά σύμφωνα με την ομοιότητα sim (γραμμή 1) και η γενικευμένη περίληψη $genSum$ εκχωρείται στην μεταβλητή $summary$ (γραμμή 2). Στη συνέχεια, οι πλειάδες του συνόλου cr (γραμμή 3), που περιέχουν μια γενικευμένη λέξη ($token_s$) της εκτιμώμενης περίληψης, μια λέξη ($token_t$) του αρχικού κειμένου και τον βαθμό ομοιότητάς τους sim, εξετάζονται σε φθίνουσα σειρά ομοιότητας (γραμμές 3-11). Εάν η λέξη $token_s$ ανήκει στο σύνολο gc (γραμμή 4), αντικαθίσταται στην περίληψη ($summary$) από τη λέξη $token_t$ (γραμμή 5) και, στη συνέχεια, η λέξη $token_s$ αφαιρείται από το σύνολο gc (γραμμή 6). Μόλις αντικατασταθούν όλες οι γενικευμένες έννοιες του συνόλου gc από συγκεκριμένες ή εξεταστούν όλες οι υποψήφιες αντικαταστάσεις του συνόλου cr, ο αλγόριθμος τερματίζει και επιστρέφει την τελική περίληψη (γραμμή 12). Το Παράδειγμα 4.9 που ακολουθεί περιγράφει περαιτέρω την άπληστη αντιστοίχιση εννοιών.

Αλγόριθμος 4.7 Άπληστη αντιστοίχιση εννοιών

Require: $genSum$, cr, gc

1: Ταξινόμηση του συνόλου cr σε φθίνουσα σειρά ταξινόμησης σύμφωνα με την ομοιότητα sim
2: $summary \leftarrow genSum$
3: **for all** $(token_s,\ token_t,\ sim) \in cr$ **do**
4: **if** $token_s \in gc$ **then**
5: $summary \leftarrow$ Αντικατάσταση της έννοιας $token_s$ με την έννοια $token_t$
6: $gc \leftarrow gc \setminus token_s$
7: **end if**
8: **if** $gc \equiv \{\}$ **then**
9: Τερματισμός δομής επανάληψης for
10: **end if**
11: **end for**
12: **return** $summary$

Παράδειγμα 4.9 (Άπληστη αντιστοίχιση εννοιών)**.**
Με δεδομένο το σύνολο των υποψήφιων αντικαταστάσεων cr του Παραδείγματος 4.8, ο άπληστος αλγόριθμος, αφού ταξινομήσει τις πλειάδες του συνόλου cr κατά φθίνουσα σειρά ανάλογα με την τιμή ομοιότητας μεταξύ των εννοιών που περιλαμβάνει κάθε πλειάδα, ταιριάζει τα ζεύγη εννοιών ως εξής: $c_{g2} - c_{c2}$ *και* $c_{g1} - c_{c3}$*. Τα ζεύγη εννοιών που προκύπτουν αποτελούν το ταίριασμα εννοιών με τον υψηλότερο βαθμό ομοιότητας σύμφωνα με την εκτέλεση του Αλγόριθμου 4.7.*

Από τα Παραδείγματα 4.8 και 4.9 που παρουσιάστηκαν παραπάνω, είναι προφανές ότι η άπληστη αντιστοίχιση μπορεί να είναι διαφορετική από τη βέλτιστη, καθώς η πρώτη βασίζεται σε έναν κατά προσέγγιση αλγόριθμο. Παρά τις διαφορές τους, και οι δύο προσεγγίσεις στην πράξη παρουσιάζουν παρόμοια απόδοση, επειδή η αποτελεσματικότητά τους εξαρτάται κυρίως από την επιλεγμένη συνάρτηση ομοιότητας, όπως θα δούμε και στο επόμενο κεφάλαιο που περιλαμβάνει το πειραματικό μέρος της προτεινόμενης προσέγγισης.

4.7.4 Υπολογιστική πολυπλοκότητα

Η υπολογιστική πολυπλοκότητα του Αλγόριθμου 4.3 επηρεάζεται σε μεγάλο βαθμό από τη μεθοδολογία αντιστοίχισης εννοιών της γραμμής 14. Στην περίπτωση χρήσης της βέλτιστης αντιστοίχισης εννοιών (Ενότητα 4.7.3), η οποία μοντελοποιείται με όρους ακέραιου γραμμικού προγραμματισμού, έχουμε ένα πρόβλημα που ανήκει στην κλάση *NP-πλήρης* (*NP-complete*, δηλ., πρόκειται για ένα δύσκολο πρόβλημα χωρίς γνωστό αλγόριθμο πολυωνυμικού χρόνου) [146]. Ωστόσο, στο πρόβλημα που εξετάζουμε, όπου ο αριθμός των γενικευμένων και υποψήφιων εννοιών είναι περιορισμένος, είναι εφικτό να βρεθεί μια βέλτιστη λύση σε λογικό χρόνο εκτέλεσης.

Σε αντίθεση με τη βέλτιστη αντιστοίχιση εννοιών, η άπληστη λύση (Αλγόριθμος 4.7), η οποία έχει περιγραφεί παραπάνω στην Ενότητα 4.7.3, αποτελεί μια εναλλακτική πιο αποδοτική (αλλά όχι μεγαλύτερης ακρίβειας) μεθοδολογία για τη λύση του προβλήματος. Ειδικότερα, υποθέτοντας ότι ο αριθμός των στοιχείων του συνόλου των υποψήφιων αντικαταστάσεων cr είναι ίσος με n (δηλ., $|cr| = n$), η υπολογιστική πολυπλοκότητα, στη χειρότερη περίπτωση, της ταξινόμησης cr είναι $O(nlogn)$ (π.χ., χρησιμοποιώντας ταξινόμηση σωρού - *heap sort*] [147] και η πολυπλοκότητα της δομής επανάληψης (γραμμές 3 - 11) είναι γραμμική ($O(n)$). Επομένως, η συνολική υπολογιστική πολυπλοκότητα του Αλγόριθμου 4.7 είναι $O(n \log n)$.

Έχοντας προσδιορίσει την υπολογιστική πολυπλοκότητα και των δύο μεθόδων αντιστοίχισης εννοιών (βέλτιστη και άπληστη μεθοδολογία), μπορούμε να εκτιμήσουμε την πολυπλοκότητα της συνολικής διαδικασίας του αλγόριθμου μετα-επεξεργασίας (Αλγόριθμος 4.3). Ο βρόχος του αλγόριθμου (γραμμές $3 - 13$), ο οποίος περιλαμβάνει και έναν εμφωλευμένο βρόχο (γραμμές 7-11) παρουσιάζει πολυπλοκότητα $O(n^2)$, καθώς εξετάζονται όλες οι λέξεις εισόδου πλήθους n της εκτιμώμενης περίληψης και όλες οι λέξεις πλήθους $k \cdot n$ του αρχικού κειμένου (στον προσδιορισμό της πολυπλοκότητας απαλείφεται η σταθερά k). Σύμφωνα με το βήμα αντιστοίχισης εννοιών της γραμμής 14, όπως αναφέρθηκε και παραπάνω, αν χρησιμοποιηθεί η βέλτιστη λύση, τότε ο Αλγόριθμος 4.3 ανήκει στην κλάση πολυπλοκότητας *NP-πλήρης*. Διαφορετικά, αν επιλεγεί η άπληστη προσέγγιση, ο Αλγόριθμος 4.3 εκτελείται με πολυπλοκότητα $O(n^2)$ (δηλ., την πολυπλοκότητα του βρόχου των γραμμών $3 - 13$, καθώς η άπληστη αντιστοίχιση εκτελείται με μικρότερη πολυπλοκότητα $O(n \log n)$). Στην πράξη, ο αντίκτυπος της υπολογιζόμενης

πολυπλοκότητας (χειρότερης περίπτωσης) στην απόδοση του συστήματος γίνεται εμφανής μόνο σε περιπτώσεις πολύ μεγάλου όγκου δεδομένων εισόδου.

Κεφάλαιο 5

Πειραματικό μέρος για την αυτόματη περίληψη Κειμένου με χρήση μηχανικής μάθησης και σημασιολογικών μετασχηματισμών περιεχομένου

5.1 Γενικά

Σε αυτό το κεφάλαιο παρουσιάζεται η πειραματική διαδικασία, η οποία έχει σκοπό τη διερεύνηση σημαντικών πτυχών του προτεινόμενου πλαισίου για την αυτόματη *ΠΚ*. Η εκτενής διαδικασία αξιολόγησης, που διεξάγεται, βασίζεται στη μεθοδολογία που ακολουθούν άλλες συναφείς εργασίες της σχετικής βιβλιογραφίας [104, 60, 59, 61, 63].

Το κεφάλαιο αυτό διαρθρώνεται ως εξής: Η Ενότητα 5.2 που ακολουθεί περιγράφει τα χρησιμοποιούμενα συνόλων δεδομένων, η Ενότητα 5.3 παρουσιάζει τις μετρικές αξιολόγησης, η Ενότητα 5.4 περιγράφει την πειραματική διαδικασία και η Ενότητα 5.5 παραθέτει την επίδοση άλλων σύγχρονων προσεγγίσεων της σχετικής βιβλιογραφίας για λόγους σύγκρισης, καθώς και την επίδοση βασικών προσεγγίσεων ως σημείο αναφοράς για την προσέγγιση που εξετάζουμε. Τα πειραματικά αποτελέσματα και η ερμηνεία τους παρουσιάζονται στις Ενότητες 5.6 και 5.7, αντίστοιχα. Το κεφάλαιο ολοκληρώνεται με τα συμπεράσματα και τις προοπτικές για μελλοντική εργασία που αναφέρονται στην Ενότητα 5.8.

5.2 Σύνολα δεδομένων

Η προτεινόμενη προσέγγιση αξιολογείται με χρήση τριών συνόλων δεδομένων τα οποία χρησιμοποιούνται ευρέως στο πεδίο της αυτόματης *ΠΚ*, το οποία είναι: το *Gigaword* [105], το *DUC 2004* [106] και το *CNN/DailyMail* [107]. Τα τρία αυτά σύνολα δεδομένων έχουν περιγραφεί

με λεπτομέρεια στην Ενότητα 3.4.1,

Πίνακας 5.1: Η κατανομή των ουσιαστικών και των ρημάτων στα σύνολα δεδομένων.

	Ουσιαστικά		Ρήματα	
Gigaword (σύνολο εκπαίδευσης)	46,989,593	69,77%	20,362,923	30,23%
Gigaword (σύνολο ελέγχου)	24,936	69,70%	10,840	30,30%
Gigaword (σύνολο επικύρωσης)	24,883	70,23%	10,548	29,77%
DUC 2004	11,023	65,57%	5,787	34,43%
CNN/DailyMail (σύνολο εκπαίδευσης)	29,165,327	59,73%	19,665,826	40,27%
CNN/DailyMail (σύνολο ελέγχου)	960,193	62,39%	578,803	37,61%
CNN/DailyMail (σύνολο επικύρωσης)	1,121,401	62,54%	671,765	37,46%

Ο Πίνακας 5.1 παρουσιάζει την κατανομή των ουσιαστικών και των ρημάτων στα τρία σύνολα δεδομένων, καθώς η προτεινόμενη προσέγγιση χρησιμοποιεί τα συγκεκριμένα μέρη του λόγου στη διαδικασία σημασιολογικού μετασχηματισμού του περιεχομένου. Η κατανομή προέκυψε σύμφωνα με τη μεθοδολογία αυτόματης επισημείωσης του μέρους του λόγου (*part-of-speech tagging - POS tagging*) που χρησιμοποιήθηκε. Είναι προφανές ότι τα ουσιαστικά κυριαρχούν σε πλήθος σε σχέση με τα ρήματα.

Τέλος, πρέπει να σημειωθεί, ότι η παρούσα εργασία διατηρεί τη μορφή των συνόλων δεδομένων όπως χρησιμοποιήθηκαν από τις σχετικές προσεγγίσεις (Πίνακες 5.3 και 5.4). Επομένως, είναι δυνατή μια άμεση σύγκριση μεταξύ των πειραματικών αποτελεσμάτων της προτεινόμενης λύσης και άλλων σχετικών προσεγγίσεων.

5.3 Μετρικές αξιολόγησης

5.3.1 Το σύνολο μετρικών Rouge

Η αξιολόγηση της προτεινόμενης προσέγγισης βασίζεται στο σύνολο μετρικών *Rouge* [109] και πιο συγκεκριμένα, χρησιμοποιούνται οι μετρικές $Rouge_1$ (βαθμός επικάλυψης λέξεων), $Rouge_2$ (επικάλυψη δύο διαδοχικών λέξεων) και $Rouge_L$ (επκάλυψη μεταξύ της μεγαλύτερης κοινής ακολουθίας). Περισσότερες λεπτομέρειες για τις χρησιμοποιούμενες μετρικές του πακέτου *Rouge* αναφέρθηκαν στην Ενότητα 3.4.2 που προηγήθηκε.

5.3.2 Ακρίβεια απόδοσης πληροφορίας

Επιπροσθέτως, για να αξιολογήσουμε τη συνέπεια των αναγραφόμενων πληροφοριών [148, 149] που περιλαμβάνει η εκτιμώμενη περίληψη, θεωρούμε μια μετρική που προσδιορίζει την *ακρίβεια απόδοσης πληροφορίας (ΑΑΠ)*. Πιο συγκεκριμένα, παρόμοια με την προσέγγιση της εργασίας [149], σχέσεις κατηγορήματος – ορίσματος (*predicate-argument relations*), όπως τριάδες της μορφής *(οντότητα, σχέση, οντότητα)* ή *(οντότητα, ιδιότητα, τιμή)* ή *(υποκείμενο, κατηγόρημα, αντικείμενο)*, ανακτώνται από το αρχικό κείμενο και την εκτιμώμενη περίληψη με σκοπό να

εκτιμηθεί ο βαθμός επικάλυψης μεταξύ των τριάδων πληροφορίας που περιλαμβάνονται στην εκτιμώμενη περίληψη και εκείνων του αρχικού κειμένου. Μπορούμε να πούμε, ότι οι τριάδες αυτές ακολουθούν το πρότυπο *RDF (Resource Description Framework)* [150] και στο εξής θα ονομάζονται *τριάδες πληροφορίας*. Η μέτρηση της επικάλυψης αυτής, (δηλ., ο βαθμός επικάλυψης των τριάδων πληροφορίας μεταξύ μιας εκτιμώμενης περίληψης και του αρχικού κειμένου), προσδιορίζει την *ΑΑΠ*, η οποία περιλαμβάνεται στο αρχικό κείμενο και αποδίδεται από την εκτιμώμενη περίληψη. Σε αυτή την κατεύθυνση, θεωρούμε τη μετρική *ΑΑΠ* (γνωστή και ως *factual accuracy* ή *factual consistency*), η οποία συμβολίζεται ως $fact_{acc}$ και ορίζεται ως ο λόγος του πλήθους των κοινών *τριάδων πληροφορίας*, οι οποίες περιέχονται ταυτόχρονα στην περίληψη και στο αρχικό κείμενο, προς το πλήθος των τριάδων πληροφορίας του αρχικού κειμένου. Δηλαδή είναι μια μετρική που βασίζεται στην ανάκληση (*recall-based*) και η τιμή της υπολογίζεται σύμφωνα με την Εξίσωση 5.1.

$$fact_{acc} = \frac{|F_t \cap F_{Se}|}{|F_t|} \tag{5.1}$$

όπου F_t είναι ένα σύνολο *τριάδων πληροφορίας* από το αρχικό κείμενο και F_{Se} είναι ένα σύνολο *τριάδων πληροφορίας* της εκτιμώμενης περίληψης. Στο πειραματικό μέρος, οι *τριάδες πληροφορίας* ανακτώνται σύμφωνα με την προσέγγιση *(OpenIE)* [151, 149] που χρησιμοποιείται για εξαγωγή πληροφορίας από κείμενα.

5.3.3 Ποσοστό νέων λέξεων

Με δεδομένο ότι η συγκεκριμένη εργασία ακολουθεί τη μέθοδο της παραγωγής κειμένου για την αυτόματη *ΠΚ*, μας ενδιαφέρει να εξετάσουμε τον βαθμό που η προτεινόμενη προσέγγιση παράγει νέο περιεχόμενο στις εκτιμώμενες περιλήψεις, το οποίο δεν περιλαμβάνεται στο αρχικό κείμενο. Σε αυτή την κατεύθυνση, η μετρική *NTR* (*New Tokens Rate*) προσδιορίζει το ποσοστό των νέων λέξεων που εμφανίζονται στην εκτιμώμενη περίληψη, αλλά δεν συμπεριλαμβάνονται στο αρχικό κείμενο. Πιο συγκεκριμένα, σύμφωνα με την εξίσωση 5.2, το πλήθους των λεκτικών μονάδων που περιλαμβάνονται σε μια περίληψη, αλλά δεν περιλαμβάνονται στο αντίστοιχο κείμενο ($|S \setminus T|$), προς το πλήθος των λεκτικών μονάδων της περίληψης ($|S|$) δίνει την τιμή της μετρικής (NTR).

$$NTR = \frac{|S \setminus T|}{|S|} \tag{5.2}$$

όπου με S συμβολίζεται το σύνολο των λέξεων μιας περίληψης και με T το σύνολο των λέξεων του αντίστοιχου κειμένου.

5.4 Πειραματική διαδικασία και βελτιστοποίηση παραμέτρων

5.4.1 Εφαρμογή γενίκευσης περιεχομένου

Η μεθοδολογία γενίκευσης περιεχομένου (Ενότητα 4.5.2) εφαρμόζεται στα δεδομένα εκπαίδευσης, ελέγχου και επικύρωσης, προκειμένου να εξεταστεί η επίδρασή της στην απόδοση

της προτεινόμενη μεθοδολογίας. Συγκεκριμένα, εξετάζονται οι στρατηγικές γενίκευσης *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* (Ενότητα 4.5.3), μαζί με τις παραλλαγές τους σε πλήρως αποσαφηνισμένο κείμενο (*ΓΒΤ-ΠΑΚ*, *ΓΟΟ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*). Για καθεμία από αυτές τις στρατηγικές, εξετάζονται διάφορες τιμές των ορίων θ_f και θ_d, προκειμένου να μελετηθεί η επίδρασή τους στα αποτελέσματα που λαμβάνονται. Τέλος, να αναφερθεί, ότι η ταξινομία εννοιών που χρησιμοποιείται είναι αυτή του ηλεκτρονικού λεξικού *WordNet* [43, 44].

Ως προς τα μέρη του λόγου, η γενίκευση περιεχομένου εφαρμόζεται είτε:

(i) σε ουσιαστικά είτε

(ii) σε ρήματα και ουσιαστικά.

Τα μέρη του λόγου που ανήκουν οι λέξεων του κειμένου προσδιορίζονται χρησιμοποιώντας μεθοδολογία αυτόματης επισημείωσης (*POS tagging*) [152].

Επιπλέον, στο πλαίσιο εφαρμογής των στρατηγικών γενίκευσης που βασίζονται σε *ΓΟΟ* έχει χρησιμοποιηθεί τεχνολογία αναγνώρισης ονοματικών οντοτήτων (*ΑΟΟ*) για αυτόματη επισημείωση (*NER tagging*) [153]. Ο Πίνακας 5.2 περιγράφει τους τύπους ονοματικών οντοτήτων που χρησιμοποιούνται σε αυτή την εργασία. Η διαδικασία *ΑΟΟ* χρησιμοποιεί ένα προ-εκπαιδευμένο μοντέλο, το οποίο έχει εκπαιδευτεί στο σύνολο δεδομένων *OntoNotes 5.0* [154]. Οι διαδικασίες επεξεργασίας φυσικής γλώσσας που χρησιμοποιήθηκαν για τον προσδιορισμό και επισημείωση των μερών του λόγου και των ονοματικών οντοτήτων βασίζονται σε σύγχρονους αλγόριθμους υψηλής απόδοσης [155, 156].

Επιπλέον, ο μηχανισμός αποσαφήνισης της έννοιας των λέξεων (*ΑΕΛ*), ο οποίος βασίζεται σε πόρους γνώσης, είναι σε συμφωνία με τη χρησιμοποιούμενη ταξινομία εννοιών (*WordNet*), καθώς τα αναγνωριστικά *ΑΕΛ* που καθορίζουν την έννοια των λέξεων (π.χ., *play.n.01*) ταυτίζονται με τους όρους της ταξινομίας εννοιών. Ως μηχανισμός *ΑΕΛ*, χρησιμοποιείται ο εκτεταμένος αλγόριθμος του *Lesk* [157, 50].

Είναι γνωστό ότι το ηλεκτρονικό λεξικό *WordNet* είναι οργανωμένο σε σύνολα συνώνυμων εννοιών (Ενότητα 2.3.3). Στην περίπτωση γενίκευσης μιας έννοιας, αν η έννοια με ευρύτερο σημασιολογικό πεδίο ανήκει σε ένα σύνολο συνωνύμων του *WordNet*, το οποίο περιλαμβάνει περισσότερες από μια έννοιες, τότε επιλέγεται η έννοια με την υψηλότερη συχνότητα παρουσίας στο σύνολο εκπαίδευσης. Αν υπάρχουν έννοιες στο εξεταζόμενο σύνολο συνώνυμων εννοιών που έχουν την ίδια συχνότητα στο σύνολο εκπαίδευση, τότε επιλέγουμε εκείνη που έχει την υψηλότερη συχνότητα χρήσης σύμφωνα με το *WordNet*. Διαφορετικά, επιλέγεται τυχαία μια έννοια από το συγκεκριμένο σύνολο των συνώνυμων εννοιών. Η επιλογή της έννοιας με την υψηλότερη συχνότητα οδηγεί στην επαρκή παρουσία των εννοιών στο σύνολο εκπαίδευσης, που είναι το ζητούμενο για την αποτελεσματική εκπαίδευση των μοντέλων μηχανικής μάθησης και τη βελτίωση των προβλέψεων.

Ως αποτέλεσμα της διαδικασίας, που περιγράφεται παραπάνω, για τις διάφορες στρατηγικές γενίκευσης που εξετάζονται, δημιουργούνται αντίστοιχες εκδόσεις των συνόλων δεδομένων που χρησιμοποιούνται στην εκπαίδευση και έλεγχο των μοντέλων μηχανικής μάθησης με σκοπό την αξιολόγηση της μεθοδολογίας. Η επίδραση των συγκεκριμένων σχημάτων γενίκευσης και η απόδοση των αντίστοιχων μοντέλων μηχανικής μάθησης περιγράφονται περαιτέρω στην Ενότητα 5.6.

Πίνακας 5.2: Οι ετικέτες ονοματικών οντοτήτων που χρησιμοποιούνται στα σχήματα γενίκευση που βασίζονται σε *ΓΟΟ*.

PERSON (Άνθρωποι)	*LAW* (Νομικά έγγραφα)
NORP (Εθνικότητες, θρησκευτικές ομάδες κ.λπ.)	*LANGUAGE* (Γλώσσες)
FAC (Εγκαταστάσεις, κτήρια, αεροδρόμια κ.λπ.)	*DATE* (Ημερομηνίες και περίοδοι)
ORG (Οργανισμοί, εταιρείες κ.λπ.)	*TIME* (Χρόνος μικρότερος από μία ημέρα)
GPE (Χώρες, πόλεις, πολιτείες)	*PERCENT* (Ποσοστά)
LOC (Τοποθεσίες εκτός *GPE*)	*MONEY* (Χρηματικές αξίες με την μονάδα)
PRODUCT (Αντικείμενα, οχήματα, τρόφιμα κ.λπ.)	*QUANTITY* (Βάρος, απόσταση κ.λπ.)
EVENT (Ονομασίες γεγονότων, εκδηλώσεις κ.λπ.)	*ORDINAL* (Πρώτο, δεύτερο κ.λπ.)
WORK_OF_ART (Βιβλία, ταινίες κ.λπ.)	*CARDINAL* (Αριθμοί, εκτός άλλων τύπων)

5.4.2 Ενσωματώσεις λέξεων

Για τη διανυσματική αναπαράσταση των λέξεων χρησιμοποιούνται ενσωματώσεις λέξεων και πιο συγκεκριμένα, το μοντέλο *Word2Vec* με 300 διαστάσεις ανά διάνυσμα. Περισσότερες λεπτομέρειες για την εκπαίδευση του μοντέλου *Word2Vec* έχουν αναφερθεί στην Ενότητα 3.2.1.

5.4.3 Εκπαίδευση μοντέλων μηχανικής μάθησης

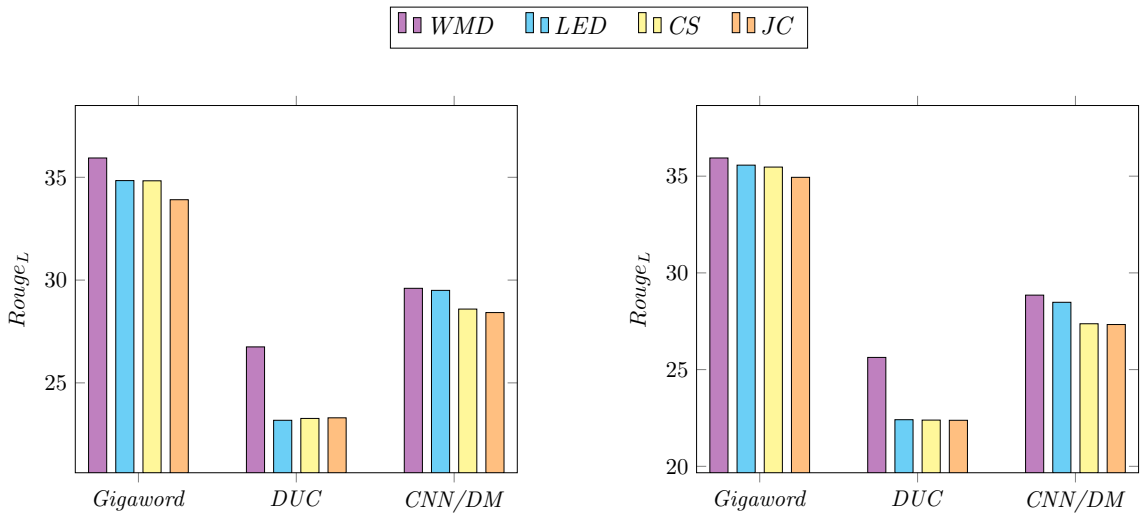
Για την εκπαίδευση των 5 μοντέλων μηχανικής μάθησης που χρησιμοποιούνται σε αυτή την εργασία, τα οποία αναφέρθηκαν στην Ενότητα 4.6.2 και παρουσιάστηκαν με λεπτομέρεια στην Ενότητα 3.3, ισχύουν οι ίδιες παραδοχές όσον αφορά την επιλογή των παραμέτρων με εκείνες που αναφέρθηκαν στην Ενότητα 3.4.3. Διατηρούνται οι παράμετροι αυτοί, καθώς η βελτιστοποίηση των μοντέλων μηχανικής μάθησης έγινε με χρήση των συνόλων δεδομένων στην αρχική τους κατάσταση, η οποία είναι σε συμφωνία με την πειραματική διαδικασία της Ενότητας 3.4.3.

5.4.4 Παραγωγή εκτιμώμενων περιλήψεων μοντέλων μηχανικής μάθησης

Για τη μεγιστοποίηση της κατανομής πιθανοτήτων στο λεξιλόγιο εξόδου της εκτιμώμενης περίληψης, χρησιμοποιούμε τον αλγόριθμο αναζήτησης δέσμης (*beam search*, Ενότητα 3.3.1) με πλάτος δέσμης ίσο με 4, όπως έχει περιγραφεί και στην Ενότητα 3.4.3.

5.4.5 Βελτιστοποίηση παραμέτρων μετα-επεξεργασίας

Η διαδικασία της μετα-επεξεργασίας που έχει περιγραφεί με λεπτομέρεια στην Ενότητα 4.7 απαιτεί μια σειρά από επιλογές παραμέτρων με σκοπό τη βελτιστοποίηση των αποτελεσμάτων. Πιο συγκεκριμένα, για το σύνολο δεδομένων *Gigaword*, η βέλτιστη τιμή των συντελεστών a_1, a_2 και a_3 του Αλγορίθμου 4.4 έχει προσδιοριστεί σε $2,0$, $1,5$ και $1,5$, αντίστοιχα, ενώ ο συντελεστής b του Αλγόριθμου 4.5 έχει οριστεί σε $2,0$ (Οι τιμές αυτές προσδιορίστηκαν μετά από διεξοδική πειραματική διαδικασία).



(α′) *ΓΒΤ* ($\theta_f = 200$ και $\theta_d = 5$) (β′) *ΓΟΟ* ($\theta_f = 200$)

Σχήμα 5.1: Οι τιμές της μετρικής αξιολόγησης $Rouge_L$ (f_1) για τις διάφορες μεθοδολογίες μέτρησης ομοιότητας ή απόστασης κειμένου (*WMD*: απόσταση μεταφοράς λέξεων, *LED*: απόσταση επεξεργασίας *Levenshtein*, *CS*: ομοιότητα συνημιτόνου, *JC*: συντελεστής *Jaccard*) που εξετάζονται στη φάση της μετα-επεξεργασίας για την αντιστοίχιση των εννοιών.

Επιπλέον, έχει εξεταστεί μια σειρά από διαφορετικές μετρικές ομοιότητας ή απόστασης κειμένου για τους Αλγόριθμους 4.4 και 4.5, όπως η *ομοιότητα συνημιτόνου (cosine similarity -CS)*, ο *Συντελεστής Jaccard (Jaccard coefficient - JC)*, η *απόσταση μεταφοράς λέξεων (word mover's distance -WMD)* [158] και η *απόσταση επεξεργασίας Levenshtein (Levenshtein edit distance - LED)* [159]. Η μετρική *WMD*, η οποία αξιοποιεί τις διανυσματικές αναπαραστάσεις των λέξεων και λαμβάνει υπόψη της συντακτικές και σημασιολογικές πτυχές του κειμένου, φαίνεται ότι επιτυγχάνει τα καλύτερα αποτελέσματα (Σχήμα 5.1).

Σχετικά με το μέγεθος των παραθύρων κειμένου γύρω από τις υποψήφιες και τις γενικευμένες έννοιες, μετά από πειραματική διαδικασία, η βέλτιστη απόδοση του αλγόριθμου μετα-επεξεργασίας επιτυγχάνεται όταν οι λέξεις των παραθύρων αυτών ορίζονται σε $w_t = 10$ και $w_s = 6$, αντίστοιχα.

Σχετικά με τα πειράματα βελτιστοποίησης του αλγόριθμου της μετα-επεξεργασίας, οι βέλτιστες τιμές των παραμέτρων προσδιορίστηκαν μέσω εξαντλητικής αναζήτησης πλέγματος για τη μεγιστοποίηση της τιμής $Rouge_L$ (f_1) στα παραδείγματα επικύρωσης του συνόλου δεδομένων *Gigaword*. Οι τιμές των παραμέτρων διατηρήθηκαν σταθερές στα σύνολα δεδομένων *DUC 2004* και *CNN/DailyMail*. Το Σχήμα 5.1 απεικονίζει τις τιμές της μετρικής $Rouge_L$ (f_1) για τις προαναφερόμενες μετρήσεις ομοιότητας κειμένου των στρατηγικών γενίκευσης κειμένου *ΓΒΤ* και *ΓΟΟ*. Η μετρική ομοιότητας κειμένου WMD τείνει να υπερτερεί των άλλων προσεγγίσεων, ειδικά στο σύνολο δεδομένων *DUC 2004*.

Τέλος, μέσα από την πειραματική διαδικασία διαπιστώθηκε ότι και οι δύο προσεγγίσεις αντιστοίχισης εννοιών, η βέλτιστη και η άπληστη (Ενότητα 4.7.3), παρουσιάζουν σχεδόν την ίδια απόδοση ως προς τις μετρικές αξιολόγησης *Rouge*. Για τον λόγο αυτό, η βέλτιστη τεχνική αντιστοίχισης εννοιών επιλέχθηκε να εφαρμοστεί στα τελικά πειράματα, σε μια προσπάθεια προσδιορισμού της

βέλτιστης απόδοσης της προτεινόμενης προσέγγισης με τη λήψη των μέγιστων δυνατών τιμών σε όρους μετρικών *Rouge*. Ωστόσο, για να δείξουμε τις διαφορές μεταξύ της βέλτιστης και άπληστης μεθοδολογίας αντιστοίχισης εννοιών, εκτελέσαμε αντίστοιχα πειράματα με τα αποτελέσματα να αναφέρονται στην Ενότητα 5.6.5.

Πίνακας 5.3: Η επίδοση σε όρους *Rouge* και *NTR* για τα σύνολα δεδομένων *Gigaword* και *DUC 2004* άλλων πρόσφατων και σχετικών προσεγγίσεων.

	Gigaword				*DUC 2004*		
Προσέγγιση	$Rouge_1$	$Rouge_2$	$Rouge_L$	*NTR* (%)	$Rouge_1$	$Rouge_2$	$Rouge_L$
ABS+ [104]	31,00	12,65	28,34	8,50	28,18	8,49	23,81
RAS-Elman [59]	33,78	15,97	31,15	-	28,97	8,26	24,06
Words-lvt2k-1sent [60]	34,97	17,17	32,70	24,15	28,35	9,46	24,59
Words-lvt5k-1sent [60]	35,30	16,64	32,62	-	28,61	9,42	25,24
Model#8 [160]	35,30	17,58	32,88	22,89	-	-	-
+CGU [73]	36,30	18,00	33,80	-	-	-	-
GLEAM [63]	36,51	16,80	33,92	-	29,51	9,78	25,60
Beam+BPNorm [161]	39,19	20,38	36,69	-	-	-	-
Prophnet [162]	39,55	20,27	36,57	-	-	-	-
SAGCopy-Indegree-1 [68]	38,84	20,39	36,27	-	-	-	-

5.5 Προσεγγίσεις σύγκρισης

Οι βασικές προσεγγίσεις που χρησιμοποιούνται για σκοπούς σύγκρισης αποτελούνται από τα προτεινόμενα μοντέλα βαθιάς μάθησης (Ενότητα 4.6.2), με τις βέλτιστες τιμές των υπερ-παραμέτρων τους, αλλά χωρίς την εφαρμογή κάποιας στρατηγικής γενίκευσης στα δεδομένα εκπαίδευσης και ελέγχου (και, συνεπώς, χωρίς την εκτέλεση της φάσης μετα-επεξεργασίας).

Επιπροσθέτως, οι Πίνακες 5.3 και 5.4 συνοψίζουν την απόδοση άλλων, σύγχρονων προσεγγίσεων στα ίδια σύνολα δεδομένων, όπως αναφέρονται στις αντίστοιχες εργασίες (οι παύλες στον πίνακα 5.3 σημαίνουν ότι οι συγκεκριμένες εργασίες δε χρησιμοποίησαν στα πειράματά τους το αναφερόμενο σύνολο δεδομένων ή την αναφερόμενη μετρική αξιολόγησης). Έχει γίνει σχετική αναφορά στα μοντέλα αυτά, που χρησιμοποιούν προσεγγίσεις μηχανικής μάθησης για την αυτόματη *ΠΚ*, στην Ενότητα 3.2. Με δεδομένο ότι οι συγγραφείς των αναφερόμενων προσεγγίσεων έχουν χρησιμοποιήσει τα ίδια σύνολα δεδομένων, με την ίδια μεθοδολογία επιλογής των παραδειγμάτων εκπαίδευση, επικύρωσης και ελέγχου, τα αποτελέσματα επίδοσης των προσεγγίσεων αυτών είναι άμεσα συγκρίσιμα με εκείνα της προτεινόμενης προσέγγισης. Να διευκρινιστεί, ότι έχει χρησιμοποιηθεί τυχαία δειγματοληψία για την επιλογή των παραδειγμάτων ελέγχου και επικύρωσης των συνόλων δεδομένων *Gigaword* και *CNN/DailyMail*, ενώ το σύνολο δεδομένων *DUC 2004* χρησιμοποιείται μόνο για σκοπούς αξιολόγησης.

Μια πρόσθετη μέτρηση που αναφέρεται για ορισμένα μοντέλα του συνόλου δεδομένων *Gigaword* (πέμπτη στήλη του πίνακα 5.3) είναι αυτή του *ποσοστού νέων λέξεων* (*NTR*) που παρατηρούνται στην περίληψη και δεν υπάρχουν στο αρχικό κείμενο.

Πίνακας 5.4: Η επίδοση σε όρους *Rouge* για το σύνολο δεδομένων *CNN/DailyMail* άλλων σύγχρονων και σχετικών προσεγγίσεων.

Προσέγγιση	$Rouge_1$	$Rouge_2$	$Rouge_L$
words-lvt2k-temp-att [60]	35,46	13,30	32,65
ML+intra-attention [78]	38,30	14,81	35,49
RL+intra-attention [78]	41,16	15,75	39,08
pointer-generator+coverage [61]	39,53	17,28	36,38
rnn-ext+abs+RL+rerank [163]	40,88	17,80	38,54
Bottom-Up Summarization [164]	41,22	18,68	38,34
ROUGESal+Ent [80]	40,43	18,00	37,10
SENECA [165]	41,52	18,36	38,09
BertSumAbs [66]	41,72	19,39	38,76
ETADS [67]	41,75	19,01	38,89
Two-Stage+RL [74]	41,71	19,49	38,79
ProphetNet [162]	43,68	20,64	40,72
SAGCopy-Outdegree [68]	42,53	19,92	39,44

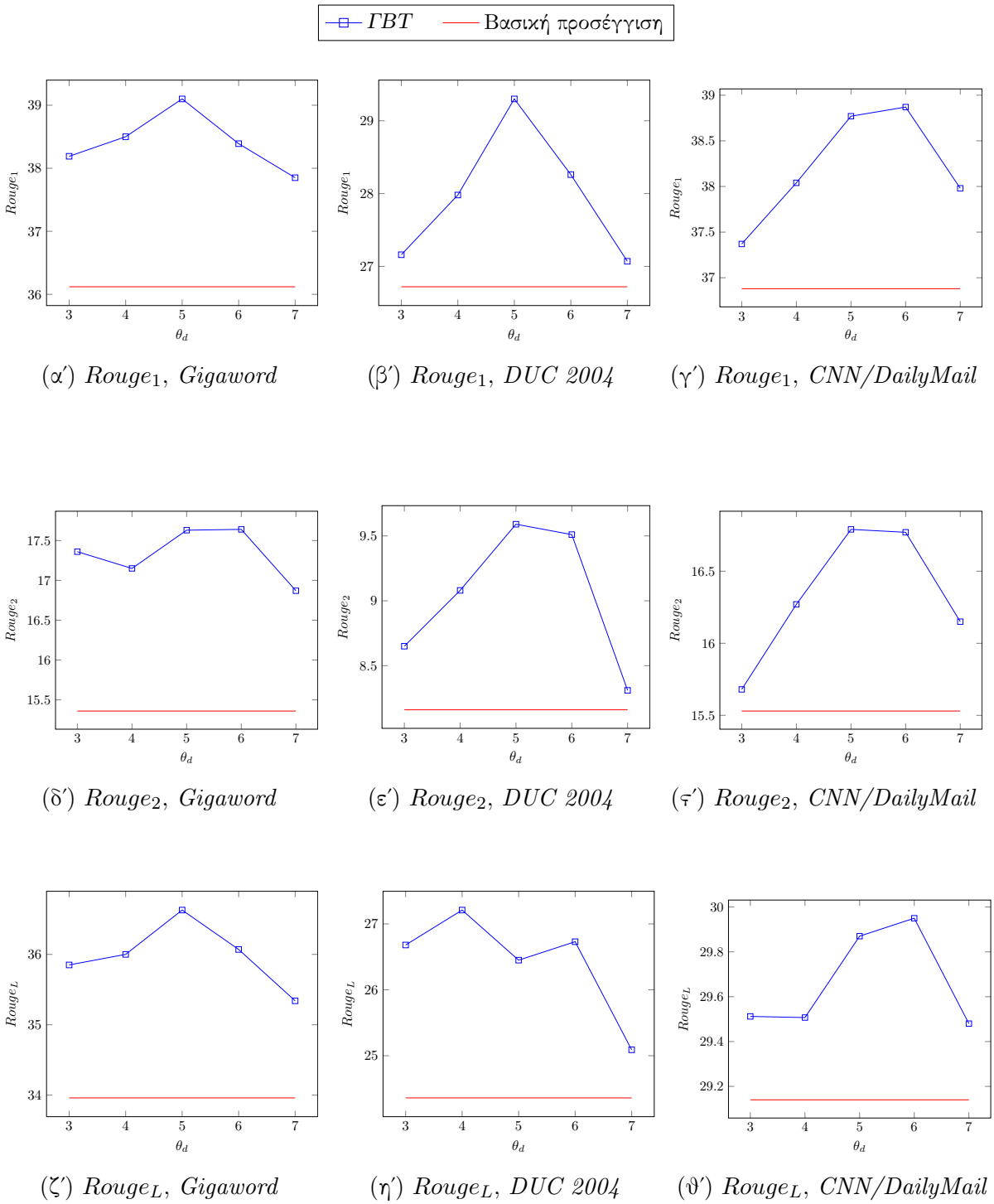
5.6 Πειραματικά αποτελέσματα

5.6.1 Επίπεδο γενίκευσης

Αρχικά, μελετάται η επίδραση που έχει η επιλογή του επιπέδου γενίκευσης στις επιδόσεις του συστήματος αυτόματης *ΠΚ*. Το Σχήμα 5.2 απεικονίζει τις τιμές των μετρικών $Rouge_1$, $Rouge_2$ και $Rouge_L$ για τα σύνολα δεδομένων *Gigaword* (Εικόνες 5.2α′, 5.2δ′, 5.2ζ′), *DUC 2004* (Εικόνες 5.2β′, 5.2ε′, 5.2η′) και *CNN/DailyMail* (Εικόνες 5.2γ′, 5.2ϛ′, 5.2θ′). Οι μετρήσεις έγιναν για διάφορες τιμές του ορίου θ_d, που προσδιορίζει το ελάχιστο βάθος ταξινομίας για την γενίκευση κειμένου με την στρατηγική *ΓΒΤ*. Για το όριο αυτό, εξετάζονται οι τιμές $\theta_d = \{3,\ 4,\ 5,\ 6,\ 7\}$. Το όριο της συχνότητάς των εννοιών που θεωρούνται υποψήφιες για γενίκευση διατηρείται σταθερό στην τιμή $\theta_f = 100$ (δηλ., όροι που έχουν συχνότητα μικρότερη ή ίση με την τιμή αυτή είναι υποψήφιοι για γενίκευση). Σύμφωνα με τα αποτελέσματα των μετρήσεων, η επίδοση της προσέγγισης θεωρούμε ότι μεγιστοποιείται για $\theta_d = 5$, καθώς σε αυτή την τιμή έχουμε τις υψηλότερες επιδόσεις σε όρους μετρικών *Rouge* για τις περισσότερες περιπτώσεις των μετρήσεων. Αυτό επιτρέπει να διατηρηθεί σταθερή η τιμή της παραμέτρου θ_d, ίση με 5 κατά την διεξαγωγή των υπόλοιπων πειραμάτων, καθώς η τιμή αυτή θεωρείται η βέλτιστη επιλογή για τον καθορισμό του επιπέδου γενίκευσης.

5.6.2 Όριο ελάχιστης συχνότητας όρων

Ο Πίνακας 5.5 παρουσιάζει τις τιμές επίδοσης των μετρικών *Rouge* για τα τρία σύνολα δεδομένων των στρατηγικών *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* και για διαφορετικές τιμές του ορίου ελάχιστης συχνότητας ($\theta_f = \{100,\ 200,\ 500,\ 1000\}$), το οποίο καθορίζει τις υποψήφιες για



(α′) $Rouge_1$, *Gigaword* (β′) $Rouge_1$, *DUC 2004* (γ′) $Rouge_1$, *CNN/DailyMail*

(δ′) $Rouge_2$, *Gigaword* (ε′) $Rouge_2$, *DUC 2004* (ϛ′) $Rouge_2$, *CNN/DailyMail*

(ζ′) $Rouge_L$, *Gigaword* (η′) $Rouge_L$, *DUC 2004* (ϑ′) $Rouge_L$, *CNN/DailyMail*

Σχήμα 5.2: Η επίδοση σε όρους μετρικών *Rouge* με χρήση των τριών συνόλων δεδομένων για διάφορα επίπεδα γενίκευσης (θ_d) της στρατηγικής *ΓΒΤ* ($\theta_f = 100$).

γενίκευση έννοιες. Τα πειράματα της ενότητας αυτής έχουν γίνει με χρήση του μοντέλου κωδικοποιητή αποκωδικοποιητή με μηχανισμό προσοχής. Η τελευταία σειρά του Πίνακα 5.5 αναφέρει την επίδοση της βασικής προσέγγισης η οποία χρησιμοποιείται για σκοπούς σύγκρισης (Ενότητα 5.5). Σε αυτή την κατηγορία πειραμάτων, εφαρμόζεται γενίκευση σε ουσιαστικά, με το ελάχιστο βάθος ταξινομίας (για γενίκευση) να περιορίζεται σε βάθος ίσο με 5, σύμφωνα με όσα περιγράφονται παραπάνω. Επιπροσθέτως, το ποσοστό νέων λέξεων (NTR) αναφέρεται για το σύνολο δεδομένων *Gigaword*, καθώς μόνο για αυτό το σύνολο δεδομένων έχουμε τέτοιες μετρήσεις από τις άλλες προσεγγίσεις. Από τα αποτελέσματα παρατηρούμε ότι όλα τα εξεταζόμενα μοντέλα της προτεινόμενης προσέγγισης ξεπερνούν σε επίδοση τη βασική προσέγγιση και, επίσης, τα περισσότερα από αυτά επιτυγχάνουν υψηλότερες τιμές σε όρους μετρικών *Rouge* από άλλες σύγχρονες προσεγγίσεις, η επίδοση των οποίων αναφέρεται, για σκοπούς σύγκρισης, στους Πίνακες 5.3 και 5.4.

Για να ελεγχθεί αν τα αποτελέσματα των τιμών *Rouge* είναι στατιστικά σημαντικά χρησιμοποιείται η *δοκιμή-t του Welch (Welch's t-test)* [166]. Για τον σκοπό αυτό, χρησιμοποιούνται οι τιμές των μετρικών *Rouge* για κάθε παράδειγμα του συνόλου ελέγχου (και για κάθε ένα από τα τρία σύνολα δεδομένων). Η δοκιμή-t πραγματοποιείται για κάθε ζεύγος μοντέλων (δηλ., εφαρμόζεται σε όλα τα μοντέλα ανά δύο) και για κάθε μετρική *Rouge*. Οι μέγιστες τιμές για το μέγεθος *p-value* που λάβαμε από τις μετρήσεις αναφέρονται στις λεζάντες των Πίνακων 5.5-5.11 και οι τιμές αυτές αποδεικνύουν ότι οι μετρήσεις με όρους μετρικών *Rouge*, είναι στατιστικά σημαντικές για όλες τις περιπτώσεις που εξετάζονται.

Πίνακας 5.5: Οι τιμές επίδοσης *Rouge* και NTR για: (i) τις στρατηγικές *ΓΒΤ*, *ΓΟΟ*, *ΓΟΟ-ΓΒΤ*, (ii) μεταβάλλοντας το όριο θ_f για σταθερή τιμή $\theta_d = 5$ και (iii) γενίκευση μόνο ουσιαστικών (δοκιμή-t: $p_{value} < 0.012$ για $Rouge_1$ και $p_{value} < 0.02$ για $Rouge_2$ και $Rouge_L$).

		Gigaword				*DUC 2004*			*CNN/DailyMail*		
Μοντέλο	θ_f	R_1	R_2	R_L	NTR (%)	R_1	R_2	R_L	R_1	R_2	R_L
ΓΒΤ-σ100-β5-ο	100	39,09	17,63	36,63	24,97	29,30	9,59	26,68	38,77	16,79	29,87
ΓΒΤ-σ200-β5-ο	200	38,23	17,29	35,94	26,36	29,02	10,06	26,75	38,36	16,24	29,62
ΓΒΤ-σ500-β5-ο	500	37,99	17,37	35,80	24,39	28,65	9,55	26,45	38,02	16,34	29,80
ΓΒΤ-σ1000-β5-ο	1000	37,60	16,52	35,21	26,51	27,94	8,78	25,09	37,34	15,82	29,26
ΓΟΟ-σ100	100	37,64	16,80	35,26	24,50	27,73	8,86	25,24	37,38	15,77	28,84
ΓΟΟ-σ200	200	38,31	17,14	35,94	24,94	28,41	8,63	25,62	37,64	15,76	28,85
ΓΟΟ-σ500	500	38,56	17,56	36,15	24,97	29,27	9,87	26,61	37,89	16,00	29,09
ΓΟΟ-σ1000	1000	37,74	17,04	35,48	24,69	27,90	8,82	24,95	37,91	15,86	28,89
ΓΟΟ-ΓΒΤ-σ100-β5-ο	100	37,24	16,39	34,95	25,72	29,21	9,94	26,66	37,31	15,19	28,90
ΓΟΟ-ΓΒΤ-σ200-β5-ο	200	37,33	16,26	34,92	26,75	29,38	9,27	26,46	37,51	14,91	28,84
ΓΟΟ-ΓΒΤ-σ500-β5-ο	500	38,36	16,80	35,77	24,14	28,77	9,26	25,83	36,89	14,74	28,43
ΓΟΟ-ΓΒΤ-σ1000-β5-ο	1000	37,60	16,73	35,26	24,49	28,41	9,58	25,57	36,33	14,08	27,98
Βασική προσέγγιση	-	36,12	15,36	33,96	28,09	26,72	8,16	24,36	36,88	15,53	29,14

Ο Πίνακας 5.6 παρουσιάζει τα πειραματικά αποτελέσματα για την περίπτωση συμμετοχής στη διαδικασία γενίκευσης και των δύο μερών του λόγου, ουσιαστικά και ρήματα, για τις περιπτώσεις *ΓΒΤ* και *ΓΟΟ-ΓΒΤ*. Το σχήμα δεδομένων *ΓΟΟ* δε λαμβάνεται υπόψη σε αυτές τις μετρήσεις, επειδή το σχήμα αυτό γενικεύει μόνο τα ουσιαστικά και δεν έχει τη δυνατότητα γενίκευσης ουσιαστικών και ρημάτων. Παρ' όλα αυτά, παρουσιάζεται η απόδοση του μικτού μοντέλου

ΓΟΟ-ΓΒΤ, καθώς περιλαμβάνει το σχήμα *ΓΒΤ* που μπορεί να εφαρμόσει γενίκευση και των δύο μερών του λόγου. Οι παράμετροι θ_d και θ_f λαμβάνουν τιμές αντίστοιχες με τα προηγούμενα πειράματα (Πίνακας 5.5). Όπως παρατηρούμε από τα αποτελέσματα, η προσθήκη των ρημάτων ως υποψήφιων εννοιών για γενίκευση οδηγεί σε περαιτέρω βελτίωση, σε όρους μετρικών *Rouge*.

