Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Αυτόματη Περίληψη Κειμένου με Χρήση Νευρωνικών Δικτύων Βαθιάς Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΛΕΞΑΝΔΡΟΣ ΝΙΚΟΠΟΥΛΟΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Παναγιώτης Κουρής
Υ.Δ. Ε.Μ.Π.

Αθήνα, Μάρτιος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Αυτόματη Περίληψη Κειμένου με Χρήση Νευρωνικών Δικτύων Βαθιάς Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΛΕΞΑΝΔΡΟΣ ΝΙΚΟΠΟΥΛΟΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Παναγιώτης Κουρής
Υ.Δ. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Μαρτίου 2019.

.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Παναγιώτης Τσανάκας Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2019

Αλέξανδρος Νικόπουλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Αυτόματη περίληψη κειμένου ονομάζεται η διαδικασία παραγωγής περίληψης με χρήση κάποιου λογισμικού, ώστε να διατηρείται το αρχικό και κύριο νόημα του κειμένου. Στην σημερινή εποχή όπου ο όγκος της πληροφορίας ολοένα και αυξάνεται, η ανάπτυξη αποτελεσματικού λογισμικού για αυτόματη περίληψη καθιστά εφικτή τη προσπέλαση μεγάλου όγκου πληροφορίας με αποδοτικό τρόπο.

Για την παραγωγή νοηματικά σωστής περίληψης κειμένου, που παράλληλα θα διατηρεί ορθή σύνταξη και γραμματική, έχουν αναπτυχθεί διάφορα εργαλεία λογισμικού τα οποία βρίσκονται ακόμα υπό έρευνα. Λόγω της εγγενώς δύσκολης φύσης του προβλήματος τα εργαλεία που έχουν αναπτυχθεί έως σήμερα απέχουν αρκετά από την παραγωγή μιας ιδανικής περίληψης. Ωστόσο, η πρόσφατη έξαρση συγκεκριμένων ευφυών τεχνικών έχει επιφέρει κάποια βελτίωση στην αυτόματη παραγωγή περίληψης.

Στην συγκεκριμένη διπλωματική εργασία σχεδιάζεται ένας μηχανισμός που βασίζεται σε νευρωνικά δίκτυα βαθιάς μάθησης για την αντιμετώπιση του προβλήματος της αυτόματης περίληψης κειμένου. Στα πλαίσια αυτού του μηχανισμού, διερευνώνται και συγκρίνονται μεταξύ τους διάφορες σχεδιαστικές επιλογές με στόχο την μεγιστοποίηση της επίδοσης. Πιο συγκεκριμένα, αρχικά πραγματοποιείται επεξεργασία των συνόλων δεδομένων που χρησιμοποιούνται με στόχο την ελαχιστοποίηση του θορύβου που περιέχουν. Στη συνέχεια παρουσιάζεται η αρχιτεκτονική του μηχανισμού παραγωγής της περίληψης και μελετάται ως προς τις διάφορες κρίσιμες παραμέτρους. Στα πλαίσια αντιμετώπισης της εγγενούς δυσκολίας που υπάρχει στο συγκεκριμένο πρόβλημα λόγω της αναγκαιότητας χειρισμού πολύ μεγάλου πλήθους λέξεων, παρουσιάζεται και αναλύεται ένας καινοφανής μηχανισμός αντιμετώπισης των πιθανών άγνωστων λέξεων που εμφανίζονται κατά την διαδικασία παραγωγής της περίληψης.

Για την εφαρμογή και αξιολόγηση του συστήματος αυτόματης περίληψης κειμένου χρησιμοποιούνται δύο γνωστά σύνολα δεδομένων. Οι μετρήσεις επίδοσης βασίζονται στην καθιερωμένη μετρική Rouge η οποία πραγματοποιεί συγκρίσεις ομοιότητας μεταξύ των παραγόμενων και των δοσμένων περιλήψεων. Από τα πειραματικά αποτελέσματα εξάγονται χρήσιμα συμπεράσματα για τη βελτίωση της απόδοσης ενός συστήματος αυτόματης περίληψης κειμένου. Τέλος, παρουσιάζονται κάποιες μελλοντικές κατευθύνσεις για την περαιτέρω προώθηση της έρευνας.

Λέξεις κλειδιά

Αυτόματη Περίληψη Κειμένου, Νευρωνικά Δίκτυα Βαθιάς Μάθησης, Αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή, Μετρήσεις Rouge

Abstract

Automatic text summarization refers to the process of shortening a text document with software, so that its main point and topics are preserved. Nowadays, the volume of information rapidly increases and effective pieces of software for text summarization can allow more information to be processed efficiently.

Many pieces of software have been manufactured which aim to produce meaningful summarizations of text, while also preserving correct syntax and grammar, but they are still under research. Due to the difficult nature of the problem these pieces of software are far from being able to produce an ideal summarization. However, due to the fact that new revolutionary and intelligent techniques have recently emerged, there has been some improvement to automatic text summarization.

In this thesis, a deep learning neural netowrk for automatic text summarization is designed. Within this design process, different configurations are implemented and compared, aiming to maximize performance. More specifically, data preprocessing procedure for noise reduction of the used datasets is described at first. After that, the summarization mechanism's architecture is described and its performance is measured, for a set of its critical parameters. Trying to overcome the inherent problem of dealing with huge amounts of different words, a novel mechanism is introduced and analyzed, that aims to handle the uknown words that appear in the automatic summarization process.

In order to evaluate the implemented text summarization mechanism correctly, two well known text summarization datasets are used. The performance metrics on these datasets are based on the well established Rouge metric for text summarization tasks. Via analyzing the best occuring results through this metric, various conclusions are extracted, aiming to improve the efficiency of text summarization systems. Finally, future directions are presented in this thesis, in an attempt to forward the research in automatic text summarization.

Key words

Automatic Text Summarization, Deep Learning Neural Networks, Encoder-Decoder Architecture, Rouge package

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού προγράμματος σπουδών της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου και σηματοδοτεί την ολοκλήρωση των σπουδών μου ενώ συγχρόνως αποτελεί το ερέθισμα για περαιτέρω έρευνα στο συγκεκριμένο αντικείμενο.

Προτού όμως αναφερθώ στη περιγραφή της εργασίας και στα αποτελέσματα που προέκυψαν, θα ήθελα να ευχαριστήσω θερμά τους ανθρώπους οι οποίοι μέσω της συνεργασίας μας, συνέβαλαν σημαντικά στην ολοκλήρωση αυτής της εργασίας.

Αρχικά θα ήθελα να απευθύνω τις ευχαριστίες μου στον επιβλέποντα κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π, ο οποίος μου προσέφερε τη δυνατότητα να εκπονήσω την διπλωματική μου σε ένα αντικείμενο ιδιαίτερα ελκυστικό και ενδιαφέρον για μένα και να διευρύνω τις επιστημονικές μου γνώσεις. Παράλληλα θα ήθελα να ευχαριστήσω τους κ.κ. Παναγιώτη Τσανάκα, Καθηγητή Ε.Μ.Π και Γεώργιο Στάμου, Αναπληρωτή Καθηγητή Ε.Μ.Π για την τιμή που μου έκαναν να είναι μέλη της επιτροπής εξέτασης της διπλωματικής εργασίας.

Επίσης οφείλω ιδιαίτερες ευχαριστίες στον κ. Παναγιώτη Κουρή, Υ.Δ. Ε.Μ.Π. για το χρόνο που αφιέρωσε και την θεμελιώδη του συνεισφορά στην εκπόνηση της συγκεκριμένης εργασίας. Η στήριξη του, επιστημονική και πνευματική, καθώς και η καθοδήγηση του σε όλη τη διάρκεια της πορείας αυτής συνέβαλαν τα μέγιστα στην επίτευξη ενός πολύ σημαντικού για εμένα στόχου. Η προθυμία του να με βοηθήσει μέσω της εμπειρίας και των γνώσεων του σε οποιαδήποτε δυσκολία συνάντησα στάθηκαν καθοριστικές και η συνεργασία μας θεωρώ πως ήταν άκρως επιτυχημένη και εποικοδομητική.

Τέλος, με εξίσου μεγάλη θέρμη θέλω να αναφερθώ και να ευχαριστήσω την οικογένεια μου , η οποία με στήριξε όλα αυτά τα χρόνια σε όλες τις δύσκολες στιγμές, καθώς και τους φίλους και τους συμφοιτητές μου, οι οποίοι στάθηκαν δίπλα μου σε όλη τη διάρκεια της ακαδημαϊκής μου πορείας, ο καθένας με τον δικό του ξεχωριστό τρόπο.

Αλέξανδρος Νικόπουλος,

Αθήνα, 13η Μαρτίου 2019

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κατάλογος Αλγορίθμων

Κεφάλαιο 1

Εισαγωγή

1.1 Αυτόματη παραγωγή περίληψης

Η παραγωγή περίληψης κειμένου αποτελούσε ανέκαθεν ένα εργαλείο ελαχιστοποίησης του όγκου της πληροφορίας, το οποίο χρησιμοποιούσαν οι άνθρωποι για να ερμηνεύσουν και να προσπελάσουν περισσότερη πληροφορία γρηγορότερα. Στην σημερινή εποχή όμως, που η τεχνολογία και το διαδίκτυο ολοένα και εξελίσσονται και ο όγκος την πληροφορίας είναι τεράστιος, η αναγκαιότητα για ελαχιστοποίηση της πληροφορίας έχει ξεπεράσει την δυνατότητα πραγματοποίησης της από ανθρώπους σε λογικό χρόνο.

Η αυτόματη περίληψη κειμένου αποτελεί πρόβλημα του τομέα επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) και στόχος της αποτελεί η μοντελοποίηση αυτόματων κανόνων για την συστηματική και ανέξοδη ελαχιστοποίηση της πληροφορίας με χρήση κάποιου λογισμικού, ώστε να μπορεί να χρησιμοποιηθεί σε μεγάλους όγκους δεδομένων. Σε μια προσπάθεια πιο αυστηρού μαθηματικού ορισμού, θα μπορούσε κανείς να ορίσει το πρόβλημα ως εξής: Δεδομένης μιας ακολουθίας λέξεων εισόδου $X = \{x_i, i = 1...N\}$, στόχος είναι να βρεθεί μια ακολουθία εξόδου $Y = \{y_i, i = 1...M\}$ με $M < N$ η οποία διατηρεί το νοηματικό περιεχόμενο της . Στην σημερινή εποχή η αναγκαιότητα για τέτοια συστήματα παραγωγής περίληψης είναι εμφανής και όλο και περισσότερα βρίσκουν καθημερινά εφαρμογή. Τέτοια παραδείγματα εφαρμογής μπορεί να είναι η χρήση λογισμικού για την εξαγωγή ενδεικτικών περιλήψεων κατά την αναζήτηση σε μηχανές αναζήτησης [Turp07] ή η χρήση λογισμικού για την αυτόματη εξαγωγή τίτλων σε κείμενα [Trip17] [Alla17] [Savo10].

Παρόλο που ο σχεδιασμός τέτοιων λογισμικών είναι πιο κρίσιμος από ποτέ στις μέρες μας, αυτός ξεκίνησε από πολύ παλιά, με τα πρώτα σημάδια αναζήτησης αυτόματων τεχνικών παραγωγής περίληψης να εμφανίζονται την δεκαετία του 1950. Πιο συγκεκριμένα, στο [Luhn58] πραγματοποιήθηκε η πρώτη προσπάθεια για ελαχιστοποίηση της πληροφορίας σε επιστημονικά κείμενα μέσω τεχνικών βαθμολόγησης του περιεχομένου των προτάσεων, με τελικό στόχο την εξαγωγή των πιο σημαντικών από αυτές. Βλέπουμε λοιπόν πως η αναγκαιότητα για ελαχιστοποίηση της πληροφορίας απασχολεί την ανθρωπότητα πολύ καιρό και ο σχεδιασμός εύρωστων τέτοιων συστημάτων έχουν πολύ ερευνητική αξία.

1.2 Προηγούμενες Εργασίες

Τα τελευταία χρόνια έχει πραγματοποιηθεί πάρα πολύ έρευνα στο αντικείμενο της αυτόματης παραγωγής περίληψης. Το συγκεκριμένο αντικείμενο είναι αρκετά αχανές και μπορεί να προσεγγιστεί με πάρα πολλούς διαφορετικούς τρόπους. Αυτό έχει ως αποτέλεσμα να υπάρχουν πολλές διαφορετικές ερευνητικές εργασίες που προσπαθούν να το προωθήσουν από διαφορετικές σκοπιές η καθεμία. Καθώς στα πλαίσια αυτής της εργασίας θα αναλυθούν τεχνικές βαθιάς μάθησης, σε αυτό το σημείο θα πραγματοποιηθεί μια μικρή ανάλυση της έρευνας που έχει ήδη πραγματοποιηθεί για την μοντελοποίηση του προβλήματος της αυτόματης παραγωγής περίληψης με χρήση βαθιάς μάθησης.

Μέχρι την ανακάλυψη των πρώτων τεχνικών βαθιάς μάθησης για την αντιμετώπιση του προβλήματος, οι περισσότερες προσεγγίσεις αφορούσαν την απόδοση βαρών στις λέξεις ή τις προτάσεις των κειμένων και στην μετέπειτα εξαγωγή των πιο σημαντικών. Στο [Bahd14] μοντελοποιήθηκε και χρησιμοποιήθηκε για πρώτη φορά ένα σύστημα βαθιάς μάθησης για την αυτόματης μετάφραση κειμένων από μία γλώσσα σε μία άλλη. Αν και αυτό το πρόβλημα είναι διαφορετικό από το πρόβλημα της αυτόματης παραγωγής περίληψης, τα δύο αυτά έχουν πολλά κοινά και μπορούν να αντιμετωπιστούν με παρόμοιους τρόπους.

Από την στιγμή που έγινε το πρώτο βήμα, άρχισαν όλο και περισσότεροι να χρησιμοποιούν τεχνικές βαθιάς μάθησης για την μοντελοποίηση παρόμοιων προβλημάτων. Στο [Rush15] παρουσιάζεται ένα πολύ βελτιωμένο σύστημα παραγωγής περίληψης εμπνευσμένο από τις τεχνικές του [Bahd14]. Σε αυτήν την εργασία παρουσιάζεται για πρώτη φορά ο επαναστατικός μηχανισμός προσοχής καθώς θέτονται και οι αρχές για την δημιουργία κοινού συνόλου δεδομένων με στόχο την δυνατότητα καλύτερης σύγκρισης διαφορετικών εργασιών. Στο [Wu16] παρουσιάζονται οι αρχές τους γνωστού συστήματος αυτόματης μετάφρασης της Google βασισμένο πάλι στις ίδιες αρχές. Στο [Nall16] προωθείται περαιτέρω η έρευνα πάνω στην αυτόματη παραγωγή περίληψης και μελετώνται διαφορετικές αρχιτεκτονικές για καλύτερη απόδοση στον χρόνο και στις υπολογιστικές απαιτήσεις. Τέλος, στο [See17] γίνεται μια προσπάθεια για αυτόματη παραγωγή περίληψης σε κείμενα μεγαλύτερου μεγέθους και παρουσιάζεται ένας πρωτοποριακός μηχανισμός για τον χειρισμό των άγνωστων λέξεων που εμφανίζονται στα κείμενα.

1.3 Αντικείμενο της διπλωματικής

Αντικείμενο της συγκεκριμένης διπλωματικής εργασίας αποτελεί ο σχεδιασμός και η υλοποίηση ενός ολοκληρωμένου συστήματος αυτόματης παραγωγής περίληψης. Το συγκεκριμένο σύστημα θα είναι σε θέση με δεδομένο κείμενο ως είσοδο, να παράξει μια συνοπτικότερη εκδοχή του, διατηρώντας ταυτόχρονα την νοηματική ουσία του. Η δυνατότητα του συγκεκριμένου συστήματος να παράξει ορθές περιλήψεις θα περιορίζεται σε κείμενα αγγλικής γλώσσας και του συγκεκριμένου συντακτικού ύφους που εμφανίζεται στα δεδομένα εκπαίδευσης που θα παρουσιαστούν στην συνέχεια. Αυτό συμβαίνει καθώς η δημιουργία ενός συστήματος παραγωγής περίληψης γενικού σκοπού αποτελεί πάρα πολύ δύσκολο πρόβλημα για το οποίο απαιτείται τεράστιος όγκος πληροφορίας.

Στα πλαίσια σχεδιασμού του προαναφερθέντος μηχανισμού, θα πραγματοποιηθεί διερεύνηση πάνω σε διάφορες σχεδιαστικές επιλογές, με στόχο την μεγιστοποίηση της επίδοσης του. Πιο συγκεκριμένα ο μηχανισμός θα αξιολογηθεί ως προς διάφορες αντικρουόμενες σχεδιαστικές αποφάσεις, ώστε να παρουσιαστεί καλύτερα στον αναγνώστη η γενικότερη συμπεριφορά του μηχανισμού κατά την αντιμετώπιση του προβλήματος της αυτόματης παραγωγής περίληψης. Γι' αυτό το σκοπό θα πραγματοποιηθούν μετρήσεις της επίδοσης του μηχανισμού με διαφορετικές παραμέτρους, με χρήση της μετρικής Rouge, και θα αναλυθούν τα τελικά αποτελέσματα.

Τα συμπεράσματα που θα προκύψουν από την τελική αυτή ανάλυση θα έχουν ως στόχο να υποδείξουν κάποιες εμπειρικά ορθές σχεδιαστικές κατευθύνσεις κατά τον σχεδιασμό παρόμοιων συστημάτων παραγωγής περίληψης. Έτσι η διερεύνηση του αντικειμένου με χρήση παρόμοιων μηχανισμών θα μπορεί να μοντελοποιηθεί και να σχεδιαστεί καλύτερα.

Στα πλαίσια της συγκεκριμένης εργασίας με στόχο την περαιτέρω προώθηση της έρευνας πάνω στο αντικείμενο, θα παρουσιαστεί και θα αναλυθεί επίσης ένας καινοφανής μηχανισμός χειρισμού των πιθανών άγνωστων λέξεων που μπορεί να εμφανιστούν κατά την εκτέλεση του λογισμικού. Το πρόβλημα αυτό, αποτελεί ένα πολύ σημαντικό πρόβλημα το οποίο πρέπει να επιλυθεί με στόχο τον σχεδιασμό ενός καθολικού συστήματος παραγωγής περίληψης.

1.4 Δομή εργασίας

Τα κεφάλαια που ακολουθούν στην συνέχεια έχουν την ακόλουθη μορφή. Στο κεφάλαιο 2 παρουσιάζεται εκτενώς η θεωρία πίσω από την οποία στηρίζεται ο μηχανισμός που θα υλοποιηθεί στην συνέχεια. Η θεωρία παρουσιάζεται με πολύ περιγραφικό τρόπο και αιτιολογείται αρκετά εκτενώς η λειτουργία του κάθε στοιχείου του μηχανισμού.

Στο κεφάλαιο 3 Παρουσιάζεται η προεπεξεργασία που έχει πραγματοποιηθεί στα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του μηχανισμού. Η προεπεξεργασία των δεδομένων αποτελεί ένα πολύ σημαντικό βήμα για την επίτευξη ενός εύρωστου συστήματος μηχανικής μάθησης και στο συγκεκριμένο κεφάλαιο δίνεται πολύ προσοχή στην περιγραφή και αιτιολόγηση όλων των αποφάσεων επεξεργασίας που πάρθηκαν.

Στο κεφάλαιο 4 παρουσιάζεται εκτενώς ο μηχανισμός που υλοποιήθηκε. Το συγκεκριμένο κεφάλαιο δεν στηρίζεται τόσο στο θεωρητικό υπόβαθρο αλλά προσεγγίζει τον μηχανισμό πιο τεχνικά και αναλύει περισσότερο διάφορες σχεδιαστικές αποφάσεις που πάρθηκαν κατά την δημιουργία του. Η περιγραφή του μηχανισμού παραλληλίζεται με την ροή της πληροφορίας μέσα στον μηχανισμό ώστε να γίνει πιο κατανοητή η λειτουργία αυτού.

Στο κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα της διερεύνησης που πραγματοποιήθηκε στα πλαίσια αυτής της εργασίας. Πραγματοποιούνται μετρήσεις για τις μετρικές Rouge 1, Rouge 2, Rouge L και με βάση αυτές πραγματοποιούνται συγκρίσεις μεταξύ διαφορετικών αρχιτεκτονικών για διερεύνηση της επίδοσής τους. Επίσης παροουσιάζεται ένας πρωτοποριακός μηχανισμός χειρισμού των άγνωστων λέξεων

Τέλος, στο κεφάλαιο 6 παρουσιάζονται τα τελικά συμπεράσματα της διερεύνησης στα πλαίσια αυτής της εργασίας καθώς και κάποιες μελλοντικές κατευθύνσεις για περαιτέρω προώθηση της έρευνας πάνω στο αντικείμενο.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

2.1 Εισαγωγικά

Όπως έχει ήδη αναφερθεί, αυτόματη παραγωγή περίληψης κειμένου ονομάζεται η διαδικασία παραγωγής περίληψης κειμένου με χρήση κάποιου λογισμικού. Παρόλο που έχουν γίνει πολλές προσπάθειες προσέγγισης του θέματος με διάφορους μηχανισμούς λογισμικού, απέχουμε ακόμα αρκετά από την δημιουργία ενός καθολικού συστήματος παραγωγής περιλήψεων γενικού σκοπού. Ο λόγος που συμβαίνει αυτό, οφείλεται στην εγγενή πολυπλοκότητα του θέματος, η οποία καθιστά δύσκολη την μοντελοποίηση του προβλήματος από υπολογιστικά συστήματα.

Πιο συγκεκριμένα, αναλύοντας κανείς το πρόβλημα, μπορεί να καταλάβει εύκολα τις προκλήσεις που μπορεί να αποφέρει αυτό σε μια υπολογιστική μηχανή. Ο υπολογιστής, ως μία ντετερμινιστική μηχανή που λειτουργεί με κανόνες δυσκολεύεται να προσεγγίσει όλες τις ιδιομορφίες που εμφανίζονται σε διαφορετικά κείμενα προς περίληψη. Οι διαφορετικές γλώσσες, οι διαφορετικές διάλεκτοι και τα διαφορετικά ύφη γραψίματος είναι μόνο η κορυφή του παγόβουνου των παραμέτρων που πρέπει να συνυπολογιστούν για την επίτευξη καλού αποτελέσματος. Ακόμη και να γίνει περιορισμός στην προσπάθεια αντιμετώπισης του προβλήματος μόνο για μία φυσική γλώσσα, ο αριθμών των διαφορετικών τρόπων έκφρασης και πάλι είναι δύσκολα υπολογίσιμος, πόσο μάλλον μοντελοποιήσιμος.

Ένα ακόμα πρόβλημα είναι ότι ο υπολογιστής δεν γνωρίζει την νοηματική συνοχή διαφόρων λέξεων και προτάσεων αλλά μόνο αναγνωρίζει τις λέξεις και τις προτάσεις ως δυαδική πληροφορία. Λόγω αυτού έχουν πραγματοποιηθεί ολόκληρες έρευνες σε μια προσπάθεια μοντελοποίησης της κατηγορηματικής πληροφορίας των διαφόρων κειμένων σε αριθμητική μορφή κατανοητή από τον υπολογιστή, ως μια προσπάθεια ο υπολογιστής να γνωρίζει όσο καλύτερα γίνεται την νοηματική ουσία των προτάσεων που επεξεργάζεται και να πραγματοποιεί καλύτερες συγκρίσεις μεταξύ τους.

Τέλος, ένα σημαντικό πρόβλημα που δυσκολεύει την προώθηση της έρευνας πάνω στο αντικείμενο, είναι η δυσκολία αξιολόγησης των παραγόμενων περιλήψεων και γενικότερα η δυσκολία εύρεσης ενός μέτρου αντικειμενικής κατάταξης των διαφόρων τρόπων παραγωγής περίληψης. Παρόλο που έχει γίνει αρκετή έρευνα πάνω στο κομμάτι αυτό και παρόλο που έχουν ορισθεί μαθηματικά μοντέλα που είναι αρκετά διαδεδομένα και χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων από διάφορες έρευνες στο αντικείμενο, απέχουν και αυτά πολύ από το να μπορούν να αξιολογήσουν νοηματικά πλήρως μια περίληψη ενός κειμένου. Συγκεκριμένα, τα πιο διαδεδομένα μέτρα σύγκρισης των διαφόρων μηχανισμών, βασίζονται στην μέτρηση εμφάνισης κοινών λέξεων και γραμμάτων μεταξύ των παραγόμενων περιλήψεων και διαφόρων περιλήψεων αναφοράς. Παρόλο που αυτά τα μέτρα αποτελούν καλό μέτρο για την κατάταξη των διάφορων μεθόδων παραγωγής περίληψης, χωλαίνουν στην δυνατότητα ανάδειξης περιλήψεων νοηματικά ορθών, που χρησιμοποιούν όμως συνώνυμα λέξεων των διαφόρων περιλήψεων αναφοράς. Αυτό σημαίνει πρακτικά ότι μια ορθή περίληψη που μάλιστα έχει καταφέρει να πιάσει πλήρως το νοηματικό περιεχόμενο του κειμένου και περιέχει συνώνυμα λέξεων του κειμένου, πιθανώς θα βαθμολογηθεί χειρότερα από άλλες πιθανές περιλήψεις που μπορεί να περιέχουν αντιγραμμένες προτάσεις από το αρχικό κείμενο.

Λόγω αυτών των δυσκολιών αλλά και άλλων πολλών, οι οποίες πάλι απορρέουν στην πολυσύνθετη φύση του προβλήματος επεξεργασίας φυσικών γλωσσών, οι προσεγγίσεις του προβλήματος που αποφέρουν τα καλύτερα αποτελέσματα έχουν υλοποιηθεί με τεχνικές μηχανικής μάθησης και τεχνικές εξόρυξης δεδομένων. Η ιδιαιτερότητα που προσφέρουν αυτές οι τεχνικές είναι ότι σε μικρό η

σε μεγάλο βαθμό μπορούν να απαλύνουν την ανάγκη του υπολογιστή για ντετερμινισμό και να του επιτρέπουν να καταλάβει μόνος του ορισμένα στοιχεία της γλώσσας, οδηγώντας εν τέλει σε πιο εύρωστα συστήματα με καλύτερη απόδοση. Στην συνέχεια θα αναλυθούν οι διάφορες κατηγορίες τέτοιων βέλτιστων τεχνικών που χρησιμοποιούνται για την προσέγγιση καλών συστημάτων αυτόματης παραγωγής περίληψης κειμένου.

2.2 Τεχνικές αυτόματης περίληψης κειμένου

Για την αντιμετώπιση του προβλήματος της αυτόματης περίληψης κειμένου υπάρχουν πολλοί διαφορετικοί μηχανισμοί που χρησιμοποιούνται, όλοι όμως από αυτούς αφορούν κάποια από τις δύο κύριες προσεγγίσεις. Η πρώτη προσέγγιση αφορά την εξαγωγή ουσιώδους πληροφορίας από το κείμενο ως μέσο δημιουργίας μιας τελικής περίληψης (extractive mechanism). Αντίθετα, η δεύτερη προσέγγιση αφορά την παραγωγή νέων προτάσεων για την δημιουργία της τελικής περίληψης, συμβουλεύοντας φυσικά την πληροφορία που υπάρχει στο αρχικό κείμενο (abstractive mechanism). Στην συνέχεια θα παρουσιαστούν λίγο πιο εκτενώς οι δύο αυτές διαφορετικές βασικές προσεγγίσεις.

- **Αυτόματη περίληψη που βασίζεται στην εξαγωγή κειμένου:** Στους μηχανισμούς εξαγωγής του κειμένου για την παραγωγή της περίληψης χρησιμοποιείται η μηχανική μάθηση και διάφορες άλλες τεχνικές εξόρυξης για την εύρεση των πιο σημαντικών προτάσεων σε ένα κείμενο ή μία συστάδα κειμένων έτσι ώστε στην συνέχεια αυτές να εξαχθούν για να σχηματίσουν την τελική περίληψη. Καθώς ο υπολογιστής όπως αναλύσαμε δυσκολεύεται να δώσει νόημα στην κατηγορηματική φύση των διάφορων προτάσεων και λέξεων, οι τεχνικές εξόρυξης που χρησιμοποιούνται έχουν ως στόχο να βαθμολογήσουν τις λέξεις και τις προτάσεις με βάση διάφορα κριτήρια, επιτρέποντας έτσι στον υπολογιστή να διαχωρίσει ευκολότερα τις πιο κρίσιμες προτάσεις. Σημειώνεται πως στους μηχανισμούς εξαγωγής πληροφορίας, καθώς η πληροφορία που χρησιμοποιείται για την εκάστοτε περίληψη έχει εξαχθεί από το αρχικό κείμενο, η διατήρηση ορθής γραμματικής και σύνταξης αν αυτές είναι επιθυμητές, είναι δεδομένη.

- **Αυτόματη περίληψη που βασίζεται στην παραγωγή κειμένου:** Στους μηχανισμούς παραγωγής του κειμένου της περίληψης, χρησιμοποιούνται τεχνικές μηχανικής μάθησης και άλλες τεχνικές εξόρυξης πληροφορίας από τα κείμενα ως ένας μέσο παραγωγής της περίληψης αυτόματα από το εκάστοτε κείμενο χωρίς προγενέστερη νοηματική επεξεργασία των προτάσεων του. Σε αυτή την περίπτωση ο υπολογιστής συνήθως μέσω της τροφοδοσίας πολλών δεδομένων, καλείται να παράξει την περίληψη του κειμένου μόνος του. Εάν η ορθή γραμματική και η σύνταξη είναι ζητούμενες, τότε ο υπολογιστής πρέπει να είναι σε θέση να τις παράξει και αυτές ορθώς μόνος του. Αυτοί οι μηχανισμοί παραγωγής περίληψης είναι πιο σύγχρονοι από τους μηχανισμούς εξαγωγής που αναφέρθηκαν παραπάνω και αν και είναι πιο περίπλοκοι και οδηγούν κατά μέσο όρο σε χειρότερα αποτελέσματα σε σχέση με αυτούς, είναι πιο υποσχόμενες στο να προσεγγίσουν ιδανικά και γενικά συστήματα παραγωγής περιλήψεων, καθώς επιτρέπουν περισσότερη αφαιρετικότητα, που σημαίνει ότι μπορούν να παράξουν περιλήψεις που διαφέρουν αρκετά από το αρχικό κείμενο και διατηρούν ωστόσο την ουσία του, ακριβώς όπως θα έκανε και ένας άνθρωπος. Ακόμη αυτά τα συστήματα απαιτούν συνήθως λιγότερους κανόνες για την παραγωγή περίληψης, και μπορούν δυνητικά να εκφράσουν μεγαλύτερο σύνολο διαλέκτων και ιδιομορφιών της ανθρώπινης γλώσσας.

Στην συγκεκριμένη εργασία, θα χρησιμοποιηθούν τεχνικές παραγωγής πληροφορίας της περίληψης, καθώς όπως περιγράφηκε πιο πάνω, αυτές είναι πιο υποσχόμενες να κυριαρχήσουν σε πιο εκλεπτυσμένα συστήματα παραγωγής περίληψης στο μέλλον. Στην συνέχεια, θα παρουσιαστεί το γενικό θεωρητικό υπόβαθρο του μηχανισμού που θα χρησιμοποιηθεί.

2.3 Θεωρητικά στοιχεία μηχανισμού

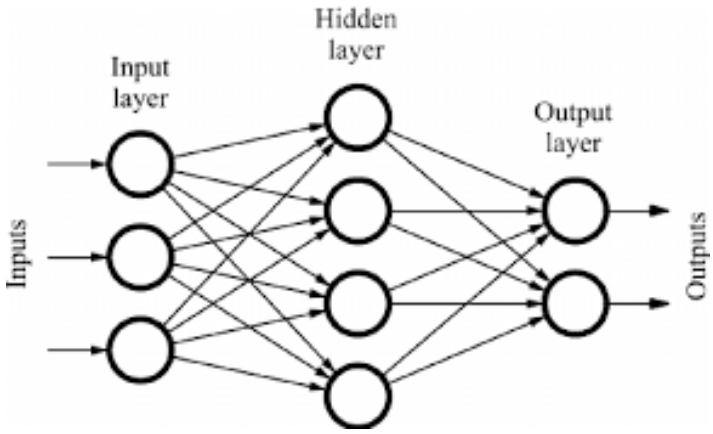
2.3.1 Γενικά

Η αυτόματη παραγωγή περίληψης κειμένου, έτσι ώστε να παραφράσεται το αρχικό κείμενο ενώ διατηρείται ορθή γραμματική και σύνταξη, αποτελεί ένα πολύ σύνθετο πρόβλημα. Με την έξαρση διάφορων μηχανισμών βαθιάς μάθησης που αποδείχθηκαν πολύ αποτελεσματικοί στην επεξεργασία φυσικών γλωσσών, άρχισαν να πραγματοποιούνται διάφορες έρευνες που έφεραν πολύ καλά αποτελέσματα σε διάφορα προβλήματα ακολουθιακής παραγωγής πληροφορίας. Προβλήματα όπως η αυτόματη μετάφραση κειμένου [Bahd14], η αυτόματη παραγωγή κειμένου από φωνή [Bahd16] και η αυτόματη περιγραφή εικόνας με κείμενο [Venu15] αντιμετωπίστηκαν για πρώτη φορά σε πολύ ικανοποιητικό βαθμό και άνοιξαν νέους ορίζοντες στην επιστήμη της επεξεργασίας φυσικών γλωσσών. Τα ίδια μοντέλα που χρησιμοποιήθηκαν για να επιλύσουν τα παραπάνω προβλήματα, χρησιμοποιήθηκαν και για την αντιμετώπιση του προβλήματος της αυτόματης παραγωγής περίληψης που αν και αποτελεί πιο σύνθετο πρόβλημα από τα παραπάνω σε επίπεδο φυσικής γλώσσας, έφεραν πολύ καλά αποτελέσματα. Πιο συγκεκριμένα, πολλές υλοποιήσεις για την αντιμετώπιση του συγκεκριμένου προβλήματος, βασίστηκαν στο ήδη ικανοποιητικά επιλυμένο πρόβλημα της αυτόματης μετάφρασης κειμένου, καθώς αυτά τα προβλήματα μοιάζουν πολύ μεταξύ τους, αφού και τα δύο επιδιώκουν να ”μετατρέψουν” μια μορφή κειμένου σε μία άλλη. Μια σημαντική διαφορά είναι ότι το μήκος του κειμένου προς πρόβλεψη στην περίπτωση της αυτόματης μετάφρασης είναι εύκολα υπολογίσιμο, σε αντίθεση με το αντίστοιχο μήκος κειμένου στην περίπτωση του προβλήματος της περίληψης.

