



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

**Αποσαφήνιση Οντοτήτων με Χρήση
Τεχνικών Επιβλεπόμενης Μάθησης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
Ηλίας Πασχάλης

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

Συνεπιβλέπων: **Γεώργιος Σιόλας**
ΕΔΙΠ Ε.Μ.Π.

Εργαστήριο Ευφυών Υπολογιστικών Συστημάτων
Αθήνα, Οκτώβριος 2018



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

Αποσαφήνιση Οντοτήτων με Χρήση Τεχνικών Επιβλεπόμενης Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ **Ηλίας Πασχάλης**

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

Συνεπιβλέπων: **Γεώργιος Σιόλας**
ΕΔΙΠ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Οκτωβρίου 2018.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

Ανδρέας-Γεώργιος
Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Γιώργος Στάμου
Αναπληρωτής καθηγητής
Ε.Μ.Π.

Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

Εργαστήριο Ευφυών Υπολογιστικών Συστημάτων
Αθήνα, Οκτώβριος 2018

Ηλίας Πασχάλης

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

Copyright © Ηλίας Πασχάλης, 2018. Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Το πρόβλημα της αποσαφήνισης ονοματικών οντοτήτων αναφέρεται στην αντιστοίχηση των ονοματικών οντοτήτων ενός κειμένου, που είναι συνήθως αμφίσημες, σε οντότητες που ανήκουν σε μια βάση γνώσης. Για την επίλυση του προβλήματος αυτού, θα χρησιμοποιήσουμε δύο προσεγγίσεις που αφορούν μεθόδους επιβλεπόμενης μάθησης.

Η πρώτη προσέγγιση χρησιμοποιεί χαρακτηριστικά τα οποία εξάγονται από τη Wikipedia ως βάση γνώσης και σύμφωνα με αυτά εκπαιδεύει ταξινομητές που αξιολογούνται σε ένα σύνολο από τυχαία άρθρα της Wikipedia.

Η δεύτερη προσέγγιση στοχεύει να εκμεταλλευτεί τις βαθύτερες τοπικές ομοιότητες μεταξύ της αναφοράς και των πιθανών οντοτήτων της με συνελικτικά νευρωνικά δίκτυα και την καθολική συνοχή μεταξύ των αναφορών ενός κειμένου ως ακολουθίες προβλέψεων μεταβλητού μήκους με χρήση αναδρομικών νευρωνικών δικτύων. Το μοντέλο έπειτα αξιολογείται σε δύο προσημειωμένα σύνολα δεδομένων και τα αποτελέσματα είναι συγκρίσιμα με τα πιο σύγχρονα συστήματα αποσαφήνισης οντοτήτων.

Λέξεις κλειδιά: Αποσαφήνιση οντοτήτων, αναγνώριση οντοτήτων, Wikipedia, ταξινόμηση, συνελικτικό νευρωνικό δίκτυο, αναδρομικό νευρωνικό δίκτυο

Abstract

The problem of named entity disambiguation refers to the act of matching named entities in a text, which are usually ambiguous, with entities which belong to a knowledge base. To solve this problem, we will use two approaches which make use of supervised learning methods.

The first approach uses features that are extracted from Wikipedia as the knowledge base and with them trains classifiers which are evaluated on a set of random Wikipedia articles.

The second approach aims to take advantage of underlying local similarities between a mention and its target entities using convolutional neural networks as well as the global coherence among the mentions in a text as variable length sequences of predictions using recurrent neural networks. This model is then evaluated in two annotated datasets and the results are comparable to state-of-the-art systems.

Key words: Entity disambiguation, entity recognition, Wikipedia, classification, convolutional neural network, recurrent neural network

Eυχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Ανδρέα Σταφυλοπάτη για την εμπιστοσύνη που μου έδειξε με την ανάθεση αυτής της εργασίας και την ευκαιρία να ασχοληθώ με το συγκεκριμένο ενδιαφέρον θέμα.

Θα ήθελα, επίσης, να ευχαριστήσω τον κ. Γιώργο Σιόλα για την συνεργασία και την καθοδήγηση που μου προσέφερε καθ' όλη τη διάρκεια της εργασίας.

Τέλος, ευχαριστώ την οικογένεια μου για την υπομονή και την συμπαράσταση της κατά τη διάρκεια των σπουδών μου και ιδιαίτερα στο διάστημα εκπόνησης της παρούσας διπλωματικής εργασίας.

Περιεχόμενα

Κατάλογος σχημάτων	15
Κατάλογος πινάκων	17
Τμήματα κώδικα	19
Κατάλογος Αλγορίθμων	21
1 Εισαγωγή	23
1.1 Αποσαφήνιση ονοματικών οντοτήτων	23
1.2 Διάρθρωση του κειμένου	25
2 Θεωρητικό Υπόβαθρο	27
2.1 Ορισμός του προβλήματος	27
2.2 Βάσεις γνώσης	28
2.3 Υποπροβλήματα της αποσαφήνισης οντοτήτων	29
2.3.1 Παραγωγή υποψήφιων οντοτήτων	30
2.3.1.1 Τεχνικές λεξικών ονομάτων	30
2.3.1.2 Τεχνικές εύρεσης εναλλακτικών ονομάτων στο κείμενο	32
2.3.1.3 Τεχνικές βασισμένες σε μηχανές αναζήτησης	33
2.3.2 Ταξινόμηση υποψήφιων οντοτήτων	33
2.3.2.1 Χαρακτηριστικά	34
2.3.3 Αναγνώριση μη αποσαφηνίσιμων αναφορών	38
2.4 Μέθοδοι επιβλεπόμενης μάθησης	41
2.4.1 Naive Bayes	41
2.4.2 Τυχαίο Δάσος	42
2.4.3 Μηχανή Διανυσμάτων Υποστήριξης	42
2.4.4 Συνελικτικά Νευρωνικά Δίκτυα	44
2.4.4.1 Συνελικτικό στρώμα	45
2.4.4.2 Στρώμα Συσσώρευσης	45
2.4.4.3 Αρχιτεκτονική δικτύου	46
2.4.5 Αναδρομικά Νευρωνικά Δίκτυα	47
2.4.5.1 Το πρόβλημα της εξαφανιζόμενης κλίσης	48
2.4.5.2 Περιφραγμένη αναδρομική μονάδα	49
2.5 Μετρικές αξιολόγησης	50
2.6 Σχετικές Εργασίες	52
2.6.1 Berkeley Entity Resolution System	52

Περιεχόμενα

2.6.2 AIDA-light	52
2.6.3 Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks	52
3 Συστήματα Αποσαφήνισης Οντοτήτων	55
3.1 Βασικό μοντέλο	55
3.1.1 Δημιουργία Dataset	57
3.1.2 Παραγωγή υποψήφιων οντοτήτων	61
3.1.3 Εξαγωγή χαρακτηριστικών	62
3.1.3.1 Εκ των προτέρων πιθανότητα	63
3.1.3.2 Πιθανότητα αναφοράς να αποτελεί σύνδεσμο	63
3.1.3.3 Σημασιολογική συνοχή	64
3.1.3.4 Ποιότητα συμφραζομένων	65
3.1.4 Ταξινόμηση υποψήφιων οντοτήτων	66
3.2 Συνδυαστικό Μοντέλο	67
3.2.1 Παραγωγή υποψήφιων οντοτήτων	67
3.2.2 Ταξινόμηση υποψήφιων οντοτήτων	69
3.2.2.1 Κωδικοποίηση	69
3.2.2.2 Τοπική ομοιότητα	72
3.2.2.3 Καθολική ομοιότητα	72
3.3 Σύγκριση των συστημάτων	75
3.3.1 Ομοιότητες	75
3.3.2 Διαφορές	75
4 Αξιολόγηση Συστημάτων	77
4.1 Βασικό μοντέλο	78
4.1.1 Στατιστικά του dataset	78
4.1.2 Ρύθμιση υπερπαραμέτρων	78
4.1.3 Απόδοση του συστήματος	79
4.1.4 Σχολιασμός αποτελεσμάτων	81
4.2 Συνδυαστικό μοντέλο	82
4.2.1 Datasets	82
4.2.1.1 ACE	82
4.2.1.2 CoNLL-YAGO	82
4.2.2 Απόδοση του συστήματος	83
4.2.3 Σχολιασμός αποτελεσμάτων	83
5 Επίλογος και Μελλοντικές Επεκτάσεις	87
Βιβλιογραφία	89
Αναφορές	89
A Άρθρα της Wikipedia	93
B Εργαλεία	101
B.1 Βασικό μοντέλο	101
B.1.1 Mediawiki API	101

Περιεχόμενα

B.1.2 Wikipedia parsers	102
B.2 Συνδυαστικό Μοντέλο	103
B.2.1 Berkeley Entity Resolution System	103
B.2.2 Word2vec	105

Κατάλογος σχημάτων

1	Παράδειγμα μέτρησης σημασιολογικής συνοχής	39
2	Λειτουργία συνάρτησης πυρήνα	43
3	Παράδειγμα υπερεπιπέδων	44
4	Δομή συνελικτικού νευρωνικού δικτύου	46
5	Δομή αναδρομικού νευρωνικού δικτύου	47
6	Περιφραγμένη αναδρομική μονάδα	50
7	Σχεδιάγραμμα αρχιτεκτονικής βασικού μοντέλου	56
8	Σχεδιάγραμμα αρχιτεκτονικής συνδυαστικού μοντέλου	68
9	Διανύσματα λέξεων Word2vec	70
10	Υπολογισμός της κατανεμημένης αναπαράστασης μιας ακολουθίας λέξεων με CNN	71
11	Υπολογισμός της καθολικής ομοιότητας με RNN	74
12	Γραφική απεικόνιση της απόδοση των μοντέλων του πρώτου συστήματος στο σύνολο αξιολόγησης.	81
13	Απόδοση του δεύτερου συστήματος συγκριτικά με state-of-the-art μοντέλα	84

Κατάλογος πινάκων

1	Παράδειγμα ζευγών κλειδιού-τιμής σε λεξικό ονομάτων	31
2	Ρύθμιση υπερπαραμέτρων τυχαίου δάσους	79
3	Ρύθμιση υπερπαραμέτρων SVM	80
4	Απόδοση των μοντέλων στο σύνολο αξιολόγησης.	80
5	Απόδοση της καθολικής ομοιότητας	83
6	Απόδοση του δεύτερου συστήματος συγκριτικά με state-of-the-art μοντέλα	83

Τμήματα κώδικα

1	Παράδειγμα άρθρου της Wikipedia σε μορφή Wikitext	57
2	Το άρθρο 1 μετά από αφαίρεση πινάκων και λιστών	60
3	Οι πρώτες σειρές του parsed dump της Wikipeδια από το wikidump	102
4	Παράδειγμα ζεύγους key-value από το dictionary των queries . . .	104

Κατάλογος Αλγορίθμων

1	Ψευδοκώδικας για τη δημιουργία των υποψήφιων οντοτήτων	62
2	Ψευδοκώδικας τον υπολογισμό της εκ των προτέρων πιθανότητας των υποψήφιων οντοτήτων μιας αναφοράς	63
3	Ψευδοκώδικας για τον υπολογισμό της πιθανότητας μιας αναφοράς να αποτελεί σύνδεσμο	64
4	Υπολογισμός βάρους μίας οντότητας των συμφραζομένων	65
5	Υπολογισμός ποιότητας των συμφραζομένων	66
6	Υπολογισμός συνοχής των υποψήφιων οντοτήτων μίας αναφοράς .	66

Κεφάλαιο 1

Εισαγωγή

Στο κεφάλαιο αυτό θα γίνει μια εισαγωγή του προβλήματος της αποσαφήνισης ονοματικών οντοτήτων με το οποίο θα ασχοληθούμε σε αυτή την εργασία. Αρχικά, στην ενότητα 1.1 παρουσιάζεται η προέλευση του όρου με την ανάγκη που οδήγησε στη διατύπωση του προβλήματος αυτού. Δίνονται, επίσης, ορισμοί κάποιων χρήσιμων βασικών όρων. Στην ενότητα 1.2 γίνεται περιγραφή της δομής των επόμενων κεφαλαίων του κειμένου.

1.1 Αποσαφήνιση ονοματικών οντοτήτων

“Michael Jordan is one of the leading figures in machine learning”. Στην πρόταση αυτή, η αναφορά (mention) “Michael Jordan” μπορεί να αναφέρεται σε πολλά πρόσωπα. Από τα συμφραζόμενα μπορεί κανείς να καταλάβει ότι αναφέρεται στον καθηγητή και επιστήμονα Michael Irwin Jordan και όχι στον γνωστό μπασκετμπολίστα Michael Jeffrey Jordan. Πολλές φορές όμως υπάρχει η ανάγκη να εξάγουμε τέτοιες πληροφορίες από κάποιο αδόμητο κείμενο χωρίς να εμπλακεί ανθρώπινη εργασία που μπορεί να είναι χρονοβόρα και δαπανηρή.

Ο όρος “ονοματική οντότητα” (named entity) που χρησιμοποιείται πλέον ευρέως στην επεξεργασία φυσικής γλώσσας, επινοήθηκε στα πλαίσια του έκτου Διεθνούς Συνεδρίου Αξιολόγησης Τεχνολογίας Εξαγωγής Πληροφορίας (Sixth Message Understanding Conference (MUC-6)). Εκείνον τον καιρό το MUC επικεντρωνόταν σε έργα εξαγωγής πληροφοριών όπου δομημένη πληροφορία σχετικά με δραστηριότητα εταιριών ή άμυνας εξαγόταν από αδόμητο κείμενο όπως άρθρα εφημερίδων. Έτσι φάνηκε επιτακτική η ανάγκη αναγνώρισης πληροφορίας που κάνει αναφορά σε ονόματα και αριθμητικές μονάδες.