Πίνακας 5.6: Οι τιμές επίδοσης *Rouge* και *NTR* για: (i) τις στρατηγικές *ΓΒΤ* και *ΓΟΟ-ΓΒΤ*, (ii) μεταβάλλοντας το όριο θ_f με σταθερή τιμή $\theta_d = 5$ και (iii) γενίκευση ουσιαστικών και ρημάτων (δοκιμή-t: $p_{value} < 0.01$).

		Gigaword				*DUC 2004*			*CNN/DailyMail*		
Μοντέλο	θ_f	R_1	R_2	R_L	NTR (%)	R_1	R_2	R_L	R_1	R_2	R_L
ΓΒΤ-σ100-β5-ο-ρ	100	39,18	17,41	36,62	24,05	29,17	8,87	26,34	38,66	16,74	30,00
ΓΒΤ-σ200-β5-ο-ρ	200	39,32	17,71	36,73	25,02	29,07	9,41	26,61	38,56	16,61	29,90
ΓΒΤ-σ500-β5-ο-ρ	500	38,81	17,32	36,31	24,70	28,46	8,20	26,71	37,91	16,05	29,51
ΓΒΤ-σ1000-β5-ο-ρ	1000	37,91	17,59	35,69	25,38	28,46	8,20	25,79	37,89	16,05	29,42
ΓΟΟ-ΓΒΤ-σ100-β5-ο-ρ	100	37,56	17,39	35,41	27,86	29,47	9,32	26,41	37,46	15,17	28,59
ΓΟΟ-ΓΒΤ-σ200-β5-ο-ρ	200	38,49	16,88	35,89	24,30	29,51	9,85	26,80	36,85	14,52	28,36
ΓΟΟ-ΓΒΤ-σ500-β5-ο-ρ	500	38,24	16,99	35,74	24,64	28,74	9,36	26,16	36,47	14,49	28,21
ΓΟΟ-ΓΒΤ-σ1000-β5-ο-ρ	1000	37,43	16,33	34,97	26,14	28,67	9,5	25,94	35,99	13,97	27,85

5.6.3 Όριο συχνότητας όρων πλήρως αποσαφηνισμένου κειμένου

Ο Πίνακας 5.7 συνοψίζει παρόμοιες μετρήσεις, με αυτές που παρουσιάστηκαν στον προηγούμενο πίνακα, για τη γενίκευση ουσιαστικών πλήρως αποσαφηνισμένου κειμένου (*ΠΑΚ*) για διάφορες τιμές του ορίου θ_f. Πιο συγκεκριμένα, οι στρατηγικές που ελέγχονται είναι οι εξής: *ΓΒΤ-ΠΑΚ*, *ΓΟΟ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*. Οι στρατηγικές αυτές εφαρμόζουν τη μεθοδολογία των σχημάτων *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* σε μια έκδοση των συνόλων δεδομένων που περιλαμβάνουν αναγνωριστικά αποσαφήνισης έννοιας λέξεων (*ΑΕΛ*) για όλα τα ουσιαστικά τους. Και εδώ, η τελευταία σειρά του Πίνακα 5.7 αναφέρει την απόδοση του βασικού μοντέλου *ΠΑΚ-ο*, το οποίο αφορά κείμενο μορφής *ΠΑΚ*, χωρίς την εφαρμογή κάποιας διαδικασίας γενίκευσης. Δηλαδή, στην περίπτωση του *ΠΑΚ-ο* (εφαρμογή σε πλήρως αποσαφηνισμένο κείμενο με αναγνωριστικά *ΑΕΛ* των ουσιαστικών, χωρίς γενίκευση περιεχομένου), το μοντέλο μηχανικής μάθησης εκπαιδεύεται με χρήση των συνόλων δεδομένων τα οποία περιλαμβάνουν αναγνωριστικά *ΑΕΛ* για όλα τα ουσιαστικά. Κατά τη διαδικασία της αξιολόγησης, το αρχικό κείμενο το οποίο περιλαμβάνει, επίσης, αναγνωριστικά *ΑΕΛ* για τα ουσιαστικά δίνεται ως είσοδο στο δίκτυο το οποίο επιστρέφει την εκτιμώμενη περίληψή σε μορφή *ΠΑΚ*. Παρόλο που στην περίπτωση του σχήματος *ΠΑΚ-ο* δεν εφαρμόζεται στρατηγική γενίκευσης, τα αναγνωριστικά *ΑΕΛ* τα οποία περιλαμβάνει η εκτιμώμενη περίληψης, χρειάζεται να μετατραπούν σε κατάλληλες λέξεις. Αυτό επιτυγχάνεται μέσω εφαρμογής της φάσης μετα-επεξεργασίας τύπου *ΓΒΤ* (Ενότητα 4.7), για την αντιστοίχιση των εννοιών της περίληψης με εκείνες του αρχικού κειμένου, με σκοπό να αντικατασταθούν τα αναγνωριστικά *ΑΕΛ* της εκτιμώμενης περίληψης με τις κατάλληλες λέξεις.

Γενικά, οι προσεγγίσεις που βασίζονται σε *ΠΑΚ* επιτυγχάνουν μειωμένες τιμές σε όρους *Rouge* και *NTR*, σε σύγκριση με αυτές που δεν εφαρμόζονται σε *ΠΑΚ* (Πίνακας 5.5). Το μόνο μοντέλο που ξεπερνά σε επίδοση την βασική προσέγγιση είναι το *ΓΒΤ-ΠΑΚ* και δείχνει ότι η προτεινόμενη

μεθοδολογία (για την περίπτωση *ΓΒΤ-ΠΑΚ*) ενισχύει την απόδοση των μοντέλων μηχανικής μάθησης ακόμη και στην περίπτωση *ΠΑΚ*. Αντίθετα, τα μοντέλα *ΓΟΟ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ* παρουσιάζουν μειωμένη απόδοση σε σύγκριση με τη βασική προσέγγιση.

Πίνακας 5.7: Οι τιμές επίδοσης *Rouge* και NTR για: (i) τις στρατηγικές *ΓΒΤ-ΠΑΚ*, *ΓΟΟ-ΠΑΚ*, *ΓΟΟ-ΓΒΤ-ΠΑΚ*, (ii) μεταβάλλοντας το όριο θ_f με σταθερή τιμή $\theta_d = 5$ και (iii) γενίκευση μόνο ουσιαστικών (δοκιμή-t: $p_{value} < 0.01$).

		Gigaword				*DUC 2004*			*CNN/DailyMail*		
Μοντέλο	θ_f	R_1	R_2	R_L	NTR (%)	R_1	R_2	R_L	R_1	R_2	R_L
ΓΒΤ-ΠΑΚ-σ100-β5-ο	100	37,00	15,21	34,30	18,77	28,35	8,66	25,78	37,69	15,69	29,15
ΓΒΤ-ΠΑΚ-σ200-β5-ο	200	36,27	15,73	33,69	18,91	28,79	8,56	26,36	37,40	15,51	29,05
ΓΒΤ-ΠΑΚ-σ500-β5-ο	500	36,09	15,45	33,90	17,67	27,99	9,42	25,59	37,23	15,27	29,04
ΓΒΤ-ΠΑΚ-σ1000-β5-ο	1000	35,53	14,23	32,97	18,71	27,28	8,12	24,52	37,19	15,02	28,75
ΓΟΟ-ΠΑΚ-σ100	100	35,43	14,34	33,00	19,27	27,18	8,34	24,31	34,50	11,52	26,45
ΓΟΟ-ΠΑΚ-σ200	200	34,71	13,89	32,19	19,38	27,75	8,12	24,84	34,39	11,34	26,38
ΓΟΟ-ΠΑΚ-σ500	500	34,33	14,00	32,17	19,58	26,77	8,15	24,57	32,17	9,30	20,93
ΓΟΟ-ΠΑΚ-σ1000	1000	33,55	13,89	31,46	19,90	26,11	7,68	23,57	32,90	9,74	20,88
ΓΟΟ-ΠΑΚ-ΓΒΤ-σ100-β5-ο	100	35,10	14,21	32,63	19,89	27,89	8,15	25,50	34,37	11,40	26,47
ΓΟΟ-ΠΑΚ-ΓΒΤ-σ200-β5-ο	200	35,14	14,04	32,71	19,22	27,88	8,23	25,43	34,28	11,18	26,50
ΓΟΟ-ΠΑΚ-ΓΒΤ-σ500-β5-ο	500	34,54	13,63	32,37	19,65	27,15	8,51	24,40	33,75	11,03	26,43
ΓΟΟ-ΠΑΚ-ΓΒΤ-σ1000-β5-ο	1000	33,74	13,51	31,66	20,17	26,65	7,85	24,27	33,56	10,79	25,90
ΠΑΚ-ο (Βασική προσέγγιση)	-	35,69	14,71	33,39	18,47	27,63	9,15	25,21	36,40	14,67	28,29

Τέλος, ο πίνακας 5.8 αναφέρει την απόδοση των σχημάτων γενίκευσης *ΓΒΤ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ* για γενίκευση ουσιαστικών και ρημάτων. Σχετικά με τις υπερ-παραμέτρους, στην προ-επεξεργασία και στη μετα-επεξεργασία, γίνονται κι εδώ παρόμοιες υποθέσεις με τις προηγούμενες περιπτώσεις. Το μοντέλο *ΠΑΚ-ο-ρ* (το οποίο κάνει αποσαφήνιση εννοιών για τα ουσιαστικά και ρήματα) χρησιμοποιείται ως η βασική προσέγγιση για λόγους σύγκρισης, με παρόμοιο τρόπο που χρησιμοποιήθηκε το μοντέλο *ΠΑΚ-ο* (το οποίο κάνει αποσαφήνιση εννοιών μόνο για τα ουσιαστικά) που αναφέρθηκε παραπάνω (Πίνακας 5.7). Παρόλο που τα περισσότερα μοντέλα *ΓΒΤ-ΠΑΚ* ξεπερνούν τη βασική προσέγγιση (*ΠΑΚ-ο-ρ*), συνολικά, παρουσιάζουν μειωμένη απόδοση, σε σύγκριση με τα μοντέλα που η επίδοσή τους παρουσιάστηκε παραπάνω (*ΓΒΤ* ή *ΓΟΟ-ΓΒΤ*, Πίνακας 5.5) ή τις σύγχρονες προσεγγίσεις (Πίνακες 5.3 και 5.4).

5.6.4 Επίδοση μηχανισμού αποσαφήνισης έννοιας των λέξεων

Στην παρούσα προσέγγιση χρησιμοποιείται μηχανισμός *ΑΕΛ* για να προσδιοριστούν οι εκάστοτε έννοιες των λέξεων του κειμένου με σκοπό την επίτευξη των σημασιολογικών μετασχηματισμών (κατά τη φάση της προ-επεξεργασίας και της μετα-επεξεργασίας) σύμφωνα με τις έννοιες των λέξεων. Για τον έλεγχο της αποτελεσματικότητας της προσέγγισης *ΑΕΛ*, που αξιοποιείται στο προτεινόμενο πλαίσιο, συγκρίνουμε τις επιδόσεις της με μια βασική μορφή *ΑΕΛ* που επιστρέφει την πιο συχνή έννοια της κάθε λέξης σύμφωνα με το *WordNet*. Τα πειράματα γίνονται με χρήση του μοντέλου μηχανικής μάθησης *ΚΑΜΠ* και τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*. Εφαρμόζεται στρατηγική γενίκευσης *ΓΒΤ* για γενίκευση ουσιαστικών, μεταβάλλοντας το όριο της συχνότητα των υποψήφιων για γενίκευση εννοιών θ_f και διατηρώντας σταθερό το ελάχιστο βάθος

Πίνακας 5.8: Οι τιμές επίδοσης *Rouge* και NTR για: (i) τις στρατηγικές *ΓΒΤ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*, (ii) μεταβάλλοντας το όριο θ_f με σταθερή τιμή $\theta_d = 5$ και (iii) γενίκευση ουσιαστικών και ρημάτων (δοκιμή-t: $p_{value} < 0.01$).

		Gigaword				*DUC 2004*			*CNN/DailyMail*		
Μοντέλο	θ_f	R_1	R_2	R_L	NTR (%)	R_1	R_2	R_L	R_1	R_2	R_L
ΓΒΤ-ΠΑΚ-σ100-β5-ο-ρ	100	34,00	12,82	31,94	22,73	26,87	7,14	24,39	34,97	13,13	27,08
ΓΒΤ-ΠΑΚ-σ200-β5-ο-ρ	200	33,43	12,18	31,30	22,52	26,82	7,80	24,41	34,07	12,66	26,55
ΓΒΤ-ΠΑΚ-σ500-β5-ο-ρ	500	33,47	11,93	31,29	23,20	26,38	7,29	24,08	34,06	12,50	26,54
ΓΒΤ-ΠΑΚ-σ1000-β5-ο-ρ	1000	33,06	12,06	30,90	23,67	26,29	7,51	23,78	33,83	12,60	26,43
ΓΟΟ-ΓΒΤ-ΠΑΚ-σ100-β5-ο-ρ	100	31,73	11,90	29,71	23,26	25,22	7,37	22,87	31,46	9,36	24,60
ΓΟΟ-ΓΒΤ-ΠΑΚ-σ200-β5-ο-ρ	200	32,21	11,50	30,14	23,43	25,44	7,65	23,40	31,44	9,20	24,59
ΓΟΟ-ΓΒΤ-ΠΑΚ-σ500-β5-ο-ρ	500	32,59	11,66	30,41	23,60	25,66	6,64	23,45	30,77	8,88	23,94
ΓΟΟ-ΓΒΤ-ΠΑΚ-σ1000-β5-ο-ρ	1000	31,52	11,24	29,55	23,98	24,86	6,44	22,55	30,68	8,66	23,90
ΠΑΚ-ο-ρ (βασική προσέγγιση)	-	33,18	12,28	31,02	22,84	25,75	6,93	23,48	33,41	12,01	26,24

γενίκευση $\theta_d = 5$, όπως παραπάνω. Ο Πίνακας 5.9 αναφέρει την επίδοση σε όρους μετρικών *Rouge* για την περίπτωση της εφαρμογής της βασικής μεθόδου *ΑΕΛ* (της πιο συχνής έννοιας) και δίπλα σε κάθε τιμή μετρικής *Rouge* (R_1, R_2 και R_L) αναγράφεται η διαφορά από τις αντίστοιχες τιμές μέτρησης σε όρους *Rouge* (ΔR_1, ΔR_2 και ΔR_L), που έχουν ληφθεί με την εφαρμογή της χρησιμοποιούμενης στο παρόν πλαίσιο *ΑΕΛ* (Πίνακας 5.5). Παρατηρούμε ότι η *ΑΕΛ* πιο συχνής έννοιας υπολείπεται σε επιδόσεις σε σχέση με την χρησιμοποιούμενη *ΑΕΛ*.

Πίνακας 5.9: Οι τιμές επίδοσης σε όρους μετρικών *Rouge* (R_1, R_2 και R_L), για την εφαρμογή *ΑΕΛ* (της πιο συχνής έννοιας) και η διαφορά των τιμών *Rouge* (ΔR_1, ΔR_2 και ΔR_L) σε σύγκριση με την εφαρμογή της χρησιμοποιούμενης μεθόδου *ΑΕΛ* στα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail* για: (i) τη στρατηγική γενίκευσης *ΓΒΤ* (ii) μεταβάλλοντας το όριο θ_f με σταθερή τιμή $\theta_d = 5$ και (iii) γενίκευση ουσιαστικών (δοκιμή-t: $p_{value} < 0.016$).

	Gigaword						*CNN/DailyMail*					
Μοντέλο	R_1	ΔR_1	R_2	ΔR_2	R_L	ΔR_L	R_1	ΔR_1	R_2	ΔR_2	R_L	ΔR_L
ΓΒΤ-σ100-β5-ο	35,22	−3,87	16,04	−1,59	33,15	−3,48	34,85	−3,92	15,25	−1,54	27,00	−2,87
ΓΒΤ-σ200-β5-ο	34,48	−3,75	15,75	−1,54	32,56	−3,38	34,56	−3,80	14,75	−1,46	26,81	−2,81
ΓΒΤ-σ500-β5-ο	34,34	−3,65	15,84	−1,53	32,51	−3,29	34,29	−3,73	14,90	−1,44	27,07	−2,73
ΓΒΤ-σ1000-β5-ο	34,18	−3,42	15,07	−1,45	32,04	−3,17	33,87	−3,47	14,44	−1,38	26,67	−2,57

5.6.5 Επίδοση άπληστης προσέγγισης αντιστοίχισης εννοιών

Τα αποτελέσματα που αναφέρονται παραπάνω έχουν ληφθεί χρησιμοποιώντας τη μεθοδολογία βέλτιστης αντιστοίχισης εννοιών. Ωστόσο, τα πειράματα έχουν επαναληφθεί και με τη χρήση της άπληστης προσέγγισης αντιστοίχισης εννοιών, με τα αποτελέσματα που προέκυψαν να μην έχουν ουσιαστικές διαφορές σε σύγκριση με τη βέλτιστη λύση. Πιο συγκεκριμένα, στα σύνολα δεδομένων *Gigaword* και *DUC 2004* (που έχουμε *ΠΚ* μικρών σε έκταση εγγράφων) οι διαφορές εντοπίζονται στο τρίτο δεκαδικό ψηφίο των μετρικών *Rouge* (η τιμή αφορά τη μέση τιμή των μετρήσεων

στο σύνολο των παραδειγμάτων ελέγχου). Αντίστοιχα, στο σύνολο δεδομένων *CNN/DailyMail* (*ΠΚ* σε επίπεδο εγγράφου), οι διαφορές είναι στο δεύτερο δεκαδικό ψηφίο. Κατά συνέπεια, συμπεραίνουμε ότι η βέλτιστη και η άπληστη αντιστοίχιση εννοιών παράγουν ουσιαστικά τις ίδιες περιλήψεις και για αυτόν τον λόγο θα ήταν πλεονασμός να αναφέρουμε ξεχωριστά αποτελέσματα για αυτές τις δύο μεθοδολογίες. Στην περίπτωση του συνόλου δεδομένων *CNN/DailyMail* (στο οποίο εντοπίζονται οι μεγαλύτερες διαφορές, καθώς πρόκειται για περιλήψεις μεγαλύτερες σε έκταση), για να παρουσιάσουμε σε πιο βαθμό διαφέρουν οι δύο μεθοδολογίες, στον Πίνακα 5.10 αναφέρονται οι τιμές *Rouge* των στρατηγικών *ΓΒΤ* και *ΓΟΟ* για την άπληστη αντιστοίχιση εννοιών. Οι τιμές αυτές συγκρίνονται με τις αντίστοιχες της βέλτιστης αντιστοίχισης εννοιών (που αναφέρθηκαν στον Πίνακα 5.5) και δίπλα σε κάθε τιμή μετρικής *Rouge* (R_1, R_2 και R_L) του Πίνακα 5.10, αναγράφεται η διαφορά από την αντίστοιχη τιμή της βέλτιστης λύσης (ΔR_1, ΔR_2 και ΔR_L), η οποία στις περισσότερες περιπτώσεις είναι αμελητέα.

Πίνακας 5.10: Οι τιμές επίδοσης σε όρους μετρικών *Rouge* (R_1, R_2 και R_L), για την άπληστη προσέγγιση αντιστοίχισης εννοιών και η διαφορά των τιμων *Rouge* (ΔR_1, ΔR_2 και ΔR_L) από τη βέλτιστη προσέγγιση στο σύνολο δεδομένων *CNN/DailyMail* για: (i) τις στρατηγικές γενίκευσης *ΓΒΤ* και *ΓΟΟ* (ii) μεταβάλλοντας το όριο θ_f με σταθερή τιμή $\theta_d = 5$ και (iii) γενίκευση ουσιαστικών (δοκιμή-t: $p_{value} < 0.01$).

Μοντέλο	R_1	ΔR_1	R_2	ΔR_2	R_L	ΔR_L
ΓΒΤ-σ100-β5-ο	38,77	0,00	16,78	−0,01	29,87	0,00
ΓΒΤ-σ200-β5-ο	38,33	−0,03	16,24	−0,00	29,60	−0,02
ΓΒΤ-σ500-β5-ο	37,93	−0,09	16,16	−0,18	29,69	−0,11
ΓΒΤ-σ1000-β5-ο	37,34	0,00	15,68	−0,14	29,21	−0,05
ΓΟΟ-σ100	37,38	0,00	15,67	−0,10	28,78	−0,06
ΓΟΟ-σ200	37,63	−0,01	15,71	−0,05	28,77	−0,08
ΓΟΟ-σ500	37,83	−0,06	15,91	−0,09	29,03	−0,06
ΓΟΟ-σ1000	37,84	−0,07	15,84	−0,02	28,81	−0,08

5.6.6 Αποτελέσματα προηγμένων μοντέλων μηχανικής μάθησης

Ο Πίνακας 5.11 παρουσιάζει την απόδοση των μοντέλων μηχανικής μάθησης *ΑΛΕΛ*, *ΕΜ*, *ΜΣ* και *ΜΣΠΚ* σε όρους μετρικών *Rouge* για τη στρατηγική *ΓΒΤ*, μεταβάλλοντας την τιμή του ορίου ελάχιστης συχνότητας λέξεων για γενίκευση ($\theta_f = \{100, 200, 500, 1000\}$). Οι τέσσερις τελευταίες σειρές του πίνακα αναφέρουν την επίδοση των βασικών μοντέλων, όπου δεν έχει εφαρμοστεί κάποια στρατηγική γενίκευσης (Ενότητα 5.5). Όπως παρατηρούμε από τα αποτελέσματα, η προτεινόμενη προσέγγιση υπερτερεί τόσο σε σχέση με τις βασικές προσεγγίσεις όσο και σε σχέση με τα αντίστοιχα μοντέλα του δικτύου κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής (Πίνακας 5.5). Η εφαρμογή της στρατηγικής *ΓΒΤ* με χρήση των δικτύων βαθιάς μάθησης *ΑΛΕΛ*, *ΕΜ*, *ΜΣ* και *ΜΣΠΚ* επιτυγχάνει θετικά αποτελέσματα, συγκρίσιμα με τις άλλες σύγχρονες προσεγγίσεις των Πινάκων 5.3 και 5.4, για τα αντίστοιχα σύνολα δεδομένων.

Πίνακας 5.11: Οι τιμές επίδοσης σε όρους μετρικών *Rouge* για τα δίκτυα βαθιάς μάθησης αντιγραφής λέξεων εκτός λεξιλογίου (*ΑΛΕΛ*), ενισχυτικής μάθησης (*ΕΜ*), μετασχηματιστών (*ΜΣ*), μετασχηματιστών προεκπαιδευμένου κωδικοποιητη (*ΜΣΠΚ*) για: (*i*) της στρατηγική γενίκευσης *ΓΒΤ*, (*ii*) μεταβάλλοντας το όριο θ_f με σταθερή τιμή $\theta_d = 5$ και (*iii*) γενίκευση ουσιαστικών (δοκιμή-t: $p_{value} < 0.01$).

	Gigaword			*DUC 2004*			*CNN/DailyMail*		
Μοντέλο	$Rouge_1$	$Rouge_2$	$Rouge_L$	$Rouge_1$	$Rouge_2$	$Rouge_L$	$Rouge_1$	$Rouge_2$	$Rouge_L$
ΓΒΤ-σ100-β5-ο (*ΑΛΕΛ*)	40,94	19,65	38,52	29,66	10,49	27,09	43,04	20,65	30,93
ΓΒΤ-σ200-β5-ο (*ΑΛΕΛ*)	40,26	19,25	37,99	29,47	09,78	26,59	42,15	20,19	30,23
ΓΒΤ-σ500-β5-ο (*ΑΛΕΛ*)	40,23	19,42	37,90	28,83	09,53	25,94	41,89	19,30	30,23
ΓΒΤ-σ1000-β5-ο (*ΑΛΕΛ*)	39,85	19,02	37,78	28,51	10,34	25,86	41,18	18,79	29,22
ΓΒΤ-σ100-β5-ο (*ΕΜ*)	41,72	20,57	39,31	29,97	10,35	27,17	43,63	20,87	31,46
ΓΒΤ-σ200-β5-ο (*ΕΜ*)	41,57	20,37	39,18	29,84	10,36	26,45	43,33	20,60	31,45
ΓΒΤ-σ500-β5-ο (*ΕΜ*)	41,11	20,05	38,83	28,92	10,23	26,20	42,36	20,01	30,48
ΓΒΤ-σ1000-β5-ο (*ΕΜ*)	40,67	19,64	38,44	28,80	10,18	26,09	41,60	19,40	29,87
ΓΒΤ-σ100-β5-ο (*ΜΣ*)	41,35	20,09	38,89	29,87	10,45	27,11	43,33	20,78	34,16
ΓΒΤ-σ200-β5-ο (*ΜΣ*)	40,91	19,84	38,49	29,59	10,08	26,50	42,70	20,38	33,81
ΓΒΤ-σ500-β5-ο (*ΜΣ*)	40,71	19,70	38,36	28,91	10,02	26,14	42,16	19,65	33,44
ΓΒΤ-σ1000-β5-ο (*ΜΣ*)	40,28	19,43	38,10	28,59	10,27	26,01	41,38	19,42	32,57
ΓΒΤ-σ100-β5-ο (*ΜΣΠΚ*)	42,12	20,76	39,50	29,91	10,34	27,32	43,93	21,07	35,96
ΓΒΤ-σ200-β5-ο (*ΜΣΠΚ*)	41,86	20,42	39,21	29,93	10,31	26,55	43,64	20,58	34,72
ΓΒΤ-σ500-β5-ο (*ΜΣΠΚ*)	41,32	20,25	38,77	29,12	10,20	26,39	42,52	20,31	33,75
ΓΒΤ-σ1000-β5-ο (*ΜΣΠΚ*)	40,59	19,68	38,12	28,60	10,09	26,14	41,72	19,54	33,71
Βασική προσέγγιση (*ΑΛΕΛ*)	39,33	18,31	37,10	27,45	09,72	25,60	41,40	19,26	29,56
Βασική προσέγγιση (*ΕΜ*)	40,57	19,72	38,36	27,82	09,79	25,75	41,67	19,34	29,42
Βασική προσέγγιση (*ΜΣ*)	40,18	18,96	37,26	27,43	09,75	25,65	41,08	18,12	33,37
Βασική προσέγγιση (*ΜΣΠΚ*)	40,95	19,78	38,61	27,91	09,85	26,05	41,79	19,38	35,62

5.6.7 Ακρίβεια απόδοσης πληροφορίας

Σε μια προσπάθειά ποιοτικής αξιολόγησης των παραγόμενων περιλήψεων, η οποία αφορά το πεδίο της αυτόματης *ΠΚ* [148, 167, 149], διερευνήσαμε τον βαθμό που η πληροφορία του αρχικού κειμένου αποδίδεται στην τελική περίληψη μέσω της μετρικής *ΑΑΠ* (Ενότητα 5.3). Ο Πίνακας 5.12 παρουσιάζει τα αποτελέσματα των μετρήσεων για τα σχήματα γενίκευσης κειμένου *ΓΒΤ*, *ΓΟΟ*, *ΓΟΟ-ΓΒΤ*, *ΓΒΤ-ΠΑΚ*, *ΓΟΟ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*. Τα πειράματα εκτελέστηκαν με χρήση της αρχιτεκτονικής δικτύου βαθιάς μάθησης κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής. Οι αναφερόμενες τιμές αντιπροσωπεύουν τη μέση τιμή της μετρικής στο σύνολο των παραδειγμάτων ελέγχου για κάθε σύνολο δεδομένων. Επιπλέον, ο ίδιος πίνακας αναφέρει την απόδοση της βασικής προσέγγισης που χρησιμοποιείται για λόγους σύγκρισης (Ενότητα 5.5) και του μοντέλου *ΠΑΚ-ο* που, όπως έχει ήδη αναφερθεί, αποτελεί το βασικό μοντέλο σύγκρισης για τα μοντέλα *ΠΑΚ*. Επίσης, η τελευταία σειρά αυτού του πίνακα αναφέρει την *ΑΑΠ* για την περίληψη αναφοράς. Σημειώνεται, ότι για την περίπτωση του συνόλου δεδομένων *DUC 2004*, η τιμή μέτρησης αντιστοιχεί στη μέση τιμή της *ΑΑΠ* των τεσσάρων περιλήψεων αναφοράς που παρέχονται για κάθε παράδειγμα χρήσης.

Με τον ίδιο τρόπο, ο Πίνακας 5.13 αναφέρει την ακρίβεια απόδοσης πληροφορίας θεωρώντας γενίκευση ουσιαστικών και ρημάτων. Οι εγγραφές που βασίζονται στο σχήμα γενίκευσης *ΓΟΟ* απουσιάζουν από αυτόν τον Πίνακα, καθώς η στρατηγική *ΓΟΟ* γενικεύει μόνο ονοματικές

Πίνακας 5.12: Ακρίβεια απόδοσης πληροφορίας για: (i) τις στρατηγικές *ΓΒΤ*, *ΓΟΟ*, *ΓΟΟ-ΓΒΤ*, *ΓΒΤ-ΠΑΚ*, *ΓΟΟ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*, (ii) μεταβάλλοντας το όριο $\theta_f = \{100, 200, 500, 1000\}$ για σταθερή τιμή $\theta_d = 5$ και (iii) θεώρηση μόνο ουσιαστικών στα σχήματα γενίκευσης.

	Gigaword	DUC	*CNN/DM*		*Gigaword*	*DUC*	*CNN/DM*
Μοντέλο	$fact_{acc}$ %			Μοντέλο	$fact_{acc}$ %		
ΓΒΤ-σ100-β5-ο	47, 21	41, 31	69, 48	ΓΒΤ-ΠΑΚ-σ100-β5-ο	49, 82	38, 48	68, 55
ΓΒΤ-σ200-β5-ο	43, 14	36, 01	67, 76	ΓΒΤ-ΠΑΚ-σ200-β5-ο	45, 35	39, 61	64, 94
ΓΒΤ-σ500-β5-ο	45, 28	39, 40	67, 98	ΓΒΤ-ΠΑΚ-σ500-β5-ο	45, 83	45, 07	66, 95
ΓΒΤ-σ1000-β5-ο	42, 20	40, 30	62, 24	ΓΒΤ-ΠΑΚ-σ1000-β5-ο	44, 87	42, 68	62, 87
ΓΟΟ-σ100	45, 74	41, 42	65, 34	ΓΟΟ-ΠΑΚ-σ100	47, 18	36, 64	35, 55
ΓΟΟ-σ200	46, 39	37, 60	67, 12	ΓΟΟ-ΠΑΚ-σ200	43, 24	38, 31	37, 28
ΓΟΟ-σ500	43, 84	39, 16	66, 85	ΓΟΟ-ΠΑΚ-σ500	42, 94	41, 25	54, 02
ΓΟΟ-σ1000	44, 96	40, 56	70, 70	ΓΟΟ-ΠΑΚ-σ1000	43, 02	37, 05	54, 42
ΓΟΟ-ΓΒΤ-σ100-β5-ο	45, 56	37, 13	61, 74	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ100-β5-ο	45, 93	39, 70	37, 22
ΓΟΟ-ΓΒΤ-σ200-β5-ο	45, 66	40, 83	61, 73	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ200-β5-ο	45, 15	37, 03	27, 84
ΓΟΟ-ΓΒΤ-σ500-β5-ο	45, 83	37, 20	63, 80	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ500-β5-ο	45, 32	35, 79	25, 47
ΓΟΟ-ΓΒΤ-σ1000-β5-ο	46, 42	41, 83	59, 36	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ1000-β5-ο	46, 95	43, 61	34, 17
Βασική προσέγγιση	41, 98	39, 55	54, 47	ΠΑΚ-ο (Βασική προσέγγιση)	45, 70	48, 27	60, 75
Περίληψη αναφοράς	32, 71	17, 36	32, 56				

Πίνακας 5.13: Ακρίβεια απόδοσης πληροφορίας για: (i) τις στρατηγικές *ΓΒΤ*, *ΓΟΟ-ΓΒΤ*, *ΓΒΤ-ΠΑΚ*και *ΓΟΟ-ΓΒΤ-ΠΑΚ*, (ii) μεταβάλλοντας το όριο $\theta_f = \{100, 200, 500, 1000\}$ για σταθερή τιμή $\theta_d = 5$ και (iii) θεώρηση ουσιαστικών και ρημάτων στα σχήματα γενίκευσης περιεχομένου.

	Gigaword	*DUC*	*CNN/DM*		*Gigaword*	*DUC*	*CNN/DM*
Μοντέλο	$fact_{acc}$ %			Μοντέλο	$fact_{acc}$ %		
ΓΒΤ-σ100-β5-ο-ρ	46, 86	41, 92	70, 42	ΓΒΤ-ΠΑΚ-σ100-β5-ο-ρ	43, 76	32, 29	24, 81
ΓΒΤ-σ200-β5-ο-ρ	44, 85	35, 20	68, 18	ΓΒΤ-ΠΑΚ-σ200-β5-ο-ρ	42, 82	33, 00	26, 87
ΓΒΤ-σ500-β5-ο-ρ	43, 89	40, 04	65, 29	ΓΒΤ-ΠΑΚ-σ500-β5-ο-ρ	45, 82	39, 08	23, 97
ΓΒΤ-σ1000-β5-ο-ρ	43, 06	39, 26	67, 68	ΓΒΤ-ΠΑΚ-σ1000-β5-ο-ρ	46, 54	38, 16	22, 39
ΓΟΟ-ΓΒΤ-σ100-β5-ο-ρ	43, 23	37, 80	63, 72	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ100-β5-ο-ρ	42, 59	38, 60	9, 05
ΓΟΟ-ΓΒΤ-σ200-β5-ο-ρ	46, 11	38, 80	59, 03	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ200-β5-ο-ρ	45, 25	35, 85	8, 09
ΓΟΟ-ΓΒΤ-σ500-β5-ο-ρ	45, 82	39, 08	58, 14	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ500-β5-ο-ρ	43, 49	35, 64	6, 61
ΓΟΟ-ΓΒΤ-σ1000-β5-ο-ρ	45, 39	37, 98	57, 02	ΓΟΟ-ΓΒΤ-ΠΑΚ-σ1000-β5-ο-ρ	43, 04	37, 31	8, 14
				ΠΑΚ-ο-ρ (βασική προσέγγιση)	45.90	40.38	17.82

οντότητες που αντιστοιχούν σε όρους ουσιαστικών, ενώ εδώ απαιτείται τόσο γενίκευση ουσιαστικών όσο και ρημάτων. Αντίστοιχα με παραπάνω, ο Πίνακας αυτός αναφέρει την απόδοση της βασικής προσέγγισης (Ενότητα 5.5) καθώς και της προσέγγισης *ΠΑΚ-ο-ρ*, που αποτελεί τη βασική προσέγγιση για τα σχήματα *ΠΑΚ*.

5.6.8 Μελέτη περίπτωσης

Με σκοπό την περαιτέρω παρουσίαση των σταδίων που περιλαμβάνει η προτεινόμενη μεθοδολογία (δηλ., της ροής εργασίας από το αρχικό κείμενο έως την τελική περίληψη), αλλά και την ανάδειξη των κύριων πτυχών της, οι Πίνακες 5.14 και 5.15 παρουσιάζουν παραδείγματα

Πίνακας 5.14: Παραδείγματα αυτόματης *ΠΚ* σε επίπεδο μικρής έκτασης αρχικού κειμένου, τα οποία παρουσιάζουν τη ροή εργασίας από το αρχικό κείμενο έως την τελική περίληψη για τις στρατηγικές γενίκευσης περιεχομένου *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* με εφαρμογή γενίκευσης ουσιαστικών.

Κείμενο εισόδου: plump , juicy , bright red tomatoes that hang heavy on the vine ; basil that grows profusely , and so many cucumbers that you run out of ideas for using them : these are just the foods to close out the hot days of summer , and to provide a final note for this column , which will cease this week .	
ΓΒΤ	**Γενικευμένο κείμενο:** plump , juicy , bright red produce.n.01 that hang heavy on the plant.n.02; flavorer.n.01 that grows profusely , and so many produce.n.01 that you run out of ideas for using them : these are just the foods to close out the hot days of summer , and to provide a final note for this column , which will cease this week . **Εκτιμώμενη περίληψη:** produce.n.01 to close out of summer **Τελική περίληψη:** tomatoes to close out of summer
ΓΟΟ	**Γενικευμένο κείμενο:** plump , juicy , bright red tomatoes that hang heavy on the vine ; basil that grows profusely , and so many cucumbers that you run out of ideas for using them : these are just the foods to close out DATE , and to provide a final note for this column , which will cease this week . **Εκτιμώμενη περίληψη:** vegetables to close this DATE **Τελική περίληψη:** vegetables to close this summer
ΓΟΟ-ΓΒΤ	**Γενικευμένο κείμενο:** plump , juicy , bright red produce.n.01 that hang heavy on the plant.n.02 ; flavorer.n.01 that grows profusely , and so many produce.n.01 that you run out of ideas for using them : these are just the foods to close out DATE , and to provide a final note for this column , which will cease this week . **Εκτιμώμενη περίληψη:** vegetable.n.01 to close this DATE **Τελική περίληψη:** foods to close this summer
Περίληψη αναφοράς: hail to summer 's end and farewell	

χρήσης αυτόματης *ΠΚ*, σύμφωνα με τις στρατηγικές *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* για γενίκευση ουσιαστικών.

Ειδικότερα, στην περίπτωση της στρατηγικής *ΓΒΤ*, κατά τη φάση της προ-επεξεργασίας (Ενότητα 4.5), γενικεύονται ορισμένες λέξεις που έχουν ανεπαρκή αριθμό παραδειγμάτων χρήσης στο σύνολο δεδομένων, χρησιμοποιώντας την ταξινομία εννοιών του ηλεκτρονικού λεξικού *WordNet* [43, 44] και τα αναγνωριστικά *ΑΕΛ* σύμφωνα, επίσης, με το *WordNet* (π.χ., *product.n.01*). Ένα μοντέλο μηχανικής μάθησης (Ενότητα 4.6.2), το οποίο έχει εκπαιδευτεί με χρήση ενός σχετικού συνόλου εκπαίδευσης, επιστρέφει την εκτιμώμενη περίληψη. Στη συνέχεια, η φάση της μετα-επεξεργασίας (Ενότητα 4.7) διαμορφώνει την τελική περίληψη.

Στο παράδειγμα περίληψης κειμένου μικρής έκτασης που παρουσιάζεται στον Πίνακα 5.14, μπορούμε να παρατηρήσουμε ότι η εκτιμώμενη περίληψη (του μοντέλου μηχανικής μάθησης) που βασίζεται στη στρατηγική *ΓΒΤ* περιέχει τη γενικευμένη έννοια *product.n.01* (“*φρέσκα*

Πίνακας 5.15: Παραδείγματα αυτόματης *ΠΚ*, έκτασης αρχικού κειμένου σε επίπεδο εγγράφου, τα οποία παρουσιάζουν τα στάδια εργασίας της ροής του κειμένου από το αρχικό κείμενο έως την τελική περίληψη, για τις στρατηγικές γενίκευσης περιεχομένου *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ*, με γενίκευση ουσιαστικών.

	Κείμενο εισόδου: firefighters responded to cries for help - from two parrots . the crew scoured a burning home in boise , idaho , searching for people shouting ‘ help ! ’ and ‘ fire ! ’ eventually , to their surprise , they found a pair of squawking birds . scroll down for video . cry for help ! this is one of the two parrots who were found in a burning home after calling for help . the tropical creatures appeared to have been alone when flames began to sweep the property . but they seemed to know what to do . both were pulled from the home and given oxygen . they are expected to survive . the fire crew in boise , idaho , thought they were chasing human voices when the found the birds . treatment : the officials treated the birds with oxygen masks and survive . according to kboi , the cause of the officers managed to contain the fire to just one room . it is being both are expected to investigated and no people were found inside . officials have yet to track down the birds ’ owners .
ΓΒΤ	**Γενικευμένο κείμενο εισόδου:** firefighters responded to cries for help - from two copycat.n.01 . the crew scoured a burning home in boise , idaho , searching for people shouting ‘ help ! ’ and ‘ fire ! ’ eventually , to their astonishment.n.01 , they found a pair of squawking birds . scroll down for video . cry for help ! this is one of the two copycat.n.01 who were found in a burning home after calling for help . [...] yet to track down the birds ’ owners . **Εκτιμώμενη περίληψη:** the two copycat.n.01 were found in a burning home in oregon , idaho . they were pulled from the home and given oxygen . they are expected to survive . **Τελική περίληψη:** the two parrot were found in a burning home in oregon , idaho . they were pulled from the home and given oxygen . they are expected to survive .
ΓΟΟ	**Γενικευμένο κείμενο:** firefighters responded to [...] a burning home in GPE , GPE , searching for people shouting [...] according to ORG , the cause of the officers managed to contain [...] the birds ’ owners . **Εκτιμώμενη περίληψη:** the fire crew in GPE , GPE , thought they were chasing human voices . they were found in a burning home in GPE , GPE . it is being investigated and no people were found inside . **Τελική περίληψη:** the fire crew in boise , idaho , thought they were chasing human voices . they were found in a burning home in boise , idaho . it is being investigated and no people were found inside .
ΓΟΟ-ΓΒΤ	**Γενικευμένο κείμενο:** firefighters responded to cries for help - from two copycat.n.01 . the crew scoured a burning home in GPE , GPE , searching for people shouting ‘ help ! ’ and ‘ fire ! ’ eventually , to their astonishment.n.01 , they found a pair of squawking birds . scroll down for video . cry for help ! this is one of the two copycat.n.01 who were found in a burning home [...]. the fire crew in GPE , GPE , thought they were chasing human voices [...] to survive . according to ORG , the cause of the officers managed to [...] to track down the birds ’ owners . **Εκτιμώμενη περίληψη:** the crew scoured a burning home in GPE , GPE , searching for people shouting ‘ help ! ’ and ‘ fire ! ’ they found the copycat.n.01 with oxygen masks and both are expected to survive . **Τελική περίληψη:** the crew scoured a burning home in boise , idaho , searching for people shouting ‘ help ! ’ and ‘ fire ! ’ they found the parrots with oxygen masks and both are expected to survive .
	Περίληψη αναφοράς: two parrots were home alone when a fire erupted in boise , idaho . started calling ‘ help ! ’ and ‘ fire ! ’ , crew thought they were human voices . both were pulled from the wreckage and treated with oxygen masks .

φρούτα και λαχανικά που καλλιεργούνται για την αγορά”, σύμφωνα με την συγκεκριμένη σημασία του *WordNet*). Στη φάση της μετα-επεξεργασίας σχηματίζεται η τελική περίληψη, με την προαναφερόμενη γενικευμένη έννοια να αντικαθίσταται από τη λέξη *tomatoes*, η οποία προέρχεται από το αρχικό κείμενο και αποτελεί ένα υπώνυμο της έννοιας *product.n.01*. Σε αυτό το παράδειγμα

της στρατηγικής *ΓΒΤ*, όλες οι λέξεις της τελικής περίληψης περιλαμβάνονται στο αρχικό κείμενο, καθώς, επίσης, και οι φράσεις "*to close out*"και "*of summer*"προέρχονται από το αρχικό κείμενο.

Αντίστοιχα, στο παράδειγμα που αφορά *ΠΚ* σε επίπεδο εγγράφου (Πίνακας 5.15), το γενικευμένο κείμενο της στρατηγικής *ΓΒΤ* περιλαμβάνει τη γενικευμένη έννοια *copycat.n.01*, η οποία εμφανίζεται, επίσης, στην εκτιμώμενη περίληψη. Στην τελική περίληψη αυτή η έννοια αντικαθίσταται από τη λέξη *parrot*. Επίσης, να σημειωθεί, ότι η τελική περίληψη του εν λόγω παραδείγματος περιέχει όχι μόνο λέξεις του αρχικού κειμένου αλλά και νέες λέξεις, όπως τη λέξη *oregon*, η οποία δεν έχει παρουσία στο αρχικό κείμενο.

Στο παράδειγμα που αφορά την στρατηγική *ΓΟΟ* (Πίνακας 5.14), η γενικευμένη οντότητα *DATE*, που εμφανίζεται στην εκτιμώμενη περίληψη, αντικαταστάθηκε με τη λέξη *summer* (η οποία προέρχεται από το αρχικό κείμενο) στη φάση της μετα-επεξεργασίας που διαμορφώνει την τελική περίληψη. Επιπλέον, στο ίδιο παράδειγμα, η λέξη *vegetables*, που εμφανίζεται τόσο στην εκτιμώμενη όσο και στην τελική περίληψη, είναι μια νέα λέξη που δεν έχει παρουσία στο αρχικό κείμενο. Η εμφάνιση νέων λέξεων στις τελικές περιλήψεις υποδηλώνει ότι το σύστημα όχι μόνο αντιγράφει περιεχόμενο από το κείμενο εισαγωγής αλλά είναι ικανό να δημιουργήσει νέες λέξεις ή φράσεις, σύμφωνα με το περιεχόμενο του αρχικού κειμένου. Παρατηρούμε, επίσης, ότι η έννοια *veget.n.01* εμφανίζεται στην εκτιμώμενη περίληψη της στρατηγικής *ΓΟΟ-ΓΒΤ*, και, στη συνέχεια, (με εφαρμογή του σταδίου μετα-επεξεργασίας), αυτή η έννοια αντικαθίσταται από τη λέξη *food*, καθώς ο αλγόριθμος προσπαθεί να επιλέξει την πιο σχετική λέξη από το αρχικό κείμενο.

Αντίστοιχα, στο παράδειγμα της στρατηγικής *ΓΟΟ* που αφορά περίληψη σε επίπεδο εγγράφου (Πίνακας 5.15), τα αναγνωριστικά ονοματικών οντοτήτων, που εμφανίζονται στην εκτιμώμενη περίληψη, έχουν αντικατασταθεί από τις κατάλληλες λέξεις του αρχικού κείμενο, προκειμένου να δημιουργηθεί η τελική περίληψη. Τέλος, στην περίπτωση της στρατηγικής *ΓΟΟ-ΓΒΤ*, γίνεται φανερό ότι έχουμε έναν συνδυασμό των μεθόδων *ΓΟΟ* και *ΓΒΤ*, καθώς το γενικευμένο κείμενο και η εκτιμώμενη περίληψη περιλαμβάνουν τόσο αναγνωριστικά ονομάτων οντοτήτων όσο και αναγνωριστικά *ΑΕΛ* (για τις γενικευμένες έννοιες), όπως παρατηρούμε στο σχετικό παράδειγμα.

5.7 Περιγραφή και ερμηνεία αποτελεσμάτων

Η πειραματική διαδικασία στοχεύει στην εξέταση και αξιολόγηση πτυχών της προτεινόμενης μεθοδολογίας οι οποίες αναλύονται περαιτέρω σε αυτή την ενότητα. Αρχικά, συζητείται η επίδραση του επιτρεπόμενου ελάχιστου βάθους ταξινομίας εννοιών στη γενίκευση περιεχομένου (Ενότητα 5.7.1) και του κατωφλίου της συχνότητας εννοιών που προσδιορίζει τις υποψήφιες για γενίκευση έννοιες (Ενότητα 5.7.2). Στη συνέχεια, εκτιμάται η συνεισφορά του μηχανισμού αποσαφήνισης εννοιών (Ενότητα 5.7.3) και η συμμετοχή των μερών του λόγου στη γενίκευση περιεχομένου (Ενότητα 5.7.4). Η ενότητα 5.7.5 παρουσιάζει τα αποτελέσματα της μέτρησης *ποσοστού νέων λέξεων* στην τελική περίληψη, ενώ η Ενότητα 5.7.7 συζητά σχετικά με την *ακρίβεια απόδοσης πληροφορίας* της τελικής περίληψης σε σχέση με το αρχικό κείμενο. Τέλος, η Ενότητα 5.7.8 σχολιάζει την προοπτική βελτίωσης των παραγόμενων περιλήψεων μέσω της προτεινόμενης προσέγγισης, σύμφωνα με την ακρίβεια των προβλέψεων που επιτυγχάνουν τα μοντέλα μηχανικής μάθησης.

5.7.1 Η επίδραση του βάθους ταξινομίας εννοιών

Οι στρατηγικές που βασίζονται στη *ΓΒΤ* είναι σε θέση να γενικεύσουν έννοιες σε ένα σημασιολογικό εύρος, από πολύ γενικές έννοιες (μικρό βάθος ταξινομίας) έως πιο συγκεκριμένες έννοιες (μεγαλύτερο βάθος ταξινομίας). Το Σχήμα 5.2 παρουσιάζει τις μετρούμενες τιμές *Rouge* για διάφορα επίπεδα γενίκευσης ($\theta_d \in \{3,\ 4,\ 5,\ 6,\ 7\}$). Το βέλτιστο βάθος ταξινομίας προσδιορίζεται ίσο με 5, καθώς σε αυτό το βάθος βελτιστοποιείται η απόδοση των περισσότερων μοντέλων, σύμφωνα με τα αποτελέσματα των μετρήσεων. Σε περιπτώσει υπερ-γενίκευσης, όταν έχουμε $\theta_d < 5$, παρατηρούμε ότι η απόδοση μειώνεται, καθώς σε αυτά τα επίπεδα γενίκευσης προκύπτουν πολύ γενικές έννοιες. Από την άλλη πλευρά, για $\theta_d > 5$, οι τιμές *Rouge* επίσης μειώνονται. Αυτό συμβαίνει, καθώς η γενίκευση σε μεγαλύτερο βάθος ταξινομίας περιορίζει το πλήθος των εννοιών που μπορούν να γενικευτούν, με αποτέλεσμα λιγότερες έννοιες να γενικεύονται, χωρίς να δημιουργούν κάποια υπολογίσιμη επίδραση στην συνολική απόδοση.