Αυτά τα παραδείγματα που αναφέρθηκαν παραπάνω, χρησιμοποιούν μια πολύ συγκεκριμένη αρχιτεκτονική βαθιάς μάθησης που ονομάζεται αρχιτεκτονική κωδικοποιητή - αποκωδικοποιητή (encoder - decoder architecture). Παρόλο που αυτή η αρχιτεκτονική έχει πολύ μεγάλο περιθώριο παραμετροποίησης θα αναλυθεί παρακάτω η γενική φιλοσοφία και η θεωρία πίσω από την λειτουργία της, η οποία καθιστά εφικτά τόσο επαναστατικά αποτελέσματα στον τομέα της επεξεργασίας φυσικών γλωσσών. Στην παρακάτω ανάλυση θα παρουσιαστούν και θα περιγραφούν εκτενώς τα θεωρητικά στοιχεία της αρχιτεκτονικής, η οποία θα χρησιμοποιηθεί στην συνέχεια για την αντιμετώπιση του προβλήματος της παραγωγής περίληψης στα πλαίσια αυτής της εργασίας.

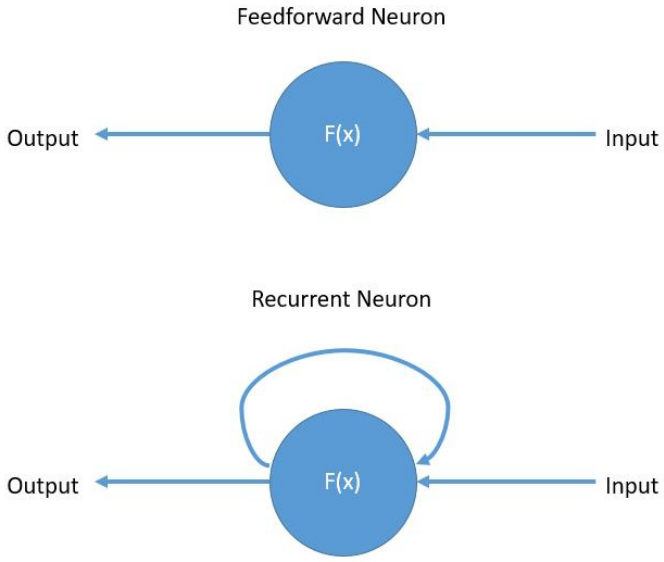
2.3.2 Αναδρομικά νευρωνικά δίκτυα

Το βασικό δομικό στοιχείο, πάνω στο οποίο βασίζεται όλη αυτή η αρχιτεκτονική ονομάζεται αναδρομικό νευρωνικό δίκτυο (recurrent neural network). Αυτό το είδος νευρωνικού δικτύου μοιάζει πολύ με το κλασσικό εμπρόσθιο είδος νευρωνικού δικτύου, το οποίο αποτελείται από τεχνητούς νευρώνες ενωμένους μεταξύ τους και η πληροφορία ρέει από το στρώμα εισόδου προς το στρώμα εξόδου, αλλάζοντας τιμές καθ' όλη την διάρκεια. Σε αυτό το δίκτυο συγκεκριμένες τιμές που ονομάζονται βάρη αλλάζουν συνεχώς κατά την διάρκεια της εκπαίδευσης ώστε οι τιμές της εξόδου να προσαρμοστούν εν τέλει στις επιθυμητές για κάθε είσοδο.



Σχήμα 2.1: Οπτική αναπαράσταση κλασσικού εμπρόσθιου νευρωνικού δικτύου

Η κύρια διαφορά του αναδρομικού νευρωνικού δικτύου με το κλασσικό, είναι ότι η έξοδος του δικτύου ανατροφοδοτείται στην είσοδο του, επιτρέποντας κάθε έξοδος του να αποτελεί συνάρτηση της τωρινής αλλά και προηγούμενων εισόδων στο δίκτυο. Η πληροφορία αυτή που ανατροφοδοτείται ονομάζεται κρυφή κατάσταση (hidden state) και είναι αυτή που χρησιμοποιείται για τυχόν προβλέψεις από το δίκτυο.



Σχήμα 2.2: Διαφορές κλασσικού νευρωνικού δικτύου και αναδρομικού νευρωνικού δικτύου

Η ιδιότητα αυτή του αναδρομικού νευρωνικού δικτύου να ανατροφοδοτεί την έξοδο στην είσοδό του είναι πολύ σημαντική και αποτελεί έναν από τους κύριους λόγους που λειτουργεί πολύ αποτελεσματικά σε προβλήματα επεξεργασίας φυσικής γλώσσας. Η ανατροφοδότηση αυτή επιτρέπει στο δίκτυο να χειρίζεται πολύ αποτελεσματικά εισόδους που αποτελούν εξαρτημένες ακολουθίες που σημαίνει ότι κάθε είσοδος της ακολουθίας εξαρτάται από τις προηγούμενες. Κάθε πρόταση φυσικής γλώσσας, αποτελείται από μια ακολουθία λέξεων οι οποίες δεν είναι ανεξάρτητες μεταξύ τους. Η κάθε λέξη σε μία πρόταση, καθορίζει σε απόλυτο βαθμό τις λέξεις που μπορεί να την ακολουθήσουν, τόσο από άποψη γραμματικής, όσο και από άποψη νοηματικού περιεχομένου. Το αναδρομικό νευρωνικό δίκτυο λοιπόν μπορεί να παράγει λέξεις της περίληψης συνυπολογίζοντας τις ήδη παραγόμενες λέξεις, δίνοντας του την δυνατότητα να μπορεί να λάβει σωστές αποφάσεις όσον αφορά το νόημα και την σύνταξη, οδηγώντας έτσι σε ορθές προτάσεις.

Παρόλο όμως που αυτή η μορφή νευρωνικού δικτύου είναι πολύ καλή στο να χειρίζεται και να επεξεργάζεται ακολουθίες, πάσχει από το πρόβλημα της βραχυπρόθεσμης μνήμης (short-term memory). Το πρόβλημα της βραχυπρόθεσμης μνήμης οφείλεται στον πολλαπλασιαστικό τρόπο διάδοσης του σφάλματος της εξόδου του δικτύου κατά την διάρκεια της εκπαίδευσης, που μπορεί να οδηγήσει σε πολύ μικρές τιμές των παραγώγων που χρησιμοποιούνται για την ενημέρωση των βαρών. Κατά την διάδοση του σφάλματος υπολογίζονται παράγωγοι της συνάρτησης κόστους η οποία μετράει την απόκλιση της εξόδου του δικτύου από την αναμενόμενη. Αυτές οι παράγωγοι σύμφωνα με τον κανόνα της αλυσίδας σχηματίζονται από όλους και περισσότερους πολλαπλασιαστικούς όρους καθώς το σφάλμα διαδίδεται προς τα πίσω. Επειδή λοιπόν στα αναδρομικά δίκτυα το σφάλμα διαδίδεται μέχρι την αρχή της ακολουθίας (backpropagation through time) οι πολλαπλασιαστικοί όροι αυξάνονται δραματικά οδηγώντας εν τέλει σε πολύ μικρούς αριθμούς ενημέρωσης των αρχικών βαρών αν πολλοί από αυτούς τους όρους προκύψουν μικροί.

$$\frac{\partial error}{\partial w1} = \frac{\partial error}{\partial output} * \frac{\partial output}{\partial hidden2} * \frac{\partial hidden2}{\partial hidden1} * \frac{\partial hidden1}{\partial w1}$$

Σχήμα 2.3: Εφαρμογή του κανόνα της αλυσίδας για την ενημέρωση τυχαίου βάρους w1

Αυτό έχει ως αποτέλεσμα, ότι όσο μεγαλώνει το μήκος της ακολουθίας εισόδου, οι πολλαπλασιασμοί αυξάνονται, και τόσο το δίκτυο δυσκολεύεται να συγκρατήσει όλο το ιστορικό της, και πολλή πληροφορία από την αρχή της ακολουθίας χάνεται. Συνεπώς σε μια πρόβλεψη που εξαρτάται από μεγάλο μήκος ακολουθίας, δεν θα δοθεί ισοδύναμο βάρος σε όλες τις εισόδους, αλλά πιο πολύ στις τελευταίες. Στην περίπτωση της αυτόματης παραγωγής περίληψης αν και οι πιο πρόσφατες λέξεις είναι πιο σημαντικές για την παραγωγή της επόμενης εξόδου από άποψη γραμματικής και σύνταξης πρέπει να δοθεί εξίσου πολύ σημασία και στις προηγούμενες για να αποφευχθούν επαναλήψεις και να προκύψει στο τέλος ένα ολοκληρωμένο νόημα.

Για την αντιμετώπιση του προβλήματος της βραχυπρόθεσμης μνήμης, έχουν κατασκευαστεί διαφορετικές εκδοχές αναδρομικών νευρωνικών δικτύων που φέρνουν πολύ καλά αποτελέσματα ανεξαρτήτως το μήκος της ακολουθίας. Τα πιο διαδεδομένα από αυτά που χρησιμοποιούνται και στις περισσότερες εφαρμογές είναι το δίκτυο LSTM (Long Short Term Memory) και το δίκτυο GRU (Gated Reccurent Unit). Παρακάτω θα αναλυθεί συνοπτικά η λειτουργία τους.

LSTM: Η κύρια διαφορά του LSTM δικτύου σε σχέση με το απλό αναδρομικό δίκτυο είναι ότι το LSTM δίκτυο ανατροφοδοτεί και μία επιπλέον κατάσταση στην είσοδο του η οποία ονομάζεται κατάσταση κυττάρου (cell state). Η κατάσταση κυττάρου θα μπορούσε να παραλληλιστεί με έναν ιμάντα μεταφοράς πληροφορίας, ο οποίος περνάει μέσα από διάφορες πύλες σε κάθε αναδρομή, οι οποίες προσθέτουν χρήσιμη ή αφαιρούν άχρηστη πληροφορία από τον ιμάντα. Η πληροφορία που βγαίνει τελικά στην έξοδο περιέχει έτσι μόνο χρήσιμη πληροφορία, η οποία μπορεί να προέρχεται από οποιοδήποτε σημείο της ακολουθίας εισόδου. Το δίκτυο καταλαβαίνει μόνο του πια πληροφορία της ακολουθίας είναι χρήσιμη και πρέπει να μπει στον ιμάντα και πια πληροφορία είναι άχρηστη και πρέπει να βγει από αυτόν, μέσω της διαδικασίας της εκπαίδευσης του. Το LSTM δίκτυο περιέχει 3 διαφορετικές πύλες που επιτελέσει τις απαραίτητες λειτουργίες του.

- **Πύλη αφαίρεσης (Forget Gate):** Η πύλη αφαίρεσης επιλέγει σε κάθε κύκλο αναδρομής πια πληροφορία θα σβηστεί από την κατάσταση κυττάρου. Η πληροφορία της προηγούμενης εξόδου του δικτύου ενώνεται με την τωρινή του είσοδο και μαζι περνάνε από μια σιγμοειδή συνάρτηση που φράζει την πληροφορία από 0 εώς 1. Η έξοδος αυτή πολλαπλασιάζεται ανά στοιχείο με την πληροφορία της κατάστασης κυττάρου. Όση πληροφορία προσεγγίζει το 0 μετά την σιγμοειδή συνάρτηση ”χάνεται” μέσω του πολλαπλασιασμού ενώ όση πληροφορία είναι κοντά στο 1 επιβιώνει.
- **Πύλη εισόδου (Input Gate):** Η πύλη εισόδου επιλέγει πια πληροφορία θα προστεθεί στην κατάσταση κυττάρου σε κάθε κύκλο της αναδρομής. Η πληροφορία της προηγούμενης εξόδου του δικτύου ενώνεται με την τωρινή του είσοδο και μαζί περνάνε από μια σιγμοειδή και από μια συνάρτηση υπερβολικής εφαπτομένης. Η πρώτη συνάρτηση φράζει την πληροφορία από το 0 εώς το 1, επιτρέποντας έτσι την επιλογή της σημαντικής πληροφορίας προς γράψιμο. Η δεύτερη συνάρτηση φράζει την πληροφορία από το -1 εώς το 1, ως ένα μέτρο κανονικοποίησης των τιμών σε συγκεκριμένο εύρος. Τα αποτελέσματα των δύο αυτών πυλών πολλαπλασιάζονται ανά σημείο μεταξύ τους έτσι ώστε να διατηρηθεί μόνο η σημαντική πληροφορία και προστίθενται στην κατάσταση κυττάρου.
- **Πύλη εξόδου (Output Gate):** Η πύλη εξόδου επιλέγει πια πληροφορία θα βγει στην έξοδο του δικτύου σε κάθε κύκλο αναδρομής. Η προηγούμενη έξοδος του δικτύου ενώνεται με την τωρινή του είσοδο και μαζί περνάνε από μια συνάρτηση υπερβολικής εφαπτομένης για λόγους κανονικοποίησης των τιμών στο -1 εώς 1. Το αποτέλεσμα της συνάρτηση πολλαπλασιάζεται ανά σημείο με την κατάσταση κυττάρου που έχει πρώτα περάσει από μια σιγμοειδή συνάρτηση και βγαίνει στην έξοδο του δικτύου.

Η μαθηματική περιγραφή της πύλης αφαίρεσης 2.1, της πύλης εισόδου 2.2 της πύλης εξόδου 2.3, της κατάστασης κυττάρου 2.4 και της κρυφής κατάστασης 2.5 του αναδρομικού δικτύου LSTM παρουσιάζονται παρακάτω:

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} * c_{t-1} + b_f) \tag{2.1}$$

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} * c_{t-1} + b_i) \tag{2.2}$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} * c_{t-1} + b_o) \tag{2.3}$$

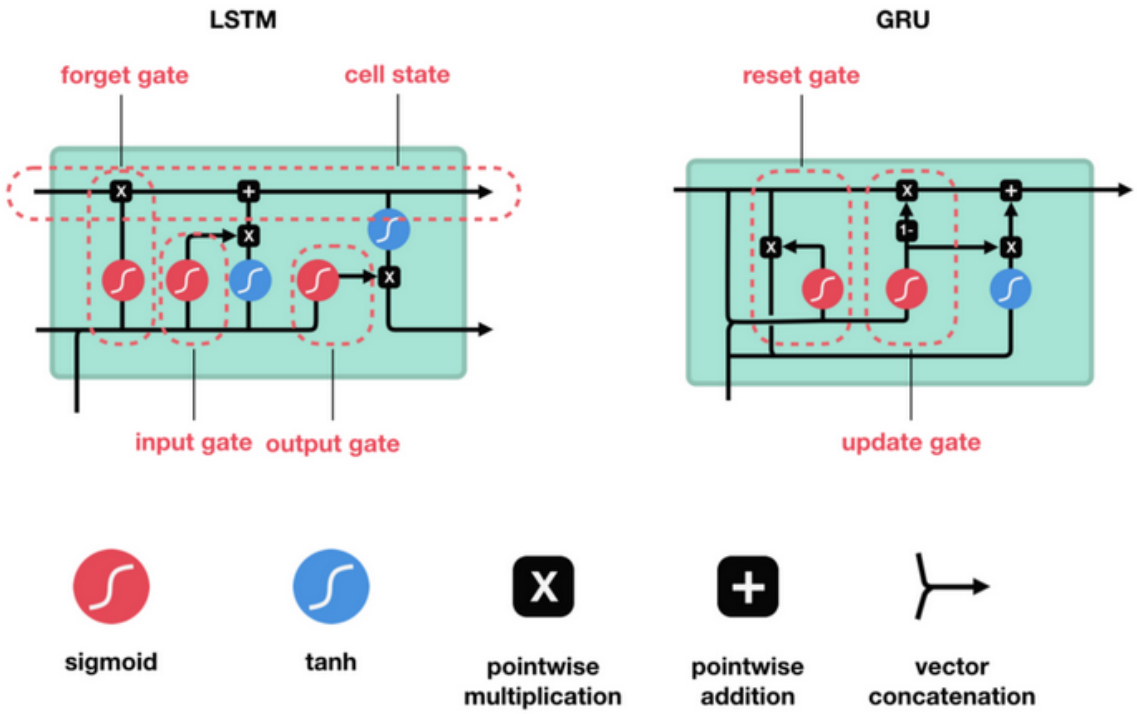
$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \tag{2.4}$$

$$h_t = o_t * \tanh(c_t) \tag{2.5}$$

όπου σ είναι η σιγμοειδής συνάρτηση και tanh η συνάρτηση υπερβολικής εφαπτομένης.

GRU: Η αρχιτεκτονική του GRU δικτύου μοιάζει αρκετά με την αρχιτεκτονική του LSTM αλλά είναι λίγο πιο απλουστευμένη. Η λειτουργία του θα μπορούσε να παραλληλιστεί πάλι με έναν ιμάντα πληροφορίας που περνάει από πύλες αλλά σε αυτή την περίπτωση τον ρόλο του ιμάντα τον αναλαμβάνει η εκάστοτε έξοδος του δικτύου (κρυφή κατάσταση). Έτσι σε κάθε κύκλο αναδρομής ανατροφοδοτείται στο δίκτυο μόνο μία ροή πληροφορίας και όχι δύο. Το GRU δίκτυο περιέχει 2 διαφορετικές πύλες για να επιτελέσει τις διάφορες λειτουργίες του, την πύλη ενημέρωσης, που χρησιμοποιείται για να ενημερώσει την πληροφορία της κρυφής κατάστασης, και την πύλη επαναφοράς, που χρησιμοποιείται για να αφαιρεθεί ασήμαντη πληροφορία. Οι λειτουργίες τους μοιάζουν πολύ με τις διάφορες λειτουργίες των πυλών του LSTM δικτύου και δεν θα αναλυθούν περαιτέρω.

Σαν μια προσπάθεια πολύ συνοπτικής εξήγησης του γιατί τα παραπάνω αυτά δίκτυα καταφέρουν να αντιμετωπίσουν το πρόβλημα της βραχυπρόθεσμης μνήμης, μπορεί να παρατηρήσει κανείς βλέποντας πως διαδίδεται το σφάλμα σε αυτά τα δίκτυα, ότι δεν υπάρχει συσσώρευση πολλαπλασιαστικών όρων αλλά συσσώρευση προσθετικών όρων. Αυτό σημαίνει πως καθώς το σφάλμα διαδίδεται όλο και πιο πίσω, αθροίζονται όλο και πιο πολλοί όροι στον τελικό όρο διάδοσης σφάλματος και δεν πολλαπλασιάζονται. Καθώς η πρόσθεση με πολύ μικρό αριθμό δεν έχει την ίδια επίπτωση στο συνολικό αποτέλεσμα σε σχέση με τον πολλαπλασιασμό με πολύ μικρό αριθμό, τα παραπάνω δίκτυα είναι πολύ ανθεκτικά στο να θυμούνται μεγαλύτερες ακολουθίες.



Σχήμα 2.4: Σχηματική αναπαράσταση LSTM και GRU δικτύου

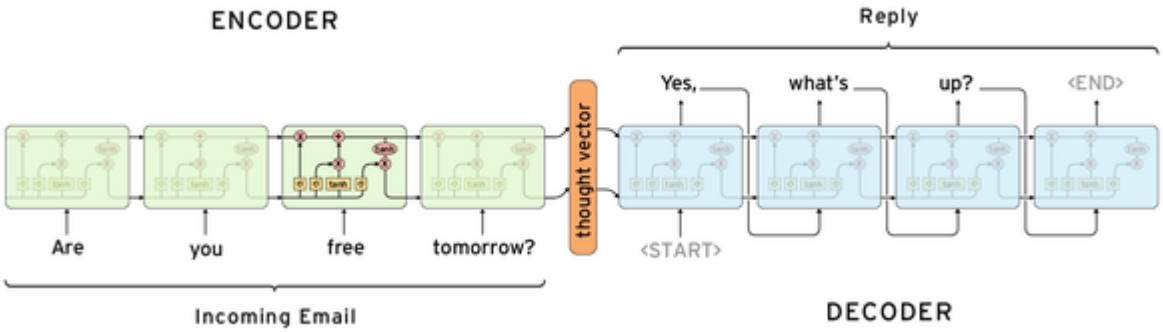
2.3.3 Αρχιτεκτονική κωδικοποιητή - αποκωδικοποιητή

Παρόλο που όπως είδαμε το αναδρομικό νευρωνικό δίκτυο και οι βελτιώσεις του μπορούν να μοντελοποιήσουν πλήρως της εξαρτήσεις σε μια ακολουθία λέξεων φυσικής γλώσσας, δεν αρκούν από μόνα τους για να αντιμετωπίσουν σύνθετα προβλήματα όπως η αυτόματη παραγωγή περίληψης κειμένου. Όταν ένας άνθρωπος πραγματοποιεί μια περίληψη ενός κειμένου, πρέπει πρώτα να διαβάσει ολόκληρο το κείμενο πριν αρχίσει να ξεχωρίζει τα σημαντικά κομμάτια του, στα οποία θα βασιστεί για την περίληψη. Δεν θα μπορούσε αντί αυτού να ξεκινήσει να διαβάζει το κείμενο ως μια ακολουθία λέξεως και να πραγματοποιεί ταυτόχρονα την περίληψη. Αντίστοιχα, στην περίπτωση παραγωγής περίληψης κειμένου μέσω υπολογιστή, δεν είναι σωστή πρακτική το αναδρομικό νευρωνικό δίκτυο να διαβάζει το κείμενο σαν ακολουθία και να πραγματοποιεί περίληψη σε αυτό. Θα πρέπει πρώτα με κάποιον τρόπο να διαβάσει ολόκληρο το κείμενο και να πραγματοποιήσει στην συνέχεια περίληψη, ώστε το τελικό αποτέλεσμα να περιέχει την ουσία του κειμένου και όχι απλή παραλλαγή των προτάσεων του.

Για την αντιμετώπιση του προβλήματος που αναλύθηκε παραπάνω έχει δημιουργηθεί και χρησιμοποιείται ευρέως με εξαιρετική επιτυχία η αρχιτεκτονική κωδικοποιητή - αποκωδικοποιητή. Όπως αναφέρει και το όνομα της, η αρχιτεκτονική αυτή νευρωνικού δικτύου αποτελείται από δύο κομμάτια που θα αναλυθούν παρακάτω.

- **Κωδικοποιητής:** Αποτελεί ουσιαστικά ένα αναδρομικό νευρωνικό δίκτυο με ένα ή παραπάνω στρώματα. Ο μοναδικός του ρόλος είναι να διαβάσει όλη την είσοδο προς επεξεργασία και βγάλει στην έξοδο του μια μαθηματική αναπαράσταση όλης της εισόδου. Η έξοδος που προκύπτει έχει μορφή διανύσματος και ονομάζεται διάνυσμα συμφραζομένων (context vector). Σε περίπτωση που η είσοδος του κωδικοποιητή είναι μια ακολουθία, το διάνυσμα θα περιέχει πληροφορία για ολόκληρη αυτή.

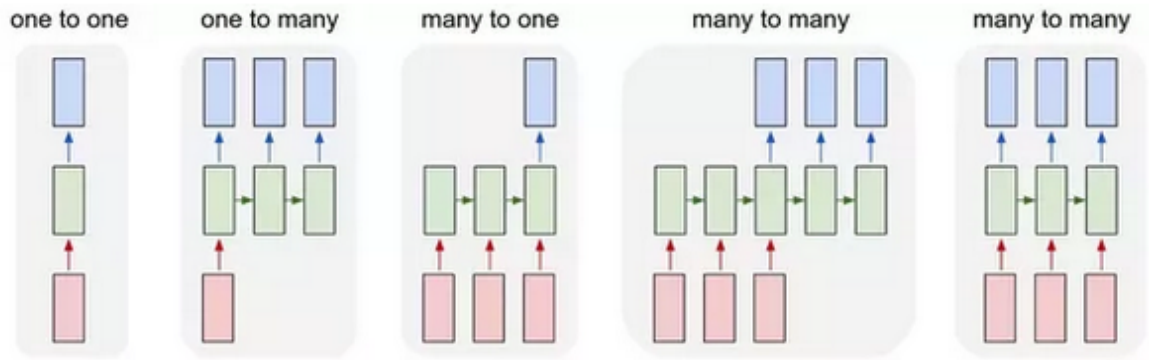
- **Αποκωδικοποιητής:** Αποτελεί πάλι ένα αναδρομικό νευρωνικό δίκτυο με ένα η παραπάνω στρώματα. Ο ρόλος του είναι να διαβάσει την πληροφορία που βγαίνει στην έξοδο του κωδικοποιητή και να παράξει στην έξοδο του κάποιο αποτέλεσμα. Σε περίπτωση που το αποτέλεσμα που αναμένεται είναι μια ακολουθία, ο αποκωδικοποιητής σε κάθε κύκλο της αναδρομής τροφοδοτεί την εκάστοτε έξοδο του ξανά στην είσοδό του για να παράξει την επόμενη έξοδο.



Σχήμα 2.5: Χρήση κωδικοποιητή - αποκωδικοποιητή για αυτόματη απάντηση σε email

Η αρχιτεκτονική κωδικοποιητή - αποκωδικοποιητή δεν είναι όμως η μόνη ευρέως διαδεδομένη αρχιτεκτονική αναδρομικών νευρωνικών δικτύων. Παρακάτω παρουσιάζονται και αναλύονται όλες οι αρχιτεκτονικές που χρησιμοποιούνται για να αντιμετωπίσουν διάφορα προβλήματα.

- **Ένα σε Ένα (One-to-One):** Μια είσοδος μία έξοδος. Ουσιαστικά σε αυτήν την περίπτωση δεν υπάρχει κάποια μορφή ακολουθίας και η αρχιτεκτονική εκφυλίζεται σε αρχιτεκτονική απλού εμπρόσθιου νευρωνικού δικτύου.
- **Ένα σε Πολλά (One-to-Many):** Μια είσοδος και μια ακολουθία ως έξοδος. Σε αυτή την περίπτωση το νευρωνικό δίκτυο καλείται με σύντομη πληροφορία εισόδου να παράξει ολόκληρη ακολουθία ως αποτέλεσμα. Ένα παράδειγμα όπου εφαρμόζεται αυτή η αρχιτεκτονική είναι η αυτόματη παραγωγή εικόνας με κείμενο.
- **Πολλά σε Ένα (Many-to-One):** Μια ακολουθία ως είσοδος και μια έξοδος. Σε αυτή την περίπτωση το νευρωνικό δίκτυο καλείται με μια ακολουθία στην είσοδο να παράξει μια έξοδο. Ένα παράδειγμα εφαρμογής αυτής της αρχιτεκτονικής είναι ο εντοπισμός συναισθήματος σε κείμενο (sentiment analysis), στο οποίο συνήθως το δίκτυο τροφοδοτείται με ένα κείμενο ως ακολουθία και εξάγει συνήθως έναν αριθμό.
- **Πολλά σε Πολλά (Many-to-Many):** Μια ακολουθία ως είσοδος και μια ακολουθία ως έξοδος. Αυτή η αρχιτεκτονική χρησιμοποιείται σε προβλήματα που το νευρωνικό δίκτυο καλείται να παράξει μια ακολουθία, δεδομένης μιας άλλης ακολουθίας στην είσοδο του. Παραδείγματα εφαρμογής αυτής της αρχιτεκτονικής αποτελούν η αυτόματη μετάφραση κειμένου καθώς και η αυτόματη παραγωγή περίληψης κειμένου. Αυτή η αρχιτεκτονική ονομάζεται και αλλιώς ακολουθία-σε-ακολουθία (sequence-to-sequence).



Σχήμα 2.6: Διαδεδομένες αρχιτεκτονικές αναδρομικών νευρωνικών δικτύων

Με τη χρήση λοιπόν της αρχιτεκτονικής ακολουθία-σε-ακολουθία, ο αποκωδικοποιητής λαμβάνει μέσω του κωδικοποιητή μια ολική αναπαράσταση του κειμένου προς περίληψη πριν αρχίσει να την παράγει. Έτσι στο τελικό αποτέλεσμα αποτυπώνεται καλύτερα η ουσία του αρχικού κειμένου και το σύστημα παραγωγής περιλήψεων αποκτά περισσότερη αξία. Η αρχιτεκτονική ακολουθία-σε-ακολουθία μπορεί να περιγραφεί αλλιώς ως ο συνδυασμός των αρχιτεκτονικών πολλά-σε-ένα και ένα-σε-πολλά.

Αυτή η αρχιτεκτονική έχει καταφέρει να φέρει την επανάσταση στον τομέα της επεξεργασίας φυσικών γλωσσών. Καθώς όμως η έρευνα στο αντικείμενο προχωράει, εμφανίζονται σιγά σιγά και άλλες βελτιώσεις πάνω της που καταφέρνουν να δώσουν ακόμη καλύτερα αποτελέσματα. Στην συνέχεια θα παρουσιαστούν τρεις τροποποιήσεις του δικτύου που αναλύθηκε παραπάνω με στόχο τα τελικά αποτελέσματα να είναι ακόμη καλύτερα και πιο ακριβή.

2.3.4 Αμφίδρομος κωδικοποιητής

Ο κλασσικός κωδικοποιητής, σε περιπτώσεις που η είσοδος έχει μορφή ακολουθίας, την διαβάζει σταδιακά και την μετατρέπει σε ένα διάνυσμα συμφραζομένων σταθερού μεγέθους, το οποίο κωδικοποιεί όλη την πληροφορία εισόδου. Μπορεί να καταλάβει κανείς ότι ο τρόπος που κωδικοποιητής επεξεργάζεται την πληροφορία εισόδου είναι σειριακός που σημαίνει ουσιαστικά ότι προσπελάζει της στοιχεία της ακολουθίας από το πρώτο εώς το τελευταίο με ορθή σειρά. Για παράδειγμα στο πρόβλημα της αυτόματης παραγωγής περίληψης ο κωδικοποιητής θα διαβάσει το κείμενο από την αρχή μέχρι το τέλος κατά την επεξεργασία του.

Καθώς όμως στον παράδειγμά μας, κάθε λέξη που εισχωρεί στο δίκτυο του κωδικοποιητή χρησιμοποιεί και το ιστορικό των προηγούμενων λέξεως της ακολουθίας εξαιτίας της αναδρομής, σε κάθε βήμα αυτής η πληροφορία εξόδου είναι μοντελοποιημένη να δείχνει την εξάρτηση της από προηγούμενες λέξεις. Όταν ο αποκωδικοποιητής λοιπόν λάβει την έξοδο του κωδικοποιητή και αρχίσει να παράγει την περίληψη, η πληροφορία ουσιαστικά που επεξεργάζεται είναι όπως αναφέρθηκε και πιο πριν μια σύνοψη όλου του κειμένου προς περίληψη αλλά και δομημένο με ορθή σειρά, έτσι ώστε να δηλώνονται καλύτερα οι εξαρτήσεις από προηγούμενες λέξεις.

Το πρόβλημα που προκύπτει με αυτόν τον μηχανισμό είναι ότι σε συγκεκριμένα προβλήματα όπου η είσοδος είναι ακολουθία όπως η αυτόματη μετάφραση κειμένου ή η αυτόματη παραγωγή περίληψης, δεν αρκεί η πληροφορία να είναι μοντελοποιημένη δείχνοντας μόνο εξαρτήσεις από προηγούμενα στοιχεία της ακολουθίας. Όταν ένας άνθρωπος παράγει μια περίληψη σε ένα κείμενο πολλές φορές χρειάζεται να κοιτάξει και πιο μπροστά στο κείμενο από το σημείο που αναλύει κάθε δεδομένη στιγμή. Αντίστοιχα ισχύει και με την μετάφραση κειμένου. Ανάλογα την σύνταξη κάθε γλώσσας, πολλές λέξεις για να μεταφραστούν σε μία πρόταση απαιτούν να έχει διαβαστεί όλη αυτή.

Σε μια προσπάθεια καλύτερης μοντελοποίησης της εξάρτησης του εκάστοτε στοιχείου της ακολουθίας από επόμενα στοιχεία του, ενώ ταυτόχρονα να διατηρείται και η εξάρτηση από προηγούμενα έχουν οριστεί οι αμφίδρομοι κωδικοποιητές. Ένας αμφίδρομος κωδικοποιητής λειτουργεί ακριβώς όπως ένας απλός κωδικοποιητής, απλώς κάνει την διπλάσια ”δουλειά” και κωδικοποιεί την ακολουθία εισόδου όχι μόνο από την αρχή προς το τέλος, αλλά και από το τέλος προς την αρχή. Τα αποτελέσματα συσσωρεύονται σε διαφορετικά διανύσματα, τα οποία μπορεί να χρησιμοποιήσει ο αποκωδικοποιητής στην συνέχεια για να εξάγει την πληροφορία του. Η αναγκαιότητα χρήσης του αμφίδρομου κωδικοποιητή γίνεται ακόμη πιο έντονη και σαφής στον μηχανισμό βελτίωσης που θα παρουσιαστεί στην συνέχεια, ο οποίος απαιτεί την ύπαρξη αμφίδρομου αποκωδικοποιητή για την βέλτιστη λειτουργία του.

2.3.5 Μηχανισμός προσοχής

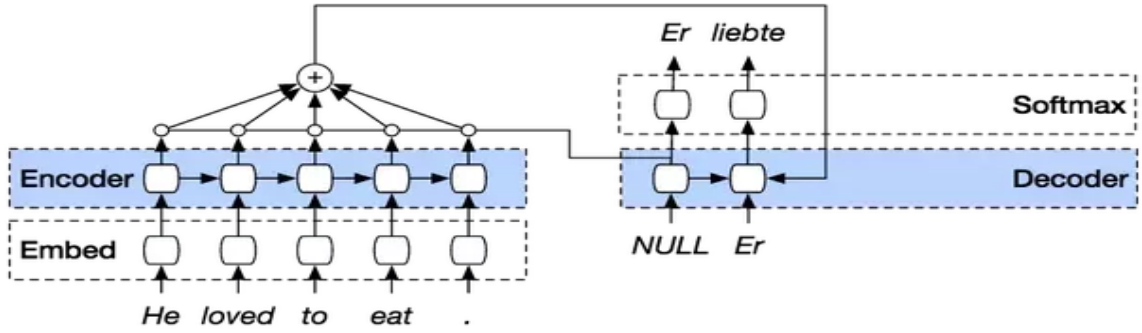
Παρατηρώντας κανείς τους εξηγήσεις που δίνονται σε αυτό το κείμενο για τους λόγους που χρησιμοποιούνται οι διάφοροι μηχανισμοί που έχουν αναφερθεί παραπάνω, μπορεί να δει ότι δίνεται πολύ έμφαση στην διαδικασία που ο άνθρωπος θα αντιμετώπιζε το ίδιο πρόβλημα που προσπαθεί να λυθεί στην εργασία αυτή με την χρήση του υπολογιστή, και στο γεγονός ότι οι μηχανισμοί αυτοί που εφαρμόζονται προσπαθούν να μιμηθούν αυτή την διαδικασία. Ανέκαθεν ο τρόπος διερεύνησης στρατηγικών μηχανικής μάθησης αφορούσε την απομίμηση της διαδικασίας που οι άνθρωποι σκέφτονται και αντιμετωπίζουν διάφορα προβλήματα ως μια προσπάθεια προσέγγισης αυτής της αφαιρετικότητας και πρωτοβουλίας που μπορεί να προσφέρει μια ανθρώπινη σκέψη. Σε αυτό το πλαίσιο θα παρουσιαστεί άλλον ένας μηχανισμός εμπνευσμένος πάλι από την ανθρώπινη διαδικασία παραγωγής περίληψης. Αυτός ο μηχανισμός, που ονομάζεται μηχανισμός προσοχής (attention mechanism), παρουσιάστηκε πρώτη φορά κατά την προσπάθεια αντιμετώπισης του προβλήματος της αυτόματης μετάφρασης κειμένου [Bahd14] και από τότε έχει διαδοθεί πάρα πολύ και έχει φέρει επαναστατικά αποτελέσματα σε πολλούς τομείς της βαθιάς μάθησης, κυρίως σε συστήματα επεξεργασίας εικόνων και κειμένων.