Στον όρο ονοματική οντότητα, η λέξη “ονοματική” στοχεύει να περιορίσει το πρόβλημα μόνο στις οντότητες αυτές που ένας ή περισσότεροι “άκαμπτοι προσδιοριστές” σημαίνουν το αναφερθέν. Οι άκαμπτοι προσδιοριστές, όπως ορίστηκαν από τον Saul Kripke, είναι οι προσδιοριστές αυτοί που προσδιορίζουν το ίδιο πράγμα σε κάθε πιθανό κόσμο και περιλαμβάνουν κυρίως ονόματα ανθρώπων, τοποθεσιών και οργανισμών (entity name expression, enamelex) και αριθμούς όπως χρόνος, ημερομηνία (time expression, timex), χρήματα και ποσοστά (numerical expression, numex) καθώς και άλλα πιο συγκεκριμένα αντικείμενα όπως προϊόντα, email, τηλέφωνα. Οι χαλαροί προσδιοριστές, από την άλλη, μπορεί να προσδιορίζουν διαφορετικά πράγματα σε διαφορετικούς κόσμους. Για παράδειγμα η φράση “The 45th President of the United States of America”, μπορεί στον κόσμο μας αναφέρεται στον Donald Trump, τα πράγματα θα ήταν όμως διαφορετικά αν είχε χάσει τις εκλογές και ήταν πρόεδρος η Hilary Clinton. Επομένως, η φράση αυτή προσδιορίζει τον Trump στον κόσμο μας, την Hilary ή και άλλους ανθρώπους σε άλλους κόσμους και είναι ένας χαλαρός προσδιοριστής.

Έχοντας καταλάβει τι σημαίνει ονοματική οντότητα, επιστρέφουμε στο πρόβλημα γύρω από αυτές. Για να γίνει εξαγωγή δομημένης πληροφορίας από ένα αδόμητο κείμενο, είναι απαραίτητα δύο πράγματα. Πρώτον, η αναγνώριση των ονοματικών οντοτήτων και δεύτερον η αποσαφήνισή τους.

Αναγνώριση ονοματικών οντοτήτων ή ταυτοποίηση οντοτήτων ή εξαγωγή οντοτήτων (named entity recognition/NER/entity identification/entity extraction) είναι η διαδικασία εντοπισμού ονοματικών οντοτήτων σε ένα αδόμητο κείμενο και, συνήθως, της κατάταξής τους σε κατηγορίες. Το κομμάτι του εντοπισμού των οντοτήτων είναι πρόβλημα τμηματοποίησης και αφορά την εύρεση λεκτικών μονάδων που αποτελούν ένα όνομα. Το δε πρόβλημα της κατάταξης τους επιδιώκει να κατηγοριοποιήσει τις ονοματικές οντότητες που βρέθηκαν προηγουμένως σε προκαθορισμένες κατηγορίες όπως ονόματα ανθρώπων, τοποθεσίες, ποσότητες και ποσοστά. Αποσαφήνιση ονοματικών οντοτήτων ή σύνδεση οντοτήτων (named entity disambiguation /NED/entity linking) είναι η διαδικασία καθορισμού της ταυτότητας των οντοτήτων ενός κειμένου, που γίνεται με την αντιστοίχηση τους σε οντότητες κάποιας βάσης γνώσης, δομή που αποθηκεύει σύνθετες πληροφορίες για τις ονοματικές οντότητες του κόσμου και των σχέσεων μεταξύ τους.

Ο συνδυασμός των δύο αυτών προβλημάτων, δηλαδή της αναγνώρισης ονοματικών οντοτήτων ενός κειμένου και της αποσαφήνισής τους ονομάζεται NERD (Named Entity Recognition and Disambiguation).

1.2 Διάρθρωση του κειμένου

Στα κεφάλαια που ακολουθούν αναλύονται δύο συστήματα αποσαφήνισης οντοτήτων και το σχετικό θεωρητικό και τεχνολογικό υπόβαθρο στο οποίο στηρίχθηκε η ανάπτυξή τους. Συγκεκριμένα:

- Στο δεύτερο κεφάλαιο με τίτλο “Θεωρητικό Υπόβαθρο” αναλύονται θεωρίες και μέθοδοι που έχουν αναπτυχθεί στον τομέα της αποσαφήνισης οντοτήτων. Θα παρουσιαστούν κάποια διαδεδομένα μοντέλα που χρησιμοποιούνται για το πρόβλημα αυτό με σκοπό να γίνει καλύτερη κατανόηση του προβλήματος, των διαφορετικών τρόπων προσέγγισής του και της συμβολής στις μεθόδους που υλοποιήθηκαν στην εργασία.
- Στο τρίτο κεφάλαιο με τίτλο ”Γενική Περιγραφή Συστημάτων” γίνεται αναλυτική περιγραφή των δύο συστημάτων που αναπτύχθηκαν. Το πρώτο σύστημα αποσαφήνισης χρησιμοποιεί ένα βασικό μοντέλο για αποσαφήνιση οντοτήτων. Αφού γίνει περιγραφή του, παρουσιάζονται τα μειονεκτήματά του και τα σημεία που βελτιώθηκαν με τα μεταγενέστερα συστήματα. Το δεύτερο σύστημα είναι ένα συνδυαστικό μοντέλο που προσθέτει πάνω σε προηγούμενη υλοποίηση. Περιγράφεται ο τρόπος κατασκευής του και οι αρχές λειτουργίας που το διέπουν.
- Στο τέταρτο κεφάλαιο με τίτλο ”Πειραματική Αξιολόγηση και Συμπεράσματα” παρουσιάζονται και σχολιάζονται τα αποτελέσματα των συστημάτων με τελικό στόχο την αξιολόγησή τους.
- Στο πέμπτο κεφάλαιο με τίτλο ”Επίλογος και Μελλοντικές Επεκτάσεις” γίνεται μια επιγραμματική ανακεφαλαίωση των δύο συστημάτων και καταγράφονται δυνατές επεκτάσεις που μπορούν να πραγματοποιηθούν στο μέλλον.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό αναλύεται το θεωρητικό υπόβαθρο της αποσαφήνισης οντοτήτων. Περιγράφεται η συνήθης διαδικασία επίλυσης του προβλήματος και κάποια μοντέλα που έχουν αναπτυχθεί με αναφορές σε παλιότερες εργασίες και τις επιλογές που έγιναν κατά τη σχεδίασή τους. Έτσι, θα γίνει ευκολότερη η επιλογή των χαρακτηριστικών ενός νέου συστήματος.

2.1 Ορισμός του προβλήματος

Δεδομένης μιας βάσης γνώσης που περιέχει ένα σύνολο οντοτήτων E και μιας συλλογής κειμένου στην οποία ένα σύνολο αναφορών ονοματικών οντοτήτων M έχουν αναγνωριστεί εκ των προτέρων, ο στόχος της αποσαφήνισης οντοτήτων είναι η σύνδεση κάθε αναφοράς $m \in M$ με την αντίστοιχη οντότητα $e \in E$ στη βάση γνώσης. Η αναφορά της ονοματικής οντότητας m είναι μια αλληλουχία συμβόλων στο κείμενο που πιθανώς αναφέρεται σε κάποια οντότητα και αναγνωρίζεται εκ των προτέρων. Είναι πιθανό κάποια οντότητα στο κείμενο να μην αντιστοιχεί σε κάποια οντότητα της δοσμένης βάσης γνώσης. Ορίζουμε αυτό το είδος των αναφορών ως μη αποσαφηνίσιμες ή μη συνδέσιμες (unlinkable) και χρησιμοποιούμε τη σήμανση NIL για να αναφερθούμε σε αυτές. Επομένως, αν η οντότητα e που αντιστοιχεί στην αναφορά m δεν υπάρχει στη βάση γνώσης ($e \notin E$) το σύστημα αποσαφήνισης θα πρέπει να επισημάνει τη m ως NIL.

2.2 Βάσεις γνώσης

Η βάση γνώσης είναι ένα θεμελιώδες στοιχείο των συστημάτων αποσαφήνισης οντοτήτων. Οι βάσεις γνώσης περιέχουν πληροφορίες σχετικά με τις οντότητες του κόσμου, τις σημασιολογικές τους κατηγορίες και τις αμοιβαίες σχέσεις μεταξύ των οντοτήτων. Στη συνέχεια, παρατίθενται σύντομες περιγραφές για κάποιες από τις πιο ευρέως χρησιμοποιημένες βάσεις γνώσης στο πεδίο της αποσαφήνισης οντοτήτων.

- **Wikipedia:** Η Wikipedia είναι μια online πολυγλωσσική εγκυκλοπαίδεια. Δημιουργήθηκε από τη συνεισφορά χιλιάδων εθελοντών από όλο τον κόσμο. Αυτή τη στιγμή η Wikipedia είναι η μεγαλύτερη και πιο δημοφιλής διαδικτυακή εγκυκλοπαίδεια στον κόσμο και συνεχίζει να αναπτύσσεται με γρήγορους ρυθμούς. Τη στιγμή εγγραφής αυτής της εργασίας η αγγλική Wikipedia εκτιμάται να έχει 5,647,193 άρθρα, ενώ η ελληνική 146,688. [1]. Η βασική καταχώρηση στη Wikipedia είναι ένα άρθρο, το οποίο ορίζει και περιγράφει μια οντότητα και ζεχωρίζει από τα υπόλοιπα με ένα μοναδικό αναγνωριστικό. Η Wikipedia έχει μεγάλη κάλυψη ονοματικών οντοτήτων και περιέχει επίσης τεράστια ποσότητα γνώσης σχετικά με αξιοσημείωτες οντότητες. Εκτός αυτού, η δομή της παρέχει ένα σύνολο χρήσιμων χαρακτηριστικών για τα συστήματα αποσαφήνισης οντοτήτων, όπως σελίδες οντοτήτων, κατηγορίες άρθρων, σελίδες ανακατεύθυνσης (redirect pages), σελίδες αποσαφήνισης (disambiguation pages) και συνδέσμους σε άλλα άρθρα (hyperlinks).
- **YAGO:** Η YAGO είναι μια βάση γνώσης που δημιουργήθηκε συνδυάζοντας τη Wikipedia και τη λεξική βάση δεδομένων Wordnet. Από τη μια πλευρά, η YAGO περιέχει ένα μεγάλο αριθμό οντοτήτων ίδιας τάξης μεγέθους με αυτό της Wikipedia. Αφετέρου, υιοθετεί από τη Wordnet την καθαρή ταξινομία των εννοιών. Περιέχει πάνω από 10 εκατομμύρια οντότητες και πάνω από 120 εκατομμύρια κατηγορήματα για τις οντότητες αυτές, όπως is-A ιεραρχία (π.χ. τύπος (type) και υποκλάση (subclassOf)) και μη ταξινομικές σχέσεις μεταξύ οντοτήτων (π.χ. ζει-σε (livesIn) και αποφοίτησε-από (graduatedFrom)).
- **DBpedia:** Η DBpedia είναι μια πολυγλωσσική βάση γνώσης που κατασκευάστηκε με την εξαγωγή δομημένων δεδομένων από τη Wikipedia. Ενώ στη Wikipedia η περισσότερη πληροφορία είναι αδόμητη, με τη μορφή φυσικής γλώσσας στο σώμα των άρθρων, η DBpedia συγκεντρώνει τα δομημένα δεδομένα όπως πληροφορίες των infoboxes, κατηγορίες των άρθρων, γεωγραφικές συντεταγμένες και σύνδεσμοι προς εξωτερικές ιστοσελίδες. Αυτή

τη στιγμή, η αγγλική DBpedia περιέχει 4.58 εκατομμύρια οντότητες, από τις οποίες 4.22 εκατομμύρια ταξινομούνται σε μία δομημένη οντολογία [2]. Επίσης, η DBpedia εξελίσσεται αυτόματα, ακολουθώντας τις αλλαγές της Wikipedia.

- **Google Knowledge Graph:** Ο Google Knowledge Graph είναι μια βάση γνώσης που χρησιμοποιεί η Google για να βελτιώσει τα αποτελέσματα της μηχανής αναζήτησής της με πληροφορίες που συλλέγει από διάφορες πηγές, όπως το CIA World Factbook, τη Wikidata και τη Wikipedia. Οι πληροφορίες που περιέχει παρουσιάζονται στους χρήστες σε ένα κουτί δεξιά από τα αποτελέσματα αναζήτησης και είναι επίσης διαθέσιμες με τη χρήση του Knowledge Graph Search API. Μέσα σε 7 μήνες από την ανακοίνωσή του στο κοινό, ο Knowledge Graph τριπλασιάστηκε σε μέγεθος και το 2012 περιείχε 720 εκατομμύρια οντότητες και 18 δισεκατομμύρια γεγονότα, ενώ το 2016 η Google ανακοίνωσε ότι είχε συγκεντρώσει πάνω από 70 δισεκατομμύρια γεγονότα.
- **Wikidata:** Η Wikidata είναι μια δωρεάν, συνεργατική, πολυγλωσσική βάση γνώσης που συλλέγει δεδομένα τα οποία μπορούν να χρησιμοποιηθούν από τη Wikipedia, άλλες wiki της Wikimedia και οποιονδήποτε άλλον το επιθυμεί. Η Wikidata είναι μια document-oriented βάση δεδομένων, όπου κάθε αντικείμενο (item) αντιπροσωπεύει ένα θέμα και αναγνωρίζεται από ένα μοναδικό αριθμό με πρόθεμα το γράμμα Q, γνωστό ως QID. Πληροφορίες προστίθενται στα αντικείμενα με τη δημιουργία δηλώσεων (statements) στη μορφή ζευγών κλειδιού-τιμής, με την κάθε δήλωση να αποτελείται από μία ιδιότητα (το κλειδί) και μια τιμή που συνδέεται με την ιδιότητα.
- **Freebase:** Η Freebase είναι μια βάση γνώσης που δημιουργήθηκε από τα μέλη της κοινότητάς της και παρείχε τη δυνατότητα σε μη προγραμματιστές να επεξεργαστούν τα δομημένα δεδομένα της. Η Freebase σταμάτησε να λειτουργεί το Μάιο του 2016 και τα δεδομένα της μετακινήθηκαν στη Wikidata ενώ το Knowledge Graph API αντικατέστησε το Freebase API.

2.3 Υποπροβλήματα της αποσαφήνισης οντοτήτων

Ένα σύστημα αποσαφήνισης οντοτήτων αποτελείται συνήθως από τρία βήματα: παραγωγή υποψήφιων οντοτήτων, ταξινόμηση υποψήφιων οντοτήτων και πρόβλεψη μη συνδέσιμων αναφορών.

2.3.1 Παραγωγή υποψήφιων οντοτήτων

Κατά το βήμα της παραγωγής υποψήφιων οντοτήτων, για κάθε αναφορά $m \in M$, το σύστημα αποσαφήνισης προσπαθεί να συμπεριλάβει πιθανές οντότητες στις οποίες μπορεί να αναφέρεται η αναφορά m από το σύνολο υποψήφιων οντοτήτων E_m . Αυτό το βήμα είναι εξίσου σημαντικό με το βήμα της ταξινόμησης υποψήφιων οντοτήτων και κρίσιμο για ένα επιτυχές σύστημα αποσαφήνισης [3].