5.7.2 Η επίδραση της συχνότητας εννοιών

Όπως έχει ήδη συζητηθεί στην Ενότητα 4.5.2, η προτεινόμενη προσέγγιση προσπαθεί να αντιμετωπίσει το πρόβλημα των λέξεων εκτός λεξιλογίου (*ΛΕΛ*) ή των σπάνιων λέξεων θεωρώντας μια ελάχιστη συχνότητα (θ_f) παρουσίας των όρων στο σύνολο εκπαίδευσης, η οποία θέτει ένα όριο συχνότητας κάτω από το οποίο οι όροι αυτοί είναι υποψήφιοι για γενίκευση. Στα πειράματα, εξετάστηκαν τέσσερα διαφορετικά όρια συχνότητας ($\theta_f \in \{100, 200, 500, 1000\}$) για κάθε στρατηγική γενίκευσης.

Σύμφωνα με τα αποτελέσματα, τα σχήματα *ΓΒΤ* και *ΓΒΤ-ΠΑΚ* επιτυγχάνουν την υψηλότερη τους επίδοση για γενίκευση εννοιών με σχετικά χαμηλή συχνότητα εμφάνισης στο σύνολο εκπαίδευσης (100 και 200, αντίστοιχα). Για παράδειγμα, στην περίπτωση της γενίκευσης ουσιαστικών των περιπτώσεων *ΓΒΤ-σ100-β5-ο* και *ΓΒΤ-ΠΑΚ-σ100-β5-ο* (Πίνακες 5.5 και 5.7) για $\theta_f = 100$ μεγιστοποιείται η τιμή της μετρικής $Rouge_1$ ή για $\theta_f = 200$ μεγιστοποιείται η τιμή της μετρικής $Rouge_L$. Το ίδιο ισχύει επίσης και για τα μοντέλα *ΓΒΤ-σ100-β5-ο-ρ* και *ΓΒΤ-ΠΑΚ-σ100-β5-ο-ρ* (Πίνακες 5.6 και 5.8) που εφαρμόζουν γενίκευση σε ουσιαστικά και ρήματα. Από τα παραπάνω συμπεραίνουμε, ότι για τα μοντέλα που βασίζονται σε *ΓΒΤ*, η γενίκευση σπάνιων λέξεων μεγιστοποιεί την απόδοσή τους, ενώ η γενίκευση συχνότερων εννοιών (π.χ. $\theta_f = 500$ ή $\theta_f = 1000$) μειώνει τις τιμές επίδοσής τους.

Στην περίπτωση της στρατηγικής *ΓΟΟ* (Πίνακας 5.5), τα μοντέλα τείνουν να παρουσιάζουν υψηλές επιδόσεις ακόμη και όταν γενικεύονται συχνές έννοιες. Συγκεκριμένα, το σχήμα *ΓΟΟ-σ500* ($\theta_f = 500$) επιτυγχάνει τις υψηλότερες τιμές σε όρους μετρικών *Rouge* και στα τρία σύνολα δεδομένων. Αυτό αποδίδεται στο γεγονός ότι οι όροι ονοματικών οντοτήτων έχουν παρόμοια λειτουργία στη γλώσσα. Επιπλέον, η φάση της μετα-επεξεργασίας επιτυγχάνει ικανοποιητικά την αντιστοίχιση των ονοματικών οντοτήτων με τις συγκεκριμένες έννοιες, ακόμη και στην περίπτωση που έχουν γενικευτεί έννοιες υψηλής συχνότητας. Εξαίρεση σε αυτό αποτελούν τα σχήματα *ΓΟΟ-ΠΑΚ* (Πίνακας 5.7) τα οποία επιτυγχάνουν την υψηλότερη επίδοση για συχνότητα $\theta_f = 100$, καθώς βασίζονται σε πλήρως αποσαφηνισμένο κείμενο (*ΠΑΚ*), το οποίο αυξάνει το μέγεθος του λεξιλογίου (η αύξηση του μεγέθους του λεξιλογίου προκαλεί την αύξηση του πλήθους των λέξεων με ανεπαρκή αριθμό εμφανίσεων στο σύνολο εκπαίδευσης). Τα σχήματα ΓΟΟ-ΓΒΤ και

ΓΟΟ-ΓΒΤ-ΠΑΚ εμφανίζουν την υψηλότερη απόδοση κυρίως για ενδιάμεσες συχνότητες ($\theta_f = 200$ ή $\theta_f = 500$), καθώς συνδυάζουν και τις δύο στρατηγικές γενίκευσης ($ΓΟΟ$ και $ΓΒΤ$). Σε αυτή την περίπτωση, παρατηρούμε μια αντιστάθμιση μεταξύ της $ΓΟΟ$ (η οποία ενισχύει την απόδοση σε περίπτωση γενίκευσης συχνών λέξεων) και της $ΓΒΤ$ (η οποία είναι πιο αποτελεσματική όταν γενικεύονται σπάνιοι όροι).

Είναι προφανές ότι η επιλογή του ορίου συχνότητας εννοιών (θ_f), οι οποίες θεωρούνται υποψήφιες για γενίκευση, επηρεάζει σε μεγάλο βαθμό την ακρίβεια των παραγόμενων περιλήψεων. Τα αποτελέσματα των πειραμάτων αποκαλύπτουν ότι η $ΓΒΤ$ λειτουργεί καλύτερα όταν γενικεύονται σπάνιες λέξεις. Από την άλλη πλευρά, η $ΓΟΟ$ προτιμά τη γενίκευση πιο συχνών λέξεων και η $ΓΟΟ$-$ΓΒΤ$ επιδιώκει μια αντιστάθμιση μεταξύ των δύο σχημάτων. Επιπλέον, όλες οι προσεγγίσεις παρουσιάζουν χαμηλότερη απόδοση όταν γενικεύονται πολύ συχνές λέξεις ($\theta_f = 1000$). Σε αυτή την περίπτωση, η υψηλή τιμή κατωφλίου συχνότητας εννοιών οδηγεί σε υπερ-γενίκευση, ομαδοποιώντας πολλές λέξεις στην ίδια ετικέτα (ή στο ίδιο αναγνωριστικό $ΑΕΛ$). Επομένως, ο μειωμένος αριθμός συνώνυμων (ή λέξεων συναφούς σημασίας) περιορίζει την ικανότητα εκτίμησης των κατάλληλων λέξεων, οι οποίες ταιριάζουν σε συγκεκριμένο σημασιολογικό πλαίσιο και, επίσης, καθιστά δύσκολο το έργο της φάσης μετα-επεξεργασίας για αντιστοίχιση των γενικευμένων εννοιών με τις κατάλληλες και συγκεκριμένες λέξεις.

5.7.3 Η επίδραση του μηχανισμού αποσαφήνισης έννοιας λέξεων

Η αποσαφήνιση της έννοιας των λέξεων ($ΑΕΛ$) χρησιμοποιείται για την αναγνώριση των διαφορετικών εννοιών που έχουν οι λέξεις του κειμένου, υποβοηθώντας τη διαδικασία γενίκευσης να εστιάσει στην κατάλληλη έννοια με σκοπό τη στοχευμένη γενίκευση περιεχομένου. Ο Πίνακας 5.5 παρουσιάζει την επίδοση των μοντέλων για γενίκευση ουσιαστικών, ενώ ο Πίνακας 5.6 αναφέρει τα αποτελέσματα θεωρώντας γενίκευση ουσιαστικών και ρημάτων. Κάποια από αυτά τα μοντέλα ξεπερνούν σε επιδόσεις άλλες σύγχρονες προσεγγίσεις (Πίνακας 5.3), σε όρους μετρικών *Rouge*. Τα θετικά πειραματικά αποτελέσματα, ειδικά στις στρατηγικές $ΓΒΤ$ και $ΓΟΟ$, οφείλονται στο γεγονός ότι ο μηχανισμός $ΑΕΛ$ επιδιώκει μια στοχευμένη και ακριβή γενίκευση περιεχομένου στη φάση της προ-επεξεργασίας. Επιπροσθέτως, στη φάση της μετα-επεξεργασίας, τα αναγνωριστικά $ΑΕΛ$ χρησιμοποιούνται για την επίτευξη μιας αποτελεσματικής αντιστοίχισης εννοιών (δηλ., αντικατάσταση των γενικευμένων εννοιών με συγκεκριμένες, δίνοντας προτεραιότητα σε συγκεκριμένες έννοιες ή σταθμίζοντας τις έννοιες που ανήκουν στο ίδιο μονοπάτι υπώνυμων ή υπερώνυμων εννοιών με τις γενικευμένες έννοιες, οι οποίες αντικαθίσταται από συγκεκριμένες).

Παρ' όλο που η $ΑΕΛ$ είναι αποτελεσματική στη διαδικασία της γενίκευσης κειμένου, η μετατροπή του αρχικού κειμένου σε μια μορφή $ΠΑΚ$ (δηλ., εφαρμογή $ΑΕΛ$ στο σύνολο του κειμένου και αποσαφήνιση ουσιαστικών ή ρημάτων ή και των δύο μερών του λόγου, καθώς και αντικατάσταση αυτών των λέξεων με τα αντίστοιχα αναγνωριστικά $ΑΕΛ$) δεν επιφέρει περαιτέρω βελτίωση. Αυτό είναι εμφανές στα αποτελέσματα των Πινάκων 5.7 και 5.8, όπου τα μοντέλα που βασίζονται σε ΠΑΚ παρουσιάζουν μειωμένες επιδόσεις, σε σύγκριση με τα υπόλοιπα μοντέλα (Πίνακες 5.5 και 5.6). Τα μοντέλα που βασίζονται σε $ΠΑΚ$ αποτυγχάνουν να επιτύχουν ικανοποιητική απόδοση, καθώς οδηγούν σε αύξηση του λεξιλογίου του συνόλου δεδομένων μειώνοντας ταυτόχρονα τον αριθμό των εμφανίσεων κάθε έννοιας στο κείμενο. Αυτό επηρεάζει την ικανότητα ενός μοντέλου μηχανικής

μάθησης να εκπαιδευτεί επαρκώς. Επιπλέον, η τρίτη φάση της μετα-επεξεργασίας δεν είναι ικανή να αντικαταστήσει το μεγάλο πλήθος εννοιών (οι οποίες αναπαρίστανται με αναγνωριστικά *ΑΕΛ*) με τις κατάλληλες λέξεις του αρχικού κειμένου. Παρ' όλα αυτά, συγκρίνοντας τα σχήματα γενίκευσης ουσιαστικών που βασίζονται σε *ΠΑΚ* (Πίνακας 5.7) και τα αντίστοιχα σχήματα γενίκευσης ουσιαστικών και ρημάτων (Πίνακας 5.8) με τις βασικές προσεγγίσεις (*ΠΑΚ-ο* και *ΠΑΚ-ο-ρ*, αντίστοιχα), γίνεται προφανές, ότι οι χρησιμοποιούμενες στρατηγικές γενίκευσης οδηγούν σε αυξημένες τιμές των μετρικών *Rouge*. Αυτό αποδίδεται στο γεγονός ότι η μεθοδολογία γενίκευσης μειώνει το μέγεθος του λεξιλογίου, αυξάνοντας ταυτόχρονα τη συχνότητα των λέξεων στο κείμενο. Αυτό αποτελεί ένδειξη ότι η προτεινόμενη μεθοδολογία βελτιώνει την απόδοση της περίληψης κειμένου, ακόμη και στην περίπτωση που το χρησιμοποιούμενο σύνολο δεδομένων αποτελεί μια έκδοση σε μορφή *ΠΑΚ*.

Στα πειράματα που διενεργήθηκαν για τη σύγκριση μεταξύ της συστηματικής μεθοδολογίας *ΑΕΛ* που ενσωματώνεται στην προτεινόμενη προσέγγιση (Ενότητα 5.4.1) και μιας βασικής μορφής *ΑΕΛ* που επιστρέφει την πιο συχνή έννοια μίας λέξης σύμφωνα με το *WordNet*, διαπιστώθηκε ότι η εφαρμογή της προτεινόμενης μεθοδολογίας *ΑΕΛ* βελτιώνει τις επιδόσεις του συστήματος (Πίνακας 5.9). Πιο συγκεκριμένα, για γενίκευση σχετικά σπάνιων όρων (π.χ., $\theta_f = 100$) φαίνεται ότι έχουμε μεγαλύτερη βελτίωση με τη χρησιμοποιούμενη *ΑΕΛ*, σε όρους μετρικών *Rouge*, σε σύγκριση με τη γενίκευση συχνότερων όρων (π.χ., $\theta_f = 1000$). Αυτό μπορεί να οφείλεται στο γεγονός ότι ο προσδιορισμός της εκάστοτε έννοιας με χρήση της πιο συχνής έννοιας του *WordNet* δεν είναι ιδιαίτερα επιτυχής σε σπάνιους όρους, καθώς και οι έννοιες που αντιστοιχούν σε αυτούς τους όρους δεν είναι ιδιαίτερα συχνές. Από την άλλη πλευρά, ο προσδιορισμός της έννοιας συχνότερων όρων μπορεί να συμπίπτει σε μεγαλύτερο βαθμό με την έννοια που έχει την μεγαλύτερη συχνότητα εμφάνισης για μια λέξη. Επομένως, το συμπέρασμά είναι ότι η βασική *ΑΕΛ*, η οποία βασίζεται στην πιο συχνή έννοια, δεν είναι κατάλληλη για τον προσδιορισμό των εννοιών, ιδιαίτερα στην περίπτωση των σπάνιων όρων, και, σε κάθε περίπτωση, η εφαρμογή μιας συστηματικής μεθόδου *ΑΕΛ* μπορεί να φέρει καλύτερα αποτελέσματα, σύμφωνα με τα αποτελέσματα των πειραμάτων.

Θα πρέπει να σημειωθεί ότι στην περίπτωση του συνόλου δεδομένων *Gigaword*, τα αριθμητικά ψηφία του κειμένου έχουν αντικατασταθεί από το σύμβολο # [104]. Με δεδομένο ότι η μέθοδος *ΑΕΛ* που βασίζεται σε γνώση και προτείνεται για την εργασία μας δεν προσδιορίζει αριθμούς, η εν λόγω αντικατάσταση δεν επηρεάζει την επίδοση του συστήματος. Σε αντίθεση με την *ΑΕΛ*, η αναγνώριση ονοματικών οντοτήτων (*ΑΟΟ*), η οποία χρησιμοποιείται στην παρούσα προσέγγιση, είναι ικανή να αναγνωρίζει αριθμητικές έννοιες (π.χ., ημερομηνία, χρόνος, ποσοστό, χρηματικό ποσό, ποσότητα), οι οποίες αντικαθίστανται από τις αντίστοιχες ετικέτες ονοματικών οντοτήτων (π.χ., *DATE, TIME, PERCENT, MONEY, QUANTITY*), σύμφωνα με την προσέγγιση *ΑΟΟ* που χρησιμοποιείται).

5.7.4 Η επίδραση των μερών του λόγου

Οι στρατηγικές που βασίζονται σε *ΓΒΤ* θεωρούν γενίκευση (*i*) ουσιαστικών ή (*ii*) ρημάτων ή (*iii*) και των δύο μερών του λόγου (ουσιαστικών και ρημάτων). Η περίπτωση της γενίκευση μόνο ρημάτων απουσιάζει από τα πειράματα λόγω της χαμηλής συχνότητας αυτής της κατηγορίας στο κείμενο (Πίνακας 5.1) και, επομένως, της μειωμένης επίδρασης που παρουσιάζει. Αντίθετα, η γενίκευση των ουσιαστικών οδηγεί σε σημαντική βελτίωση της απόδοσης λόγω της υψηλής

συχνότητας παρουσίας αυτού του μέρους του λόγου στο κείμενο. Όταν ληφθούν υπόψη και τα δύο μέρη του λόγου (ουσιαστικά και ρήματα), η απόδοση του συστήματος βελτιώνεται περαιτέρω σύμφωνα με τις τιμές των μετρικών *Rouge* (Πίνακες 5.5 και 5.6). Στο ίδιο συμπέρασμα δεν μπορούμε να καταλήξουμε για μοντέλα που βασίζονται σε *ΠΑΚ* (Πίνακες 5.7 και 5.8), καθώς αυτά δεν παρουσιάζουν ικανοποιητική απόδοση για να μπορούν να γίνουν αξιόπιστες συγκρίσεις. Από τις μετρήσεις που έγιναν μπορούμε να συμπεράνουμε ότι τα μοντέλα που βασίζονται σε *ΓΒΤ*, τα οποία γενικεύουν και τα δύο μέρη του λόγου, ουσιαστικά και ρήματα με σχετικά χαμηλή συχνότητα στο σύνολο εκπαίδευσης, είναι αυτά που επιτυγχάνουν τις υψηλότερες τιμές σε όρους μετρικών *Rouge*.

5.7.5 Νέες λέξεις στην τελική περίληψη

Το ποσοστό των νέων λέξεων σε μια παραγόμενη περίληψη, οι οποίες δεν έχουν παρουσία στο αρχικό κείμενο, προσδιορίζεται μέσω της μετρικής NTR. Η μετρική αυτή ποσοτικοποιεί τον βαθμό που η προτεινόμενη προσέγγιση παράγει νέο κείμενο. Από τα αποτελέσματα γίνεται φανερό ότι οι μετρήσεις σε όρους NTR για τα προτεινόμενα μοντέλα είναι σε συμφωνία με τις αντίστοιχες μετρήσεις των σχετικών προσεγγίσεων *Words-lvt2k-1sent* και *Model #8* (Πίνακας 5.3). Πιο συγκεκριμένα, στον Πίνακα 5.5 παρατηρούμε ότι η βασική προσέγγιση παρουσιάζει την καλύτερη απόδοση σε όρους NTR, με τις νέες λέξεις να εμφανίζονται με συχνότητα: μια νέα λέξη για κάθε τέσσερις λέξεις της τελικής περίληψης. Σε μοντέλα που βασίζονται σε *ΠΑΚ*, η τιμή της μετρικής NTR μειώνεται σε μία νέα λέξη για κάθε πέντε λέξεις της περίληψης, λόγω της συγκεκριμένης σημασίας κάθε όρου (δηλ., χρήση αναγνωριστικού *ΑΕΛ* για κάθε όρο) στα δεδομένα, που περιορίζει το μοντέλο βαθιάς μάθησης και το οδηγεί στη διάκριση της συγκεκριμένης λειτουργίας του κάθε όρου. Γενικά, τα σχήματα που έχουν πολλές λέξεις με παρόμοια σημασία (δηλ., συνώνυμα), όπως η βασική προσέγγιση, επιτυγχάνουν αυξημένη τιμή σε όρους NTR. Αντίθετα, μοντέλα με λίγα συνώνυμα (π.χ., μοντέλα *ΠΑΚ*) παρουσιάζουν μειωμένη τιμή σε όρους NTR. Τα επίπεδα τιμών της μετρικής NTR μπορεί, επίσης, να σχετίζονται με την ποσότητα των εναλλακτικών επιλογών που έχει ένα μοντέλο για την εκτίμηση μιας λέξης κατά τη σύνθεση μιας περίληψης. Πιο συγκεκριμένα το συμπέρασμα εδώ είναι ότι όταν μια έννοια μπορεί να περιγραφεί με πολλές λέξεις ή υπάρχουν πολλά συνώνυμα στο κείμενο (δηλ., στο σύνολο εκπαίδευσης), τότε μπορεί να παρατηρηθούν υψηλότερες τιμές σε όρους NTR. Από την άλλη πλευρά, στην περίπτωση του περιορισμένου αριθμού συνώνυμων λέξεων, αυτό οδηγεί σε μειωμένη εμφάνιση νέων λέξεων στην παραγόμενη περίληψη.

5.7.6 Άπληστη και βέλτιστη προσέγγιση αντιστοίχισης εννοιών

Οι μικρές διαφορές μεταξύ της άπληστης και βέλτιστης αντιστοίχισης εννοιών (Πίνακας 5.10) οφείλονται κυρίως στο γεγονός ότι αυτές οι προσεγγίσεις υποβοηθούνται από τον προσδιορισμό της ομοιότητας κειμένου (Αλγόριθμοι 4.4 και 4.5). Συγκεκριμένα, αυτοί οι αλγόριθμοι διακρίνουν τις υποψήφιες έννοιες σύμφωνα με τα σημασιολογικά μονοπάτια των υπώνυμων ή υπερώνυμων τους (για στρατηγικές που βασίζονται σε *ΓΒΤ*) ή την καθορισμένη ετικέτα ονοματικών οντοτήτων (για τις στρατηγικές που βασίζονται σε *ΓΟΟ*). Επομένως, η μεθοδολογία αντιστοίχισης εννοιών προσπαθεί να επιλύσει τις συγκρούσεις μεταξύ δύο ή περισσότερων εννοιών που είναι υποψήφιες για την

αντικατάσταση της ίδιας γενικευμένης έννοιας. Δεδομένου ότι αυτές οι συγκρούσεις εμφανίζονται συχνότερα σε μεγαλύτερα κείμενα (π.χ., σε κείμενα του συνόλου δεδομένων *CNN/DailyMail*), οι διαφορές μεταξύ της άπληστης και της βέλτιστης αντιστοίχισης εννοιών γίνονται πιο εμφανείς εκεί, ωστόσο παραμένουν ασήμαντες όπως δείχνουν τα πειραματικά αποτελέσματα (Πίνακας 5.10). Τέλος, η άπληστη αντιστοίχιση εννοιών είναι μια απλή και αποτελεσματική μέθοδος που θα μπορούσε να χρησιμοποιηθεί ως εναλλακτική λύση αντικαθιστώντας τη βέλτιστη, ειδικά σε εκείνες τις περιπτώσεις που δεν παρέχεται αποτελεσματικό λογισμικό επίλυσης προβλημάτων ακέραιου γραμμικού προγραμματισμού ή χρειάζεται να γίνει η περίληψη ενός μεγάλου όγκου εγγράφων.

5.7.7 Ακρίβεια απόδοσης πληροφορίας

Για να διερευνηθεί η συνέπεια της πληροφορίας που περιλαμβάνουν οι παραγόμενες περιλήψεις σε σχέση με το αρχικό κείμενο, χρησιμοποιείται η μετρική *ΑΑΠ*, όπως περιγράφεται στην Ενότητα 5.3. Σύμφωνα με τα αποτελέσματα (Πίνακες 5.12 και 5.13) που παρουσιάζονται στην Ενότητα 5.6.7, οι τελικές περιλήψεις για το σύνολο δεδομένων *CNN/DailyMail* παρουσιάζουν αυξημένες τιμές *ΑΑΠ* σε σύγκριση με τα άλλα δύο σύνολα δεδομένων (*Gigaword* και *DUC 2004*). Αυτό αποδίδεται στο γεγονός ότι στην *ΠΚ* σε έκταση εγγράφου, οι παραγόμενες περιλήψεις καλύπτουν περισσότερα γεγονότα ή πληροφορίες του αρχικού κειμένου, ενώ στην *ΠΚ* εγγράφων μικρής έκτασης, οι σύντομες περιλήψεις δεν είναι ικανές να αναφέρουν επαρκώς τα γεγονότα ή τις πληροφορίες του αρχικού κειμένου. Αυτό είναι ιδιαίτερα εμφανές στα αποτελέσματα των στρατηγικών *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* σε *ΠΚ* έκτασης εγγράφου. Μια άλλη παρατήρηση είναι ότι, ενώ οι περιλήψεις του συνόλου δεδομένων *CNN/DailyMail* περιέχουν περισσότερες έννοιες από αυτές των άλλων συνόλων δεδομένων (δηλ., περισσότερες γενικευμένες έννοιες που χρειάζονται αντιστοίχιση με συγκεκριμένες στη φάση της μετα-επεξεργασίας), η μετρούμενη *ΑΑΠ* δεν μειώνεται. Κατά συνέπεια, η *ΑΑΠ* των εκτιμώμενων περιλήψεων δεν επηρεάζεται από την αύξηση του αριθμού των γενικευμένων εννοιών που χρειάζεται να αντικατασταθούν από πιο συγκεκριμένες έννοιες ή από την αύξηση του μήκους του αρχικού κειμένου ή της περίληψης. Αντίθετα, οι μεγαλύτερες σε έκταση περιλήψεις (δηλ., αυτές του συνόλου δεδομένων *CNN/DailyMail*) φαίνεται να επιτυγχάνουν υψηλότερες τιμές σε όρους *ΑΑΠ* από τις συντομότερες (π.χ., εκείνες του συνόλου δεδομένων *Gigaword*).

Αντίστοιχα με τις τιμές των μετρικών *Rouge*, η *ΑΑΠ* της στρατηγικής *ΓΒΤ* λαμβάνει τη μέγιστη τιμή της για $\theta_f = 100$ (δηλ., σε γενίκευση σχετικά σπάνιων εννοιών), ενώ στη στρατηγική *ΓΟΟ*, η *ΑΑΠ* τείνει σε υψηλότερες τιμές στην περίπτωση γενίκευσης συχνών όρων (π.χ., το μοντέλο *ΓΟΟ-σ1000-ο* παρουσιάζει την υψηλότερη τιμή *ΑΑΠ* στο σύνολο δεδομένων *CNN/DailyMail*). Συγκρίνοντας την *ΑΑΠ* στην περίπτωση γενίκευσης ουσιαστικών (Πίνακας 5.12) με αυτήν της γενίκευσης ουσιαστικών και ρημάτων (Πίνακας 5.13), συμπεραίνουμε ότι η προσθήκη των ρημάτων στη φάση γενίκευσης του περιεχομένου δημιουργεί μικρές διαφορές για τα μοντέλα *ΓΒΤ*, *ΓΟΟ* και *ΓΟΟ-ΓΒΤ* με τη βελτίωση της *ΑΑΠ* να μην θεωρείται σημαντική. Από την άλλη πλευρά, η γενίκευση ουσιαστικών και ρημάτων σε σχήματα που βασίζονται σε *ΠΑΚ* (*ΓΒΤ-ΠΑΚ*, *ΓΟΟ-ΠΑΚ* και *ΓΟΟ-ΓΒΤ-ΠΑΚ*) επηρεάζει αρνητικά την *ΑΑΠ*. Στις περισσότερες περιπτώσεις, τα σχήματα που βασίζονται σε *ΠΑΚ* αποτυγχάνουν να δημιουργήσουν σημασιολογικά συνεπείς περιλήψεις για τους ίδιους λόγους που αποτυγχάνουν να επιτύχουν ικανοποιητική απόδοση σε όρους μετρικών *Rouge*, όπως αναφέρθηκε παραπάνω.

Επίσης, να αναφερθεί ότι η *ΑΑΠ* των περιλήψεων που προκύπτουν από την προτεινόμενη προσέγγιση είναι βελτιωμένη σε σύγκριση με εκείνη των βασικών προσεγγίσεων και των περιλήψεων αναφοράς. Αυτό δε σημαίνει απαραίτητα ότι οι εκτιμώμενες περιλήψεις είναι καλύτερες από εκείνες που γράφτηκαν από άνθρωπο (περιλήψεις αναφοράς), καθώς οι συγγραφείς είναι σε θέση να γράψουν περιλήψεις με αναδιατυπωμένο περιεχόμενο σε τέτοιο βαθμό που η μετρική *ΑΑΠ* αποτυγχάνει να λειτουργήσει αξιόπιστα. Σε περιλήψεις που παράγονται από μηχανές, από την άλλη πλευρά, το σύστημα τείνει να αντιγράφει γεγονότα και πληροφορίες από το αρχικό κείμενο (π.χ. για το μοντέλο *ΓΒΤ-σ100-β5-ο* έχουμε $NTR = 24,97\%$, Πίνακας 5.5), επιτρέποντας στην *ΑΑΠ* (που μετρά την επικάλυψη των γεγονότων ή πληροφοριών μεταξύ μιας περίληψης και του αντίστοιχου αρχικού κειμένου) να επιτύχει υψηλές τιμές. Επομένως, η μειωμένη τιμή σε όρους *ΠΑΚ* των περιλήψεων αναφοράς δεν αποτελεί ένδειξη ότι αυτές οι περιλήψεις είναι κακής ποιότητας. Αντίθετα, οι υψηλές τιμές *ΑΑΠ*, ειδικά στην περίπτωση *ΠΚ* σε έκταση εγγράφου (δηλ., για το σύνολο δεδομένων *CNN/DailyMail*), αποτελούν ισχυρή ένδειξη ότι η προτεινόμενη μεθοδολογία παράγει σημασιολογικά συνεπείς περιλήψεις.

5.7.8 Βελτίωση προβλέψεων μηχανικής μάθησης

Ο κύριος σκοπός αυτής της εργασίας είναι να παρουσιάσει μιας νέα προσέγγιση, ικανή να ενισχύσει την απόδοση των μοντέλων βαθιάς μάθησης για την αυτόματη *ΠΚ*. Σε αυτή την κατεύθυνση, η εργασία αυτή προσπαθεί να αντιμετωπίσει δύο θεμελιώδη προβλήματα που επηρεάζουν αρνητικά τις προσεγγίσεις μηχανικής μάθησης, τα οποία είναι τα εξής: (*i*) η παροχή ενός επαρκή αριθμού παραδειγμάτων χρήσης στη φάση της εκπαίδευσης ενός μοντέλου μηχανικής μάθησης για προβλέψεις με μεγαλύτερη ακρίβεια και (*ii*) η εμφάνιση νέων παραδειγμάτων κατά τη διάρκεια των προβλέψεων, τα οποία περιλαμβάνουν κείμενο, νέες λέξεις ή φράσεις που δεν περιλαμβάνονται στο σύνολο εκπαίδευσης. Η εκτεταμένη πειραματική διαδικασία που διεξήχθη, η οποία βασίζεται σε προβλέψεις μηχανικής μάθησης (Ενότητα 4.6.2), οδηγεί στο συμπέρασμα ότι το προτεινόμενο πλαίσιο είναι ικανό να βελτιώσει τις επιδόσεις ενός τέτοιου μοντέλου. Αυτό επιβεβαιώνεται τόσο από τη χρήση της βασικής αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής (Ενότητα 3.3.1) όσο και από την αξιοποίηση των τεσσάρων σύγχρονων αρχιτεκτονικών βαθιάς μάθησης που εξετάστηκαν, το δίκτυο αντιγραφής άγνωστων λέξεων (Ενότητα 3.3.2) το μοντέλο ενισχυτικής μάθησης (Ενότητα 3.3.3), το μοντέλο μετασχηματιστών (Ενότητα 3.3.4) και την αρχιτεκτονική μετασχηματιστών προ-εκπαιδευμένου κωδικοποιητή (Ενότητα 3.3.4). Επομένως, σύμφωνα με τα πειραματικά αποτελέσματα, το προτεινόμενο πλαίσιο, μέσω του σημασιολογικού μετασχηματισμού των δεδομένων, που οδηγεί στη δημιουργία ενός περισσότερο ισορροπημένου συνόλου δεδομένων (δηλ., οι λεκτικές μονάδες έχουν επαρκή παρουσία στο σύνολο εκπαίδευσης), συνεισφέρει στη βελτίωση των προβλέψεων μηχανικής μάθησης που οδηγούν, με τη σειρά τους, στη βελτίωση των παραγομένων περιλήψεων.

5.8 Συμπεράσματα και μελλοντικές επεκτάσεις

Η ανάλυση που παρουσιάστηκε μέχρι τώρα αποδεικνύει ότι η προτεινόμενη προσέγγιση αποτελεί μια αποτελεσματική λύση για την αυτόματη *ΠΚ* με τη μέθοδο της παραγωγής κειμένου, καθώς παρουσιάζει ικανοποιητικές επιδόσεις σε σύγκριση με τις συναφείς ερευνητικές εργασίες. Ειδικά

στην περίπτωση των προηγμένων μοντέλων μηχανικής μάθησης, όπως το μοντέλο ενισχυτικής μάθησης ή οι προσεγγίσεις που βασίζονται σε αρχιτεκτονικές μετασχηματιστών (Πίνακας 5.11), διαπιστώθηκαν οι αυξημένες επιδόσεις της προτεινόμενης μεθοδολογίας τόσο σε όρους *Rouge* όσο και σε όρους *ΑΑΠ*. Όπως είδαμε και παραπάνω, το επίπεδο γενίκευσης επηρεάζει τα αποτελέσματα και, επίσης, ο καθορισμός του κατάλληλου ορίου για τη συχνότητα των εννοιών, οι οποίες είναι υποψήφιες για γενίκευση, μπορεί να βελτιώσει την απόδοση του συστήματος.

Η μειωμένη απόδοση των μοντέλων που βασίζονται σε *ΠΑΚ* αποδίδεται στους ακόλουθους λόγους: (*i*) Το σύστημα βαθιάς μάθησης αποτυγχάνει να εκτιμήσει την κατάλληλη ακολουθία λέξεων (δηλ., την περίληψη), λόγω του μεγάλου αριθμού διακριτών εννοιών (δηλαδή, όροι με συγκεκριμένη σημασία που αυξάνουν το μέγεθος του λεξιλογίου), και (*ii*) η φάση της μετα-επεξεργασίας δεν είναι ικανή τα αντιστοιχίσει τα αναγνωριστικά *ΑΕΛ* με τις κατάλληλους όρους, λόγω του μεγάλου αριθμού εννοιών, που απαιτείται να αντικατασταθούν από κατάλληλες λέξεις για να διαμορφωθεί η τελική περίληψη. Επιπροσθέτως, τα μοντέλα που βασίζονται σε γενίκευση μόνο ουσιαστικών παρουσιάζουν αυξημένες επιδόσεις, με περαιτέρω βελτίωση στην περίπτωση που λαμβάνονται υπόψη και τα δύο μέρη του λόγου, ουσιαστικά και ρήματα, στον μετασχηματισμό των δεδομένων. Επιπλέον, το ποσοστό νέων λέξεων (*NTR*) μειώνεται όταν τα συνώνυμα στο κείμενο είναι περιορισμένα (π.χ., σε μοντέλα που βασίζονται σε *ΠΑΚ*), ενώ αυτή η μέτρηση μεγιστοποιείται στην περίπτωση των βασικών προσεγγίσεων στις οποίες δεν έχει εφαρμοστεί κάποια στρατηγική γενίκευσης και περιλαμβάνουν αυξημένο αριθμό λέξεων με παρόμοια σημασία. Τέλος, στην περίπτωση της *ΠΚ* έκτασης εγγράφου παρατηρήσαμε υψηλότερες τιμές *ΑΑΠ* από την περίπτωση της *ΠΚ* μικρής έκτασης κειμένου. Αυτό σημαίνει ότι η *ΑΑΠ* δεν επηρεάζεται αρνητικά από την αύξηση της έκτασης του κειμένου αλλά αντίθετα σε περιλήψεις μεγαλύτερου σε έκταση κειμένου έχουμε βελτίωση της *ΑΑΠ*.

Παρ' όλο που αυτή η προσέγγιση παρουσιάζει ήδη ικανοποιητική απόδοση, θα μπορούσε να ενισχυθεί περαιτέρω, ιδίως όσον αφορά την εκτίμηση των βέλτιστων τιμών των παραμέτρων. Η αξιοποίηση της γνώσης που προσφέρει η παρούσα εργασία για τις παραμέτρους που επηρεάζουν την ακρίβεια των εκτιμώμενων περιλήψεων, καθώς και η περαιτέρω εξέταση των πτυχών που επηρεάζουν την απόδοση ενός τέτοιου συστήματος, θα μπορούσαν να οδηγήσουν στη δημιουργία ενός θεωρητικού μοντέλου για την εκτίμηση των βέλτιστων τιμών των παραμέτρων για περαιτέρω βελτίωση. Μια ακόμη επέκταση θα μπορούσε να βασίζεται στην παρατήρησή μας, ότι η ένταξη του δικτύου αντιγραφής λέξεων εκτός λεξιλογίου στο πλαίσιο της ενισχυτικής μάθησης επέφερε σημαντική βελτίωση στις επιδόσεις σε σχέση με την επίδοση του μοντέλου αυτού εκτός πλαισίου ενισχυτικής μάθησης. Με δεδομένο ότι τα δίκτυα μετασχηματιστών παρουσιάζουν βελτιωμένες επιδόσεις σε σχέση με τα άλλα δίκτυα, θα μπορούσε να εξεταστεί αν η ένταξη των δικτύων αυτών σε μια αρχιτεκτονική ενισχυτικής μάθησης θα επέφερε περαιτέρω βελτίωση. Τέλος, μια σημαντική προοπτική αφορά τη σημασιολογική αναπαράσταση των δεδομένων και την περαιτέρω διερεύνηση των σημασιολογικών μετασχηματισμών με σκοπό τη δημιουργία ενός περισσότερο ισορροπημένου συνόλου δεδομένων και, ταυτόχρονα, την αποφυγή της απώλειας πληροφορίας κατά την παραγωγή των τελικών περιλήψεων.

Κεφάλαιο 6

Αυτόματη περίληψη κειμένου με χρήση βαθιάς μάθησης και σημασιολογικών γραφημάτων

6.1 Γενικά

Όπως έχει ήδη αναφερθεί και στην Ενότητα 4.2, κάποιες από τις πρώτες εργασίες στον τομέα της αυτόματης περίληψης κειμένου με τη μέθοδο της παραγωγής κειμένου ακολουθούσαν την προσέγγιση της μετατροπής του αρχικού κειμένου σε γράφημα [116, 117]. Με χρήση του γραφήματος αυτού, μέσω της συγχώνευσης ή της απόρριψης κόμβων και ταυτόχρονα της διατήρησης των πιο σημαντικών από αυτούς, προέκυπτε μια σύνοψη του αρχικού γραφήματος (δηλ., ένα γράφημα μικρότερης διάστασης από το αρχικό), το οποίο αντιστοιχούσε στο γράφημα της περίληψης. Το γράφημα αυτό αποτελούσε μια ενδιάμεση κατάσταση, από το οποίο μπορούσε να εξαχθεί η περίληψη σε μορφή κειμένου. Στις προσεγγίσεις αυτές η αναπαράσταση των γραφημάτων περιελάμβανε την ακολουθία των λέξεων του κειμένου και οι κοινές λέξεις των προτάσεων αποτελούσαν κοινούς κόμβους για περισσότερες από μία προτάσεις. Δηλαδή, μπορούσε σχετικά εύκολα να γίνει η μετατροπή από κείμενο σε γράφημα και το αντίστροφο. Ωστόσο, αυτή η μεθοδολογία δημιουργούσε ένα αρκετά μεγάλο αρχικό γράφημα που ήταν δύσκολα διαχειρίσιμο στη συνέχεια, και επίσης, τα γραφήματα αυτά δεν ενσωμάτωναν κάποια αναπαράσταση σημασιολογίας παρά μόνο σύνταξης και γραμματικής. Για να ξεπεραστούν τα προβλήματα αυτά, θα μπορούσε να χρησιμοποιηθεί κάποια εναλλακτική αναπαράσταση του κειμένου σε μορφή γραφήματος, η οποία, όμως, ενσωματώνει σημασιολογία και χρησιμοποιεί κάποιο είδος αφαίρεσης προκειμένου να μειωθεί το μέγεθος του αρχικού γραφήματος. Μια τέτοια προοπτική θα μπορούσε να επιτευχθεί με τη χρήση σημασιολογικών γραφημάτων. Ένας τύπος τέτοιου γραφήματος αποτελεί το μοντέλο αναπαράστασης αφηρημένης έννοιας (*abstract meaning representation - AMR*) [168], το οποίο εστιάζει στη σημασιολογική αναπαράσταση κειμένου.

Τα γραφήματα *AMR*, τα οποία παρουσιάζονται στην εργασία [168], παρέχουν μια σημασιολογική αναπαράσταση των προτάσεων ενός κειμένου μέσω ενός κατευθυνόμενου άκυκλου γραφήματος το οποίο περιέχει μια ρίζα (δηλ., γράφημα τύπου δέντρο). Οι κόμβοι ενός γραφήματος *AMR*

αναπαριστούν όρους του κειμένου ή έννοιες (π.χ., λέξεις, σημασιολογικούς ρόλους ή λέξεις κλειδιά) και οι ακμές αναπαριστούν σημασιολογικές σχέσεις μεταξύ των κόμβων. Τα σύνολα των ρόλων και των σημασιολογικών σχέσεων προέρχονται από το σώμα κειμένων *PropBank (proposition bank)*, το οποίο βασίζεται σε επισημείωση σημασιολογικών ρόλων (δηλ., επισημείωση σχέσεων κατηγορήματος-ορίσματος) [169]. Ουσιαστικά, τα γραφήματα *AMR* προσπαθούν να αναπαραστήσουν εκφράσεις όπως "*ποιος κάνει κάτι σε ποιον*", σύμφωνα με τις πληροφορίες που περιλαμβάνονται στο κείμενο. Επίσης, τα γραφήματα αυτά δημιουργούν τις ίδιες αναπαραστάσεις για τις προτάσεις που έχουν το ίδιο νόημα (π.χ., οι προτάσεις "*he eats apples*" και "*apples are eaten by him*" έχουν την ίδια αναπαράσταση σε μορφή *AMR*), καθώς τα γραφήματα εστιάζουν περισσότερο στη σημασιολογία και λιγότερο στη σύνταξη και στη γραμματική. Για την ανάκτηση των γραφημάτων *AMR* από ένα κείμενο, έχουν αναπτυχθεί ορισμένες προσεγγίσεις (AMR parsers) που κάνουν εφικτή την αυτόματη ανάκτηση τέτοιων γραφημάτων [170, 171, 172, 173, 174].

Η παρούσα εργασία αποτελεί μια προσπάθεια συνδυασμού προβλέψεων μηχανικής μάθησης και τεχνικών σημασιολογικής αναπαράστασης κειμένου για την αυτόματη *ΠΚ*. Ένας τέτοιος συνδυασμός μεθοδολογίας δεν έχει διερευνηθεί επαρκώς στη σχετική βιβλιογραφία, καθώς η έρευνα σε αυτόν τον τομέα επικεντρώνεται κυρίως είτε σε προσεγγίσεις μηχανικής μάθησης είτε σε τεχνικές που βασίζονται στη γνώση και στη σημασιολογία, οι οποίες αντιμετωπίζονται ως ξεχωριστές μεθοδολογίες [9]. Σε αυτή την κατεύθυνση, συνδυασμού πτυχών και χαρακτηριστικών προβλέψεων μηχανικής μάθησης και σημασιολογικών τεχνικών στον τομέα της αυτόματης *ΠΚ*, έχει καταγραφεί κάποια προηγούμενη ερευνητική δραστηριότητα [175, 176, 177, 178, 140]. Οι περισσότερο σχετικές από αυτές τις εργασίες, όπως οι [175, 176, 178], οι οποίες βασίζονται σε αναπαράσταση σημασιολογικών γραφημάτων για την αυτόματη *ΠΚ*, περιγράφονται στην ενότητα: 6.2 που ακολουθεί. Παρά την προηγούμενη ερευνητική δραστηριότητα, υπάρχει χώρος για περαιτέρω διερεύνηση, όπως θα δούμε στη συνέχεια της παρούσας εργασίας. Με σκοπό την περαιτέρω διερεύνηση, το προτεινόμενο πλαίσιο εστιάζει στην αξιοποίηση μηχανικής μάθησης και τεχνικών αναπαράστασης περιεχομένου σε μορφή γραφήματος, σε μια προσπάθεια συνεισφοράς στον τομέα της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου.

Πιο συγκεκριμένα, η παρούσα εργασία εστιάζει στη βελτίωση των προσεγγίσεων που βασίζονται σε σημασιολογικά γραφήματα για την εκτίμηση περιλήψεων ενός εγγράφου με τη μέθοδο της παραγωγής κειμένου. Οι θετικές προοπτικές που παρουσιάζουν αυτές οι προσεγγίσεις, όπως η σημασιολογική αναπαράσταση σε μια δομημένη, αναγνώσιμη από τις μηχανές και συνοπτική μορφή του περιεχομένου, αποτελούν βασικό κίνητρο για περαιτέρω έρευνα στον συγκεκριμένο τομέα. Επιπλέον, αυτές οι μεθοδολογίες μειώνουν τον πλεονασμό και παράγουν σημασιολογικά συναφείς προτάσεις που διαμορφώνουν μία περίληψη. Η πρόσφατη έρευνα σε αυτόν τον τομέα έχει οδηγήσει στην ανάπτυξη προσεγγίσεων που αξιοποιούν *AMR* γραφήματα, διαμορφώνοντας ένα νέο πεδίο έρευνας [179, 180], το οποίο, όμως, απαιτεί περαιτέρω διερεύνηση, καθώς η χρήση σημασιολογικών αναπαραστάσεων στο πλαίσιο της αυτόματης *ΠΚ* δεν έχει μελετηθεί επαρκώς. Επίσης, θα πρέπει να σημειωθεί ότι οι προσεγγίσεις που βασίζονται σε σημασιολογικά γραφήματα εξακολουθούν να παρουσιάζουν μειωμένη απόδοση σε σύγκριση με εκείνες που βασίζονται μόνο σε βαθιά μάθηση [181] (δηλ., συστήματα που βασίζονται σε τεχνικές βαθιάς μάθησης χωρίς να χρησιμοποιούν κάποια σημασιολογική αναπαράσταση). Ωστόσο, οι προσεγγίσεις που αξιοποιούν τη σημασιολογική αναπαράσταση του περιεχομένου σε μορφή γραφήματος βελτιώνονται συνεχώς, σύμφωνα με τη σχετική βιβλιογραφία [182, 183, 184, 185, 176, 178]. Σε αυτή την κατεύθυνση, η μεθοδολογία που περιγράφεται στην παρούσα εργασία εμπίπτει στην κατηγορία των προσεγγίσεων που βασίζονται σε

σημασιολογικά γραφήματα για την αυτόματη *ΠΚ*. Στην ενότητα 6.2 που ακολουθεί, παρουσιάζεται η σχετική βιβλιογραφία, περιγράφοντας τις διαφορές και τα πλεονεκτήματα της προτεινόμενης μεθοδολογίας σε σχέση με τις συναφείς εργασίες.

Στην κατεύθυνση αξιοποίησης προβλέψεων μηχανικής μάθησης για την εκτίμηση μιας περίληψης σε μορφή κειμένου ενός σημασιολογικού γραφήματος εισόδου, στη συνέχεια της παρούσας έρευνας, επιδιώκουμε την αξιοποίηση των χαρακτηριστικών των *AMR* γραφημάτων με σκοπό να εξετάσουμε την προοπτική αυτή στο πλαίσιο της αυτόματης *ΠΚ*. Συγκεκριμένα, διερευνούμε εάν μια σημασιολογική αναπαράσταση ενός αρχικού κειμένου σε μορφή γραφήματος, ως ενδιάμεσο βήμα, θα μπορούσε να αποτελέσει έναν αποτελεσματικό παράγοντα για την εκτίμηση της περίληψης του αρχικού κειμένου. Πιο συγκεκριμένα, αυτή η προσέγγιση εστιάζει στην εκτίμηση μιας περίληψης ενός αρχικού κειμένου ακολουθώντας δύο κύρια βήματα: (*i*) την ανάκτηση ενός σημασιολογικού γραφήματος από το αρχικό κείμενο και (*ii*) τη χρήση του σημασιολογικού γραφήματος για την εκτίμηση της περίληψης σε μορφή κειμένου. Στο πρώτο βήμα, ένας αποτελεσματικός αναλυτής (*AMR parser*) θα μπορούσε να χρησιμοποιηθεί για την ανάκτηση των σημασιολογικών γραφημάτων, τα οποία χρειάζεται να είναι διαθέσιμα για τη συνέχεια. Στο δεύτερο βήμα, αντιμετωπίζουμε το πρόβλημα ως πρόβλημα μάθησης από γράφημα-σε-κείμενο (*graph-to-text*) ή, πιο συγκεκριμένα, από γράφημα-σε-περίληψη (*graph-to-summary*), αξιοποιώντας τεχνικές μηχανικής μάθησης για την εκτίμηση της περίληψης ενός υποψήφιου για περίληψη κειμένου.

Στο πλαίσιο που περιγράφεται παραπάνω, η κύρια συνεισφορά της παρούσας εργασίας συνοψίζεται στη διερεύνηση τεσσάρων προοπτικών, στις οποίες περιλαμβάνεται: (*i*) η αντιμετώπιση του προβλήματος της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου ως ένα πρόβλημα μάθησης από γράφημα-σε-περίληψη, χρησιμοποιώντας τεχνικές βαθιάς μάθησης, (*ii*) η εξέταση μιας σειράς από διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων βαθιάς μάθησης, (*iii*) η διερεύνηση σχημάτων σημασιολογικής αναπαράστασης περιεχομένου που βασίζονται σε σημασιολογικά γραφήματα και (*iv*) η εισαγωγή ενός συνόλου μετρικών αξιολόγησης που εστιάζει στην παροχή ποιοτικής αξιολόγηση για την αυτόματη *ΠΚ*.