Όπως έχει αναφερθεί μέχρι στιγμής, το μοντέλο κωδικοποιητή-αποκωδικοποιητή επιτρέπει στον αποκωδικοποιητή να έχει επίγνωση ολόκληρης της ακολουθίας εισόδου κατά την επεξεργασία του. Έτσι κατά την παραγωγή περιλήψεων μπορεί να συνυπολογίζει όλο το κείμενο για την εξαγωγή κάθε λέξης, οδηγώντας σε καλύτερα αποτελέσματα. Κατά την παραγωγή όμως μια περίληψης ενός κειμένου, δεν αρκεί να υπάρχει πλήρης επίγνωση αυτού, αλλά θα πρέπει να υπάρχει και ένας μηχανισμός που θα υποδεικνύει πιο είναι το πιο σημαντικό κομμάτι του κειμένου εισόδου για την εξαγωγή κάθε λέξης της περίληψης. Αν σκεφτεί κανείς ξανά πως ο άνθρωπος πραγματοποιεί μια περίληψης, μπορεί να παρατηρήσει, ότι παρόλο που ο άνθρωπος θέλει να έχει πλήρη επίγνωση όλου του κειμένου για να έχει καλύτερα αποτελέσματα, κάθε στιγμή θα στρέφει την προσοχή του σε συγκεκριμένα σημεία αυτού, συγκεκριμένα στα σημεία που αναφέρεται στην περίληψη του κάθε στιγμή. Ουσιαστικά, ο λόγος που θέλει να γνωρίζει όλο το κείμενο πριν αρχίσει την διαδικασία παραγωγής, είναι για να ξέρεις που είναι τα κρίσιμα σημεία που θα πρέπει να εστιάσει στην συνέχεια.

Ο μηχανισμός προσοχής έχει δημιουργηθεί λοιπόν για να επιτρέπει στον αποκωδικοποιητή να εστιάζει σε συγκεκριμένα σημεία της ακολουθίας εισόδου κατά την εξαγωγή της εξόδου του. Μέσω αυτής της ”προσοχής” που μπορεί να αποδώσει, τα αποτελέσματα είναι πολύ καλύτερα, καθώς ο αποκωδικοποιητής μπορεί να συμβουλευτεί μικρότερες ακολουθίες και όχι όλη την ακολουθία εισόδου, σε κάθε έξοδό του.

Για την λειτουργία του μηχανισμού προσοχής, η κλασσική αρχιτεκτονική του κωδικοποιητή-αποκωδικοποιητή αλλάζει δραστικά. Πλέον ο αποκωδικοποιητής δεν λαμβάνει μόνο της τελευταία κατάσταση του κωδικοποιητή ως είσοδο, αλλά όλες τις καταστάσεις του που προκύπτουν κατά το διάβασμα της ακολουθίας εισόδου. Η φιλοσοφία είναι ότι κάθε κατάσταση που συμβουλεύεται ο αποκωδικοποιητής θα πρέπει να εστιάζει σε μία λέξη της ακολουθίας εισόδου, αλλά να περιέχει και όλη την ακολουθία. Για να γίνει αυτό εφικτό, χρησιμοποιείται αμφίδρομος κωδικοποιητής ο οποίος επεξεργάζεται την ακολουθία και από τις δύο κατευθύνσεις και για κάθε στοιχείο αυτής, δημιουργεί μια συνολική κατάσταση που ισούται με την ένωση των δύο αμφίδρομων καταστάσεων που καταλήγουν στο εκάστοτε στοιχείο. Έτσι η συνολική κατάσταση για κάθε στοιχείο της ακολουθίας περιέχει δύο διανύσματα, που αφορούν τόσο το διάβασμα από την αρχή της ακολουθίας μέχρι το εκάστοτε στοιχείο, αλλά και το διάβασμα από το τέλος της ακολουθίας μέχρι το εκάστοτε στοιχείο. Λόγω του φαινομένου ότι το κάθε διάνυσμα από αυτά εστιάζει στις πιο πρόσφατες εισόδους του, ο συνδυασμός των δύο διανυσμάτων προκύπτει να είναι ένα διάνυσμα που περιλαμβάνει όλη την ακολουθία εισόδου και εστιάζει στην εκάστοτε είσοδο αναφοράς.

Αφού λοιπόν έχουν υπολογιστεί τα διανύσματα που εστιάζουν σε κάθε στοιχείο της ακολουθίας εισόδου, υπολογίζεται για κάθε ένα από αυτά τα διανύσματα μια μετρική που δείχνει πόσο ”κοντά” είναι το εκάστοτε διάνυσμα με την κάθε έξοδο του δικτύου. Ο υπολογισμός αυτής της μετρικής πραγματοποιείται με κάποιο νευρωνικό δίκτυο και περιέχει μεταβλητές που μπορούν να εκπαιδευτούν, επιτρέποντας στο σύστημα να αποφαίνεται ορθά μόνο του πια διανύσματα είναι πιο κοντά στην κάθε έξοδο του. Έτσι ο αποκωδικοποιητής μετά την εκπαίδευση μπορεί συμβουλεύοντας την μετρική, να συμβουλευτεί τα κατάλληλα διανύσματα κατά την εξαγωγή κάθε στοιχείου ακολουθίας και τα αποτελέσματα προσεγγίζουν καλύτερα τα ζητούμενα.



Σχήμα 2.7: Οπτική αναπαράσταση του μηχανισμού προσοχής, κατά την διαδικασία αυτόματης μετάφρασης κειμένου

Η μαθηματική έκφραση του μηχανισμού προσοχής που χρησιμοποιείται στα πλαίσια αυτής της εργασίας, όπως περιγράφηκε πρώτη φορά στο [Bahd14] φαίνεται παρακάτω:

Στην αρχή υπολογίζονται οι αμφίδρομες κωδικοποιημένες καταστάσεις h_j για κάθε στοιχείο j της ακολουθίας εισόδου. Από αυτές της καταστάσεις υπολογίζεται το διάνυσμα περιεχομένου για κάθε λέξη i της ακολουθίας εξόδου ως σταθμισμένο άθροισμα των κωδικοποιημένων καταστάσεων όπως φαίνεται στην 2.6.

$$c_i = \sum_j a_{ij} h_j \tag{2.6}$$

Τα βάρη αυτού του σταθμισμένου αθροίσματος προκύπτουν από τις εξισώσεις 2.7, 2.8:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})} \tag{2.7}$$

$$e_{ij} = \alpha(s_{i-1}, h_j) \tag{2.8}$$

όπου η συνάρτηση α αποτελεί μια συνάρτηση βαθμολόγησης που καθορίζει πόσο σημαντική είναι η εκάστοτε κωδικοποιημένη κατάσταση h_j για την εξαγωγή της εξόδου y_t του αποκωδικοποιητή δεδομένης της κρυφής κατάστασης s_{t-1} του αποκωδικοποιητή. Η συγκεκριμένη συνάρτηση είναι εκπαιδευόμενη και συνήθως χρησιμοποιείται ένα απλό νευρωνικό δίκτυο για την μοντελοποίηση της.

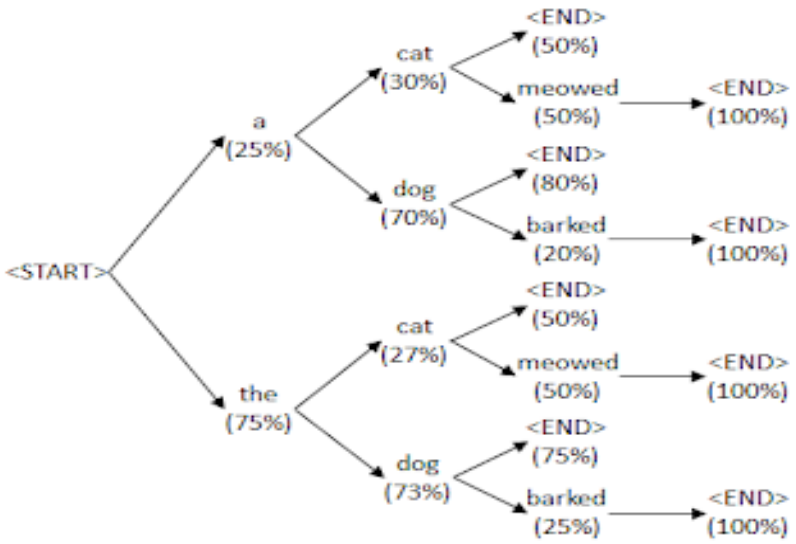
Σε κάθε βήμα της διαδικασίας παραγωγής της περίληψης από τον αποκωδικοποιητή λοιπόν, αυτός πλέον δεν τροφοδοτείται μόνο με την προηγούμενη κρυφή κατάσταση του, αλλά και με το διάνυσμα περιεχομένου που του υποδεικνύει τα πιο σημαντικά στοιχεία του κειμένου για την εξαγωγή της κάθε λέξης περίληψης

2.3.6 Ακτινική αναζήτηση

Μια ακόμη βελτίωση του μηχανισμού κωδικοποιητή-αποκωδικοποιητή, αφορά την διαδικασία παραγωγής του στοιχείου εξόδου από τον αποκωδικοποιητή. Όλα τα νευρωνικά δίκτυα αναπαριστούν αριθμητικές πράξεις και τα τελικά αποτελέσματα προκύπτουν πάντα σε αριθμητική μορφή. Σε προβλήματα επεξεργασίας φυσικών γλωσσών υπάρχει η αναγκαιότητα για μετατροπή της εξόδου του νευρωνικού δικτύου σε κάποια λέξη. Στις περισσότερες περιπτώσεις τέτοιων προβλημάτων υπάρχει κάποιο συγκεκριμένο μήκους λεξιλογίου, δηλαδή πλήθος λέξεων που μπορεί να προκύψουν στο αποτέλεσμα μας. Μέσα στο δίκτυο το πλήθος αυτό παριστάνεται στην έξοδο του δικτύου με ακμές εξόδου ίσες σε πλήθος με το μέγεθος του λεξιλογίου. Όλες αυτές οι ακμές λαμβάνουν κάποια τιμή και το ζήτημα είναι πια από αυτές τις ακμές θα επιλεχθεί και θα μεταφραστεί τελικά ως λέξη.

Ο κλασσικός αλγόριθμος που πραγματοποιεί αυτή την επιλογή, είναι ένας άπληστος αλγόριθμος που απλώς επιλέγει την ακμή με την μεγαλύτερη τιμή. Η φιλοσοφία πίσω από αυτό είναι πως οι ακμές εξόδου παριστάνουν μια πιθανοτική κατανομή των λέξεων και η ακμή με την μεγαλύτερη τιμή δίνει την πιο πιθανή βέλτιστη λέξη σε κάθε βήμα την αναδρομής του αποκωδικοποιητή. Ύστερα από αυτό η μετατροπή στην αντίστοιχη λέξη είναι εύκολη καθώς απλά επιλέγεται η λέξη που έχει τον ίδιο αύξων αριθμό με την ακμή. Το πρόβλημα με αυτή τη στρατηγική είναι η επιλογή της ακμής με την μεγαλύτερη τιμή σε κάθε βήμα δεν εξασφαλίζει ότι το συνολικό σκορ στο τέλος της ακολουθίας θα είναι το μέγιστο (το σκορ συνήθως σε αυτή την περίπτωση ορίζεται ως το γινόμενο των τιμών των επιλεγμένων ακμών, αφού έχουν περάσει πρώτα από λογαριθμική συνάρτηση για εξομάλυνσή τους). Αυτό ακριβώς το πρόβλημα επιδιώκει να λύση η ακτινική αναζήτηση (beam search).

Η ακτινική αναζήτηση στην πραγματικότητα δεν διαφέρει πολύ από τον κλασσικό άπληστο αλγόριθμο επιλογής. Η μόνη διαφορά είναι πως αντί να εξερευνεί μόνο ένα μονοπάτι λύσης κάθε φορά, μελετάει λύσεις βάθους k. Αυτό ουσιαστικά σημαίνει πως σε κάθε έξοδο του αποκωδικοποιητή, επιλέγει τις k μεγαλύτερες τιμές ακμών και στην συνέχεια εξερευνεί κάθε μία από αυτές, υπολογίζοντας το σκορ σε κάθε βήμα και επιλέγοντας ξανά και ξανά τις k πρώτες επιλογές. Στο τέλος της πρόβλεψης, η ακτινική αναζήτηση επιστρέφει την ακολουθία με το συνολικό μεγαλύτερο σκορ. Μπορεί να παρατηρήσει κανείς ότι ουσιαστικά ο κλασσικός άπληστος αλγόριθμος επιλογής αποτελεί και αυτός μια μορφή ακτινικής αναζήτησης με $k = 1$.



Σχήμα 2.8: Οπτική απεικόνιση ακτινικής αναζήτησης

2.3.7 Αποφυγή υπερεκπαίδευσης

Μια ολόκληρη κατηγορία τεχνικών που χρησιμοποιούνται εκτενώς σε όλα είδη των μοντέλων μηχανικής μάθησης και είναι πολύ σημαντική για την ορθή λειτουργίας τους είναι η κατηγορία των τεχνικών συστηματοποίησης (Regularization techniques). Ο κύριος λόγος χρήσης των τεχνικών συστηματοποίησης είναι η προσπάθεια μείωσης του κινδύνου υπερεκπαίδευσης που μπορεί να προκύψει σε ένα μοντέλο μηχανικής μάθησης, η οποία μπορεί να το οδηγήσει σε αδυναμία γενίκευσης σε άγνωστα δεδομένα εισόδου. Παρακάτω θα παρουσιαστούν συνοπτικά οι πιο κλασσικές τεχνικές που χρησιμοποιούνται για συστηματοποίηση ενώ θα δοθούν και εναλλακτικές χρηστικότητες αυτών.

Η πιο κλασσική τεχνική συστηματοποίησης αφορά την εισαγωγή επιπρόσθετου όρου συστηματοποίησης την συνάρτηση κόστους. Ο όρος αυτός αποτελεί συνάρτηση μόνο των μεταβλητών εκπαίδευσης του εκάστοτε μοντέλου και όχι των δεδομένων εισόδου και ρόλος του αποτελεί να οδηγήσει το εκάστοτε μοντέλο σε απλούστερες λύσεις. Όπως είναι γνωστό, τα περισσότερα μοντέλα μηχανικής μάθησης προσπαθούν να προσεγγίσουν μια μαθηματική συνάρτηση με βάση τα δεδομένα εκπαίδευσης και στην συνέχεια να προβλέψουν με βάση νέα δεδομένα. Για κάθε σύνολο δεδομένων εισόδου όμως πιθανώς δεν υπάρχει μοναδική συνάρτηση που τα περιγράφει βέλτιστα. Μέσω αυτού του όρου λοιπόν σε περίπτωση πολλαπλών λύσεων, προτιμώνται οι πιο απλοικές. Φυσικά υπάρχει διαφορετική ερμηνεία για το τι σημαίνει απλοική λύση και αυτή εξαρτάται από το είδος του προβλήματος και των δεδομένων. Θα πρέπει να δοθεί προσοχή λοιπόν κατά την επιλογή χρήσης τέτοιου όρου. Δύο πολύ κλασσικοί όροι συστηματοποίησης που χρησιμοποιούνται ευρέως είναι ο όρος L1 και ο όρος L2. Αξίζει να σημειωθεί ότι οι όροι συστηματοποίησης δεν χρησιμοποιούνται τόσο πολύ στα νευρωνικά δίκτυα, αλλά περισσότερο σε συστήματα στατιστικής ανάλυσης.

Μια άλλη πολύ γνωστή τεχνική συστηματοποίησης που χρησιμοποιείται είναι η τεχνική της κανονικοποίησης των τεμαχίων εισόδου (batch normalization). Αυτή η τεχνική είναι πολύ ισχυρή και έφερε την επανάσταση στον τομέα της όρασης υπολογιστή. Σε αυτό τον τομέα τα νευρωνικά δίκτυα που απαιτούνται ώστε τα αποτελέσματα να είναι καλά είναι πολύ μεγάλα. Ένα πρόβλημα που παρουσιαζόταν πριν την χρήση αυτής της τεχνικής ήταν ότι η τιμή των παραγώγων ενημέρωσης των βαρών δεν επιβίωνε από το τέλος μέχρι την αρχή του δικτύου, σε πολύ βαθιά δίκτυα. Όπως έχει αναλυθεί και πιο πριν, μεγάλο πλήθος πολύ μικρών όρων στην πολλαπλασιαστική αυτή διαδικασία ενημέρωσης των βαρών μπορούν οριακά να μηδενίσουν την πληροφορία ενημέρωσης στα πρώτα στρώματα. Αυτή η τεχνική βοήθησε πολύ στην αντιμετώπιση αυτού του προβλήματος εισάγοντας ειδικά στρώματα σε νευρωνικά δίκτυα, τα οποία κανονικοποιούν τα δεδομένα που περνάνε από αυτά ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Μέσω αυτής της κανονικοποίησης των δεδομένων σε διάφορα σημεία του δικτύου, εξασφαλίζεται πως αυτά έχουν ποικιλόμορφες τιμές και πως θα υπάρχουν παράγωγοι που θα διαδοθούν σίγουρα προς τα πίσω.

Τέλος, μια πολύ διαδεδομένα τακτική συστηματοποίησης που χρησιμοποιείται σε νευρωνικά δίκτυα είναι η τυχαία πτώση κόμβων (dropout). Σύμφωνα με αυτήν την τεχνική, σε κάθε εμπρόσθιο πέρασμα του δικτύου, τυχαίοι κόμβοι αποκόπτονται από το δίκτυο και συνεισφέρουν στο τελικό αποτέλεσμα της εξόδου. Κατά το οπίσθιο πέρασμα, αυτοί οι κόμβοι φυσικά δεν ενημερώνονται. Μέσω αυτού το τελικώς εκπαιδευμένο δίκτυο που θα προκύψει αποτελεί κατά μία έννοια μια συνάθροιση πολλαπλών διαφορετικών δικτύων με διαφορετικές παραμέτρους το καθένα. Αυτή η διαδικασία καθιστά το δίκτυο πολύ ανθεκτικό σε υπερεκπαίδευση, ενώ παράλληλα μειώνεται κατά πολύ ο χρόνος εκπαίδευσής του.

Στις αρχιτεκτονικές νευρωνικών δικτύων τύπου κωδικοποιητή-αποκωδικοποιητή κυρίως χρησιμοποιείται η τυχαία πτώση κόμβων ως τεχνική συστηματοποίησης. Έχουν γίνει προσπάθειες και για χρήση τεχνικών κανονικοποίησης τεμαχίων εισόδου, αλλά καθώς αυτές μοντελοποιείται δύσκολα σε προβλήματα όπου τα δεδομένα αποτελούν ακολουθίες, δεν προτιμούνται συχνά.

2.3.8 Συνάρτηση κόστους σε προβλήματα ταξινόμησης

Προβλήματα ταξινόμησης (classification problems) ονομάζονται τα προβλήματα της μηχανικής μάθησης στα οποία τα δεδομένα εισόδου πρέπει να καταταχθούν σε ορισμένες γνωστές εκ των προτέρων κατηγορίες. Σε αυτά τα προβλήματα, δοσμένου ενός στοιχείου εισόδου, πρέπει το εκάστοτε μοντέλο μηχανικής μάθησης να προβλέψει σε πια ή πιες κατηγορίες ανήκει. Τα προβλήματα αυτά ανήκουν στην κλάδο της επιβλεπόμενης μάθησης και για την εκπαίδευση τέτοιων μοντέλων απαιτούνται σύνολα δεδομένων που θα ορίζουν για διαφορετικά στοιχεία εισόδου τις κατηγορίες τους.

Κατά την εκπαίδευση λοιπόν προβλέπεται για κάθε στοιχείο εισόδου η κατηγορία στόχος του και στην συνέχεια υπολογίζεται η απόκλιση από την πραγματική κατηγορία (κόστος) ώστε να προσαρμοστούν οι μεταβλητές εκπαίδευσης. Το ερώτημα που προκύπτει φυσικά είναι πια συνάρτηση κόστους μπορεί να εκφράσει καλά την ιδιομορφία αυτών των μοντέλων που πραγματοποιούν προβλέψεις σε διακριτές, πεπερασμένες κατηγορίες.

Μία πολύ διαδεδομένη συνάρτηση κόστους που εφαρμόζεται ευρέως σε προβλήματα κατάταξης, ειδικά σε περίπτωση που τα προβλήματα αυτά αφορούν νευρωνικά δίκτυα, είναι η συνάρτηση απώλειας εντροπίας (crossentropy loss). Στην συνέχεια θα περιγράφει πως πραγματοποιείται ο υπολογισμός του κόστους μέσω της συνάρτησης απώλειας εντροπίας υποθέτοντας ένα πρόβλημα κανονικοποίησης n κλάσεων.

Σε ένα τέτοιο πρόβλημα ή έξοδος του μοντέλου μηχανικής μάθησης θα είναι ένα διάνυσμα μήκους n με αριθμητικές τιμές σε όλες τις θέσεις του, κάθε μια από τις οποίες αντιστοιχεί σε κάποια κλάση. Σκοπός είναι η μεγιστοποίηση της τιμής που βρίσκεται στην θέση της ορθής κλάσης και η ελαχιστοποίηση των άλλων όρων. Για τον υπολογισμό του κόστους μέσω της συνάρτησης μας θα υπολογιστούν καταρχάς οι κανονικοποιημένοι εκθετικοί όροι του διανύσματος εξόδου (softmax function). Στόχος αυτού του υπολογισμού είναι η κανονικοποίηση των τιμών εξόδου στο εύρος 0 εώς 1 έτσι ώστε το άθροισμά τους να ισούται με 1. Η συνάρτηση που το επιτυγχάνει αυτό φαίνεται στην εξίσωση 2.9

$$S(y_i) = \frac{e^{y_i}}{\sum e^{y_i}} \tag{2.9}$$

Με βάση αυτή την συνάρτηση ορίζεται το κόστος απώλειας εντροπίας όπως φαίνεται στην 2.10

$$J = -\frac{1}{N}(\sum y_t * \log(y_p)) \tag{2.10}$$

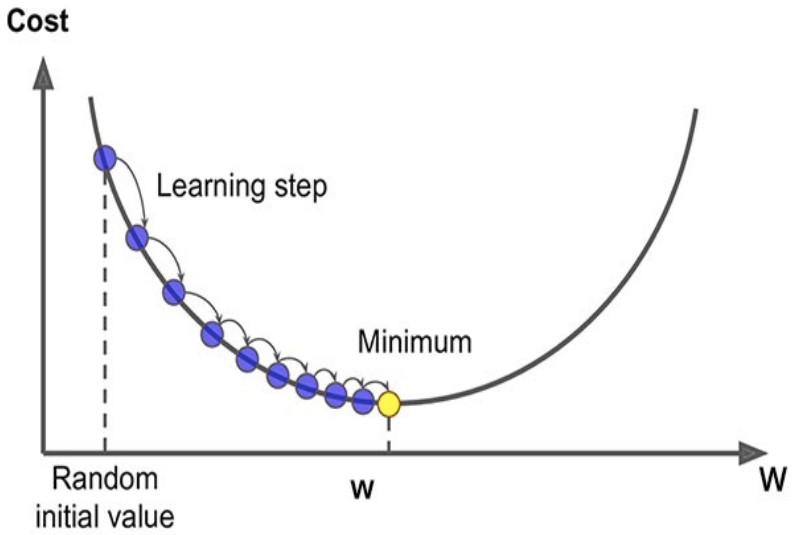
όπου yt είναι το διάνυσμα με τις πραγματικές κατηγορίες και yp το διάνυσμα πρόβλεψης. Καθώς το διάνυσμα με τις πραγματικές κατηγορίες έχει 1 μόνο στην ορθή κατηγορία και 0 στις άλλες το τελικό κόστος ισούται με τον αρνητικό μέσο όρο της λογαριθμικής πρόβλεψης της ορθής κατηγορίας για όλες τις εισόδους.

Η συνάρτηση αυτή υπολογισμού κόστους σε προβλήματα κατάταξης είναι πολύ σημαντική για το πρόβλημα της αυτόματης παραγωγής περίληψης με χρήση βαθιάς μάθησης καθώς στο τελικό στάδιό του, κάθε υλοποιημένο μοντέλο που αντιμετωπίζει αυτό το πρόβλημα ανάγεται σε πρόβλημα κατάταξης καθώς πρέπει να κατατάξει την έξοδο του μοντέλου σε τόσες κατηγορίες ίσες με το πλήθος του λεξιλογίου του αποκωδικοποιητή, κάθε μια από τις οποίες αντιστοιχεί σε μία λέξη εξόδου.

2.3.9 Βελτιστοποίηση μοντέλων μηχανικής μάθησης

Η βελτιστοποίηση μοντέλων μηχανικής μάθησης, αποτελεί μια πολύ σημαντική διαδικασία κατά την διάρκεια εκπαίδευσης μοντέλων και επηρεάζει πολύ τα τελικά αποτελέσματα και το κατά πόσο, το τελικώς εκπαιδευμένο σύστημα θα επιτυγχάνει τον αναμενόμενο σκοπό του. Μέσω της βελτιστοποίησης, τα διάφορα μοντέλα μηχανικής μάθησης μπορούν να αλλάζουν τις μεταβλητές τους προς εκπαίδευση ώστε οι προβλέψεις τους να προσεγγίζουν καλύτερα τις αναμενόμενες τιμές σε κάθε βήμα της εκπαίδευσης. Συμβουλεύοντας την τιμή κόστους που παράγει η συνάρτηση κόστους, ο εκάστοτε βελτιστοποιητής πρέπει να μεταδώσει αυτό το κόστος σε μορφή παραγώγων μεταβολής, μέσω της διαδικασίας της αλυσιδωτής παραγωγισης και να ενημερώσει όλες τις εκπαιδευόμενες μεταβλητές, σύμφωνα με τον κανόνα ενημέρωσης του.

Η διαδικασία της βελτιστοποίησης έχει περάσει πολλαπλά στάδια εξέλιξης. Τα πρώτα στάδια αυτής ξεκίνησαν με τον απλό αλγόριθμο κατάβασης της παραγώγου (gradient descent). Αυτός ο αλγόριθμος ήταν ο πρώτος αλγόριθμος που εισήγαγε την νοοτροπία εκπαίδευσης με χρήση παραγώγων όπως την ξέρουμε σήμερα. Η λογική πάνω στην οποία βασίστηκε, στηρίζεται στο γεγονός ότι η παράγωγος την συνάρτησης κόστους, η οποία υπολογίζει την απόκλιση από τις επιθυμητές τιμές σε κάθε βήμα της διαδικασίας, τείνει να ελαχιστοποιήσει την συνάρτηση κόστους, καθώς υποδηλώνει την κλίση της προς την ελάχιστη τιμή της. Γι' αυτό τον λόγω οι τιμές της παραγώγου της συνάρτησης κόστους, θα είναι μεγαλύτερες όσο πιο μακριά από το ελάχιστο της συνάρτησης κόστους βρίσκονται και θα εξομαλύνονται σιγά σιγά καθώς πλησιάζουν αυτό το ελάχιστο. Μέσω αυτής της μαθηματικής ιδιότητας μπορεί να εκφραστεί ένα μοντέλο εκπαίδευσης που θα χρησιμοποιεί τις αριθμητικές τιμές της παραγώγου της συνάρτησης κόστους ως προς όλες τις μεταβλητές εκπαίδευσης, ως βήμα εκπαίδευσης για την κάθε μεταβλητή. Μέσω αυτού του μοντέλου εκπαίδευσης τα βήματα για την ελαχιστοποίηση της συνάρτησης κόστους και για την τελική εκπαίδευση του μοντέλου θα είναι μεγαλύτερα όσο οι προβλέψεις αποκλίνουν πολύ από τις αναμενόμενες τιμές και θα μειώνονται όσο οι προβλέψεις τις προσεγγίζουν.



Σχήμα 2.9: Οπτική απεικόνιση αλγορίθμου κατάβασης παραγώγου

Καθώς η διαδικασία βελτιστοποίησης εξελίχτηκε, ορίστηκαν και χρησιμοποιούνται πλέον διαφορετικοί αλγόριθμοι βελτιστοποίησης, οι οποίοι προσεγγίζουν με λίγο διαφορετικό τρόπο την νοοτροπία εκπαίδευσης, ή προσπαθούν να αντιμετωπίσουν άμεσα, διάφορα προβλήματα που προκύπτουν από τον αλγόριθμο της κατάβασης παραγώγου. Στην συνέχεια θα παρουσιαστούν κάποιες εναλλακτικές μορφές βελτιστοποιητών που χρησιμοποιούνται πολύ σήμερα.

- **Στοχαστική κατάβαση παραγώγου:** Αυτός ο αλγόριθμος εισάγει έννοια της εκπαίδευσης με χρήση τεμαχίων και όχι αναφορικά με όλα τα δεδομένα εισόδου σε κάθε βήμα της διαδικασίας. Αυτή η τεχνική εκπαίδευσης χρησιμοποιείται πάρα πολύ σήμερα, ειδικά με την ανάπτυξη και την διάδοση των βαθιών νευρωνικών δικτύων.

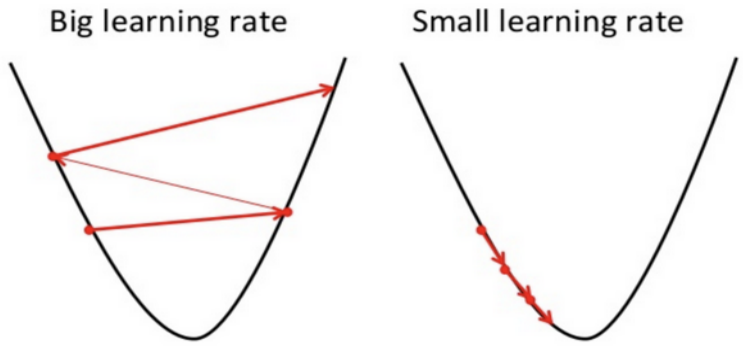
- **Στοχαστική κατάβαση με ορμή:** Η ορμή αποτελεί μια μαθηματική έννοια που ορίζει ένα μέτρο αδράνειας τους αλγορίθμους βελτιστοποίησης. Η στοχαστική κατάβαση παραγώγου με ορμή προσπαθεί να αντιμετωπίσει φαινόμενα τοπικών ελαχίστων καθώς και φαινόμενα ύπαρξης επίπεδων περιοχών χωρίς κλίση, τα οποία απαγορεύουν στον κλασσικό αλγόριθμο κατάβασης να τα υπερπηδήσει και αν συνεχίσει την αναζήτηση του για το ολικό ελάχιστο (αφού έχουν παράγωγο 0). Μέσω της ορμής, συνυπολογίζεται και η συνολική ταχύτητα πτώσης της παραγώγου σε κάθε βήμα, επιτρέποντας καλύτερη εκπαίδευση και αποφυγή τοπικών μη απότομων σημείων.

- **Προσαρμοστικός αλγόριθμος παραγώγου:** Ο προσαρμοστικός αλγόριθμος παραγώγου (adaptive gradient algorithm - Adagrad) χρησιμοποιεί μια διαφορετική προσέγγιση στην ενημέρωση των μεταβλητών προς βελτιστοποίηση. Για κάθε μεταβλητή, κρατάει ένα συνολικό άθροισμα των τετραγώνων των παραγώγων της συνάρτησης κόστους που χρησιμοποιούνται για την ενημέρωση της. Αυτό λειτουργεί σαν ιστορικό διαδρομής για κάθε μεταβλητή και έχει ανάλογη λειτουργία με αυτήν την ορμής αλλά προσαρμόζεται και στην κλίση πτώσης της συνάρτησης κόστους ως προς κάθε μεταβλητή, ώστε να διατηρεί σχετικά σταθερά βήματα βελτιστοποίησης σε κάθε βήμα εκπαίδευσης. Αυτό επιτυγχάνεται μέσω της διαίρεσης του εκάστοτε όρου μεταβολής της κάθε μεταβλητής με την τετραγωνική ρίζα του συνολικό αυτό άθροισμά των τετραγώνων των παραγώγων. Μέσω αυτού, αν η συνάρτηση κόστους έχει μικρή κλίση ως προς μια μεταβλητή (άρα μικρές παραγώγους) η μικρή αυτή μεταβολή διαιρείται με μια μικρή τιμή και ο συνολικός όρος μεγαλώνει. Αντίστοιχα, αν η μεταβολή της συνάρτησης κόστους ως προς μία μεταβλητή είναι μεγάλη, αυτή διαιρείται με μια μεγάλη τιμή και ο συνολικός όρος μικραίνει. Το πρόβλημα αυτού του αλγορίθμου είναι πως καθώς το άθροισμα παραγώγων ως προς κάθε μεταβλητή αυξάνεται συνεχόμενα, τα βήματα εκπαίδευσης γίνονται όλο και πιο μικρά κατά την διάρκεια αυτής, οδηγώντας εύκολα προβλήματα εκπαίδευσης σε περίπτωση κακού σχεδιασμού.