2.3.1.1 Τεχνικές λεξικών ονομάτων

Η δομή της Wikipedia παρέχει ένα σύνολο χρήσιμων χαρακτηριστικών για την παραγωγή υποψήφιων οντοτήτων όπως σελίδες οντοτήτων, σελίδες ανακατεύθυνσης, σελίδες αποσαφήνισης, φράσεις με έντονα γράμματα στις πρώτες παραγράφους και συνδέσμους προς άλλα άρθρα (hyperlinks). Πολλά συστήματα αποσαφήνισης χρησιμοποιούν διαφορετικούς συνδυασμούς των χαρακτηριστικών αυτών για να κατασκευάσουν ένα λεξικό ονομάτων D . Πιο συγκεκριμένα, το λεξικό ονομάτων D είναι μια συλλογή από ζεύγη κλειδιού-τιμής ($key, value$), όπου κάποιο κλειδί k είναι μια αναφορά και η τιμή $k.value$ που του αντιστοιχεί είναι το σύνολο των ονοματικών οντοτήτων στις οποίες μπορούμε να αναφερόμαστε χρησιμοποιώντας την αναφορά k . Οι τεχνικές λεξικών ονομάτων είναι η κύρια μέθοδος παραγωγής υποψήφιων οντοτήτων και χρησιμοποιείται από πολλά συστήματα αποσαφήνισης οντοτήτων [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Στη συνέχεια, γίνεται πιο αναλυτική περιγραφή των χαρακτηριστικών της Wikipedia που αξιοποιούνται για την κατασκευή του λεξικού D .

- **Σελίδες οντοτήτων:** Κάθε σελίδα οντότητας (entity page) στη Wikipedia περιέχει πληροφορίες που επικεντρώνονται γύρω από την οντότητα αυτή. Ο τίτλος της σελίδας είναι γενικά ο πιο συνηθισμένος τρόπος να αναφερόμαστε στην οντότητα. Έτσι, ο τίτλος της σελίδας για κάποια οντότητα προστίθεται στα κλειδιά του λεξικού D ως κάποια αναφορά k και η οντότητα που περιγράφεται στη σελίδα προστίθεται στις τιμές του, $k.value$.
- **Σελίδες ανακατεύθυνσης:** Οι σελίδες ανακατεύθυνσης (redirect pages) υπάρχουν για εναλλακτικά ονόματα με τα οποία μπορούμε να αναφερθούμε σε μια ήδη υπάρχουσα οντότητα της Wikipedia. Συνήθως, περιέχουν συνώνυμους όρους, συντομογραφίες ή άλλες παραλλαγές των ονομάτων. Επομένως,

Key	Value
Michael Jordan	Michael J. Jordan Michael Jordan (footballer) Michael I. Jordan Michael B. Jordan ...
AI	Artificial intelligence A.i. (band) The American Interest Ai (2014 film) ...
Network	Computer network Neural network Social network Network (1976 film) ...

Πίνακας 1: Παράδειγμα ζευγών κλειδιού-τιμής σε λεξικό ονομάτων

ο τίτλος κάποιας σελίδας ανακατεύθυνσης προστίθεται στα κλειδιά του λεξικού D ως αναφορά k και η οντότητα στην οποία αναφέρεται προστίθεται στις τιμές $k.value$.

- **Σελίδες αποσαφήνισης:** Οι σελίδες αποσαφήνισης (disambiguation pages) υπάρχουν για τα ονόματα αυτά που μπορούν να αναφέρονται σε περισσότερες από μια οντότητες και περιέχουν μια λίστα με τις σελίδες των οντοτήτων αυτών. Είναι χρήσιμες για την εξαγωγή συντομογραφιών και εναλλακτικών ονομάτων των οντοτήτων. Άρα, ο τίτλος μιας σελίδας αποσαφήνισης προστίθεται στα κλειδιά του λεξικού D ως αναφορά k και οι οντότητες που απαριθμούνται στη σελίδα αυτή προστίθενται στις τιμές $k.value$.
- **Έντονες φράσεις στις πρώτες παραγράφους:** Γενικά, η πρώτη παράγραφος στα άρθρα της Wikipedia είναι μια περίληψη του άρθρου και μερικές φορές περιέχει φράσεις με έντονα γράμματα. Οι φράσεις αυτές κατά κανόνα είναι ψευδώνυμα ή πλήρη ονόματα των οντοτήτων που περιγράφονται στη σελίδα. Έτσι, αν K είναι το σύνολο των έντονων φράσεων που εμφανίζονται στην πρώτη παράγραφο μιας σελίδας στη Wikipedia, κάθε αναφορά $k \in K$ προστίθεται στο λεξικό D σαν κλειδί και η οντότητα που περιγράφεται στη σελίδα προστίθεται στις τιμές $k.value$.

- **Σύνδεσμοι προς άλλα άρθρα:** Η Wikipedia συνήθως περιέχει συνδέσμους προς άλλα άρθρα (hyperlinks) οντοτήτων που αναφέρονται στο τρέχον άρθρο. Το κείμενο του συνδέσμου (anchor text) που οδηγεί σε μια σελίδα οντότητας αποτελεί χρήσιμη πηγή συνωνύμων και παραλλαγών του ονόματος της οντότητας στη οποία οδηγεί. Άρα το κείμενο του συνδέσμου μπορεί να προστεθεί στα κλειδιά του λεξικού D ως αναφορά k και η οντότητα στην οποία οδηγεί στις τιμές $k.value$.

Χρησιμοποιώντας τα χαρακτηριστικά που περιγράφηκαν παραπάνω, τα συστήματα αποσαφήνισης μπορούν να κατασκευάσουν το λεξικό D . Με βάση αυτό, ο πιο απλός τρόπος παραγωγής των υποψήφιων οντοτήτων E_m για την αναφορά $m \in M$ είναι να ελέγχουν αν η συμβολοσειρά κάποιου κλειδιού k είναι ίδια με της αναφοράς m , και τότε οι οντότητες $k.value$ προστίθενται στο σύνολο υποψήφιων οντοτήτων E_m . Εκτός από την ακριβή σύγκριση συμβολοσειρών (exact string matching), κάποιες μέθοδοι χρησιμοποιούν μερική σύγκριση (partial string matching) [14, 16, 17] με κανόνες όπως το όνομα της οντότητας να περιέχει την αναφορά, ή να περιέχεται σε αυτή, το όνομα της οντότητας να έχει αρκετές κοινές λέξεις με την αναφορά ή να έχει ισχυρή ομοιότητα συμβολοσειράς με την αναφορά.

2.3.1.2 Τεχνικές εύρεσης εναλλακτικών ονομάτων στο κείμενο

Πολλές φορές, σε ένα κείμενο υπάρχουν αναφορές στην ίδια οντότητα με διαφορετικά ονόματα. Αν μπορούμε να βρούμε τις αναφορές αυτές, που μπορεί να είναι πλήρες όνομα, ακρώνυμα, ή άλλα εναλλακτικά ονόματα για την ίδια οντότητα, μπορούμε να τις εκμεταλλευτούμε για να διευρύνουμε το λεξικό ονομάτων που κατασκευάστηκε με τις τεχνικές που περιγράφαμε προηγουμένως. Οι τεχνικές εύρεσης εναλλακτικών ονομάτων χωρίζονται σε ευριστικές μεθόδους και μεθόδους επιβλεπόμενης μάθησης.

- **Ευριστικές μέθοδοι:** Κάποιες προσεγγίσεις [8, 14, 16] ευριστικών μεθόδων χρησιμοποιούν ευριστική αντιστοίχηση μοτίβων (heuristic pattern matching) στο κείμενο γύρω από την αναφορά της οντότητας, με τα πιο συνηθισμένα μοτίβα να είναι παρενθέσεις δίπλα από την αναφορά που περιέχουν το πλήρες όνομά ή ένα ακρώνυμό της. Άλλη μια προσέγγιση [10, 13, 15] είναι η χρήση N-Gram για την εύρεση του πλήρους ονόματος, όπου γίνεται έλεγχος για την ύπαρξη N συνεχόμενων λέξεων στο κείμενο, μετά την αφαίρεση των πιο κοινών λέξεων (stop words) μεταξύ τους, που να έχουν τα ίδια αρχικά με το ακρώνυμο της οντότητας. Επίσης, κάποια μοντέλα [11, 24] χρησιμοποιούν υπάρχοντα συστήματα αναγνώρισης οντοτήτων στο κείμενο και

ελέγχουν αν κάποια εντοπισμένη οντότητα περιέχει τη συμβολοσειρά της αναφοράς και τότε θεωρούν πως η οντότητα αυτή είναι εναλλακτικό όνομα της αναφοράς.

- **Μέθοδοι επιβλεπόμενης μάθησης:** Οι προηγούμενες μέθοδοι που βασίζονται σε ευριστικές μεθόδους δεν μπορούν να εντοπίσουν το πλήρες όνομα για πιο σύνθετα ακρώνυμα στα οποία κάποια γράμματα έχουν διαφορετική σειρά ή λείπουν. Μία μέθοδος επιβλεπόμενης μάθησης που προτάθηκε [19] για την εύρεση του πλήρους ονόματος σύνθετων ακρωνύμων οδήγησε σε 15.1% βελτίωση της ακρίβειας σε σχέση με τις τότε state-of-the-art μεθόδους. Συγκεκριμένα, εντόπιζε πιθανά πλήρη ονόματα από το έγγραφο μέσω ορισμένων προκαθορισμένων στρατηγικών, και έπειτα σε κάθε ζεύγος αναφοράς και πλήρους ονόματος εφαρμοζόταν μία Μηχανή Διανυσμάτων Υποστήριξης (support vector machine - SVM) για την εξαγωγή ενός βαθμού εμπιστοσύνης. Για κάθε ακρώνυμο επιλέγεται το πλήρες όνομα με το μεγαλύτερο βαθμό εμπιστοσύνης.

2.3.1.3 Τεχνικές βασισμένες σε μηχανές αναζήτησης

Ορισμένα συστήματα αποσαφήνισης οντοτήτων προσπαθούν να αξιοποιήσουν ολόκληρη την πληροφορία του διαδικτύου για τον εντοπισμό των υποψηφίων οντοτήτων μέσω μηχανών αναζήτησης. Αυτό συνήθως γίνεται υποβάλλοντας τη συμβολοσειρά της αναφοράς, με ή χωρίς κάποια σύντομα συμφραζόμενα, στο Google API και έπειτα φίλτραροντας από όλα τα αποτελέσματα τις οντότητες αυτές των οποίων το άρθρο στην αγγλική Wikipedia εμφανίζεται στα πρώτα αποτελέσματα, ή έχει τίτλο που δε διαφέρει πολύ από το κείμενο αναζήτησης [8, 16, 20, 27]. Εκτός από τη μηχανή αναζήτησης της Google, αξιοποιείται επίσης και η μηχανή αναζήτησης της Wikipedia η οποία μπορεί να επιστρέψει μια λίστα σχετικών σελίδων οντοτήτων της Wikipedia όταν γίνει αναζήτηση βάσει λέξεων-κλειδιών. Αυτή η λειτουργία χρησιμοποιήθηκε για να παραχθούν υποψήφιες οντότητες σπάνια αναφερόμενων οντοτήτων κάνοντας αναζήτηση με τη συμβολοσειρά της αναφοράς [12].

2.3.2 Ταξινόμηση υποψήφιων οντοτήτων

Στην προηγούμενη ενότητα αναλύθηκαν μέθοδοι με τις οποίες μπορεί να παραχθεί το σύνολο υποψήφιων οντοτήτων E_m για κάθε αναφορά m . Δηλώνουμε το μέγεθος του E_m ως $|E_m|$ και τις υποψήφιες οντότητες του συνόλου ως e_i , όπου

$1 \leq i \leq |E_m|$. Στις περισσότερες περιπτώσεις, το σύνολο E_m περιέχει περισσότερες από μία υποψήφια οντότητα ($|E_m| \geq 1$). Ενδεικτικά, στο ConLL dataset ο μέσος αριθμός υποψήφιων οντοτήτων βρέθηκε να είναι 73 [28]. Επομένως, πρέπει να λυθεί το υπόλοιπο μέρος του προβλήματος, το οποίο είναι η εύρεση κατάλληλων κριτηρίων για να επιλεχθεί η πιο σωστή οντότητα για την αναφορά m από το σύνολο E_m . Οι μέθοδοι ταξινόμησης των υποψήφιων οντοτήτων μπορεί αν είναι επιβλεπόμενες ή μη επιβλεπόμενες. Στη συνέχεια, θα εξετάσουμε κάποια χρήσιμα χαρακτηριστικά που εκμεταλλεύονται τα συστήματα αποσαφήνισης και τις κύριες τεχνικές που χρησιμοποιούνται.

2.3.2.1 Χαρακτηριστικά

Τα χαρακτηριστικά που χρησιμοποιούνται για την ταξινόμηση των υποψήφιων οντοτήτων μπορούν να κατηγοριοποιηθούν σε ανεξάρτητα από τα συμφραζόμενα και εξαρτώμενα από τα συμφραζόμενα.

Τα ανεξάρτητα από τα συμφραζόμενα, ή τοπικά, χαρακτηριστικά είναι αυτά που αξιοποιούν μόνο τη συμβολοσειρά της αναφοράς και τις γνώσεις που έχουμε για τις ανεξάρτητες οντότητες. Παρακάτω περιγράφονται τα κύρια χαρακτηριστικά που ανήκουν σε αυτή την κατηγορία.