Σύμφωνα με τη βιβλιογραφική έρευνα που έγινε, κανένα από τα προαναφερθέντα σημεία της συνεισφοράς μας δεν έχει μελετηθεί στη σχετική βιβλιογραφία. Προβλέψεις από σημασιολογικό γράφημα-σε-περίληψη με μοντέλα βαθιάς μάθησης (δηλ., μοντέλα των οποίων η είσοδος είναι μόνο μια αναπαράσταση γραφήματος και η εκτιμώμενη περίληψη παράγεται σε μορφή κειμένου, χωρίς άλλα ενδιάμεσα βήματα) χρειάζεται να διερευνηθούν για την αυτόματη *ΠΚ*. Ωστόσο, όπως περιγράφεται με μεγαλύτερη λεπτομέρεια στην Ενότητα 6.2 που ακολουθεί, έχουν προταθεί ορισμένες προσεγγίσεις που βασίζονται σε ένα σύνολο μετατροπών από ένα σημασιολογικό γράφημα του αρχικού κειμένου σε ένα σημασιολογικό γράφημα της περίληψης και, στη συνέχεια, το τελευταίο γράφημα χρησιμοποιείται για τη λήψη της αντίστοιχης περίληψης σε μορφή κειμένου [183, 182, 184, 185]. Επιπλέον, άλλες προσεγγίσεις χρησιμοποιούν το κείμενο εισόδου μαζί με τη σημασιολογική αναπαράστασή του για την παραγωγή μιας περίληψης, με πρόσθετο υπολογιστικό φορτίο [175, 176]. Σε αντίθεση με τις προαναφερθείσες προσεγγίσεις, η παρούσα εργασία βασίζεται σε προβλέψεις μηχανικής μάθησης από γράφημα-σε-περίληψη με την προτεινόμενη λύση να ξεπερνά σε επιδόσεις τις προαναφερθείσες προσεγγίσεις (Ενότητα 7.6), αποφεύγοντας το πρόσθετο υπολογιστικό φορτίο της χρήσης του αρχικού κειμένου στη φάση της μηχανικής μάθησης και, συγχρόνως, περιορίζοντας την απώλεια πληροφορίας λόγω διαδοχικών μετατροπών από το αρχικό κείμενο έως την περίληψη. Είναι γνωστό ότι έχει αναπτυχθεί μεθοδολογία για παραγωγή κειμένου από *AMR* γράφημα

(*graph-to-text generators*) [186, 173]. Ωστόσο, η προοπτική της παραγωγής μιας περίληψης από ένα σημασιολογικό γράφημα ενός αρχικού κειμένου δεν έχει μελετηθεί, σύμφωνα με τη σχετική βιβλιογραφία, και γι΄ αυτό αποτελεί το αντικείμενο διερεύνησης της παρούσας εργασίας. Επιπλέον, εξετάζουμε μια σειρά από αρχιτεκτονικές νευρωνικών δικτύων βαθιάς μάθησης, διερευνώντας τόσο την προσαρμογή τους στο προτεινόμενο πλαίσιο όσο και τα πλεονεκτήματα ή τα μειονεκτήματα που παρουσιάζουν. Οι προσεγγίσεις βαθιάς μάθησης περιλαμβάνουν αναδρομικά νευρωνικά δίκτυα αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή, ενισχυτική μάθηση, δίκτυα μετασχηματιστών και προεκπαιδευμένα μοντέλα γλωσσικής αναπαράστασης. Επιπλέον, διερευνώνται διάφορες μορφές σημασιολογικής αναπαράστασης σε μορφή γραφήματος, οι οποίες αποτελούν την είσοδο για ένα μοντέλο βαθιάς μάθησης, προσδιορίζοντας τις πιο αποτελεσματικές από αυτές. Να σημειωθεί ότι τα χρησιμοποιούμενα μοντέλα βαθιάς μάθησης, σε συνδυασμό με τις προτεινόμενες αναπαραστάσεις σημασιολογικών γραφημάτων, δεν έχουν μελετηθεί από τη σχετική βιβλιογραφία των προσεγγίσεων που βασίζονται σε *AMR* γραφήματα και η διερεύνησή τους αποτελεί αντικείμενο της παρούσας εργασίας.

Μια ακόμη σημαντική συνεισφορά αποτελεί η εισαγωγή ενός νέου συνόλου μετρικών που εστιάζει στην ποιοτική αξιολόγηση των παραγόμενων περιλήψεων. Ειδικότερα, προτείνεται μια επέκταση της μετρικής προσδιορισμού της *ακρίβειας απόδοσης πληροφορίας* (*ΑΑΠ*), η οποία παρουσιάστηκε στην Ενότητα 5.3. Το νέο σύνολο μετρικών, το οποίο περιγράφεται με λεπτομέρεια στην Ενότητα 7.3, προσδιορίζει τη *συνέπεια απόδοσης πληροφορίας* (*ΣΑΠ*) και λαμβάνει υπόψη του την έκταση μιας περίληψης σε σχέση με την έκταση του αρχικού κειμένου, σε μια προσπάθεια σταθμισμένων εκτιμήσεων για τον προσδιορισμό της *ΣΑΠ* των παραγόμενων περιλήψεων.

Για την αξιολόγηση της παρούσας προσέγγισης, διεξήχθη μια εκτεταμένη πειραματική διαδικασία, η οποία παρουσιάζεται στο Κεφάλαιο 7. Στο πλαίσιο της αξιολόγησης εξετάζονται σημαντικές πτυχές του προτεινόμενου πλαισίου χρησιμοποιώντας δύο δημοφιλή σύνολα δεδομένων. Τα πειραματικά αποτελέσματα, καθώς και η σύγκριση της επίδοσης του πλαισίου που παρουσιάζεται σε αυτή την εργασία με την επίδοση των πιο σχετικών προσεγγίσεων της βιβλιογραφίας, επιβεβαιώνουν την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας.

Το υπόλοιπο αυτού του κεφαλαίου οργανώνεται ως εξής: Η ενότητα 6.2 παρέχει μια επισκόπηση της σχετικής βιβλιογραφίας, ενώ η ενότητα 6.3 κάνει μια εισαγωγή στα σημασιολογικά γραφήματα. Η ενότητα 6.4 παρουσιάζει το προτεινόμενο πλαίσιο που περιλαμβάνει την ανάκτηση των σημασιολογικών γραφημάτων (Ενότητα 6.5), την κατασκευή σημασιολογικού γραφήματος (Ενότητα 6.6) και τις τεχνικές μετασχηματισμού των σημασιολογικών γραφημάτων (Ενότητα 6.7). Τέλος, η ενότητα 6.8 εισάγει τη μεθοδολογία μηχανικής μάθησης για την εκτίμηση των περιλήψεων.

6.2 Σχετικές εργασίες

Η έρευνα που έχει διεξαχθεί έως σήμερα στον τομέα της αυτόματης *ΠΚ* ενός εγγράφου έχει επιδείξει κάποια πρόοδο στην αξιοποίηση σημασιολογικών γραφημάτων. Στην εργασία [187], οι συγγραφείς προτείνουν μια μέθοδο που βασίζεται σε εννοιολογικά γραφήματα [188], η οποία εστιάζει στην κατάταξη των κόμβων ενός γραφήματος σύμφωνα με τη σημαντικότητά τους, και στη συνέχεια, με αφαίρεση κόμβων ή κλάδεμα μέρους ενός γραφήματος προκύπτει ένα γράφημα μικρότερης διάστασης από το αρχικό. Σε αυτή την κατεύθυνση, η σχετική μεθοδολογία

περιλαμβάνει σταθμισμένα εννοιολογικά γραφήματα και σημασιολογικά μοτίβα από τα ηλεκτρονικά λεξικά *WordNet* [43, 44] και *VerbNet* [189], για τον προσδιορισμό συνεκτικών δομών μεταξύ εννοιών που διατηρούνται για τον σχηματισμό μιας περίληψης. Στην εργασία [190] χρησιμοποιείται το μοντέλο αναπαράστασης αφηρημένης έννοιας (*AMR*) και άλλοι σημασιολογικοί πόροι όπως το *WordNet* και το *PropBank* [169] για την επιλογή των πιο σημαντικών εννοιών που δημιουργούν μια περίληψη. Αυτή η προσέγγιση συνδυάζει λεκτικές οντότητες της θεωρίας ρητορικής δομής (*rhetorical structure theory - RST*) [191] και τον γνωστό αλγόριθμο γραφημάτων *PageRank* [192] για τον προσδιορισμό των πιο σχετικών εννοιών που συνθέτουν μια περίληψη. Για τη σύνθεση της περίληψης, χρησιμοποιείται το εργαλείο λογισμικού παραγωγής φυσικής γλώσσας *SimpleNLG* [193]. Σε αντίθεση με τις προαναφερθείσες προσεγγίσεις, οι οποίες χρησιμοποιούν ευρετικούς αλγόριθμους για τη μείωση της διάστασης ενός αρχικού γραφήματος που οδηγεί στο γράφημα της περίληψης, η προτεινόμενη προσέγγιση αξιοποιεί μοντέλα μηχανικής μάθησης, που δέχονται ως είσοδο κατάλληλες αναπαραστάσεις σημασιολογικών γραφημάτων του αρχικού κειμένου, προκειμένου τα μοντέλα αυτά να εκτιμήσουν μια περίληψη σε μορφή κειμένου.

Έχουν αναπτυχθεί ορισμένες προσεγγίσεις που βασίζονται σε σημασιολογικά γραφήματα, οι οποίες ακολουθούν μια σειρά από βήματα για την εκτίμηση μιας περίληψης ενός κειμένου. Σύμφωνα με αυτές τις προσεγγίσεις [183, 182, 184, 185], στο πρώτο βήμα, από ένα υποψήφιο για περίληψη κείμενο ανακτάται ένα γράφημα τύπου *AMR*, το οποίο στη συνέχεια αξιοποιείται για την εκτίμηση ενός νέου γραφήματος μικρότερης διάστασης που, με τη σειρά του, χρησιμοποιείται για τη δημιουργία μιας περίληψης σε μορφή κειμένου. Πιο συγκεκριμένα, στην εργασία [183], οι συγγραφείς προτείνουν ένα πλαίσιο για τη δημιουργία ενός ενιαίου σημασιολογικού γραφήματος ενός εγγράφου συνδυάζοντας τα επιμέρους γραφήματα κάθε πρότασης του εγγράφου. Η προσέγγισή τους εστιάζει στην εκτίμηση ενός γραφήματος μικρότερης διάστασης που αποτελεί το γράφημα της περίληψης. Το πρόβλημα αυτό διατυπώνεται ως ένα πρόβλημα από γράφημα-σε-γράφημα (*graph-to-graph*) και επιλύεται με ακέραιο γραμμικό προγραμματισμό και αξιοποίηση εποπτευόμενης μάθησης για την εκτίμηση των παραμέτρων. Σε μια παρόμοια προσέγγιση [182], οι συγγραφείς προτείνουν μια σειρά βημάτων για την επιλογή των πιο σημαντικών προτάσεων ενός κειμένου, ανακτώντας ένα υπο-γράφημα τύπου *AMR* για κάθε επιλεγμένη πρόταση. Σε μια προσέγγιση που στοχεύει στην επίλυση προβλημάτων αναφοράς, η εργασία [184] προτείνει έναν αλγόριθμο που εξάγει ένα σημασιολογικό γράφημα ενός εγγράφου, χρησιμοποιώντας μεθοδολογία ανάλυσης συν-αναφοράς για τη συγχώνευση των υπο-γραφημάτων των επιμέρους προτάσεων. Στη συνέχεια, επιχειρείται ο προσδιορισμός των πιο σημαντικών σχέσεων μεταξύ των κόμβων, αξιοποιώντας πληροφορία που εξάγεται λαμβάνοντας υπόψη το περιβάλλον κάθε μονοπατιού στο γράφημα. Η μεθοδολογία αυτή οδηγεί στην εξαγωγή ενός νέου γραφήματος μικρότερης διάστασης. Επιπλέον, στην εργασία [185], περιγράφεται μια μεθοδολογία που περιλαμβάνει ανάλυση συν-αναφοράς πριν από την ανάκτηση των σημασιολογικών γραφημάτων των επιμέρους προτάσεων. Τα σημασιολογικά γραφήματα χρησιμοποιούνται από έναν αλγόριθμο που προτείνεται για τη συγχώνευση των *AMR* γραφημάτων με σκοπό το σχηματισμό του συνοπτικού γραφήματος. Για την παραγωγή της περίληψης σε μορφή κειμένου, οι τρεις τελευταίες προσεγγίσεις που περιγράφονται παραπάνω, [182], [184] και [185], χρησιμοποιούν προϋπάρχουσα μεθοδολογία μετατροπής *AMR* γραφήματος σε κείμενο [186, 173].

Όλες οι προσεγγίσεις που περιγράφονται στην προηγούμενη παράγραφο χρησιμοποιούν μια σειρά από μετασχηματισμούς, οι οποίοι περιλαμβάνουν μετατροπές από το αρχικό κείμενο σε γράφημα, από το γράφημα του κειμένου στο γράφημα της περίληψης και από το γράφημα της

περίληψης στην περίληψη. Οι μετασχηματισμοί αυτοί αυξάνουν το υπολογιστικό φορτίο των παραπάνω προσεγγίσεων και, επίσης, οι πολλαπλές μετατροπές ενδέχεται να οδηγήσουν σε σημαντική απώλεια πληροφορίας κατά τη διάρκεια των διαδοχικών βημάτων που ακολουθούνται. Σε αντίθεση με την παραπάνω μεθοδολογία, η λύση που προτείνουμε στην παρούσα εργασία βασίζεται στην αξιοποίηση των πλεονεκτημάτων που παρέχει η σημασιολογική αναπαράστασης του περιεχομένου ενός υποψήφιου για περίληψη κειμένου, όπως κάνουν οι παραπάνω προσεγγίσεις στο πρώτο βήμα, για την εκτίμηση μιας περίληψης σε μορφή κειμένου χωρίς άλλους ενδιάμεσους μετασχηματισμούς. Συγκεκριμένα, η προσέγγισή μας αντιμετωπίζει το πρόβλημα ως ένα πρόβλημα μάθησης από γράφημα-σε-περίληψη, αποφεύγοντας περαιτέρω μετασχηματισμούς και μετατροπές που μπορεί να οδηγήσουν σε μειωμένη απόδοση.

Εφόσον η εργασία αυτή περιλαμβάνει μεθοδολογία για την ανάκτηση σημασιολογικών γραφημάτων από κείμενα καθώς και την κατασκευή και τον μετασχηματισμό των γραφημάτων αυτών, οι συγγραφείς στην εργασία [194] παρουσιάζουν μια ανάλυση σχετικά με μεθοδολογία κατασκευής *AMR* γραφημάτων στο πλαίσιο της αυτόματης *ΠΚ*. Στην εργασία αυτή εξετάζονται στρατηγικές που βασίζονται στη συγχώνευση γραφημάτων και διερευνάται η απόδοση των μεθόδων επιλογής περιεχομένου. Στο πλαίσιο αυτό, προτείνεται μια μέθοδος συγχώνευσης των κόμβων ενός σημασιολογικού γραφήματος, αξιοποιώντας μεθοδολογία επίλυσης των προβλημάτων αναφοράς για τη στοχευμένη συγχώνευση εννοιών. Σε μια παρόμοια κατεύθυνση, στην εργασία [195] προτείνεται μια προσέγγιση για τη συγχώνευση των γραφημάτων των επιμέρους προτάσεων με την αξιοποίηση αναγνώρισης ονοματικών οντοτήτων και αντωνυμικών αναφορών, η οποία αποσκοπεί στη δημιουργία ενός ενιαίου γραφήματος ενός εγγράφου. Παρόμοια με τις παραπάνω προσεγγίσεις, στην παρούσα εργασία εξετάζουμε τεχνικές συγχώνευσης των σημασιολογικών γραφημάτων των επιμέρους προτάσεων ενός κειμένου προκειμένου να δημιουργηθεί ένα ενιαίο σημασιολογικό γράφημα για ένα κείμενο.

Στο πλαίσιο της μεθοδολογίας που συνδυάζει σημασιολογική αναπαράσταση κειμένου με προβλέψεις μηχανικής μάθησης για την αυτόματη *ΠΚ*, οι συγγραφείς της εργασίας [175] επεκτείνουν το μοντέλο νευρωνικών δικτύων αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής, που παρουσιάζεται στην εργασία [104] και χρησιμοποιείται ως βασικό μοντέλο, ενσωματώνοντας έναν κωδικοποιητή *AMR* γραφημάτων, ο οποίος βασίζεται σε μια δενδροειδή αρχιτεκτονική μονάδων *LSTM* [196]. Σε αυτή την προσέγγιση εφαρμόζεται ένα σχήμα εκπαίδευσης δύο σταδίων. Στο πρώτο βήμα, το βασικό μοντέλο εκπαιδεύεται χρησιμοποιώντας ζεύγη κειμένου-περίληψης ως παραδείγματα χρήσης και, στη δεύτερη φάση, τα *AMR* γραφήματα που έχουν ανακτηθεί από το κείμενο χρησιμοποιούνται για περαιτέρω εκπαίδευση και προσαρμογή των παραμέτρων του *AMR* κωδικοποιητή, που οδηγεί στη βελτίωση του αρχικού μοντέλου. Αυτή η προσέγγιση συνδυάζει σημασιολογικές και συντακτικές πτυχές του αρχικού κειμένου, βελτιώνοντας τις παραγόμενες περιλήψεις. Στην εργασία [176], οι συγγραφείς προτείνουν μια προσέγγιση που χρησιμοποιεί πληροφορίες από το αρχικό κείμενο για την υποβοήθηση της διαδικασίας παραγωγής φυσικής γλώσσας από *AMR* γράφημα με σκοπό τη σύνθεση μιας περίληψης. Η προσέγγιση βασίζεται σε ένα μοντέλο μηχανικής μάθησης ακολουθία-σε-ακολουθία [71] (δηλ., με δεδομένη μια ακολουθία εισόδου προβλέπει μια ακολουθία εξόδου) για την εκτίμηση μιας περίληψης με δεδομένο ένα γράφημα *AMR* του αρχικού κειμένου. Αυτή η προσέγγιση χρησιμοποιεί πληροφορίες του αρχικού κειμένου που δεν υπάρχουν στην *AMR* αναπαράστασή του, για τη βελτίωση της ποιότητας μιας εκτιμώμενης περίληψης. Επιπλέον, η εργασία [178] προτείνει μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή που λαμβάνει ως είσοδο ένα αρχικό

κείμενο μαζί με το σημασιολογικό του γράφημα για την εκτίμηση της περίληψής του. Οι προσεγγίσεις που αναφέρονται παραπάνω σε αυτή την παράγραφο έχουν συνάφεια με την εργασία μας, καθώς βασίζονται σε αναπαραστάσεις σημασιολογικών γραφημάτων του αρχικού κειμένου και προβλέψεις μηχανικής μάθησης για την εκτίμηση μιας περίληψης. Σε αυτή την κατεύθυνση με σκοπό την περαιτέρω βελτίωση, η προσέγγιση που παρουσιάζουμε εξετάζει μια σειρά μοντέλων μηχανικής μάθησης για προβλέψεις τύπου ακολουθία-σε-ακολουθία που περιλαμβάνουν αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή, ενισχυτική μάθηση (*EM*), δίκτυα μετασχηματιστών και προ-εκπαιδευμένα μοντέλα αναπαράστασης φυσικής γλώσσας, όπως περιγράφουμε λεπτομερώς στις επόμενες ενότητες. Τέλος, σημειώνουμε ότι σε αντίθεση με τη μεθοδολογία των προσεγγίσεων [176] και [178] που αναφέρονται παραπάνω, τα μοντέλα βαθιάς μάθησης που προτείνουμε στην παρούσα εργασία λαμβάνουν ως είσοδο την αναπαράσταση του αρχικού κειμένου σε μορφή σημασιολογικού γραφήματος, χωρίς να απαιτείται το υποψήφιο για περίληψη κείμενο στη φάση των προβλέψεων μηχανικής μάθησης για την εκτίμηση μιας περίληψης.

Με αφετηρία τις προσεγγίσεις που αναφέρθηκαν παραπάνω σε αυτή την ενότητα, η παρούσα εργασία επιχειρεί στη συνέχεια να παρουσιάσει ένα νέο πλαίσιο για την αυτόματη *ΠΚ*, το οποίο αξιοποιεί τεχνικές βαθιάς μάθησης και μεθοδολογία σημασιολογικής αναπαράστασης περιεχομένου σε μορφή γραφήματος. Σε αυτή την κατεύθυνση, η έρευνα που διεξάγεται εστιάζει σε μια προσπάθεια διερεύνησης των επιλογών εκείνων που οδηγούν στη βέλτιστη εφαρμογή της προτεινόμενης μεθοδολογίας.

6.3 Σημασιολογικά γραφήματα

Η τυπική μορφή της εισόδου ενός συστήματος αυτόματης *ΠΚ* ενός εγγράφου αποτελείται από κείμενο φυσικής γλώσσας που αντιστοιχεί σε μη δομημένη πληροφορία. Ο μετασχηματισμός του κειμένου σε μορφή δομημένης πληροφορίας μπορεί να αποτελέσει μια θετική προοπτική, η οποία μπορεί να χρησιμοποιηθεί για τον αποτελεσματικό χειρισμό, έλεγχο και αξιοποίηση των δεδομένων εισόδου ενός συστήματος αυτόματης *ΠΚ*. Ένας τέτοιος μετασχηματισμός μπορεί να επιτευχθεί με την αξιοποίηση σημασιολογικών γραφημάτων αναπαράστασης κειμένου. Μια αναπαράσταση κειμένου που βασίζεται σε γράφημα αποτυπώνει το περιεχόμενο ενός κειμένου σε μια δομημένη, συνοπτική και αναγνώσιμη από τις υπολογιστικές μηχανές μορφή. Στην παρούσα εργασία, όπως αναφέρεται με λεπτομέρεια παρακάτω, αξιοποιούμε σημασιολογικά γραφήματα, τα οποία ανακτώνται από μη δομημένο κείμενο εισόδου. Σε αυτή την κατεύθυνση, αρχικά, εξετάζουμε τα εννοιολογικά γραφήματα ως μία γενική προσέγγιση σημασιολογικού γραφήματος, και στη συνέχεια, εστιάζουμε σε ένα συγκεκριμένο τύπο γραφημάτων που ονομάζονται γραφήματα *αναπαράστασης αφηρημένης* έννοιας (*abstract meaning representation*), γνωστά και ως *AMR* γραφήματα σύμφωνα με τη σχετική βιβλιογραφία [168]. Από θεωρητική σκοπιά, η αξιοποίηση της αναπαράστασης σε μορφή γραφήματος εντάσσεται στο πεδίο γνώσης και σημασιολογίας, καθώς και στον τομέα της εξαγωγής συμπερασμάτων της τεχνητής νοημοσύνης [197, 198].

Στη συνέχεια γίνεται μια εισαγωγή στα σημασιολογικά γραφήματα, καθώς το πλαίσιο που προτείνουμε για την αυτόματη *ΠΚ* εστιάζει σε αναπαράσταση δεδομένων κειμένου με χρήση τέτοιων γραφημάτων. Αρχικά περιγράφονται τα εννοιολογικά γραφήματα ως μια γενική μορφή σημασιολογικών αναπαραστάσεων (Ενότητα 6.3.1) και στη συνέχεια τα γραφήματα τύπου *AMR*

(Ενότητα 6.3.2).

6.3.1 Εννοιολογικά γραφήματα

Τα εννοιολογικά γραφήματα, τα οποία βασίζονται στη λογική πρώτης τάξης, διευκολύνουν τη μετάβαση είτε από φυσική γλώσσα σε γράφημα είτε από γράφημα σε φυσική γλώσσα [188, 198]. Σύμφωνα με τον Ορισμό 6.1, ένα εννοιολογικό γράφημα CG αποτελείται από τύπους εννοιών C, σχέσεις R και άτομα I για να αποδώσει τη σημασιολογία της φυσικής γλώσσας. Σε μια αναπαράσταση σε μορφή γραφήματος, οι τύποι εννοιών και τα άτομα αναπαρίστανται με χρήση κόμβων και οι σχέσεις αντιστοιχούν σε ακμές μεταξύ των κόμβων. Οι τύποι εννοιών μπορεί να είναι κλάσεις ή οντότητες, ενώ τα άτομα φέρουν συγκεκριμένες τιμές ή ονόματα που αντιστοιχούν σε συγκεκριμένους τύπους εννοιών, όπως το *"κίτρινο"* μπορεί να είναι άτομο της κλάσης *"χρώμα"* ή το όνομα *"Άγγελος"* μπορεί να αποτελεί άτομο της κλάσης *"πρόσωπο"*.

Ορισμός 6.1 (Εννοιολογικό γράφημα)**.** *Ένα εννοιολογικό γράφημα αναπαριστά μια σημασιολογική αντιστοίχιση από ένα κείμενο T σε ένα γράφημα $CG = \{N, E\}$ που αποτελείται από ένα σύνολο κόμβων $N = C \cup I$ και ένα σύνολο ακμών $E = R$ μεταξύ των κόμβων, όπου με C συμβολίζεται ένα σύνολο που περιλαμβάνει τύπους εννοιών, R είναι ένα σύνολο σχέσεων και I είναι ένα σύνολο ατόμων.*

Ένα εννοιολογικό γράφημα μπορεί να αναπαραστήσει το περιεχόμενο ενός κειμένου χρησιμοποιώντας δύο μορφές συμβολισμού: (i) αναπαράσταση σε μορφή σχηματικού γραφήματος και (ii) αναπαράσταση σε μορφή λογικής έκφρασης. Για να γίνει σαφής ο τρόπος αναπαράστασης ενός εννοιολογικού γραφήματος, το παράδειγμα 6.1 περιλαμβάνει την αναπαράσταση μιας πρότασης με χρήση ενός εννοιολογικού γραφήματος, σύμφωνα με τις δύο προαναφερθείσες μορφές.

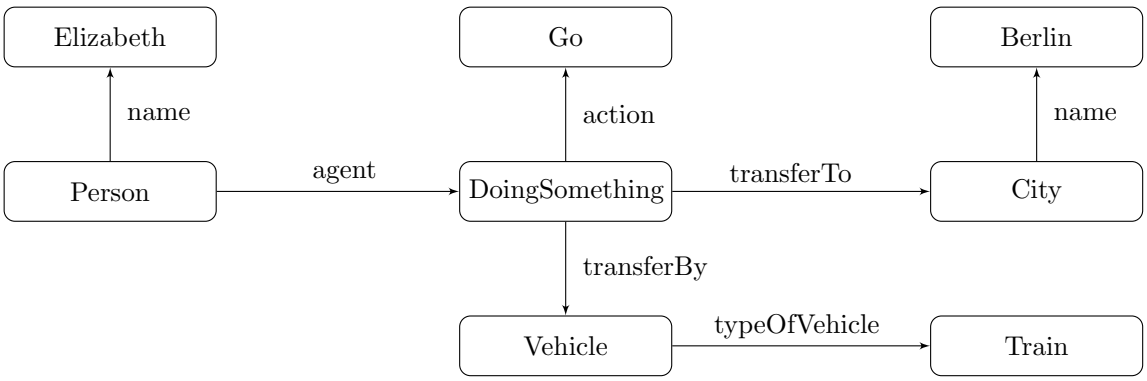
Παράδειγμα 6.1 (Αναπαράσταση εννοιολογικού γραφήματος)**.**
Με δεδομένη την πρόταση "Elizabeth is going to Berlin by train",

(i) Το εννοιολογικό γράφημα σε σχηματική μορφή απεικονίζεται στο Σχήμα 6.1, όπου οι κλάσεις "Person", "DoingSomething" και "City" συνδέονται με συγκεκριμένα άτομα που είναι τα εξής: "Elizabeth", "Go" και "Berlin", αντίστοιχα. Οι κόμβοι του γραφήματος συνδέονται με κατάλληλες σχέσεις μεταξύ τους και το φύλλο "Train" του γραφήματος δενδροειδούς μορφής μπορεί να θεωρηθεί μια υπο-κλάση της κλάσης "Vehicle", χωρίς να συνδέεται με κάποιο άτομο της συγκεκριμένης κλάσης.

(ii) Το εννοιολογικό γράφημα σε μορφή λογικής έκφρασης είναι:

$$\exists p,\ d,\ c,\ e,\ g,\ b,\ v,\ t:\ instance(p,\ Person)\ \wedge\ instance(d,\ DoingSomething)\ \wedge$$
$$instance(c,\ City)\ \wedge\ instance(e,\ Elizabeth)\ \wedge\ instance(g,\ Go)\ \wedge$$
$$instance(b,\ Berlin)\ \wedge\ instance(v,\ Vehicle)\ \wedge\ instance(t,\ Train)\ \wedge$$
$$name(p,\ e)\ \wedge\ agent(p,\ d)\ \wedge\ action(d,\ g)\ \wedge\ transferTo(d,\ c)\ \wedge$$
$$name(c,\ b)\ \wedge\ transferBy(d,\ v)\ \wedge\ typeOfVehicle(v,\ t)$$

Όπως μπορούμε να δούμε παραπάνω, στην αναπαράσταση του εννοιολογικού γραφήματος σε μορφή λογικής έκφρασης χρησιμοποιούνται μεταβλητές που αντιστοιχούν σε συγκεκριμένους τύπους εννοιών ή άτομα (π.χ., p, d, c κλπ.).



Σχήμα 6.1: Το εννοιολογικό γράφημα της πρότασης *"Elizabeth is going to Berlin by train"*.

6.3.2 Αναπαράσταση αφηρημένης έννοιας

Η αναπαράσταση αφηρημένη έννοιας ή *AMR* γράφημα όπως είναι γνωστό, αποτελεί ένα μοντέλο σημασιολογικής αναπαράστασης μιας πρότασης, το οποίο αποτυπώνει κυρίως την ερώτηση *"ποιος κάνει κάτι σε ποιον"* [168]. Αυτή είναι μια ερώτηση που μπορεί να απαντηθεί εύκολα από τον άνθρωπο. Ωστόσο, είναι πολύ δύσκολο για τις μηχανές να αναλύσουν το περιεχόμενο μιας πρότασης και να απαντήσουν σε ένα τέτοιο ερώτημα. Πιο συγκεκριμένα, μια αναπαράσταση τύπου *AMR* αντιστοιχεί σε ένα κατευθυνόμενο άκυκλο γράφημα (*directed acyclic graph - DAG*) που φέρει ρίζα και ετικέτες στις ακμές και στους κόμβους, το οποίο αποτυπώνει το νόημα μιας πρότασης. Το γράφημα αυτό είναι σε μορφή εύκολα αναγνώσιμη από τις μηχανές και, επίσης, σχετικά εύκολα μπορεί να αναγνωσθεί από τον άνθρωπο. Η αναπαράσταση *AMR* είναι προσαρμοσμένη περισσότερο για την αγγλική γλώσσα παρά για άλλες γλώσσες. Επιπλέον, δυο διαφορετικές προτάσεις με το ίδιο νόημα μπορεί να έχουν την ίδια αναπαράσταση (π.χ., οι προτάσεις *"ο Γιάννης θέλει ο φίλος του να τον πιστέψει"* και *"ο Γιάννης έχει την επιθυμία να τον πιστέψει ο φίλος του"* μπορεί να έχουν την ίδια αναπαράσταση *AMR*). Επιπλέον, μία *AMR* αναπαράσταση αποτυπώνει περισσότερο τη λογική και τη σημασιολογία ενός κειμένου και λιγότερο τη σύνταξη και τη γραμματική, παραλείποντας μορφολογικά χαρακτηριστικά κλίσεων, τους χρόνους, τα άρθρα και τα σημεία στίξης.

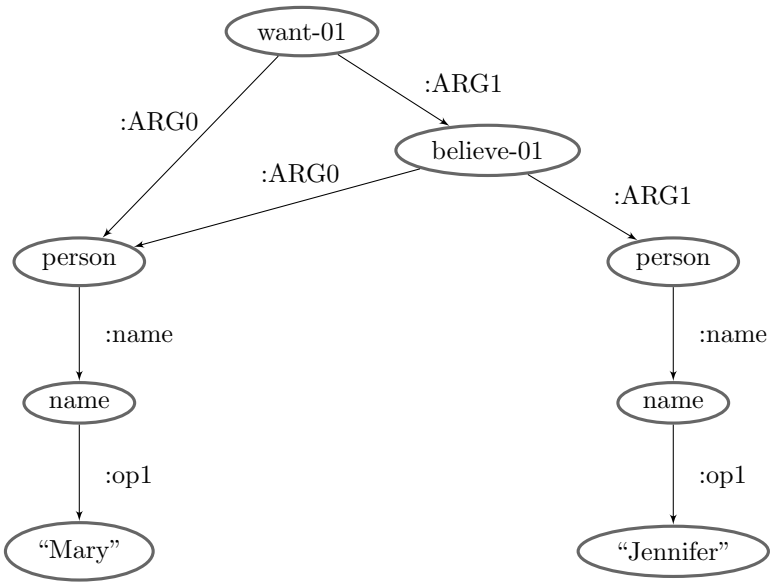
Το παράδειγμα 6.2 παρουσιάζει μια *AMR* αναπαράσταση μιας πρότασης σε τρεις μορφές: (*i*) σε μορφή σχηματικού γραφήματος, (*ii*) σε μορφή λογικής έκφρασης και (*iii*) σε μορφή κειμένου. Η τρίτη μορφή βασίζεται στην εργασία PENMAN [199].

Ένα *AMR* γράφημα περιλαμβάνει κόμβους που αντιστοιχούν σε μεταβλητές (π.χ., *a*, *b*, κλπ.), οι οποίες αναπαριστούν έννοιες, οντότητες, γεγονότα, ιδιότητες και καταστάσεις. Οι κόμβοι συνδέονται μεταξύ τους με σχέσεις σχηματίζοντας ένα γράφημα με ρίζα. Οι έννοιες μπορεί να είναι αγγλικές λέξεις (π.χ., *person*), σύνολα εννοιών των επισημειωμένων σωμάτων κειμένου *PropBank* [169] (π.χ., *want-01*) ή λέξεις-κλειδιά που περιλαμβάνουν τύπους οντοτήτων (π.χ., *itdate-entity, world-region*), ποσότητες (π.χ., *distance-quantity*) και λογικούς τελεστές (π.χ., *and, or*). Το μοντέλο *AMR* χρησιμοποιεί σχέσεις που ακολουθούν τις συμβάσεις της εργασίας *PropBank* [169] (π.χ., *:ARG0, :ARG1, :ARG2* Κλπ.). Επιπλέον, χρησιμοποιούνται περίπου 100 ακόμη σχέσεις για ποσότητες (π.χ., *:quant, :unit* κλπ.), οντότητες ημερομηνίας (π.χ., *:day, :month, :year, :time* κλπ.), λίστες (π.χ., *:op1, :op2, :op3* κλπ.) και γενικές σημασιολογικές σχέσεις (π.χ., *:age, :cause, :compared-to, :consist-of, :domain, :location, :name, :polarity, :topic*, κλπ.). Επίσης, περιλαμβάνεται μια αντιστροφή για κάθε σχέση (π.χ., *:ARG0-0f, :location-of*, κλπ.).

Περισσότερες λεπτομέρειες αναφέρονται στις προδιαγραφές του μοντέλου *AMR* (βλέπε την αρχική έκδοση [200] και την αναθεωρημένη έκδοση [201] των προδιαγραφών *AMR*). Τέλος, μετά από διαδικασία επισημείωσης που έγινε στο πλαίσιο της εργασίας [202], ένα σύνολο προτάσεων κειμένου στην αγγλική γλώσσα είναι διαθέσιμο σε μορφή *AMR* γραφημάτων, ως παραδείγματα χρήσης.

Παράδειγμα 6.2 (Αναπαράσταση *AMR*)**.**
Με δεδομένη την πρόταση "Mary wants Jennifer to believe her",

(i) Το AMR γράφημα απεικονίζεται στο σχήμα 6.2.



Σχήμα 6.2: Το *AMR* γράφημα της πρότασης *"Mary wants Jennifer to believe her".*

(ii) Η AMR αναπαράσταση με χρήση λογικής έκφρασης:

$$\exists w,\ b,\ p1,\ p2,\ n1,\ n2 :\ instance(w,\ want\text{-}01)\ \wedge\ instance(p1,\ person)\ \wedge\ instance(n1,\ name)\ \wedge\ instance(b,\ believe\text{-}01)\ \wedge\ instance(p2,\ person)\ \wedge\ instance(n2,\ name)\ \wedge\ ARG0(w,\ p1)\ \wedge\ name(p1,\ n1)\ \wedge op1(n1,\ \text{“}Mary\text{”})\ \wedge\ ARG1(w,\ b)\ \wedge\ ARG0(b,\ p1)\ \wedge ARG1(b,\ p2)\ \wedge\ name(p2,\ n2)\ \wedge\ op1(n2,\ \text{“}Jennifer\text{”})$$

(iii) Η AMR αναπαράσταση με χρήση κειμένου:

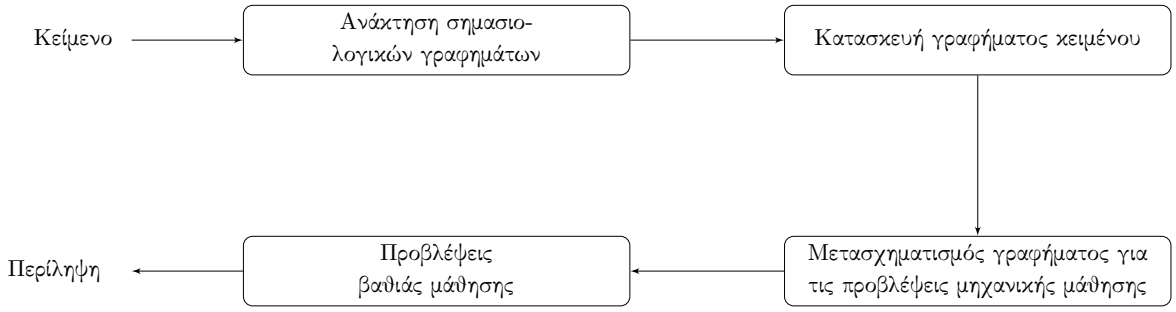
$$\begin{aligned}&(a\ /\ want\text{-}01\\ &\quad :ARG0\ (b1\ /\ person\\ &\qquad :name\ (n1\ /\ name\ \ :op1\ \text{”}Mary\text{”}))\\ &\quad :ARG1\ (c\ /\ believe\text{-}01\\ &\qquad :ARG0\ (b2\ /\ person\\ &\qquad\qquad :name\ (n2\ /\ name\ \ :op1\ \text{”}Jennifer\text{”}))\\ &\qquad :ARG1\ b1))\end{aligned}$$

Όπως θα δούμε και στη συνέχεια, η εργασία αυτή στο πειραματικό της μέρος (Κεφάλαιο 7) αξιοποιεί το σημασιολογικό μοντέλο των *AMR* γραφημάτων για την αναπαράσταση κειμένου. Να σημειωθεί ότι η χρήση του μοντέλου αυτού διερευνάται διεξοδικά σε ερευνητικές περιοχές της επεξεργασίας φυσικής γλώσσας, όπως ειδικότερα γίνεται και στον τομέα της αυτόματης *ΠΚ* που εξετάζουμε (βλέπε την περιγραφή των σχετικών εργασιών στην Ενότητα 6.2), λόγω των θετικών προοπτικών που δημιουργεί.

6.4 Εισαγωγή στην αρχιτεκτονική της προτεινόμενης προσέγγισης

Η αρχιτεκτονική του προτεινόμενου πλαισίου απεικονίζεται στο Σχήμα 6.3. Όπως παρατηρούμε στο σχήμα αυτό, η είσοδος του συστήματος αποτελείται από ένα υποψήφιο για περίληψη έγγραφο κειμένου (αρχικό κείμενο), ενώ η έξοδος είναι η εκτιμώμενη περίληψη του αρχικού κειμένου. Τα κύρια μέρη της αρχιτεκτονικής του προτεινόμενου συστήματος είναι τέσσερα ξεκινώντας από τη βαθμίδα της ανάκτησης σημασιολογικών γραφημάτων, σκοπός της οποίας είναι η ανάκτηση ενός σημασιολογικού γραφήματος για κάθε πρόταση του αρχικού κειμένου. Το σύνολο των σημασιολογικών γραφημάτων που ανακτώνται χρησιμοποιείται στη δεύτερη βαθμίδα που αντιστοιχεί στην κατασκευή του σημασιολογικού γραφήματος του αρχικού κειμένου. Στο τρίτο βήμα έχουμε τον μετασχηματισμό του σημασιολογικού γραφήματος σε μια κατάλληλη μορφή που χρησιμοποιείται ως είσοδος σε ένα μοντέλο μηχανικής μάθησης. Το τελευταίο βήμα περιλαμβάνει τις προβλέψεις μηχανικής μάθησης, σύμφωνα με τις οποίες, ένα μοντέλο βαθιάς μάθησης, έχοντας εκπαιδευτεί σε ένα σώμα παραδειγμάτων χρήσης που αποτελείται από ζεύγη τύπου σημασιολογικό γράφημα-περίληψη, εκτιμά μια περίληψη σε μορφή κειμένου για ένα νέο υποψήφιο για περίληψη κείμενο.

Στις επόμενες ενότητες παρουσιάζουμε τη διαδικασία ανάκτησης σημασιολογικών γραφημάτων (Ενότητα 6.5), καθορίζουμε τις διαδικασίες κατασκευής σημασιολογικού γραφήματος που αναπαριστά ένα αρχικό κείμενο (Ενότητα 6.6) και τη μεθοδολογία μετασχηματισμού ενός γραφήματος σε κατάλληλη μορφή για τις προβλέψεις μηχανικής μάθησης (Ενότητα 6.7). Τέλος, η φάση των προβλέψεων βαθιάς μάθησης παρουσιάζεται στην ενότητα 6.8.



Σχήμα 6.3: Διάγραμμα ροής του προτεινόμενου πλαισίου για την αυτόματη περίληψη κειμένου με χρήση βαθιάς μάθησης και σημασιολογικών γραφημάτων

6.5 Ανάκτηση σημασιολογικών γραφημάτων

Σύμφωνα με τη διαδικασία ανάκτησης σημασιολογικών γραφημάτων που ακολουθούμε, με δεδομένη μια πρόταση κειμένου S, η οποία αποτελείται από μια ακολουθία λέξεων $S = (w_1,\ w_2,\ ...\ w_k)$, η ανάκτηση ενός σημασιολογικού γραφήματος αποσκοπεί στην εξαγωγή μιας σημασιολογικής αναπαράστασης της πρότασης αυτής σε μορφή γραφήματος $G = (N,\ E)$, όπου με N υποδηλώνεται ένα σύνολο εννοιών που αντιστοιχούν στους κόμβους του γραφήματος και με E ένα σύνολο ακμών μεταξύ των κόμβων. Ένα τέτοιο γράφημα μπορεί να αναπαρασταθεί με ένα σύνολο τριάδων της μορφής $G = \{(u_1,\ u_2,\ e_{12}),\ ...,\ (u_i,\ u_j,\ e_{ij}), ...\}$, όπου $u_i \in N$ και $u_j \in N$ αντιστοιχούν σε κόμβους του γραφήματος και με $e_{ij} \in E$ συμβολίζεται μια ακμή μεταξύ δύο κόμβων, στην προκειμένη περίπτωση πρόκειται για την ακμή μεταξύ των κόμβων u_i και u_j. Ουσιαστικά, η διαδικασία ανάκτησης ενός σημασιολογικού γραφήματος βασίζεται σε μια συνάρτηση f_p που εκτελεί μια αντιστοίχιση μεταξύ του περιεχομένου ενός κειμένου T και του αντίστοιχου σημασιολογικού γραφήματος ($f_p : T \rightarrow G$).

Για να δείξουμε τη διαδικασία ανάκτησης σημασιολογικών γραφημάτων, υποθέτουμε ότι έχουμε ένα κείμενο T που περιλαμβάνει μια ακολουθία προτάσεων $T = (S_1,\ S_2,\ ...,\ S_n)$, με την κάθε πρόταση να αποτελείται, με τη σειρά της, από μια ακολουθία λέξεων $S_i = (w_1,\ w_2,\ ...,\ w_k)$. Η διαδικασία ανάκτησης των σημασιολογικών γραφημάτων στοχεύει στην εξαγωγή ενός γραφήματος για κάθε πρόταση του κειμένου. Επομένως, με δεδομένη μια ακολουθία προτάσεων ενός κειμένου, η διαδικασία αυτή επιστρέφει μια ακολουθία γραφημάτων $\mathsf{G_T} = (G_1,\ G_2,\ ...,\ G_n)$. Η διαδικασία αυτή που αποτελεί την πρώτη βαθμίδα του προτεινόμενου πλαισίου (Σχήμα 6.3) ανακτά μια ακολουθία σημασιολογικών γραφημάτων για κάθε κείμενο που δίνεται ως είσοδος στο σύστημα. Στη συνέχεια, η ακολουθία αυτή των γραφημάτων δίνεται ως είσοδος στην επόμενη βαθμίδα του συστήματος που κατασκευάζει ένα συνολικό σημασιολογικό γράφημα για κάθε κείμενο, όπως εξηγούμε με λεπτομέρεια παρακάτω.

Να σημειώσουμε ότι για τη σημασιολογική αναπαράσταση ενός κειμένου σε μορφή σημασιολογικού γραφήματος, υιοθετούμε το μοντέλο γραφημάτων *AMR* (Ενότητα 6.3.2) για το οποίο έχουν αναπτυχθεί διάφοροι αναλυτές (*parsers*) [186, 203, 173, 204] που μπορούν να ενταχθούν στο προτεινόμενο πλαίσιο, καθώς είναι διαθέσιμοι για χρήση. Επομένως, όπως θα δούμε στο επόμενο κεφάλαιο, στο πλαίσιο της πειραματικής διαδικασίας που διεξάγεται για την αξιολόγηση του προτεινόμενου πλαισίου, αξιοποιούμε έναν ήδη ανεπτυγμένο αναλυτή σημασιολογικών γραφημάτων τύπου *AMR*.

6.6 Κατασκευή σημασιολογικού γραφήματος

Η διαδικασία της ανάκτησης των σημασιολογικών γραφημάτων, η οποία αναφέρθηκε παραπάνω, στοχεύει στην ανάκτηση ενός γραφήματος για κάθε πρόταση ή περίοδο ενός κειμένου. Θεωρώντας ότι έχουν ανακτηθεί και είναι διαθέσιμα τα επιμέρους γραφήματα ενός κειμένου, τα οποία στη συνέχεια θα ονομάζονται υπο-γραφήματα, εξετάζουμε δύο στρατηγικές για τη σημασιολογική αναπαράσταση του κειμένου. Σύμφωνα με την πρώτη στρατηγική, θεωρούμε ότι η σημασιολογική αναπαράσταση του κειμένου προκύπτει από την ακολουθία των υπο-γραφημάτων, ενώ σύμφωνα με τη δεύτερη στρατηγική, το σημασιολογικό γράφημα ενός κειμένου κατασκευάζεται συνδυάζοντας

τα επιμέρους υπο-γραφήματα, δημιουργώντας ένα γράφημα. Ακολουθεί η περιγραφή των δύο αυτών μεθόδων.

6.6.1 Σημασιολογικό γράφημα ως ακολουθία υπο-γραφημάτων

Σύμφωνα με τη στρατηγική δημιουργίας της σημασιολογικής αναπαράστασης ενός κειμένου σε μορφή γραφήματος ως ακολουθία υπο-γραφημάτων, το γράφημα ενός κειμένου $\mathsf{G_T}$ αποτελείται από μια πλειάδα που περιέχει την ακολουθία των υπο-γραφημάτων $G_i \in G_T$ ενός κειμένου T, έτσι ώστε $\mathsf{G_T} = (G_1,\ G_2,\ ...,\ G_n)$, όπου G_i με $i \in (1,\ 2,\ ...\ n)$ αναπαριστά το σημασιολογικό υπο-γράφημα που αντιστοιχεί στην πρόταση S_i του κείμενο T, το οποίο αποτελείται από μια ακολουθία προτάσεων $T = (S_1,\ S_2,\ ...,\ S_n)$. Σε αυτή την αναπαράσταση, τα υπο-γραφήματα παραμένουν ανεξάρτητα μεταξύ τους (δηλ., η αναπαράσταση του κειμένου αποτελείται από μεμονωμένα γραφήματα χωρίς να υποδηλώνεται κάποια σύνδεση ή ακμή μεταξύ των κόμβων τους).

Συνεπώς, το σύνολο των υπο-γραφήματων σε διάταξη ακολουθίας που αντιστοιχεί στην ακολουθία των προτάσεων ενός κειμένου υποδηλώνει τη σημασιολογική αναπαράσταση του κειμένου αυτού. Σύμφωνα με τη μεθοδολογία, το κείμενο που αναπαρίσταται ως ακολουθία υπο-γραφημάτων είναι ένα υποψήφιο για περίληψη κείμενο.

6.6.2 Σημασιολογικό γράφημα ως συνδυασμός υπο-γραφημάτων

Σύμφωνα με τη δεύτερη στρατηγική αναπαράστασης ενός κειμένου, τα υπο-γραφήματα που αντιστοιχούν στις επιμέρους προτάσεις ή περιόδους ενός κειμένου συνδυάζονται προκειμένου να κατασκευαστεί ένα ενιαίο σημασιολογικό γράφημα $\mathsf{G_C}$ του συνολικού κειμένου. Το γράφημα αυτό παραμένει ένα άκυκλο σημασιολογικό γράφημα, το οποίο φέρει ρίζα. Η μεθοδολογία αυτή μπορεί να οδηγήσει σε ένα μικρότερης διάσταση γράφημα, σε σύγκριση με τη διάσταση του συνόλου των υπο-γραφημάτων, καθώς μειώνεται ή εξαλείφεται ο πλεονασμός στο τελικό γράφημα. Θα μπορούσαμε να εκφράσουμε το σημασιολογικό γράφημα $\mathsf{G_C}$ που προκύπτει ως ένωση των υπο-γραφημάτων ως εξής $\mathsf{G_C} = G_1 \cup G_2 \cup ... \cup G_n$.

Πιο συγκεκριμένα, για να συνδυάσουμε δύο ή περισσότερα σημασιολογικά υπο-γραφήματα, ακολουθούμε μια παρόμοια προσέγγιση με την εργασία [183] και, επιπροσθέτως, αξιοποιούμε το πλαίσιο περιγραφής πόρων (*resource description framework - RDF*) [205] προκειμένου να σχεδιάσουμε έναν Αλγόριθμο για τον σκοπό αυτό. Ο Αλγόριθμος 6.1, τον οποίο προτείνουμε, παρέχει έναν συστηματικό τρόπο για τον συνδυασμό των επιμέρους υπο-γραφημάτων σε ένα γράφημα, το οποίο αναπαριστά σημασιολογικά ένα κείμενο που περιέχει περισσότερες από μία προτάσεις.

Ο Αλγόριθμος 6.1 λαμβάνει ως είσοδο μια πλειάδα των επιμέρους σημασιολογικών υπο-γραφημάτων ενός κειμένου ($\mathsf{G_T} = (G_1,\ G_2,\ ...,\ G_n)$, όπου G_i είναι ένα υπο-γράφημα που αντιστοιχεί στην πρόταση S_i ενός κειμένου). Η διαδικασία ξεκινά με την αρχικοποίηση του συνόλου RDF_C σε ένα κενό σύνολο (γραμμή 1). Το σύνολο RDF_C περιέχει τον συνδυασμό των τριάδων τύπου RDF του συνόλου των υπο-γραφημάτων. Στη συνέχεια, ο αλγόριθμος, στον βρόχο των

Αλγόριθμος 6.1 Κατασκευή ενός γραφήματος με συνδυασμό υπο-γραφημάτων για τη σημασιολογική αναπαράσταση ενός κειμένου με περισσότερες από μια προτάσεις.