- **Αλγόριθμος διάδοσης μέσω μέσης τετραγωνικής ρίζας:** Ο αλγόριθμος διάδοσης μέσω τετραγωνικής ρίζας (root mean squared propagation - Rmsprop), μοιάζει πολύ με τον προσαρμοστικό αλγόριθμο παραγώγου, αλλά προσπαθεί να αντιμετωπίσει το φαινόμενο της ατέρμονης συσσώρευσης όρων στο άθροισμα κανονικοποίησης. Για να το επιτύχει αυτό, εισάγει έναν παράγοντα φθοράς, ο οποίος είναι μικρότερος του 1 και σε κάθε ενημέρωση τείνει να μειώσει το συνολικό άθροισμα. Έτσι η μείωση των βημάτων εκπαίδευσης μέσω της συσσώρευσης των όρων του αθροίσματος αντισταθμίζεται από αυτόν τον παράγοντα κάνοντας την εκπαίδευση λιγότερο περιοριστική.

- **Προσαρμοστικός εκτιμητής με ορμή:** Ο προσαρμοστικός εκτιμητής με ορμή (Adaptive momentum estimator - Adam) συνδυάζει τις παραπάνω διαφορετικές προσεγγίσεις που αναλύθηκαν σε έναν έυρωστο μηχανισμό βελτιστοποίησης. Αυτός ο μηχανισμός συνδυάζει την τεχνική της ορμής καθώς και την τεχνική κανονικοποίησης μέσω της τετραγωνικής ρίζας του αθροίσματος των τετραγώνων του ιστορικού των παραγώγων κρατώντας τα θετικά και από τις δύο μεθόδους. Ο αλγόριθμος αυτός, αποτελεί σήμερα έναν πολύ δυνατό αλγόριθμο βελτιστοποίησης με πολύ λίγες αδυναμίες σε σχέση με τους υπόλοιπους που περιγράφηκαν και αποτελεί μια πολύ καλή εισαγωγή σε σχεδόν όλα τα προβλήματα μηχανικής μάθησης.

Σημειώνεται ότι όλοι αυτοί οι αλγόριθμοι βελτιστοποίησης χρησιμοποιούν μια σταθερή τιμή βελτιστοποίησης η οποία λέγεται ρυθμός μάθησης (learning rate). Ο ρυθμός μάθησης αποτελεί την σταθερή τιμή που πολλαπλασιάζεται με το εκάστοτε βήμα ενημέρωσης που υπολογίζεται,και μαζί συνθέτουν την τελική τιμή ενημέρωσης. Η τιμή του ρυθμού μάθησης θέλει πολύ προσοχή κατά την προσαρμογή της και αποτελεί την πρώτη υπερπαράμετρο που πρέπει κανείς να διερευνά κατά την ανάπτυξη συστημάτων μηχανικής μάθησης. Πρέπει να προσέξει κανείς, ότι πολύ μεγάλες τιμές του ρυθμού μάθησης δυσκολεύουν το μοντέλο να συγκλίνει σε στο ελάχιστο της συνάρτησης κόστους, ενώ πολύ μικρές τιμές αυξάνουν δραματικά τους χρόνους εκπαίδευσης για την επίτευξη καλών αποτελεσμάτων.



Σχήμα 2.10: Προβλήματα που προκύπτουν σε πολύ μεγάλες και πολύ μικρές τιμές ρυθμού μάθησης

Κεφάλαιο 3

Σύνολα δεδομένων και Προεπεξεργασία

3.1 Γενικά

Η επιστήμη της μηχανικής μάθησης αποτελεί ουσιαστικά επιστήμη μελέτης των δεδομένων. Είναι πραγματικά αξιοσημείωτο πόση δύναμη μπορεί να έχει ένα μεγάλο πλήθος δεδομένων και πόση πληροφορία κρύβει μέσα του. Όσο η τεχνολογία εξελίσσεται και η συλλογή αλλά και η επεξεργασία δεδομένων γίνεται όλο και πιο γρήγορη και αποδοτική, φαίνεται πιο πολύ το πλήθος των τομέων στους οποίους οι επιστήμες των δεδομένων έχουν επίδραση. Μέσω μηχανικής μάθησης και στατιστικών αναλύσεων, μπορούν να προκύψουν αποτελέσματα και συμπεράσματα τα οποία ούτε μπορούν εύκολα να μοντελοποιηθούν με κάποιον συγκεκριμένο ντετερμινιστικό αλγόριθμο, ούτε να γίνουν εμφανή με τεχνικές ανθρώπινης επεξεργασίας. Σήμερα, λόγω του ότι η συλλογή δεδομένων έχει αποκτήσει τόση δύναμη, έχουν αρχίσει να εμφανίζονται ανησυχίες και αμφιβολίες από πολλούς για την ασφάλεια την ιδιωτικότητας τους, καθώς γίνεται όλο και πιο πολύ εμφανές ότι ο έλεγχος πολλών δεδομένων, αποτελεί πολύ ισχυρό χαρτί στην σημερινή κοινωνία.

Με την έξαρση της βαθιάς μάθησης, η ανάγκη για δεδομένα έγινε ακόμη πιο έντονη. Μέσω αυτής, είναι πλέον εφικτή η επίλυση τεράστιου πλήθους προβλημάτων, δεδομένου του κατάλληλου πλήθους δεδομένων. Η αρχιτεκτονική αυτών των δικτύων είναι τόσο γενική, που καθιστά δυνατή την προσέγγιση θεωρητικά οποιασδήποτε συνάρτησης, που θα μπορεί να παράξει τα ζητούμενα αποτελέσματα δοσμένης κατάλληλης εισόδου. Όσο όμως τα προβλήματα προς αντιμετώπιση γίνονται πιο σύνθετα, τόσο πιο σύνθετες γίνονται τέτοιες αρχιτεκτονικές και οι απαιτήσεις τους σε δεδομένα αυξάνονται συνεχώς.

Όσο όμως δύναμη και να κρύβουν τα δεδομένα, αυτή δεν θα γίνει ποτέ εμφανής κατά την μελέτη τους αν δεν έχει προηγηθεί κατάλληλη επεξεργασία αυτών. Συγκεκριμένα για την επιστήμη της μηχανικής μάθησης, η προεπεξεργασία των δεδομένων έτσι ώστε να είναι κατάλληλη μορφή για τον αλγόριθμο μάθησης είναι πάρα πολύ κρίσιμη και μπορεί να κάνει την διαφορά στην προσπάθεια δημιουργίας ενός εύρωστου συστήματος. Καθώς ο όγκος των δεδομένων μεγαλώνει συνέχεια, η προεπεξεργασία αυτών, γίνεται και αυτή με την σειρά της πιο απαιτητική. Σε πραγματικά συστήματα, τα δεδομένα που συλλέγονται δεν είναι κανονικοποιημένα και περιέχουν πολλά σφάλματα. Η προσπάθεια κανονικοποίησης και ”καθαρισμού” αυτών των δεδομένων αποτελεί πραγματική πρόκληση και απασχολεί μεγάλο μέρος επιστημόνων σήμερα.

Σε αυτό το κεφάλαιο θα γίνει μια παρουσίαση των συνόλων δεδομένων που θα χρησιμοποιηθούν για την αντιμετώπιση του προβλήματος της αυτόματης περίληψης κειμένου. Συγκεκριμένα θα αναλυθεί που βρέθηκαν τα σύνολα που θα χρησιμοποιηθούν, γιατί επιλέχθηκαν και θα περιγραφεί τυχόν προεπεξεργασία που είχε γίνει ήδη σε αυτά. Στην συνέχεια θα παρουσιαστεί η προεπεξεργασία τους που πραγματοποιήθηκε σε αυτή την εργασία ως μια προσπάθεια βελτιστοποίησης των αποτελεσμάτων που θα προκύψουν. Η προεπεξεργασία αυτη που πραγματοποιήθηκε, επέδρασε πολύ δραστικά στις τελικές προβλέψεις των υλοποιημένων μοντέλων, όπως θα φανεί και στην συνέχεια της εργασίας.

3.2 Σύνολα δεδομένων

- **Σύνολο Gigaword:** Το σύνολο δεδομένων Gigaword (Gigaword corpus) αποτελεί ένα από τα πιο διαδεδομένα σύνολα για την αντιμετώπιση του προβλήματος αυτόματης παραγωγής περίληψης. Η αρχική του επεξεργασία και εφαρμογή στο συγκεκριμένο πρόβλημα πραγματοποιήθηκε στο [Rush15] και από τότε έχει χρησιμοποιηθεί σε πολλές επιστημονικές εργασίες που αφορούν το συγκεκριμένο θέμα. Ο λόγος που επιλέχθηκε να χρησιμοποιηθεί το συγκεκριμένο σύνολο, αποτελεί το γεγονός ότι οι περισσότερες και πιο σύγχρονες προσεγγίσεις για την αντιμετώπιση του προβλήματος παραγωγής περίληψης αφαιρετικού χαρακτήρα το χρησιμοποιούν και συγκρίνουν την απόδοση των μοντέλων τους σε σχέση με άλλες επιστημονικές εργασίες πάνω σε αυτό.

  Πρέπει να σημειωθεί πως η δημιουργία ενός καλού συνόλου δεδομένων για την εκπαίδευση ενός συστήματος παραγωγής περίληψης δεν είναι εύκολη υπόθεση. Το εκάστοτε σύνολο απαιτείται να περιέχει αντιστοιχίες κειμένων περιλήψεων, πράγμα που είναι δύσκολο να συμβεί χωρίς την επέμβαση ανθρώπινου παράγοντα που θα παράγει τις περιλήψεις χειροκίνητα. Καθώς το πρόβλημα της παραγωγής περίληψης αποτελεί ένα πολύ δύσκολο πρόβλημα βαθιάς μάθησης, το ελάχιστο μέγεθος δικτύου που απαιτείται για την αντιμετώπισή του, είναι αρκετά μεγάλο, που σημαίνει ότι έχει αυξημένες απαιτήσεις σε όγκο δεδομένων για την αποτελεσματική εκπαίδευσή του. Αυτό το φαινόμενο δυσκολεύει πολύ την εισαγωγή ανθρώπινου παράγοντα για την δημιουργία των περιλήψεων στόχων του συνόλου.

  Το σύνολο δεδομένων Gigaword που παρουσιάστηκε πρώτη φορά, κατόρθωσε να περιέχει πάνω από 4 εκατομμύρια αντιστοιχίες κειμένων-περιλήψεων. Η φιλοσοφία του συνόλου αφορά αντιστοιχίες μικρών κειμένων (κατά μέσω όρο μήκους 31.3 λέξεων) και μικρών περιλήψεων τους (κατά μέσω όρο μήκους 8.3 λέξεων). Οι αντιστοιχίες αυτές παράχθηκαν με αυτόματο τρόπο από το σύνολο άρθρων του Gigaword συνδυάζοντας την πρώτη πρόταση κάθε άρθρου με τον τίτλο του. Η θεωρία πίσω από αυτό είναι ότι στην αρχή κάθε άρθρου γίνεται μια γενική περιγραφή του τι θα ακολουθήσει και επομένως η πρώτη πρόταση πιάνει το νόημα του κάθε άρθρου καλύτερα από οποιαδήποτε άλλη. Αυτή η αυτόματη συνθήκη μπορεί να μην είναι πλήρως αυστηρή και ορθή όσον αφορά τις αντιστοιχίες αλλά κατόρθωσε να δημιουργήσει ένα πολύ μεγάλο σύνολο δεδομένων και να ανοίξει τους ορίζοντες της έρευνας στο αντικείμενο.

  Για την προεπεξεργασία του συνόλου, ο δημιουργός του ξεκίνησε 9.5 εκατομμύρια αντιστοιχίες κειμένων περιλήψεων που δημιουργήθηκαν σύμφωνα με την διαδικασία που περιγράφηκε παραπάνω. Μέσω των παρακάτω πολύ απλών κανόνων, κατόρθωσαν να ρίξουν τον τελικό όγκο σε 4 εκατομμύρια αντιστοιχίες.

  1. Αφαίρεση αντιστοιχιών που περιέχουν εξαιρούμενες λέξεις (βρισιές)
  2. Αφαίρεση αντιστοιχιών, που στον τίτλο περιέχουν επιπλέον πληροφορίες εκτός νοηματικού περιεχομένου όπως για παράδειγμα ο τίτλος του συγγραφέα
  3. Αφαίρεση αντιστοιχιών, που στον τίτλο περιέχουν ερωτηματικό ή τελεία

  Η τελική επεξεργασία των δεδομένων πραγματοποιήθηκε στην συνέχεια στις εναπομείνασες αντιστοιχίες και αφορούσε τα εξής

  1. μετατροπή όλων των γραμμάτων σε πεζά για λόγους κανονικοποίησης
  2. σπάσιμο συντμήσεων, ώστε να εμφανίζονται ως διαφορετικές λέξεις (π.χ. is n't)
  3. αντικατάσταση όλων των αριθμητικών ψηφίων με τον χαρακτήρα δίεσης (#)
  4. αντικατάσταση όλων των λέξεων που εμφανίζονται κάτω από 5 φορές με την λέξη 'unk'
  5. αφαίρεση τυχόν άρθρων που περιλαμβάνονται στα σύνολα δεδομένων DUC, ώστε να μην υπάρχουν επικαλύψεις.

- **Σύνολο DUC 2004:** Το συγκεκριμένο σύνολο δεδομένων, προέρχεται από μια σειρά συνεδρίων (Document Understadting Conferences) με στόχο την συνεχή βελτίωση του τομέα της αυτόματης παραγωγής περίληψης κειμένου. Τα συγκεκριμένα συνέδρια, οργανώνονταν από το διεθνές ινστιτούτο τεχνολογίας NIST μέχρι το 2007 και στην συνέχεια αφομοιώθηκαν με το συνέδριο ανάλυσης κειμένου (TAC). Πλέον, είναι διαθέσιμα διάφορα σύνολα δεδομένων που χρησιμοποιήθηκαν σε αυτά τα συνέδρια ανά τα χρόνια. Τα επίσημα διαθέσιμα σύνολα δεδομένων κάθε χρονιάς περιέχουν τα εξής δεδομένα:

  1. τα άρθρα προς περίληψη
  2. περιλήψεις των άρθρων που έχουν παραχθεί από ανθρώπους
  3. περιλήψεις που έχουν παραχθεί από αυτόματα από υπολογιστή
  4. περιλήψεις που έχουν παραχθεί από διάφορους συμμετέχοντες των συνεδρίων
  5. διάφορα αποτελέσματα αξιολόγησης των περιλήψεων

  Το σύνολο δεδομένων που θα χρησιμοποιηθεί σε αυτή την εργασία, αποτελεί μια προεπεξεργασμένη εκδοχή των αποτελεσμάτων που έγιναν διαθέσιμα κατά το συνέδριο του 2004. Τα δεδομένα που έγιναν διαθέσιμα εκείνη την χρονιά προήλθαν από τις συλλογές TDT και TREC σύμφωνα με την επίσημη αναφορά. Το πλήθος των άρθρων είναι μόλις 500 και συνεπώς αυτό το σύνολο δεδομένων θα χρησιμοποιηθεί μόνο για αξιολόγηση του μηχανισμού και όχι για την εκπαίδευσή του. Από τα αρχικά άρθρα έχει διατηρηθεί για το σύνολο δεδομένων μας μόνο η πρώτη τους πρόταση, ενώ για κάθε ένα από αυτά τα άρθρα υπάρχουν τέσσερις αντιστοιχισμένες περιλήψεις οι οποίες έχουν παραχθεί από ανθρώπους.

  Το σύνολο δεδομένων DUC 2004 αποτελεί ένα σύνολο αρκετά διαδεδομένα και χρησιμοποιείται πολύ στις επιστημονικές έρευνες που αφορούν την αυτόματη παραγωγή περιλήψεων κειμένου, όπως και το σύνολο Gigaword. Όπως αναλύθηκε και πιο πριν λοιπόν τα σύνολα επιλέχθηκαν καθώς είναι αρκετά διαδεδομένα, έτσι ώστε να είναι εφικτή η σύγκριση του μηχανισμού μας με άλλους αντίστοιχους. Καθώς η μηχανική μάθηση είναι όπως αναφέραμε και πιο πριν μια επιστήμη μελέτης των δεδομένων, έρευνες με διαφορετικά δεδομένα μπορεί να αποφέρουν πολύ διαφορετικά αποτελέσματα. Χρησιμοποιώντας έτσι τα ίδια δεδομένα με εργασίες τις οποίες θα συγκριθεί η συγκεκριμένη, υπάρχει καλύτερη εποπτεία για την απόδοση του μηχανισμού αυτής της εργασίας.

Όπως αναλύθηκε λοιπόν, τα σύνολα δεδομένων που θα χρησιμοποιηθούν έχουν ήδη υποστεί αρκετή επεξεργασία έτσι ώστε να έρθουν σε ορθή μορφή αντιστοιχίσεων από άρθρα και περιλήψεις. Σε αυτή την εργασία ωστόσο δόθηκε αρκετή επιπλέον προσοχή στην προεπεξεργασία των δεδομένων, ειδικότερα στο σύνολο δεδομένων Gigaword, σε μια προσπάθεια ο μηχανισμός παραγωγής περιλήψεων να βγάλει τα καλύτερα δυνατά αποτελέσματα. Όσο καλός και να είναι ένας υλοποιημένος μηχανισμός, δεν γίνεται να αποδώσει μέγιστα, αν τα σύνολα δεδομένων δεν είναι ορθά καθαρισμένα και συνεπώς μαθαίνει λανθασμένες συσχετίσεις μεταξύ τους. Σε έναν τόσο μεγάλο όγκο δεδομένων όσο αυτόν του συνόλου Gigaword, οι λιγοστοί κανόνες επεξεργασίας που προηγήθηκαν από τον δημιουργό του συνόλου σίγουρα δεν αρκούν για να μετασχηματίσουν όλα αυτά τα δεδομένα στην βέλτιστη μορφή τους. Ακόμη, είναι δυνατόν να υπάρχουν κανόνες προεπεξεργασίας των δεδομένων που βρίσκουν νόημα μόνο εν γνώση της αρχιτεκτονικής του εκάστοτε μοντέλου που χρησιμοποιείται. Στην συνέχεια της εργασίας θα παρουσιαστούν οι διάφοροι κανόνες προεπεξεργασίας που χρησιμοποιήθηκαν για να έρθουν τα δεδομένα στην τελική βελτιστοποιημένη μορφή, ενώ θα παρουσιαστεί και η φιλοσοφία πίσω από την χρήση διάφορων από αυτούς τους κανόνες, αναλύοντας γιατί τα υλοποιημένα μοντέλα στην συνέχεια λειτουργούν καλύτερα με την χρήση αυτών.

3.3 Προεπεξεργασία δεδομένων

Όπως αναλύθηκε και στην προηγούμενη υποενότητα, τα σύνολα δεδομένων που χρησιμοποιήθηκαν στα πλαίσια αυτής της εργασίας πέρασαν από διάφορα στάδια προεπεξεργασίας σε μια προσπάθεια η μορφή τους να είναι όσο πιο ιδανική γίνεται για τα διάφορα μοντέλα που θα τα αναλύσουν.

Για να εξηγηθούν καλύτερα διάφορες από τις αποφάσεις που πάρθηκαν στο στάδιο της προεπεξεργασίας, πρέπει να τονιστεί η ανάγκη για σταθερά μήκη των διάφορων ακολουθιών εισόδου καθώς και των διάφορων ακολουθιών εξόδου κατά την ανάλυση του μοντέλου μας. Στο πρόβλημα της αυτόματης παραγωγής περίληψης ένα από τα ζητήματα που καλείται να αντιμετωπιστεί είναι ο χειρισμός των πολύ διαφορετικών μηκών ακολουθίας που προκύπτουν μεταξύ διαφορετικών κειμένων και περιλήψεων τους. Αυτή η διαφορά στα μήκη των ακολουθιών είναι πρόβλημα καθώς διάφορες μαθηματικές διαδικασίες που πραγματοποιούνται κατά την εκπαίδευση και την πρόβλεψη περιλαμβάνουν πολλαπλασιασμούς πινάκων, που αποτελούνται από πολλαπλές ακολουθίες δεδομένων και πρέπει να ορίζονται σε ένα συγκεκριμένο μήκος, ώστε αυτές οι πράξεις να είναι εφικτές. Για να αντιμετωπιστεί αυτό το φαινόμενο, κατά την μοντελοποίηση των ακολουθιών εισόδου και εξόδου για την εισαγωγή τους στο εκάστοτε μοντέλο, αυτές περιορίζονται σε ένα σταθερό μήκος με μία διαδικασία που θα αναλυθεί καλύτερα στο επόμενο κεφάλαιο. Αυτό που πρέπει να συγκρατηθεί είναι πως κατά την προεπεξεργασία έγινε προσπάθεια να αφαιρεθούν ”ανούσια” στοιχεία ακολουθίας από τις εισόδους και τις εξόδους, ώστε στον πολύτιμο περιορισμένο χώρο που έχει κάθε ακολουθία, τα στοιχεία που έχουν απομείνει να έχουν μεγαλύτερη αξία και ώστε κατά την πρόβλεψη περιλήψεων να αυξηθεί η πιθανότητα να παραχθούν στοιχεία με ουσιαστική αξία, ώστε να μην σπαταλιέται χώρος.

Άλλη μια γενική φιλοσοφία πάνω στην οποία βασίστηκαν διάφορες αποφάσεις που πάρθηκαν κατά την προεπεξεργασία των συνόλων δεδομένων αφορά την προσπάθεια για περιορισμό του λεξιλογίου όσο αυτό είναι δυνατό. Όπως έχει ήδη αναφερθεί κατά την διαδικασία εκπαίδευσης και πρόβλεψης, ο κωδικοποιητής και ο αποκωδικοποιητής γνωρίζουν συγκεκριμένο πλήθος λέξεων, τις οποίες μπορούν να χειριστούν. Συγκεκριμένα για τον αποκωδικοποιητή, είχε αναφερθεί πως στην τελική έξοδο του δικτύου του υπάρχει ένα νευρωνικό δίκτυο με πλήθος ακμών εξόδου, όσες και το μήκος του λεξιλογίου πρόβλεψης του αποκωδικοποιητή. Όσο μεγαλώνει το πλήθος του λεξιλογίου από ανούσια στοιχεία πρότασης, τόσο αυξάνεται το χρονικό κόστος της εκπαίδευσης και της πρόβλεψης, ενώ ταυτόχρονα αυξάνεται και η πιθανότητα λάθος πρόβλεψης από τον αποκωδικοποιητή, αφού το πλήθος των δυνατών επιλογών είναι αυξημένο.

Ένα ακόμη στοιχείο που παρατηρήθηκε στο σύνολο Gigaword και για το οποίο λήφθηκαν μέτρα, ήταν η ύπαρξη πολλαπλών ίδιων εγγραφών. Συγκεκριμένα παρατηρήθηκε ότι πολλαπλές αντιστοιχίες κείμενων - περιλήψεων εμφανίζονταν πάνω από μια φορά. Μερικές από αυτές μάλιστα εμφανίζονταν ακριβώς ίδιες σε συχνότητα τάξης 700 αλλά και περισσότερη. Σε μια προσπάθεια, όλες οι εγγραφές να έχουν ισοδύναμη αξία και να αποφευχθεί η υπερεκπαίδευση (overfitting), δόθηκε πολύ σημασία στην εξάλειψη αυτών των πολλαπλών εγγραφών από το σύνολο.

Τέλος, ένα στοιχείο που παρατηρήθηκε στο σύνολο Gigaword, ήταν η ύπαρξη αντιστοιχιών χωρίς ουσιαστικό νοηματικό περιεχόμενο. Πολλές ακολουθίες δηλαδή φάνηκε να περιέχουν αυτοματοποιημένα μηνύματα που δίνουν συγκεκριμένες πληροφορίες για το άρθρο, αλλά δεν σχετίζονται με το νόημα του. Τα αυτοματοποιημένα αυτά μηνύματα, εμφανίζονταν είτε σε σχετικά απαλή μορφή, για παράδειγμα σε διάφορα άρθρα υπήρχαν πληροφορίες για την ημερομηνίας εγγραφής τους, την τοποθεσία και τον συγγραφέα, είστε σε πιο σκληρή μορφή, για παράδειγμα συγκεκριμένα άρθρα δεν είχαν καν ορθή αντιστοιχία με κάποια περίληψη αλλά η περίληψη που τους είχε αντιστοιχηθεί αποτελούνταν εξ ολοκλήρου από κάποιο αυτοματοποιημένο μήνυμα.

Στην συνέχεια παρουσιάζονται με περισσότερη λεπτομέρεια τα στάδια της προεπεξεργασίας που ακολουθήθηκαν για την παραγωγή των τελικών συνόλων δεδομένων:

3.3.1 Στάδια προεπεξεργασίας δεδομένων

Τα στάδια που ακολουθήθηκαν κατά την προεπεξεργασίας των δεδομένων σε αυτή την εργασία παρουσιάζονται παρακάτω.

1. **Αντικατάσταση όλων των αριθμών με την κωδική λέξη NUMBER.** Οι αριθμοί στο σύνολο Gigaword εμφανίζονται ως ακολουθίες διέσεων, ενώ στο σύνολο DUC 2004 έχουν την πραγματική τους μορφή. Με στόχο το διαφορετικό μήκος και τα διαφορετικά ψηφία των αριθμών να μην ορίζουν διαφορετικές λέξεις στο λεξιλόγιο του μοντέλου, αντικαταστάθηκαν με την λέξη NUMBER που δεν εμφανίζεται απο πριν στα σύνολα.

2. **Αφαίρεση παρενθέσεων και των περιεχομένων τους.** Στα σύνολα δεδομένων εμφανίζονται παρενθέσεις που ακολουθούν την μορφή ”-lrb- πληροφορία -rrb-”. Η πληροφορία που εμφανίζεται μέσα στις παρενθέσεις καθώς και τα σύμβολα -lrb-, -rrb- δεν προσφέρουν επιπλέον πληροφορία στο νοηματικό περιεχόμενο των ακολουθιών και αφαιρέθηκαν για την εξοικονόμηση χώρου και όγκου λεξιλογίου.

3. **Αντικατάσταση συντμήσεων με την ολική τους μορφή.** Στα σύνολα δεδομένων εμφανίζονται διάφορες συντμήσεις αντί της ολικής τους μορφής. Σε μια προσπάθεια επίτευξης καθολικότητας, αυτές αντικαταστάθηκαν με την ολική τους μορφή. Αυτή η διαδικασία μειώνει τον όγκο του λεξιλογίου ενώ παράλληλα επιτρέπει στο μοντέλο να διδαχθεί καλύτερα τις εξαρτήσεις τους με άλλες λέξεις. Παραδείγματα συντμήσεων που αντικαταστάθηκαν είναι: n’t -> not, ’m -> am, ’ve -> have, ενώ οι συντμήσεις ’d, ’ll, ’s λόγω της αμφιλεγόμενης υπόστασης τους, αντικαταστάθηκαν με χρήση του εκπαιδευμένου μοντέλου pycontractions.

4. **Σπάσιμο των ενωμένων λέξεων.** Πολλές λέξεις στα σύνολα δεδομένων εμφανίζονται ενωμένες με μια παύλα στην μέση. Με στόχο το μοντέλο να μην καταλαβαίνει την ένωση αυτών ως διαφορετική λέξη, οι παύλες αφαιρέθηκαν και οι λέξεις σπάσανε σε δύο διαφορετικές.

5. **Αντικατάσταση όλων των εισαγωγικών χαρακτήρων με έναν καθολικό.** Στα σύνολα δεδομένων εμφανίζονται διάφοροι χαρακτήρες εισαγωγικού(’, ”, ‘) . Για λόγους συνοχής αντικαταστάθηκαν όλοι με τον βασικό χαρακτήρα εισαγωγικού (’).

6. **Αφαίρεση πολλαπλών εγγραφών.** Όπως αναφέρθηκε και πιο πριν, αφαιρέθηκαν από τα σύνολα δεδομένων όλες οι αντιστοιχίες που εμφανίζονταν πάνω από μια φορά και παρέμεινε ουσιαστικά μια μοναδική από αυτές. Αφαιρέθηκαν μόνο οι εγγραφές των οποίων και το κείμενο και η περίληψη αναφοράς ήταν ακριβώς ίδιες με άλλες.

7. **Αφαίρεση αυτοματοποιημένων μηνυμάτων.** Όπως πάλι αναφέρθηκε και πιο πάνω, πολλές αντιστοιχίες περιείχαν αυτοματοποιημένα μηνύματα νοηματικά κενά σε σχέση με το κείμενο. Μέσω διάφορων κανονικών εκφράσεων έγινε πολύ προσπάθεια αφαίρεσης αυτών και αν και δεν γίνεται να είναι σίγουρη η εξάλειψη όλων αυτών λόγω του τεράστιου όγκου των δεδομένων, τα τελικά αποτελέσματα απέκτησαν πολύ καλύτερη μορφή από την αρχική τους.

8. **Αφαίρεση αντιστοιχιών στις οποίες το μήκος της περίληψης ξεπερνά το μήκος του κειμένου.** Σαν μια προσπάθεια γενικότερης εξασφάλισης ότι τα δεδομένα έχουν σωστή μορφή αφαιρέθηκαν οι αντιστοιχίες στις οποίες η περίληψη αναφοράς, είναι μεγαλύτερη από το κείμενο, καθώς απ’ ότι παρατηρήθηκε αυτό ίσχυε κατά κύριο λόγω σε λανθασμένες αντιστοιχίες.

Ύστερα από αυτή την προεπεξεργασία, προέκυψαν εν τέλει περίπου 3 εκατομμύρια μοναδικές αντιστοιχίες άρθρων περιλήψεων στο σύνολο Gigaword, πάνω στις οποίες πραγματοποιήθηκε η εκπαίδευση και ο έλεγχος της απόδοσης. Συγκεκριμένα χρησιμοποιήθηκαν για την εκπαίδευση 2.9 εκατομμύρια και για τον έλεγχο 150 χιλιάδες αντιστοιχίες. Ο όγκος του συνόλου DUC παρέμεινε ίδιος και χρησιμοποιήθηκε όπως αναφέρθηκε και πριν μόνο για έλεγχο της απόδοσης.

Κεφάλαιο 4

Υλοποίηση λογισμικού αυτόματης περίληψης κειμένου

4.1 Γενικά

Σε αυτό το κεφάλαιο θα παρουσιαστεί λεπτομερώς η υλοποίηση του συνολικού μηχανισμού που σχεδιάστηκε στα πλαίσια αυτής της εργασίας για την αντιμετώπιση του προβλήματος της αυτόματης παραγωγής περίληψης. Πιο συγκεκριμένα, στην αρχή θα αναλυθούν όλα τα βήματα της διαδικασίας και θα εξηγηθεί αναλυτικά η σημασία τους. Η συνολική ανάλυση που θα πραγματοποιηθεί δεν θα περιλαμβάνει μόνο τον μηχανισμό βαθιάς μάθησης που χρησιμοποιήθηκε, αλλά θα παρουσιάζει το ολικό σύστημα, ξεκινώντας από τον αρχική επεξεργασία δεδομένων με σκοπό την ορθή εισαγωγή τους στο μοντέλο βαθιάς μάθησης, συνεχίζοντας με την περιγραφή του μοντέλου και των διάφορων επιλογών που πάρθηκαν σε αυτό και τελειώνοντας με την εξαγωγή των αποτελεσμάτων και την τελική τους ανάλυση και επεξεργασία τους με στόχο την βελτιστοποίηση των αποτελεσμάτων.

Στα πλαίσια αυτής της εργασίας, έγινε προσπάθεια να μελετηθεί η γενικότερη συμπεριφορά και η αξία του μοντέλου κωδικοποιητή - αποκωδικοποιητή στην αντιμετώπιση του προβλήματος της αυτόματης παραγωγής περίληψης. Γι' αυτό το σκοπό, υλοποιήθηκαν και θα αναλυθούν στην συνέχεια διαφορετικές εκδοχές μοντέλων βαθιάς μάθησης. Σε κάθε ένα από αυτά τα μοντέλα έχει γίνει προσπάθεια να μελετηθεί η συμπεριφορά του ως προς κάποια σημαντική παράμετρο ή ως προς κάποια γενικότερη εξάρτηση του. Σε αυτή την ενότητα θα αναλυθεί επίσης ο λόγος για την επιλογή της εκάστοτε παραμέτρου προς διερεύνηση. Τα συνολικά αποτελέσματα που θα προκύψουν θα αναλυθούν εκτενώς στο επόμενο κεφάλαιο.

Η οργάνωση του συγκεκριμένου κεφαλαίο όπως περιγράφηκε και παραπάνω θα είναι ακολουθιακή που σημαίνει ότι θα σέβεται την ροή της πληροφορίας από την αρχή προς την έξοδο του συστήματος. Ο λόγος που συμβαίνει αυτό, αποτελεί η καλύτερη κατανόηση της λειτουργίας του μηχανισμού από τον αναγνώστη. Όπου απαιτείται ελάχιστο θεωρητικό υπόβαθρο για την καλύτερη κατανόηση οποιουδήποτε στοιχείου του μηχανισμού, αυτό θα παρουσιάζεται. Όπως αναφέρθηκε και πριν, θα δοθεί πολύ προσοχή στην πρακτική αιτιολόγηση της ύπαρξης κάθε στοιχείου, στο γιατί δηλαδή επιλέχθηκε να χρησιμοποιηθεί.

4.2 Στάδια λειτουργίας λογισμικού

4.2.1 Δημιουργία Λεξιλογίου

Ο υπολογιστής, ως ένα σύστημα πραγματοποίησης πεπερασμένων διαδικασιών και λειτουργιών έχει περιορισμούς όσον αφορά την χωρητικότητα και τις δυνατότητές του. Στο σύστημα αυτόματης παραγωγής περίληψης, καλείται αυτός να παράξει σε κάθε έξοδο της αναδρομής μια λέξη περίληψης του εκάστοτε κειμένου εισόδου. Από που προέρχονται όμως αυτές οι λέξεις; Για τον υπολογιστή είναι αδύνατο να παράξει μόνος του λέξεις μιας συγκεκριμένης γλώσσας, ή να χρησιμοποιήσει ολόκληρα λεξικά ως βάση παραγωγής λέξεων καθώς τότε το χρονικό κόστος θα ήταν τεράστιο. Θα πρέπει κατά την διάρκεια της διαδικασίας να έχει οριστεί αυστηρά ένα λεξιλόγιο από το οποίο θα επιλέγονται οι λέξεις που θα παράγονται. Ακόμη, καθώς οι λέξεις πρέπει να κωδικοποιηθούν σε αριθμητική μορφή πριν την είσοδο τους στο μοντέλο, το λεξιλόγιο μπορεί να λειτουργήσει ως ένα μέσο αναφοράς και μετατροπής των γνωστών λέξεων στην τελική μορφή τους.