- **Σύγκριση συμβολοσειράς ονόματος:** Αυτό είναι το πιο βασικό χαρακτηριστικό που μπορεί να χρησιμοποιήσει κανείς. Οι πιο συνηθισμένοι τρόποι σύγκρισης συμβολοσειράς ονόματος μεταξύ μιας αναφοράς m και μιας υποψήφιας οντότητας e περιλαμβάνουν ελέγχους όπως η m να ταιριάζει ακριβώς με την e , η e να ξεκινάει ή να τελειώνει με την m , η m να περιέχεται στην e , πόσες κοινές λέξεις υπάρχουν μεταξύ τους και άλλα μέτρα ομοιότητας συμβολοσειρών όπως edit distance, character Dice, skip bigram Dice, Hamming Distance [15, 16, 20, 27].
- **Εκ των προτέρων πιθανότητα:** Ένα άλλο ανεξάρτητο από τα συμφραζόμενα χαρακτηριστικό που αποδείχθηκε πολύ χρήσιμο είναι η δημοτικότητα της υποψήφιας οντότητας σχετικά με την αναφορά της, η οποία ουσιαστικά πρόκειται για την εκ των προτέρων πιθανότητα (prior probability) εμφάνισης της υποψήφιας οντότητας δοσμένης της αναφοράς αυτής. Χρησιμοποιώντας τις πληροφορίες της Wikipedia, έχοντας μια αναφορά m και μια υποψήφια οντότητα $e_i \in E_m$ η εκ των προτέρων πιθανότητα της οντότητας είναι η αναλογία του αριθμού των συνδέσμων με κείμενο τους (anchor text) την αναφορά m που οδηγούν στη σελίδα της οντότητας e_i προς τον αριθμό όλων των

συνδέσμων με anchor text m . Αν δηλώσουμε το σύνολο όλων των συνδέσμων με κείμενο την αναφορά m ως

$$L_m = \{ \text{Wikipedia link} \mid \text{anchor(link)} = m \wedge \exists e_i \in E_m : \text{link} \mapsto e_i \}$$

τότε η εκ των προτέρων πιθανότητα της υποψήφιας οντότητας e_i για την αναφορά m είναι:

$$Pr(e_i|m) = \frac{\text{count}_m(e_i)}{\sum_{\forall e_j \in E_m} \text{count}_m(e_j)}$$

όπου

$$\text{count}_m(e_j) = \sum_{\forall l \in L(m)} \begin{cases} 1, & l \mapsto e_i \\ 0, & l \not\mapsto e_i \end{cases}$$

Στις περισσότερες περιπτώσεις η αναφορά m αναφέρεται στην οντότητα με τη μεγαλύτερη εκ των προτέρων πιθανότητα. Για παράδειγμα η αναφορά "Michael Jordan" σε ένα κείμενο είναι πιο πιθανό να αναφέρεται στον μπασκετμπολίστα Michael Jeffrey Jordan παρά στον καθηγητή Michael Irwin Jordan. Πολλά συστήματα αποσαφήνισης την εισάγουν στα χαρακτηριστικά τους [4, 9, 20, 23, 25, 26, 29].

- **Τύπος οντότητας:** Αυτό το χαρακτηριστικό υποδεικνύει αν ο τύπος της αναφοράς είναι ίδιος με τον τύπο της υποψήφιας οντότητας (π.χ. άνθρωπος, τοποθεσία, οργανισμός). Ο εντοπισμός του τύπου της αναφοράς τις περισσότερες φορές γίνεται από το σύστημα αναγνώρισης οντοτήτων στο κείμενο, ενώ για τον εντοπισμό του τύπου της υποψήφιας οντότητας χρησιμοποιούνται οι πληροφορίες που περιέχονται στη βάση γνώσης. Για τις οντότητες των οποίων ο τύπος δεν βρίσκεται στη βάση γνώσης, μπορεί να χρησιμοποιηθεί άλλη πηγή όπως τα infoboxes της Wikipedia και η DBpedia, ή να εντοπιστεί από κάποιο σύστημα αναγνώρισης οντοτήτων όπως το CiceroLite. Με τον τρόπο αυτό βρίσκεται ο τύπος της πλειοψηφίας των υποψήφιων οντοτήτων του συνόλου E_m και μπορούν να αποκλειστούν ή να δοθεί μικρότερη βαρύτητα σε αυτές των οποίων ο τύπος δεν ταιριάζει με αυτόν της οντότητας m . Για παράδειγμα, αν ξέρουμε ότι η αναφορά "Washington" στο κείμενο αναφέρεται σε τοποθεσία, μπορούμε να δώσουμε λιγότερο βάρος στην οντότητα του προέδρου George Washington, στην οντότητα του Washington College και άλλες οντότητες διαφορετικού τύπου στο E_m . Το χαρακτηριστικό αυτό χρησιμοποιήθηκε στις εργασίες [16, 20, 27].

Αν και τα ανεξάρτητα από τα συμφραζόμενα χαρακτηριστικά είναι χρήσιμα, οι πληροφορίες που παρέχουν αφορούν μόνο την αναφορά και την υποψήφια οντότητα. Εκτός αυτών, είναι ιδιαίτερα αναγκαία η χρήση χαρακτηριστικών που σχετίζονται με τα συμφραζόμενα όπου εμφανίζεται η αναφορά της οντότητας. Ακολουθεί, λοιπόν, η περιγραφή εξαρτωμένων από τα συμφραζόμενα, ή αλλιώς καθολικών, χαρακτηριστικών.

- **Ομοιότητα κειμένου συμφραζομένων:** Το πιο βασικό χαρακτηριστικό σχετικό με τα συμφραζόμενα είναι ο υπολογισμός της ομοιότητας του κειμένου γύρω από την αν φορά και του κειμένου του άρθρου της υποψήφιας οντότητας. Δύο μορφές για την αναπαράσταση των συμφραζομένων είναι οι εξής:
 - **Bag of words.** Για κάθε αναφορά, τα συμφραζόμενα αναπαρίστανται ως ένα σύνολο λέξεων (bag of words) από ολόκληρο το κείμενο του άρθρου στο οποίο εμφανίζεται η αναφορά [4, 13, 14, 17], ή από ένα κατάλληλο παράθυρο γύρο από την αναφορά [6, 9, 22, 23, 26], που σημαίνει πως αν A το σύνολο των λέξεων όλου του άρθρου και W το μέγεθος του παραθύρου, το σύνολο των λέξεων των συμφραζομένων είναι $\{w \in A \mid 0 < distance(w, m) \leq W/2\}$. Για κάθε υποψήφια οντότητα, τα συμφραζόμενα συνήθως αναπαρίστανται σαν bag of words των λέξεων από ολόκληρο το κείμενο του άρθρου της οντότητας στη Wikipedia [6, 9, 13, 17, 22], την πρώτη παράγραφο του άρθρου της Wikipedia [9], ένα παράθυρο γύρο από κάθε εμφάνισή της οντότητας στο corpus των σελίδων της Wikipedia [26] μεταξύ άλλων μεθόδων.
 - **Διανύσματα εννοιών.** Στο το άρθρο όπου εμφανίζεται η αναφορά και στο άρθρο της Wikipedia της υποψήφιας οντότητας, τα συστήματα εξάγουν κάποιες λέξεις κλειδιά [28], anchor texts [9], περιγραφικές ετικέτες [5], κατηγορίες [7, 27], ή συγγενικές οντότητες, γνωρίσματά τους και πληροφορίες από το Wikipedia infobox [14, 16, 27], για να κατασκευάσουν ένα διάνυσμα εννοιών που αναπαριστά τη σημασιολογικά συμφραζόμενα του άρθρου.

Με βάση αυτά τα διαφορετικά είδη αναπαράστασης, το κείμενο γύρω από την αναφορά ή το κείμενο που σχετίζεται με την υποψήφια οντότητα μπορεί να μετατραπεί σε ένα διάνυσμα. Έχοντας τα διανύσματα αυτά, μπορεί να υπολογιστεί ένας βαθμός ομοιότητας μεταξύ των συμφραζομένων της αναφοράς και του κειμένου που περιγράφει την υποψήφια οντότητα. Για το σκοπό αυτό έχουν χρησιμοποιηθεί διάφορες τεχνικές, που συμπεριλαμβάνουν εσωτερικό γινόμενο [4, 9, 22], ομοιότητα συνημιτόνου [6, 9, 12, 13,

15, 17, 23, 26, 27], επικάλυψη λέξεων [28], απόκλιση Kullback–Leibler [28], μετρήσεις βασισμένες σε n-gram [28] και ομοιότητα Jaccard [9].

- **Συνοχή μεταξύ των οντοτήτων του κειμένου:** Αναμφίβολα, τα συμφραζόμενα γύρω από την αναφορά διαδραματίζουν καθοριστικό ρόλο στην αποσαφήνιση οντοτήτων. Αλλά εκτός αυτού, οι άλλες αναφορές που χρήζουν αποσαφήνισης είναι επίσης σημαντικές. Πολλά state-of-the-art συστήματα αποσαφήνισης κάνουν την υπόθεση ότι το ένα άρθρο αναφέρεται σε οντότητες που έχουν συνοχή μεταξύ τους από ένα ή λίγα σχετικά μεταξύ τους θέματα και αυτή η τοπική συνοχή μπορεί να αξιοποιηθεί για τη συλλογική αποσαφήνιση των οντοτήτων του ίδιου κειμένου [4, 7, 9, 22, 23, 25, 26].

Για να μετρήσουν τη συνοχή μεταξύ των οντοτήτων του κειμένου, κάποιες προσεγγίσεις [9, 22, 23, 25, 26, 28] ακολουθούν μια μέθοδο βασισμένη στους συνδέσμους της Wikipedia, που ονομάζεται WLM (Wikipedia Link-based Measure) και προτάθηκε από τους Milne και Witten [29]. Η μέθοδος αυτή αποτελεί μια έκφραση του Normalized Google Distance [30] και στηρίζεται στην υπόθεση ότι δύο οντότητες της Wikipedia είναι σημασιολογικά συγγενικές εάν υπάρχουν πολλά άρθρα της Wikipedia που οδηγούν και στις δύο. Συγκεκριμένα, αν e, e' δύο οντότητες της Wikipedia, η τοπική συνοχή μεταξύ τους ορίζεται ως εξής:

$$WLM(e, e') = \frac{\log(\max(|IL|, |IL'|)) - \log(IL \cap IL')}{\log(|WP|) - \log(\min(|IL|, |IL'|))}$$

όπου IL και IL' είναι το σύνολο των άρθρων της Wikipedia με συνδέσμους προς τις e και e' αντίστοιχα και WP το σύνολο όλων των άρθρων της Wikipedia. Εκτός του Google Distance μοντέλου, έχει εφαρμοστεί μετρική PMI (Point-wise Mutual Information) [23] που υπολογίζεται ως εξής:

$$PMI(e, e') = \frac{|LI \cap LI'| / |WP|}{|LI| / |WP| \cdot |LI'| / |WP|}$$

καθώς και Jaccard distance [4]:

$$J(e, e') = \frac{|LI \cap LI'|}{|LI \cup LI'|}$$

Για παράδειγμα, στην Εικόνα 1 φαίνεται η σημασιολογική συνοχή για δύο οντότητες της Wikipedia μέσα από τμήμα των εισερχόμενων συνδέσμων προς αυτές (backlink). Ο αριθμός των εισερχόμενων συνδέσμων για την οντότητα Artificial intelligence είναι 4613, ενώ για την οντότητα Ethics που αναφέρεται στον κλάδο της φιλοσοφίας είναι 2979. Από αυτούς, 101

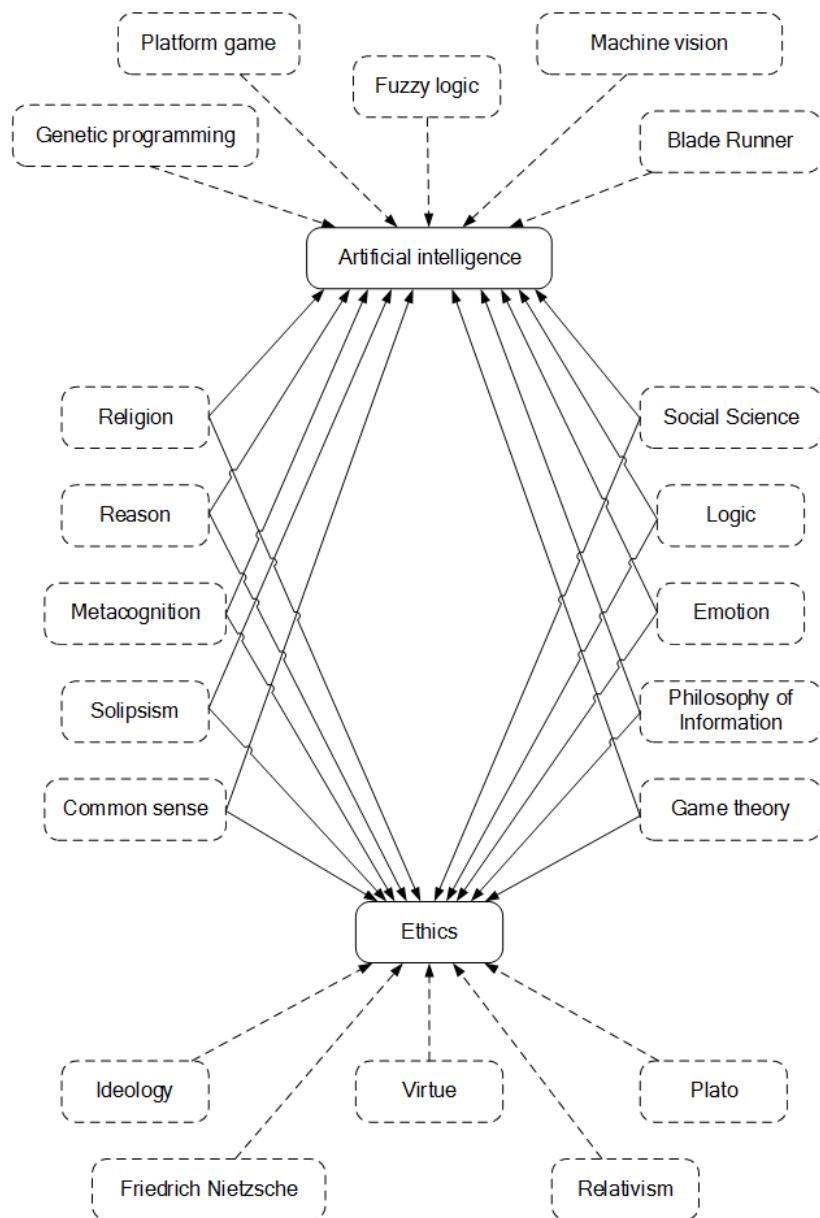
σύνδεσμοι είναι κοινοί και για τις δύο οντότητες και η WLM μεταξύ τους υπολογίζεται να είναι 0.4937, που σημαίνει ότι έχουν αρκετή σημασιολογική συνοχή. Αντιθέτως, αν δοκιμάσουμε να βρούμε τη συνοχή μεταξύ Artificial Intelligence και της οντότητας Ethics (journal) που αναφέρεται σε ακαδημαϊκή εφημερίδα, η συνοχή μεταξύ τους είναι αρκετά μικρότερη και ίση με 0.2669. Άρα με βάση αυτό το χαρακτηριστικό, η αναφορά ethics σε αυτό το πλαίσιο είναι πιθανότερο να αναφέρεται στον κλάδο της φιλοσοφίας παρά στην εφημερίδα.