Require: $\mathsf{G_T}$

1: $RDF_C \leftarrow \{\}$
2: **for all** $G_i \in \mathsf{G_T}$ **do**
3: $\quad G_i \leftarrow$ προσθήκη ρίζας στο γράφημα G_i
4: $\quad RDF_i \leftarrow$ τριάδες τύπου RDF του γραφήματος G_i
5: $\quad RDF_C \leftarrow RDF_C \cup RDF_i$
6: **end for**
7: $\mathsf{G_C} \leftarrow$ Κατασκευή του γραφήματος με χρήση των τριάδων RDF του συνόλου RDF_C
8: **return** $\mathsf{G_C}$

γραμμών $2-6$, διατρέχει τα επιμέρους υπο-γραφήματα ($G_i \in \mathsf{G_T}$) προσθέτοντας έναν κόμβο ρίζας σε κάθε υπο-γράφημα (γραμμή 3) και, επίσης, ανακτώντας τις RDF τριάδες για κάθε υπο-γράφημα G_i οι οποίες ανατίθενται στο σύνολο RDF_i (γραμμή 4). Το σύνολο RDF_i περιλαμβάνει το σύνολο των RDF τριάδων του τρέχοντος υπο-γραφήματος. Το σύνολο RDF_C ανανεώνεται μετά την ένωσή του με τα επιμέρους σύνολα RDF_i σε κάθε επανάληψη του βρόχου (γραμμή 5). Ο αλγόριθμος κατασκευάζει το επιθυμητό σημασιολογικό γράφημα G_C του κειμένου χρησιμοποιώντας τις RDF τριάδες του συνόλου RDF_C (γραμμή 7). Τέλος, ο αλγόριθμος επιστρέφει το σημασιολογικό γράφημα $\mathsf{G_C}$ (π.χ., σε μορφή AMR), το οποίο αποτελείται από τον συνδυασμό των υπο-γραφημάτων ενός κειμένου.

Για να δικαιολογήσουμε την προσθήκη του κόμβου ρίζας σε κάθε υπο-γράφημα (γραμμή 3, Αλγόριθμος 6.1), θα πρέπει να διευκρινιστεί ότι ένα κείμενο μπορεί να περιλαμβάνει προτάσεις οι οποίες αντιστοιχούν σε υπο-γραφήματα τα οποία δεν περιλαμβάνουν κάποιο κοινό κόμβο μεταξύ τους (δηλ., υπο-γραφήματα τα οποία δεν μπορούν να συνενωθούν κατά τη διαδικασία συνδυασμού των υπο-γραφημάτων, λόγω έλλειψης επικαλυπτόμενης πληροφορίας μεταξύ των επιμέρους προτάσεων). Για τον λόγο αυτό, προσθέτουμε έναν κόμβο ως ρίζα σε κάθε υπο-γράφημα, ο οποίος, επίσης, αποτελεί τον κόμβο ρίζας του συνολικού σημασιολογικού γραφήματος που δημιουργείται κατά τη διαδικασία συνένωσης των υπο-γραφημάτων. Η διαδικασία αυτή γίνεται περισσότερο σαφής μέσω του Παραδείγματος 6.3 που παρουσιάζεται παρακάτω.

Πιο συγκεκριμένα, το Παράδειγμα 6.3, το οποίο ακολουθεί, επεξηγεί περαιτέρω τη μεθοδολογία κατασκευής ενός σημασιολογικού γραφήματος ως συνδυασμό υπο-γραφημάτων. Σε αυτό το παράδειγμα, θεωρούμε ότι έχουμε ένα κείμενο με δύο προτάσεις, από το οποίο έχουν ανακτηθεί τα σημασιολογικά γραφήματα των επιμέρους προτάσεων σε μορφή AMR που αντιστοιχούν στο περιεχόμενο του κειμένου.

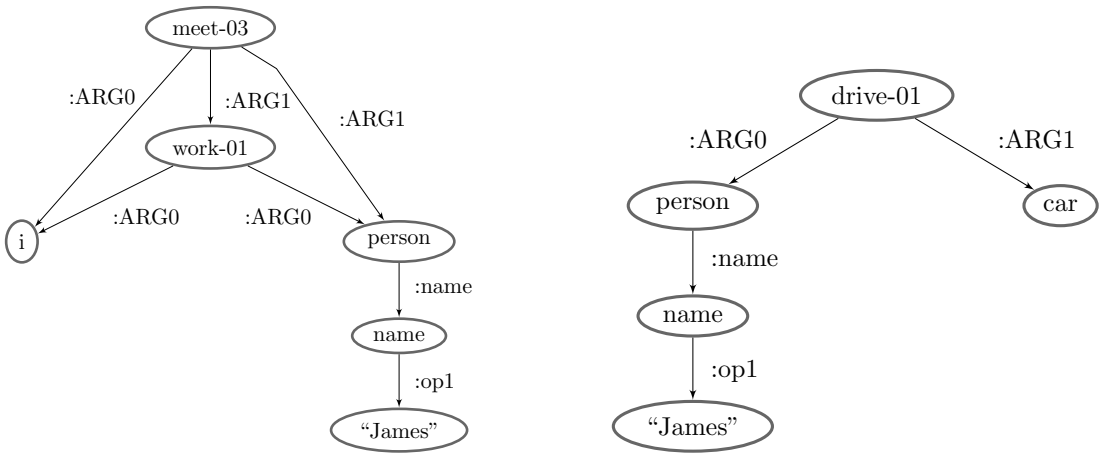
Παράδειγμα 6.3 (Κατασκευή σημασιολογικού γραφήματος ως συνδυασμός υπο-γραφημάτων)**.**

Με δεδομένες τις προτάσεις S_1 και S_2 ενός κειμένου:

S_1*: I met James, who was going to work.*

S_2*: James was driving a car.*

Τα AMR υπο-γραφήματα των προτάσεων S_1 και S_2 αποτυπώνονται στα Σχήματα 6.4α′ και 6.4β′, αντίστοιχα.



(α′) (β′)

Σχήμα 6.4: (6.4α′) Το *AMR* γράφημα της πρότασης *“I met James, who was going to work”* και (6.4β′) το *AMR* γράφημα της πρότασης *“James was driving his car”*.

Σύμφωνα με τον Αλγόριθμο 6.1, οι RDF τριάδες της πρώτης πρότασης (RDF_{S_1}) και της δεύτερης πρότασης (RDF_{S_2}) είναι όπως ακολούθως.

$$RDF_{S_1} = \{(root,\ :ROOT,\ meet\text{-}03),\ (meet\text{-}03,\ :ARG0,\ i),\ (meet\text{-}03,\ :ARG1,\ work\text{-}01),\ (meet\text{-}03,\ :ARG1,\ person),\ (work\text{-}01,\ :ARG0,\ i),\ (work\text{-}01,\ :ARG0,\ person),\ (person,\ :name,\ name),\ (name,\ :op1,\ \text{“}James\text{”})\}$$

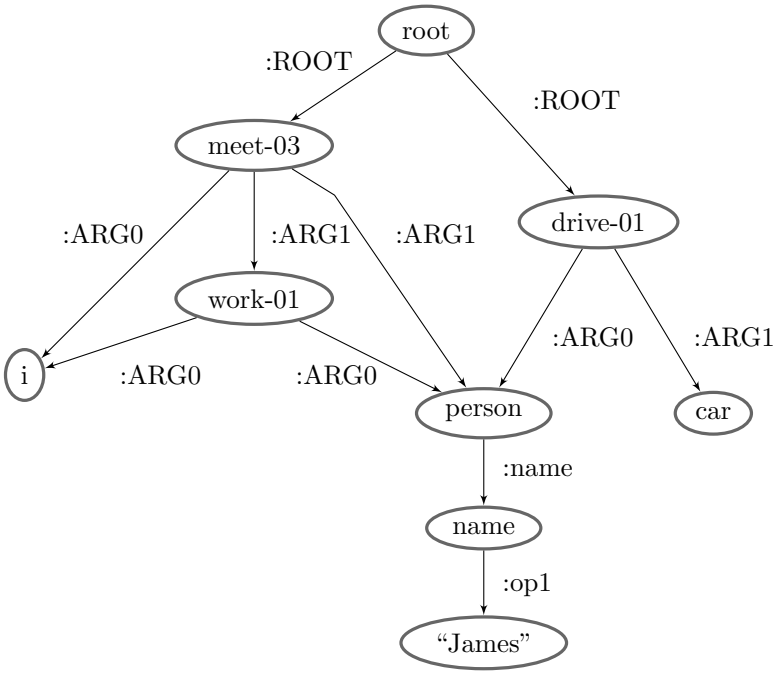
$$RDF_{S_2} = \{(root,\ :ROOT,\ drive\text{-}01),\ (drive\text{-}01,\ :ARG0,\ person),\ (drive\text{-}01,\ :ARG1,\ car),\ (person,\ :name,\ name),\ (name,\ :op1,\ \text{“}James\text{”})\}$$

Το σύνολο με τις RDF τριάδες του κειμένου (RDF_C) προκύπτει από την ένωση των δύο επιμέρους συνόλων, RDF_{S_1} και RDF_{S_2}, με τις RDF τριάδες της κάθε πρότασης, όπως ακολούθως.

$$\begin{aligned} RDF_C = & RDF_{S_1} \cup RDF_{S_2} = \{(root,\ :ROOT,\ meet\text{-}03),\ (root,\ :ROOT,\ drive\text{-}01),\\ & (meet\text{-}03,\ :ARG0,\ i),\ (meet\text{-}03,\ :ARG1,\ work\text{-}01),\ (meet\text{-}03,\ :ARG1,\ person),\\ & (work\text{-}01,\ :ARG0,\ i),\ (work\text{-}01,\ :ARG0,\ person),\ (person,\ :name,\ name),\\ & (name,\ :op1,\ \text{“}James\text{”}),\ (drive\text{-}01,\ :ARG0,\ person),\ (drive\text{-}01,\ :ARG1,\ car)\} \end{aligned}$$

Σύμφωνα με τις RDF τριάδες που προκύπτουν από τον συνδυασμό των επιμέρους προτάσεων (RDF_C), το σημασιολογικό γράφημα του κειμένου των δύο παραπάνω προτάσεων απεικονίζεται στο Σχήμα 6.5.

Τα σημασιολογικά γραφήματα, με τη σειρά τους, θα πρέπει να μετασχηματιστούν σε κατάλληλη μορφή για να δοθούν ως είσοδος σε ένα μοντέλο μηχανικής μάθησης. Στην επόμενη ενότητα, παρουσιάζουμε τις μεθόδους μετασχηματισμού των σημασιολογικών γραφημάτων σε κατάλληλη



Σχήμα 6.5: Το γράφημα που συνδυάζει τα δύο υπό-γραφήματα των προτάσεων *"I met James, who was going to work"* και *"James was driving a car"*.

μορφή για τη χρήση τους στη φάση της μηχανικής μάθησης, όπου ένα μοντέλο βαθιάς μάθησης εκπαιδεύεται με χρήση ζευγών σημασιολογικού γραφήματος-περίληψης, για να είναι σε θέσει να εκτιμήσει την περίληψη ενός νέου κειμένου.

6.7 Μετασχηματισμοί γραφημάτων για τη μηχανική μάθηση

Τα μοντέλα μηχανικής μάθησης που αξιοποιούνται στο πλαίσιο της εργασίας αυτής (Ενότητα 6.8) βασίζονται κυρίως σε αρχιτεκτονικές βαθιάς μάθησης που εκτελούν προβλέψεις τύπου ακολουθία-σε-ακολουθία. Αυτές οι αρχιτεκτονικές λαμβάνουν ως είσοδο μια ακολουθία στοιχείων ή λεκτικών μονάδων που αντιπροσωπεύουν ένα σημασιολογικό γράφημα ενός κειμένου εισόδου για την εκτίμηση μιας ακολουθίας λεκτικών μονάδων που αποτελούν την περίληψη. Επομένως, με δεδομένο το σημασιολογικό γράφημα ενός κειμένου (Ενότητα 6.6), χρειάζεται να εφαρμοστεί κάποιος μετασχηματισμός που θα μετατρέψει το γράφημα αναπαράστασης ενός κειμένου εισόδου σε μια ακολουθία στοιχείων που περιγράφουν το γράφημα. Για τον σκοπό αυτό, η χρήση των γραφημάτων τύπου *AMR* σε μορφή κειμένου διευκολύνει τη μετατροπή σε μια ακολουθία λεκτικών μονάδων που χρησιμοποιείται για την αναπαράσταση τέτοιων γραφημάτων. Σε αυτή την κατεύθυνση, εξετάζουμε κάποιες μεθόδους μετασχηματισμού ενός σημασιολογικού γραφήματος σε μια ακολουθία λεκτικών μονάδων. Οι μετασχηματισμοί αυτοί ακολουθούν.

Αρχικό σημασιολογικό γράφημα (ΑΡΣΓ): Πρόκειται για μια εκδοχή ενός γραφήματος *AMR* σε μορφή κειμένου μιας γραμμής. Για παράδειγμα, η μορφή κειμένου του γραφήματος *AMR* του Παραδείγματος 6.2 μετασχηματίζεται σε μια ακολουθία λεκτικών μονάδων

ως εξής:

(a / $want$-01 : $ARG0$ ($b1$ / $person$: $name$ ($n1$ / $name$: $op1$ "$Mary$")) : $ARG1$ (c / $believe$-01 : $ARG0$ ($b2$ / $person$: $name$ ($n2$ / $name$: $op1$ "$Jennifer$")) : $ARG1$ $b1$))

Αρχικό σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών (ΑΡΣΓΧΕ): Σε αυτή την αναπαράσταση, αφαιρούνται οι αριθμοί διάκρισης των εννοιών ενός γραφήματος τύπου *AMR* για λόγους μείωσης του λεξιλογίου (δηλ., μείωση του αριθμού των διακριτών λεκτικών μονάδων) αναπαράστασης των σημασιολογικών γραφημάτων (π.χ., το αναγνωριστικό "*want*-01" που αντιστοιχεί στην έννοια *"01"* της λέξης *"want"* αλλάξει σε *"want"*). Σύμφωνα με αυτόν τον μετασχηματισμό, το γράφημα *AMR* του Παραδείγματος 6.2 αναγράφεται σε μια γραμμή ως εξής:

(a / $want$: $ARG0$ ($b1$ / $person$: $name$ ($n1$ / $name$: $op1$ "$Mary$")) : $ARG1$ (c / $believe$: $ARG0$ ($b2$ / $person$: $name$ ($n2$ / $name$: $op1$ "$Jennifer$")) : $ARG1$ $b1$))

Απλοποιημένο σημασιολογικό γράφημα (ΑΠΣΓ): Σε αυτή την έκδοση του γραφήματος σε μορφή κειμένου μίας γραμμής, αφαιρούνται οι μεταβλητές και το σύμβολο *"/"* της *AMR* αναπαράστασης (π.χ., *"b1 / person"* μετατρέπεται σε *"person"*) ή οι μεταβλητές αντικαθίστανται με τις οντότητες που αντιστοιχούν σε αυτές (π.χ., *"b1"* αντικαθίσταται από *"person"*). Επίσης, οι παρενθέσεις στην αρχή και στο τέλος της αναπαράστασης σε μορφή κειμένου αφαιρούνται. Αυτός ο μετασχηματισμός μειώνει το μήκος τη αναπαράστασης γραφήματος. Για παράδειγμα, το γράφημα *AMR* του Παραδείγματος 6.2 μετασχηματίζεται ως εξής:

$want$-01 : $ARG0$ ($person$: $name$ ($name$: $op1$ "$Mary$")) : $ARG1$ ($believe$-01 : $ARG0$ ($person$: $name$ ($name$: $op1$ "$Jennifer$")) : $ARG1$ $person$

Απλοποιημένο σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών (ΑΠΣΓΧΕ): Συμπληρωματικά με την προηγούμενη αναπαράσταση, αφαιρούνται οι αριθμοί που προσδιορίζουν την έννοια των λέξεων ενός γραφήματος *AMR*. Η αναπαράσταση *AMR* του παραδείγματος 6.2 μετασχηματίζεται ως εξής.

$want$: $ARG0$ ($person$: $name$ ($name$: $op1$ "$Mary$")) : $ARG1$ ($believe$: $ARG0$ ($person$: $name$ ($name$: $op1$ "$Jennifer$")) : $ARG1$ $person$

Οι απλοποιημένες μορφές των παραπάνω σημασιολογικών αναπαραστάσεων παρέχουν μια συνοπτική, μικρότερης διάστασης αναπαράσταση του σημασιολογικού γραφήματος, που οδηγεί στη μείωση του υπολογιστικού φορτίου της συνολικής προσέγγισης. Η διαγραφή των αριθμών των αναγνωριστικών αποσαφήνισης εννοιών των αναπαραστάσεων τύπου *AMR* οδηγεί σε μείωση του μεγέθους του λεξιλογίου, που, με τη σειρά του, μειώνει επίσης τον χώρο αναζήτησης των

προβλέψεων μηχανικής μάθησης. Η επίδραση κάθε αναπαράστασης τύπου *AMR*, σε μορφή κειμένου μιας γραμμής, διερευνάται στην πειραματική διαδικασία (Κεφάλαιο 7) που διεξάγεται στο πλαίσιο της παρούσας διατριβής.

6.7.1 Συνδυασμός μεθοδολογίας κατασκευής και μετασχηματισμού γραφήματος

Στην περίπτωση που έχει εφαρμοστεί η μέθοδος κατασκευής γραφήματος ως ακολουθία υπο-γραφημάτων (Ενότητα 6.6.1) και στην περίπτωση κειμένου με περισσότερες από μία προτάσεις, του οποίου η σημασιολογική αναπαράσταση περιλαμβάνει περισσότερα από ένα υπο-γραφήματα, η αναπαράσταση του σημασιολογικού γραφήματος σε μια γραμμή αποτελείται από μια ακολουθία υπο-γραφημάτων. Τα υπο-γραφήματα αυτά χωρίζονται μεταξύ τους από μια προκαθορισμένη λεκτική μονάδα (π.χ., *[υπο-γράφημα σε μια γραμμή της πρότασης 1] [EOG] [υπο-γράφημα σε μια γραμμή της πρότασης 2][EOG]...[EOG] [υπο-γράφημα σε μια γραμμή της πρότασης n]*, όπου το διακριτικό *[EOG]* δηλώνει το τέλος ενός υπο-γραφήματος).

Στην περίπτωση σημασιολογικού γραφήματος που έχει προκύψει ως συνδυασμός των επιμέρους υπο-γραφημάτων (Ενότητα 6.6.2), ο κόμβος που έχει προστεθεί ως ρίζα του γραφήματος χρησιμοποιείται ως κόμβος έναρξης της αναζήτησης για την εξαγωγή όλων των μονοπατιών του γραφήματος (π.χ., χρησιμοποιώντας τον αλγόριθμο *DFS* για αναζήτηση κατά βάθος). Τα επιμέρους μονοπάτια από τη ρίζα προς τα φύλλα που ανακτώνται διατάσσονται σε μια ακολουθία μονοπατιών σε μορφή κειμένου, τα οποία αποτελούν τη σημασιολογική αναπαράσταση του κειμένου με χρήση λεκτικών μονάδων σε μια γραμμή. Στη συνέχεια, ο κόμβος ρίζα αφαιρείται από την αναπαράσταση σε μορφή κειμένου μιας γραμμής του σημασιολογικού γραφήματος. Ο κόμβος αυτός αφαιρείται, καθώς ως πρόσθετος κόμβος (δηλ., κόμβος που έχει προστεθεί ως ρίζα για να συνενώσει μεταξύ τους ανεξάρτητα γραφήματα) δεν φέρει κάποια χρήσιμη πληροφορία για το περιεχόμενο του κειμένου. Επίσης, ο κόμβος αυτός αυξάνει τη διάσταση της αναπαράστασης και επιβαρύνει τα μοντέλα μηχανικής μάθησης με επιπρόσθετο υπολογιστικό κόστος.

Η παρούσα εργασία προτείνει δύο τεχνικές κατασκευής γραφήματος και τέσσερις μεθόδους μετασχηματισμού γραφήματος που η εφαρμογή τους μπορεί να οδηγήσει σε οχτώ διαφορετικά σχήματα δεδομένων. Η αποτελεσματικότητα των σχημάτων δεδομένων, τα οποία προκύπτουν με συνδυασμό των προαναφερόμενων τεχνικών, διερευνάται στην πειραματική διαδικασία (Κεφάλαιο 7). Πιο συγκεκριμένα, τα επιμέρους σχήματα δεδομένων αναφέρονται στην Ενότητα 7.2 και ειδικότερα συνοψίζονται στον Πίνακα 7.1, σύμφωνα με τα σύνολα δεδομένων που χρησιμοποιούνται στην πειραματική διαδικασία, όπως θα δούμε στο επόμενο κεφάλαιο.

6.8 Προβλέψεις βαθιάς μάθησης

Με δεδομένο ένα σημασιολογικό γράφημα σε κατάλληλη μορφή, όπως έχει περιγραφεί στην Ενότητα 6.7, η φάση της μηχανικής μάθησης του προτεινόμενου πλαισίου εκτελείται σε δύο βήματα. Στο πρώτο βήμα, οι λεκτικές μονάδες αναπαράστασης των γραφημάτων του συνόλου εκπαίδευσης αναπαρίστανται σε έναν συνεχή διανυσματικό χώρο, χρησιμοποιώντας είτε ενσωματώσεις λέξεων ανεξάρτητες από τα συμφραζόμενα (π.χ., τα μοντέλα *word2vec, glove* κλπ.) [85] είτε διανυσματικές

αναπαραστάσεις λέξεων που εξαρτώνται από τα συμφραζόμενα (π.χ. τα μοντέλα *ELMO, BERT* κλπ.) [86]. Στη συνέχεια, οι διανυσματικές αναπαραστάσεις των λέξεων παρέχονται ως είσοδος σε ένα μοντέλο βαθιάς μάθησης. Στο δεύτερο βήμα, ένα μοντέλο βαθιάς μάθησης (Ενότητα 6.8.2) εκπαιδεύεται για την εκτίμηση μιας περίληψης ενός σημασιολογικού γραφήματος που αντιστοιχεί σε ένα αρχικό κείμενο. Σημειώνεται, ότι κατά τη διαδικασία εκπαίδευσης ενός μοντέλου μηχανικής μάθησης, οι διανυσματικές αναπαραστάσεις των λέξεων δε θεωρούνται σταθερές αλλά τα διανύσματα μπορούν να προσαρμόζονται περαιτέρω σύμφωνα με τα χρησιμοποιούμενα παραδείγματα χρήσης ενός συνόλου εκπαίδευσης.

6.8.1 Διανυσματική αναπαράσταση λεκτικών μονάδων σημασιολογικών γραφημάτων και περιλήψεων

Το προτεινόμενο πλαίσιο στη φάση εκπαίδευσης μοντέλων μηχανικής μάθησης (Ενότητα 6.8.2) χρησιμοποιεί τόσο ενσωματώσεις λέξεων ανεξάρτητες των συμφραζομένων όσο και διανυσματικές αναπαραστάσεις λέξεων που βασίζονται στα συμφραζόμενα. Οι προσεγγίσεις αυτές έχουν περιγραφεί στην Ενότητα 3.2.1. Σημειώνεται, ότι μπορεί να χρησιμοποιηθεί οποιαδήποτε μεθοδολογία διανυσματικής αναπαράστασης λέξεων, με την προϋπόθεση ότι τα διανύσματα διατηρούν τις σημασιολογικές σχέσεις μεταξύ των λέξεων στον διανυσματικό χώρο. Στην πειραματική διαδικασία αυτής της εργασίας και σύμφωνα με τα χρησιμοποιούμενα μοντέλα βαθιάς μάθησης που παρουσιάζονται παρακάτω (Ενότητα 6.8.2), αξιοποιούμε τα μοντέλα *Word2Vec* [82, 137] και *BERT (bidirectional encoder representations from transformers)* [102], για ενσωματώσεις λέξεων ανεξάρτητες των συμφραζομένων και εξαρτώμενες από τα συμφραζόμενα, αντίστοιχα.

Για τη διαμόρφωση των διανυσματικών αναπαραστάσεων τύπου ενσωμάτωσης λέξεων ανεξάρτητων των συμφραζομένων, εκπαιδεύουμε ένα κατάλληλο μοντέλο χρησιμοποιώντας τα παραδείγματα χρήσης ενός συνόλου εκπαίδευσης, τα οποία περιλαμβάνουν ζεύγη σημασιολογικού γραφήματος - περίληψης. Ένα σημασιολογικό γράφημα αναπαρίσταται από μια ακολουθία λεκτικών μονάδων, σύμφωνα με τη χρησιμοποιούμενη αναπαράσταση γραφήματος σε μορφή κειμένου μιας γραμμής, όπως αναφέρθηκε στην Ενότητα 6.7. Μια περίληψη, η οποία είναι σε μορφή κειμένου, επίσης, αναπαρίσταται από μια ακολουθία λεκτικών μονάδων. Οι δύο προαναφερθείσες ακολουθίες λεκτικών μονάδων ενώνονται δημιουργώντας μια ακολουθία λεκτικών μονάδων που αντιστοιχεί σε ένα παράδειγμα χρήσης για την εκπαίδευση ενός μοντέλου ενσωμάτωσης λέξεων.

Η ακολουθία των λεκτικών μονάδων, η οποία αναπαριστά ένα σημασιολογικό γράφημα ενός υποψήφιου για περίληψη κειμένου, περιέχει λέξεις που είναι κοινές με λέξεις που περιλαμβάνονται και στο λεξιλόγιο των περιλήψεων (π.χ., *boy, train, person* κλπ.). Τα διανύσματα αυτών των λέξεων διαμορφώνονται χρησιμοποιώντας τις ακολουθίες των λεκτικών μονάδων που αναπαριστούν ένα σημασιολογικό γράφημα και τις ακολουθίες των λεκτικών μονάδων που σχηματίζουν τις περιλήψεις. Επίσης, τα σημασιολογικά γραφήματα περιέχουν αναγνωριστικά αποσαφήνισης εννοιών (π.χ., *want-01, say-01* κλπ.) και λέξεις-κλειδιά για τον προσδιορισμό των σχέσεων μεταξύ των κόμβων ενός γραφήματος (π.χ., *:ARG0, :ARG1, :name* κλπ. για γραφήματα τύπου *AMR*). Οι δύο προαναφερόμενες μορφές λεκτικών μονάδων, οι έννοιες και οι λέξεις-κλειδιά των σχέσεων, εμφανίζονται μόνο στα σημασιολογικά γραφήματα του αρχικού κειμένου, ενώ, όπως έχει αναφερθεί ήδη, υπάρχουν και κοινές λέξεις που εμφανίζονται τόσο σε ένα σημασιολογικό γράφημα όσο και σε

μία περίληψη. Επομένως, οι έννοιες και οι σχέσεις αναπαρίστανται από διανύσματα που ανακτώνται χρησιμοποιώντας μόνο τα σημασιολογικά γραφήματα των αρχικών κειμένων κατά τη διαδικασία εκπαίδευσης ενός μοντέλου ενσωμάτωσης λέξεων. Σε αυτό το σημείο θα πρέπει να σημειωθεί ότι τα διανύσματα ενσωμάτωσης λέξεων που έχουν αρχικά ανακτηθεί με την εκπαίδευση ενός κατάλληλου μοντέλου (π.χ., *Word2Vec*), προσαρμόζονται περαιτέρω κατά τη διαδικασία της εκπαίδευσης ενός μοντέλου μηχανικής μάθησης για την αυτόματη *ΠΚ*. Επίσης, πρέπει να αναφερθεί ότι το βήμα της εκπαίδευσης ενός μοντέλου ενσωμάτωσης λέξεων μπορεί να παραληφθεί, αρχικοποιώντας την αναπαράσταση κάθε λέξης σε ένα τυχαίο διάνυσμα D διαστάσεων. Τα διανύσματα αυτά, στη συνέχεια, διαμορφώνονται κατά τη φάση εκπαίδευσης ενός μοντέλου μηχανικής μάθησης για την αυτόματη *ΠΚ*. Παρά το γεγονός ότι είναι προαιρετική η παρουσία προ-εκπαιδευμένων διανυσμάτων αναπαράστασης λέξεων, η διαδικασία εκπαίδευσης ενός μοντέλου προβλέψεων μηχανικής μάθησης για την αυτόματη *ΠΚ* μπορεί να επιταχυνθεί στην περίπτωση που υπάρχει διαθέσιμο ένα ήδη εκπαιδευμένο μοντέλο ενσωμάτωσης λέξεων.

Για τη διανυσματική αναπαράσταση των λέξεων που βασίζεται στα συμφραζόμενα, χρησιμοποιούμε το μοντέλο *BERT* [102], το οποίο χρησιμοποιεί ενσωματώσεις λέξεων ή ενσωματώσεις τμημάτων μιας λέξης (δηλ., μια λέξη μπορεί να χωριστεί σε επιμέρους τμήματα που συνθέτουν τη λέξη αυτή σύμφωνα με το λεξιλόγιο του προ-εκπαιδευμένου μοντέλου). Για να συμπεριλάβουμε στο λεξιλόγιο του προ-εκπαιδευμένου μοντέλου τις λέξεις-κλειδιά που υποδηλώνουν τις σχέσεις των σημασιολογικών γραφημάτων, προσθέτουμε αυτές τις λέξεις-κλειδιά στο λεξιλόγιο του προ-εκπαιδευμένου μοντέλου ως λέξεις που δεν διασπώνται σε τμήματα λέξεων. Τα διανύσματα αυτών των λέξεων-κλειδιών αρχικοποιούνται τυχαία και προσαρμόζονται ανάλογα κατά τη διαδικασία εκπαίδευσης ενός μοντέλου βαθιάς μάθησης για την αυτόματη *ΠΚ* (Ενότητα 6.8.2).

6.8.2 Μοντέλα βαθιάς μάθησης

Με δεδομένη μια ακολουθία λεκτικών μονάδων για την αναπαράσταση ενός σημασιολογικού γραφήματος $G = (g_1,\ g_2, ...,\ g_k)$ που αντιστοιχεί σε ένα υποψήφιο για περίληψη κείμενο, ένα μοντέλο βαθιάς μάθησης, το οποίο εκπαιδεύεται σε ζεύγη σημασιολογικού γραφήματος-περίληψης, χρησιμοποιείται για την εκτίμηση μιας ακολουθίας λεκτικών μονάδων $Y' = (y'_1,\ y'_2, ...)$ που διαμορφώνουν την περίληψη. Στο πλαίσιο αυτό, εξετάζουμε πέντε μοντέλα βαθιάς μάθησης αρχιτεκτονικής:

(*i*) Κωδικοποιητή-αποκωδικοποιητή με μηχανισμούς προσοχής και αντιγραφής λέξεων εκτός λεξιλογίου,

(*ii*) Ενισχυτικής μάθησης,

(*iii*) Μετασχηματιστών,

(*iv*) Μετασχηματιστών με ενσωματώσεις λέξεων εξαρτώμενες από τα συμφραζόμενα και

(*v*) Μετασχηματιστών προ-εκπαιδευμένου κωδικοποιητή.

Τα παραπάνω μοντέλα βασίζονται ή αποτελούν παραλλαγή των αρχιτεκτονικών βαθιάς μάθησης που παρουσιάστηκαν στην Ενότητα 3.3. Η περιγραφή που ακολουθεί σχετικά με τις αρχιτεκτονικές

βαθιάς μάθησης παρουσιάζει τις διαφορές μεταξύ της βασικής αρχιτεκτονικής των εν λόγω μοντέλων και της προσαρμογής τους για αξιοποίηση στο προτεινόμενο πλαίσιο, το οποίο βασίζεται σε προβλέψεις μηχανικής μάθησης της μορφής σημασιολογικό γράφημα-σε-περίληψη.

Μοντέλο κωδικοποιητή-αποκωδικοποιητή με μηχανισμούς προσοχής και αντιγραφής λέξεων εκτός λεξιλογίου (*ΚΑΜΠΑΛ*)

Σε αυτή την ενότητα παρουσιάζονται οι διαφορές από τη βασική αρχιτεκτονική του εν λόγω μοντέλου που έχει ήδη παρουσιαστεί στην Ενότητα 3.3.2.

Επίπεδο διανυσματικής αναπαράστασης λέξεων: Η κύρια διαφορά του μοντέλου μηχανικής μάθησης που παρουσιάστηκε στην Ενότητα 3.3.2 με το αντίστοιχο μοντέλο που χρησιμοποιείται στο προτεινόμενο πλαίσιο βρίσκεται στο στρώμα διανυσματικής αναπαράστασης των λέξεων. Το στρώμα αυτό δέχεται τις λεκτικές μονάδες αναπαράστασης ενός σημασιολογικού γραφήματος, οι οποίες αντιστοιχούν σε ένα υποψήφιο προς περίληψη κείμενο, αντικαθιστώντας κάθε λεκτική μονάδα αναπαράστασης του γραφήματος με ένα διάνυσμα, σύμφωνα με το χρησιμοποιούμενο μοντέλο ενσωμάτωσης λέξεων. Οι ενσωματώσεις λέξεων μπορεί να είναι διαθέσιμες κατά την έναρξη της εκπαίδευσης ενός νευρωνικού δικτύου και, στη συνέχεια, κατά τη διαδικασία εκπαίδευσης, να επιτυγχάνεται η περαιτέρω προσαρμογή των διανυσμάτων αυτών, όπως έχουμε ήδη εξηγήσει στην Ενότητα 6.8.1.

Επίπεδο αμφίδρομων μονάδων *LSTM* κωδικοποιητή: Το δεύτερο επίπεδο δέχεται τις διανυσματικές αναπαραστάσεις μιας ακολουθίας λεκτικών στοιχείων που αναπαριστούν ένα σημασιολογικό γράφημα $G = (g_1,\ g_2, ...,\ g_k)$ (λαμβάνοντας ένα διάνυσμα σε κάθε χρονικό βήμα του κωδικοποιητή). Το στρώμα αυτό αποτελείται από μια ή περισσότερες αμφίδρομες μονάδες *LSTM* [88, 23] που παράγουν στην έξοδό του μια κρυφή κατάσταση $H_t = bi\_lstm(g_t, H_{t-1})$, σύμφωνα με την εξίσωση 6.1, όπως έχει περιγραφεί με μεγαλύτερη λεπτομέρεια στην Ενότητα 3.3.1.

$$\begin{aligned} \overrightarrow{h_t} &= lstm(g_i,\ \overrightarrow{h_{t-1}}),\ g_i \in (g_1,\ g_2,\ ...,\ g_{n-1},\ g_k) \\ \overleftarrow{h_t} &= lstm(g_j,\ \overleftarrow{h_{t-1}}),\ g_j \in (g_k,\ g_{n-1}\ ,...,\ g_2,\ \ g_1) \\ H_t &= [\overrightarrow{h_t} + \overleftarrow{h_t}] \end{aligned} \tag{6.1}$$

Επίπεδο προσοχής: Ο μηχανισμός προσοχής που χρησιμοποιείται είναι αυτός που παρουσιάστηκε στην Ενότητα 3.3.1. Η προσθήκη του μηχανισμού αυτού στο παρόν δίκτυο οδηγεί σε βελτίωση της ακρίβειας των προβλέψεων μηχανικής μάθησης, εστιάζοντας σε σχετικές λέξεις που περιλαμβάνονται σε ένα σημασιολογικό γράφημα εισόδου. Με δεδομένο ότι υπάρχουν λέξεις μεταξύ των λεκτικών μονάδων ενός σημασιολογικού γραφήματος που περιλαμβάνονται και στο λεξιλόγιο των περιλήψεων, όπως έχουμε ήδη εξηγήσει στην Ενότητα 6.3, μόνο αυτές οι λέξεις μπορούν να εμφανιστούν σε μια περίληψη. Για τον λόγο αυτό, ο μηχανισμός προσοχής εστιάζει στις λέξεις εκείνες, της αναπαράστασης ενός σημασιολογικού γραφήματος, που δυνητικά θα μπορούσαν να εμφανιστούν σε μια περίληψη. Από την άλλη πλευρά, λεκτικές μονάδες που δεν περιλαμβάνονται ποτέ σε μια περίληψη, όπως οι λέξεις-κλειδιά που αναπαριστούν τις σχέσεις σε ένα σημασιολογικό γράφημα, δεν ενισχύονται από τον μηχανισμό προσοχής με αποτέλεσμα να αποτρέπεται η εμφάνισή τους σε μια εκτιμώμενη περίληψη.

Επίπεδο μονάδων *LSTM* αποκωδικοποιητή: Η αρχιτεκτονική του αποκωδικοποιητή, ο οποίος αποτελείται από μονάδες *LSTM* μονής κατεύθυνσης, είναι αυτή που έχει περιγραφεί στην Ενότητα 3.3.1.

Επίπεδο κανονικοποιημένης εκθετικής συνάρτησης (softmax): Το επίπεδο αυτό χρησιμοποιείται για τον υπολογισμό μιας κατανομής πιθανοτήτων των λέξεων του χρησιμοποιούμενου λεξιλογίου, σε κάθε χρονικό βήμα λειτουργίας του νευρωνικού δικτύου. Η πιθανότητα κάθε υποψήφιας λέξης y_i του λεξιλογίου Y να παρουσιαστεί σε μια εκτιμώμενη περίληψη υπολογίζεται σύμφωνα με την Εξίσωση 6.2.

$$p_t(y_i|G,\ y_{t-1}) = \frac{e^{h_i^T w_i + b_i}}{\sum_{j=1}^{k} e^{h_t^T w_j + b_j}} \tag{6.2}$$

όπου G, y_{t-1} και h_t είναι το σημασιολογικό γράφημα εισόδου, η προηγούμενη εκτιμώμενη λέξη και η κρυφή κατάσταση του αποκωδικοποιητή, αντίστοιχα. Οι παράμετροι w και b αντιστοιχούν στα βάρη και τις πολώσεις που προσαρμόζονται κατά τη διαδικασία εκπαίδευσης του δικτύου. Το άθροισμα των πιθανοτήτων στο σύνολο των υποψήφιων λέξεων είναι ίσο με τη μονάδα (Εξίσωση 6.3).

$$\sum_{i=1}^{k} p_t(y_i|G,\ y_{t-1}) = 1 \tag{6.3}$$

Μηχανισμός αντιγραφής λέξεων εκτός λεξιλογίου: Το παραπάνω περιγραφόμενο δίκτυο είναι, επίσης, εξοπλισμένο με έναν μηχανισμό αντιγραφής λέξεων εκτός λεξιλογίου (*ΛΕΛ*) [61], ο οποίος επιτρέπει την αντιγραφή των *ΛΕΛ* από από ένα σημασιολογικό γράφημα στην περίληψη. Τα βασικά στοιχεία της αρχιτεκτονικής του δικτύου αυτού έχουν περιγραφεί στην Ενότητα 3.3.2, εδώ θα αναφέρουμε κάποιες διαφοροποιήσεις, καθώς σε αυτή την περίπτωση χρησιμοποιούμε την αναπαράσταση ενός κειμένου σε μορφή γραφήματος ως είσοδο στο δίκτυο. Αυτός ο μηχανισμός υπολογίζει τα βάρη παρουσίας μιας λέξης στην περίληψη, χρησιμοποιώντας το σταθερό λεξιλόγιο του συνόλου εκπαίδευσης και τις *ΛΕΛ* που ενδεχομένως εμφανίζονται στην αναπαράσταση του γραφήματος εισόδου. Επομένως, το λεξιλόγιο επεκτείνεται ώστε να περιλαμβάνει τις *ΛΕΛ* που ενδεχομένως εμφανίζονται στο γράφημα εισόδου. Διευκρινίζεται ότι οι προκαθορισμένες λέξεις-κλειδιά ή οι λεκτικές μονάδες, που χρησιμοποιούνται για την αναπαράσταση ενός σημασιολογικού γραφήματος και δεν εμφανίζονται ποτέ σε μια περίληψη, δεν μπορούν να θεωρηθούν *ΛΕΛ* και δεν λαμβάνονται υπόψη στους υπολογισμούς του μηχανισμού αντιγραφής *ΛΕΛ*, καθώς δεν εμφανίζονται ποτέ στην έξοδο. Πιο συγκεκριμένα, σε κάθε χρονικό βήμα t, η κατανομή πιθανότητας της γεννήτριας υπολογίζεται από την Εξίσωση 6.4.

$$P_{g,t} = \sigma(w_c \cdot c_t + w_s \cdot s_t + w_g \cdot g_t + b_p) \tag{6.4}$$

όπου σ υποδηλώνει τη σιγμοειδή συνάρτηση, c_t είναι το διάνυσμα περιβάλλοντος όπως υπολογίζεται από τον μηχανισμό προσοχής (Εξίσωση 3.3), s_t είναι η κατάσταση του αποκωδικοποιητή και g_t είναι το διάνυσμα εισόδου. Τα βάρη w_c, w_s και w_g προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης όπως και η πόλωση b_p του μηχανισμού προσοχής. Η πιθανότητα $P_t(i)$ παρουσίας της λέξης i στην

εκτιμώμενη περίληψη κατά το χρονικό βήμα t υπολογίζεται σύμφωνα με την εξίσωση 6.5, η οποία χρησιμοποιείται για τον υπολογισμό της κατανομής πιθανοτήτων στο εκτεταμένο λεξιλόγιο.

$$P_t(i) = P_{g,t} \cdot P_v(i) + (1 - P_{g,t}) \cdot a_{i,t} \tag{6.5}$$

όπου $P_v(i)$ είναι η πιθανότητα παρουσίασης μιας λέξης i στην ακολουθία των λέξεων εξόδου, σύμφωνα με την κατανομή πιθανοτήτων P_v των λέξεων του σταθερού λεξιλογίου του συνόλου εκπαίδευσης, όπως υπολογίζεται από το επίπεδο κανονικοποιημένης εκθετικής συνάρτησης (Εξίσωση 6.2). Η παράμετρος $a_{i,t}$ αντιπροσωπεύει το βάρος της λέξης i στο χρονικό βήμα t, όπως υπολογίζεται από τον μηχανισμό προσοχής. Εάν η λέξη i είναι *ΛΕΛ*, τότε $P_v(i) = 0$. Διαφορετικά, εάν μια λέξη δεν εμφανίζεται στην αναπαράσταση του γραφήματος εισόδου, τότε $a_{i,t} = 0$. Με αυτόν τον τρόπο, το δίκτυο εκτιμά την κατανομή πιθανοτήτων στο εκτεταμένο λεξιλόγιο ή στο σταθερό λεξιλόγιο σε περίπτωση που δεν υπάρχουν *ΛΕΛ*.

Μηχανισμός κάλυψης: Επιπλέον, το μοντέλο ενσωματώνει έναν μηχανισμό κάλυψης για την αποφυγή της επανάληψης των ίδιων λέξεων στην έξοδο. Για τον σκοπό αυτό προσαρμόζουμε τη λύση που προτείνεται στην εργασία [98] για το πρόβλημα της μηχανικής μετάφρασης, ως προσθήκη στο δίκτυο, όπως έχει περιγραφεί στην Ενότητα 3.3.2.

Το μοντέλο βαθιάς μάθησης εκπαιδεύεται από άκρο-σε-άκρο σύμφωνα με τη διαδικασία εποπτευόμενης μάθησης, χρησιμοποιώντας ένα σύνολο εκπαίδευσης με παραδείγματα χρήσης αποτελούμενα από ζεύγη σημασιολογικού γραφήματος-περίληψης. Για τη σύγκλιση των βαρών του δικτύου χρησιμοποιείται στοχαστική κατάβαση κλίσης, ελαχιστοποιώντας την αρνητική λογαριθμική πιθανοφάνεια μίας λέξης-στόχου y_t (Εξίσωση 6.6) που χρησιμοποιείται ως συνάρτηση σφάλματος.

$$Loss = -\sum_{t=1}^{T} logP(y_t|G) \tag{6.6}$$

όπου $P(y_t|G)$ είναι η υπό συνθήκη πιθανότητα της λέξης-στόχου y_t στο χρονικό βήμα $t \in (1,\ 2, ...,\ T)$ των T λέξεων της περίληψης, δεδομένου ενός σημασιολογικού γραφήματος εισόδου G.

Επιπλέον, χρησιμοποιείται η τεχνική απόρριψης συνδέσεων μεταξύ κόμβων του δικτύου (*dropout*) [92, 94] για την αποφυγή της υπερ-προσαρμογής (*overfitting*) κατά τη διαδικασία εκπαίδευσης.

Για τη βελτιστοποίηση της εκτιμώμενης ακολουθίας λέξεων που θα αποτελέσει την περίληψη, χρησιμοποιείται ο αλγόριθμος αναζήτησης δέσμης *beam search* [95, 96], ο οποίος έχει περιγραφεί στην Ενότητα 3.3.1.

Μοντέλο ενίσχυσης μάθησης (*EM*)

Στο πλαίσιο της *ΕΜ*, το παραπάνω μοντέλο (*ΚΑΜΠΑΛ*) χρησιμοποιείται ως πράκτορας που αλληλεπιδρά με το περιβάλλον του. Η μόνη διαφορά της αρχιτεκτονικής *ΕΜ* που χρησιμοποιείται στο προτεινόμενο πλαίσιο με αυτή που έχει περιγραφεί στην Ενότητα 3.3.3 είναι η διαφοροποίηση της εισόδου στο δίκτυο που χρησιμοποιείται ως πράκτορας. Η είσοδος σε αυτή την περίπτωση είναι η

ακολουθία των λεκτικών μονάδων που αναπαριστούν το σημασιολογικό γράφημα εισόδου G και όχι η ακολουθία των λέξεων του υποψήφιου προς περίληψη κειμένου X. Αντίστοιχα, η εκτίμηση μιας περίληψης γίνεται για δεδομένο σημασιολογικό γράφημα εισόδου. Συνεπώς, το χρησιμοποιούμενο μοντέλο *EM* διατηρεί την αρχιτεκτονική και τα χαρακτηριστικά του μοντέλου που περιγράφηκε στην Ενότητα 3.3.3, λαμβάνοντας υπόψη την προαναφερθείσα διαφοροποίηση.

Μοντέλα που βασίζονται σε μετασχηματιστές (*ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ*)

Με δεδομένο ότι τα μοντέλα που βασίζονται σε μετασχηματιστές [65, 74, 67, 66, 68] θεωρούνται προσεγγίσεις αιχμής σε εφαρμογές της επεξεργασίας φυσικής γλώσσας [69], διερευνούμε την αποτελεσματικότητα τριών τέτοιων αρχιτεκτονικών σε συνδυασμό με το προτεινόμενο πλαίσιο. Σε αυτή την κατεύθυνση, αξιοποιούμε τρία μοντέλα που βασίζονται σε αρχιτεκτονική μετασχηματιστών, τα οποία είναι τα ακόλουθα: (*i*) μοντέλο μετασχηματιστών (*ΜΣ*), (*ii*) μοντέλο μετασχηματιστών με ενσωματώσεις λέξεων που βασίζονται στα συμφραζόμενα (*ΜΣΕΛ*) και (*iii*) ένα μοντέλο μετασχηματιστών με προ-εκπαιδευμένο κωδικοποιητή (*ΜΣΠΚ*).

Το πρώτο και το τρίτο μοντέλο είναι εκείνα που έχουν περιγραφεί με λεπτομέρεια στην Ενότητα 3.3.4. Το δεύτερο μοντέλο (*ΜΣΕΛ*) είναι πανομοιότυπο με το δίκτυο *ΜΣ*, με τη διαφορά ότι η διανυσματική αναπαράσταση των λέξεων δεν αρχικοποιείται τυχαία, αλλά τα διανύσματα ανακτώνται με χρήση κάποιου μοντέλου ενσωμάτωσης λέξεων που βασίζεται στα συμφραζόμενα (π.χ., *BERT*). Και σε αυτή την περίπτωση, η αρχιτεκτονική και τα χαρακτηριστικά των δικτύων μετασχηματιστών παραμένουν σύμφωνα με την περιγραφή που έχει γίνει στην Ενότητα 3.3.4, με τη διαφοροποίηση ότι η είσοδος σε ένα δίκτυο αποτελείται από την ακολουθία των λεκτικών μονάδων που αναπαριστούν ένα σημασιολογικό γράφημα ενός υποψηφίου για περίληψη κειμένου, και, επίσης, η εκτίμηση των περιλήψεων γίνεται για δεδομένο σημασιολογικό γράφημα εισόδου.

Κεφάλαιο 7

Πειραματικό μέρος για την αυτόματη περίληψη κειμένου με χρήση βαθιάς μάθησης και σημασιολογικών γραφημάτων

7.1 Γενικά

Σε αυτό το κεφάλαιο παρουσιάζεται το πειραματικό μέρος για την αξιολόγηση του προτεινόμενου πλαισίου που έχει περιγραφεί στο κεφάλαιο 6, το οποίο αξιοποιεί μηχανική μάθηση και σημασιολογική αναπαράσταση περιεχομένου σε μορφή γραφήματος για την αυτόματη *ΠΚ*. Στο πλαίσιο της πειραματικής διαδικασίας, εξετάζονται μία σειρά από σημαντικές πτυχές της προτεινόμενης μεθοδολογίας, όπως αναφέρουμε με λεπτομέρεια παρακάτω.

Στη συνέχεια ακολουθεί η περιγραφή των χρησιμοποιούμενων συνόλων δεδομένων (Ενότητα 7.2) και η παρουσίαση των μετρικών αξιολόγησης (Ενότητα 7.3), όπου γίνεται και η εισαγωγή ενός νέου συνόλου μετρικών για τον προσδιορισμό της συνέπειας απόδοσης πληροφορίας των παραγόμενων περιλήψεων, ως επέκταση της μέτρησης ακρίβειας απόδοσης πληροφορίας που έχει ήδη παρουσιαστεί στην Ενότητα 5.3. Επίσης, περιγράφεται η πειραματική διαδικασία και οι επιλογές βελτιστοποίησης των χρησιμοποιούμενων παραμέτρων (Ενότητα 7.4). Η επίδοση άλλων συναφών εργασιών, για λόγους σύγκρισης με την παρούσα προσέγγιση, αναφέρεται στην Ενότητα 7.5. Τα πειραματικά αποτελέσματα παρουσιάζονται στην Ενότητα 7.6 και η Ενότητα 7.7 περιλαμβάνει τη μελέτη περίπτωσης παραδειγμάτων χρήσης για την αποτύπωση της ροής εργασίας από το αρχικό κείμενο έως την εκτιμώμενη περίληψη. Στη συνέχεια, στην Ενότητα 7.8 επιχειρείται μια ερμηνεία των αποτελεσμάτων. Το Κεφάλαιο αυτό ολοκληρώνεται με την Ενότητα 7.9, στην οποία αναφέρονται κάποια γενικά συμπεράσματα που προέκυψαν από τη διερεύνηση της προτεινόμενης μεθοδολογίας, και τέλος, στην ίδια ενότητα προτείνονται κάποιες κατευθύνσεις για μελλοντική εργασία, η διερεύνηση των οποίων θα μπορούσαν να επιφέρει περαιτέρω βελτίωση στο προτεινόμενο πλαίσιο.