Στο μοντέλο κωδικοποιητή - αποκωδικοποιητή λοιπόν, υπάρχει η αναγκαιότητα ύπαρξης κάποιου λεξιλογίου, συγκεκριμένου μήκους, το οποίο ο κωδικοποιητής θα χρησιμοποιεί για την τελική αναπαράσταση των λέξεων σε αριθμητική μορφή πριν την είσοδο σε αυτόν, και ο αποκωδικοποιητής θα το χρησιμοποιεί πάλι για μετατροπή λέξεων σε αριθμητική μορφή, καθώς και για την τελική πρόβλεψη λέξης στην έξοδό του κατά την διαδικασία εκπαίδευσης ή ελέγχου. Το λεξιλόγιο αυτό δεν χρειάζεται να είναι το ίδιο για τον κωδικοποιητή και για τον αποκωδικοποιητή αλλά πρέπει να είναι αυστηρώς ορισμένο και όσο πιο περιορισμένο γίνεται για την μείωση του χρονικού κόστους των υπολογισμών, καθώς μειωμένος όγκος λεξιλογίου συνεπάγεται με μειωμένους χρόνους αναζήτησης σε αυτό.

Συνήθης τακτική που χρησιμοποιείται πολύ σε εργασίες επεξεργασίας φυσικών γλωσσών και που χρησιμοποιήθηκε σε αυτήν, είναι το λεξιλόγιο να παράγεται από το σύνολο λέξεων που υπάρχουν στα σύνολα δεδομένων. Πρέπει να δοθεί προσοχή, ώστε το λεξιλόγιο που θα παραχθεί να αποτελείται μόνο από λέξεις του συνόλου εκπαίδευσης και όχι από λέξεις του συνόλου ελέγχου, καθώς το σύνολο ελέγχου αναπαριστά δεδομένα που είναι τελείως άγνωστα από το μοντέλο κατά την διάρκεια της εκπαίδευσης, ως μια προσπάθεια για έλεγχο αυτού σε πραγματικές συνθήκες. Έτσι το λεξιλόγιο μπορεί να αντιπροσωπεύει καλύτερα τον χαρακτήρα και το λεξιλόγιο των διάφορων δεδομένων που χρησιμοποιούμε για εκπαίδευση και μπορεί να αποδώσει πολύ καλά σε κείμενα με παρόμοιο λεξιλόγιο κατά τον έλεγχο του μοντέλου. Φυσικά σε περίπτωση ελέγχου κειμένου τελείως διαφορετικού λεξιλογίου από αυτό του συνόλου εκπαίδευσης θα υπάρχει πρόβλημα με την ορθή αναπαράσταση της πληροφορίας μέσα στο δίκτυο, αλλά καθώς το πρόβλημα της γενίκευσης του προβλήματος της παραγωγής περίληψης σε γενικά δεδομένα παραμένει άλυτο και μακρινό, η εργασία αυτή δεν θα απασχοληθεί με πιθανή αντιμετώπισή του.

Για τα πλαίσια αυτής της εργασίας λοιπόν, καθώς το μόνο σύνολο εκπαίδευσης που χρησιμοποιήθηκε είναι ένα μέρος του συνόλου Gigaword, αυτό χρησιμοποιήθηκε για την παραγωγή του λεξιλογίου. Οι προτάσεις που εμφανίζονται σε αυτό χωρίστηκαν σε λέξεις (tokenization), στις οποίες πραγματοποιήθηκε μέτρησης συχνότητας εμφάνισης. Η πολιτική που ακολουθήθηκε είναι, ανάλογα με το μήκος του λεξιλογίου, να επιλέγονται οι πιο συχνά εμφανιζόμενες λέξεις στα κείμενα και να εισάγονται σε αυτό. Το λεξιλόγιο που επιλέχθηκε μετά από μελέτη είναι κοινό για τον κωδικοποιητή και τον αποκωδικοποιητή και περιέχει 50000 λέξεις. Στα αποτελέσματα που ακολουθούν θα συγκριθούν αποτελέσματα αυτού του μήκους λεξιλογίου με αντίστοιχα αποτελέσματα του ολικού λεξιλογίου (στο οποίο δεν έχει αφαιρεθεί καμία λέξη) και θα αιτιολογηθεί καλύτερα η επιλογή του.

4.2.2 Αριθμητική αναπαράσταση δεδομένων εισόδου

Όπως έχει αναφερθεί ήδη σε προηγούμενο κεφάλαιο, τα σύνολα δεδομένων που χρησιμοποιήθηκαν έχουν υποστεί πολύ προεπεξεργασία, με στόχο την μετατροπή τους στη καταλληλότερη δυνατή μορφή για τον μηχανισμό αυτόματης παραγωγής περίληψης. Η προεπεξεργασία αυτή των δεδομένων δεν αρκεί όμως για την ορθή εισαγωγή των δεδομένων στο μοντέλο. Ο ρόλος των μοντέλων μηχανικής και βαθιάς μάθησης είναι να προσεγγίσουν μια θεωρητική μαθηματική συνάρτηση που θα συνδέει την εκάστοτε είσοδο του μοντέλου στην επιθυμητή έξοδο. Για να γίνει εφικτό αυτό, τα βάρη του μοντέλου προσαρμόζονται σύμφωνα με κάποιον αλγόριθμο βελτιστοποίησης. Πρέπει συνεπώς όλα τα δεδομένα που εισέρχονται στο μοντέλο να είναι σε αριθμητική μορφή ώστε η εκπαίδευση και οι διάφορες προβλέψεις να μπορούν να λειτουργήσουν. Προκύπτει λοιπόν η αναγκαιότητα για μετατροπή των λέξεων του κειμένου προς περίληψη σε αναπαράσταση με αριθμητικές τιμές.

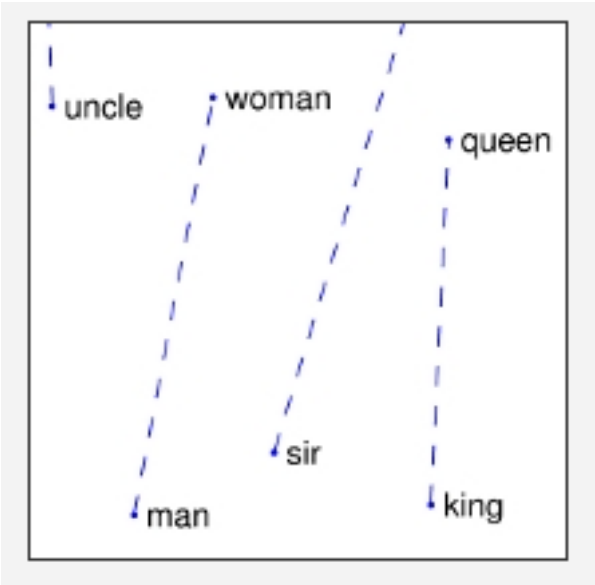
Το πρόβλημα που προκύπτει από τα παραπάνω, είναι πως η μετατροπή λέξεων σε αριθμητικές τιμές δεν είναι απλή υπόθεση. Καθώς οι λέξεις περιέχουν σημασιολογικές διαφορές και πολλαπλά νοήματα, θα πρέπει το μαθηματικό μοντέλο που τις περιγράφει να διατηρεί αυτά, αλλιώς οι παραγόμενες περιλήψεις θα προκύπτουν νοηματικά κενές. Στις μέρες μας, όσο ο όγκος της πληροφορίας αυξάνεται, ο κλάδος της επεξεργασίας φυσικών γλωσσών γίνεται όλο και πιο σπουδαίος καθώς δημιουργείται ανάγκη για περισσότερη και πιο ορθή επεξεργασία φυσικών γλωσσών για διάφορους σκοπούς. Στα πλαίσια αυτά, έχουν πραγματοποιηθεί πολλές προσπάθειες μοντελοποίησης των κατηγορηματικών δεδομένων εισόδου σε αριθμητικές τιμές, ώστε να διατηρούνται τα νοήματα και οι σχέσεις μεταξύ διαφορετικών λέξεων.

Στην συγκεκριμένη εργασία, για την αντιμετώπιση του προβλήματος μετατροπής των λέξεων επεξεργασίας σε αριθμητικές παραστάσεις, χρησιμοποιήθηκαν οι ήδη εκπαιδευμένες με μηχανική μάθηση αναπαραστάσεις Glove, οι οποίες αποτελούν από διανύσματα μήκους 300, που έχουν δημιουργηθεί από το Stanford και χρησιμοποιούνται ευρέως στις μέρες μας. Πιο συγκεκριμένα, για όλες τις λέξεις του λεξιλογίου αναφοράς, αποθηκεύτηκε για μελλοντική χρήση το ήδη εκπαιδευμένο διάνυσμα της. Ο αλγόριθμος εξαγωγής αριθμητικών αναπαραστάσεων Glove, χρησιμοποιεί την συχνότητα εμφάνισης μιας λέξης στην ”περιοχή” μιας άλλης, ως μετρική, για την δημιουργία των τελικών αναπαραστάσεων. Στην ουσία, δημιουργείται ένας πίνακας με γραμμές και στήλες ίσες με το πλήθος των λέξεων προς εκπαίδευση. Η τιμή σε κάθε κελί του πίνακα υποδηλώνει την συχνότητα εμφάνισης της λέξης που αντιστοιχεί στην γραμμή, στην περιοχή της λέξης στόχου που αντιστοιχεί στην στήλη, όπου το μέγεθος της περιοχής δηλώνει την ποσότητα των λέξεων που που έπονται ή ακολουθούν την λέξη στόχο. Με βάση τα παραπάνω μεγέθη του πίνακα ορίζεται η συνάρτηση κόστους που φαίνεται στην εξίσωση 4.1.

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T * w_j + b_i + b_j - \log(X_{ij}))^2 \tag{4.1}$$

Σε αυτή την συνάρτηση τα w_i, w_j δηλώνουν τα διανύσματα αριθμητικών παραστάσεων των λέξεων i, j του πίνακα και η τιμή X_{ij} υποδηλώνει την ποσότητα εμφάνισης που ορίστηκε παραπάνω. Τέλος τα b_i, b_j αποτελούν παράγοντες πόλωσης (bias). Η ελαχιστοποίηση αυτής της συνάρτησης κόστους μέσω της εκπαίδευσης, ελαχιστοποιεί το αποτέλεσμα της δεύτερης παρένθεσης καθώς προσαρμόζονται τα βάρη w και οι παράγοντες πόλωσης, έχοντας ως αποτέλεσμα το γινόμενο των διανυσμάτων να συγκλίνει στην λογαριθμική έκφραση της συχνότητας εμφάνισης X_{ij}. Τα τελικά w που προκύπτουν από την εκπαίδευση αποτελούν τις τελικές παραστάσεις των λέξεων.

Μέσω αυτής της διαδικασίας, παράγονται εν τέλει διανυσματικές αναπαραστάσεις λέξεων, οι οποίες συνδέουν τις νοηματικές ερμηνείες αυτών μέσω διανυσματικών πράξεων. Για παράδειγμα, μέσω της άθροισης διανυσμάτων διαφόρων λέξεων προκύπτει ένα τελικό διάνυσμα η ερμηνεία του οποίου είναι άρρηκτα συνδεδεμένη με τις λέξεις που αθροίστηκαν. Αυτή η ιδιότητα είναι πολύ σημαντική καθώς προκύπτει μια κανονικοποιημένη μετατροπή των λέξεων σε παραστάσεις επιτρέποντας στο μοντέλο παραγωγής περίληψης να καταλαβαίνει καλύτερα νοηματικές εξαρτήσεις μεταξύ λέξεων.



Σχήμα 4.1: Παραδείγματα νοηματικά κοντινών λέξεων. Σε αυτή την φωτογραφία απεικονίζεται ότι διάφορες λέξεις αντίθετες μεταξύ τους ως προς τον παράγοντα φύλο απέχουν διανυσματικά σχεδόν ίσες αποστάσεις μεταξύ τους.

Αξίζει να σημειωθεί, ότι αντί να χρησιμοποιηθούν έτοιμες, ήδη εκπαιδευμένες αναπαραστάσεις λέξεων, θα μπορούσε να πραγματοποιηθεί εκπαίδευση στον αλγόριθμο του Glove με το σύνολο των προτάσεων που εμφανίζονται στα σύνολα δεδομένων που χρησιμοποιήθηκαν. Η φιλοσοφία πίσω από αυτό είναι πως έτσι οι τελικές αναπαραστάσεις θα είναι καλύτερα συνδεδεμένες με τα συγκεκριμένα δεδομένα και με το συγκεκριμένο πρόβλημα οδηγώντας σε βελτίωση των τελικών αποτελεσμάτων. Παρά όμως αυτή την θεώρηση, αποδείχθηκε στην πραγματικότητα πως η εκπαίδευση στα σύνολα δεδομένων της εργασίας δεν είναι επαρκής και δεν οδήγησε σε καλά αποτελέσματα καθώς το πλήθος διαφορετικών λέξεων και προτάσεων δεν ήταν επαρκώς μεγάλα ώστε οι αναπαραστάσεις να είναι σχετικά ”αντικειμενικές”. Συνεπώς προτιμήθηκαν οι έτοιμες αναπαραστάσεις του Glove που είναι εκπαιδευμένες σε μεγαλύτερο σύνολο δεδομένων.

4.2.3 Τελική προετοιμασία εισόδου

Η αντιμετώπιση του προβλήματος της αριθμητικής αναπαράστασης των λέξεων ώστε να διατηρείται το νοηματικό περιεχόμενο όσο καλύτερα γίνεται, δεν αρκεί για την ορθή εισαγωγή των δεδομένων στο μοντέλο εκπαίδευσης καθώς και στο μοντέλο πρόβλεψης. Πρέπει ακόμα οι διάφορες είσοδοι στο μοντέλο να έρθουν στα σωστά μεγέθη και μορφή όπως αυτό τα αναμένει για επεξεργασία. Το μοντέλο που επεξεργάζεται αυτά τα δεδομένα πραγματοποιεί πολλαπλές πράξεις πινάκων και συνεπώς τα δεδομένα πρέπει να έχουν προβλέψιμες διαστάσεις και συγκεκριμένους περιορισμούς ώστε να υπάρχει κανονικότητα στο σύνολο αυτών. Παρακάτω παρουσιάζονται τα βήματα που αφορούν την τελική επεξεργασία των δεδομένων με στόχο την ομαλή και ομοιόμορφη εισαγωγή τους στα μοντέλα. Για κάθε ένα από αυτά τα βήματα θα δοθεί μια συνοπτική εξήγηση της αναγκαιότητας της λειτουργίας του.

1. **Αντικατάσταση άγνωστων λέξεων με ειδική λέξη 'unk':** Όπως αναλύθηκε παραπάνω κατά την διαδικασία της εκπαίδευσης και του ελέγχου του μοντέλου υπάρχουν συγκεκριμένο μέγεθος λεξιλογίου που χρησιμοποιείται για την μετατροπή των λέξεων σε μαθηματικές αναπαραστάσεις κατανοητές από το μοντέλο. Σε περίπτωση που σε κάποια είσοδο βρεθεί μια λέξη για την οποία δεν υπάρχει αντιστοιχία με διάνυσμα, αυτή δεν γίνεται να εισέλθει στο σύστημα, καθώς δεν έχει ορθή αναπαράσταση. Σε αυτό το βήμα της διαδικασίας, άγνωστες τέτοιες λέξεις αντικαθιστώνται με την ειδική λέξη 'unk' στην οποία έχει δοθεί χειροκίνητα μια μαθηματική έκφραση. Η μέθοδος αυτή καθιστά το μοντέλο πιο εύρωστο, καθώς μπορεί να λειτουργήσει και σε περιπτώσεις άγνωστων λέξεων.

2. **Περικοπή παραπανήσιου μήκους ακολουθιών:** Το μοντέλο παραγωγής περιλήψεων πρέπει να δέχεται ακολουθίες συγκεκριμένου μήκους ως είσοδο με στόχο τον ορθό υπολογισμών των εξόδων. Αυτά τα μήκη έχουν οριστεί σε αυτή την εργασία στις 50 λέξεις για το κείμενο και στις 15 λέξεις για την περίληψη. Η επιλογή αυτών των μεγεθών, βασίστηκε στην αρχική αναφορά κατά την δημιουργία του συνόλου δεδομένων Gigaword στο [Rush15], στο οποίο αναφέρονται οι μέσοι όροι λέξεων των κειμένων και των περιλήψεων. Σε αυτούς τους μέσους όρους δόθηκε και παραπάνω παράθυρο για επιπλέον αφαιρετικότητα του μοντέλου.

   Καθώς τα σύνολα δεδομένων όμως μπορεί να περιέχουν παραπάνω λέξεις στις ακολουθίες από την επιτρεπτή τιμή, σε αυτό το βήμα της επεξεργασίας, όλες οι ακολουθίες εισόδου περικόπτονται στο μέγιστο μήκος τους ώστε να υπάρχει μια κανονικότητα μεταξύ των εισόδων. Μέσω αυτής της τεχνικής αποτρέπονται και ακολουθίες τεράστιου μήκους που πιθανώς θα μπορούσαν να παραπλανήσουν το μοντέλο από την πραγματική λειτουργία του.

3. **Εισαγωγή ειδικών λέξεων για αρχή <s> και τέλος </s> ακολουθιών:** Όπως έχει ήδη αναλυθεί, η δουλειά του αποκωδικοποιητή είναι να μαντεύει το επόμενο στοιχείο της ακολουθίας δεδομένου του προηγούμενου σε αυτήν. Γι' αυτό τον λόγο, κατά την διαδικασία της εκπαίδευσης, θα πρέπει για κάθε περίληψη, να δημιουργηθούν δύο ακολουθίες για τον αποκωδικοποιητή, η ακολουθία εισόδου σε αυτόν, και η ακολουθία ελέγχου για σύγκριση της εξόδου αυτού. Και οι δύο αυτές ακολουθίες θα περιέχουν την περίληψη στόχο που αντιστοιχίζεται με το κείμενο εισόδου, η ακολουθία εισόδου όμως θα περιέχει στην αρχή της το σύμβολο αρχής (<s>) και η ακολουθία εξόδου θα περιέχει στο τέλος της το σύμβολο τέλους ακολουθίας (</s>). Η λογική πίσω από αυτό είναι η ακολουθία ελέγχου να βρίσκεται ένα χρονικό βήμα αναδρομής πιο μπροστά από την ακολουθία εισόδου ώστε να μπορεί να εκπαιδευτεί το δίκτυο να προβλέπει το επόμενο στοιχείο της ακολουθίας. Θα γίνει και περισσότερη ανάλυση αυτής της λειτουργία στην περιγραφή του υλοποιημένου αποκωδικοποιητή.

4. **Εισαγωγή χαρακτήρων παραγέμισης <padding> σε μικρότερες ακολουθίες:** Ακολουθώντας την ίδια νοοτροπία ότι τα μήκη των ακολουθιών πρέπει να έχουν σταθερά μήκη, σε αυτό το στάδιο της επεξεργασίας, αν οι ακολουθίες περιέχουν μήκος μικρότερο του αποδεκτού, τότε παραγεμίζονται με τον ειδικό χαρακτήρα <padding> για τον οποίο έχει οριστεί χειροκίνητα η μαθηματική αναπαράστασή του. Αυτοί οι χαρακτήρες εισάγονται στις ακολουθίες μόνο για την διατήρηση των ορθών διαστάσεων και δεν επηρεάζουν το τελικό αποτέλεσμα, καθώς το σύστημα περιέχει γνώση για τις πραγματικές διαστάσεις των ακολουθιών κατά την εκτέλεση των πράξεων με πίνακες από αυτό.

5. **Τελική αντικατάσταση με τιμών με αριθμητικές αναπαραστάσεις τους:** Στο τελευταίο βήμα αυτής της διαδικασίας, οι παραπάνω επεξεργασμένες ακολουθίες αντικαθιστώνται με τις αριθμητικές παραστάσεις Glove που αναλύθηκαν παραπάνω. Οι ειδικοί χαρακτήρες ελέγχου (<s>, </s>, unk, <padding>) αντικαθιστώνται και αυτοί με αριθμητικές παραστάσεις ίδιου μεγέθους για λόγους ομοιομορφίας. Τα δεδομένα πλέον είναι έτοιμα να εισέλθουν στο μοντέλο για επεξεργασία

4.2.4 Επεξεργασία από τον κωδικοποιητή

Όπως έχει ήδη αναφερθεί ο κωδικοποιητής αποτελεί το πρώτο στάδιο της αρχιτεκτονικής ακολουθία-σε-ακολουθία. Αποτελεί ουσιαστικά ένα δίκτυο αρχιτεκτονικής πολλά-σε-ένα και ο ρόλος του είναι να μετατρέπει το κείμενο εισόδου σε μαζεμένες αριθμητικές αναπαραστάσεις ώστε ο κωδικοποιητής να μπορεί στην συνέχεια να το αναλύσει έχοντας την πλήρη γνώση του. Ο παραδοσιακός κωδικοποιητής αποτελείται από ένα νευρωνικό δίκτυο το οποίο σε κάθε χρονικό βήμα τροφοδοτείται από μια είσοδο της ακολουθίας και παράγει μια έξοδο η οποία ανατροφοδοτείται στην είσοδο του. Η τελική έξοδος, που προκύπτει όταν ολόκληρη η ακολουθία έχει περάσει από τον κωδικοποιητή, αποτελεί την τελική αναπαράσταση στόχο που θα χρησιμοποιηθεί στην συνέχεια.

Ο κωδικοποιητής που χρησιμοποιείται όμως στην περίπτωση μας είναι λόγω διαφορετικός λόγω του μηχανισμού προσοχής (attention mechanism) που χρησιμοποιείται στον αποκωδικοποιητή. Ο μηχανισμός προσοχής, όπως αναλύθηκε και πριν αποτελεί μια επαναστατική βελτίωση στην λειτουργία του μοντέλου κωδικοποιητή-αποκωδικοποιητή για την επεξεργασία φυσικών γλωσσών. Ο μηχανισμός αυτός λειτουργεί κατά την διαδικασία πρόβλεψης του αποκωδικοποιητή ”συμβουλεύοντας” τον πίνακα περιεχομένου του (context vector) για το πιο στοιχείο της ακολουθίας εισόδου να συμβουλευτεί περισσότερο. Για να είναι εφικτό αυτό χρησιμοποιούμε έναν τροποποιημένο κωδικοποιητή που αντί να κρατάει μόνο την τελευταία έξοδο του, κρατάει και όλες τις ενδιάμεσες. Ακόμη, για να είναι συμμετρική η πληροφορία σε όλες τις εξόδους καθώς και να περιέχεται πληροφορία για όλο το κείμενο σε αυτές, ο κωδικοποιητής που χρησιμοποιείται είναι και αμφίδρομος. Η τελική έξοδο σε κάθε χρονικό βήμα του κωδικοποιητή είναι ο συνδυασμός της πληροφορίας των αμφίδρομων προσπελάσεων που καταλήγουν στο εκάστοτε στοιχείο ακολουθίας. Έτσι ο μηχανισμός προσοχής καταφέρνει να έχει πολύ καλή γνώση των στοιχείων της ακολουθίας αναφορικά με όλη την ακολουθία εισόδου. Στην συνέχεια θα παρουσιαστούν οι διαφορετικές σχεδιαστικές αποφάσεις που πάρθηκαν κατά την υλοποίηση του κωδικοποιητή.

Οι σχεδιαστικές αποφάσεις που πάρθηκαν κατά τις διάφορες υλοποιήσεις του κωδικοποιητή αφορούσαν την προσπάθεια διερεύνησης της επίπτωσης του μεγέθους του κωδικοποιητή στην απόδοση αυτού, έχοντας ως γνώμονα και το χρονικό κόστος της εκπαίδευσης. Αυτό σημαίνει πως διερευνήθηκαν διαφορετικά μεγέθη κωδικοποιητών και οι συγκρίσεις που θα πραγματοποιηθούν στην συνέχεια θα παρουσιάζουν την επίδοση του καθενός, συγκριτικά με το χρονικό του κόστος.

Τα διαφορετικά μεγέθη κωδικοποιητών που επιλέχθηκαν να αναλυθούν πιο εξονυχιστικά και θα παρουσιαστούν στα αποτελέσματα έχουν μεγέθη κρυφής κατάστασης 128, 256 και 384. Αυτό συνεπάγεται φυσικά πως ο κωδικοποιητής περνάει στον αποκωδικοποιητή για κάθε στοιχείο ακολουθίας πληροφορία ίση με το διπλάσιο μέγεθος της κρυφής κατάστασής του λόγω της αμφίδρομης φύσης του.

Οι επιλογές αυτές των μεγεθών πάρθηκαν καθώς φάνηκε από πειράματα ότι αυτό το εύρος τιμών παρουσιάζει καλά την επίπτωση του μεγέθους κρυφής κατάστασης στην απόδοση χωρίς να ξεφεύγουν πολύ οι χρόνοι εκπαίδευσης.

4.2.5 Επεξεργασία από τον αποκωδικοποιητή

Αφού ο κωδικοποιητής επεξεργαστεί όλη την ακολουθία εισόδου και παράξει τις τελικές αναπαραστάσεις της, είναι πλέον η σειρά του αποκωδικοποιητή να τις χρησιμοποιήσει μέσω του μηχανισμού προσοχής για να προβλέψει την ακολουθία εισόδου. Ο αποκωδικοποιητής, αποτελείται επίσης από ένα αναδρομικό νευρωνικό δίκτυο, δομημένο σε αρχιτεκτονική ένα-σε-πολλά και ρόλος του είναι να χρησιμοποιήσει την πληροφορία από την έξοδο του κωδικοποιητή για να παράξει την ακολουθία εξόδου του. Ξεκινώντας λοιπόν από ένα αρχικό σύμβολο, σε κάθε βήμα της διαδικασίας αυτός παράγει ένα στοιχείο ακολουθίας, το οποίο στην συνέχεια ανατροφοδοτείται στην είσοδό του ώστε να παράξει το επόμενο. Αυτό δημιουργεί μια αλυσιδωτή αντίδραση στην οποία δοσμένου ενός αρχικού συμβόλου, ο αποκωδικοποιητής μπορεί να παράξει ολόκληρη την ακολουθία εξόδου μέχρι και το τελικό σύμβολο. Μέσω του μηχανισμού προσοχής, ο αποκωδικοποιητής σε κάθε έξοδο της διαδικασίας του δεν συμβουλεύεται μόνο την προηγούμενη έξοδό του, αλλά και την πληροφορία του διανύσματος περιεχομένου (context vector) που περιέχει πληροφορία για όλα τα στοιχεία της ακολουθίας εισόδου. Έτσι μπορεί να επιλέγει πια στοιχεία της ακολουθίας εισόδου είναι τα πιο σημαντικά κάθε φορά.

Καθώς ο αποκωδικοποιητής με μηχανισμό προσοχής πρέπει σε κάθε βήμα της διαδικασίας να συμβουλεύεται το διάνυσμα περιεχομένου, πρέπει οι διαστάσεις του κρυφού του στρώματος του να ισούνται με τις διαστάσεις του διανύσματος, αφού εκεί τροφοδοτείται αυτή η πληροφορία. Καθώς όμως το διάνυσμα περιεχομένου έχει διπλάσια διάσταση από αυτή του κρυφού στρώματος του κωδικοποιητή λόγω της αμφίδρομης φύσης του, αυτό συνεπάγεται πως ο αποκωδικοποιητής θα πρέπει να έχει διπλάσιο μέγεθος από αυτό του κωδικοποιητή. Αυτό μπορεί να δημιουργήσει χρονικά προβλήματα καθώς και προβλήματα μνήμης αν το μέγεθος του κωδικοποιητή είναι πολύ μεγάλο. Για να αντιμετωπιστεί αυτό, μπορεί να τοποθετηθεί ένα πυκνό στρώμα ανάμεσα στην έξοδο του κωδικοποιητή και του αποκωδικοποιητή το οποίο θα ρίχνει τις διαστάσεις της εξόδου του κωδικοποιητή στις επιθυμητές. Μέσω αυτής της τεχνικής, μπορεί εν τέλει να ελεγχθεί το μέγεθος του αποκωδικοποιηη στην επιθυμητή τιμή του.

Αυτή η αρχιτεκτονική που περιγράφηκε παραπάνω ωστόσο δεν αρκεί για την πλήρη λειτουργία του αποκωδικοποιητή. Πρέπει ακόμα να περιγραφεί πως ο αποκωδικοποιητής αποφασίζει πια λέξη ακολουθίας θα παράξει στην έξοδο του. Όπως έχει αναλυθεί, οι λέξεις στόχοι των περιλήψεων είναι κατηγορηματικά δεδομένα και τα νευρωνικά δίκτυα τροφοδοτούνται και παράγουν στην έξοδο τους μόνο αριθμητικά. Γι' αυτό το σκοπό, η έξοδος του αναδρομικού νευρωνικού δικτύου του αποκωδικοποιητή περνάει από ένα πυκνό στρώμα με πλήθος εξόδων που ισούται με το μέγεθος λεξιλογίου του αποκωδικοποιητή. Με βάση αυτή την αρχιτεκτονική, ο αποκωδικοποιητής σε κάθε έξοδο του δίνει ένα διάνυσμα με μήκος όσο το πλήθος του λεξιλογίου, και με τιμές σε όλες τις θέσεις του. Στόχος του, είναι να μεγιστοποιήσει την τιμή του διανύσματος εξόδου στην θέση του που αντιστοιχεί στην ορθή λέξη πρόβλεψης. Παρατηρώντας έτσι σε πια θέση του εμφανίζεται η μέγιστη τιμή (softmax function) και παίρνοντας την αντίστοιχη λέξη του λεξιλογίου επιτυγχάνεται η μετατροπή της εξόδου του αποκωδικοποιητή σε λέξεις για την περίληψη.

Τα διαφορετικά μεγέθη αποκωδικοποιητών που επιλέχθηκαν να μελετηθούν έχουν μεγέθη κρυφής κατάστασης 256, 512 και 768 (το διπλάσιο μέγεθος της κρυφής κατάστασης του κωδικοποιητή) ενώ παράλληλα επιλέχθηκε να μελετηθεί και το μέγεθος κρυφής κατάστασης 384, για μέγεθος κωδικοποιητή 384 χρησιμοποιώντας ενδιάμεσο κρυφό στρώμα, όπως αναλύθηκε παραπάνω.

4.2.6 Διαδικασία υπολογισμού κόστους

Για να εκπαιδευτούν τα μοντέλα που περιγράφηκαν παραπάνω ώστε να καταφέρουν να παράξουν ορθές περιλήψεις από κείμενα, πρέπει να υπολογιστεί κάποιο κόστος απόκλισης της εξόδου του από τα αναμενόμενα αποτελέσματα και να βελτιωθεί με βάση αυτό, ώστε να προσεγγίσει καλύτερα επιθυμητές εξόδους. Ανάλογα όμως το είδος του προβλήματος μηχανικής μάθησης, υπάρχουν διαφορετικές συναρτήσεις κόστους που το εκφράζουν καλύτερα και οδηγούν σε καλύτερα αποτελέσματα. Όπως έχει περιγραφεί παραπάνω, το πρόβλημα της αυτόματης παραγωγής περίληψης καταλήγει σε πρόβλημα κατάταξης (classification) στην έξοδο του. Καθώς λοιπόν όπως έχει περιγραφεί παραπάνω, στα προβλήματα κατάταξης, μια συνάρτηση κόστους που τα εκφράζει πολύ καλά και οδηγεί καλά αποτελέσματα, είναι η συνάρτηση απώλειας εντροπίας (crossentropy loss function) χρησιμοποιείται αυτή για την έκφραση της απόκλισης των υλοποιημένων μοντέλων από τα επιθυμητά αποτελέσματα.

Στην ανάπτυξη μοντέλων μηχανικής μάθησης, αποτελεί σημαντικό ζήτημα η απόφαση για το πόσα δεδομένα εισόδου θα χρησιμοποιηθούν για τον υπολογισμό του κόστους και την εκπαίδευση του εκάστοτε μηχανισμού σε κάθε βήμα της διαδικασίας εκπαίδευσης. Το ιδανικό σε αυτή την περίπτωση θα ήταν να υπολογίζεται για όλα τα δεδομένα εισόδου η απόκλιση τους από τα επιθυμητά αποτελέσματα με βάση την συνάρτηση κόστους και να υπολογίζεται η συνολική απώλεια με χρήση όλων αυτών. Έτσι το εκάστοτε μοντέλο θα μπορεί σε κάθε βήμα εκπαίδευσης να λαμβάνει πλήρη γνώση για την απόκλιση των προβλέψεων των δεδομένων εκπαίδευσης από τις επιθυμητές εξόδους και να προσαρμόζεται ανάλογα.

Αυτή η στρατηγική ενημέρωσης των μεταβλητών του μηχανισμού, αν και ιδανική σε πραγματικές συνθήκες δεν είναι πολύ ρεαλιστική. Ειδικά τα τελευταία χρόνια, που η βαθιά μάθηση έχει αρχίσει να χρησιμοποιείται σε όλο και περισσότερες εφαρμογές, αυτή η στρατηγική ενημέρωσης βρίσκει εφαρμογή όλο και λιγότερο λόγω του μεγάλου όγκου δεδομένων που απαιτείται για την εκπαίδευση βαθιών συστημάτων. Αντί αυτού, σε κάθε βήμα της διαδικασίας, συνηθίζεται να χρησιμοποιείται ένα δείγμα των συνολικών δεδομένων (batch) και να πραγματοποιείται εκπαίδευση με βάση αυτό. Αυτή η στρατηγική εφαρμόζεται πολύ συχνά, κυρίως κατά την εκπαίδευση μοντέλων βαθιάς μάθησης και οδηγεί σε πολύ γρήγορα και ορθά αποτελέσματα.

Στην συγκεκριμένα εργασία επιλέχθηκε να χρησιμοποιούνται 64 δεδομένα εισόδου ανά τεμάχιο, για κάθε βήμα της εκπαίδευσης. Για τα εκάστοτε 64 δεδομένα εισόδου υπολογίζεται η έξοδος τους από τον μηχανισμό, μετριέται η απόκλιση τους από τις αναμενόμενες τιμές μέσω της συνάρτησης κόστους και υπολογίζεται το συνολικό κόστος ως η μέση τιμή αυτών των αποτελεσμάτων.

4.2.7 Διαδικασία βελτιστοποίησης

Αφού υπολογιστεί το κόστος απόκλισης σε κάθε βήμα της διαδικασίας μέσω της συνάρτησης κόστους, είναι η ώρα του βελτιστοποιητή (optimizer) να ενημερώσει τις μεταβλητές του συστήματος σύμφωνα με την διαδικασία που περιγράφηκε και πιο πάνω. Η επιλογή του βελτιστοποιητή σε όλα τα προβλήματα της μηχανικής μάθησης είναι πολύ σημαντική και μπορεί να μπορεί να καθορίσει μια ορθή και γρήγορη εκπαίδευση.