Αν και το χαρακτηριστικό της συνοχής των οντοτήτων του κειμένου έχει μεγάλη αποτελεσματικότητα στο πρόβλημα αποσαφήνισης οντοτήτων, ο υπολογισμός του δεν είναι εύκολος. Για μια μόνο αναφορά απαιτεί τη σύγκριση των υποψήφιων οντοτήτων της με τις οντότητες όλων των άλλων αναφορών του κειμένου. Όμως οι οντότητες που αντιστοιχούν στις άλλες αναφορές δεν είναι γνωστές εκ των προτέρων και πρέπει επίσης να βρεθούν από το σύστημα. Επομένως, οι αναθέσεις οντοτήτων στις αναφορές είναι αλληλοεξαρτώμενες μεταξύ τους. Σύμφωνα με κάποιες εργασίες [28, 9] η βελτιστοποίηση αυτού του προβλήματος είναι NP-hard, κάτι που κάνει το χαρακτηριστικό αυτό πολύ υπολογιστικά ακριβό και χρονοβόρο για καθημερινές εφαρμογές.

Ο μεγάλος αριθμός χαρακτηριστικών που παρουσιάστηκαν αντικατοπτρίζει το μεγάλο αριθμό πτυχών που πρέπει να ληφθούν υπόψη για το πρόβλημα της αποσαφήνισης οντοτήτων. Δυστυχώς, οι έρευνες που συγκρίνουν την αποτελεσματικότητα των διαφόρων χαρακτηριστικών είναι λίγες. Ωστόσο, αξίζει να δοθεί έμφαση στο ότι κανένα χαρακτηριστικό δεν είναι καλύτερο από τα υπόλοιπα σε όλα τα είδη των dataset, αφού χαρακτηριστικά που έχουν υψηλή επίδοση σε μερικά dataset μπορεί να είναι μη αποτελεσματικά σε άλλα. Ως εκ τούτου, κατά την επιλογή των χαρακτηριστικών που θα χρησιμοποιεί το σύστημα αποσαφήνισης, πρέπει να ληφθούν αποφάσεις για ζητήματα όπως ο συμβιβασμός που θα γίνει ανάμεσα σε διαφορετικές μετρικές αξιολόγησης των αποτελεσμάτων και τα χαρακτηριστικά του εφαρμοσμένου dataset.

2.3.3 Αναγνώριση μη αποσαφηνίσιμων αναφορών

Στην προηγούμενη ενότητα περιγράφηκαν τα κυριότερα χαρακτηριστικά που χρησιμοποιούνται για την ταξινόμηση των υποψήφιων οντοτήτων του συνόλου E_m . Με βάση αυτά, τα συστήματα αποσαφήνισης μπορούν να επιλέξουν την πιο υψηλά βαθμολογημένη οντότητα e_{top} του E_m ως την οντότητα που αντιστοιχεί



Σχήμα 1: Παράδειγμα μέτρησης σημασιολογικής συνοχής για τις οντότητες Artificial intelligence και Ethics από εισερχόμενους συνδέσμους της Wikipedia.

στην αναφορά m . Υπάρχουν όμως περιπτώσεις στις οποίες η οντότητα που αντιστοιχεί στην αναφορά δεν υπάρχει στο E_m , δηλαδή δεν έχει δική της καταχώρηση στη βάση γνώσης. Αυτές οι περιπτώσεις είναι αναμενόμενες δεδομένου του γρήγορου ρυθμού εμφάνισης νέων οντοτήτων και της αδυναμίας των βάσεων γνώσης να περιλαμβάνουν όλες τις λιγότερο γνωστές οντότητες του κόσμου.

Αρκετές εργασίες [7, 9, 22] χρησιμοποιούν την πιο απλή προσέγγιση και θεωρούν ότι η βάση γνώσης περιέχει όλες τις οντότητες που αντιστοιχούν στις αναφορές και άρα αγνοούν το πρόβλημα των μη αποσαφηνίσιμων αναφορών. Μερικές άλλες προσεγγίσεις [10, 14] χρησιμοποιούν μια απλή ευριστική μέθοδο για να τις προβλέψουν. Αν το σύνολο E_m που δημιουργήθηκε κατά τη διαδικασία παραγωγής υποψήφιων οντοτήτων για την οντότητα m είναι άδειο, προβλέπουν ότι η m είναι μη αποσαφηνίσιμη και επιστρέφουν NIL.

Εκτός από αυτές τις μεθόδους, πολλά συστήματα αποσαφήνισης [6, 8, 16, 24, 25, 26] χρησιμοποιούν ένα κατώφλι για την ανάθεση του NIL, το οποίο συνήθως υπολογίζεται κατά την εκπαίδευση του συστήματος. Συγκεκριμένα, η βαθμολογία της επικρατέστερης οντότητας e_{top} συγκρίνεται με την τιμή κατωφλίου και αν είναι μικρότερη επιστρέφεται NIL για την αναφορά m και τη θεωρούν μη αποσαφηνίσιμη. Διαφορετικά επιστρέφουν την οντότητα e_{top} ως σωστή αντιστοίχηση της αναφοράς m .

Ένας μεγάλος ακόμη αριθμός συστημάτων αποσαφήνισης [13, 15, 16, 17, 19, 20, 21, 23, 27] προβλέπουν τις μη αποσαφηνίσιμες αναφορές με τεχνικές επιβλεπόμενης μάθησης. Ο μεγαλύτερος αριθμός αυτών περιλαμβάνει έναν binary classifier σε μορφή SVM, ο οποίος με δεδομένο ένα ζεύγος αναφοράς και την πιο υψηλά βαθμολογημένη οντότητα $\langle m, e_{top} \rangle$, μπορεί να προβλέψει αν η e_{top} είναι η σωστή οντότητα της m , διαφορετικά επιστρέφει NIL.

Η πρόβλεψη των μη αποσαφηνίσιμων οντοτήτων, όμως, μπορεί να εισαχθεί στο βήμα της βαθμολόγησης της ταξινόμησης των υποψήφιων οντοτήτων, όπως έγινε στις εργασίες [21, 27, 31]. Αν η οντότητα NIL εισαχθεί στο σύνολο των υποψήφιων οντοτήτων ως ξεχωριστός υποψήφιος στόχος της αναφοράς, το νέο σύνολο υποψήφιων οντοτήτων $E_m \cup \{NIL\}$ θα καλύπτει πλέον την περίπτωση που η αναφορά m δεν μπορεί να αποσαφηνιστεί, καθώς αν επικρατέστερη οντότητα βρεθεί να είναι η NIL, η αναφορά θεωρείται μη αποσαφηνίσιμη, διαφορετικά επιστρέφεται η οντότητα e_{top} ως σωστή αντιστοίχηση της αναφοράς.

2.4 Μέθοδοι επιβλεπώμενης μάθησης

Στην υποενότητα αυτή γίνεται μια συνοπτική θεωρητική περιγραφή των μεθόδων επιβλεπώμενης μάθησης (supervised learning) που χρησιμοποιούνται στην εργασία. Οι μέθοδοι που χρησιμοποιούνται είναι ταξινομητές (classifiers) Naive Bayes, τυχαίο δάσος και μηχανή διανυσμάτων υποστήριξης καθώς και συνελικτικά και αναδρομικά νευρωνικά δίκτυα.

2.4.1 Naive Bayes

Ο Naive Bayes ταξινομητής είναι μια μέθοδος μάθησης που ακολουθεί πιθανοθεωρητική προσέγγιση βασισμένη στο θεώρημα του Bayes. Βασίζεται στην υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, δεδομένης της ετικέτας κατηγορίας y . Δοσμένου ενός παραδείγματος προς ταξινόμηση, το οποίο αναπαρίσταται από ένα διάνυσμα $x = (x_1, \dots, x_n)$ με n ανεξάρτητα χαρακτηριστικά, αποδίδει σε αυτό το παράδειγμα πιθανότητες $p(C_k|x_1, \dots, x_n)$ για καθένα από τα πιθανά αποτελέσματα ή κλάσεις C_k που δίνονται από τη σχέση:

$$p(C_k|x_1, \dots, x_n) = p(C_k) \prod_{i=1}^n p(x_i|C_k)$$

Από αυτό το πιθανοτικό μοντέλο μπορεί να κατασκευαστεί ένας ταξινομητής έχοντας ως κανόνα απόφασης την επιλογή της κλάσης με ταμπέλα $\hat{y} = C_k$ για κάποιο k έτσι ώστε:

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^n p(x_i|C_k)$$

Μια άλλη εκδοχή του Naive Bayes ταξινομητή, είναι ο Gaussian Naive Bayes. Σε αυτή την εκδοχή, γίνεται η υπόθεση ότι η πιθανότητα των χαρακτηριστικών ακολουθούν γκαουσιανή κατανομή. Έστω μ_k η μέση τιμή και σ_k^2 η διακύμανση των τιμών του x_i που σχετίζονται με την κλάση C_k , τότε η κατανομή πιθανότητας του x_i δοσμένης της κλάσης C_k είναι:

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right)$$

2.4.2 Τυχαίο Δάσος

Τα τυχαία δάση ανήκουν στις συνδυαστικές μεθόδους ταξινόμησης και αποτελούν ουσιαστικά μια συλλογή από δέντρα απόφασης (decision trees).

Κάθε δέντρο απόφασης δέχεται ως είσοδο στη ρίζα του ένα σύνολο δειγμάτων εκπαίδευσης και με βάση αυτά κατασκευάζει ένα σύνολο κανόνων αποφάσεων (decision rules) με σκοπό την ταξινόμηση μελλοντικών δειγμάτων στις κλάσεις αυτές. Κάθε ενδιάμεσος κόμβος (node) του δέντρου περιέχει μια τέτοια απόφαση που γεννά δύο ή περισσότερα κλαδιά (branches) τα οποία χωρίζουν το υποσύνολο των δειγμάτων του σε μικρότερα υποσύνολα στο επόμενο επίπεδο. Όταν καταλήξει σε μία κλάση ή μια πιθανοτική πρόβλεψη, αυτή βρίσκεται σε ένα φύλλο (leaf) του δέντρου.

Για την δημιουργία του συνόλου εκπαίδευσης ενός δάσους, χρησιμοποιείται η τεχνική bootstrap aggregating ή bagging, κατά την οποία από το συνολικό πλήθος των παραδειγμάτων εκπαίδευσης μεγέθους n , δημιουργούνται B καινούργια σύνολα εκπαίδευσης, επιλέγοντας για καθένα από αυτά n φορές ένα παράδειγμα από το αρχικό σύνολο. Στα νέα αυτά σύνολα εκπαίδευσης, ένα παράδειγμα του αρχικού συνόλου μπορεί να εμφανίζεται παραπάνω από μία φορά. Με τη διαδικασία αυτή τα δέντρα του δάσους αποσυσχετίζονται μιας και το καθένα χρησιμοποιεί διαφορετικό σύνολο δειγμάτων εκπαίδευσης.

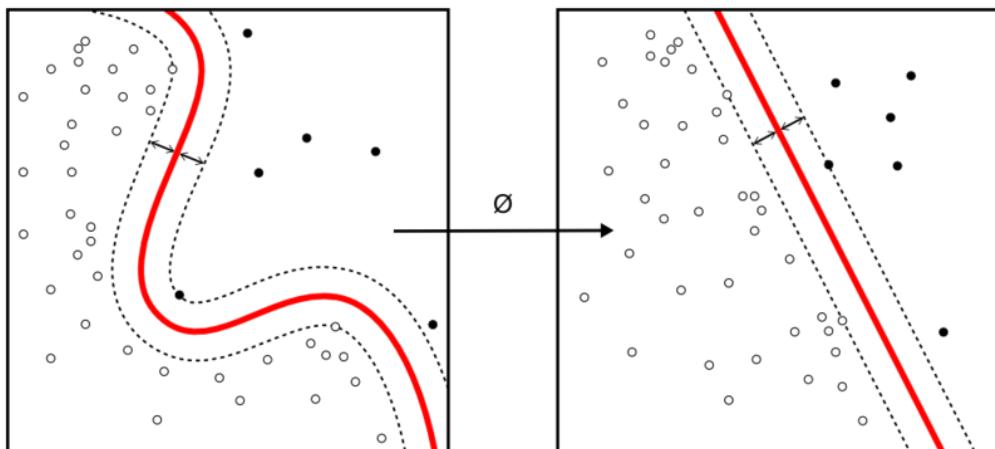
Μετά την εκπαίδευση, η πρόβλεψη y' των δειγμάτων αξιολόγησης x' πραγματοποιείται υπολογίζοντας τον μέσο όρο των πιθανοτικών προβλέψεων όλων των δέντρων f_b του δάσους μεγέθους B .

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

2.4.3 Μηχανή Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (support vector machines or SVMs) είναι μοντέλα επιβλεπόμενης μάθησης τα οποία αναλύουν δεδομένα για ταξινόμηση και ανάλυση παλινδρόμησης (regression analysis). Θα αναφερθούμε στα binary classification SVMs αλλά το μοντέλο αυτό μπορεί να επεκταθεί για multiclass classification πρόβλημα, το οποίο μπορεί να θεωρηθεί ως πολλαπλά binary classification προβλήματα.

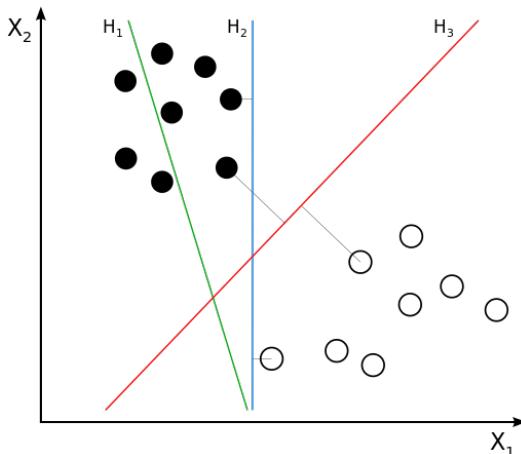
Δοσμένου ενός συνόλου δειγμάτων εκπαίδευσης, όπου το καθένα μπορεί να ανήκει σε μία από τις δύο πιθανές κλάσεις, ο αλγόριθμος εκπαίδευσης SVM επιδιώκει να κτίσει ένα μοντέλο το οποίο είναι αναπαράσταση των παραδειγμάτων στο χώρο με τέτοιο τρόπο ώστε τα παραδείγματα των διαφορετικών κλάσεων χωρίζονται από ευδιάκριτο, όσο δυνατόν μεγαλύτερο κενό. Τα σημεία που βρίσκονται στα όρια του κενού αυτού ονομάζονται διανύσματα υποστήριξης. Για να είναι ευκολότερος ο διαχωρισμός αυτός, ο αρχικό χώρος χαρακτηριστικών (feature space) του SVM προβάλλεται σε ένα χώρο χαρακτηριστικών περισσότερων διαστάσεων. Προκειμένου η υπολογιστική πολυπλοκότητα να είναι σε λογικά επίπεδα, οι προβολές που χρησιμοποιεί το SVM είναι σχεδιασμένες ώστε να εξασφαλίζεται ο εύκολος υπολογισμός των εσωτερικών γινομένων που απαιτούνται, με χρήση μιας επιλεγμένης συνάρτησης πυρήνα (kernel function) $k(x, y)$.