7.2 Σύνολα δεδομένων

Για την αξιολόγηση του προτεινόμενου πλαισίου χρησιμοποιούνται δύο δημοφιλή σύνολα δεδομένων, το *Gigaword* και το *CNN/DailyMail*, τα οποία έχουν παρουσιαστεί με λεπτομέρεια στην Ενότητα 3.4.1.

Από τα δύο χρησιμοποιούμενα σύνολα δεδομένων, *Gigaword* και *CNN/DailyMail*, τα οποία περιλαμβάνουν ζεύγη κειμένου-περίληψης, ανακτώνται τα *AMR* γραφήματα των κειμένων με χρήση ενός αναλυτή *AMR* γραφημάτων [170]. Ο αναλυτής *AMR* γραφημάτων επιστρέφει ένα σχετικό γράφημα για κάθε πρόταση ενός κειμένου. Μετά τη διαδικασία αυτή, της ανάκτησης των σημασιολογικών γραφημάτων, εφαρμόζουμε τις διαδικασίες κατασκευής και μετασχηματισμού γραφημάτων που έχουν περιγραφεί στις Ενότητες 6.6 και 6.7, αντίστοιχα. Σύμφωνα με τη διαδικασία κατασκευής γραφήματος, δημιουργείται ένα σημασιολογικό γράφημα ενός κειμένου είτε (*i*) ως μια ακολουθία υπο-γραφημάτων είτε (*ii*) ως ένας συνδυασμός υπο-γραφημάτων (δηλ., συγχώνευση των επιμέρους υπο-γραφημάτων των προτάσεων του κειμένου για τη δημιουργία μιας συνολικής αναπαράστασης του κειμένου σε μορφή γραφήματος). Στο επόμενο βήμα, εφαρμόζεται η μεθοδολογία μετασχηματισμού γραφήματος, η οποία περιλαμβάνει τέσσερις εναλλακτικές μεθόδους που είναι οι εξής: (*i*) αρχικό σημασιολογικό γράφημα (*ΑΡΣΓ*), (*ii*) αρχικό σημασιολογικό γράφημα χωρίς αναγνωριστικά εννοιών (*ΑΡΣΓΧΕ*), (*iii*) απλοποιημένο σημασιολογικό γράφημα (*ΑΠΣΓ*) και (*iv*) απλοποιημένο σημασιολογικό γράφημα χωρίς αναγνωριστικά εννοιών (*ΑΠΣΓΧΕ*). Ο πίνακας 7.1 περιγράφει το σύνολο των εναλλακτικών που αντιστοιχούν σε σχήματα αναπαράστασης των σημασιολογικών γραφημάτων, οι οποίες αντιστοιχούν σε ισάριθμες εκδόσεις ενός συνόλου δεδομένων στο οποίο εφαρμόζονται τεχνικές κατασκευής και μετασχηματισμού γραφημάτων. Τα οχτώ εναλλακτικά σχήματα αναπαράστασης των δεδομένων που προκύπτουν εφαρμόζονται στο σύνολο δεδομένων *CNN/DailyMail*, καθώς περιέχει κείμενα πολλαπλών προτάσεων και, κατά συνέπεια, είναι διαθέσιμα περισσότερα από ένα σημασιολογικά γραφήματα για κάθε παράδειγμα χρήσης. Συνεπώς, για το σύνολο δεδομένων *CNN/DailyMail* υπάρχει η δυνατότητα εφαρμογής της μεθοδολογίας τόσο για την κατασκευή όσο και για τον μετασχηματισμό ενός γραφήματος. Στην περίπτωση του συνόλου δεδομένων *Gigaword*, εξετάζουμε μόνο τις τέσσερις εναλλακτικές μεθόδους μετασχηματισμού γραφήματος, καθώς αυτό το σύνολο δεδομένων περιέχει κείμενα μιας πρότασης και το γράφημα *AMR* που λαμβάνεται για κάθε κείμενο δεν απαιτεί καμία διαδικασία κατασκευής σημασιολογικού γραφήματος για το υποψήφιο για περίληψη κείμενο.

Για τα προτεινόμενα σχήματα δεδομένων, Ο Πίνακας 7.2 αναφέρει στατιστικά στοιχεία σχετικά με το μέγεθος του λεξιλογίου και το μέσο μήκος της αναπαράστασης των σημασιολογικών γραφημάτων σε μορφή *AMR*, καθώς και το μέσο μήκος των αντίστοιχων περιλήψεων για τα δύο χρησιμοποιούμενα σύνολα δεδομένων, *Gigaword* και *CNN/DailyMail*. Το μέσο μήκος των σημασιολογικών γραφημάτων και των περιλήψεων, για ένα συγκεκριμένο σύνολο δεδομένων και σχήμα δεδομένων, αντιστοιχεί στο μέσο πλήθος των λεκτικών στοιχείων για την αναπαράσταση ενός γραφήματος ή μίας περίληψης, αντίστοιχα. Επίσης, έχει υπολογιστεί ο συνολικός αριθμός των λεκτικών στοιχείων και ο αριθμός των διακριτών λεκτικών στοιχείων (δηλ., το μέγεθος λεξιλογίου) για κάθε σχήμα δεδομένων. Το σχήμα δεδομένων *ΑΠΣΓΧΕ* παρουσιάζει το μικρότερο μέγεθος λεξιλογίου, καθώς και το μικρότερο μέσο μήκος ενός σημασιολογικού γραφήματος. Από την άλλη πλευρά, τα σχήματα δεδομένων που διατηρούν τα αναγνωριστικά αποσαφήνισης εννοιών (*ΑΡΣΓ* και *ΑΠΣΓ*) παρουσιάζουν αυξημένο μέγεθος λεξιλογίου και, επίσης, αυξημένο μέσο μήκος της

Πίνακας 7.1: Οι συνδυασμοί των τεχνικών κατασκευής και μετασχηματισμού γραφήματος ως σχήματα δεδομένων.

Κατασκευή γραφήματος	**Μετασχηματισμός γραφήματος**	**Σχήμα δεδομένων**
Ακολουθία υπο-γραφημάτων	Αρχικό σημασιολογικό γράφημα	*Α-ΑΡΣΓ*
	Αρχικό σημασιολογικό γράφημα χωρίς έννοιες	*Α-ΑΡΣΓΧΕ*
	Απλοποιημένο σημασιολογικό γράφημα	*Α-ΑΠΣΓ*
	Απλοποιημένο σημασιολογικό γράφημα χωρίς έννοιες	*Α-ΑΠΣΓΧΕ*
Συνδυασμός υπο-γραφημάτων	Αρχικό σημασιολογικό γράφημα	*Σ-ΑΡΣΓ*
	Αρχικό σημασιολογικό γράφημα χωρίς έννοιες	*Σ-ΑΡΣΓΧΕ*
	Απλοποιημένο σημασιολογικό γράφημα	*Σ-ΑΠΣΓ*
	Απλοποιημένο σημασιολογικό γράφημα χωρίς έννοιες	*Σ-ΑΠΣΓΧΕ*

αναπαράστασης ενός σημασιολογικού γραφήματος. Σύμφωνα με τα χαρακτηριστικά των συνόλων δεδομένων, όπως περιγράφονται στην Ενότητα 3.4.1, το αρχικό κείμενο έχει το ελάχιστο μήκος αλλά παρουσιάζει το μέγιστο μέγεθος λεξιλογίου σε σύγκριση με τα σχήματα δεδομένων σε μορφή σημασιολογικών γραφημάτων μορφής *AMR*. Το αυξημένο μέγεθος λεξιλογίου αναμένεται να επηρεάζει αρνητικά την ακρίβεια των προβλέψεων μηχανικής μάθησης και αποτελεί έναν παράγοντα που θα διερευνηθεί στη συνέχεια.

7.3 Μετρικές αξιολόγησης

7.3.1 Το σύνολο μετρικών Rouge

Για την αξιολόγηση της απόδοσης του προτεινόμενου πλαισίου χρησιμοποιείται το σύνολο μετρικών *Rouge* [109] και συγκεκριμένα, υπολογίζουμε την τιμή του αρμονικού μέσου (f_1) των μετρικών $Rouge_1$ (επικάλυψη λέξεων), $Rouge_2$ (επικάλυψη δύο διαδοχικών λέξεων) και $Rouge_L$ (επικάλυψη της μεγαλύτερης κοινής ακολουθίας λέξεων), καθώς η χρήση αυτών των μετρικών αποτελεί μια τυπική πρακτική για την αξιολόγηση συστημάτων αυτόματης *ΠΚ* [60, 59, 63]. Οι μετρήσεις γίνονται με χρήση του συνόλου ελέγχου για κάθε ένα από τα χρησιμοποιούμενα σύνολα δεδομένων. Οι μετρικές *Rouge* έχουν παρουσιαστεί με λεπτομέρεια στην Ενότητα 3.4.2.

7.3.2 Συνέπεια απόδοσης πληροφορίας

Σε μια προσπάθεια για μια περισσότερο ποιοτική και λιγότερο ποσοτική αξιολόγηση, προσδιορίζουμε τη συνέπεια απόδοσης πληροφορίας μιας εκτιμώμενης περίληψης σε σχέση με το αρχικό κείμενο, καθώς μια τέτοια μέτρηση παρουσιάζει ενδιαφέρον στον τομέα της αυτόματης *ΠΚ* [148, 149]. Για τον σκοπό αυτό, επεκτείνουμε τη μετρική που ονομάσαμε *ακρίβεια απόδοσης πληροφορίας* (*ΑΑΠ*), η οποία έχει παρουσιαστεί στην Ενότητα 5.3 και στη δημοσιευμένη εργασία μας [140], ορίζοντας μια νέα μετρική που την ονομάζουμε *συνέπεια απόδοσης πληροφορίας* (*ΣΑΠ*)

Πίνακας 7.2: Το μέγεθος του λεξιλογίου και το μέσο μήκος της αναπαράστασης των σημασιολογικών γραφημάτων σε μορφή *AMR* και των περιλήψεων για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail* (*CNN/DM*), καθώς και για τα προτεινόμενα σχήματα δεδομένων, σύμφωνα με τη μεθοδολογία κατασκευής και μετασχηματισμού των γραφημάτων.

Σύνολο δεδομένων	Σχήμα δεδομένων	Μέσο μήκος		Αριθμός λεκτικών στοιχείων		Αριθμός διακριτών λεκτικών στοιχείων	
		Γράφημα	Περίληψη	Γράφημα	Περίληψη	Γράφημα	Περίληψη
Gigaword	*ΑΡΣΓ*	$128,4$	$8,2$	$441,7M$	$28,2M$	84.627	68.882
	ΑΡΣΓΧΕ	$128,4$		$441,7M$		82.029	
	ΑΠΣΓ	$66,1$		$227,3M$		84.400	
	ΑΠΣΧΧΕ	$66,1$		$227,3M$		81.806	
CNN/DM	*Α-ΑΡΣΓ*	$849,1$	$61,08$	$277,4M$	$17,5M$	253.961	195.208
	Α-ΑΡΣΓΧΕ	$849,1$		$277,4M$		245.296	
	Α-ΑΠΣΓ	$494,4$		$142,0M$		253.165	
	Α-ΑΠΣΓΧΕ	$494,4$		$142,0M$		244.388	
	Σ-ΑΡΣΓ	$830,7$		$272,5M$		253.961	
	Σ-ΑΡΣΓΧΕ	$830,7$		$272,5M$		245.296	
	Σ-ΑΠΣΓ	$478,6$		$140,3M$		253.165	
	Σ-ΑΠΣΓΧΕ	$478,6$		$140,3M$		244.388	

και περιλαμβάνει τις παραλλαγές υπολογισμού της ακρίβειας (*precision*), της ανάκλησης (*recall*) και της τιμής f_β [206] για τον υπολογισμό των αντίστοιχων τιμών της μετρικής *ΣΑΠ*. Πιο συγκεκριμένα, όπως έχουμε αναφέρει με μεγαλύτερη λεπτομέρεια στην ενότητα 5.3, ανακτώνται τριάδες πληροφορίας από μια εκτιμώμενη περίληψη και το αντίστοιχο υποψήφιο για περίληψη κείμενο, οι οποίες ακολουθούν το πρότυπο *RDF* (*resource description framework*) [150]. Σημειώνεται ότι στο πλαίσιο της πειραματικής διαδικασίας, οι τριάδες πληροφορίας ανακτώνται χρησιμοποιώντας το εργαλείο λογισμικού *(OpenIE)* [151, 149]. Η επικάλυψη των ανακτώμενων τριάδων πληροφορίας μεταξύ μιας περίληψης και του αντίστοιχου κειμένου καθορίζει τη μετρούμενη *ΣΑΠ* (FC) που βασίζεται στην ακρίβεια (FC_p), στην ανάκληση (FC_r) και στην τιμή f_β (FC_{f_β}). Οι υπολογισμοί των τριών παραλλαγών της μετρικής *ΣΑΠ* γίνονται σύμφωνα με τις Εξισώσεις 7.1.

$$\begin{aligned} FC_p &= \frac{|F_t \cap F_s|}{|F_s|} \\ FC_r &= \frac{|F_t \cap F_s|}{|F_t|} \\ Fc_{f_\beta} &= \frac{(1+\beta^2)\cdot FC_p \cdot FC_r}{\beta^2 \cdot FC_p + FC_r} \end{aligned} \tag{7.1}$$

όπου F_t είναι ένα σύνολο τριάδων πληροφορίας από το αρχικό κείμενο, F_s είναι ένα σύνολο τριάδων πληροφορίας της αντίστοιχης εκτιμώμενης περίληψης και β είναι ο συντελεστής που υποδηλώνει πόσο σημαντικότερη είναι η μέτρηση που βασίζεται στην ανάκληση (FC_r) σε σχέση με εκείνη που

βασίζεται στην ακρίβεια FC_p. Για να γίνει σαφής ο λόγος χρήσης της τιμής FC_{f_β}, θα πρέπει να σημειώσουμε ότι στην περίπτωση που το αρχικό κείμενο είναι πολύ μεγαλύτερης διάστασης (δηλ., περιλαμβάνει πολύ περισσότερες λέξεις) από τη διάσταση της περίληψης (π.χ., όπως συμβαίνει στο σύνολο δεδομένων *CNN/DailyMail*), η *ΣΑΠ* που βασίζεται στην ανάκληση αναμένεται να πάρει πολύ μικρές τιμές. Καθώς σε αυτή την περίπτωση, το αρχικό κείμενο περιέχει πολύ περισσότερες τριάδες πληροφορίας από αυτές της περίληψής του. Για τον λόγο αυτό χρησιμοποιούμε τη μετρική f_β, η οποία παρέχει μια σταθμισμένη τιμή υπολογισμού της *ΣΑΠ*, λαμβάνοντας υπόψη την τιμή της ακρίβειας και της ανάκλησης, σύμφωνα με το σχετικό μήκος μεταξύ ενός αρχικού κειμένου και της περίληψής του. Για τον υπολογισμό της τιμής FC_{f_β}, υπολογίζουμε το πηλίκο του μήκους της περίληψης (L_S) προς το μήκος του κειμένου (L_T), το οποίο αντιστοιχεί στην τιμή του συντελεστή β, σύμφωνα με την εξίσωση 7.2.

$$\beta = \frac{L_S}{L_T} \tag{7.2}$$

Για παράδειγμα, αν ο αριθμός των λεκτικών στοιχείων ενός κειμένου και της περίληψής του είναι $L_T = 400$ και $L_S = 100$, αντίστοιχα, τότε $\beta = 100/400 = 0,25$. Επίσης, η τιμή β μπορεί να καθοριστεί σύμφωνα με το μέσο μήκος των κειμένων και των περιλήψεων των παραδειγμάτων χρήσης ενός συνόλου δεδομένων. Επομένως, η τιμή του συντελεστή f_β παρέχει μια σταθμισμένη τιμή της *ΣΑΠ* που υποδηλώνει τον βαθμό κάλυψης της πληροφορίας που περιλαμβάνει η εκτιμώμενη περίληψη σε σχέση με το αρχικό κείμενο, χωρίς η τιμή της μετρικής FC_{f_β} να επηρεάζεται ή να εξαρτάται από το σχετικό μήκος μεταξύ ενός αρχικού κειμένου και μιας περίληψης. Τέλος, σημειώνουμε ότι σε αντίθεση με τις μετρικές *Rouge* που παρέχουν μια ποσοτική αξιολόγηση, οι μετρήσεις σε όρους *ΣΑΠ* εστιάζουν σε μια περισσότερο ποιοτική παρά ποσοτική αξιολόγηση.

7.3.3 Ποσοστό νέων λέξεων

Με δεδομένο ότι το προτεινόμενο πλαίσιο αποτελεί μια προσέγγιση της αυτόματης *ΠΚ* που ακολουθεί τη μέθοδο της παραγωγής κειμένου, διερευνάται ο βαθμός που τα εξεταζόμενα μοντέλα βαθιάς μάθησης σε συνδυασμό με τα προτεινόμενα σχήματα δεδομένων παράγουν νέες λέξεις ή φράσεις στις εκτιμώμενες περιλήψεις. Νέες λέξεις θεωρούμε εκείνες που δεν αναφέρονται στο αρχικό κείμενο. Επομένως, η μετρική *NTR (new tokens rate)*, η οποία παρουσιάστηκε στην Ενότητα 5.3, υπολογίζει στο ποσοστό των λεκτικών μονάδων που εμφανίζονται σε μία περίληψη, οι οποίες δεν έχουν παρουσία στο αρχικό κείμενο. Σημειώνουμε ότι μια παρόμοια εκδοχή αυτής της μετρικής έχει χρησιμοποιηθεί σε προηγούμενες εργασίες [104, 160] του χώρου της αυτόματης *ΠΚ*.

7.4 Πειραματική διαδικασία και βελτιστοποίηση παραμέτρων

Η πειραματική διαδικασία απαιτεί τη διενέργεια των επιμέρους εργασιών που περιλαμβάνουν οι επιμέρους βαθμίδες του προτεινόμενου πλαισίου (Ενότητα 6.4). Οι εργασίες αυτές περιλαμβάνουν την ανάκτηση των *AMR* γραφημάτων, την κατασκευή και τον μετασχηματισμό των σημασιολογικών

γραφημάτων του συνόλου των παραδειγμάτων χρήσης. Η ολοκλήρωση των διαδικασιών αυτών δημιούργησε πολλές εκδόσεις των συνόλων δεδομένων που αντιστοιχούν σε διαφορετικά σχήματα δεδομένων, σύμφωνα με την Ενότητα 7.2. Οι διαφορετικές εκδοχές των συνόλων δεδομένων που προέκυψαν για τα σύνολα εκπαίδευσης, επικύρωσης και ελέγχου, οι οποίες περιλαμβάνουν ζεύγη αναπαράστασης σημασιολογικού γραφήματος-περίληψης, χρησιμοποιούνται στη φάση εκπαίδευσης και ελέγχου του προτεινόμενου πλαισίου. Ακολουθούν οι επιλογές των παραμέτρων που οδήγησαν στη βελτιστοποίηση των μοντέλων βαθιάς μάθησης, ενώ η απόδοση των εξεταζόμενων μοντέλων βαθιάς μάθησης σε συνδυασμό με τα προτεινόμενα σχήματα δεδομένων, καθώς και η ερμηνεία των αποτελεσμάτων παρουσιάζονται στις Ενότητες 7.6 και 7.8, αντίστοιχα.

7.4.1 Διανυσματική αναπαράσταση λεκτικών μονάδων

Για τη διανυσματική αναπαράσταση των λεκτικών μονάδων χρησιμοποιούνται ενσωματώσεις λέξεων είτε ανεξάρτητες είτε εξαρτώμενες από τα συμφραζόμενα, όπως έχει αναφερθεί στην Ενότητα 6.8.1. Πιο συγκεκριμένα, για την περίπτωση των διανυσμάτων ενσωμάτωσης λέξεων ανεξάρτητων από τα συμφραζόμενα, χρησιμοποιείται η προσέγγιση *Word2Vec* [82] για τα μοντέλα βαθιάς μάθησης *ΚΑΜΠΑΛ* και *ΕΜ*. Πιο συγκεκριμένα, εκπαιδεύονται τόσα μοντέλα *Word2Vec* αρχιτεκτονικής *CBOW* όσα είναι τα διαφορετικά σχήματα δεδομένων που εξετάζουμε στο πλαίσιο αυτής της εργασίας. Μέσα από τη διαδικασία εκπαίδευσης, η οποία διαρκεί 10 εποχές με ρυθμό μάθησης που μειώνεται σταδιακά από $0,02$ έως $0,001$ και μέγεθος παραθύρου ίσο με 5, διαμορφώνονται οι διανυσματικές αναπαραστάσεις των λέξεων, 300 διαστάσεων. Από την άλλη πλευρά, στην περίπτωση των διανυσμάτων ενσωμάτωσης λέξεων που βασίζονται στα συμφραζόμενα, τα οποία αξιοποιούνται στις αρχιτεκτονικές μετασχηματιστών (*ΜΣΕΛ* και *ΜΣΠΚ*), χρησιμοποιείται το προ-εκπαιδευμένο μοντέλο *BERT* (Ενότητα 6.8.1).

7.4.2 Εκπαίδευση μοντέλων βαθιάς μάθησης

Εκπαίδευση του μοντέλου *ΚΑΜΠΑΛ*

Το δίκτυο *ΚΑΜΠΑΛ*, το οποίο παρουσιάστηκε στην Ενότητα 6.8.2, εκπαιδεύεται με χρήση των διαφορετικών εκδόσεων των συνόλων εκπαίδευσης δημιουργώντας ισάριθμα εκπαιδευμένα μοντέλα μηχανικής μάθησης. Οι βέλτιστες τιμές των παραμέτρων αυτού του μοντέλου έχουν προσδιοριστεί μέσα από την πειραματική διαδικασία, χρησιμοποιώντας το σύνολο επικύρωσης κάθε συνόλου δεδομένων. Πιο συγκεκριμένα, ο κωδικοποιητής περιέχει δύο στρώματα αμφίδρομων μονάδων *LSTM* 256 διαστάσεων τα καθένα και ο αποκωδικοποιητής περιλαμβάνει ένα στρώμα μονάδας *LSTM* 512 διαστάσεων. Σε κάθε εποχή εκπαίδευσης τα παραδείγματα χρήσης αναδιατάσσονται τυχαία, το μέγεθος δέσμης ορίζεται σε 64 για το σύνολο δεδομένων *Gigaword* και 32 για το *CNN/DailyMail*, ενώ ο ρυθμός μάθησης έχει οριστεί σε $0,001$. Επιπλέον, ως μέθοδος βελτιστοποίηση χρησιμοποιείται η *Adam* [21] με ψαλιδισμό στην κατάβαση κλίσης [110] και ως συνάρτηση σφάλματος χρησιμοποιείται η αρνητική λογαριθμική πιθανοφάνεια [111]. Επιπλέον, χρησιμοποιείται μεθοδολογία απόρριψης συνδέσεων του δικτύου (*dropout*) με πιθανότητα απόρριψης $p = 20\%$. Το λεξιλόγιο περιορίζεται, χρησιμοποιώντας τις 150.000 πιο συχνά χρησιμοποιούμενες λεκτικές μονάδες του συνόλου εκπαίδευσης. Για την εκπαίδευση των μοντέλων χρησιμοποιούνται μονάδες γραφικής επεξεργασίας (*NVidia K40 GPU*) που οδηγούν σε επιτάχυνση της διαδικασίας

λόγω του παραλληλισμού των υπολογισμών. Η εκπαίδευση διαρκεί περίπου 15 εποχές μέχρι τα μοντέλα να συγκλίνουν επαρκώς.

Εκπαίδευση του μοντέλου *ΕΜ*

Εφόσον το μοντέλο *ΕΜ*, το οποίο παρουσιάστηκε στην Ενότητα 6.8.2, χρησιμοποιεί το δίκτυο *ΚΑΜΠΑΛ* ως πράκτορα, οι διαστάσεις των μονάδων *LSTM* παραμένουν ίδιες με αυτές του μοντέλου *ΚΑΜΠΑΛ*, όπως περιγράφηκε παραπάνω. Ο ρυθμός μάθησης έχει οριστεί σε 10^{-4} και το μέγεθος δέσμης σε 32 και 16 για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*, αντίστοιχα. Για τις υπόλοιπες παραμέτρους, ισχύουν οι ίδιες παραδοχές όπως και στο μοντέλο ΚΑΜΠΑΛ που περιγράφηκε παραπάνω.

Εκπαίδευση του μοντέλου *ΜΣ*

Ο κωδικοποιητής και ο αποκωδικοποιητής της αρχιτεκτονικής *ΜΣ* που παρουσιάστηκε στην Ενότητα 6.8.2 αποτελούνται από 6 πανομοιότυπα επίπεδα ο καθένας. Η διάσταση του μοντέλου για τον κωδικοποιητή και τον αποκωδικοποιητή έχει οριστεί σε 512 και το εσωτερικό στρώμα σε 2.048. Επιπλέον, ο αριθμός των κεφαλών ορίζεται σε 8 (δηλ., ο μηχανισμός προσοχής εκτελεί τους υπολογισμούς του σε 8 παράλληλα επίπεδα, μειώνοντας τη διάσταση του μοντέλου για κάθε επίπεδο σε $512/8 = 64$). Το μέγεθος δέσμης έχει οριστεί σε 64 και για τα δύο σύνολα δεδομένων, *Gigaword* και *CNN/DailyMail*, και η πιθανότητα απόρριψης συνδέσεων (*dropout*) τίθεται σε $p = 10\%$. Ως αλγόριθμος βελτιστοποίησης χρησιμοποιείται ο *Adam* με παραμέτρους $\beta_1 = 0,9$ και $\beta_2 = 0,999$. Ο ρυθμός μάθησης προσαρμόζεται κατά τη διαδικασία της εκπαίδευσης σύμφωνα με την Εξίσωση 7.3, όπου $warmupSteps = 10.000$ και $a = 0,05$. Σύμφωνα με αυτή την εξίσωση, ο ρυθμός μάθησης αυξάνεται για τα πρώτα $warmupSteps$ βήματα εκπαίδευσης και στη συνέχεια μειώνεται.

$$lr = a \cdot \min\{step^{-0.5},\ step \cdot warmupSteps^{-1.5}\} \tag{7.3}$$

Εκπαίδευση του μοντέλου *ΜΣΕΛ*

Η διαφορά του μοντέλου *ΜΣΕΛ* (Ενότητα 6.8.2) από το μοντέλο *ΜΣ* που αναφέρθηκε παραπάνω είναι ότι το μοντέλο *ΜΣΕΛ* χρησιμοποιεί ενσωματώσεις λέξεων τύπου *BERT*, ως είσοδο στο στρώμα διανυσματικής αναπαράστασης των λεκτικών μονάδων για την αρχικοποίηση των αντίστοιχων διανυσμάτων, ενώ το μοντέλο *ΜΣ* εκπαιδεύεται με τυχαία αρχικοποιημένα διανύσματα αναπαράστασης των λεκτικών μονάδων. Κατά τη διάρκεια της εκπαίδευσης του μοντέλου βαθιάς μάθησης, τα διανύσματα αναπαράστασης των λεκτικών μονάδων προσαρμόζονται περαιτέρω, σύμφωνα με τα παραδείγματα χρήσης του χρησιμοποιούμενου συνόλου δεδομένων.

Εκπαίδευση του μοντέλου *ΜΣΠΚ*

Το μοντέλο *ΜΣΠΚ* (Ενότητα 6.8.2) χρησιμοποιεί έναν προ-εκπαιδευμένο κωδικοποιητή τύπου *BERT* και έναν αποκωδικοποιητή τύπου μετασχηματιστή 6 πανομοιότυπων στρωμάτων, όπως και στην περίπτωση του αποκωδικοποιητή του μοντέλου *ΜΣ* που αναφέρθηκε παραπάνω. Οι διαστάσεις του μοντέλου για τον προ-εκπαιδευμένο κωδικοποιητή είναι 768. Επιπλέον, το μοντέλο

χρησιμοποιεί δύο διαδικασίες βελτιστοποίησης τύπου *Adam*, μια για τον κωδικοποιητή και μία για τον αποκωδικοποιητή, με σκοπό μια πιο ομαλή προσαρμογή των βαρών του δικτύου. Οι παράμετροι για κάθε αλγόριθμο βελτιστοποίησης είναι $\beta_1 = 0,9$ και $\beta_2 = 0,99$. Οι διαδικασίες βελτιστοποίησης έχουν διαφορετικό ρυθμό μάθησης που προσαρμόζεται σύμφωνα με την Εξίσωση 7.3, όπου $a_{enc} = 0,002$, $a_{dec} = 0,1$, $warmupSteps_{enc} = 20.000$ και $warmupSteps_{dec} = 10.000$. Οι διαφορετικοί ρυθμοί μάθησης στοχεύουν σε μια ομαλή προσαρμογή των βαρών του μοντέλου, καθώς για τον κωδικοποιητή τύπου *BERT* χρησιμοποιείται μικρότερος ρυθμός μάθησης για πιο ομαλή προσαρμογή των βαρών του από εκείνη του αποκωδικοποιητή, αποφεύγοντας την υπερ-προσαρμογή ή την υπο-προσαρμογή του κωδικοποιητή ή του αποκωδικοποιητή [66]. Οι υπόλοιπες παράμετροι παραμένουν ίδιες με τα μοντέλα *ΜΣ* και *ΜΣΕΛ*.

Βελτιστοποίηση εκτιμώμενων περιλήψεων

Για τη βελτιστοποίηση της ακολουθίας των λέξεων που αποτελεί την εκτιμώμενη περίληψη, σύμφωνα με την εκτιμώμενη πιθανότητα παρουσίας κάθε λέξης στην εκτιμώμενη περίληψη, χρησιμοποιείται ο αλγόριθμος αναζήτησης δέσμης (*beam search*) [95, 96]. Το πλάτος αναζήτησης του αλγόριθμου αυτού ορίζεται ίσο με 5 για όλα τα χρησιμοποιούμενα μοντέλα βαθιάς μάθησης.

7.5 Συναφείς προσεγγίσεις σύγκρισης

Για λόγους σύγκρισης της απόδοσης του προτεινόμενου πλαισίου με άλλες συναφείς προσεγγίσεις [175, 182, 185], στην Ενότητα 7.6, που παρουσιάζει τα αποτελέσματα της πειραματικής διαδικασίας αναφέρεται και η επίδοση άλλων σχετικών συστημάτων. Ως σχετικές προσεγγίσεις, θεωρούνται τα συστήματα αυτόματης *ΠΚ* που βασίζονται σε σημασιολογική αναπαράσταση σύμφωνα με το μοντέλο *AMR* γραφημάτων και επίσης, τα συστήματα αυτά είναι ικανά να διαχειριστούν τα δεδομένα ενός συνόλου δεδομένων μεγάλης κλίμακας, όπως το *Gigaword* ή το *CNN/DailyMail*. Με δεδομένο ότι η προτεινόμενη μεθοδολογία βασίζεται σε μοντέλα βαθιάς μάθησης, αυτά τα μοντέλα απαιτούν μεγάλο όγκο δεδομένων για να εκπαιδευτούν επαρκώς. Επομένως, από τη σύγκριση αποκλείουμε ορισμένες προσεγγίσεις [184, 183, 176] που έχουν αναφερθεί ως σχετικές εργασίες στην Ενότητα 6.2 και δεν ικανοποιούν αυτό το κριτήριο, διαχείρισης συνόλων δεδομένων μεγάλης κλίμακας. Οι προσεγγίσεις αυτές έχουν αξιολογηθεί με χρήση περιορισμένου αριθμού παραδειγμάτων χρήσης (π.χ., όπως είναι το τμήμα *proxy report* του συνόλου δεδομένων *AMR Bank* [207]). Συνεπώς, εφόσον τα μοντέλα βαθιάς μάθησης που χρησιμοποιούμε δεν μπορούν να εκπαιδευτούν επαρκώς σε ένα τόσο περιορισμένο σύνολο δεδομένων, οι προαναφερόμενες προσεγγίσεις αποκλείονται από την διαδικασία σύγκρισης. Όπως μπορούμε να δούμε στην ανάλυσή μας παρακάτω, αυτή η σύγκριση βασίζεται στις τιμές της μετρικής *Rouge*, οι οποίες έχουν αναφερθεί στις αντίστοιχες εργασίες παρουσίασης των προσεγγίσεων σύγκρισης. Υποθέτουμε ότι τα πειραματικά αποτελέσματα των άλλων προσεγγίσεων είναι άμεσα συγκρίσιμα με τα δικά μας, καθώς χρησιμοποιούν τα ίδια σύνολα δεδομένων, *Gigaword* και *CNN/DailyMail*, ακολουθώντας την ίδια μεθοδολογία προ-επεξεργασίας για την ανάκτηση των παραδειγμάτων χρήσης.

7.6 Πειραματικά αποτελέσματα

Στον πίνακα 7.3 αναφέρεται η απόδοση των μοντέλων *ΚΑΜΠΑΛ*, *ΕΜ*, *ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ* σε όρους μετρικών *Rouge* για το σύνολο δεδομένων *Gigaword*. Οι μετρήσεις αυτές αφορούν τα τέσσερα σχήματα δεδομένων που βασίζονται στις μεθόδους μετασχηματισμού γραφήματος (Ενότητα 6.7), τα οποία εφαρμόζονται για τη διαμόρφωση των ζευγών σημασιολογικών γραφημάτων-περίληψης που σχηματίζουν τα παραδείγματα χρήσης των συνόλων εκπαίδευσης, επικύρωσης και ελέγχου. Επιπλέον, αναφέρεται το ποσοστό των νέων λεκτικών μονάδων (NTR) (Ενότητα 7.3) για το σύνολο δεδομένων *Gigaword*. Όπως παρατηρούμε, οι υψηλότερες τιμές που λαμβάνουν οι μετρικές *Rouge* επιτυγχάνονται για τη μέθοδο μετασχηματισμού γραφήματος *ΑΠΣΓΧΕ*. Για τον λόγο αυτό, στην περίπτωση των μοντέλων που βασίζονται σε μετασχηματιστές (*ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ*), αναφέρουμε την επίδοση σε όρους *Rouge* για το σχήμα δεδομένων *ΑΠΣΓΧΕ*. Επίσης, για τον ίδιο λόγο, στα υπόλοιπα πειράματα χρησιμοποιούμε την ίδια μεθοδολογία (*ΑΠΣΓΧΕ*) ως μέθοδο μετασχηματισμού γραφήματος. Οι τελευταίες σειρές του Πίνακα 7.3 περιλαμβάνουν την επίδοση των σχετικών προσεγγίσεων που χρησιμοποιούνται για λόγους σύγκρισης (Ενότητα 7.5). Όπως παρατηρούμε, τα μοντέλα *ΜΣ* και *ΜΣΠΚ* λαμβάνουν τις υψηλότερες τιμές σε όρους μετρικών *Rouge*.

Πίνακας 7.3: Οι τιμές σε όρους μετρικών *Rouge* για το σύνολο δεδομένων *Gigaword* των μοντέλων βαθιάς μάθησης *ΚΑΜΠΑΛ*, *ΕΜ*, *ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ* και τα σχήματα δεδομένων: (i) αρχικών σημασιολογικών γραφημάτων ($ΑΡΣΓ$), (ii) αρχικών σημασιολογικών γραφημάτων χωρίς έννοιες ($ΑΡΣΓΧΕ$), (iii) απλοποιημένων σημασιολογικών γραφημάτων ($ΑΠΣΓ$) και (iv) απλοποιημένων σημασιολογικών γραφημάτων χωρίς έννοιες ($ΑΠΣΓΧΕ$) ($p_{value} < 0,01$).

Μοντέλο	Σχήμα δεδομένων	$Rouge_1$	$Rouge_2$	$Rouge_L$	NTR (%)
ΚΑΜΠΑΛ	*ΑΡΣΓ*	34,62	12,69	31,89	37,48
	ΑΡΣΓΧΕ	35,24	13,36	32,81	36,81
	ΑΠΣΓ	36,84	14,12	33,73	35,83
	ΑΠΣΓΧΕ	37,93	14,36	34,84	35,68
ΕΜ	*ΑΡΣΓ*	35,14	13,31	32,65	36,75
	ΑΡΣΓΧΕ	35,44	13,63	32,87	37,21
	ΑΠΣΓ	37,18	14,12	34,22	34,07
	ΑΠΣΓΧΕ	38,17	15,24	35,40	34,11
ΜΣ	*ΑΠΣΓΧΕ*	36,14	13,02	33,04	41,35
ΜΣΕΛ	*ΑΠΣΓΧΕ*	36,68	12,85	33,57	**41,46**
ΜΣΠΚ	*ΑΠΣΓΧΕ*	**38,97**	**15,87**	**36,17**	41,17
ABS+AMR [175]		31,64	12,94	28,54	
SemSUM [178]		38,78	19,75	36,09	

Επιπλέον, προσδιορίζουμε τη στατιστική σημαντικότητα των τιμών *Rouge* για κάθε μοντέλο με χρήση της t-δοκιμής του *Welch* [166]. Για να υπολογίσουμε τη στατιστική σημαντικότητα, χρησιμοποιούμε τις τιμές των μετρικών *Rouge* για κάθε επιμέρους σχήμα δεδομένων του συνόλου

ελέγχου. Οι υπολογισμοί της t-δοκιμής εφαρμόζονται σε κάθε ζεύγος σχημάτων δεδομένων για κάθε μετρική *Rouge* ($Rouge_1$, $Rouge_2$ και $Rouge_L$). Οι τιμές προσδιορισμού της στατιστικής σημαντικότητας που λαμβάνονται (p_{value}), οι οποίες αναφέρονται στη λεζάντα του Πίνακα 7.3, αποδεικνύουν ότι τα αποτελέσματα είναι στατιστικά σημαντικά για όλες τις περιπτώσεις που εξετάσαμε. Η ίδια μεθοδολογία ακολουθείται για τον προσδιορισμό της στατιστικής σημαντικότητας και στις υπόλοιπες μετρήσεις που αναφέρονται παρακάτω.

Πίνακας 7.4: Οι τιμές σε όρους μετρικών *Rouge* για το σύνολο δεδομένων *CNN/DailyMail* των μοντέλων βαθιάς μάθησης *ΚΑΜΠΑΛ*, *ΕΜ*, *ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ* για τις μεθόδους κατασκευής γραφήματος: (i) ακολουθία υπο-γραφημάτων και (ii) συνδυασμός υπο-γραφημάτων, καθώς και τεχνική μετασχηματισμού απλοποιημένου σημασιολογικού γραφήματος χωρίς έννοιες (*ΑΠΣΓΧΕ*) ($p_{value} < 0,01$).

Μοντέλο	Ακολουθία υπο-γραφημάτων				Συνδυασμός υπο-γραφημάτων			
	R_1	R_2	R_L	NTR (%)	R_1	R_2	R_L	NTR
ΚΑΜΠΑΛ	36,44	11,25	29,98	25,26	37,69	11,21	30,52	27,34
ΕΜ	37,11	**11,49**	**31,20**	26,16	**39,83**	**11,95**	**31,65**	28,30
ΜΣ	36,32	7,00	25,53	45,28	31,57	4,88	22,73	44,16
ΜΣΕΛ	39,49	8,70	26,61	44,81	34,36	6,57	23,91	**45,15**
ΜΣΠΚ	**40,75**	11,18	27,67	**45,38**	35,50	8,46	24,52	44,92
Lead-3-AMR [182]	31,70	5,80	16,80					

Παρόμοια με τα αποτελέσματα που περιγράφηκαν παραπάνω, ο Πίνακας 7.4 αναφέρει την απόδοση των χρησιμοποιούμενων μοντέλων σε όρους μετρικών *Rouge* για το σύνολο δεδομένων *CNN/DailyMail*. Ο πίνακας αυτός περιλαμβάνει τα αποτελέσματα για τις δύο μεθόδους κατασκευής γραφήματος (Ενότητα 6.6) και τα απλοποιημένα σημασιολογικά γραφήματα χωρίς έννοιες (*ΑΠΣΓΧΕ*) ως μέθοδο μετασχηματισμού των γραφημάτων. Επιπλέον, αναφέρονται οι τιμές της μετρικής NTR και η στατιστική σημαντικότητα (p_{value}) των αποτελεσμάτων στη λεζάντα του Πίνακα 7.4. Για λόγους σύγκρισης, η τελευταία σειρά του πίνακα αυτού περιλαμβάνει την επίδοση της σχετικής προσέγγισης που χρησιμοποιεί το ίδιο σύνολο δεδομένων. Σύμφωνα με τις μετρήσεις, τα μοντέλα βαθιάς μάθησης *ΕΜ* και *ΜΣΠΚ* είναι εκείνα που επιτυγχάνουν τις καλύτερες επιδόσεις σε όρους μετρικών *Rouge* και, επίσης, τα μοντέλα μετασχηματιστών (*ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ*) παράγουν περιλήψεις με τις περισσότερες νέες λέξεις, οι οποίες δεν συμπεριλαμβάνονται στο αρχικό κείμενο σύμφωνα με τη μετρική NTR.

Με δεδομένο ότι ο προσδιορισμός της συνέπειας απόδοσης πληροφορίας (*ΣΑΠ*, Ενότητα 7.3) αποτελεί μια ενδιαφέρουσα μέτρηση στον τομέα της αυτόματης *ΠΚ* [148, 167, 149, 140], οι Πίνακες 7.5 και 7.6 αναφέρουν τα αποτελέσματα των μετρήσεων αυτών για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*, αντίστοιχα. Για τον υπολογισμό της τιμής F_β της *ΣΑΠ*, ο συντελεστής β, ορίστηκε σε $\beta = 0,264$ και $\beta = 0,158$ για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*, αντίστοιχα. Οι τιμές αυτές του συντελεστή β αντιστοιχούν στο μέσο μήκους μιας περίληψης προς το μέσο μήκος ενός κειμένου για κάθε ένα από τα δύο σύνολα δεδομένων (η χρήση του συντελεστή β έχει εξηγηθεί λεπτομερώς στην Ενότητα 7.3). Οι αναφερόμενες τιμές των μετρήσεων της *ΣΑΠ* αντιστοιχούν στη μέση τιμή της κάθε μετρικής στο σύνολο των παραδειγμάτων χρήσης του συνόλου ελέγχου για κάθε ένα από τα σύνολα δεδομένων. Να αναφερθεί ότι η τελευταία σειρά αυτών

Πίνακας 7.5: Οι τιμές των μετρικών της συνέπειας απόδοσης πληροφορίας για το σύνολο δεδομένων *Gigaword* με χρήση των μοντέλων βαθιάς μάθησης (*ΚΑΜΠΑΛ*, *ΕΜ*, *ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ*) και την εφαρμογή των διαφορετικών μεθόδων μετασχηματισμού γραφήματος (*ΑΡΣΓ*, *ΑΡΣΓΧΕ*, *ΑΠΣΓ* και *ΑΠΣΓΧΕ*) ως σχήματα δεδομένων, για τις μετρήσεις FC_{f_β} τίθεται $\beta = 0,264$ ($p_{value} < 0,02$).

Μοντέλο	Σχήμα δεδομένων	FC_p (%)	FC_r (%)	FC_{f_β} (%)
ΚΑΜΠΑΛ	*ΑΡΣΓ*	74,03	24,93	65,60
	ΑΡΣΓΧΕ	77,92	28,93	70,16
	ΑΠΣΓ	81,93	26,60	72,13
	ΑΠΣΓΧΕ	83,23	24,67	72,06
ΕΜ	*ΑΡΣΓ*	80,11	24,48	69,75
	ΑΡΣΓΧΕ	81,67	25,21	71,25
	ΑΠΣΓ	81,12	25,58	71,04
	ΑΠΣΓΧΕ	**83,48**	25,96	72,92
ΜΣ	*ΑΠΣΓΧΕ*	79,17	27,69	70,60
ΜΣΕΛ	*ΑΠΣΓΧΕ*	61,55	32,13	58,08
ΜΣΠΚ	*ΑΠΣΓΧΕ*	79,01	**39,38**	**74,14**
Περιλήψεις αναφοράς		76,87	32,65	70,63

Πίνακας 7.6: Οι τιμές των μετρικών της συνέπειας απόδοσης πληροφορίας για το σύνολο δεδομένων *CNN/DailyMail* με χρήση των μοντέλων βαθιάς μάθησης (*ΚΑΜΠΑΛ*, *ΕΜ*, *ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ*), τις δύο τεχνικές κατασκευής γραφημάτων και μετασχηματισμό απλοποιημένου σημασιολογικού γραφήματος χωρίς έννοιες (*ΑΠΣΓΧΕ*), για τις μετρήσεις FC_{f_β} τίθεται $\beta = 0,158$ ($p_{value} < 0,025$).

Μοντέλο	Ακολουθία υπο-γραφημάτων			Συνδυασμός υπο-γραφημάτων		
	FC_p (%)	FC_r (%)	FC_{f_β} (%)	FC_p (%)	FC_r (%)	FC_{f_β} (%)
ΚΑΜΠΑΛ	64,96	2,68	41,45	58,97	2,71	39,16
ΕΜ	**68,79**	**3,00**	**44,82**	**63,87**	2,66	40,91
ΜΣ	57,61	2,65	38,26	61,09	1,83	34,14
ΜΣΕΛ	62,72	2,65	40,38	54,55	2,28	35,00
ΜΣΠΚ	65,45	2,97	43,27	58,89	**3,33**	**41,86**
Περιλήψεις αναφοράς	65,44	4,8	50,05			

των πινάκων περιλαμβάνει τα αποτελέσματα των μετρικών της *ΣΑΠ* για την περίληψη αναφοράς. Παρόμοια με τις τιμές *Rouge* των δύο περιπτώσεων, της *ΠΚ* σε κείμενο μικρής έκτασης (σύνολο δεδομένων *Gigaword*) και της *ΠΚ* σε επίπεδο εγγράφου (σύνολο δεδομένων *CNN/DailyMail*), τα μοντέλα βαθιάς μάθησης *ΕΜ* και *ΜΣΠΚ* παρουσιάζουν τις υψηλότερες τιμές σε όρους μετρικών *ΣΑΠ*.

7.7 Μελέτη περίπτωσης

Πίνακας 7.7: Παραδείγματα αυτόματης *ΠΚ* σε επίπεδο κειμένου μικρής έκτασης που αποτυπώνουν τη ροή εργασίας από το κείμενο εισόδου έως την εκτιμώμενη περίληψη για τις μεθόδους μετασχηματισμού γραφήματος *ΑΡΣΓ, ΑΡΣΓΧΕ, ΑΠΣΓ* και *ΑΠΣΓΧΕ*.

Κείμενο εισόδου: more than one million chinese have studied abroad over the last decade, an official with the ministry of education said here on monday .	
ΑΡΣΓ	**Αναπαράσταση σημασιολογικού γραφήματος:** (s / say-01 :arg0 (p / person :arg0-of (h2 / have-org-role-91 :arg1 (m3 / ministry) :arg2 (o / official))) :arg1 (s3 / study-01 :arg0 (p2 / person :mod (c / country :name (n / name :op1 “china”) :wiki “china”) :quant (m4 / more)) :arg1 (e / educate-01) :location (a / abroad) :time (s2 / since :time (l / late))) :arg2 (m / monday) :location (h / here)) **Περίληψη συστήματος:** more than chinese students studying abroad
ΑΡΣΓΧΕ	**Αναπαράσταση σημασιολογικού γραφήματος:** (s / say :arg0 (p / person :arg0-of (h2 / have-org-role :arg1 (m3 / ministry) :arg2 (o / official))) :arg1 (s3 / study :arg0 (p2 / person :mod (c / country :name (n / name :op1 “china”) :wiki “china”) :quant (m4 / more)) :arg1 (e / educate) :location (a / abroad) :time (s2 / since :time (l / late))) :arg2 (m / monday) :location (h / here)) **Περίληψη συστήματος:** more chinese students studying abroad
ΑΠΣΓ	**Αναπαράσταση σημασιολογικού γραφήματος:** say-01 :arg0 (person :arg0-of (have-org-role-91 :arg1 ministry :arg2 official)) :arg1 (study-01 :arg0 (person :mod (country :name (name :op1 china) :wiki china) :quant more) :arg1 educate-01 :location abroad :time (since :time late)) :arg2 monday :location here **Περίληψη συστήματος:** more chinese students abroad study
ΑΠΣΓΧΕ	**Αναπαράσταση σημασιολογικού γραφήματος:** say :arg0 (person :arg0-of (have-org-role :arg1 ministry :arg2 official)) :arg1 (study :arg0 (person :mod (country :name (name :op1 china) :wiki china) :quant more) :arg1 educate :location abroad :time (since :time late)) :arg2 monday :location here **Πιρίληψη συστήματος:** more than one million chinese students studying abroad
Περίληψη αναφοράς: number of chinese students abroad exceeds one million	

Σε μια προσπάθεια περαιτέρω παρουσίασης της ροής εργασίας και της ανάδειξης των κύριων πτυχών που παρουσιάζουν τα επιμέρους βήματα της προτεινόμενης μεθοδολογίας, από το αρχικό κείμενο έως την εκτίμηση των περιλήψεων, οι Πίνακες 7.7 και 7.9 παρουσιάζουν σχετικά παραδείγματα αυτόματης *ΠΚ* σε επίπεδο κειμένου μικρής έκτασης και σε επίπεδο εγγράφου,

αντίστοιχα.

Πίνακας 7.8: Παράδειγμα χρήσης για την *ΠΚ* σε επίπεδο εγγράφου που περιλαμβάνει ένα κείμενο εισόδου και την αντίστοιχη περίληψη αναφοράς.

Κείμενο εισόδου: a family trip to a nebraska zoo turned terrifying for one family after the gorilla they were looking at leaped toward the exhibit window , cracking it . kevin cave caught the incident on video that he posted on his reddit page . it has already been viewed more than 1 million times . cave said when his family first arrived at omaha 's henry doorly zoo gorilla exhibit , he noticed one of the gorillas had a cut below his eye that was " bleeding a little bit . [...] even with the crack , the public was never in danger , he said , because the window has multiple layers of both glass and acrylic . kijoto is a 20 year old western lowland gorilla , according to a release on the zoo 's website . he weighs 375 pounds .
Περίληψη αναφοράς: gorilla leaps toward exhibit window and hits it , sending family running . zoo says patrons were never in danger .