Στην συγκεκριμένη εργασία επιλέχθηκε ως βελτιστοποιητής σε όλες τις εκτελέσεις ο προσαρμοστικός εκτιμητής με ορμή (adaptive momentum estimator - Adam). Όπως αναλύθηκε και πιο πάνω ο βελτιστοποιητής αυτός, αποτελεί έναν πολύ εύρωστο μηχανισμό με λίγες αδυναμίες και με γρήγορους χρόνους σύγκλισης σε ορθά αποτελέσματα πρόβλεψης. Σε αυτόν τον βελτιστοποιητή ορίστηκε επίσης σε όλες τις εκτελέσεις, ο ρυθμός μάθησης (learning rate) να ισούται με 0.001. Σημειώνεται ότι ο ρυθμός μάθησης αντιστοιχεί στην σταθερή τιμή, με την οποία πολλαπλασιάζεται το βήμα εκπαίδευσης που έχει υπολογιστεί μέσω του αλγορίθμου βελτιστοποίησης, πριν την ενημέρωση της εκάστοτε μεταβλητής. Παρατηρήθηκε ότι με αυτόν τον ρυθμό μάθησης, το δίκτυο καταφέρνει να συγκλίνει στις επιθυμητές εξόδους σχετικά γρήγορα.

4.2.8 Ακτινική αναζήτηση κατά την πρόβλεψη

Όπως έχει ήδη αναλυθεί, κάθε φορά που προβλέπεται μια λέξη ακολουθίας στην έξοδο των μηχανισμών βαθιάς μάθησης, δεν είναι υποχρεωτικό να οδηγεί αυτή στην βέλτιστη τελική ακολουθία. Μπορεί για παράδειγμα μια λέξη που δεν είχε την μεγαλύτερη τιμή στην έξοδο του δικτύου να οδηγεί εν τέλει σε καλύτερη τελική περίληψη κειμένου. Για να αντιμετωπιστεί αυτό στην έξοδο του των διάφορων μηχανισμών που υλοποιήθηκαν χρησιμοποιήθηκε ακτινική αναζήτηση με μέγεθος 5. Αυτό σημαίνει, πως σε κάθε στάδιο πρόβλεψης δεν επιλέγεται μια λέξη για την έξοδο αλλά επιλέγονται οι 5 πιο πιθανές λέξεις με βάση το σκορ τους στο πυκνό στρώμα εξόδου. Για κάθε μια από αυτές τις λέξεις συνεχίζεται η ακολουθία με άλλες 5 λέξεις και αυτή η διαδικασία επαναλαμβάνεται μέχρι να έχουν ολοκληρωθεί όλες οι ακολουθίες. Στο τέλος, η τελική περίληψη του εκάστοτε κειμένου επιλέγεται να είναι εκείνη που έχει μαζέψει το συνολικό περισσότερο σκορ (το σκορ όλων των λέξεων που χρησιμοποιεί, όπως αυτό προέκυψε στην έξοδο του δικτύου για την εκάστοτε λέξη).

4.2.9 Χειρισμός άγνωστων λέξεων

Όπως έχει ήδη περιγραφεί ο μηχανισμός που χρησιμοποιείται για την αυτόματη παραγωγή περιλήψεων από κείμενα χρησιμοποιεί συγκεκριμένο λεξιλόγιο, το οποίο ο κωδικοποιητής χρησιμοποιεί για να μετατρέψει τις λέξεις εισόδου σε αριθμητικές αναπαραστάσεις κατανοητές από αυτόν και ο αποκωδικοποιητής το χρησιμοποιεί για να μετατρέψει με την σειρά του τις λέξεις στόχο της εκάστοτε περίληψης σε αριθμητικές παραστάσεις κατά την εκπαίδευση. Το λεξιλόγιο είναι επίσης σημαντικό για να γνωρίζει ο αποκωδικοποιητής τις πιθανές λέξεις που μπορεί να παράξει στην έξοδο του, κατά την πρόβλεψή του. Φυσικά, το λεξιλόγιο του κωδικοποιητή και του αποκωδικοποιητή μπορεί να είναι διαφορετικά σε μέγεθος.

Ένα πρόβλημα που πρέπει να σκεφτεί κανείς κατά τον σχεδιασμό ενός τέτοιου συστήματος αποτελεί το μέγεθος αυτού του λεξιλογίου. Το μέγεθος αυτό δεν μπορεί να είναι πάρα πολύ μεγάλο καθώς κάτι τέτοιο θα αύξανε πολύ τις απαιτήσεις μνήμης και θα καθυστερούσε πολύ τον χρόνο εκπαίδευσης καθώς και τον χρόνο πρόβλεψης, κάτι το οποίο δεν είναι ποτέ επιθυμητό. Το μέγεθος επίσης δεν μπορεί να είναι πολύ μικρό, καθώς τότε το σύστημα μηχανικής δεν θα είχε ικανοποιητικό πλήθος λέξεων να επιλέξει και θα αδυνατούσε να γενικεύσει πλήρως σε νέα δεδομένα εισόδου με διαφορετικό λεξιλόγιο. Είναι σημαντικό λοιπόν να δοθεί πού προσοχή στην επιλογή αυτού του μεγέθους.

Ορίζοντας όμως ένα συγκεκριμένο μέγεθος λεξιλογίου και απαγορεύοντας την εισαγωγή νέων λέξεων σε αυτό για λόγους ταχύτητας και απαιτήσεων μνήμης δημιουργεί το πρόβλημα των άγνωστων λέξεων. Σε ένα τέτοιο σύστημα θα πρέπει να προβλεφθεί πολύ προσεκτικά πως το σύστημα θα συμπεριφέρεται σε άγνωστες λέξεις καθώς δεν γίνεται να εξασφαλιστεί ότι τα κείμενα εισόδου δεν θα περιέχουν άγνωστες λέξεις και ούτε γίνεται αυτές να αγνοηθούν. Ένα σύστημα που αρνείται να δεχθεί εισόδους που περιέχουν άγνωστες λέξεις δεν έχει καμία εφαρμογή στον πραγματικό κόσμο.

Όπως είχε αναφερθεί και πιο πριν λοιπόν, για την αντιμετώπιση άγνωστων λέξεων που το λεξιλόγιο δεν μπορεί να μεταφράσει σε αριθμητικές μορφές κατάλληλες για το δίκτυο, έχει εισαχθεί η ειδική λέξη unk. Για αυτή την λέξη έχει δεσμευτεί μια συγκεκριμένη αριθμητική αναπαράσταση και όσες λέξεις δεν μπορούν να αντιστοιχιστούν με μία αντιστοιχίζονται με αυτής. Αυτή η πρακτική βοηθάει στην ευρωστία του συστήματος καθώς καλύπτει πιθανές εξαιρέσεις στην είσοδο του αλλά δεν αρκεί για την πλήρη αντιμετώπιση του θέματος. Για να γίνει κατανοητό αυτό θα δοθεί ένα παράδειγμα στην συνέχεια.

Ας υποθέσουμε ότι θέλουμε να πραγματοποιηθεί περίληψη σε ένα κείμενο που αφορά ένα διάσημο πολιτικό πρόσωπο. Σε μία γενική εκπαίδευση ενός μοντέλου αυτόματης παραγωγής περίληψης δεν μπορεί να εξασφαλιστεί ότι το όνομα αυτού του προσώπου θα υπάρχει στο λεξιλόγιο του μοντέλου μας. Αν όμως αυτό το πρόσωπο είναι πολύ σημαντικό και αναφέρεται συχνά στο συγκεκριμένο κείμενο, πιθανώς το πρέπει η περίληψη να αναφερθεί σε αυτό. Αντικαθιστώντας το όνομα του με unk δεν θα μας λύσει αυτό το πρόβλημα καθώς η πληροφορία για το όνομα του θα χαθεί τελείως και στην καλύτερη το όνομα του θα εμφανιστεί ως unk στην τελική περίληψη.

Είναι αναγκαία λοιπόν η ύπαρξη ενός μηχανισμού χειρισμού άγνωστων λέξεων, έτσι ώστε αυτές να μπορούν να εμφανιστούν στην τελική περίληψη αν είναι απαραίτητο. Αυτό το πρόβλημα αποτελεί πολύ σημαντικό πρόβλημα, και η αντιμετώπιση του είναι πολύ σημαντικό με στόχο την επίτευξη ενός συστήματος παραγωγής περιλήψεων γενικού σκοπού. Στο επιστημονικό άρθρο [See17] έγινε μια προσπάθεια να αντιμετωπιστεί αυτό το πρόβλημα με χρήση υβριδικών δικτύων παραγωγής-αντιγραφής (pointer - generator networks). Σε αυτές τις αρχιτεκτονικές αποφασίζεται μέσω κάποιων μεταβλητών που μπορούν να εκπαιδευτούν σε κάθε στάδιο πρόβλεψης του αποκωδικοποιητή, αν αυτός θα παράξει λέξη στην έξοδο του ή θα αντιγραφεί κάποια λέξη από το αρχικό κείμενο. Αυτή η απόφαση αντιγραφής μπορεί να επιλύσει το πρόβλημα των άγνωστων λέξεων καθώς δεν απαιτείται η εκάστοτε άγνωστη λέξη να υπάρχει στο λεξιλόγιο και να υπάρχει αριθμητική αναπαράστασή της, αλλά αυτή αντιγράφεται κατευθείαν στην τελική έξοδο. Αντίστοιχες στρατηγικές αντιγραφής έχουν χρησιμοποιηθεί και σε άλλες εργασίες και χρησιμοποιούνται στο σύστημα αυτόματης μετάφρασης της google (google translate) [Wu16].

Στα πλαίσια αυτής της εργασίας έχει υλοποιηθεί ένας διαφορετικός μηχανισμός που χρησιμοποιεί δύο τεχνικές για να εισάγει τις άγνωστες λέξεις του αρχικού κειμένου στα τελικά αποτελέσματα, όπου αυτό είναι απαραίτητο. Ο μηχανισμός αυτός βασίζεται στην νοοτροπία ότι σε περίπτωση αντικατάστασης άγνωστης λέξης με την λέξη unk κατά την εισαγωγή ακολουθίας στο μοντέλο παραγωγής περίληψης, αν αυτή η λέξη είναι σημαντική για το νόημα του τελικού κειμένου θα εμφανιστεί στην τελική περίληψη η λέξη unk. Φυσικά στην τελική περίληψη μπορεί να εμφανιστούν περισσότερα από μία unk λέξη, ενώ παράλληλα μπορεί και πολλαπλές διαφορετικές λέξεις στο κείμενο να αλλαχθούν με την λέξη unk.

Με βάση αυτή την παραδοχή, ο μηχανισμός χειρισμού άγνωστων λέξεων ελέγχει στην τελική ακολουθία εξόδου για πιθανές εμφανίσεις unk. Στόχος του είναι να αντικαταστήσει αυτές τις λέξεις με λέξεις από το κείμενο τις οποίες δεν γνωρίζει. Αξίζει να σημειωθεί ότι είναι πολύ εύκολο για τον μηχανισμό να αποφανθεί αν γνωρίζει μια λέξη, αφού αρκεί να ελέγξει αν υπάρχει καταχώρηση για αυτήν στο λεξιλόγιο του συστήματος. Φυσικά πρέπει να αποφασιστεί, σε περίπτωση πολλών άγνωστων λέξεων του κειμένου και σε περίπτωση πολλαπλών unk στην τελική περίληψη, πια λέξη θα αντικατασταθεί με πιο unk.

Για την απόφαση αυτή, στον μηχανισμού εμφανίζεται η στρατηγική διερεύνησης περιοχής λέξης. Μέσω αυτής ορίζονται κοντές περιοχές τόσο στις άγνωστες λέξεις του κειμένου εισόδου καθώς και στις αναπαραστάσεις unk της τελικής περίληψης. Ως κοντινή περιοχή (ή παράθυρο) μιας λέξης ορίζεται ένα πλήθος λέξεων που βρίσκονται πριν και μετά από αυτήν σε μια πρόταση. Το μήκος περιοχής (ή μήκος παραθύρου) καθορίζει το πόσες λέξεις πριν ή μετά από αυτήν θεωρούνται ότι ανήκουν στην περιοχή. Στόχος είναι, για κάθε ζεύγος άγνωστης λέξης κειμένου - αναπαράστασης unk στην περίληψη, να βρεθεί κατά πόσο οι γειτονικές περιοχές τους μοιάζουν. Η άγνωστη λέξη που ταιριάζει περισσότερο σε κάθε περιοχή της περίληψης, αντικαθιστά την αναπαράσταση unk και παίρνει την θέση της. Παρακάτω αναλύεται καλύτερα ο αλγόριθμος αυτός αντικατάστασης:

Αλγόριθμος 1 Αλγόριθμος αντικατάστασης άγνωστων λέξεων στην τελική περίληψη

1: Επέλεξε αλγόριθμο μέτρησης ομοιότητας παραθύρου (συνάρτηση score)
2: **Για** i... στις λέξεις unk της τελικής περίληψη:
3: Θέσε $maxscore = 0$
4: **Για** j... στις άγνωστες λέξεων του κειμένου εισόδου:
5: Υπολόγισε την ομοιότητα παραθύρου της λέξης j και της λέξης i $score(i, j)$
6: **Αν** $score(i, j) > maxscore$:
7: θέσε $maxscore = score(i, j)$
8: θέσε $unknown\_word = j$
9: **Τέλος Αν**
10: **Τέλος Για**
11: θέσε την $unknown\_word$ στην θέση της j στην τελική περίληψη
12: **Τέλος Για**

Στην συνέχεια θα πραγματοποιηθεί μια μικρή ανάλυση για τους αλγορίθμους ομοιότητας που χρησιμοποιήθηκαν.

Ο πρώτος αλγόριθμος ομοιότητας χρησιμοποιεί την ομοιότητα συνημιτόνου (cosine similarity) για να αποφασίσει ποιες περιοχές μοιάζουν πιο πολύ μεταξύ τους. Η ομοιότητα συνιμοτόνου εφαρμόζεται πάνω σε διανύσματα και μετράει το συνημίτονο της μεταξύ τους γωνίας στον ν-διάστατο χώρο που αυτά ορίζονται. Όσο μεγαλύτερο προκύψει αυτό το συνημίτονο τόσο πιο πολύ μοιάζουν αυτά τα διανύσματα καθώς η γωνία μεταξύ τους μικραίνει. Αν το συνημίτονο προκύψει 1, αυτό σημαίνει πως τα διανύσματα είναι ακριβώς τα ίδια. Η μαθηματική έκφραση της ομοιότητας συνημιτόνου για δύο διανύσματα A, B φαίνεται στην σχέση 4.2:

$$similarity = \cos(\theta) = \frac{\mathbf{AB}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4.2}$$

Αξίζει να σημειωθεί, πως όπως είχε αναλυθεί και πριν οι διανυσματικές αναπαραστάσεις των λέξεων μέσω των αναπαραστάσεων Glove αποκτούν μαθηματικές εξαρτήσεις που σημαίνει ότι νοηματικά κοινές λέξεις απεικονίζεται από όμοια διανύσματα. Αυτή η ιδιότητα είναι πολύ σημαντική για την παραπάνω μετρική ομοιότητας καθώς επιβραβεύει νοηματικά όμοιες λέξεις, καθώς οδηγούν σε μεγαλύτερη ομοιότητα συνημιτόνου από άλλες νοηματικά διαφορετικές λέξεις. Αυτό σημαίνει πως η ορθή περιοχή αντικατάστασης μπορεί να βρεθεί παρόλο που μπορεί να μην υπάρχουν οι ίδιες λέξεις στην περιοχή του κειμένου και της περίληψης αλλά υπάρχει νοηματικό περιεχόμενο.

Μέσω αυτού του μηχανισμού, η ομοιότητα συνημιτόνου των λέξεων των δύο περιοχών ελέγχεται κάθετα, έτσι ώστε η κάθε λέξη στην εκάστοτε περιοχή να ελέγχεται μόνο με την αντίστοιχη λέξη στην άλλη περιοχή, η οποία βρίσκεται στην ίδια θέση παραθύρου. Από τα συνολικές μετρικές που υπολογίζονται για κάθε κάθετο έλεγχο προκύπτει τελικά η συνολική μετρική ομοιότητας παραθύρου ως ο μέσος όρος αυτών. Έτσι προκύπτει μια συνολική έκφραση της ομοιότητας σύμφωνα με τον αλγόριθμο της ομοιότητας συνημιτόνου, η οποία είναι αρκετά ανθεκτική σε νοηματικά συνώνυμες λέξεις.

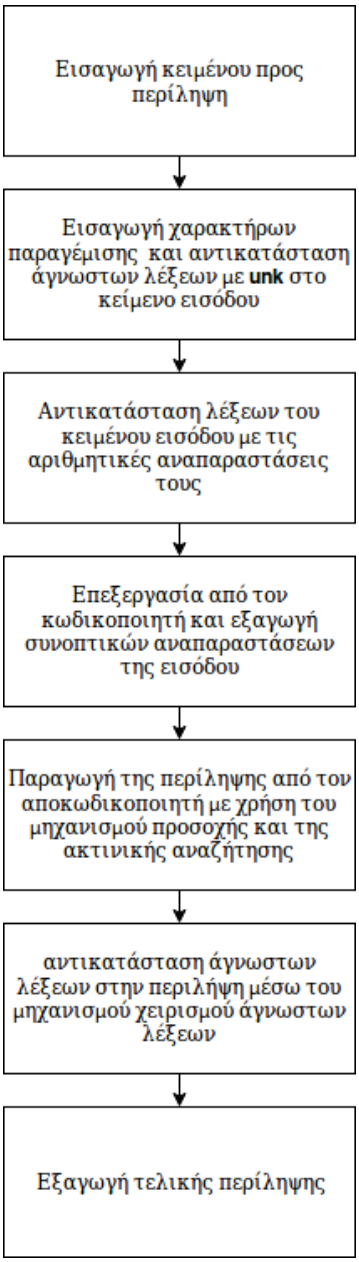
Ο δεύτερος αλγόριθμος που χρησιμοποιήθηκε για μέτρηση του σκορ ομοιότητας χρησιμοποιεί την ομοιότητα Jaccart (Jaccart similarity). Αυτή μετρική ελέγχει για τα δύο σύνολα των λέξεων που εκπροσωπούν οι περιοχές σε κάθε περίπτωση, πόσες λέξεις είναι κοινές και στα δύο αυτά. Είναι εμφανές ότι αυτή η τεχνική δεν είναι το ίδιο ανθεκτική σε συνώνυμες λέξεις αλλά απαιτεί να υπάρχουν ακριβώς ίδιες λέξεις στα δύο παράθυρα. Η μαθηματική έκφραση της ομοιότητας Jaccart για τις δύο περιοχές A,B φαίνεται στην σχέση 4.3

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{4.3}$$

Τα αποτελέσματα εφαρμογής αυτών των μεθόδων και η σύγκριση τους θα πραγματοποιηθεί στην επόμενη ενότητα.

4.2.10 Διάγραμμα ροής συστήματος αυτόματης περίληψης κειμένου

Για την καλύτερη κατανόηση της λειτουργίας του μηχανισμού θα παρουσιαστεί η λειτουργία του για δεδομένο κείμενο εισόδου σε ένα διάγραμμα ροής παρακάτω:



Σχήμα 4.2: Διάγραμμα ροής λειτουργίας του μηχανισμού παραγωγής περίληψης

4.2.11 Συνοπτική αναπαράσταση παραμέτρων

Σε αυτό το σημείο θα παρουσιαστούν συνοπτικά σε μορφή πίνακα όλες οι αριθμητικές τιμές των παραμέτρων του μηχανισμού παραγωγής περίληψης με στόχο τον ευκολότερο εντοπισμό τους από τον αναγνώστη. Αν μια παράμετρος διερευνάται για παραπάνω από μία αριθμητικές τιμές, αυτές θα διαχωρίζονται με κόμμα (,).

Πίνακας 4.1: Συνοπτική παρουσίαση αριθμητικών παραμέτρων μηχανισμού

	Αριθμητικές τιμές παραμέτρων
Μέγεθος λεξιλογίου (*1000)	50, 103 (απεριόριστο)
Μέγιστο μήκος ακολουθίας εισόδου	50
Μέγιστο μήκος ακολουθίας εξόδου	15
Μέγεθος κρυφής κατάστασης κωδικοποιητή	128, 256, 384
Μέγεθος κρυφής κατάστασης αποκωδικοποιητή	256, 384, 512, 768
Μέγεθος δέσμης εκπαίδευσης (batch size)	64
Ρυθμός εκπαίδευσης (learning rate)	0.001
Μέγεθος ακτινικής αναζήτησης	5
Μέγεθος παραθύρου μηχανισμού άγνωστων λέξεων	1, 2, 3

Κεφάλαιο 5

Εφαρμογή και αξιολόγηση

5.1 Γενικά

Αφού αναλύθηκαν και αιτιολογήθηκαν οι σχεδιαστικές επιλογές των υλοποιημένων μοντέλων, σε αυτό το κεφάλαιο θα παρουσιαστούν τα αποτελέσματα από τα διαφορετικά μοντέλα που σχεδιάστηκαν. Αξίζει να σημειωθεί πως η επιλογή των παραμέτρων των μοντέλων πραγματοποιήθηκε με τέτοιον τρόπο, ώστε τα τελικά αποτελέσματα να έχουν διερευνητικό χαρακτήρα, δηλαδή να διερευνούν πως επηρεάζουν διαφορετικές τιμές παραμέτρων τα τελικά αποτελέσματα και ποιες επιλογές τελικά αξίζουν περισσότερο συνυπολογίζοντας και το χρονικό κόστος της εκπαίδευσης. Καθώς το χρονικό κόστος της εκπαίδευσης ωστόσο δεν επηρεάζει την μετέπειτα λειτουργία του εκπαιδευμένου μοντέλου και καθώς η συγκεκριμένη εργασία κινείται σε ερευνητικά πλαίσια, κατά την σύγκριση θα δίνεται περισσότερο βάρος στην τελική απόδοση και αυτή θα σχολιάζεται περισσότερο.

Με στόχο η σύγκριση των υλοποιημένων μοντέλων να είναι όσο πιο αντικειμενική γίνεται κάθε μοντέλο εκπαιδεύτηκε για 10 εποχές, παρόλο που όπως θα δούμε κάποια συνέκλιναν πολύ πιο πριν από την τελευταία. Στην ανάλυση που ακολουθεί, θα γίνει αρχικά μια περιγραφή του μηχανισμού που θα χρησιμοποιηθεί για έλεγχο της απόδοσης του εκάστοτε μοντέλου. Θα περιγραφεί γιατί επιλέχθηκε ο συγκεκριμένος μηχανισμός και θα παρουσιαστεί ο αλγόριθμος που χρησιμοποιεί για τις μετρήσεις του. Με χρήση αυτού του μηχανισμού, θα παρουσιαστούν στην συνέχεια τα αποτελέσματα των 10 εποχών για όλα τα υλοποιημένα μοντέλα και από αυτά θα επιλεχθούν τα καλύτερα. Με χρήση των καλύτερα αυτών μοντέλων θα πραγματοποιηθούν στην συνέχεια συγκρίσεις με στόχο την διερεύνηση της επίδρασης διάφορων σχεδιαστικών επιλογών στην αντιμετώπιση του προβλήματος την αυτόματης παραγωγής περίληψης. Στο τέλος του κεφαλαίου, θα παρουσιαστούν και τα αποτελέσματα των μηχανισμών χειρισμού άγνωστων λέξεων, με στόχο την ανάλυση της επίδρασης τους στον συνολικό μηχανισμό, καθώς και για να αποφασιστεί ποιος είναι ο καλύτερος από αυτούς, μέσω της μεταξύ τους σύγκρισης.

Αξίζει να σημειωθεί, ότι ρόλος της συγκεκριμένης εργασίας, δεν αποτελεί απλά η διερεύνηση την επίδρασης συγκεκριμένων παραμέτρων στα υλοποιημένα μοντέλα. Σε αυτή την εργασία γίνεται και μια προσπάθεια για βελτίωση ήδη υπάρχοντων μηχανισμών με στόχο την προώθηση της έρευνας στο αντικείμενο της αυτόματης παραγωγής περίληψης. Γι' αυτό το λόγο, όπως αναλύθηκε και σε προηγούμενο κεφάλαιο, τα σύνολα δεδομένων που επιλέχθηκαν για την εκπαίδευση, είναι αυτά στα οποία έχει πραγματοποιηθεί ήδη έρευνα, με στόχο την δυνατότητα σύγκρισης των αποτελεσμάτων αυτής της εργασίας με άλλες ερευνητικές εργασίες. Αντίστοιχα, ο μηχανισμός μέτρησης της απόδοσης επιλέχθηκε για τον ίδιο λόγο. Στα πλαίσια αυτής της εργασίας, το σύστημα που υλοποιήθηκε στηρίχθηκε στις αρχές των συστήματος που παρουσιάστηκαν στα [Bahd14] και [Nall16], ενώ παράλληλα χρησιμοποιήθηκαν και τα ίδιο σύνολα δεδομένων με αυτά. Συνεπώς στα αποτελέσματα μας θα πραγματοποιηθεί και σύγκριση με τα αποτελέσματα αυτών των μεθόδων, στην οποία θα φανεί η βελτίωση που επιτεύχθηκε.

5.2 Μετρικές επίδοσης

5.2.1 Γενικά

Η μετρική που χρησιμοποιήθηκε στα πλαίσια αυτής της εργασίας για την μέτρηση της επίδοσης των μοντέλων, ονομάζεται μετρική Rouge (Recall-Oriented Understudy for Gisting Evaluation). Η μετρική Rouge χρησιμοποιείται πολύ συχνά σε συστήματα επεξεργασίας φυσικών γλωσσών ως μέτρο σύγκρισης της επίδοσης ερευνητικών συστημάτων. Η μετρική αυτή αποτελείται από ένα πλήθος αλγορίθμων που προσπαθούν να εντοπίσουν ομοιότητες ανάμεσα προτάσεις φυσικής γλώσσας. Όσο μεγαλύτερο σκορ συλλέγουν αυτοί οι αλγόριθμοι, τόσο πιο πολλές ομοιότητες έχουν οι προτάσεις που συγκρίνονται.

Στα πλαίσια αυτής της εργασίας, η επίδοση θέλουμε να υπολογιστεί ανάμεσα στις περιλήψεις που παράγονται από τα μοντέλα και στις ”ιδανικές” περιλήψεις που περιέχουν τα σύνολα δεδομένων μας. Στόχος είναι να βαθμολογηθούν ως προς την ομοιότητα τους, καθώς ως ορθές περιλήψεις θεωρούνται οι περιλήψεις των συνόλων δεδομένων. Ο όρος ορθή περίληψη είναι λίγο αμφιλεγόμενος και δύσκολα μπορεί να οριστεί μαθηματικά σωστά, ώστε η μετρική επίδοσης να αποτελεί μια αντικειμενική συνάρτηση επίδοσης για τις παραγόμενες περιλήψεις. Καθώς όμως τα μοντέλα έχουν εκπαιδευτεί σε παρόμοια δεδομένα με αυτά των συνόλων μέτρησης της επίδοσης, θεωρείται ότι προσπαθούν να προσεγγίσουν περιλήψεις που μοιάζουν στο ύφος με τις περιλήψεις σύγκρισης, οπότε θεωρούνται αυτές ιδανικές.

Σε ένα διαφορετικό σύνολο εκπαίδευσης φυσικά, οι ιδανικές περιλήψεις θα ήταν διαφορετικές τόσο ως προς το ύφος, αλλά και ως προς το λεξιλόγιο. Το γεγονός ότι τα μοντέλα μαθαίνουν ουσιαστικά να παράγουν περιλήψεις σύμφωνα με συγκεκριμένα στυλ των δεδομένων εκπαίδευσης δυσκολεύει φυσικά την έρευνα του τομέα της αυτόματης παραγωγής περίληψης καθώς γίνεται πολύ δύσκολο να φτιαχτεί ένα καθολικό σύστημα περίληψης που θα γενικεύει σε διαφορετικά είδη κειμένων. Αυτό το φαινόμενο θα φανεί και καλύτερα και παρακάτω που θα μετρηθεί η επίδοση στο σύνολο δεδομένων DUC στο οποίο οι περιλήψεις στόχου έχουν παραχθεί με διαφορετικό τρόπο από ότι στο σύνολο εκπαίδευσης που χρησιμοποιήθηκε (Gigaword corpus). Στην συνέχεια θα παρουσιαστούν κάποιοι βασικοί αλγόριθμοι Rouge, που χρησιμοποιήθηκαν για την μέτρηση της απόδοσης στα πλαίσια αυτής της εργασίας.

5.2.2 Μετρικές Rouge

Όλοι οι διαφορετικοί αλγόριθμοι Rouge που θα παρουσιαστούν σε αυτήν την υποενότητα, έχουν ως στόχο να μοντελοποιήσουν την σύγκριση των περιλήψεων που απαιτείται, με μαθηματικό τρόπο. Γι’ αυτό το σκοπό, όλοι αυτοί οι αλγόριθμοι προσπαθούν να συγκρίνουν το κατά πόσο διαφορετικά είδη υποακολουθιών των περιλήψεων εμφανίζονται και στις δύο περιλήψεις. Στα πλαίσια αυτής της εργασίας, χρησιμοποιούνται 3 αλγόριθμοι Rouge για σύγκριση των ακολουθιών, και παρουσιάζονται παρακάτω. Σημειώνεται πως όλοι αυτοί οι αλγόριθμοι πραγματοποιούν μετρήσεις σε συγκεκριμένες μετρικές, οι οποίες θα παρουσιαστούν και αυτές.

- **Αλγόριθμος ROUGE-1:** Ο αλγόριθμος ROUGE-1 προσπαθεί να μοντελοποιήσει μαθηματικά την ομοιότητα ακολουθιών λέξεων, μετρώντας πόσες λέξεις εμφανίζονται κοινές στις εκάστοτε δύο περιλήψεις. Μέσω αυτού υπολογίζεται κατά πόσο το νοηματικό περιεχόμενο των δύο περιλήψεων κινείται σε κοντινά πλαίσια.

- **Αλγόριθμος ROUGE-2:** Ο αλγόριθμος ROUGE-2 κινείται στα ίδια πλαίσια, αλλά μετράει πόσα ζεύγη λέξεων εμφανίζονται κοινά στις εκάστοτε περιλήψεις. Μέσω αυτής της μετρικής, δίνεται περισσότερη σημασία, στους συντακτικούς κανόνες και ελέγχεται κατά πόσο η περίληψη που έχει προβλεφθεί από το εκάστοτε μοντέλο ακολουθεί και σωστή συντακτική δομή.

- **Αλγόριθμος ROUGE-L:** Ο αλγόριθμος ROUGE-L έχει λίγο διαφορετική φιλοσοφία σύγκρισης και προσπαθεί να εντοπίσει την πιο μεγάλη κοινή ακολουθία λέξεων (lonest common sequence). Αποτελεί άλλο ένα μέτρο που προσπαθεί να μοντελοποιήσει την ομοιότητα δύο ακολουθιών λέξεων, απλώς από άλλη οπτική.

Όλοι αυτοί οι αλγόριθμοι υπολογίζουν τις ομοιότητες των ακολουθιών σύμφωνα με τους κανόνες που περιγράφηκαν παραπάνω, και πραγματοποιούν μετρήσεις στις εξής μετρικές:

- **Ανάκληση (Recall):** Στην μετρική της ανάκλησης, για τον εκάστοτε αλγόριθμο Rouge υπολογίζεται το πλήθος των κοινών ακολουθιών που βρέθηκαν, σύμφωνα με την λειτουργία του καθενός, δια το συνολικό μήκος της ”ορθής” περίληψης. Ουσιαστικά αυτή η μετρική μετράει πόσο μέρος της ”ορθής” περίληψης εμφανίζεται και σε αυτήν που έχει προβλεφθεί από τον μηχανισμό.

$$Recall = \frac{number\_of\_overlaping\_sequence\_tokens}{total\_tokens\_in\_reference\_summary} \tag{5.1}$$

- **Ακρίβεια (Precision):** Η μετρική της ακρίβειας προσπαθεί ποινικοποιήσει πολύ μεγάλες προβλεπόμενες περιλήψεις που μπορεί να περιέχουν πολλές κοινές λέξεις με την περίληψη αναφοράς λόγω του μεγάλου μήκους τους κάνοντας την μετρική ανάκλησης να προκύψει πολύ μεγάλη. Η μετρική της ακρίβειας γι αυτό τον λόγο υπολογίζει το πλήθος των κοινών ακολουθιών των περιλήψεων, δια το συνολικό μήκος της προβλεπόμενης περίληψης.

$$Precision = \frac{number\_of\_overlaping\_sequence\_tokens}{total\_tokens\_in\_predicted\_summary} \tag{5.2}$$

- **Μετρική F1 (F1 score):** Η μετρική F1 προσπαθεί να μοντελοποιήσει τις δύο προηγούμενες μετρικές σε μια κοινή, ώστε να μεγιστοποιείται όταν μεγιστοποιούνται και οι άλλες. Αποτελεί πιο πολύ ένας μαθηματικός συνδυασμός τους χωρίς εμφανές πρακτικό νόημα. Χρησιμοποιείται παρόλα αυτά πολύ συχνά καθώς καταφέρνει να αναπαριστά και τις δύο προηγούμενες μετρικές σε έναν τελικό αριθμό.

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.3}$$

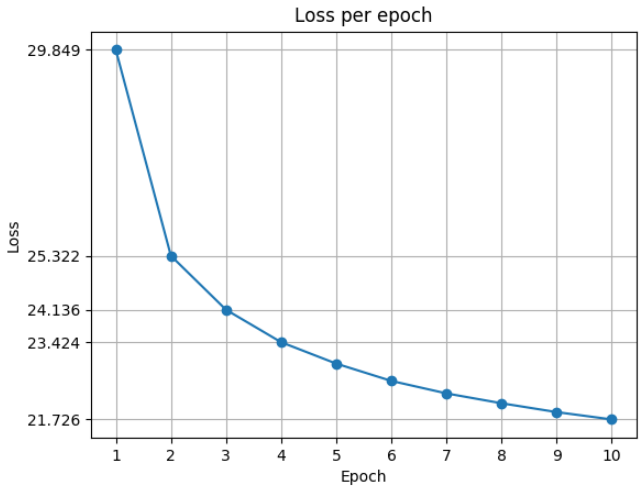
Αξίζει να σημειωθεί ότι όλες οι παραπάνω μετρικές βαθμολογούν σε ποσοστιαία κλίμακα στο εύρος 0-100, όπου σκορ 100 υποδηλώνει πως οι περιλήψεις που συγκρίθηκαν είναι πανομοιότυπες ως προς την εκάστοτε μετρική. Όταν η μετρική F1 είναι 100, αυτό σημαίνει πως και η ακρίβεια και η ανάκληση έχουν σκορ 100, και υποδηλώνει πως οι περιλήψεις είναι ολόιδιες μεταξύ τους. Στην συνέχεια του κεφαλαίου θα παρουσιαστεί η επίδοση όλων των μοντέλων που υλοποιήθηκαν για όλες τις εποχές εκπαίδευσης, με στόχο να επιλεχθούν τα καλύτερα για περαιτέρω μελέτη στην συνέχεια.