Σχήμα 2: Λειτουργία συνάρτησης πυρήνα. Το μη γραμμικά διαχωρίσιμο πρόβλημα στον αρχικό χώρο χαρακτηριστικών γίνεται γραμμικά διαχωρίσιμο σε ένα χώρο χαρακτηριστικών υψηλότερης διάστασης, άρα μπορούμε να βρούμε απλούστερα υπερεπίπεδα.

Πιο συγκεκριμένα, το SVM μοντέλο κατασκευάζει ένα υπερεπίπεδο ή ένα σύνολο υπερεπιπέδων στον πολυδιάστατο χώρο, τα οποία χρησιμοποιούνται για το διαχωρισμό των παραδειγμάτων σε κατηγορίες. Τα υπερεπίπεδα αυτά στον πολυδιάστατο χώρο ορίζονται ως οι γεωμετρικοί τόποι των σημείων για τα οποία το εσωτερικό γινόμενό τους με ένα διάνυσμα του χώρου αυτού είναι σταθερό. Τα διανύσματα που ορίζουν υπερεπίπεδα μπορούν να επιλεχθούν με τέτοιο τρόπο ώστε να είναι γραμμικοί συνδυασμοί των διανυσμάτων εισόδου x_i με παραμέτρους a_i . Με αυτή την επιλογή υπερεπιπέδου, τα σημεία x του αρχικού χώρου που προβάλλονται στο υπερεπίπεδο ορίζονται από τη σχέση: $\sum_i \alpha_i k(x_i, x) = \text{constant}$.

Στην περίπτωση ενός γραμμικού SVM, όπου οι κλάσεις y_i είναι 1 ή -1, κάθε υπερεπίπεδο μπορεί να περιγραφεί από το σύνολο σημείων x που ικανοποιούν τη



Σχήμα 3: Παράδειγμα υπερεπιπέδων. Το H_1 δεν διαχωρίζει τις κλάσεις. Το H_2 τις διαχωρίζει αλλά με μικρό κενό. Το H_3 τις διαχωρίζει με το μέγιστο κενό.

σχέση:

$$w \cdot x - b = 0$$

όπου w είναι το κάθετο διάνυσμα στο υπερεπίπεδο. Για την εύρεση του υπερεπιπέδου μεγίστου κενού επιδιώκεται η ελαχιστοποίηση του $\|w\|$ δεδομένου ότι $y_i(w \cdot x_i - b) \geq 1$. Τα w και b που λύνουν αυτό το πρόβλημα καθορίζουν τον ταξινομητή.

Στην περίπτωση μη γραμμικών προβλημάτων, το εσωτερικό γινόμενο αντικαθίσταται από μια μη-γραμμική συνάρτηση πυρήνα και αυτό επιτρέπει να ευρεθεί ένα υπερεπίπεδο μεγίστου κενού σε ένα μετασχηματισμένο feature space.

2.4.4 Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα είναι μια κατηγορία βαθιών (deep), εμπρόσθιας τροφοδότησης (feed-forward) νευρωνικών δικτύων. Η πιο συνηθισμένη χρήση τους αφορά την ανάλυση ψηφιακής εικόνας, αλλά εφαρμόζονται με επιτυχία σε πολλά άλλα προβλήματα όπως αναγνώριση φωνής και επεξεργασία φυσικής γλώσσας.

Ένα συνελικτικό νευρωνικό δίκτυο αποτελείται από ένα στρώμα εισόδου (input layer), ένα στρώμα εξόδου (output layer) και πολλαπλά κρυφά στρώματα (hidden layers) τα πιο κοινά των οποίων είναι τα συνελικτικά στρώματα (convolution layers) και τα στρώματα συσσώρευσης (pooling layers).

2.4.4.1 Συνελικτικό στρώμα

Το συνελικτικό στρώμα είναι η βασική δομική μονάδα των συνελικτικών νευρωνικών δικτύων. Κατά το εμπρόσθιο πέρασμα από το στρώμα, η είσοδος συνελίσσεται με μια σειρά από φίλτρα κατά μήκος του πλάτους και του ύψους της και για κάθε φίλτρο παράγεται ένας χάρτης χαρακτηριστικών (feature map). Τα υπό μάθηση βάρη των φίλτρων αποτελούν παραμέτρους του συνελικτικού στρώματος. Για μια δισδιάστατη συνέλιξη, η λειτουργία του συνελικτικού στρώματος για κάθε φίλτρο μπορεί να περιγραφεί από τον παρακάτω μαθηματικό τύπο:

$$C(i, j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W_f(m, n) \cdot I(i - m, j - n)$$

όπου f είναι ο δείκτης του φίλτρου, W_f το διάνυσμα των βαρών του με διαστάσεις $M \times N$. Συνήθως στο αποτέλεσμα αυτό εφαρμόζεται έπειτα μια μη γραμμική συνάρτηση ενεργοποίησης, με την πιο διαδεδομένη να είναι η ReLU. Το φίλτρο μπορεί είτε να κυλίεται πάνω σε κάθε θέση της εισόδου, ή να προσπερνά κάποιες θέσεις, κάτι που ορίζεται από την παράμετρο που ονομάζεται βήμα (stride). Επίσης, γύρω από τα όρια της εισόδου συχνά απαιτείται η χρήση γεμίσματος (padding) με μηδενικά έτσι ώστε να μη μειώνονται οι διαστάσεις της εξόδου. Το βήμα, το γέμισμα και ο αριθμός των φίλτρων αποτελούν τις υπερπαραμέτρους του συνελικτικού στρώματος.

Δίνοντας είσοδο διαστάσεων $W \times H$ σε συνελικτικό στρώμα με φίλτρα διαστάσεων $M \times N$, με βήμα = S και P ποσότητα γεμίσματος, οι διαστάσεις της εξόδου του μπορούν να υπολογιστούν ως εξής:

$$W_c = \frac{W - M + 2P}{S} + 1, \quad H_c = \frac{H - N + 2P}{S} + 1$$

2.4.4.2 Στρώμα Συσσώρευσης

Η χρήση στρώματος συσσώρευσης ανάμεσα σε συνελικτικά στρώματα είναι κοινή πρακτική στην αρχιτεκτονική των συνελικτικών νευρωνικών δικτύων. Σκοπός του είναι η μείωση της διάστασης του μοντέλου, του αριθμού των παραμέτρων και του υπολογιστικού φόρτου που απαιτείται οδηγώντας σε καλύτερη απόδοση και αποφυγή της υπερεκπαίδευσης (overfitting). Στο στρώμα συσσώρευσης, η έξοδος του προηγούμενου στρώματος υφίσταται δειγματοληψία (downsampling) με τη χρήση μιας συνάρτησης συσσώρευσης σε υποπεριοχές της. Πιο συνηθισμένη είναι η συνάρτηση μεγίστου (max pooling), ενώ άλλες

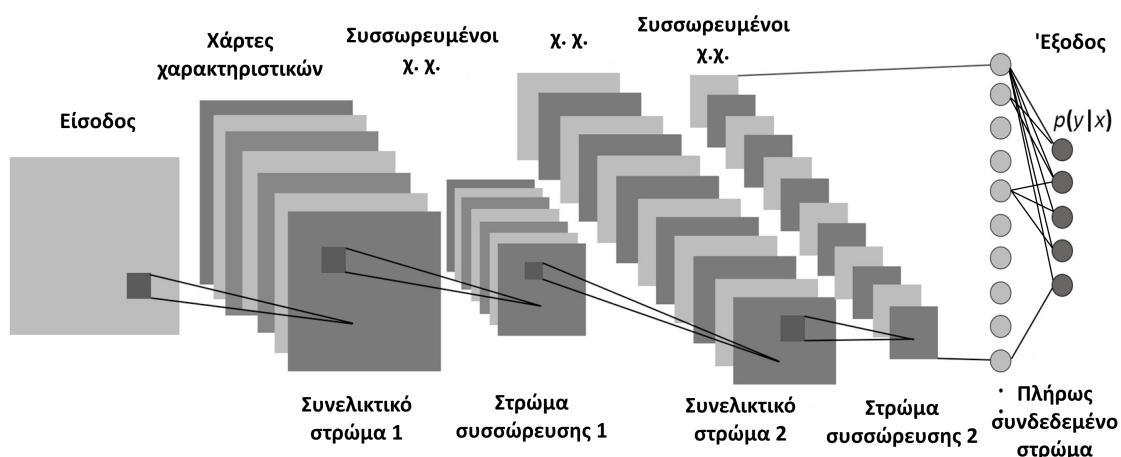
συνηθισμένες συναρτήσεις είναι αυτή του μέσου (average pooling) και του αθροίσματος (sum pooling).

Το στρώμα συσσώρευσης δεν περιέχει παραμέτρους προς μάθηση, καθώς εφαρμόζει μια προκαθορισμένη συνάρτηση στην είσοδο του. Έχει δύο υπερπαραμέτρους, τη διάσταση και το βήμα του φίλτρου. Η χρήση γεμίσματος με μηδενικά δεν συνηθίζεται.

Έχοντας ένα στρώμα συσσώρευσης με φίλτρο διαστάσεων $M \times N$ και βήμα S σε ένα δίκτυο δισδιάστατης συνέλιξης, τότε αν αυτό δεχθεί είσοδο διαστάσεων $W \times H$ θα παράξει για κάθε χάρτη χαρακτηριστικών της εισόδου μία έξοδο διαστάσεων $W_p \times H_p$ όπου:

$$W_p = \frac{W - M}{S} + 1, \quad H_p = \frac{H - N}{S} + 1$$

2.4.4.3 Αρχιτεκτονική δικτύου



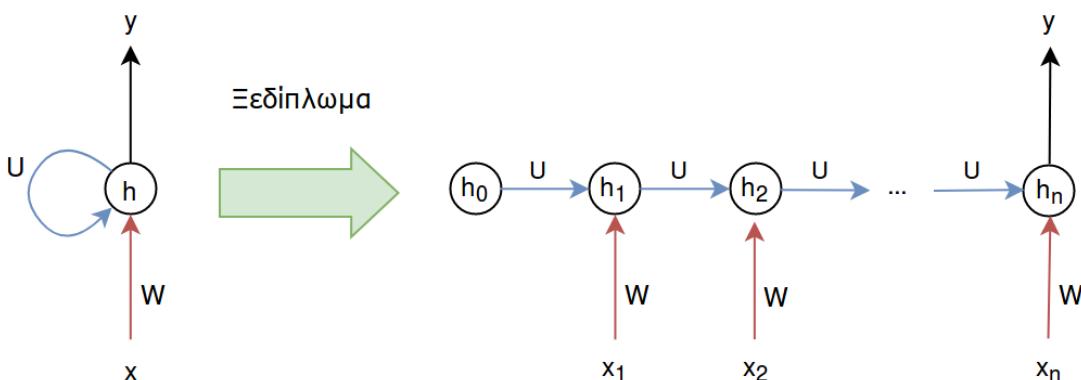
Σχήμα 4: Δομή συνελικτικού νευρωνικού δικτύου.

Η πιο συνηθισμένη δομή του συνελικτικού νευρωνικού δικτύου είναι η επαναλαμβανόμενη αλληλουχία ενός συνελικτικού στρώματος και ενός στρώματος συσσώρευσης μέχρι η διάσταση της αρχικής εισόδου να μειωθεί αρκετά. Με τη χρήση των συνελικτικών στρωμάτων, εξάγονται από την είσοδο χάρτες χαρακτηριστικών, καθένας από τους οποίους μαθαίνει να εστιάζει σε κάποιο διαφορετικό χαρακτηριστικό, όπως για παράδειγμα γωνίες ή καμπύλες σε εικόνα. Στη συνέχεια το στρώμα συσσώρευσης μειώνει τις διαστάσεις κάθε χάρτη χαρακτηριστικών. Εκτός αυτού, δίνει τη δυνατότητα στο δίκτυο να μην επηρεάζεται από μικρές μετατοπίσεις και διαστρεβλώσεις της εισόδου, καθώς η έξοδος του στρώματος συσσώρευσης που λαμβάνει υπόψιν τη μέγιστη/μέση τιμή σε μία υποπεριοχή δειγματοληψίας θα είναι σχεδόν ίδια.

Μετά από κάποιο αριθμό συνελικτικών και συσσωρευτικών στρωμάτων, το δίκτυο συνήθως χρησιμοποιεί ένα πλήρως συνδεδεμένο στρώμα νευρώνων (fully connected layer). Οι νευρώνες στο πλήρως συνδεδεμένο στρώμα έχουν συνδέσεις με κάθε στοιχείο του προηγούμενου στρώματος και οι ενεργοποιήσεις του υπολογίζονται με πολλαπλασιασμό πινάκων της εισόδου με τον πίνακα βαρών και έπειτα πρόσθεση κατωφλίου (bias).

2.4.5 Αναδρομικά Νευρωνικά Δίκτυα

Τα αναδρομικά νευρωνικά δίκτυα είναι μια κατηγορία νευρωνικών δικτύων που επεξεργάζονται ακολουθιακά δεδομένα. Αυτό τα κάνει ιδανικά για προβλήματα όπως αναγνώριση ομιλίας, αναγνώριση γραφικού χαρακτήρα και γενικώς προβλήματα επεξεργασίας φυσικής γλώσσας.



Σχήμα 5: Δομή αναδρομικού νευρωνικού δικτύου. Με τον όρο "ξεδίπλωμα" αναφερόμαστε στην παρουσίαση του για ολόκληρη την ακολουθία. Για παράδειγμα αν η ακολουθία περιέχει 5 στοιχεία το δίκτυο ξεδιπλώνεται σε 5 στρώματα.