Συγκεκριμένα, μετά την εκτέλεση της εργασίας ανάκτησης των *AMR* γραφημάτων από το αρχικό κείμενο (Ενότητα 6.5), το κείμενο εισόδου μετατρέπεται σε μια σημασιολογική αναπαράσταση σε μορφή γραφήματος που αναπαρίσταται με μια ακολουθία λεκτικών μονάδων. Στην περίπτωση της *ΠΚ* περιορισμένης έκτασης (Πίνακας 7.7), το σημασιολογικό γράφημα του αρχικού κειμένου αντιστοιχεί στην *AMR* αναπαράσταση που ανακτήθηκε από αυτό, καθώς πρόκειται για κείμενο μιας περιόδου. Σύμφωνα με την περίπτωση αυτή, με δεδομένο το *AMR* γράφημα του αρχικού κειμένου, εφαρμόζεται κάποια από τις τεχνικές μετασχηματισμού γραφήματος (Ενότητα 6.7). Από την άλλη πλευρά, στην περίπτωση της *ΠΚ* σε επίπεδο εγγράφου, εφαρμόζεται μεθοδολογία τόσο για την κατασκευή γραφήματος (Ενότητα 6.6) όσο και για τον μετασχηματισμό γραφήματος (Ενότητα 6.7), καθώς σε αυτή την περίπτωση ανακτώνται περισσότερα από ένα *AMR* υπο-γραφήματα που αντιστοιχούν στις περιόδους του κειμένου. Ο μετασχηματισμός γραφήματος οδηγεί στην αναπαράσταση του γραφήματος σε μορφή ακολουθίας λεκτικών μονάδων, η οποία αποτελεί κατάλληλη μορφή για τη φάση των προβλέψεων μηχανικής μάθησης. Το χρησιμοποιούμενο μοντέλο βαθιάς μάθησης (σε αυτή την περίπτωση έχει χρησιμοποιηθεί το μοντέλο *ΚΑΜΠΑΛ* της Ενότητας 6.8.2) εκπαιδεύεται από άκρο-σε-άκρο με χρήση κατάλληλων παραδειγμάτων χρήσης, τα οποία περιλαμβάνουν ζεύγη τύπου σημασιολογικό γράφημα-περίληψη. Στη συνέχεια, το μοντέλο βαθιάς μάθησης είναι σε θέση να εκτιμήσει μια περίληψη (δηλ., την περίληψη συστήματος) ενός νέου κειμένου.

Για την περίπτωση της *ΠΚ* μικρής έκτασης, το αρχικό κείμενο και η περίληψη αναφοράς αναγράφονται στον Πίνακα 7.7. Στην περίπτωση της *ΠΚ* σε επίπεδο εγγράφου, το αρχικό κείμενο και η περίληψη αναφοράς περιλαμβάνονται στον Πίνακα 7.8.

Στο παράδειγμα της *ΠΚ* μικρής έκτασης που παρουσιάζεται στον Πίνακα 7.7, μπορούμε να παρατηρήσουμε ότι οι περιλήψεις συστήματος διαφέρουν μεταξύ των διαφορετικών σχημάτων δεδομένων. Παρατηρούμε ότι η περίληψη βελτιώνεται σταδιακά και λαμβάνει την περισσότερο ενημερωτική μορφή για το σχήμα δεδομένων *ΑΠΣΓΧΕ*. Επιπλέον, γίνεται φανερό ότι οι εκτιμώμενες περιλήψεις περιλαμβάνουν λέξεις που δεν εμφανίζονται στο κείμενο εισόδου. Για παράδειγμα, στην περίπτωση του σχήματος δεδομένων *ΑΠΣΓΧΕ*, οι λέξεις "*number*" και "*exceeds*"

Πίνακας 7.9: Παραδείγματα αυτόματης *ΠΚ* σε επίπεδο εγγράφου που αποτυπώνουν τη ροή εργασίας από το κείμενο εισόδου του Πίνακα 7.8 έως την περίληψη συστήματος για τις προτεινόμενες μεθόδους κατασκευής γραφήματος (*i*) ακολουθία υπο-γραφημάτων (*Α-ΑΠΣΓΧΕ*) και (*ii*) συνδυασμός υπο-γραφημάτων (*Σ-ΑΠΣΓΧΕ*), με εφαρμογή μετασχηματισμού απλοποιημένου σημασιολογικού γραφήματος χωρίς έννοιες (*ΑΠΣΓΧΕ*).

Α-ΑΠΣΓΧΕ	**Αναπαράσταση σημασιολογικού γραφήματος:** nebraska :arg0-of (trip :arg1 family :arg1-of (crack :arg0 (gorilla :arg0-of (turn :arg3 (terrify :arg1 (family :arg0-of (exhibit :arg1 window))) :time (after :op1 (leap))) :arg1-of (look :arg0 they)))) :mod-of zoo [EOG] reddit :poss (he :arg1-of (post :arg0 (cave :arg0-of (catch :arg1 (incident :location video)) :mod kevin) :arg2 page [EOG] view :arg0 1 :arg1 more :time already [EOG] [...] say :arg1 (release :arg1 (website :poss zoo :arg2 (gorilla :domain kijoto :mod lowland :mod (world-region : name (name :op1 west) :wiki western_world) :mod (old :quant (temporal-quantity :quant 20 :unit year [EOG] weigh :arg0 he :arg1 375 [EOG] **Περίληψη συστήματος:** kevin cave captured the incident in henry doorly zoo in nebraska . the gorilla was captured in a video of a cave gorillas leaped toward the exhibit window. the video has already been more than 1 million views .
Σ-ΑΠΣΓΧΕ	**Αναπαράσταση σημασιολογικού γραφήματος:** nebraska :arg0-of (trip :arg1 (family :arg0-of (exhibit :arg1 window)) :arg1-of (crack :arg0 (gorilla :arg0-of (turn :arg3 (terrify :arg1 family) :time (after :op1 leap)) :arg1-of (look :arg0 they) :domain kijoto :mod lowland :mod (world-region :name (name :op1 west) :wiki western_world) :mod (old :quant (temporal-quantity :quant (20 :unit year))))) :mod-of (zoo :arg2 (gorilla)) (say :arg1 (release :arg1 (website :poss zoo))) [...] (reddit :poss (he :arg1-of (post :arg0 (cave :arg0-of (catch :arg1 (incident :location video)) :mod kevin) :arg2 page)))(view :arg0 1 :arg1 more :time already [EOG] **Περίληψη συστήματος:** kevin cave captured the incident in henry doorly zoo in nebraska . the gorilla was spotted terrifying a family at the henry zoo in nebraska . the gorilla was overheard fighting with one another .

δεν υπάρχουν ούτε στο αρχικό κείμενο ούτε στο σημασιολογικό γράφημα εισόδου. Ωστόσο, αυτές οι λέξεις είναι σε συμφωνία με το περιεχόμενο και τη σημασιολογία του αρχικού κειμένου της συγκεκριμένης περίπτωσης. Αυτό δείχνει τη δυναμική του προτεινόμενου πλαισίου να δημιουργεί νέο κείμενο ή να παραφράζει το περιεχόμενο του αρχικού κειμένου, καθώς πρόκειται για μια προσέγγιση αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου.

Στην περίπτωση του παραδείγματος της *ΠΚ* σε επίπεδο εγγράφου, το οποίο παρουσιάζεται στους Πίνακες 7.8 και 7.9, αναφέρονται παραδείγματα εκτίμησης περιλήψεων συστήματος για τις δύο μεθόδους κατασκευής γραφήματος και για το απλοποιημένο σημασιολογικό γράφημα χωρίς έννοιες (*ΑΠΣΓΧΕ*), ως σχήμα μετασχηματισμού γραφήματος. Στην περίπτωση κατασκευής του σημασιολογικού γραφήματος ως ακολουθίας υπο-γραφημάτων (*Α-ΑΠΣΓΧΕ*), παρατηρούμε ότι

οι διαδοχικές αναπαραστάσεις των υπο-γραφημάτων διαχωρίζονται μεταξύ τους με τη λεκτική μονάδα [*EOG*] (τέλος γραφήματος ή υπο-γραφήματος). Ενώ στην περίπτωση συνδυασμού των υπο-γραφημάτων (*Σ-ΑΠΣΓΧΕ*), έχουμε μια ενιαία αναπαράσταση του συνολικού γραφήματος για ολόκληρο το κείμενο. Όπως παρατηρούμε, ενώ οι περιλήψεις που παράγονται διαφέρουν μεταξύ τους, και στις δύο περιπτώσεις, *Α-ΑΠΣΓΧΕ* και *Σ-ΑΟΣΓΧΕ*, οι περιλήψεις περιέχουν τρεις περιόδους. Επίσης, και σε αυτή την περίπτωση, οι περιλήψεις περιέχουν λέξεις ή φράσεις που δεν αναφέρονται στο αρχικό κείμενο, παραφράζοντας το περιεχόμενο του αρχικού κειμένου. Όπως μπορούμε να δούμε, παρά τα λίγα συντακτικά και γραμματικά λάθη, οι περιλήψεις που δημιουργούνται μπορεί να θεωρηθεί ότι αποτυπώνουν ένα μέρος του περιεχομένου του αρχικού κειμένου, το οποίο περιλαμβάνει σημαντικές πληροφορίες σύμφωνα με το περιεχόμενο του κειμένου που δίνεται στην είσοδο.

7.8 Ερμηνεία αποτελεσμάτων

Στην παρούσα ενότητα γίνεται μια προσπάθεια περιγραφής και ερμηνείας των πειραματικών αποτελεσμάτων που παρουσιάστηκαν παραπάνω. Η πειραματική διαδικασία στοχεύει στην εξέταση σημαντικών πτυχών του προτεινόμενου πλαισίου. Συγκεκριμένα, διερευνάται η επίδραση διαφορετικών αρχιτεκτονικών βαθιάς μάθησης σε συνδυασμό με τις προτεινόμενες μεθόδους κατασκευής και μετασχηματισμού γραφήματος (Ενότητες 7.8.1, 7.8.2 και 7.8.3). Στην Ενότητα 7.8.4 περιγράφονται τα αποτελέσματα της μετρικής *NTR* που αφορά τον προσδιορισμό της παρουσίας νέου περιεχομένου στις εκτιμώμενες περιλήψεις. Τέλος, στην Ενότητα 7.8.5 εξετάζεται η συνέπεια απόδοσης πληροφορίας (*ΣΑΠ*) ως μέτρηση που εστιάζει σε μια ποιοτική αξιολόγηση των παραγόμενων περιλήψεων.

7.8.1 Η επίδραση των αρχιτεκτονικών βαθιάς μάθησης

Στο πλαίσιο της προτεινόμενης προσέγγισης, εξετάσαμε πέντε μοντέλα βαθιάς μάθησης (Ενότητα 6.8.2) που περιλαμβάνουν ένα δίκτυο κωδικοποιητή-αποκωδικοποιητή με μηχανισμούς προσοχής και αντιγραφής λέξεων εκτός λεξιλογίου (*ΚΑΜΠΑΛ*), ένα μοντέλο *ΕΜ* και αρχιτεκτονικές που βασίζονται σε μετασχηματιστές (*ΜΣ*, *ΜΣΕΛ* και *ΜΣΠΚ*). Τα πειραματικά αποτελέσματα επιβεβαιώνουν ότι όλα τα εξεταζόμενα μοντέλα επιτυγχάνουν επιδόσεις αιχμής στον τομέα της αυτόματης *ΠΚ* που βασίζεται σε σημασιολογικά γραφήματα τύπου *AMR*. Η προσέγγιση *ΕΜ*, η οποία χρησιμοποιεί ως πράκτορα το δίκτυο *ΚΑΜΠΑΛ*, υπερέχει του μοντέλου *ΚΑΜΠΑΛ* (δηλ., το μοντέλο *ΕΜ* υπερέχει της περίπτωσης που το μοντέλο *ΚΑΜΠΑΛ* χρησιμοποιείται αυτόνομα, εκτός του περιβάλλοντος της *ΕΜ*). Τα θετικά αποτελέσματα της *ΕΜ* μπορεί να αποδοθούν στο γεγονός ότι επιδιώκεται η βελτιστοποίηση της μετρικής $Rouge_L$ που αποτελεί μια συγκεκριμένη μέτρηση για την αυτόματη *ΠΚ*, σε αντίθεση με το μοντέλο *ΚΑΜΠΑΛ* που ελαχιστοποιεί μια συνάρτηση σφάλματος (δηλ., τη συνάρτηση σφάλματος αρνητικής λογαριθμικής πιθανοφάνειας), η οποία δεν αποτελεί μετρική αξιολόγησης για την αυτόματη *ΠΚ*. Το δίκτυο *ΜΣΕΛ*, το οποίο χρησιμοποιεί το μοντέλο *BERT* για την αρχικοποίηση των διανυσματικών αναπαραστάσεων των λεκτικών μονάδων, παρουσιάζει καλύτερα αποτελέσματα από το μοντέλο *ΜΣ*, το οποίο αρχικοποιεί τις ενσωματώσεις λέξεων σε τυχαία διανύσματα. Το μοντέλο *ΜΣΠΚ*, το οποίο βασίζεται σε έναν προ-εκπαιδευμένο κωδικοποιητή τύπου *BERT*, επιτυγχάνει υψηλότερες επιδόσεις

τόσο από τις άλλες αρχιτεκτονικές που βασίζονται σε μετασχηματιστές όσο και από το μοντέλο *ΚΑΜΠΑΛ*. Τέλος, διαπιστώθηκε ότι οι προσεγγίσεις *ΕΜ* και *ΜΣΠΚ* είναι αυτές που ξεχωρίζουν σε σύγκριση με τις υπόλοιπες αρχιτεκτονικές, καθώς επιτυγχάνουν τις καλύτερες επιδόσεις σε όρους μετρικών *Rouge* και *ΣΑΠ*, για τα δύο σύνολα δεδομένων, *Gigaword* και *CNN/DailyMail*.

7.8.2 Η επίδραση των μεθόδων κατασκευής γραφήματος

Οι προτεινόμενη μεθοδολογία κατασκευής γραφήματος (Ενότητα 6.6) εφαρμόζεται στις περιπτώσεις που το αρχικό κείμενο περιέχει περισσότερες από μία προτάσεις, με στόχο τη δημιουργία της σημασιολογικής αναπαράστασης του συνολικού κειμένου σε μορφή γραφήματος, σύμφωνα με τα υπο-γραφήματα που έχουν ανακτηθεί και αντιστοιχούν στις επιμέρους προτάσεις του αρχικού κειμένου. Στην περίπτωση των μοντέλων βαθιάς μάθησης που βασίζονται σε μετασχηματιστές, η ακολουθία των υπο-γραφημάτων, ως μέθοδος κατασκευής γραφήματος, είναι πιο αποτελεσματική από τον συνδυασμό των υπο-γραφημάτων. Από την άλλη πλευρά, οι αρχιτεκτονικές που βασίζονται σε αναδρομικά νευρωνικά δίκτυα (δηλ., δίκτυο *ΚΑΜΠΑΛ* και η αρχιτεκτονική *ΕΜ* που χρησιμοποιεί το δίκτυο *ΚΑΜΠΑΛ* ως πράκτορα) τείνουν να παρουσιάζουν καλύτερες επιδόσεις στην περίπτωση εφαρμογής του συνδυασμού υπο-γραφημάτων, ως μεθοδολογία κατασκευής γραφήματος.

Πιο συγκεκριμένα, σύμφωνα με τις τιμές των μετρικών *Rouge* και *ΣΑΠ* (Πίνακες 7.4 και 7.6), μπορούμε να συμπεράνουμε ότι οι επιδόσεις των δύο μεθόδων κατασκευής γραφήματος είναι συγκρίσιμες. Επομένως, η μέθοδος της ακολουθίας των υπο-γραφημάτων θα μπορούσε να προτιμάται για εφαρμογή στο προτεινόμενο πλαίσιο, καθώς αποτελεί την απλούστερη μέθοδο δημιουργίας ενός σημασιολογικού γραφήματος, χωρίς αυτή να υπολείπεται σε επιδόσεις. Αυτή η τεχνική προτείνεται ως η επικρατέστερη καθώς δεν απαιτεί κάποια συστηματική μεθοδολογία για τη συνένωση ή τον συνδυασμό των επιμέρους υπο-γραφημάτων, αποφεύγοντας το επιπρόσθετο υπολογιστικό κόστος που απαιτείται στη δεύτερη περίπτωση, του συνδυασμού των υπο-γραφημάτων.

7.8.3 Η επίδραση των τεχνικών μετασχηματισμού γραφήματος

Εξετάσαμε τέσσερις τεχνικές μετασχηματισμού γραφήματος (Ενότητα 6.7) που αντιστοιχούν σε τέσσερις εκδόσεις των συνόλων δεδομένων (Ενότητα 7.2). Οι τεχνικές αυτές στοχεύουν στη δημιουργία μιας κατάλληλης αναπαράστασης ενός σημασιολογικού γραφήματος για τη χρήση του ως είσοδο σε ένα μοντέλο μηχανικής μάθησης. Οι μεθοδολογία μετασχηματισμού γραφήματος καλύπτει ένα εύρος μετασχηματισμών από μια αναπαράσταση του σημασιολογικού γραφήματος σε μορφή κειμένου χωρίς κάποια απλοποίηση (*ΑΡΣΓ*) έως μια απλοποιημένη έκδοση ενός σημασιολογικού γραφήματος (*ΑΠΣΓΧΕ*).

Συγκεκριμένα, το μήκος της αναπαράστασης των σημασιολογικών γραφημάτων (πλήθος λεκτικών μονάδων αναπαράστασης γραφήματος) και το μέγεθος του λεξιλογίου (αριθμός διακριτών λεκτικών μονάδων) μειώνονται καθώς η αναπαράσταση γραφήματος απλοποιείται (Πίνακας 7.2). Τα σχήματα δεδομένων, με το ελάχιστο μήκος αναπαράστασης γραφήματος εισόδου και, επίσης, το

ελάχιστο μέγεθος λεξιλογίου επιτυγχάνουν τις καλύτερες επιδόσεις, σύμφωνα με τα πειραματικά αποτελέσματα. Αυτό μπορεί να οφείλεται στο γεγονός ότι ένα μοντέλο μηχανικής μάθησης, τύπου προβλέψεων ακολουθία-σε-ακολουθία, εκπαιδεύεται πιο αποτελεσματικά χρησιμοποιώντας ακολουθίες λεκτικών μονάδων περιορισμένου μήκους και μειωμένου μεγέθους λεξιλογίου. Στην περίπτωση περιορισμού του λεξιλογίου, κάθε λεκτική μονάδα εισόδου αντιστοιχεί σε περισσότερα παραδείγματα χρήσης, τα οποία χρησιμοποιούνται στη φάση εκπαίδευσης ενός μοντέλου βαθιάς μάθησης. Όπως είναι αναμενόμενο, το γεγονός αυτό οδηγεί στη βελτίωση της ακρίβειας των προβλέψεων μηχανικής μάθησης. Τα σχήματα δεδομένων που διατηρούν τα αναγνωριστικά αποσαφήνισης εννοιών (π.χ., ο όρος "*say-01*" στα σχήματα δεδομένων *ΑΡΣΓ* και *ΑΠΣΓ*), παρουσιάζουν αυξημένο λεξιλόγιο (δηλ., περισσότερες λεκτικές μονάδες) για την αναπαράσταση των σημασιολογικών γραφημάτων από εκείνα τα σχήματα δεδομένων που μετατρέπουν τις έννοιες σε λέξεις (π.χ., η έννοια "*say-01*" μετατρέπεται σε "*say*" στα σχήματα δεδομένων *ΑΡΣΓΧΕ* και *ΑΠΣΓΧΕ*). Σύμφωνα με τα παραπάνω, τα πειραματικά αποτελέσματα επιβεβαιώνουν, ότι η πιο αποτελεσματική μέθοδος μετασχηματισμού γραφήματος είναι η *ΑΠΣΓΧΕ* που προτείνεται για χρήση στο προτεινόμενο πλαίσιο, καθώς παρουσιάζει βελτιωμένες επιδόσεις σε σχέση με τα υπόλοιπα σχήματα δεδομένων.

7.8.4 Νέο περιεχόμενο στις παραγόμενες περιλήψεις

Η μετρική *NTR* αποτυπώνει το επίπεδο δημιουργίας νέου περιεχομένου στις παραγόμενες περιλήψεις, καθώς υπολογίζει το ποσοστό των νέων λεκτικών μονάδων μιας εκτιμώμενης περίληψης, οι οποίες δεν περιλαμβάνονται στο αντίστοιχο αρχικό κείμενο. Όπως παρατηρούμε στους Πίνακες 7.3 και 7.4, για τα σύνολα δεδομένων *Gigaword* και *CNN/DailyMail*, αντίστοιχα, η τιμή *NTR* τείνει να είναι αντιστρόφως ανάλογη με το μέγεθος του λεξιλογίου. Τα σχήματα δεδομένων με περισσότερες διακριτές λεκτικές μονάδες (π.χ. *ΑΡΣΓ*) στην αναπαράσταση σημασιολογικών γραφημάτων επιτυγχάνουν υψηλότερες τιμές σε όρους *NTR* από αυτά με μειωμένο μέγεθος λεξιλογίου (π.χ. *ΑΠΣΓΧΕ*). Επίσης, το επίπεδο των τιμών *NTR* διαφέρει μεταξύ των μοντέλων βαθιάς μάθησης για το ίδιο σχήμα δεδομένων. Οι προσεγγίσεις *ΚΑΜΠΑΛ* και *ΕΜ* παρουσιάζουν μικρές διαφορές μεταξύ, σε όρους *NTR*, όταν συγκρίνονται για το ίδιο σχήμα δεδομένων. Τέλος, τα μοντέλα που βασίζονται σε μετασχηματιστές διαπιστώνουμε ότι παρουσιάζουν τη μεγαλύτερη αφαίρεση κατά την παραγωγή των περιλήψεων, παραφράζοντας το περιεχόμενο σε μεγαλύτερο βαθμό από τα άλλα μοντέλα προβλέψεων.

7.8.5 Συνέπεια απόδοσης πληροφορίας

Στο πλαίσιο της αξιολόγησης του προτεινόμενου πλαισίου, αποτιμάται η συνέπεια απόδοσης πληροφορίας (*ΣΑΠ*) (μια μετρική που παρουσιάστηκε στην Ενότητα 7.3) ως μια προσπάθεια ποιοτικής αξιολόγησης των παραγόμενων περιλήψεων, διερευνώντας τον βαθμό που μία περίληψη αποδίδει το περιεχόμενο του αρχικού κειμένου. Σύμφωνα με τα πειραματικά αποτελέσματα (Πίνακες 7.5 και 7.6), στην περίπτωση της *ΠΚ* μικρής έκτασης (δηλ., σύνολο δεδομένων *Gigaword*), οι περιλήψεις που παράγονται παρουσιάζουν υψηλότερες τιμές σε όρους *ΣΑΠ* από την περίπτωση της αυτόματης *ΠΚ* σε επίπεδο εγγράφου (δηλ., σύνολο δεδομένων *CNN/DailyMail*). Αυτό μπορεί να αποδοθεί στο γεγονός ότι στην περίπτωση της *ΠΚ* σύντομων εγγράφων, οι εκτιμώμενες

περιλήψεις καλύπτουν μεγαλύτερο μέρος της παρεχόμενης πληροφορίας του αρχικού κειμένου. Από την άλλη πλευρά, στην περίπτωση της *ΠΚ* σε επίπεδο εγγράφου, οι περιλήψεις δεν είναι ικανές να συμπεριλάβουν επαρκώς το περιεχόμενο του αρχικού κειμένου. Για να διευκρινίσουμε, στην περίπτωση του συνόλου δεδομένων *Gigaword*, το μέσο μήκος ενός κειμένου είναι περίπου τέσσερις φορές μεγαλύτερο από αυτό μιας περίληψης, ενώ στο σύνολο δεδομένων *CNN/DailyMail*, η περίληψη είναι σχεδόν επτά φορές μικρότερη από το κείμενο. Επομένως, όταν ένα κείμενο είναι πολύ μεγαλύτερης έκτασης από την έκταση της περίληψής του, η *ΣΑΠ* αναμένεται να λάβει μειωμένες τιμές λόγω της περιορισμένης πληροφορίας που είναι ικανή να συμπεριλάβει μια παραγόμενη περίληψη.

Παρόμοια με τα πειραματικά αποτελέσματα των μετρικών *Rouge* που αναφέρονται παραπάνω, οι τιμές επίδοσης της μετρικής *ΣΑΠ* μεγιστοποιούνται για τη μέθοδο μετασχηματισμού γραφήματος *ΑΠΣΓΧΕ*. Επιπλέον, σε μια σύγκριση μεταξύ των μεθόδων κατασκευής γραφήματος, οι υψηλότερες τιμές για την μετρική *ΣΑΠ* λαμβάνονται για την περίπτωση της κατασκευής γραφήματος ως ακολουθία υπο-γραφημάτων. Επίσης, τα πειραματικά αποτελέσματα επιβεβαιώνουν ότι τα μοντέλα βαθιάς μάθησης *ΕΜ* και *ΜΣΠΚ* εμφανίζουν τις καλύτερες επιδόσεις σε όρους *ΣΑΠ*.

Όπως παρατηρούμε στα πειραματικά αποτελέσματα, για ορισμένα μοντέλα μηχανικής μάθησης σε συνδυασμό με κάποια σχήματα δεδομένων, οι τιμές μετρήσεων της *ΣΑΠ* των παραγόμενων περιλήψεων φαίνεται να είναι βελτιωμένες σε σύγκριση με αυτή των περιλήψεων αναφοράς. Αυτό δεν υπονοεί ότι οι εκτιμώμενες περιλήψεις είναι καλύτερες από τις αντίστοιχες περιλήψεις αναφοράς. Καθώς οι περιλήψεις που γράφονται από άνθρωπο μπορεί να αποτυπώνουν το περιεχόμενο του αρχικού κειμένου παραφράζοντας το, σε βαθμό τέτοιο που μια μετρική τύπου *ΣΑΠ* αδυνατεί να αποτυπώσει αξιόπιστα την επικάλυψη του περιεχομένου των παραγόμενων περιλήψεων σε σχέση με το αρχικό κείμενο. Αντίθετα, ένα σύστημα αυτόματης *ΠΚ* τείνει να αντιγράφει λέξεις ή φράσεις από ένα αρχικό κείμενο διαμορφώνοντας την παραγόμενη περίληψη και επιτρέποντας στη μετρική *ΣΑΠ* να επιτύχει υψηλές τιμές. Επομένως, οι περιορισμένες τιμές σε όρους μετρικών *ΣΑΠ*, στην περίπτωση των περιλήψεων αναφοράς, δεν υποδηλώνουν ότι αυτές οι περιλήψεις παρουσιάζουν αδυναμίες. Από την άλλη πλευρά, το γεγονός ότι μετρήθηκαν υψηλές τιμές σε όρους *ΣΑΠ*, ειδικά στην περίπτωση της *ΠΚ* σύντομων εγγράφων (σύνολο δεδομένων *Gigaword*), μπορεί να θεωρηθεί ότι αποτελεί μια ισχυρή ένδειξη παραγωγής περιλήψεων από το σύστημα που αποδίδουν επαρκώς το περιεχόμενο του αρχικού κειμένου.

7.9 Συμπεράσματα και μελλοντική εργασία

Σε αυτή την εργασία, παρουσιάστηκε μια νέα προσέγγιση για την αυτόματη *ΠΚ* με τη μέθοδο της παραγωγής κειμένου, η οποία συνδυάζει μεθοδολογία μηχανικής μάθησης και τεχνικές σημασιολογικής αναπαράστασης περιεχομένου. Στο πλαίσιο αυτό, καθορίζεται η μεθοδολογία κατασκευής σημασιολογικών γραφημάτων και παρουσιάζονται τεχνικές μετασχηματισμού των γραφημάτων, που απαιτούνται για την μετατροπή τους σε κατάλληλες αναπαραστάσεις για είσοδο σε ένα από τα προτεινόμενα μοντέλα βαθιάς μάθησης. Η ροή εργασίας περιλαμβάνει την ανάκτηση σημασιολογικών γραφημάτων, την κατασκευή γραφήματος κειμένου, τον μετασχηματισμό ενός σημασιολογικού γραφήματος για τα μοντέλα βαθιάς μάθησης και τις εκτιμήσεις μηχανικής μάθησης. Σε αυτή την κατεύθυνση, εξετάστηκαν διάφορες αρχιτεκτονικές βαθιάς μάθησης για την εκτίμηση

μιας περίληψης, με δεδομένο ένα σημασιολογικό γράφημα ενός κειμένου εισόδου. Στο πλαίσιο της αξιολόγησης, παρουσιάσαμε ένα σύνολο μετρικών που αξιολογεί τη συνέπεια απόδοσης πληροφορίας σε μια προσπάθεια ποιοτικής αξιολόγησης. Διεξήχθη μια εκτεταμένη πειραματική διαδικασία με τη χρήση δύο δημοφιλών συνόλων δεδομένων, προκειμένου να αξιολογηθεί η απόδοση του προτεινόμενου πλαισίου, και εξετάστηκαν διάφορες πτυχές της προσέγγισης αυτής.

Ειδικότερα, η ανάλυση που έχει παρουσιαστεί παραπάνω δείχνει ότι το προτεινόμενο πλαίσιο μπορεί να αποτελέσει μια αποτελεσματική λύση, η οποία αξιοποιεί σημασιολογικά γραφήματα στο πεδίο της αυτόματης *ΠΚ* με τη μέθοδο της παραγωγής κειμένου. Διαπιστώθηκε, μέσα από την πειραματική διαδικασία, ότι η μεθοδολογία που προτείνεται υπερτερεί σε επιδόσεις συγκρινόμενη με άλλες προσεγγίσεις, ειδικά στις περιπτώσεις των μοντέλων *ΕΜ* και των αρχιτεκτονικών που βασίζονται σε μετασχηματιστές. Για την κατασκευή του σημασιολογικού γραφήματος ενός κειμένου, η ακολουθία των υπο-γραφημάτων, ως μέθοδος κατασκευής γραφήματος, αποτελεί την απλούστερη λύση, η οποία φαίνεται να αποδίδει επαρκώς σε σύγκριση με τη δεύτερη μέθοδο του συνδυασμού των υπο-γραφημάτων. Συνεπώς, η κατασκευή γραφήματος ως ακολουθία υπο-γραφημάτων αποτελεί τη μέθοδο που προτείνεται για λόγους αποφυγής της πολύπλοκης και υπολογιστικά απαιτητικής διαδικασίας συγχώνευσης των υπο-γραφημάτων. Για τη δημιουργία της αναπαράσταση ενός σημασιολογικού γραφήματος ως ακολουθία λεκτικών μονάδων, η οποία απαιτείται ως είσοδος σε ένα μοντέλο μηχανικής μάθησης, η μέθοδος μετασχηματισμού γραφήματος, που αντιστοιχεί στην πιο συνοπτική μορφή μιας σημασιολογικής αναπαράστασης γραφήματος και, επίσης, οδηγεί στο ελάχιστο μέγεθος λεξιλογίου, επιτυγχάνει τις καλύτερες επιδόσεις. Η συγκεκριμένη μέθοδος μετασχηματισμού γραφήματος, η οποία προτείνεται, αντιστοιχεί στα απλοποιημένα σημασιολογικά γραφήματα χωρίς έννοιες (*ΑΠΣΓΧΕ*). Σύμφωνα με τα πειραματικά αποτελέσματα, αυτή η μέθοδος (*ΑΠΣΓΧΕ*) είναι η πιο αποτελεσματική. Επιπλέον, οι μετρήσεις που δείχνουν το ποσοστό των νέων λέξεων στις εκτιμώμενες περιλήψεις (NTR) επηρεάζονται από το χρησιμοποιούμενο σχήμα δεδομένων και το μοντέλο μηχανικής εκμάθησης. Οι τιμές της μετρικής NTR μειώνονται όταν μειώνεται το μέγεθος του λεξιλογίου (π.χ., στο σχήμα δεδομένων *ΑΠΣΓΧΕ*) και λαμβάνουν τις υψηλότερες τιμές όταν χρησιμοποιούνται μοντέλα μηχανικής μάθησης που βασίζονται σε μετασχηματιστές. Τέλος, κατά την αξιολόγηση σε όρους *ΣΑΠ*, η μετρική που προτείνεται λαμβάνει τις υψηλότερες τιμές στην περίπτωση της *ΠΚ* μικρής έκτασης σε σύγκριση με την περίπτωση της *ΠΚ* σε επίπεδο εγγράφου. Τα πειραματικά αποτελέσματα, σύμφωνα με την μετρική *ΣΑΠ*, επιβεβαιώνουν ότι οι παραγόμενες περιλήψεις αποδίδουν επαρκώς το περιεχόμενο των κειμένων εισόδου.

Τα θετικά πειραματικά αποτελέσματα που προέκυψαν μπορεί να αποδοθούν στην κατάλληλη σημασιολογική αναπαράσταση των δεδομένων, στις κατάλληλες αρχιτεκτονικές βαθιάς μάθησης, καθώς και στη βελτιστοποίηση των παραμέτρων των μοντέλων μηχανικής μάθησης.

Το προτεινόμενο πλαίσιο παρουσιάζει ικανοποιητικές επιδόσεις συγκρινόμενο με άλλες συναφείς προσεγγίσεις που βασίζονται σε σημασιολογική αναπαράσταση περιεχομένου. Ωστόσο, θα μπορούσε να επεκταθεί περαιτέρω με την εξέταση προοπτικών που μπορεί να περιλαμβάνουν τη βελτίωση της σημασιολογικής αναπαράστασης του κειμένου εισόδου ή την περαιτέρω διερεύνηση αρχιτεκτονικών βαθιάς μάθησης. Ειδικότερα, η σημασιολογική αναπαράσταση του αρχικού κειμένου χρειάζεται να μελετηθεί περαιτέρω, καθώς η ποιότητα των παραγόμενων περιλήψεων βασίζεται στην αναπαράσταση του περιεχομένου. Επιπλέον, η διερεύνηση αρχιτεκτονικών νευρωνικών δικτύων και, ειδικότερα, η διερεύνηση των προ-εκπαιδευμένων γλωσσικών μοντέλων ή μοντέλων

ΕΜ σε συνδυασμό με αρχιτεκτονικές μετασχηματιστών ως πράκτορες στο περιβάλλον της *ΕΜ*, θα μπορούσε να επιφέρει περαιτέρω βελτίωση στην αυτόματη *ΠΚ* και, πιο συγκεκριμένα, στην αυτόματη *ΠΚ* που βασίζεται σε σημασιολογική αναπαράσταση περιεχομένου.

Κεφάλαιο 8

Γενικά συμπεράσματα και μελλοντικές κατευθύνσεις έρευνας

8.1 Γενικά συμπεράσματα

Στο πλαίσιο της παρούσας διδακτορικής διατριβής, αρχικά, εξετάστηκαν αρχιτεκτονικές νευρωνικών δικτύων βαθιάς μάθησης για την αυτόματη περίληψη κειμένου και, στη συνέχεια, παρουσιάστηκε νέα μεθοδολογία που συνδυάζει μηχανική μάθηση και σημασιολογικές τεχνικές με σκοπό την περαιτέρω βελτίωση των παραγόμενων περιλήψεων. Ιδιαίτερη έμφαση δόθηκε στην ανάπτυξη μεθοδολογίας σημασιολογικών μετασχηματισμών του περιεχομένου με κύριο σκοπό την αντιμετώπιση του προβλήματος της διαχείρισης των νέων υποψήφιων για περίληψη κειμένων, που ενδεχομένως περιλαμβάνουν περιεχόμενο χωρίς επαρκή παρουσία στο σύνολο εκπαίδευσης ενός μοντέλου μηχανικής μάθησης. Εξίσου σημαντική θεωρείται η διερεύνηση της προοπτικής σχετικά με τη σημασιολογική αναπαράσταση του περιεχομένου σε μορφή γραφήματος, καθώς η προσπάθεια αυτή εστιάζει στην οργάνωση της μη δομημένης πληροφορίας κειμένου σε μια συνοπτική και αναγνώσιμη από τις μηχανές μορφή, η οποία οδηγεί στη παραγωγή περιλήψεων με σημασιολογική συνάφεια περιεχομένου.

Η ερευνητική διαδικασία, σχετικά με τη διερεύνησή των μοντέλων μηχανικής μάθησης για την αυτόματη *ΠΚ*, ανέδειξε σημαντικές πτυχές, πλεονεκτήματα ή αδυναμίες των εξεταζόμενων προσεγγίσεων βαθιάς μάθησης. Οι αρχιτεκτονικές των νευρωνικών δικτύων που ξεχωρίζουν για τις επιδόσεις τους στον τομέα της αυτόματης *ΠΚ* είναι αυτές της *ΕΜ* και των μοντέλων που βασίζονται σε μετασχηματιστές, όπως έχει αναφερθεί με λεπτομέρεια στο Κεφάλαιο 3.

Το νέο πλαίσιο σημασιολογικών μετασχηματισμών του περιεχομένου, που παρουσιάστηκε (Κεφάλαιο 4), αξιοποιεί μεθοδολογία μηχανικής μάθησης για την εκτίμηση περιλήψεων με τη μέθοδο της παραγωγής κειμένου. Ειδικότερα, το πλαίσιο που προτείνεται βασίζεται κυρίως σε πόρους γνώσης, σε ιεραρχικές ταξινομίες εννοιών, σε μεθοδολογία αποσαφήνισης έννοιας λέξεων, σε αναγνώριση ονοματικών οντοτήτων και σε μεθοδολογία αντιστοίχισης εννοιών, για να επιτευχθούν οι σημασιολογικοί μετασχηματισμοί του περιεχομένου. Συνολικά, η μεθοδολογία αντιμετωπίζει το πρόβλημα των λέξεων εκτός λεξιλογίου ή των σπάνιων λέξεων και στοχεύει στη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης για την αυτόματη *ΠΚ*. Μέσα

από μία εκτεταμένη πειραματική διαδικασία αναδείχτηκαν σημαντικές πτυχές της προτεινόμενης μεθοδολογίας και προσδιορίστηκαν οι παράγοντες εκείνοι που μπορούν να επιφέρουν βελτίωση των εκτιμώμενων περιλήψεων, όπως αναλυτικά έχει αναφερθεί στο Κεφάλαιο 5, στο οποίο παρουσιάστηκε η πειραματική διαδικασία, τα πειραματικά αποτελέσματα και η ερμηνεία τους, καθώς και τα αναλυτικά συμπεράσματα που προέκυψαν για το προτεινόμενο πλαίσιο.

Ειδικότερα, θα αναφερθούν κάποια από τα σημαντικότερα συμπεράσματα που προέκυψαν σχετικά με το προαναφερόμενο πλαίσιο. Σχετικά με τη μεθοδολογία γενίκευσης του περιεχομένου, θεωρήσαμε δύο σημαντικούς παράγοντες, οι οποίοι αφορούν το επίπεδο σημασιολογικής γενίκευσης των εννοιών και το όριο της συχνότητας εμφάνισης των όρων σε ένα σύνολο εκπαίδευσης. Διαπιστώθηκε ότι τόσο το σημασιολογικό εύρος μιας έννοιας όσο και η συχνότητα εμφάνισης του αντίστοιχου όρου επηρεάζουν τα αποτελέσματα. Ο προσδιορισμός των κατάλληλων τιμών για τις δύο προαναφερόμενες παραμέτρους μπορεί να επιφέρει βελτιώσεις. Επιπλέον, η εφαρμογή της μεθοδολογίας των σημασιολογικών μετασχηματισμών σε συνδυασμό με διαφορετικές αρχιτεκτονικές βαθιάς μάθησης αποκάλυψε τις αδυναμίες ή την υπεροχή των εξεταζόμενων αρχιτεκτονικών. Διαπιστώθηκε ότι το μοντέλο ενισχυτικής μάθησης ή οι προσεγγίσεις που βασίζονται σε αρχιτεκτονικές μετασχηματιστών παρουσιάζουν αυξημένες επιδόσεις σε σύγκριση με άλλες αρχιτεκτονικές ή προσεγγίσεις της σχετικής βιβλιογραφίας.

Επιπροσθέτως, διερευνήθηκε ο βαθμός που οι παραγόμενες περιλήψεις περιλαμβάνουν νέες λέξεις ή φράσεις, οι οποίες δεν έχουν παρουσία στο αρχικό κείμενο. Διαπιστώθηκε ότι μειώνονται οι νέες λέξεις σε μια παραγόμενη περίληψη όταν τα συνώνυμα στο κείμενο είναι περιορισμένα (π.χ., σε εφαρμογή της γενίκευσης του περιεχομένου), ενώ ο αριθμός των νέων λέξεων μεγιστοποιείται στην περίπτωση των βασικών προσεγγίσεων, στις οποίες δεν έχει εφαρμοστεί κάποια στρατηγική γενίκευσης. Σε μια ακόμη μέτρηση για τον προσδιορισμό της ακρίβειας απόδοσης πληροφορίας της εκτιμώμενης περίληψης σε σχέση με το αρχικό κείμενο, στην περίπτωση της περίληψης κειμένου μεγαλύτερης έκτασης (π.χ., έγγραφο 500 λέξεων και περίληψη 100 λέξεων) παρατηρήσαμε υψηλότερες τιμές σε σχέση με την περίπτωση της περίληψης εγγράφων μικρότερης έκτασης (π.χ., έγγραφο 50 λέξεων και περίληψη 10 λέξεων). Αυτό σημαίνει ότι η σημασιολογική επικάλυψη της πληροφορίας που περιλαμβάνει μια περίληψη σε σχέση με το αρχικό κείμενο δεν επηρεάζεται αρνητικά από την αύξηση της έκτασης του κειμένου αλλά αντίθετα οι περιλήψεις μεγαλύτερου σε έκταση κειμένου ανταποκρίνονται σε μεγαλύτερο βαθμό στο περιεχόμενο του αρχικού κειμένου. Ένα τελευταίο συμπέρασμα αφορά την εφαρμογή του προτεινόμενου πλαισίου σε πλήρες αποσαφηνισμένο κείμενο, δηλαδή σε κείμενο που όλοι οι όροι του έχουν προσδιοριστεί σημασιολογικά. Σε αυτή την περίπτωση παρατηρήθηκε μειωμένη απόδοση καθώς τα μοντέλα μηχανικής μάθησης δεν ήταν εφικτό να εκπαιδευτούν επαρκώς λόγω του μεγάλου αριθμού διακριτών εννοιών. Επίσης, με δεδομένο ότι οι όροι του κειμένου που συμμετέχουν στη γενίκευση του περιεχομένου ανήκουν σε συγκεκριμένα μέρη του λόγου, διερευνήθηκε η συμβολή των ουσιαστικών και των ρημάτων και διαπιστώθηκε ότι η συμμετοχή και των δύο μερών του λόγου στην διαδικασία γενίκευση επιφέρει περαιτέρω βελτίωση για την αυτόματη περίληψη κειμένου.

Στη συνέχεια παρουσιάστηκε ένα νέο πλαίσιο που αξιοποιεί τη σημασιολογική αναπαράσταση κειμένου σε μορφή γραφήματος για την εκτίμηση περιλήψεων με χρήση βαθιάς μάθησης (Κεφάλαιο 6). Αυτή είναι μια προσέγγιση που εστιάζει στη δομημένη σημασιολογική αναπαράσταση του αρχικού κειμένου με σκοπό τη βελτίωση των προβλέψεων μηχανικής μάθησης, επιδιώκοντας τη μοντελοποίηση του προβλήματος της αυτόματης *ΠΚ* ως ένα πρόβλημα μάθησης από

γράφημα-σε-κείμενο. Η μεθοδολογία που προτείνεται τόσο στο επίπεδο της σημασιολογικής αναπαράστασης του περιεχομένου όσο και στο επίπεδο των προβλέψεων μηχανικής μάθησης διερευνήθηκε σε βάθος με σκοπό την ανάδειξη των πτυχών εκείνων που οδηγούν στις βέλτιστες επιλογές για την παραγωγή ποιοτικών περιλήψεων. Τα συμπεράσματα που προέκυψαν μέσα από μια καλά οργανωμένη πειραματική διαδικασία έχουν παρουσιαστεί με λεπτομέρεια στο Κεφάλαιο 7.

Προς την κατεύθυνση της παροχής μιας ποιοτικής αξιολόγησης για την αυτόματη ΠΚ, η παρούσα διατριβή παρουσίασε ένα νέο σύνολο μετρικών (Ενότητα 7.3.2), οι οποίες προσδιορίζουν τη συνέπεια απόδοσης πληροφορίας των παραγόμενων περιλήψεων σε σχέση με το αρχικό κείμενο. Το σύνολο αυτό των μετρικών, συμπληρωματικά με άλλες μετρικές που αναφέρονται στη διατριβή, χρησιμοποιούνται στο πειραματικό μέρος που διεξήχθη για την αξιολόγηση του πλαισίου που αξιοποιεί τη σημασιολογικής αναπαράστασης του περιεχομένου σε μορφή γραφήματος (Κεφάλαιο 7). Οι εν λόγω μετρικές παρέχουν μια σταθμισμένη τιμή αξιολόγησης, σύμφωνα με την έκταση του αρχικού κειμένου και της περίληψης συστήματος, προσδιορίζοντας τη σημασιολογική επικάλυψη μεταξύ της πληροφορίας που περιλαμβάνει η παραγόμενη περίληψη σε σχέση με το αρχικό κείμενο. Το νέο σύνολο μετρικών μπορεί να συνεισφέρει στην αξιολόγηση και βελτίωση των συστημάτων αυτόματης περίληψης κειμένου.

Ειδικότερα, ακολουθεί μια σύνοψη των βασικότερων συμπερασμάτων σύμφωνα με την πειραματική διαδικασία που διενεργήθηκε σχετικά με το πλαίσιο σημασιολογικής αναπαράστασης του περιεχομένου σε μορφή γραφήματος (Κεφάλαιο 7). Σχετικά με την κατασκευή του σημασιολογικού γραφήματος ενός κειμένου, προτάθηκαν και διερευνήθηκαν δύο λύσεις, η ακολουθία των υπο-γραφημάτων και ο συνδυασμός των υπο-γραφημάτων. Διευκρινίζεται ότι ένα υπο-γράφημα αντιστοιχεί στο σημασιολογικό γράφημα μιας περιόδου του αρχικού κειμένου. Η ακολουθία των υπο-γραφημάτων, ως μέθοδος κατασκευής γραφήματος, αποτελεί την απλούστερη λύση, η οποία φαίνεται να αποδίδει επαρκώς σε σύγκριση με τη δεύτερη μέθοδο του συνδυασμού των υπο-γραφημάτων. Συνεπώς, η μέθοδος αυτή προτείνεται για λόγους αποφυγής της πολύπλοκης και υπολογιστικά απαιτητικής διαδικασίας συγχώνευσης των υπο-γραφημάτων. Επίσης, προτείνονται λύσεις μετασχηματισμού γραφήματος για τη δημιουργία της αναπαράσταση ενός σημασιολογικού γραφήματος ως ακολουθία λεκτικών μονάδων, η οποία απαιτείται ως είσοδος σε ένα μοντέλο μηχανικής μάθησης. Η μέθοδος μετασχηματισμού γραφήματος, που αντιστοιχεί στην πιο συνοπτική μορφή μιας σημασιολογικής αναπαράστασης γραφήματος και, επίσης, οδηγεί στο ελάχιστο μέγεθος διακριτών λεκτικών μονάδων αναπαράστασης (δηλ., μικρότερο μέγεθος λεξιλογίου), επιτυγχάνει τις καλύτερες επιδόσεις. Επιπλέον, οι μετρήσεις που δείχνουν το ποσοστό των νέων λέξεων στις εκτιμώμενες περιλήψεις επηρεάζονται από το χρησιμοποιούμενο σχήμα δεδομένων και το μοντέλο μηχανικής εκμάθησης. Οι νέες λέξεις σε μια εκτιμώμενη περίληψη μειώνονται όταν μειώνεται το μέγεθος του λεξιλογίου και λαμβάνουν το μεγαλύτερο πλήθος όταν χρησιμοποιούνται μοντέλα μηχανικής μάθησης που βασίζονται σε μετασχηματιστές. Σχετικά με τη διερεύνηση των αρχιτεκτονικών βαθιάς μάθησης, διαπιστώσαμε και εδώ ότι τα μοντέλα ενισχυτικής μάθησης και εκείνα που βασίζονται σε μετασχηματιστές παρουσιάζουν τις καλύτερες επιδόσεις. Επιπροσθέτως, τα πειραματικά αποτελέσματα, σύμφωνα με τον προσδιορισμό της συνέπειας απόδοσης πληροφορίας, επιβεβαιώνουν ότι οι παραγόμενες περιλήψεις αποδίδουν σε ικανοποιητικό βαθμό το περιεχόμενο των αρχικών κειμένων.

Η έρευνα που διεξήχθη στο πλαίσιο της παρούσας διδακτορικής διατριβής, οδήγησε στην ανάπτυξη νέων προσεγγίσεων για την αυτόματη *ΠΚ*, οι οποίες, όπως διαπιστώθηκε, μέσα από

την πειραματική διαδικασία και την ανάλυση, παρουσιάζουν θετικά αποτελέσματα. Συγκεκριμένα, σε αρκετές περιπτώσεις διαπιστώθηκε βελτίωση των επιδόσεων της προτεινόμενης μεθοδολογίας σε σύγκριση είτε με βασικές προσεγγίσεις είτε με άλλες σύγχρονες προσεγγίσεις της σχετικής βιβλιογραφίας. Τα θετικά αποτελέσματα αποδίδονται κυρίως στους κατάλληλους σημασιολογικούς μετασχηματισμούς των δεδομένων, στην κατάλληλη σημασιολογική αναπαράσταση του περιεχομένου, στην αξιοποίηση κατάλληλων αρχιτεκτονικών μηχανικής μάθησης, καθώς και στη βελτιστοποίηση των μοντέλων αυτών. Σε ένα γενικότερο πλαίσιο, θα μπορούσαμε να αναφέρουμε ότι οι θετικές επιδόσεις που καταγράψαμε αποδίδονται κυρίως στην ικανότητα της προτεινόμενης μεθοδολογίας να αντιμετωπίσει θεμελιώδη προβλήματα του πεδίου εφαρμογής των μεθόδων μηχανικής μάθησης. Πιο συγκεκριμένα, τα προβλήματα αυτά αφορούν στη διαχείριση μιας νέας εισόδου ενός συστήματος μηχανικής μάθησης, για την οποία το αντίστοιχο μοντέλο δεν έχει αποκτήσει επαρκή γνώση κατά την διαδικασία εκπαίδευσής του. Στην περίπτωση της αυτόματης *ΠΚ*, το πρόβλημα αυτό αντιστοιχεί στη διαχείριση των άγνωστων ή σπάνιων λέξεων που αντιμετωπίζει ένα μέρος της προτεινόμενης μεθοδολογίας, αυξάνοντας τον αριθμό των παραδειγμάτων χρήσης για κάθε λεκτική μονάδα του συνόλου εκπαίδευσης. Ένα δεύτερο και εξίσου σημαντικό πρόβλημα αφορά τη διαχείριση της μη δομημένης πληροφορίας από ένα υπολογιστικό σύστημα και της αναπαράστασης της πληροφορίας αυτής σε ένα σημασιολογικό πλαίσιο. Σε αυτό το πρόβλημα, επίσης, εστιάζει η διατριβή αυτή, καθώς παρουσιάστηκε ένα πλαίσιο σημασιολογικής αναπαράστασης των δεδομένων εισόδου για την αυτόματη *ΠΚ*, το οποίο οργανώνει τη μη δομημένη πληροφορία και προσδίδει χαρακτηριστικά σημασιολογίας στο περιεχόμενο.