5.3 Μοντέλα εκπαίδευσης

Με στόχο την σύγκριση των διαφορετικών μοντέλων που υλοποιήθηκαν, πρέπει σαν πρώτο βήμα να αποφασιστεί πια εκδοχή αυτών των μοντέλων είναι η βέλτιστη. Παρόλο που κάθε ένα από αυτά τα μοντέλα έχει εκπαιδευτεί για 10 εποχές, μπορεί η βέλτιστη απόδοση του να μην στην τελικώς εκπαιδευμένη εκδοχή του, λόγω του προβλήματος της υπερεκπαίδευσης. Για να αποφασιστεί πια εκδοχή θα επιλεχθεί σε κάθε διαφορετικό είδους μοντέλο, θα παρουσιαστούν στατιστικά, που θα απεικονίζουν την βαθμολογία ROUGE για όλους τους αλγορίθμους που περιγράφηκαν παραπάνω σε ένα μικρό σύνολο δεδομένων ελέγχου (validation set) μεγέθους 2000 αντιστοιχιών άρθρου - περίληψης. Ταυτόχρονα θα παρουσιαστούν στατιστικά για την πτώση του κόστους (loss) σε κάθε εποχή για να παρουσιαστεί καλύτερα η σύγκλιση του εκάστοτε μοντέλου μέσω της εκπαίδευσης. Έτσι θα γίνει πιο σαφής η επιλογή των τελικών μοντέλων που θα συγκριθούν για εξαγωγή συμπερασμάτων καθώς και για την μελέτη της συμπεριφοράς διάφορων υπερπαραμέτρων στο μοντέλο αφαιρετικής παραγωγής περίληψης με χρήση βαθιάς μάθησης. Σημειώνεται πως η τελική επιλογή της καλύτερης εκδοχής κάθε μοντέλου θα βασιστεί κυρίως στις τιμές των μετρικών Rouge καθώς η επιλογή με βάση την πτώση της συνάρτησης κόστους μπορεί να είναι παραπλανητική λόγω του φαινόμενου της υπερεκπαίδευσης. Τέλος αξίζει να αναφερθεί, πως σε όλα τα παρακάτω μοντέλα, εκτός αν έχει ορισθεί διαφορετικά, το μέγεθος κρυφής κατάστασης του κωδικοποιητή είναι 384, το μέγεθος κρυφής κατάστασης του αποκωδικοποιητή είναι το διπλάσιο και το μέγεθος του λεξιλογίου είναι 50000.

5.3.1 Μοντέλο με απεριόριστο μέγεθος λεξιλογίου

Με τον όρο απεριόριστο μέγεθος λεξιλογίου, υπονοείται πως στο συγκεκριμένο μοντέλο δεν υπάρχει περιορισμός στο πλήθος των λέξεων που μπορούν να χρησιμοποιηθούν ως λεξιλόγιο, αλλά επιλέγονται όλες οι λέξεις που εμφανίζονται στα σύνολα εκπαίδευσης. Με αυτό τον τρόπο εξασφαλίζεται ότι το μοντέλο θα μπορεί να χρησιμοποιήσει μεγαλύτερο πλήθος λέξεων περιορίζοντας έτσι το ποσοστό των άγνωστων λέξεων που εμφανίζονται. Όσο μεγαλώνει το πλήθος λεξιλογίου φυσικά τόσο μεγαλώνει και ο χρόνος της εκπαίδευσης και πρόβλεψης, καθώς μεγαλώνει το μέγεθος του στρώματος εξόδου του μοντέλου, αυξάνοντας την συνολική πολυπλοκότητα του συστήματος. Η μελέτη του συγκεκριμένου μοντέλου γίνεται με σκοπό να διερευνηθεί στην συνέχεια η επίδραση του μεγέθους του λεξιλογίου στην συνολική απόδοση του συστήματος, καθώς όπως θα φανεί και στην συνέχεια, μεγαλύτερο πλήθος λεξιλογίου δεν σημαίνει υποχρεωτικά καλύτερη απόδοση καθώς όσο αυξάνονται οι πιθανές λέξεις προς πρόβλεψη, τόσο αυξάνεται και η πιθανότητα λάθος πρόβλεψης.



Σχήμα 5.1: Μέση τιμής της συνάρτησης κόστους για το μοντέλο με απεριόριστο μέγεθος λεξιλογίου

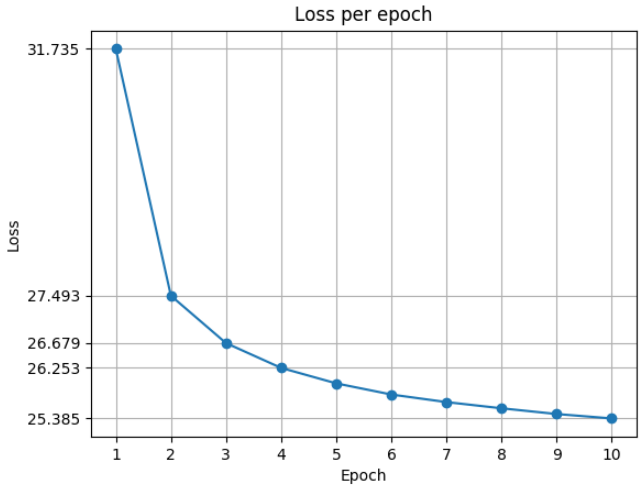
Πίνακας 5.1: Πίνακας μετρικών Rouge για το μοντέλο με απεριόριστο μέγεθος λεξιλογίου

Εποχή	Rouge 1 F1	Rouge 2 F1	Rouge L F1
1	0,42813	0,20802	0,40106
2	0,42725	0,21243	0,40141
3	0,43533	0,21721	0,40936
4	0,43704	0,22035	0,40984
5	0,43008	0,21003	0,40340
6	**0,43992**	**0,22155**	**0,41226**
7	0,43554	0,22188	0,40945
8	0,43411	0,21739	0,40806
9	0,43816	0,21904	0,40868
10	0,43895	0,22129	0,41157

Όπως μπορεί να φανεί παρόλο που η συνάρτηση κόστους μειώνεται συνεχόμενα, για όλες τις εποχές εκπαίδευσης, οι μετρικές Rouge λαμβάνουν τις καλύτερες τιμές τους για το σύνολο ελέγχου στην έκτη εποχή και συνεπώς αυτή θα χρησιμοποιηθεί στην συνέχεια για την σύγκριση των μοντέλων. Φυσικά, δεν γίνεται να εγγυηθούμε πως το συνολικά καλύτερο μοντέλο για όλα τα δεδομένα ελέγχου είναι αυτό που επιλέχθηκε από το μικρό δείγμα στο οποίο πραγματοποιήσαμε μετρήσεις, αλλά αποτελεί μια αρκετά ασφαλή εκτίμηση της απόδοσης για όλες τις εποχές.

5.3.2 Μοντέλο με μέγεθος κρυφής κατάστασης κωδικοποιητή 128

Το συγκεκριμένο μοντέλο μελετήθηκε στα πλαίσια διερεύνησης της επίδρασης της κρυφής κατάστασης στην συνολική επίδοση του μηχανισμού που θα παρουσιαστεί στην συνέχεια. Το μέγεθος που αποκωδικοποιητή σε αυτή την περίπτωση, ισούται με το διπλάσιο του κωδικοποιητή, δηλαδή με 256.



Σχήμα 5.2: Μέση τιμής της συνάρτησης κόστους για το μοντέλο με μέγεθος κρυφής κατάστασης 128

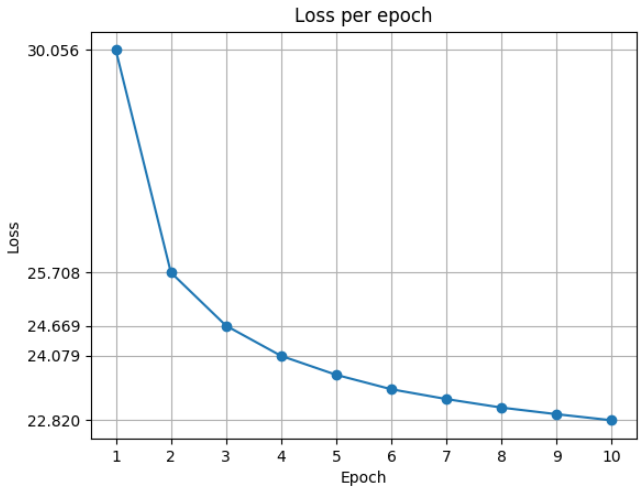
Πίνακας 5.2: Πίνακας μετρικών Rouge για το μοντέλο με μέγεθος κρυφής κατάστασης κωδικοποιητή 128

Εποχή	Rouge 1 F1	Rouge 2 F1	Rouge L F1
1	0,41029	0,19932	0,38821
2	0,41996	0,20831	0,39596
3	0,41611	0,20508	0,39211
4	0,42277	0,21015	0,39971
5	0,42684	0,21234	0,40281
6	0,43136	0,21534	0,40555
7	0,42592	0,21199	0,39979
8	0,42901	0,21637	0,40436
9	**0,43197**	**0,21559**	**0,40555**
10	0,42648	0,20965	0,40118

Σε αυτή την περίπτωση φαίνεται πως τα συνολικά καλύτερα αποτελέσματα για το σύνολο ελέγχου υπολογίστηκαν στην ένατη εποχή και συνεπώς αυτό το μοντέλο θα χρησιμοποιηθεί στην συνέχεια για τις περαιτέρω συγκρίσεις.

5.3.3 Μοντέλο με μέγεθος κρυφής κατάστασης κωδικοποιητή 256

Το μοντέλο αυτό μελετήθηκε όπως και το προηγούμενο, στα πλαίσια διερεύνησης της επίδρασης της κρυφής κατάστασης στην συνολική επίδοση του μηχανισμού. Το μέγεθος που αποκωδικοποιητή σε αυτή την περίπτωση, ισούται πάλι με το διπλάσιο του κωδικοποιητή, δηλαδή με 512.



Σχήμα 5.3: Μέση τιμής της συνάρτησης κόστους για το μοντέλο με μέγεθος κρυφής κατάστασης 256

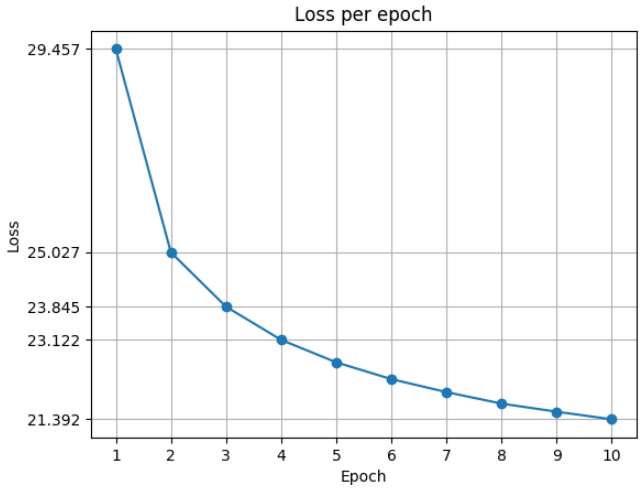
Πίνακας 5.3: Πίνακας μετρικών Rouge για το μοντέλο με μέγεθος κρυφής κατάστασης κωδικοποιητή 256

Εποχή	Rouge 1 F1	Rouge 2 F1	Rouge L F1
1	0,43126	0,21033	0,40602
2	0,43493	0,21526	0,40826
3	0,43276	0,21352	0,40665
4	0,43293	0,21485	0,40653
5	0,43679	0,21774	0,41070
6	0,43009	0,21342	0,40356
7	0,43281	0,21457	0,40450
8	**0,43871**	**0,22045**	**0,41209**
9	0,43748	0,21633	0,41139
10	0,43618	0,21547	0,41011

Παρατηρούμε ότι σε αυτή την περίπτωση τα βέλτιστα αποτελέσματα συλλέχθηκαν την 8η εποχή και συνεπώς το μοντέλο της 8ης εποχής επιλέγεται ως το βέλτιστο που θα συγκριθεί στην συνέχεια με τα υπόλοιπα. Όσον αφορά την τιμή της συνάρτησης κόστους, και σε αυτή την περίπτωση βλέπουμε ότι πέφτει ομαλά σε κάθε εποχή όπως είναι και το αναμενόμενο.

5.3.4 Μοντέλο με μέγεθος κρυφής κατάστασης κωδικοποιητή 384

Το μοντέλο αυτό, μελετήθηκε επίσης στα πλαίσια διερεύνησης του μεγέθους κρυφής κατάστασης στην επίδοση του συστήματος. Σε αντίθεση όμως με τα υπόλοιπα μοντέλα της ίδιας κατηγορίας, το συγκεκριμένο μοντέλο έχει επιλεχθεί ως μοντέλο αναφοράς, και θα χρησιμοποιηθεί ως κοινός παρονομαστής σε πολλαπλές συγκρίσεις στην συνέχεια, έτσι ώστε να υπάρχει ένας κοινός παράγοντας σε όλες τις μετέπειτα συγκρίσεις ώστε να παρουσιάζονται καλύτερα οι σχέσεις μεταξύ των διαφόρων μοντέλων. Το μέγεθος κρυφής κατάστασης του αποκωδικοποιητή σε αυτό το μοντέλο είναι πάλι ίσο με το διπλάσιο του μεγέθους της κρυφής κατάστασης του κωδικοποιητή, δηλαδή 768.



Σχήμα 5.4: Μέση τιμής της συνάρτησης κόστους για το μοντέλο με μέγεθος κρυφής κατάστασης 384

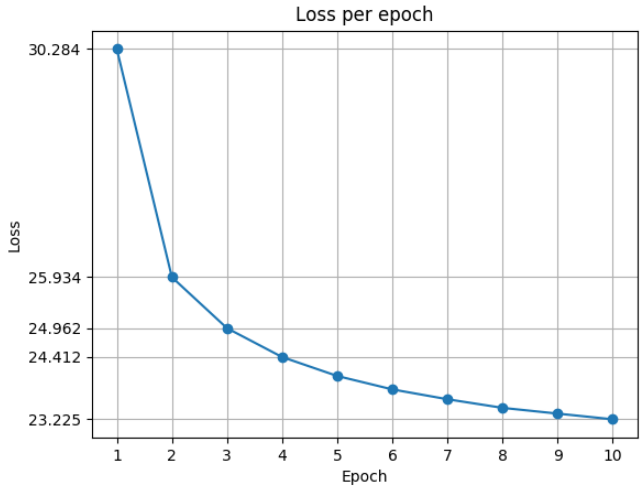
Πίνακας 5.4: Πίνακας μετρικών Rouge για το μοντέλο με μέγεθος κρυφής κατάστασης κωδικοποιητή 384

Εποχή	Rouge 1 F1	Rouge 2 F1	Rouge L F1
1	0,43012	0,21190	0,40499
2	0,43452	0,21814	0,40726
3	0,43316	0,21845	0,40701
4	0,43461	0,21533	0,40653
5	0,44150	0,21691	0,41256
6	0,43281	0,21900	0,40704
7	0,43588	0,22092	0,41014
8	0,43598	0,21333	0,40668
9	**0,44429**	**0,22455**	**0,41730**
10	0,43598	0,21333	0,40668

Παρατηρείται και σε αυτή την περίπτωση ότι η τιμή της συνάρτησης κόστους μειώνεται ομαλά με την πάροδο των εποχών. Από τις μετρικές Rouge φαίνεται πως το καλύτερο μοντέλο σύμφωνα με το σύνολο δεδομένων ελέγχου είναι αυτό της εποχής 9 και επιλέγεται για να χρησιμοποιηθεί στην συνέχεια.

5.3.5 Μοντέλο με μειωμένο μέγεθος αποκωδικοποιητή

Στο συγκεκριμένο μοντέλο χρησιμοποιείται η τεχνική για μείωση του μεγέθους της κρυφής κατάστασης του αποκωδικοποιητή που είχε αναλυθεί σε προηγούμενο κεφάλαιο ως μια προσπάθεια διερεύνησης της απόδοσης της. Ουσιαστικά, σε αυτή την περίπτωση, η έξοδος του αμφίδρομου κωδικοποιητή συνδέεται με ένα πυκνό στρώμα που έχει ως στόχο να ελαττώσει το μέγεθος των δεδομένων εξόδου, ώστε να μειωθεί εν τέλει το μέγεθος του αποκωδικοποιητή που θα τα λάβει ως είσοδο και με στόχο του ταχύτερους χρόνου εκπαίδευσης. Σε αυτό το μοντέλο το μέγεθος κρυφής κατάστασης του κωδικοποιητή που χρησιμοποιείται είναι 384 και ισούται με το μέγεθος κρυφής κατάστασης του αποκωδικοποιητή.



Σχήμα 5.5: Μέση τιμής της συνάρτησης κόστους για το μοντέλο με μειωμένο μέγεθος κρυφής κατάστασης αποκωδικοποιητή

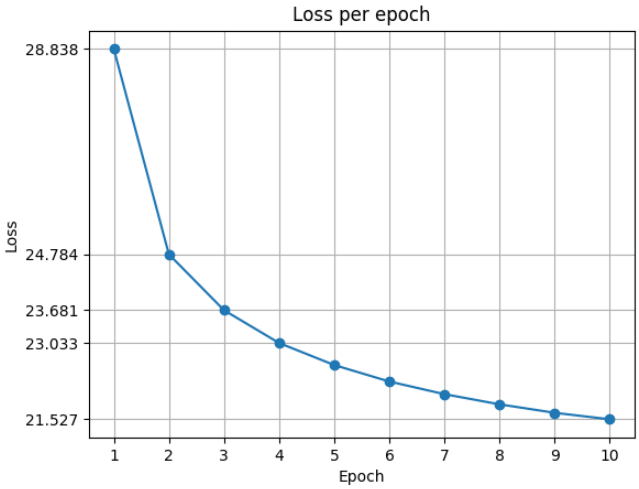
Πίνακας 5.5: Πίνακας μετρικών Rouge για το μοντέλο με μειωμένο μέγεθος κρυφής κατάστασης αποκωδικοποιητή

Εποχή	Rouge 1 F1	Rouge 2 F1	Rouge L F1
1	0,42664	0,20809	0,40037
2	**0,44030**	**0,22159**	**0,41424**
3	0,43676	0,22023	0,41150
4	0,43768	0,22125	0,41184
5	0,43260	0,21632	0,40768
6	0,43121	0,21574	0,40572
7	0,43776	0,21747	0,41124
8	0,43335	0,21317	0,40526
9	0,43180	0,21539	0,41523
10	0,43650	0,21791	0,40934

Για άλλη μια φορά παρατηρούμε ότι η συνάρτηση κόστους μειώνεται σταδιακά ανά εποχή και συνεπώς το μοντέλο έχει εκπαιδευτεί ορθά. Από τις μετρικές Rouge παρατηρούμε ότι στην δεύτερη εποχή παρατηρούνται τα καλύτερα αποτελέσματα και συνεπώς αυτό το μοντέλο επιλέγεται να χρησιμοποιηθεί στην συνέχεια.

5.3.6 Μοντέλο με ανεπεξέργαστα δεδομένα εκπαίδευσης

Το συγκεκριμένο μοντέλο εκπαιδεύτηκε για να χρησιμοποιηθεί στην συνέχεια ως ένας τρόπος μέτρησης της επίδοσης της διαδικασίας προεπεξεργασίας των δεδομένων που χρησιμοποιήθηκε σε αυτή την εργασία. Σε αυτό το μοντέλο, τα δεδομένα εκπαίδευσης δεν έχουν υποστεί καμία περαιτέρω επεξεργασία και έχουν χρησιμοποιηθεί ακριβώς στην μορφή που ορίσθηκαν και περιγράφηκαν για πρώτη φορά στο [Rush15]. Το μέγεθος κρυφής κατάστασης του κωδικοποιητή σε αυτή την περίπτωση είναι 384 και του αποκωδικοποιητή το διπλάσιο. Σημειώνεται πως σε αυτή την περίπτωση το σύνολο δεδομένων ελέγχου που χρησιμοποιήθηκε για την μέτρηση της επίδοσης του μοντέλου ανά εποχή είναι διαφορετικό από αυτό που είχε χρησιμοποιηθεί στα προηγούμενα μοντέλα λόγω των διαφορών που υπάρχουν στα σύνολα δεδομένων εκπαίδευσης των άλλων μοντέλων σε σχέση με αυτό.



Σχήμα 5.6: Μέση τιμής της συνάρτησης κόστους για το μοντέλο με ανεπεξέργαστα δεδομένα εκπαίδευσης

Πίνακας 5.6: Πίνακας μετρικών Rouge για το μοντέλο με ανεπεξέργαστα δεδομένα εκπαίδευσης

Εποχή	Rouge 1 F1	Rouge 2 F1	Rouge L F1
1	0,38978	0,18549	0,36831
2	0,39908	0,19368	0,37727
3	0,39685	0,18781	0,37239
4	0,40769	0,19825	0,38421
5	0,40872	0,19652	0,38422
6	**0,40815**	**0,20008**	**0,38469**
7	0,40319	0,19740	0,38000
8	0,40576	0,19823	0,38126
9	0,40192	0,19400	0,37836
10	0,40427	0,19833	0,38137

Όπως παρατηρείται, παρά του αυξημένου θορύβου από τα μη επεξεργασμένα δεδομένα εκπαίδευσης, το μοντέλο εκπαιδεύεται κανονικά και η τιμή της συνάρτησης κόστους μειώνεται αρμονικά με το πέρασμα των εποχών. Από τις μετρικές Rouge φαίνεται πως το καλύτερο μοντέλο παρουσιάζεται στην 6η εποχή και συνεπώς θα χρησιμοποιηθεί αυτό στην συνέχεια.

5.3.7 Συνοπτική παρουσίαση επιλεγμένων μοντέλων

Σε αυτό το σημείο θα παρουσιαστεί ένας συνολικός πίνακας που θα παρουσιάζει συνολικά τα υλοποιημένα μοντέλα καθώς και τις μετρικές Rouge για την εκάστοτε καλύτερη εκδοχή που επιλέχθηκε παραπάνω.

Πίνακας 5.7: Αποτελέσματα Αξιολόγησης Δέντρων Αποφάσεων

Μοντέλα	M1	M2	M3	M4	M5	M6
encoder hidden size	384	128	256	384	384	384
decoder hidden size	768	256	512	768	384	768
vocabulary size (*1000)	103	50	50	50	50	50
data preprocess	yes	yes	yes	yes	yes	no
best epoch chosen	6	9	8	9	2	6
Rouge 1 F1 best epoch	0,43992	0,43197	0,43871	0,44429	0,44030	0,40815
Rouge 2 F1 best epoch	0,22155	0,21559	0,22045	0,22455	0,22159	0,20008
Rouge L F1 best epoch	0,41226	0,40555	0,41209	0,41730	0,41424	0,38469

5.4 Διερεύνηση βέλτιστων παραμέτρων

Σε αυτή την ενότητα θα πραγματοποιηθούν συγκρίσεις μεταξύ των καλύτερων μοντέλων που επιλέχθηκαν πριν ώστε να γίνει εμφανές πως επηρεάζουν διάφορες παράμετροι την λειτουργία του μηχανισμού αυτόματης παραγωγής περίληψης. Στόχος αυτής της διερεύνησης αποτελεί η εξακρίβωση της συμπεριφοράς του μηχανισμού πάνω στο συγκεκριμένο πρόβλημα, ώστε να μπορεί να συστηματοποιηθεί καλύτερα, μετέπειτα έρευνα που μπορεί να πραγματοποιηθεί στο αντικείμενο.

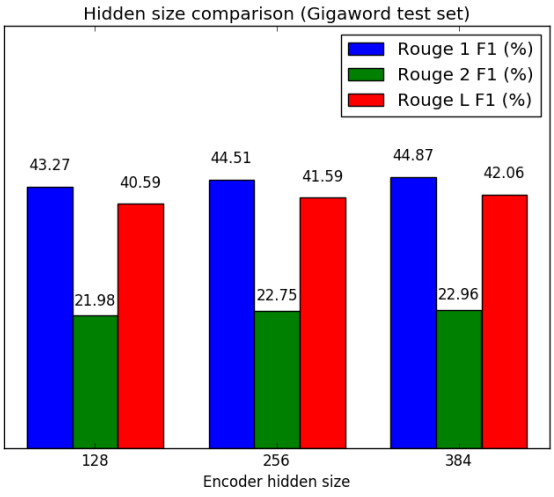
Για να πραγματοποιηθεί ορθά αυτή η διερεύνηση, τα καλύτερα μοντέλα που επιλέχθηκαν στην προηγούμενη ενότητα θα συγκριθούν μεταξύ τους, μέσω της χρήσης δύο διαφορετικών συνόλων δεδομένων. Το πρώτο σύνολο δεδομένων αποτελείται από 2000 τυχαίες αντιστοιχίες άρθων - περιλήψεων αρχικού συνόλου ελέγχου gigaword, ενώ το δεύτερο σύνολο δεδομένων αποτελείται από 500 αντιστοιχίες άρθων - περιλήψεων του συνόλου DUC. Γι αυτές τις αντιστοιχίες θα συλλεχθούν μετρικές σύγκρισης από τους αλγόριθμους Rouge, και θα χρησιμοποιηθούν για να κριθεί η επίδοση των διάφορων μοντέλων. Σε κάθε σύγκριση, θα παρουσιάζονται κάθε φορά και οι αντίστοιχοι χρόνοι εκπαίδευσης των μοντέλων. Παρόλο που δεν θα δοθεί πολύ σημασία στους χρόνους εκπαίδευσης για την επιλογή των καλύτερων μοντέλων, αυτοί θα σχολιάζονται κάθε φορά, καθώς σε καταστάσεις έλλειψης πόρων ή χρόνου, πρέπει να συνυπολογιστούν από τον εκάστοτε ερευνητή>

Στα πλαίσια αυτής της διερεύνηση θα πραγματοποιηθούν τέσσερις διαφορετικές συγκρίσεις, οι οποίες θα παρουσιαστούν στην συνέχεια μαζί με τα αποτελέσματά τους. Στις συγκρίσεις αυτές, λόγω του χρονικού κόστους της εκπαίδευσης, θα παρουσιαστούν λίγα σημεία σύγκρισης, αυτά που αποφασίστηκαν ότι περιέχουν το περισσότερο ενδιαφέρον κατά τον πειραματισμό με τον μηχανισμό.

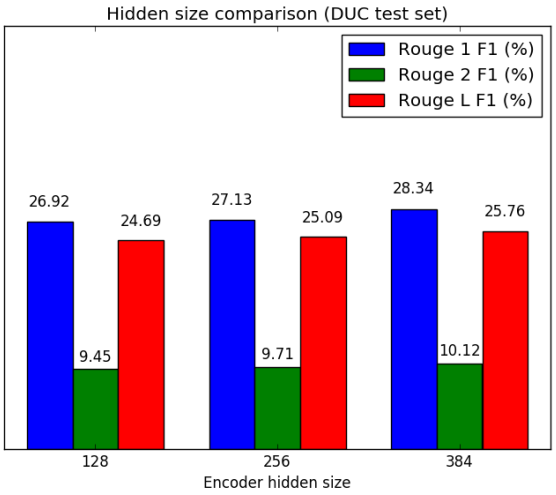
5.4.1 Διερεύνηση μεγέθους κρυφής κατάστασης

Μια πολύ σημαντική παράμετρος του μηχανισμού κωδικοποιητή - αποκωδικοποιητή για την αυτόματη παραγωγή περίληψης, αποτελεί το μέγεθος κρυφής κατάστασης του κωδικοποιητη και του αποκωδικοποιητή. Αυτή η παράμετρος καθορίζει ουσιαστικά το μέγεθος του μηχανισμού και μπορεί να επηρεάσει σε πολύ σημαντικό βαθμό την απόδοσή του, γι αυτό η επιλογή του μεγέθους της πρέπει να πραγματοποιηθεί προσεκτικά. Πολύ μικρές τιμές μεγέθους κρυφής κατάστασης μπορεί να αποτρέψουν τον μηχανισμό να συγκλίνει σε ορθά και γενικευμένα αποτελέσματα. Αντίθετα, πολύ μεγάλες τιμές μεγέθους κρυφής μνήμης μπορεί να αυξήσουν δραματικά τις απαιτήσεις χρόνου και μνήμης για την εκπαίδευση, ενώ μπορούν να οδηγήσουν και σε φαινόμενα υπερεκπαίδευσης μειώνονται πάλι την δυνατότητα γενίκευσης. Στην συνέχεια παρουσιάζονται οι μετρικές Rouge για μεγέθη 128, 256, 384 κρυφής κατάστασης κωδικοποιητή, ενώ η κατάσταση του αποκωδικοποιητή ορίζεται ως το διπλάσιο της κρυφής κατάστασης του κωδικοποιητή σε όλες τις περιπτώσεις.

Σχήμα 5.7: Σύγκριση μεγέθους κρυφής κατάστασης κωδικοποιητή για το σύνολο ελέγχου Gigaword



Σχήμα 5.8: Σύγκριση μεγέθους κρυφής κατάστασης κωδικοποιητή για το σύνολο ελέγχου DUC



Πίνακας 5.8: Χρόνοι εκπαίδευσης ανά μέγεθος κρυφής κατάστασης

encoder hidden size	1000 batches train time (s)
128	138
256	164
384	210

Όπως μπορεί να φανεί από τις γραφικές παραστάσεις, το μοντέλο με μέγεθος κατάστασης 384 σύλλεξε τα μεγαλύτερα σκορ για όλους τους αλγορίθμους Rouge και στα δύο σύνολα ελέγχου. Αυτό σημαίνει πως το πρόβλημα που προσπαθεί να επιλυθεί από τον μηχανισμό είναι πολύ σύνθετο και δεν μπορεί να αναπαρασταθεί καλά σε μικρότερα μεγέθη κρυφής κατάστασης. Φυσικά μεγαλύτερα μεγέθη κρυφής κατάστασης μπορεί να οδηγήσουν και σε καλύτερα ακόμα αποτελέσματα, αλλά όπως μπορεί να παρατηρηθεί κατά την μετάβαση από μέγεθος κρυφής κατάστασης 256 σε μέγεθος 384, το σκορ ανεβαίνει ελάχιστα. Μεταβάσεις σε μεγαλύτερα μεγέθη κρυφής κατάστασης θα έχουν πολύ μικρή επίπτωση στην επίδοση, ενώ θα καταστήσουν τους χρόνους εκπαίδευσης απαγορευτικούς.

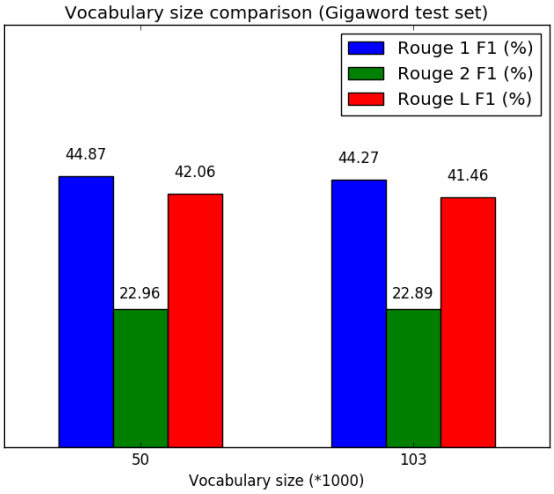
5.4.2 Διερεύνηση επιρροής μεγέθους λεξιλογίου

Μια άλλη πολύ σημαντική παράμετρος στους μηχανισμούς αυτόματης παραγωγής περίληψης αποτελεί το μέγεθος του λεξιλογίου κωδικοποιητή και αποκωδικοποιητή. Το μέγεθος του λεξιλογίου καθορίζει πόσες λέξεις γνωρίζει και μπορεί να αναγνωρίζει ο κωδικοποιητής και πόσες λέγεις γνωρίζει και μπορεί να παράξει ο αποκωδικοποιητής. Καθώς ο αποκωδικοποιητής περιέχει μια έξοδο για κάθε λέξη του λεξιλογίου, η ορθή επιλογή του μεγέθους αυτού είναι πολύ σημαντική καθώς μια πολύ μεγάλη επιλογή μπορεί να οδηγήσει σε τεράστιες υπολογιστικές απαιτήσεις.

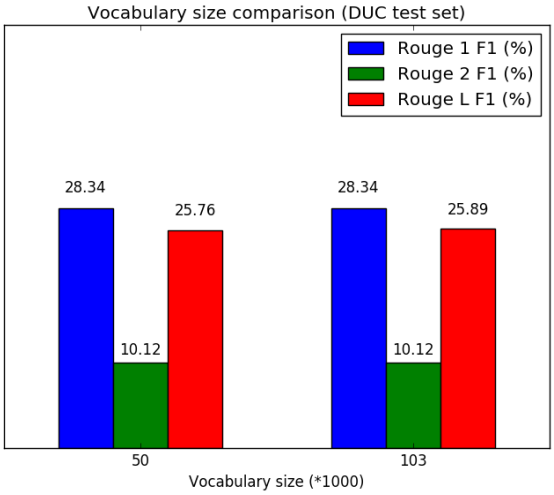
Αν κάποιος σκεφτεί την ουσιαστική επίπτωση που έχει το λεξιλόγιο στην αύξηση της ανθεκτικότητας του μηχανισμού μέσω της αναγνώρισης και του χειρισμού πολλαπλών λέξεων, μπορεί να υποθέσει πως το μεγαλύτερο μέγεθος του λεξιλογίου οδηγεί σε καλύτερα αποτελέσματα. Αυτό που πρέπει να αναλογιστεί όμως, είναι πως όσο μεγαλώνει το μέγεθος του λεξιλογίου τόσο μεγαλώνουν και οι πιθανές αποφάσεις του αποκωδικοποιητή αυξάνοντας έτσι την πιθανότητα αυτός να προβλέψει λανθασμένη έξοδο. Η μετέπειτα σύγκριση θα πραγματοποιηθεί μεγέθη λεξιλογίου (κωδικοποιητή και αποκωδικοποιητή) 50000 καθώς και 103000 που αντιστοιχεί στο πλήθος των συνολικών λέξεων που εμφανίστηκαν στα σύνολα εκπαίδευσης. Για το λεξιλόγιο με μέγεθος 50000 επιλέχθηκαν οι 50000 πιο συχνά εμφανιζόμενες λέξεις που εμφανίστηκαν πάλι στα σύνολα εκπαίδευσης. Τα μοντέλα που

συγκρίνονται στην συνέχεια έχουν μεγέθη κρυφής κατάστασης 384/768 για τον κωδικοποιητή και τον αποκωδικοποιητή αντίστοιχα.