Σε αντίθεση με τα δίκτυα εμπρόσθιας τροφοδότησης, τα αναδρομικά νευρωνικά δίκτυα διαθέτουν μία "μνήμη" που περιέχει πληροφορία για ό,τι έχει προηγηθεί μέχρι εκείνη τη στιγμή. Με βάση αυτή τη μνήμη, η έξοδος του επόμενου στοιχείου της ακολουθίας εξαρτάται από τους προηγούμενους υπολογισμούς που πραγματοποιήθηκαν για τα στοιχεία της ακολουθίας αυτής. Αν h_t είναι η έξοδος ή κρυφή κατάσταση (hidden state) για το στοιχείο x_t της ακολουθίας τη χρονική στιγμή t και ϕ η αναδρομική συνάρτηση του δικτύου τότε η λειτουργία του δικτύου περιγράφεται από τη σχέση:

$$h_t = \phi(x_t, h_{t-1})$$

Η αναδρομική συνάρτηση ϕ περιέχει τις παραμέτρους υπό μάθηση του δικτύου. Σε ένα απλό αναδρομικό δίκτυο (δίκτυο Elman) αυτές είναι τα βάρη W , U και το

κατώφλι b . Το βάρος W μετασχηματίζει την προηγούμενη κρυφή κατάσταση του δικτύου, ενώ το βάρος U εφαρμόζεται στην τρέχουσα είσοδο της ακολουθίας. Με βάση αυτά και αν f μια μη γραμμική συνάρτηση ενεργοποίησης, η εξίσωση ενός απλού αναδρομικού δικτύου είναι η εξής:

$$h_t = f(Wx_t + Uh_{t-1} + b)$$

2.4.5.1 Το πρόβλημα της εξαφανιζόμενης κλίσης

Όμως, τα απλά αναδρομικά νευρωνικά δίκτυα δεν είναι αποτελεσματικά στη μοντελοποίηση εξαρτήσεων μεγάλης διαρκείας λόγω του προβλήματος της εξαφανιζόμενης κλίσης (vanishing gradient problem). Ο αλγόριθμος backpropagation αναζητάει το ελάχιστο της συνάρτησης σφάλματος σε σχέση με τα βάρη W και U χρησιμοποιώντας τη μέθοδο της στοχαστικής κατάβασης δυναμικού (stochastic gradient descent). Έτσι, ο backpropagation αλγόριθμος πρέπει να υπολογίσει την κλίση του σφάλματος $Loss$ σε σχέση με το βάρος U και η ολική κλίση θα είναι το άθροισμα των κλίσεων σε κάθε χρονική στιγμή της ακολουθίας εκπαίδευσης. Αν h_t η κρυφή κατάσταση του αναδρομικού δικτύου όπως ορίστηκε παραπάνω και $y_t = Vh_t$ η έξοδος την τελευταία χρονική στιγμή t , η κλίση του σφάλματος είναι:

$$\frac{\partial Loss}{\partial U} = \sum_{k=1}^t \frac{\partial Loss}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial U}$$

Όμως, το h_t εξαρτάται από το h_{t-1} , το οποίο εξαρτάται από το h_{t-2} και ούτω καθεξής. Επομένως, ο όρος $\frac{\partial h_t}{\partial h_k}$ υπολογίζεται με τον κανόνα αλυσίδας ως εξής:

$$\frac{\partial h_t}{\partial h_k} = \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_{k+1}}{\partial h_k} = \prod_{i=k+1}^t \frac{\partial h_i}{\partial h_{i-1}} \quad (2.1)$$

Η παράγωγος του h_{k-1} ως προς το h_k , αν η $diag$ μετατρέπει ένα διάνυσμα σε διαγώνιο πίνακα, θα είναι:

$$\frac{\partial h_i}{\partial h_{i-1}} = U^T diag(f'(h_{i-1})) \quad (2.2)$$

Άρα από 2.1 και 2.2 έχουμε:

$$\frac{\partial h_t}{\partial h_k} = \prod_{i=k+1}^t U^T diag(f'(h_{i-1}))$$

Για την ανάλυση του γινομένου των Ιακωβιανών πινάκων, χρησιμοποιείται η επαναληπτικού υπολογισμού (power iteration method). Η παράγωγος της συνάρτησης ενεργοποίησης f' στην ιδιοτιμή $\|diag(f'(h_{i-1}))\|$ έχει ως άνω όριο γ_h (1 για την $tanh$ και 0.25 για τη σιγμοειδή) και η ιδιοτιμή $\|U^T\|$ έχει ως άνω όριο το γ_U . Τότε η 2-νόρμα του Ιακωβιανού πίνακα που υπολογίζεται από την 2.2 έχει ως άνω όριο το γινόμενο των νορμών των δύο πινάκων U^T και $diag(f'(h_{i-1}))$ δηλαδή:

$$\left\| \frac{\partial h_i}{\partial h_{i-1}} \right\| \leq \|U^T\| \|diag(f'(h_{i-1}))\| \leq \gamma_U \gamma_h$$

Παρατηρούμε ότι για τιμές $\gamma_U < \frac{1}{\gamma_h}$ θα ισχύει $\left\| \frac{\partial h_i}{\partial h_{i-1}} \right\| < 1$. Τότε η νόρμα της 2.1 που μπορεί να εκφραστεί ως:

$$\left\| \frac{\partial h_t}{\partial h_k} \right\| = \left\| \prod_{i=k+1}^t \frac{\partial h_i}{\partial h_{i-1}} \right\| \leq (\gamma_U \gamma_h)^{t-k}$$

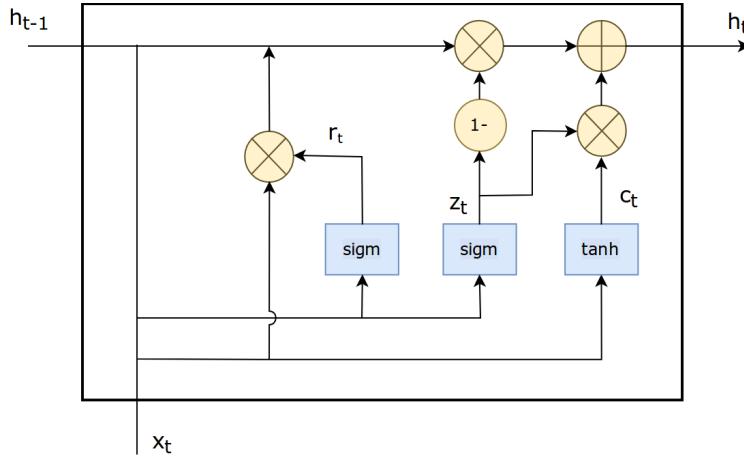
Θα μειώνεται συνεχώς και με εκθετικούς ρυθμούς θα γίνει 0. Με την ίδια λογική αν τα βάρη έχουν μεγαλύτερες τιμές και $\gamma_U > \frac{1}{\gamma_h}$, η κλίση θα εκτοξευθεί και δημιουργείται το πρόβλημα της εκτοξεύομενης κλίσης. Αυτά τα προβλήματα αποτρέπουν την είσοδο στην παρελθοντική χρονική στιγμή k από το να έχει επιρροή στην έξοδο την παρούσα χρονική στιγμή t .

2.4.5.2 Περιφραγμένη αναδρομική μονάδα

Για την επίλυση του προβλήματος της εξαφανιζόμενης κλίσης, προτάθηκε από τον Cho και τους συνεργάτες του [32] ένα είδος αναδρομικού νευρωνικού δικτύου που το ονόμασαν περιφραγμένη αναδρομική μονάδα (gated recurrent unit ή GRU). Το GRU μπορεί να θεωρηθεί ως μια παραλλαγή του δικτύου μακράς-βραχείας μνήμης (long short-term memory ή LSTM) που χρησιμοποιείται για τον ίδιο σκοπό καθώς έχουν ομοιότητες στο σχεδιασμό τους και παρόμοια επίδοση.

Για την επίλυση του προβλήματος της εξαφανιζόμενης κλίσης, το GRU χρησιμοποιεί τη θύρα ενημέρωσης (update gate) z_t και τη θύρα επαναφοράς (reset gate) r_t . Οι εξισώσεις κατάστασης που χαρακτηρίζουν μια μονάδα GRU είναι οι εξής:

$$\begin{aligned} r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \\ z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \\ c_t &= \tanh(W_{xc}x_t + r_t \odot (W_{hc}h_{t-1}) + b_c) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot c_t \end{aligned}$$



Σχήμα 6: Απεικόνιση της περιφραγμένης αναδρομικής μονάδας.

όπου x_t η είσοδος τη χρονική στιγμή t , b το διάνυσμα πόλωσης, \odot το γινόμενο Hadamard, σ η σιγμοειδής συνάρτηση και \tanh η υπερβολική εφαπτομένη. Η λειτουργία του μοντέλου περιγράφεται ως εξής. Η θύρα ενημέρωσης z_t ελέγχει πόση παρελθοντική πληροφορία χρειάζεται να διατηρηθεί και να προωθηθεί στα επόμενα βήματα της ακολουθίας. Η δε θύρα επαναφοράς r_t χρησιμοποιείται για να διευκρινιστεί από το μοντέλο πόση παρελθοντική πληροφορία δεν χρησιμεύει στην πρόβλεψη και άρα πρέπει να ξεχαστεί. Το τρέχον περιεχόμενο μνήμης c_t χρησιμοποιεί την πύλη επαναφοράς για να αποθηκεύσει την χρήσιμη παρελθοντική πληροφορία, αφαιρώντας την περιττή πληροφορία μέσω του γινομένου Hadamard. Τέλος το διάνυσμα h_t περιέχει την πληροφορία της τρέχουσας μονάδας και την προωθεί στο δίκτυο. Για το σκοπό αυτό χρειάζεται η θύρα ενημέρωσης που καθορίζει τι θα κρατηθεί από την τρέχουσα μνήμη c_t και τι από τη μνήμη των προηγούμενων βημάτων h_{t-1} . Το μοντέλο μπορεί να εκπαιδευτεί να θέτει το διάνυσμα z_t κοντά στο 0 και έτσι από την εξίσωση φαίνεται πως αφού το $(1 - z_t)$ θα είναι κοντά στο 1 θα κρατάει την περισσότερη παλιά πληροφορία, αλλά μικρό μόνο τμήμα της τρέχουσας.

2.5 Μετρικές αξιολόγησης

Για την αξιολόγηση των συστημάτων αποσαφήνισης χρησιμοποιούνται datasets κειμένων στα οποία η πληροφορία αποσαφήνισης είναι διαθέσιμη αφού έχει βρεθεί και επαληθευτεί από ανθρώπινο παράγοντα (gold standard). Οι πιο δημοφιλείς μετρικές αξιολόγησης της επίδοσης του συστήματος είναι οι precision, recall και F_1 score (γνωστό και ως F-score ή F-measure). Ο υπολογισμός αυτών των μετρικών χρησιμοποιεί τα παρακάτω σύνολα:

- True Positive (TP), είναι το σύνολο των αναφορών εκείνων που αντιστοιχή-θηκαν σωστά στην οντότητα που ορίζει το gold standard.
- False Postitive (FP), είναι το σύνολο των αναφορών που αντιστοιχήθηκαν σε λάθος οντότητα από αυτή που ορίζει το gold standard.
- True Negative (TN), είναι το σύνολο των αναφορών των οποίων η οντότητα καλώς αναγνωρίστηκε ότι δεν βρίσκεται στο σύνολο οντοτήτων του dataset.
- False Negative (FN), είναι το σύνολο των αναφορών των οποίων η οντότητα που ορίζει το gold standard δεν επιστράφηκε από το σύστημα.

Με βάση αυτά, μπορούμε να ορίσουμε τους τύπους των precision, recall και F_1 score. Η μετρική precision λαμβάνει υπόψη όλες τις αναφορές που αντιστοιχήθηκαν σε κάποια οντότητα και δίνει το ποσοστό αυτών που αποσαφηνίστηκαν σωστά.

$$\text{precision} = \frac{TP}{TP + FP}$$

Η μετρική recall χρησιμοποιείται συνήθως μαζί με την precision και δίνει το ποσοστό των σωστά αποσαφηνισμένων αναφορών από το σύνολο των αναφορών που μπορούσαν να αποσαφηνιστούν.

$$\text{recall} = \frac{TP}{TP + FN}$$

Αυτές οι δύο μετρικές μπορούν να συνδυαστούν και να δώσουν το F_1 score που είναι ο αρμονικός μέσος των precision και recall.

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Πολλά συστήματα αποσαφήνισης μπορούν να πάρουν ως είσοδο τις ονοματικές αναφορές προς αποσαφήνιση, έτσι ώστε ο αριθμός των αναφορών που αποσαφηνίστηκαν να ισούται με τον αριθμό των αναφορών που χρήζουν αποσαφήνισης. Σε αυτές τις περιπτώσεις αρκεί η χρήση μιας μόνο μετρικής αξιολόγησης. Η μετρική αυτή ονομάζεται accuracy και υπολογίζει το ποσοστό των σωστά αποσαφηνισμένων αναφορών από όλες τις αναφορές προς αποσαφήνιση. Άρα, τότε ισχύει:

$$\text{accuracy} = \text{recall} = F_1 = \text{precision}$$

2.6 Σχετικές Εργασίες

Σε αυτό το σημείο έχει καταστεί σαφές το θέμα ενασχόλησης της διπλωματικής εργασίας και θεωρείται σκόπιμο να γίνει αναφορά σε σχετικές εργασίες αποσαφήνισης οντοτήτων που αποτέλεσαν πηγή έμπνευσης, ή αντικείμενο σύγκρισης του συστήματος που υλοποιείται σε αυτή την εργασία.

2.6.1 Berkeley Entity Resolution System

To Berkeley Entity Resolution System παρουσιάζεται στην εργασία [33] των Durrett και Klein. Ασχολείται με τρεις πτυχές της ανάλυσης οντοτήτων: επίλυση συναναφοράς (coreference resolution), αναγνώριση ονοματικών οντοτήτων και αποσαφήνιση ονοματικών οντοτήτων. Το μοντέλο αυτό πρόκειται για ένα υπό συνθήκη τυχαίο πεδίο (conditional random field). Μοναδιαίοι συντελεστές (unary factors) κωδικοποιούν τοπικά χαρακτηριστικά για κάθε ένα από τα τρία υποπροβλήματα, ενώ δυαδικοί και υψηλότερης τάξης συντελεστές εντοπίζουν αλληλεπιδράσεις μεταξύ των υποπροβλημάτων, όπως το γεγονός ότι συναναφορικές αναφορές έχουν κοινό σημασιολογικό τύπο.