8.2 Μελλοντικές κατευθύνσεις έρευνας

Με δεδομένο ότι το πρόβλημα της αυτόματης *ΠΚ* παραμένει ένα ανοιχτό πρόβλημα, υπάρχει αρκετός διαθέσιμος ερευνητικός χώρος για περαιτέρω έρευνα στον τομέα αυτό. Οι περιοχές έρευνας που αγγίζει η παρούσα διατριβή σχετικά με την αυτόματη *ΠΚ*, όπως είναι τα μοντέλα μηχανικής μάθησης, οι προσεγγίσεις που βασίζονται στη σημασιολογία, ο συνδυασμός μηχανικής μάθησης και σημασιολογίας ή οι μετρικές αξιολόγησης, επιδέχονται περαιτέρω διερεύνηση με σκοπό τη βελτίωση των συστημάτων αυτών. Πέρα από τις συγκεκριμένες κατευθύνσεις για μελλοντική έρευνα που αναφέρθηκαν στις Ενότητες 3.5, 5.8 και 7.9 για την εκάστοτε προτεινόμενη προσέγγιση, στη συνέχεια θα επιχειρήσουμε να αναφέρουμε κάποιες γενικότερες κατευθύνσεις για μελλοντική εργασία.

Σχετικά με τα μοντέλα μηχανικής μάθησης, τα οποία βελτιώνονται διαρκώς, απαιτείται να γίνει ερευνητική δουλειά σχετικά με την ερμηνευσιμότητα των μοντέλων που θα οδηγούσε στην κατανόηση των αρχών εκείνων που διέπουν τα εν λόγω μοντέλα. Αυτό θα οδηγούσε στην ανάπτυξη αρχιτεκτονικών οι οποίες θα μπορούν να προσαρμοστούν αποτελεσματικότερα στην επίλυση του συγκεκριμένου προβλήματος. Μια τέτοια γνώση θα οδηγούσε στην εύστοχη διερεύνηση και ανάδειξη των παραγόντων εκείνων που επηρεάζουν την απόδοση των συστημάτων μηχανικής μάθησης σε συνδυασμό με τη συγκεκριμένη εφαρμογή. Ταυτόχρονα, η διερεύνηση των ορίων που παρουσιάζουν τα συστήματα μηχανικής μάθησης (δηλ., τα όρια της ποιότητας ή της ακρίβειας που μπορούν να επιτύχουν τα συστήματα αυτά κατά τη διαδικασία των προβλέψεων) θα μπορούσε να μας δώσει χρήσιμες απαντήσεις για τον βαθμό αξιοποίησής τους.

Σχετικά με τους σημασιολογικούς μετασχηματισμούς του περιεχομένου που αφορούν ένα μεγάλο μέρος της έρευνας που παρουσιάζεται σε αυτή τη διατριβή, απαιτείται περαιτέρω ερευνητική προσπάθεια σχετικά με την αναπαράσταση των δεδομένων που θα οδηγούσε σε σημασιολογικές αναπαραστάσεις, οι οποίες θα έλυναν θεμελιώδη προβλήματα στην επεξεργασία φυσικής γλώσσας, όπως είναι τα προβλήματα της αμφισημίας, της αναφοράς ή του πλεονασμού. Επίσης, οι σημασιολογικοί μετασχηματισμοί του περιεχομένου, όπως είναι το πλαίσιο γενίκευσης περιεχομένου ή τα σημασιολογικά γραφήματα που παρουσιάστηκαν στο ερευνητικό μέρος αυτής της διατριβής, δεν αφορούν μόνο την μετατροπή από κείμενο σε σημασιολογική αναπαράσταση αλλά και το αντίστροφο. Η διατριβή αυτή αποκάλυψε τις δυσκολίες μετατροπής του γενικευμένου περιεχομένου σε συγκεκριμένο αλλά και τους περιορισμούς της μετατροπής των σημασιολογικών γραφημάτων σε κείμενο. Συνεπώς, η διερεύνηση και σχεδίαση κατάλληλων σημασιολογικών μετασχηματισμών, οι οποίοι είναι αμφίδρομοι, θα οδηγούσε σε βελτίωση τόσο σε προσεγγίσεις της αυτόματης *ΠΚ*, όπως αυτές που παρουσιάστηκαν στην παρούσα διατριβή, όσο και σε άλλες εφαρμογές της επεξεργασίας φυσικής γλώσσας που βασίζονται στη σημασιολογία.

Στο επίπεδο της αξιολόγησης των προσεγγίσεων του πεδίου της αυτόματης *ΠΚ*, υπάρχει έλλειψη αξιόπιστων μετρικών, οι οποίες θα μπορούσαν να οδηγήσουν σε κατακόρυφη βελτίωση της επίδοσης αυτών των συστημάτων. Παρά τις φιλότιμες προσπάθειες έρευνας που γίνονται στο πεδίο αυτό, όπως και η παρούσα εργασία παρουσιάζει ένα νέο σύνολο μετρικών (Ενότητα 7.3.2), καμία μετρική αυτόματης αξιολόγησης δεν μπορεί να συγκριθεί με την αξιολόγηση από άνθρωπο, η οποία, όμως, είναι μια απαιτητική και δύσκολα εφαρμόσιμη διαδικασία. Να σημειωθεί ότι η αξιολόγηση από άνθρωπο θεωρείται αξιόπιστη εφόσον χρησιμοποιηθεί σωστά αλλά παρουσιάζει δυσκολίες στην εφαρμογή της, καθώς απαιτείται ένας αριθμός αξιολογητών, οι οποίοι γνωρίζουν σε βάθος τη συγκεκριμένη γλώσσα (π.χ., αξιολόγηση παραδειγμάτων χρήσης στη μητρική τους γλώσσα) και αξιολογούν ένα σύνολο παραδειγμάτων χρήσης σύμφωνα με κάποια προκαθορισμένα κριτήρια. Αυτή η διαδικασία απαιτεί χρόνο και πόρους (π.χ., χώρος εργασίας, συντονισμός ομάδας, χρηματική αποζημίωση των αξιολογητών κλπ.). Η δημιουργία κατάλληλων μετρικών αξιολόγησης για την αυτόματης *ΠΚ*, αρχικά, απαιτεί να διερευνηθούν διεξοδικά και σε βάθος οι παράγοντες εκείνοι, ο οποίοι θα μπορούσαν να είναι μετρήσιμοι και μας δίνουν την πληροφορία για τον βαθμό που μία περίληψη ικανοποιεί τον προορισμό της. Δηλαδή, σε πρώτη φάση, θα βοηθούσε η διερεύνηση και η καταγραφή των παραγόντων που προσδιορίζουν την ποιότητα των παραγόμενων περιλήψεων. Στη συνέχεια, η γνώση αυτών των παραγόντων θα οδηγούσε στην ανάπτυξη καταλληλότερων μετρικών αξιολόγησης για την αυτόματη *ΠΚ*. Συνεπώς, η κατεύθυνση έρευνας που αφορά την περαιτέρω διερεύνηση των μετρικών αξιολόγησης αποτελεί βασική συνιστώσα συνεισφοράς στο πεδίο της αυτόματης *ΠΚ*.

Ένα ακόμη θέμα μελλοντικής εργασίας, το οποίο, σύμφωνα με τη βιβλιογραφική μελέτη που έγινε στο πλαίσιο αυτής της διατριβής, δεν περιλαμβάνεται στις κατεύθυνσης έρευνας του συγκεκριμένου πεδίου γνώσης, είναι ο συνδυασμός των προσεγγίσεων της αυτόματης *ΠΚ* με το πεδίο της αλληλεπίδρασης ανθρώπου - μηχανής. Αυτός ο συνδυασμός των δύο πεδίων γνώσης θα μπορούσε να διευρύνει τα όρια επίδοσης των συστημάτων αυτόματης *ΠΚ*, με τη βελτίωση των επιδόσεων τους μέσα από την αλληλεπίδραση των αυτόματων συστημάτων με τον άνθρωπο. Σε αυτή την περίπτωση, ο χρήστης θα χρειαζόταν να καταβάλει πολύ λιγότερη προσπάθεια σε σχέση με αυτή που θα απαιτούσε η συγγραφή μιας περίληψης από τον ίδιο. Η κατεύθυνση αυτή αποτελεί πρόκληση για περαιτέρω έρευνα, καθώς θα μπορούσε να αποτελέσει τον ενδιάμεσο σταθμό παραγωγής περιλήψεων υψηλής ποιότητας, πριν αυτό επιτευχθεί, ή αν ποτέ επιτευχθεί, από τα

πλήρως αυτοματοποιημένα συστήματα.

Σύμφωνα με τα παραπάνω, ο χώρος της αυτόματης *ΠΚ* αναμένει ερευνητική δουλειά σε πολλούς τομείς του με σκοπό την εξέλιξή του και, ίσως, την ανάδειξη προσεγγίσεων που θα ξεχωρίσουν. Μια δυσκολία που συναντά κανείς στον χώρο αυτό είναι η απαίτηση εμπλοκής διαφορετικών κατευθύνσεων έρευνας της επεξεργασίας φυσικής γλώσσας (π.χ., παραγωγή φυσικής γλώσσας, αναγνώριση ονοματικών οντοτήτων, αποσαφήνιση έννοιας λέξεων κλπ.) για τη δημιουργία προσεγγίσεων σύνθεσης των επιθυμητών περιλήψεων. Η ενασχόληση με ένα τόσο ευρύ πεδίο γνώσης δεν λειτούργησε αποθαρρυντικά αλλά αποτέλεσε πρόκληση για την ερευνητική δουλειά που διεξήχθη στο πλαίσιο αυτής της διατριβής. Συνεπώς, μια ευρεία γνώση του πεδίου της επεξεργασίας φυσικής γλώσσας, η δεδομένη χρησιμότητα εφαρμογής της αυτόματης *ΠΚ* και ο ιδιαίτερα ανοιχτός ερευνητικός χώρος θα μπορούσαν να αποτελέσουν λόγους ενθάρρυνσης για περαιτέρω έρευνα στο πεδίο της αυτόματης *ΠΚ*.

Βιβλιογραφία

[1] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 2017.

[2] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[3] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[4] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining text data*, pp. 43–76, Springer, 2012.

[5] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[6] J.-g. Yao, X. Wan, and J. Xiao, "Recent advances in document summarization," *Knowledge and Information Systems*, vol. 53, no. 2, pp. 297–336, 2017.

[7] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[8] A. Joshi, E. Fernández, and E. Alegre, "Deep learning based text summarization: Approaches databases and evaluation measures," in *International Conference of Applications of Intelligent Systems*, 2018.

[9] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Systems with Applications*, vol. 121, pp. 49 – 65, 2019.

[10] M. Bhandari, P. Gour, A. Ashfaq, P. Liu, and G. Neubig, "Re-evaluating evaluation in text summarization," *arXiv preprint arXiv:2010.07100*, 2020.

[11] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 04 2021.

[12] T. M. Mitchell and T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997.

[13] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168, 2006.

[14] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Springer, 2016.

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[16] K. Gurney, *An introduction to neural networks*. CRC press, 2018.

[17] B. Ding, H. Qian, and J. Zhou, "Activation functions and their characteristics in deep neural networks," in *2018 Chinese control and decision conference (CCDC)*, pp. 1836–1841, IEEE, 2018.

[18] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[19] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[20] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, no. 8, p. 2, 2012.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] R. Rojas, "The backpropagation algorithm," in *Neural networks*, pp. 149–182, Springer, 1996.

[23] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[24] K. M. Tarwani and S. Edem, "Survey on recurrent neural network in natural language processing," *Int. J. Eng. Trends Technol*, vol. 48, pp. 301–304, 2017.

[25] E. D. Liddy, "Natural language processing," *In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.*, 2001.

[26] K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[27] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, pp. 1–32, 2022.

[28] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," *arXiv preprint arXiv:2003.01200*, 2020.

[29] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," *arXiv preprint arXiv:2010.00711*, 2020.

[30] E. D. Liddy, "Enhanced text retrieval using natural language processing," *Bulletin of the American Society for Information Science and Technology*, vol. 24, no. 4, pp. 14–16, 1998.

[31] S. Feldman, "Nlp meets the jabberwocky: Natural language processing in information retrieval," *ONLINE-WESTON THEN WILTON-*, vol. 23, pp. 62–73, 1999.

[32] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[33] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*, 1992.

[34] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, no. 1, pp. 45–55, 1992.

[35] P. Willett, "The porter stemming algorithm: then and now," *Program*, 2006.

[36] A. G. Jivani *et al.*, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[37] V. Balakrishnan and L.-Y. Ethel, "Stemming and lemmatization: A comparison of retrieval performances," *Lecture Notes on Software Engineering*, vol. 2, pp. 262–267, 01 2014.

[38] D. Kumawat and V. Jain, "Pos tagging approaches: a comparison," *International Journal of Computer Applications*, vol. 118, no. 6, 2015.

[39] S. G. Kanakaraddi and S. S. Nandyal, "Survey on parts of speech tagger techniques," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1–6, IEEE, 2018.

[40] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, and P. Szekely, "Dimensions of commonsense knowledge," *Knowledge-Based Systems*, vol. 229, p. 107347, 2021.

[41] Y. Xie and P. Pu, "How commonsense knowledge helps with natural language tasks: a survey of recent resources and methodologies," *arXiv preprint arXiv:2108.04674*, 2021.

[42] A. A. Abubakar, "A survey on knowledge and commonsense reasoning for natural language processing," *Scientific and practical cyber security journal*, 2022.

[43] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[44] C. Fellbaum, *WordNet: An electronic lexical database*. MIT press, 1998.

[45] "Princeton university." https://www.princeton.edu/. Accessed: 2021-11-01.

[46] "Wordnet." https://wordnet.princeton.edu/. Accessed: 2021-11-01.

[47] R. Navigli, "Word sense disambiguation: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 2, pp. 1–69, 2009.

[48] A. R. Pal and D. Saha, "Word sense disambiguation: A survey," *arXiv preprint arXiv:1508.01346*, 2015.

[49] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in *Proceedings of the 5th annual international conference on Systems documentation*, pp. 24–26, 1986.

[50] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *Ijcai*, vol. 3, pp. 805–810, 2003.

[51] A. Raganato, C. D. Bovi, and R. Navigli, "Neural sequence learning models for word sense disambiguation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1156–1167, 2017.

[52] S. Nithyanandan and C. Raseek, "Deep learning models for word sense disambiguation: A comparative study," in *Proceedings of the International Conference on Systems, Energy & Environment (ICSEE), Kerala, India*, 2019.

[53] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[54] G. K. Palshikar, "Techniques for named entity recognition: a survey," in *Bioinformatics: Concepts, Methodologies, Tools, and Applications*, pp. 400–426, IGI Global, 2013.

[55] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," *Computer Science Review*, vol. 29, pp. 21–43, 2018.

[56] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," *arXiv preprint arXiv:1910.11470*, 2019.

[57] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[58] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, 2014.

[59] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 93–98, 2016.

[60] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, (Berlin, Germany), pp. 280–290, Association for Computational Linguistics, Aug. 2016.

[61] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[62] S. Song, H. Huang, and T. Ruan, "Abstractive text summarization using lstm-cnn based deep learning," *Multimedia Tools and Applications*, pp. 1–19, 2018.

[63] Y. Gao, Y. Wang, L. Liu, Y. Guo, and H. Huang, "Neural abstractive summarization fusing by global generative topics," *Neural Computing and Applications*, vol. 32, no. 9, pp. 5049–5058, 2020.

[64] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9815–9822, 2019.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[66] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," *arXiv preprint arXiv:1908.08345*, 2019.

[67] Y. You, W. Jia, T. Liu, and W. Yang, "Improving abstractive document summarization with salient information modeling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2132–2141, 2019.

[68] S. Xu, H. Li, P. Yuan, Y. Wu, X. He, and B. Zhou, "Self-attention guided copy mechanism for abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1355–1362, 2020.

[69] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

[70] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[71] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[72] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," *arXiv preprint arXiv:1804.05685*, 2018.

[73] J. Lin, X. Sun, S. Ma, and Q. Su, "Global encoding for abstractive summarization," *arXiv preprint arXiv:1805.03989*, 2018.

[74] H. Zhang, J. Xu, and J. Wang, "Pretraining-based natural language generation for text summarization," *arXiv preprint arXiv:1902.09243*, 2019.

[75] J. Pilault, R. Li, S. Subramanian, and C. Pal, "On extractive and abstractive neural document summarization with transformer language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9308–9319, 2020.

[76] Y. Li, "Deep reinforcement learning," *arXiv preprint arXiv:1810.06339*, 2018.

[77] Y. Keneshloo, T. Shi, N. Ramakrishnan, and C. K. Reddy, "Deep reinforcement learning for sequence-to-sequence models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2469–2489, 2020.

[78] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *International Conference on Learning Representations*, 2018.

[79] A. Celikyilmaz, A. Bosselut, X. He, and Y. Choi, "Deep communicating agents for abstractive summarization," *arXiv preprint arXiv:1803.10357*, 2018.

[80] R. Pasunuru and M. Bansal, "Multi-reward reinforced summarization with saliency and entailment," *arXiv preprint arXiv:1804.06451*, 2018.

[81] Y. Li and T. Yang, *Word Embedding for Understanding Natural Language: A Survey*, pp. 83–104. Cham: Springer International Publishing, 2018.

[82] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[83] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[84] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[85] F. Almeida and G. Xexéo, "Word embeddings: A survey," *arXiv preprint arXiv:1901.09069*, 2019.

[86] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," *arXiv preprint arXiv:2003.07278*, 2020.

[87] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[88] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 273–278, IEEE, 2013.

[89] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with lstm," in *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, pp. 850–855 vol.2, Sep. 1999.

[90] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.

[91] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[92] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[93] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[94] N. Watt and M. C. du Plessis, "Dropout algorithms for recurrent neural networks," in *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, pp. 72–78, ACM, 2018.

[95] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[96] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Audio chord recognition with recurrent neural networks.," in *ISMIR*, pp. 335–340, Citeseer, 2013.

[97] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," *arXiv preprint arXiv:1506.03134*, 2015.

[98] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," *arXiv preprint arXiv:1601.04811*, 2016.

[99] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, 2017.

[100] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[101] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[102] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[103] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

[104] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[105] C. Napoles, M. Gormley, and B. Van Durme, "Annotated gigaword," in *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pp. 95–100, Association for Computational Linguistics, 2012.

[106] P. Over, H. Dang, and D. Harman, "Duc in context," *Information Processing & Management*, vol. 43, no. 6, pp. 1506–1520, 2007.

[107] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in neural information processing systems*, pp. 1693–1701, 2015.

[108] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequence-to-sequence models," *arXiv preprint arXiv:1812.02303*, 2018.

[109] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[110] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, pp. 1310–1318, 2013.

[111] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: a theoretical and experimental comparison.," in *Interspeech*, vol. 13, pp. 1756–1760, 2013.

[112] K. Knight and D. Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artificial Intelligence*, vol. 139, no. 1, pp. 91–107, 2002.

[113] T. Cohn and M. Lapata, "Sentence compression beyond word deletion," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 137–144, Association for Computational Linguistics, 2008.

[114] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," *Computational Linguistics*, vol. 31, no. 3, pp. 297–328, 2005.

[115] E. Marsi and E. Krahmer, "Explorations in sentence fusion," in *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, 2005.

[116] K. Filippova and M. Strube, "Sentence fusion via dependency graph compression," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 177–185, Association for Computational Linguistics, 2008.

[117] K. Filippova, "Multi-sentence compression: Finding shortest paths in word graphs," in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 322–330, Association for Computational Linguistics, 2010.

[118] N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," in *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*, pp. 1–7, IEEE, 2016.

[119] M. J. Mohan, C. Sunitha, A. Ganesh, and A. Jaya, "A study on ontology based abstractive summarization," *Procedia Computer Science*, vol. 87, pp. 32–37, 2016.

[120] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," *Web Semantics: science, services and agents on the world wide web*, vol. 7, no. 3, pp. 154–165, 2009.

[121] C.-S. Lee, Z.-W. Jian, and L.-K. Huang, "A fuzzy ontology and its application to news summarization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 5, pp. 859–880, 2005.

[122] L. Hennig, W. Umbrath, and R. Wetzker, "An ontology-based approach to text summarization," in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, pp. 291–294, IEEE, 2008.

[123] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, and S. Shah, "Multi-document summarization based on the yago ontology," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6976–6984, 2013.

[124] P. Hípola, J. A. Senso, A. Leiva-Mederos, and S. Domínguez-Velasco, "Ontology-based text summarization. the case of texminer," *Library Hi Tech*, vol. 32, no. 2, pp. 229–248, 2014.

[125] A. Khan, N. Salim, H. Farman, M. Khan, B. Jan, A. Ahmad, I. Ahmed, and A. Paul, "Abstractive text summarization based on improved semantic graph approach," *International Journal of Parallel Programming*, pp. 1–25, 2018.

[126] M. Joshi, H. Wang, and S. McClean, "Dense semantic graph and its application in single document summarisation," in *Emerging Ideas on Information Filtering and Retrieval*, pp. 55–67, Springer, 2018.

[127] I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive text summarization," in *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*, pp. 132–138, IEEE, 2012.

[128] P.-E. Genest and G. Lapalme, "Framework for abstractive summarization using text-to-text generation," in *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pp. 64–73, Association for Computational Linguistics, 2011.

[129] S. Alshaina, A. John, and A. G. Nath, "Multi-document abstractive summarization based on predicate argument structure," in *Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2017 IEEE International Conference on*, pp. 1–6, IEEE, 2017.

[130] J. Zhang, Y. Zhou, and C. Zong, "Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1842–1853, 2016.

[131] G. Stratogiannis, P. Kouris, G. Alexandridis, G. Siolas, G. Stamou, and A. Stafylopatis, "Semantic enrichment of documents: a classification perspective for ontology-based imbalanced semantic descriptions," *Knowledge and Information Systems*, pp. 1–39, 2021.

[132] P. P. Borah, G. Talukdar, and A. Baruah, "Approaches for word sense disambiguation–a survey," *International Journal of Recent Technology and Engineering*, vol. 3, no. 1, pp. 35–38, 2014.

[133] R. Navigli, "A quick tour of word sense disambiguation, induction and related approaches," in *International Conference on Current Trends in Theory and Practice of Computer Science*, pp. 115–129, Springer, 2012.

[134] D. S. Chaplot and R. Salakhutdinov, "Knowledge-based word sense disambiguation using topic models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[135] A. Raganato, J. Camacho-Collados, and R. Navigli, "Word sense disambiguation: A unified evaluation framework and empirical comparison," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 99–110, 2017.

[136] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," *arXiv preprint arXiv:1712.09405*, 2017.

[137] X. Rong, "word2vec parameter learning explained," *arXiv preprint arXiv:1411.2738*, 2014.

[138] P. Kouris, I. Varlamis, and G. Alexandridis, "A package recommendation framework based on collaborative filtering and preference score maximization," in *International Conference on Engineering Applications of Neural Networks*, pp. 477–489, Springer, 2017.

[139] P. Kouris, I. Varlamis, G. Alexandridis, and A. Stafylopatis, "A versatile package recommendation framework aiming at preference score maximization," *Evolving Systems*, pp. 1–19, 2018.

[140] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive text summarization: enhancing sequence to sequence models using word sense disambiguation and semantic content generalization," *Computational Linguistics*, pp. 1–41, 2021.

[141] B. B. Hansen and S. O. Klopfer, "Optimal full matching and related designs via network flows," *Journal of computational and Graphical Statistics*, vol. 15, no. 3, pp. 609–627, 2006.

[142] S. Dasgupta, C. H. Papadimitriou, and U. V. Vazirani, *Algorithms*. McGraw-Hill, 2008.

[143] T. Brunsch, K. Cornelissen, B. Manthey, and H. Röglin, "Smoothed analysis of belief propagation for minimum-cost flow and matching," in *International Workshop on Algorithms and Computation*, pp. 182–193, Springer, 2013.

[144] D. P. Bertsekas, *Network optimization: continuous and discrete models*. Belmont: Athena Scientific., 1998.

[145] P. Kovács, "Minimum-cost flow algorithms: an experimental evaluation," *Optimization Methods and Software*, vol. 30, no. 1, pp. 94–127, 2015.

[146] J. Matousek and B. Gärtner, *Understanding and using linear programming*. Springer Science & Business Media, 2007.

[147] A. D. Mishra and D. Garg, "Selection of best sorting algorithm," *International Journal of intelligent information Processing*, vol. 2, no. 2, pp. 363–368, 2008.

[148] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," *arXiv preprint arXiv:1910.12840*, 2019.

[149] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh, "Assessing the factual accuracy of generated text," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 166–175, 2019.

[150] O. Lassila, R. R. Swick, W. Wide, and W. Consortium, "Resource description framework (rdf) model and syntax specification," 1998.

[151] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 344–354, 2015.

[152] "Spacy part-of-speech tagger." https://spacy.io/usage/linguistic-features#pos-tagging. Accessed: 2021-11-20.

[153] "Spacy named entity recognition." https://spacy.io/usage/linguistic-features#named-entities. Accessed: 2021-11-20.

[154] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, *et al.*, "Ontonotes release 5.0 ldc2013t19," *Linguistic Data Consortium, Philadelphia, PA*, vol. 23, 2013.

[155] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1373–1378, 2015.

[156] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," *arXiv preprint arXiv:1603.06042*, 2016.

[157] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *International conference on intelligent text processing and computational linguistics*, pp. 136–145, Springer, 2002.

[158] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, pp. 957–966, 2015.

[159] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[160] R. Nallapati, B. Xiang, and B. Zhou, "Sequence-to-sequence rnns for text summarization," *CoRR*, vol. abs/1602.06023, 2016.

[161] K. Song, B. Wang, Z. Feng, L. Ren, and F. Liu, "Controlling the amount of verbatim copying in abstractive summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[162] Y. Yan, W. Qi, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training," *arXiv preprint arXiv:2001.04063*, 2020.

[163] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," *arXiv preprint arXiv:1805.11080*, 2018.

[164] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," *arXiv preprint arXiv:1808.10792*, 2018.

[165] E. Sharma, L. Huang, Z. Hu, and L. Wang, "An entity-driven framework for abstractive summarization," *arXiv preprint arXiv:1909.02059*, 2019.

[166] T. Sakai, "Two sample t-tests for ir evaluation: Student or welch?," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1045–1048, 2016.

[167] Y. Zhang, "Evaluating the factual correctness for abstractive summarization," *CS230 Project*, 2019.

[168] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation for sembanking," in *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pp. 178–186, 2013.

[169] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[170] J. Flanigan, S. Thomson, J. G. Carbonell, C. Dyer, and N. A. Smith, "A discriminative graph-based parser for the abstract meaning representation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1436, 2014.

[171] J. Zhou, F. Xu, H. Uszkoreit, W. Qu, R. Li, and Y. Gu, "Amr parsing with an incremental joint model," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 680–689, 2016.

[172] C. Wang, S. Pradhan, X. Pan, H. Ji, and N. Xue, "Camr at semeval-2016 task 8: An extended transition-based amr parser," in *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pp. 1173–1178, 2016.

[173] I. Konstas, S. Iyer, M. Yatskar, Y. Choi, and L. Zettlemoyer, "Neural amr: Sequence-to-sequence models for parsing and generation," *arXiv preprint arXiv:1704.08381*, 2017.

[174] X. Peng, C. Wang, D. Gildea, and N. Xue, "Addressing the data sparsity issue in neural amr parsing," *arXiv preprint arXiv:1702.05053*, 2017.

[175] S. Takase, J. Suzuki, N. Okazaki, T. Hirao, and M. Nagata, "Neural headline generation on abstract meaning representation," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1054–1059, 2016.

[176] A. Vlachos *et al.*, "Guided neural language generation for abstractive summarization using abstract meaning representation," *arXiv preprint arXiv:1808.09160*, 2018.

[177] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive text summarization based on deep learning and semantic content generalization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 5082–5092, Association for Computational Linguistics, July 2019.

[178] H. Jin, T. Wang, and X. Wan, "Semsum: Semantic dependency guided neural abstractive summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8026–8033, 2020.

[179] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[180] K. Sindhu and K. Seshadri, "Text summarization: A technical overview and research perspectives," *Handbook of Intelligent Computing and Optimization for Sustainable Development*, pp. 261–286, 2022.

[181] D. Suleiman and A. Awajan, "Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges," *Mathematical problems in engineering*, vol. 2020, 2020.

[182] S. Dohare, H. Karnick, and V. Gupta, "Text summarization using abstract meaning representation," *arXiv preprint arXiv:1706.01678*, 2017.

[183] F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith, "Toward abstractive summarization using semantic representations," *arXiv preprint arXiv:1805.10399*, 2018.

[184] S. Dohare, V. Gupta, and H. Karnick, "Unsupervised semantic abstractive summarization," in *Proceedings of ACL 2018, Student Research Workshop*, pp. 74–83, 2018.

[185] R. Mishra and T. Gayen, "Automatic lossless-summarization of news articles with abstract meaning representation," *Procedia Computer Science*, vol. 135, pp. 178–185, 2018.

[186] J. Flanigan, C. Dyer, N. A. Smith, and J. G. Carbonell, "Generation from abstract meaning representation using tree transducers," in *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 731–739, 2016.

[187] S. Miranda-Jiménez, A. Gelbukh, and G. Sidorov, "Summarizing conceptual graphs for automatic summarization task," in *International Conference on Conceptual Structures*, pp. 245–253, Springer, 2013.

[188] J. F. Sowa, "Conceptual graphs," *Foundations of Artificial Intelligence*, vol. 3, pp. 213–237, 2008.

[189] K. K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon.* University of Pennsylvania, 2005.

[190] G. C. V. Vilca and M. A. S. Cabezudo, "A study of abstractive summarization using semantic representations and discourse level information," in *International Conference on Text, Speech, and Dialogue*, pp. 482–490, Springer, 2017.

[191] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[192] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[193] A. Gatt and E. Reiter, "Simplenlg: A realisation engine for practical applications," in *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pp. 90–93, 2009.

[194] F.-T. Lee, C. Kedzie, N. Verma, and K. McKeown, "An analysis of document graph construction methods for amr summarization," *arXiv preprint arXiv:2111.13993*, 2021.

[195] T. Naseem, A. Blodgett, S. Kumaravel, T. O'Gorman, Y.-S. Lee, J. Flanigan, R. F. Astudillo, R. Florian, S. Roukos, and N. Schneider, "Docamr: Multi-sentence amr representation and evaluation," *arXiv preprint arXiv:2112.08513*, 2021.

[196] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[197] F. Van Harmelen, V. Lifschitz, and B. Porter, *Handbook of knowledge representation.* Elsevier, 2008.

[198] K. Trentelman, "Survey of knowledge representation and reasoning systems," *DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION EDINBURGH (AUSTRALIA)*, 2009.

[199] J. A. Bateman, R. T. Kasper, J. D. Moore, and R. A. Whitney, "A general organization of knowledge for natural language processing: the penman upper model," tech. rep., Technical report, USC/Information Sciences Institute, Marina del Rey, CA, 1990.

[200] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation (amr) 1.0 specification," in *Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL*, pp. 1533–1544, 2012.

[201] "Abstract meaning representation (amr) 1.2.6 specification." https://github.com/amrisi/amr-guidelines/blob/master/amr.md. Accessed: 2022-07-24.

[202] K. Knight, B. Badarau, L. Baranescu, C. Bonial, M. Bardocz, K. Griffitt, U. Hermjakob, D. Marcu, M. Palmer, T. O'Gorman, *et al.*, "Abstract meaning representation (amr) annotation release 3.0," *LDC2020T02. Web Download. Philadelphia: Linguistic Data Consortium*, 2020.

[203] M. Damonte, S. B. Cohen, and G. Satta, "An incremental parser for abstract meaning representation," *arXiv preprint arXiv:1608.06111*, 2016.

[204] W. Foland and J. H. Martin, "Abstract meaning representation parsing using lstm recurrent neural networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 463–472, 2017.

[205] J. Z. Pan, "Resource description framework," in *Handbook on ontologies*, pp. 71–90, Springer, 2009.

[206] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[207] K. Knight, L. Baranescu, C. Bonial, M. Georgescu, K. Griffitt, U. Hermjakob, D. Marcu, M. Palmer, and N. Schneider, "Abstract meaning representation (amr) annotation release 1.0 ldc2014t12," *Web Download. Philadelphia: Linguistic Data Consortium*, 2014.

Συντομογραφίες - Ακρωνύμια

ΑΑΠ	Ακρίβεια απόδοσης πληροφορίας
ΑΕΛ	Αποσαφήνιση έννοιας λέξεων
ΑΛΕΛ	Αντιγραφή λέξεων εκτός λεξιλογίου
ΑΟΟ	Αναγνώριση ονοματικών οντοτήτων
ΑΠΣΓ	Απλοποιημένο σημασιολογικό γράφημα
ΑΠΣΓΧΕ	Απλοποιημένο σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών
ΑΡΣΓ	Αρχικό σημασιολογικό γράφημα
ΑΡΣΓΧΕ	Αρχικό σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών
Α-ΑΠΣΓ	Απλοποιημένο σημασιολογικό γράφημα ως ακολουθία υπο-γραφημάτων
Α-ΑΠΣΓΧΕ	Απλοποιημένο σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών ως ακολουθία υπο-γραφημάτων
Α-ΑΡΣΓ	Αρχικό σημασιολογικό γράφημα ως ακολουθία υπο-γραφημάτων
Α-ΑΡΣΓΧΕ	Αρχικό σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών ως ακολουθία υπο-γραφημάτων
ΕΜ	Ενισχυτική μάθηση
ΕΦΓ	Επεξεργασία φυσικής γλώσσας
ΚΑΜΠ	Κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής
ΚΑΜΠΑΛ	Κωδικοποιητή-αποκωδικοποιητή με μηχανισμούς προσοχής και αντιγραφής λέξεων εκτός λεξιλογίου
ΚΦΓ	Κατανόηση φυσικής γλώσσας
ΛΕΛ	Λέξεις εκτός λεξιλογίου
ΜΜ	Μηχανική μάθηση
ΜΣ	Μετασχηματιστές
ΜΣΕΛ	Μετασχηματιστές με ενσωματώσεις λέξεων που βασίζονται στα συμφραζόμενα
ΜΣΠΚ	Μετασχηματιστές με προ-εκπαιδευμένο κωδικοποιητή
ΠΚ	Περίληψη κειμένου
Σ-ΑΠΣΓ	Απλοποιημένο σημασιολογικό γράφημα ως συνδυασμός υπο-γραφημάτων
Σ-ΑΠΣΓΧΕ	Απλοποιημένο σημασιολογικό γράφημα χωρίς αναγνωριστικά αποσαφήνισης εννοιών ως συνδυασμός υπο-γραφημάτων
Σ-ΑΡΣΓ	Αρχικό σημασιολογικό γράφημα ως συνδυασμός υπο-γραφημάτων
Σ-ΑΡΣΓΧΕ	Αρχικό σημασιολογικό γράφημα χωρίς αναγνωριστικά

	αποσαφήνισης εννοιών ως συδνυασμός υπο-γραφημάτων
ΤΝΔ	Τεχνητά νευρωνικά δίκτυα

ADAM	Adaptive moment estimation
AMR	Abstractive text summarization
ANN	Artificial neural networks
BERT	Bidirectional encoder representations from transformers
BFS	Breadth-first search
CBOW	Continuous bag of words
CNN	Convolutional neural networks
CS	Cosine similarity
DAG	Directed acyclic graph
DFS	Depth first search
ELU	Exponential linear unit
EOG	End of graph
GPU	Graphics processing unit
GRU	Gated recurrent units
JC	Jaccard coefficient
LED	Levenshtein edit distance
LSTM	Long short term memory
ML	Machine learning
NER	Named-entity recognition
NLP	Natural language processing
NTR	New tokens rate
NLU	Natural language understanding
OOV	Out-of-vocabulary
POS	Part-of-speech
PropBank	Proposition bank
RDF	Resource description framework
ReLU	Rectified linear unit
RL	Reinforcement learning
RMSProp	Root mean square propagation
RNN	Recurrent neural network
RST	Rhetorical structure theory
SGD	Stochastic gradient descent
WMD	Word mover's distance
WSD	Word sense disambiguation

Γλωσσάρι

Abstract meaning representation	Αναπαράσταση αφηρημένης έννοιας
Abstractive text summarization	Περίληψη κειμένου με παραγωγή κειμένου
Action	Δράση
Activation function	Συνάρτηση ενεργοποίησης
Adaptive Moment Estimation	Εκτίμηση προσαρμοστικών ροπών
Association analysis	Ανάλυση συσχετίσεων
Attention mechanism	Μηχανισμός προσοχής
Automatic text summarization	Αυτόματη περίληψη κειμένου
Backpropagation error	Οπισθοδιάδοση σφάλματος
Batch gradient descent	Κατάβαση κλίσης στο σύνολο των δεδομένων
Batch size	Μέγεθος δέσμης παραδειγμάτων χρήσης
Beam search	Αλγόριθμος αναζήτησης δέσμης
Bidirectional encoder representations from transformers	Αναπαραστάσεις αμφίδρομου κωδικοποιητή από μετασχηματιστές
Breadth-first search	Αναζήτηση κατά πλάτος
Class balancing	Ισορρόπηση των κλάσεων
Classification	Ταξινόμηση
Clustering	Ομαδοποίηση ή συσταδοποίηση
Coefficient	Συντελεστής
Convolutional neural networks	Συνελικτικά νευρωνικά δίκτυα
Corpora	Σώμα κειμένου
Cosine similarity	Ομοιότητα συνημιτόνου
Coverage mechanism	Μηχανισμός κάλυψης αποφυγής επανάληψης λέξεων
Cross entropy loss function	Συνάρτηση σφάλματος διασταυρούμενης εντροπίας
Depth first search	Αλγόριθμος αναζήτησης κατά βάθος
Directed acyclic graph	Κατευθυνόμενο άκυκλο γράφημα
Dropout	Απόρριψη συνδέσεων μεταξύ κόμβων ενός δικτύου
End-to-end	Άκρο-σε-άκρο
Error function	Συνάρτηση σφάλματος
Exponential linear unit	Εκθετική γραμμική μονάδα
Exposure bias	Έκθεση σε πόλωση ή μεροληψία σε γνωστή

	έξοδο κατά την εκπαίδευση ενός μοντέλου
Extractive text summarization	Περίληψη κειμένου με τη μέθοδο της εξαγωγής κειμένου
Features	Χαρακτηριστικά ή γνωρίσματα
First moment	Ροπή πρώτης τάξης
Fully connected feed-forward network	Πλήρως διασυνδεδεμένο δίκτυο εμπρόσθιας τροφοδότησης
Gated recurrent units	Αναδρομική μονάδα με πύλη
Gloss	Ορισμός έννοιας λέξης
Gradient descent	Κατάβαση κλίσης
Gradient norm clipping	Μέθοδος κατάβασης κλίσης με ψαλιδισμό
Graph to graph	Γράφημα σε γράφημα
Graph to summary	Γράφημα σε περίληψη
Graph to text	Γράφημα σε κείμενο
Graphics processing unit	Μονάδα επεξεργασίας γραφικών
Holonym	Ολώνυμο
Hyperbolic tangent	Υπερβολική εφαπτομένη
Hypernym	Υπερώνυμο
Hyponym	Υπώνυμο
Information items	Στοιχεία πληροφορίας
Intra-attention	Ενδο-προσοχή
Layer normalization	Κανονικοποίηση στρώματος
Lemmatization	Λημματοποίηση
Long short term memory	Μακρά και βραχεία μνήμη
Loss function	Συνάρτηση σφάλματος
Mean squared error	Σφάλμα μέσων τετραγώνων
Meronym	Μερώνυμο
Mini-batch gradient descent	Κατάβαση κλήσης σε δέσμη παραδειγμάτων
Momentum optimization	Βελτιστοποίηση ορμής
Multi-documents text summarization	Περίληψη κειμένου πολλαπλών εγγράφων
Multi-head self-attention mechanism	Μηχανισμός αυτο-προσοχής πολλαπλών κεφαλών
Natural language processing	Επεξεργασία φυσικής γλώσσας
Natural language understanding	Κατανόηση φυσικής γλώσσας
Negative log likelihood function	Συνάρτηση αρνητικής λογαριθμικής πιθανοφάνειας
New tokens rate	Ποσοστό νέων λέξεων
Overfitting	Υπερπροσαρμογή
Part-of-speech tagging	Επισημείωση μέρους του λόγου
Pointer-generator network	Δίκτυο γεννήτριας δεικτών
Policy	Πολιτική
Position embeddings	Διανυσματικές αναπαραστάσεις θέσης
Position wise	Γνώση της θέσης ή ως προς τη θέση
Precision	Ακρίβεια

Predicate-argument relations	Σχέσεις κατηγορήματος ορίσματος
Query-based text summarization	Περίληψη κειμένου που βασίζεται σε ερωτήσεις
Raw corpora	Σώμα κειμένου μη δομημένης πληροφορίας
Recall	Ανάκληση
Recommender systems	Συστήματα συστάσεων
Rectified linear unit	Ανορθωμένη γραμμική μονάδα
Recurrent neural networks	Αναδρομικά νευρωνικά δίκτυα
Regression	Παλιδρόμηση
Reinforcement learning	Ενισχυτική μάθηση
Residual connections	Υπολειμματικές συνδέσεις
Resource description framework	Πλαίσιο περιγραφής πόρων
Reward	Ανταμοιβή
Rhetorical structure theory	Θεωρία ρητορικής δομής
Root mean square propagation	Διάδοση ρίζας μέσων τετραγώνων
Second moment	Ροπή δεύτερης τάξης
Segmentation embeddings	Διανυσματική αναπαράσταση για διαχωρισμό προτάσεων
Self-attention	Αυτο-προσοχή
Self-critical sequence training	Εκπαίδευση εκτίμησης ακολουθίας με αυτοκριτική
Sense-annotated corpora	Σώμα κειμένου με επισημείωση εννοιών
Sentence segmentation	Διαχωρισμός προτάσεων
Sequence-to-sequence	Ακολουθία-σε-ακολουθία
Sigmoid function	Σιγμοειδής συνάρτηση
Single-document text summarization	Περίληψη κειμένου ενός-εγγράφου
Sinusoid positional embeddings	Ημιτονοειδείς διανυσματικές αναπαραστάσεις θέσης
Softmax	Κανονικοποιημένη εκθετική συνάρτηση
State	Κατάσταση
Stemming	Αποκοπή κατάληξης ή αποκατάληξη
Stochastic gradient descent	Στοχαστική κατάβαση κλίσης
Stop words	Κοινές λέξεις
Supervised learning	Επιβλεπόμενη μάθηση
Synset	Σύνολο συνώνυμων λέξεων
Test set	Σύνολο ελέγχου
Text summarization	Περίληψη κειμένου
Tokenization	Διαχωρισμός λεκτικών μονάδων κειμένου
Training set	Σύνολο εκπαίδευσης
Tranformers	Μετασχηματιστές
Uderfitting	Υποπροσαρμογή
Unsupervised learning	Μη-επιβλεπόμενη μάθηση
Validation set	Σύνολο επικύρωσης
Word embeddings	Ενσωματώσεις λέξεων

Word mover's distance	Απόσταση μεταφοράς λέξεων

Βιογραφικό σημείωμα του συγγραφέα

Ο Παναγιώτης Κουρής έχει λάβει Δίπλωμα Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών, με κατεύθυνση την Πληροφορική, από το Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ) και Μεταπτυχιακό Δίπλωμα Ειδίκευσης στην Πληροφορική και Τηλεματική, με κατεύθυνση τις Υπολογιστικές και Διαδικτυακές Τεχνολογίες και Εφαρμογές, από το Χαροκόπειο Πανεπιστήμιο. Από το 2016 είναι υποψήφιος διδάκτορας στο Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης του τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (ΣΗΜΜΥ) του ΕΜΠ. Η Διδακτορική του Διατριβή ολοκληρώθηκε τον Δεκέμβριο του 2022. Στο πλαίσιο των Μεταπτυχιακών-Διδακτορικών του σπουδών έχει δημοσιεύσει μια σειρά από άρθρα σε ερευνητικά περιοδικά και πρακτικά συνεδρίων με κρίση. Έχει πάρει μέρος στο ερευνητικό πρόγραμμα με τίτλο «Η αισθησιοκινητική βάση της αιτιακότητας και του ποιού ενεργείας και η δήλωσή τους στα ρήματα ώθησης, έλξης, κρούσης και δαρμού της Νέας Ελληνικής» του Ερευνητικού Κέντρου «Αθηνά», το οποίο ολοκληρώθηκε τον Απρίλιο του 2020. Συμμετέχει ως κριτής άρθρων σε ερευνητικά περιοδικά και συνέδρια και έχει παρακολουθήσει πλήθος επιστημονικών συνεδρίων. Έχει προσφέρει επικουρικό έργο στο πλαίσιο του εργαστηριακού μέρους συναφών με τα ερευνητικά του ενδιαφέροντα μαθημάτων του προπτυχιακού και μεταπτυχιακού κύκλου σπουδών του ΕΜΠ και έχει παρακολουθήσει την πορεία διπλωματικών εργασιών που εκπονήθηκαν στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης της ΣΗΜΜΥ του ΕΜΠ.

Τα ερευνητικά του ενδιαφέροντα περιλαμβάνουν τις περιοχές της τεχνητής νοημοσύνης, της μηχανικής μάθησης, της επεξεργασίας φυσικής γλώσσας, της αυτόματης περίληψη κειμένου και των συστημάτων συστάσεων.
Σύνδεσμος δημοσιεύσεων: https://scholar.google.gr/citations?user=7tpY4DUAAAAJ

Κατάλογος δημοσιεύσεων του συγγραφέα

Δημοσιεύσεις σχετικές με τη διατριβή

Περιοδικά με κρίση

Panagiotis Kouris, Georgios Alexandridis, Andreas Stafylopatis. Text summarization based on semantic graphs: An abstract meaning representation graph-to-text deep learning approach. Preprint (version 1, under peer review by a journal) available at *Research Square*, 2022. doi: https://doi.org/10.21203/rs.3.rs-1938526/v1

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. Abstractive text summarization: enhancing sequence to sequence models using word sense disambiguation and semantic content generalization. In: *Computational Linguistics*, pp. 1-41, 2021. doi: https://doi.org/10.1162/coli_a_00417

Georgios Stratogiannis, Panagiotis Kouris, Georgios Alexandridis, Georgios Siolas, Giorgos Stamou, and Andreas Stafylopatis. Semantic enrichment of documents: a classification perspective for ontology-based imbalanced semantic descriptions. In: *Knowledge and Information Systems*, pp. 1–39, 2021. doi: https://doi.org/10.1007/s10115-021-01615-y

Panagiotis Kouris, Iraklis Varlamis, Georgios Alexandridis, and Andreas Stafylopatis. A versatile package recommendation framework aiming at preference score maximization. In: *Evolving Systems*, pp. 1-19, 2018. doi: https://doi.org/10.1007/s12530-018-9231-2

Συνέδρια με κρίση

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. Abstractive text summarization based on deep learning and semantic content generalization. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5082-5092, 2019. doi: http://doi.org/10.18653/v1/P19-1501

Panagiotis Kouris, Iraklis Varlamis, and Georgios Alexandridis. A package recommendation framework based on collaborative filtering and preference score maximization. In: *International Conference on Engineering Applications of Neural Networks*. Springer. pp. 477-489, 2017. doi: https://doi.org/10.1007/978-3-319-65172-9_40

Δημοσιεύσεις εκτός διατριβής

Marietta Sionti, Panagiotis Kouris, Chrysovalantis Korfitis, Vasiliki Moutzouri, and Stella Markantonatou. Relations Among MOCAP and Textual Data of Motion Verbs: A Distance Calculation Perspective. In: *International Journal of Art, Culture and Design Technologies (IJACDT)*. (9.2), pp. 45-62, 2020. doi: https://doi.org/10.4018/ijacdt.2020070104

Panagiotis Kouris, Marietta Sionti, Chrysovalantis Korfitis, and Stella Markantonatou. Motion Capture of Modern Greek Verbs: Measuring aspects and relations among actions. In: *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. pp. 1-7, 2020. doi: https://doi.org/10.1109/PerComWorkshops48775.2020.9156254

Stella Markantonatou, Panagiotis Kouris, K. Selimi, D. Stasinou, and Yanis Maistros. Ψυχή άσπρη σαν το χιόνι but never ψυχή άσπρη σαν το γάλα: semasio-syntactic comments on the fixed similes of Modern Greek. *Proceedings of the Annual Meeting of the Department of Linguistics, Aristotle University of Thessaloniki, Studies in Greek Linguistics.* (39), pp. 639-650, 2019. url: http://ins.web.auth.gr/index.php?option=com_content&view=article&id=1214

Stella Markantonatou, Panagiotis Kouris, and Yanis Maistros. Fixed Similes: Measuring aspects of the relation between MWE idiomatic semantics and syntactic flexibility. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. pp. 51-61, 2018. url: https://aclanthology.org/W18-4908