Σχήμα 5.9: Σύγκριση μεγέθους λεξιλογίου για το σύνολο ελέγχου Gigaword



Σχήμα 5.10: Σύγκριση μεγέθους λεξιλογίου για το σύνολο ελέγχου DUC



Πίνακας 5.9: Χρόνοι εκπαίδευσης ανά μέγεθος λεξιλογίου

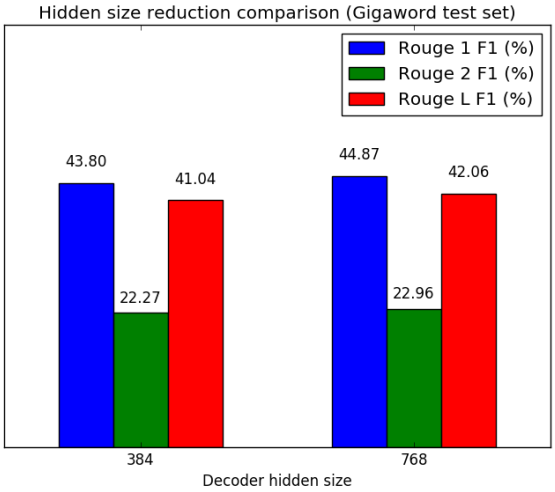
Vocabulary size (*1000)	1000 batches train time (s)
50	210
103	289

Όπως μπορεί να φανεί από τις γραφικές παραστάσεις, το μοντέλο με μέγεθος λεξιλογίου 50000 αποφέρει ελάχιστα καλύτερα αποτελέσματα από το μοντέλο με το πλήρες λεξιλόγιο. Παρόλο που η διαφορά στα σκορ δεν είναι σημαντική, αν κοιτάξει κανείς του χρόνους εκπαίδευσης, μπορεί να παρατηρήσει ότι το μοντέλο με το πλήρες λεξιλόγιο έχει πολύ μεγαλύτερο χρονικό κόστος ενώ αποφέρει ουσιαστικά τα ίδια αποτελέσματα. Ενάντια στην πρώτη αντίληψη επομένως, ένα μικρότερο μέγεθος λεξιλογίου μπορεί να αποδώσει και καλύτερα από ένα μεγαλύτερο μέγεθος λεξιλογίου και πιο αποδοτικά.

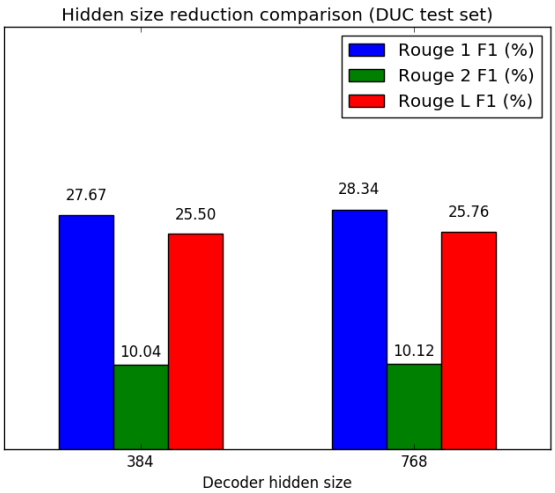
5.4.3 Διερεύνηση επιρροής μείωσης της διάστασης της κρυφής κατάστασης

Σε αυτή την ενότητα θα διερευνηθεί η απόδοση του μηχανισμού για μείωση της κρυφής κατάστασης του αποκωδικοποιητή στο μέγεθος της κρυφής κατάστασης του κωδικοποιητή. Όπως έχει αναλυθεί και σε προηγούμενο κεφάλαιο, ο μηχανισμός αυτός προσθέτει ένα επιπλέον πυκνό στρώμα ανάμεσα στον κωδικοποιητή και τον αποκωδικοποιητή, με στόχο να ρίξει την διάσταση του. Ο μηχανισμός αυτός φυσικά, μειώνει πάρα πολύ το μέγεθος του δικτύου οδηγώντας σε καλύτερους χρόνους εκπαίδευσης και σύγκλισης. Στην παρακάτω σύγκριση, και τα δύο μοντέλα που συγκρίνονται έχουν μέγεθος κρυφής κατάστασης κωδικοποιητή 384.

Σχήμα 5.11: Διερεύνηση επιρροής μείωσης της κρυφής κατάστασης για το σύνολο ελέγχου Gigaword



Σχήμα 5.12: Διερεύνηση επιρροής μείωσης της κρυφής κατάστασης για το σύνολο ελέγχου DUC



Πίνακας 5.10: Χρόνοι εκπαίδευσης μέγεθος κρυφής κατάστασης αποκωδικοποιητή

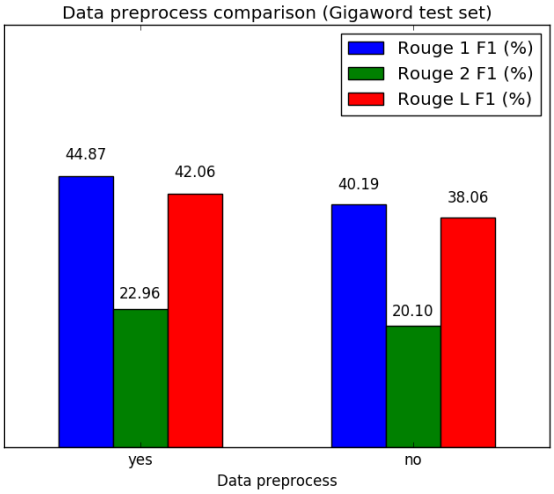
Decoder hidden size	1000 batches train time (s)
384	155
768	210

Όπως μπορεί να παρατηρηθεί από τις γραφικές παραστάσεις, το μοντέλο χωρίς τον μηχανισμό αποδίδει καλύτερα απ' ότι αυτό με τον μηχανισμό μείωσης της κρυφής κατάστασης. Αυτό το φαινόμενο είναι εν μέρη λογικό, καθώς το πυκνό στρώμα ανάμεσα στον κωδικοποιητή και τον αποκωδικοποιητή αλλοιώνει την πληροφορία που μεταφέρεται ανάμεσα τους. Η δύναμη αυτού του μηχανισμού μπορεί να φανεί αν κοιτάξει κανείς τους χρόνους εκπαίδευσης και σύγκλισης. Αν κοιτάξει κανείς τα αποτελέσματα της διερεύνησης των βέλτιστων μοντέλων που πραγματοποιήθηκε, μπορεί να δει πως η βέλτιστη εκδοχή του μοντέλου χωρίς τον μηχανισμό, προέκυψε στην 9η εποχή, ενώ η βέλτιστη εκδοχή του μοντέλου με τον μηχανισμό προέκυψε στην 2η εποχή. Ακόμη, μπορεί να φανεί και από τους χρόνους εκπαίδευσης, πως το μοντέλο με τον μηχανισμό εκπαιδεύεται πολύ πιο γρήγορα από το ανταγωνιστικό του. Συνεπώς ο μηχανισμός ελάττωσης της κρυφής κατάστασης μπορεί να αποβεί πολύ σημαντικός για τον σχεδιασμό μοντέλων σε συνθήκες περιορισμένων πόρων και χρόνου.

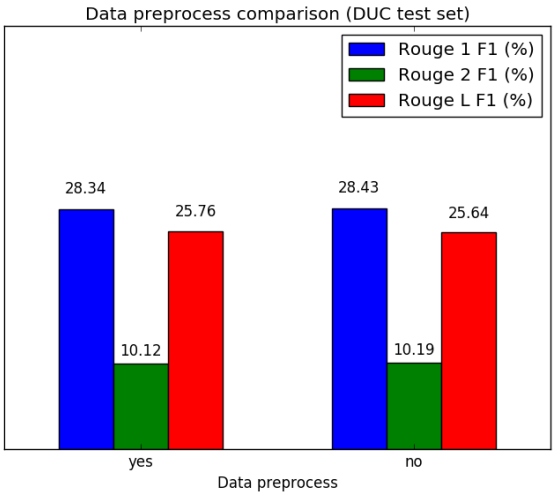
5.4.4 Διερεύνηση επίδοσης της προεπεξεργασίας δεδομένων

Όπως έχει αναλυθεί και πιο πριν, τα σύνολα δεδομένων που χρησιμοποιήθηκαν πέρασαν πρώτα από μια διαδικασία προεπεξεργασίας. Στόχος αυτής της προεπεξεργασίας αποτέλεσε η προσπάθεια αφαίρεσης του "θόρυβου" από τα σύνολα δεδομένων ώστε κάθε λέξη που παράγεται από τον αποκωδικοποιητή να είναι όσο πιο ουσιαστική γίνεται. Η συγκεκριμένη διερεύνηση χρησιμοποιεί δύο μοντέλα, το ένα εκ των οποίων χρησιμοποιεί προεπεξεργασμένα δεδομένα εκπαίδευσης και ελέγχου ενώ το άλλο χρησιμοποιεί τα αυθεντικά, χωρίς καμία προπεπεξεργασία στα πλαίσια αυτής της εργασίας. Στόχος της διερεύνησης αποτελεί ο μαθηματικός υπολογισμός της βελτίωσης που επιτυγχάνεται από την διαδικασία της προεπεξεργασίας. Τα μοντέλα που θα αναλυθούν στην συνέχεια έχουν μεγέθη κρυφής κατάστασης 384/768 για τον κωδικοποιητή και τον αποκωδικοποιητή και μέγεθος λεξιλογιου 50000. Σημειώνεται πως για το μοντέλο χωρίς προεπεξεργασία, οι υπολογισμοί από τους αλγόριθμους Rouge έγιναν σε διαφορετικό σύνολο δεδομένων μεγέθους 2000 στοιχείων Gigaword λόγω των έντονων διαφορών της προεπεξεργασίας που απαιτούσαν την δημιουργία νέου συνόλου από τα ανεπεξέργαστα δεδομένα.

Σχήμα 5.13: Διερεύνηση επίδρασης της προεπεξεργασίας των δεδομένων για το σύνολο ελέγχου Gigaword



Σχήμα 5.14: Διερεύνηση επίδρασης της προεπεξεργασίας των δεδομένων για το σύνολο ελέγχου DUC



Όπως μπορεί να φανεί από τις γραφικές παραστάσεις, η διαδικασία της προεπεξεργασίας επιδράει πάρα κατά την χρήση του συνόλου δεδομένων Gigaword καθώς αυτό περιέχει μέσα πολύ θόρυβο λόγω της αυτοματοποιημένης διαδικασίας παραγωγής του και του τεράστιου όγκου του. Κατά την χρήση του συνόλου δεδομένων DUC οι διαφορές στην επίδοση είναι ελάχιστες καθώς αυτό το σύνολο δεδομένων είναι πολύ πιο καθαρό και ουσιαστικό. Οι 4 μονάδες διαφορά που εμφανίζονται κατά την χρήση του συνόλου Gigaword είναι πάρα πολύ σημαντικές και δείχνουν την αξία την προεπεξεργασίας στην προσπάθεια δημιουργίας ενός ανθεκτικού συστήματος παραγωγής περίληψης.

Σε αυτή την διερεύνηση, καθώς οι αρχιτεκτονικές των δύο μοντέλων είναι ίδιες δεν παρουσιάστηκαν ούτε σχολιάστηκαν οι χρόνοι της εκπαίδευσης καθώς δεν μπορεί να προκύψει κανένα πρακτικό συμπέρασμα από αυτούς.

5.5 Μηχανισμός χειρισμού άγνωστων λέξεων

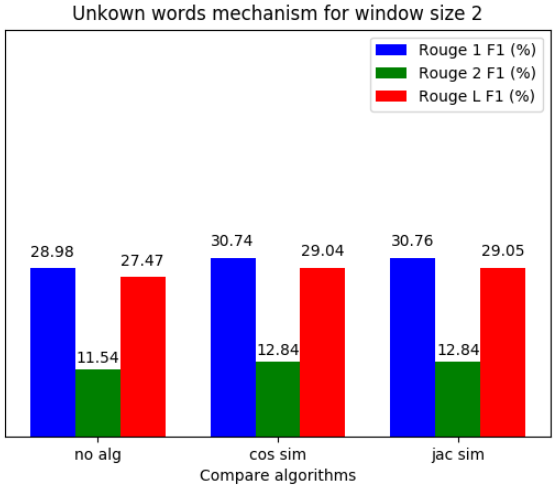
Σε αυτό το σημείο της εργασίας θα αναλυθεί η επίδοση του μηχανισμού χειρισμού άγνωστων λέξεων. Όπως έχει παρουσιαστεί και σε προηγούμενο κεφάλαιο ο μηχανισμός αυτός προσπαθεί να αντιμετωπίσει το εγγενές πρόβλημα που εμφανίζεται σε συστήματα παραγωγής περίληψης περιορισμένου λεξιλογίου για χειρισμό των λέξεων που δεν εμφανίζονται στο λεξιλόγιο. Η αντιμετώπιση αυτού του προβλήματος είναι πολύ σημαντική για την επίτευξη ενός ισχυρού καθολικού συστήματος παραγωγής περίληψης, καθώς αυτό το σύστημα θα πρέπει να είναι σε θέση να αντιμετωπίζει διαφορετικές κατηγορίες κειμένων, με διαφορετικά λεξιλόγια.

Όπως είχε αναλυθεί πιο πριν στην περιγραφή του μηχανισμού αυτού (4), η λειτουργία του βασίζεται στην χρήση ολισθαίνοντος παραθύρου γύρω από τις περιοχές που εμφανίζονται άγνωστες λέξεις στο κείμενο και στην εκάστοτε προβλεπόμενη περίληψη. Στο εκάστοτε ολισθαίνον παράθυρο εφαρμόζεται κάποιος αλγόριθμος ομοιότητας (cosine similarity, jaccard similarity) ώστε να βρεθεί πιο παράθυρο του αρχικού κειμένου μοιάζει πιο πολύ με το παράθυρο της εκάστοτε άγνωστης λέξης στην περίληψη. Η λέξη του κειμένου που αντικαθίσταται τελικά στην περίληψη είναι εκείνη της οποίας το παράθυρο λέξεων γύρω της έχει την μεγαλύτερη ομοιότητα με το παράθυρο της άγνωστης λέξης στην περίληψη.

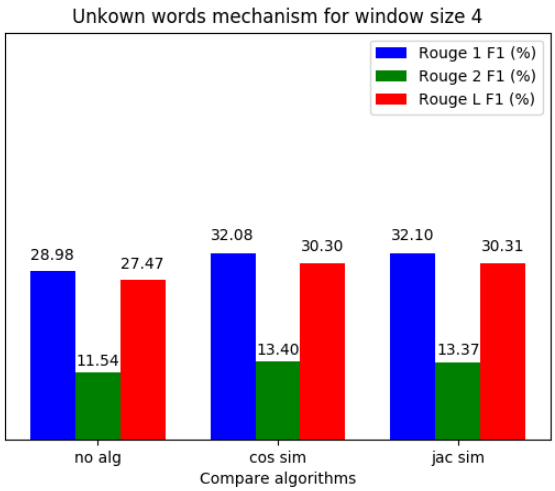
Για να αναλυθεί λοιπόν και να φανεί καλύτερα η επίδραση του μηχανισμού χειρισμού άγνωστων λέξεων, θα μετρηθεί η μετρική Rouge, στις παραγόμενες περιλήψεις του συνόλου ελέγχου Gigaword που περιέχουν άγνωστες λέξεις (δηλαδή την λέξη unk). Επιλέγεται να μελετηθούν μόνο οι περιλήψεις που περιέχουν άγνωστες λέξεις ώστε να φανεί καλύτερα η βαρύτητα του μηχανισμού σε αυτές, χωρίς να αλλοιώνεται η μετρική από τις υπόλοιπες περιλήψεις. Γι' αυτό το σκοπό θα χρησιμοποιηθεί το μοντέλο που πέτυχε τα καλύτερα αποτελέσματα στην προηγούμενη ανάλυση, δηλαδή αυτό με μεγέθη κρυφής κατάστασης 384/768 για τον κωδικοποιητή και τον αποκωδικοποιητή και μέγεθος λεξιλογίου 50000. Έτσι συγκρίνοντας την μετρική σε αυτές τις περιλήψεις πριν και μετά την εφαρμογή του μηχανισμού θα φανεί η αξία και η σημασία του.

Στην ανάλυση που ακολουθεί θα παρουσιαστούν γραφήματα για μεγέθη παραθύρου 2, 4, 6 και για τους δύο αλγόριθμους ομοιότητας (Jaccard similarity, Cosine similarity), από τα οποία θα προκύψουν και τα τελικά συμπεράσματα.

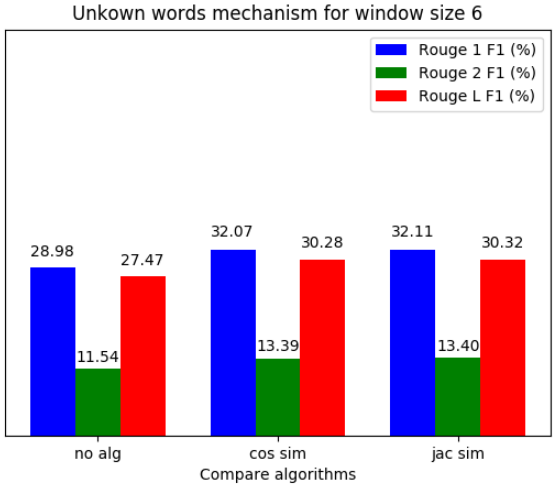
Σχήμα 5.15: Αποτελέσματα μηχανισμού χειρισμού άγνωστων λέξεων για μέγεθος παραθύρου 2



Σχήμα 5.16: Αποτελέσματα μηχανισμού χειρισμού άγνωστων λέξεων για μέγεθος παραθύρου 4



Σχήμα 5.17: Αποτελέσματα μηχανισμού χειρισμού άγνωστων λέξεων για μέγεθος παραθύρου 6



Όπως μπορεί να φανεί μετά το μέγεθος παραθύρου 2 τα αποτελέσματα σταθεροποιούνται και είναι σχεδόν ίδια. Αξίζει να παρατηρηθεί ότι και οι δύο μηχανισμοί βελτιώνουν τα αποτελέσματα κατά 3 μονάδες Rouge για παράθυρα μεγαλύτερα του 2, πράγμα το οποίο είναι πολύ σημαντικό και δείχνει την αξία αυτού.

Μπορεί να παρατηρήσει επίσης κάποιος πως τα αποτελέσματα των μετρικών είναι αισθητά μικρότερα από τα προηγούμενα που παρουσιάστηκαν. Αυτό συμβαίνει γιατί στις περιλήψεις που περιέχουν άγνωστες λέξεις (unk) το μοντέλο προσπαθεί να μαντέψει επόμενες λέξεις συνυπολογίζοντας τις άγνωστες και αποκλίνει έτσι από τα επιθυμητά νοήματα καθώς η αναπαράσταση που έχει μάθει για την λέξη unk και τα συμφραζόμενα που αυτή θα πρέπει να έχει, είναι λανθασμένα, λόγω τους εύρους της χρήσης της.

Τέλος αξίζει να τονιστεί πως τα αποτελέσματα των δύο αλγορίθμων ομοιότητας είναι τόσο κοντινά καθώς λόγω της απλότητας του συνόλου ελέγχου η μεταξύ τους σύγκριση δεν μπορεί να είναι αντικειμενική. Στο συγκεκριμένο σύνολο δεδομένων υπάρχουν πολύ μικρά κείμενα και περιλήψεις με μοναδικές εμφανίζεις άγνωστης λέξης στο κείμενο και στην αντίστοιχη περίληψή του. Σε αυτές τις περιπτώσεις οι αλγόριθμοι αποφαίνονται το ίδιο συλλέγοντας το ίδιο σκορ.

5.6 Σύγκριση με παρόμοιες υλοποιήσεις

Όπως έχει αναφερθεί και σε προηγούμενα κεφάλαια, ο κυριότερος λόγος για την επιλογή της χρήσης των συγκεκριμένων σύνολων δεδομένων καθώς και της μετρικής Rouge, αποτελεί η δυνατότητα σύγκρισης των αποτελεσμάτων με άλλες παρόμοιες ερευνητικές εργασίες. Η συγκεκριμένη εργασία έχει εμπνευστεί και στηριχθεί στις εργασίες των [Rush15], [Chop16] στις οποίες χρησιμοποιήθηκαν τα ίδια δεδομένα και μετρική αξιολόγησης, και συνεπώς τα αποτελέσματα που παρουσιάστηκαν είναι άμεσα συγκρίσιμα με τα δικά τους. Σημειώνεται, πως για την παρακάτω σύγκριση που θα πραγματοποιηθεί, τα αποτελέσματα που θα χρησιμοποιηθούν από τις επιστημονικές εργασίες που αναφέρθηκαν, έχουν εξαχθεί κατευθείαν από τις μετρήσεις τους, όπως αυτές παρουσιάστηκαν στα πρωτότυπα κείμενα τους. Ακόμη αξίζει να σημειωθεί ότι κατά την μέτρηση της επίδοσης του μηχανισμού της παρούσας εργασίας δόθηκε πολύ προσοχή στην προσπάθεια αντιγραφής των συνθηκών μέτρησης της επίδοσης που χρησιμοποιήθηκαν στις δύο αυτές επιστημονικές εργασίες.

Στην συνέχεια θα παρουσιαστεί ένας πίνακας που θα περιέχει τα βέλτιστα αποτελέσματα των τριών εργασιών για τα σύνολα δεδομένων Gigaword, DUC, μέσω της μετρικής Rouge, με στόχο την καλύτερη μεταξύ τους σύγκριση.

Πίνακας 5.11: Πίνακας σύγκρισης των βέλτιστων αποτελεσμάτων της παρούσας εργασίας με άλλες παρόμοιες

	Gigaword			**DUC**		
	R1 F1	**R2 F1**	**RL F1**	**R1 F1**	**R2 F1**	**RL F1**
Rush et al. (2015)	0,3100	0,1265	0,2834	0,2818	0,0849	0,2381
Chopra et al. (2016)	0,3378	0,1597	0,3115	**0,2897**	0,0826	0,2406
Παρούσα εργασία	**0,4487**	**0,2296**	**0,4206**	0,2834	**0,1012**	**0,2576**

Και στις τρεις εργασίες τα σύνολα ελέγχου για το σύνολο Gigaword έχουν επιλεγεί με τυχαίο τρόπο και έχουν μέγεθος 2000 αντιστοιχίες άρθρων - περιλήψεων.

Όπως φαίνεται από τον πίνακα, τα αποτελέσματα σε αυτή την εργασία για το σύνολο δεδομένων Gigaword έχουν ξεπεράσει κατά πολύ τα αποτελέσματα των εργασιών στις οποίες στηρίχθηκε. Αυτή η πολύ σημαντική διαφορά στις μετρικές, οφείλεται αφενός στην πολύ καλή προεπεξεργασία των δεδομένων που προηγήθηκε και αφετέρου στην ιδιαίτερη προσοχή που δόθηκε στην επιλογή των παραμέτρων κατά τον σχεδιασμό του μοντέλου.

Όσον αφορά το σύνολο DUC, μπορεί να παρατηρήσει κανείς ότι τα αποτελέσματα αυτής της εργασίας είναι πάλι τα καλύτερα αλλά οι διαφορές είναι πολύ πιο μικρές. Μια εικασία για τον λόγο που συμβαίνει αυτό το φαινόμενο αφορά το γεγονός πως το σύνολο δεδομένων DUC δεν έχει παραχθεί με τον ίδιο τρόπο με αυτόν που παράχθηκε το σύνολο Gigaword (το οποίο είναι και το σύνολο εκπαίδευσης). Αυτό σημαίνει πως πιθανή βελτίωση του μοντέλου προς το σύνολο εκπαίδευσης δεν θα προκαλέσει υποχρεωτικά βελτίωση στο σύνολο αυτό καθώς διέπεται από διαφορετικό ύφος και λεξιλόγιο. Φυσικά, στην πολύ μικρή διαφορά παίζει ρόλο και η καθαρότητα του συνόλου DUC, που ευνοεί και μοντέλα που δεν έχουν δώσει τόση προσοχή στην επεξεργασία και στον καθαρισμό των δεδομένων όπως αυτή η εργασία.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Κατευθύνσεις

6.1 Συμπεράσματα

Σε αυτό το σημείο θα παρουσιαστούν τα τελικά συμπεράσματα που προκύπτουν από την παραπάνω ανάλυση, για όλα τα στάδια του μηχανισμού.

Παρατηρώντας τα αποτελέσματα από τις μετρικές Rouge κατά την διερεύνηση της επίδοσης της προεπεξεργασίας των δεδομένων, μπορεί να φανεί η πολύ μεγάλη αξία που έχει η διαδικασία αυτή στην βελτίωση του μηχανισμού αυτόματης περίληψης. Όπως αναλύθηκε σε προηγούμενο σημείο, η φιλοσοφία πίσω από την εφαρμογή των κανόνων προεπεξεργασίας βασίστηκε στην προσπάθεια ελαχιστοποίησης του θορύβου που υπάρχει στα σύνολα δεδομένων που χρησιμοποιήθηκαν, καθώς λόγω της αναδρομικής φύσης του μηχανισμού, το εκάστοτε στοιχείο θορύβου επηρεάζει σε συλλογικό βαθμό την ακολουθία εξόδου. Είναι πολύ σημαντικό λοιπόν, κατά την δημιουργία κάποιου παρόμοιου συστήματος αυτόματης περίληψης να δίνεται πολύ προσοχή στην ποιότητα των δεδομένων που χρησιμοποιούνται, καθώς υπάρχουν πολύ ισχυρές εξαρτήσεις σε αυτά και τυχόν θόρυβος μπορεί να υπονομεύσει την ορθή λειτουργία του.

Παρατηρώντας τις μετρήσεις Rouge από την διερεύνηση της επίδρασης των σημαντικών παραμέτρων στην επίδοση του μηχανισμού αυτόματης περίληψης, μπορεί να φανεί ότι δημιουργούνται διάφορα σχεδιαστικά διλήμματα κατά τον σχεδιασμό παρόμοιων συστημάτων. Κόντρα στην πρώτη αντίληψη που μπορεί να έχει κάποιος, το μεγαλύτερο μέγεθος λεξιλογίου δεν οδηγεί υποχρεωτικά σε καλύτερη επίδοση. Παρόλα αυτά, μπορεί να ενισχύσει την δυνατότητα γενίκευσης σε κείμενα διαφορετικού λεξιλογίου. Ακόμη, ένα άλλο σχεδιαστικό σημείο που πρέπει να δώσει κανείς προσοχή αφορά την σχέση ανταλλαγής μεταξύ της μεγιστοποίησης της επίδοσης και της ελαχιστοποίησης του χρονικού κόστους λειτουργίας. Μέσω της μείωσης τους μεγέθους της κρυφής κατάστασης των αναδρομικών στοιχείων, μπορεί να επιτευχθεί σημαντική επιτάχυνση στις ταχύτητες εκπαίδευσης, πρόβλεψης καθώς και στην ταχύτητα σύγκλισης του μηχανισμού παραγωγής περίληψης. Καλό είναι λοιπόν να έχουν οριστεί με ακρίβεια οι σχεδιαστικοί στόχοι πριν τον σχεδιασμό αντίστοιχων συστημάτων ώστε να εκπληρωθούν καλύτερα οι ζητούμενες απαιτήσεις.

Ο μηχανισμός χειρισμού άγνωστων λέξεων που παρουσιάστηκε σε αυτήν την εργασία αποτελεί μια πρόταση αντιμετώπισης του συγκεκριμένου προβλήματος, το οποίο αποτελεί ένα από τα μεγαλύτερα εμπόδια κατά τον σχεδιασμό συστημάτων βαθιάς μάθησης για αυτόματη περίληψη κειμένου. Όπως φάνηκε από τα αποτελέσματα μέτρησης της επίδοσης του, ο συγκεκριμένος μηχανισμός βελτίωσε αισθητά τις περιλήψεις που περιείχαν άγνωστες λέξεις, δίνοντας τους ακριβέστερη νοηματική συνοχή, χωρίς να κατορθώσει όμως να τις ανυψώσει στο επίπεδο των υπολοίπων. Το γεγονός αυτό οφείλεται στις νοηματικές αποκλίσεις που δημιουργούνται κατά την αναδρομική διαδικασία παραγωγής της περίληψης, ύστερα από την πρόβλεψη της λέξης γενικού περιεχομένου 'unk'. Σε κάθε περίπτωση, ο μηχανισμός αυτός μπορεί να βοηθήσει στην προώθηση της έρευνας πάνω στο αντικείμενο της αυτόματης παραγωγής περίληψης και στην επίτευξη ισχυρότερων τέτοιων συστημάτων.

6.2 Μελλοντικές Κατευθύνσεις

Στην συγκεκριμένη εργασία πραγματοποιήθηκε εκτενής διερεύνηση ως προς συγκεκριμένες παραμέτρους του μηχανισμού αυτόματης περίληψης κειμένου ενώ αναλύθηκε και ένας καινοφανής μηχανισμός χειρισμού των άγνωστων λέξεων που εμφανίζονται στα κείμενα προς περίληψη. Παρόλα αυτά, η διερεύνηση που πραγματοποιήθηκε δεν μπορεί να χαρακτηριστεί ως πλήρης λόγω διάφορων περιορισμών (κυρίως χρονικών περιορισμών εκπαίδευσης) που προέκυψαν. Στην συνέχεια θα προταθούν κάποιες κατευθύνσεις για περαιτέρω προώθησης της συγκεκριμένης εργασίας με στόχο την εξαγωγή καλύτερων ερευνητικών συμπερασμάτων.

Μια κατεύθυνση που περιέχει αρκετό ερευνητικό ενδιαφέρον αφορά την χρήση διαφορετικών μεγεθών λεξιλογίου για τον κωδικοποιητή και τον αποκωδικοποιητή. Καθώς ο κωδικοποιητής πρέπει να είναι σε θέση να αναγνωρίσει πολλαπλές διαφορετικές λέξεις κειμένου, είθισται να χρησιμοποιεί μεγάλο μέγεθος λεξιλογίου. Αντίθετα ο αποκωδικοποιητής μπορεί να χρησιμοποιήσει μικρότερο μέγεθος λεξιλογίου, το οποίο μπορεί να αποτελεί και υποσύνολο αυτού του κωδικοποιητή. Με αυτό τον τρόπο η λειτουργία του μηχανισμού μπορεί να επιταχυνθεί σε πολύ σημαντικό βαθμό με πιθανώς ελάχιστη επίδραση στην απόδοση του.

Ένα άλλο σημείο που αξίζει περαιτέρω διερεύνηση αφορά την χρήση διαφορετικών συνόλων εκπαίδευσης για την μελέτη της συμπεριφοράς του μηχανισμού σε διαφορετικά είδη φυσικής γλώσσας. Αν και το σύνολο δεδομένων Gigaword που χρησιμοποιήθηκε αποτελεί ένα από τα πιο διαδεδομένα σύνολα που χρησιμοποιούνται για την αντιμετώπιση του προβλήματος της αυτόματης περίληψης, ότι συμπέρασμα έχει προκύψει στα πλαίσια αυτής της εργασίας αφορά την συμπεριφορά γύρω από αυτό, και δεν έχει αντικειμενική υπόσταση. Με περαιτέρω διερεύνηση, μέσω εκπαίδευσης σε διαφορετικά σύνολα δεδομένων, θα μπορούν να προκύψουν πιο γενικευμένα συμπεράσματα και να μοντελοποιηθεί καλύτερα η μεθοδολογία σχεδιασμού παρόμοιων συστημάτων.

Ένα τελικό σημείο διερεύνησης που αξίζει να σημειωθεί αφορά την μελέτη περισσότερων διαφορετικών παραμέτρων κατά τον σχεδιασμό παρόμοιων μηχανισμών βαθιάς μάθησης. Ο συγκεκριμένος μηχανισμός, αποτελείται από πολλαπλά περίπλοκα στοιχεία πάνω στα οποία μπορεί να πραγματοποιηθεί πολύ μεγάλος πειραματισμός. Μερικές ιδέες για τέτοιο πειραματισμό είναι οι εξής:

- χρήση αναδρομικών νευρωνικών δικτύων διαφορετικού τύπου (π.χ. GRU).
- πειραματισμός με μηχανισμούς προσοχής διαφορετικού τύπου.
- πειραματισμός με εναλλακτικές στρατηγικές επιλογής της βέλτιστης πρόβλεψης, σε κάθε βήμα παραγωγής της περίληψης, αντί της ακτινικής αναζήτησης.

Βιβλιογραφία

[Alla17] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez and Krys Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques", *arXiv preprint arXiv:1707.02919*, 2017.

[Bahd14] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv preprint arXiv:1409.0473*, 2014.

[Bahd16] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition", in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 4945–4949, IEEE, 2016.

[Chop16] Sumit Chopra, Michael Auli and Alexander 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.

[Luhn58] Hans Peter Luhn, "The automatic creation of literature abstracts", *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[Nall16] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang et al., "Abstractive text summarization using sequence-to-sequence rnns and beyond", *arXiv preprint arXiv:1602.06023*, 2016.

[Rush15] Alexander M Rush, Sumit Chopra and Jason Weston, "A neural attention model for abstractive sentence summarization", *arXiv preprint arXiv:1509.00685*, 2015.

[Savo10] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler and Christopher G Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[See17] Abigail See, Peter J Liu and Christopher D Manning, "Get to the point: Summarization with pointer-generator networks", *arXiv preprint arXiv:1704.04368*, 2017.

[Trip17] Elizabeth D Trippe, Jacob B Aguilar, Yi H Yan, Mustafa V Nural, Jessica A Brady, Mehdi Assefi, Saeid Safaei, Mehdi Allahyari, Seyedamin Pouriyeh, Mary R Galinski et al., "A vision for health informatics: Introducing the sked framework. an extensible architecture for scientific knowledge extraction from data", *arXiv preprint arXiv:1706.07992*, 2017.

[Turp07] Andrew Turpin, Yohannes Tsegay, David Hawking and Hugh E Williams, "Fast generation of result snippets in web search", in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 127–134, ACM, 2007.

[Venu15] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell and Kate Saenko, "Sequence to sequence-video to text", in *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542, 2015.

[Wu16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey et al., "Google's neural machine translation system: Bridging the gap between human and machine translation", *arXiv preprint arXiv:1609.08144*, 2016.