2.6.2 AIDA-light

To AIDA-light παρουσιάζεται στην εργασία [34] του Nguyen και των συνεργατών του. Χρησιμοποιεί έναν αλγόριθμο χαρτογράφησης δύο σταδίων. Αρχικά, προσδιορίζει ένα σύνολο "εύκολων" αναφορών με χαμηλή αμφισημότητα και τις συνδέει αποτελεσματικά με οντότητες. Αυτό το στάδιο καθορίζει επίσης το θεματικό τομέα του κειμένου που αποτελεί σημαντικό χαρακτηριστικό του μοντέλου. Στη συνέχεια, το δεύτερο στάδιο χρησιμοποιεί την αξιόπιστη διασύνδεση των "εύκολων" αναφορών για να δημιουργήσει ένα σύνολο από συμφραζόμενα για την αποσαφήνιση των υπόλοιπων αναφορών.

2.6.3 Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks

Σε αυτή την εργασία του Francis και των συνεργατών του [35] παρουσιάζεται ένα μοντέλο αποσαφήνισης οντοτήτων που χρησιμοποιεί συνελικτικά νευρωνικά δίκτυα προκειμένου να εντοπίσει τη σημασιολογική συνοχή μεταξύ των συμφραζόμενων μιας αναφοράς και των υποψήφιων οντοτήτων της. Αυτά τα συνελικτικά

δίκτυα λειτουργούν σε πολλαπλά επίπεδα για να εκμεταλλευτούν διάφορα είδη πληροφορίας και έχουν τη δυνατότητα να μάθουν ποια n-grams είναι χαρακτηριστικά διαφορετικών θεμάτων. Στη συνέχεια, τα συνελικτικά δίκτυα συνδέονται με ένα γραμμικό μοντέλο αραιών χαρακτηριστικών.

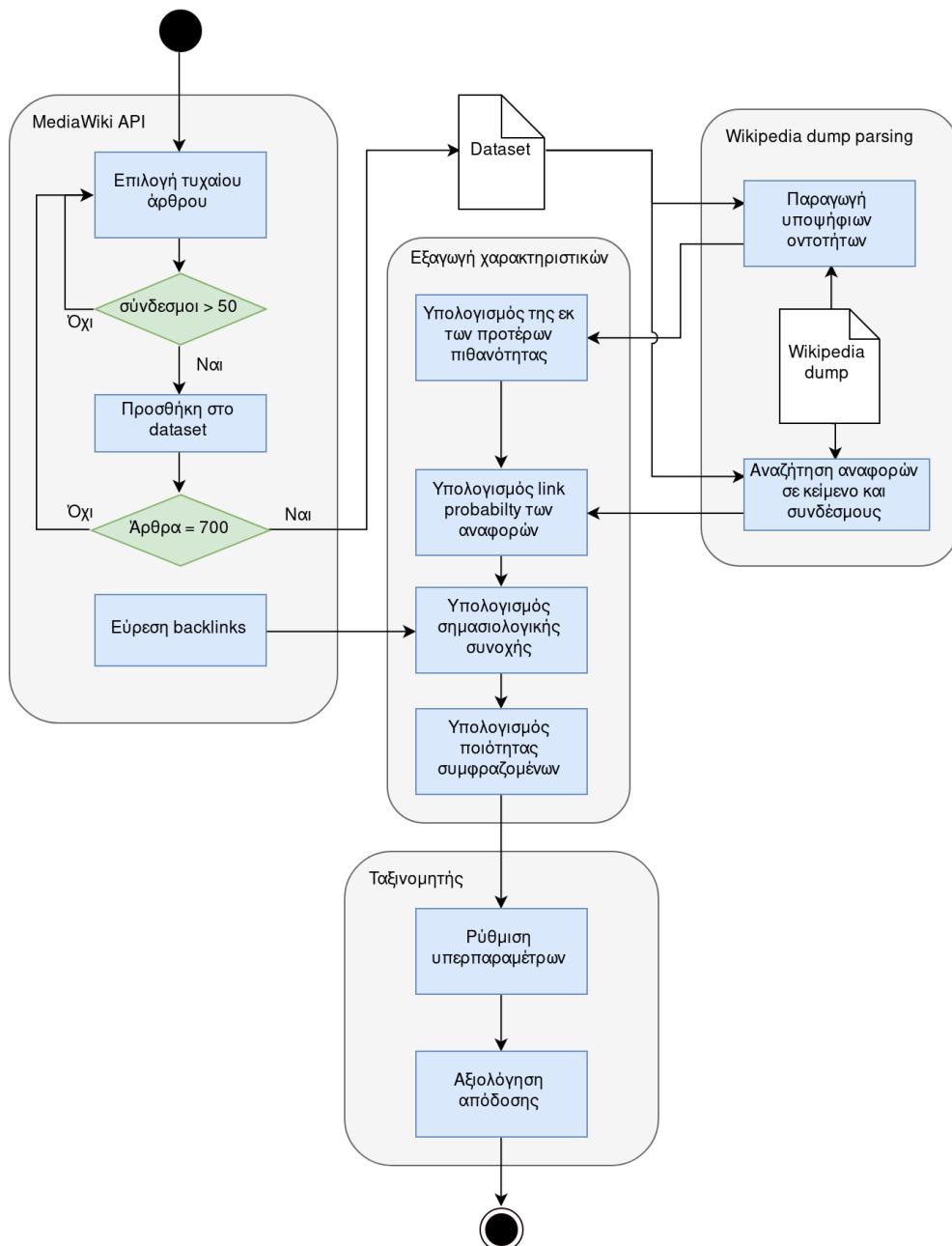
Κεφάλαιο 3

Συστήματα Αποσαφήνισης Οντοτήτων

Στο κεφάλαιο αυτό θα γίνει γενική περιγραφή των δύο συστημάτων που υλοποιήθηκαν, της γενικής λειτουργίας τους, των συγκεκριμένων σκοπών που εξυπηρετούν και της σημασίας εκπλήρωσής τους. Επιπλέον θα δοθεί η ανάλυση των κρίσιμων αποφάσεων σχετικά με τα μέσα που χρησιμοποιήθηκαν, τις αρχιτεκτονικές επιλογές και τον τρόπο χειρισμού επιμέρους εργασιών. Λόγω των διαφορετικών προσεγγίσεων, κρίθηκε απαραίτητο η περιγραφή να χωριστεί σε δύο υποενότητες, μια για κάθε σύστημα.

3.1 Βασικό μοντέλο

Το πρώτο σύστημα που υλοποιήθηκε χρησιμοποιεί την πληροφορία της Wikipedia για να εξάγει χαρακτηριστικά για τις αναφορές και τις υποψήφιες οντότητες. Το σύστημα αυτό είναι βασισμένο στο μοντέλο που χρησιμοποιήθηκε στην εργασία των Milne και Witten [29] αποτελεί τμήμα ενός συνδυαστικού μοντέλου αναγνώρισης και αποσαφήνισης ονοματικών οντοτήτων. Επειδή η παρούσα εργασία ασχολείται με το θέμα αποσαφήνισης οντοτήτων, θα αναπαραχθεί αυτό το κομμάτι του μοντέλου και θα αξιολογηθεί σε σύγχρονο dataset.



Σχήμα 7: Σχεδιάγραμμα της αρχιτεκτονικής του βασικού μοντέλου.

3.1.1 Δημιουργία Dataset

Το τμήμα της αποσαφήνισης του συστήματος αυτού χρησιμοποιεί τους συνδέσμους της Wikipedia ως ονομαστικές οντότητες προς αποσαφήνιση. Για τη δημιουργία του dataset, χρησιμοποιήθηκε το Mediawiki API. Επιλέχθηκαν 700 τυχαία άρθρα από τη Wikipedia, με την προϋπόθεση να έχουν πάνω από 50 συνδέσμους, έτσι ώστε να υπάρχουν αρκετά δεδομένα για την εκπαίδευση και την αξιολόγηση του συστήματος. Πριν την καταμέτρηση των συνδέσμων κρίθηκε σκόπιμο να να αφαιρεθούν λίστες και πίνακες καθώς συχνά η ύπαρξη τους οδηγούσε σε επιλογή σελίδων με ελάχιστο αδόμητο κείμενο. Από αυτά τα άρθρα, 600 χρησιμοποιήθηκαν για την εκπαίδευση (training set) και 100 για την αξιολόγηση (test set). Στο dataset, τα άρθρα της Wikipedia βρίσκονται σε μορφή Wikitext, που χρησιμοποιεί συγκεκριμένη σύνταξη και λέξεις κλειδιά για τη διαμόρφωση του κειμένου.

Τμήμα κώδικα 1: Παράδειγμα άρθρου της Wikipedia σε μορφή Wikitext

```
 {{Use dmy dates|date=January 2018}}
 {{Use British English|date=January 2018}}
 {{Refimprove|date=April 2011}}
 [[Image:David Hughes.jpg|thumb|David Hughes (Geoffrey Paddison)]]

'''David Hughes''' (born '''Geoffrey Paddison'''; 11 October 1925&nbsp;- 19 October 1972)<ref name="British Hit Singles & Albums"/> was an English [[popular music|pop]] and opera singer.

==The popular tenor==
Paddison was born in [[Bournbrook]], [[Birmingham]], [[England]] to an English mother and Welsh father. As a child he listened to [[gramophone record|records]] by [[Enrico Caruso|Caruso]]. He found work as a railway clerk at Curzon Street, Birmingham. During his time there he was invited to sing "[[On the Road to Mandalay (song)|On the Road to Mandalay]]" at an office [[concert]]. This was so well received that he started taking professional singing lessons. He also was an RAF cadet and took flying lessons at Wythall airfield near Birmingham. In 1945 he joined the [[Royal Air Force|RAF]]. The war in Europe was over and he was despatched on a ship to the Pacific which was intended to construct an airfield in support of the war against Japan. Following the fall of Japan his ship was diverted to Hong Kong as part of Operation Ethelred under Rear-Admiral Harcourt in order to receive the surrender of the Japanese. Paddison found himself acting as a policeman. While here he regularly sang on ZBW in [[Kowloon]], the armed forces' [[radio station]]. At this time he was singing [[Bing Crosby]] [[song]]s. After returning to the UK and being demobilised in 1947 he received a grant to study singing at Wigmore Hall, London.
[[File:48 Alton Road.jpg|thumb|48 Alton Road, Bournbrook - David Hughes's birthplace]]
```

After that he studied at the Royal Academy of Music and Dramatic Art. He had an early break in 1948 appearing with [[Ginger Rogers]] and [[Lizbeth Webb]] in ''[[Carissima (musical)|Carissima]]'', a [[West End theatre|West End]] [[musical theatre|musical]]. In 1951 he appeared on [[Henry Hall (bandleader)|Henry Hall]]'s "Guest Night". He was introduced by Hall as "the young Welsh tenor", as his stage name "David Hughes" which was his father's Christian names was typically Welsh. He appeared often in the 1950s on [[television]] and [[radio]]. These shows included "Presenting David Hughes", ''[[Sunday Night at the London Palladium]]'' and 2 series of his own show "Make Mine Music" (1959). In 1954, while touring [[Australia]], he arranged for his fiancée Anne Sullivan to join him there, and were [[marriage|married]]. He appeared in the stage show "Summer Song" in 1956, a biographical musical about [[Antonín Dvořák]]'s visit to the [[United States]]. [[Sally Ann Howes]] was the female lead. In 1956 he had his only [[hit record|hit]] in the [[UK Singles Chart]], "By The Fountains of Rome".<ref name="British Hit Singles & Albums">{{cite book

| first= David
| last= Roberts
| year= 2006
| title= British Hit Singles & Albums
| edition= 19th
| publisher= Guinness World Records Limited
| location= London
| isbn= 1-904994-10-5
| page= 262}}</ref> The composer was [[Mátyás Seiber]] and the lyrics by [[Norman Newell]], who also wrote hits for Ken Dodd ("Promises", 1966), [[Shirley Bassey]] ("Never Never Never", 1973) and [[Matt Monro]] ("Portrait of My Love", 1960).

He also co-starred in the musical Scapa at the Adelphi theatre alongside Pete Murray and Edward Woodward. He participated in the [[United Kingdom|UK]] heat of the [[Eurovision Song Contest]] in 1960, [[A Song For Europe]], finishing in second place with the song "Mi Amor".

==The opera singer==

Shortly after appearing in the musical "Seagulls Over Sorrento" (1962) and whilst appearing in a summer season show at Torquay he had his first heart attack.

Whilst recovering he decided to undertake his first love and retrain as an opera singer. This was a major undertaking and meant starting at the bottom.

his first role was as the high priest in Idomeneo at Glyndebourne in 1963. [[opera]]. He sang at [[Glyndebourne]] and with [[Sadler's Wells]], English National and [[Welsh National]] operas. [[John Barbirolli|Sir John Barbirolli]] conducted Verdi's "Requiem" several times, with Hughes singing. He earned a reputation as a thorough professional, popular with colleagues. His most famous role was as Don José in Carmen, a role he performed over 100 times. He also performed the role of Lieutenant Pinkerton in "Madama Butterfly" many times.'

He never forgot his fans of popular music, continuing to do Sunday concerts in theatres all over the UK. His popularity also saw him presenting his show "Make Mine Music" with the Midland Light Orchestra on BBC Radio.

He was on "Desert Island Discs" twice . On 21st May 1956 as a pop singer and on the 23rd November 1970 as a tenor.

On 8 October 1972, he fell ill while singing the part of Pinkerton at the [[London Coliseum]]. He collapsed in the wings near the end but managed to complete the final scene. He died the following day, from [[heart failure]]. Just before the ambulance men took him out of the theatre he said "I didn't let them down, did I?"{{Citation needed|date=June 2008}}

==Namesake==

- * He chose his father's two Christian names to use as his stage name. His father was David Hughes Paddison.

==Discography==

====Popular songs====

- * "By the Fountains of Rome" ([[single (music)|single]]) (1956) [[Record chart|Charted]] at #27
- * "'Here in My Heart'" ([[compilation album|compilation]])
- * "'The Best of David Hughes'" (compilation)
- * "'Great British Song Stylist'" (compilation)

====Stage musicals====

- * "Summer Song" (1956)
- * "Plain and Fancy"
- * "Here in My Heart"

====Classical music====

- * "16th-18th Century Songs of Love"
- * "Favourite Opera/ Operetta Arias and Songs You Love" EMI TWO 319, City of Birmingham Symphony Orchestra, conducted by Louis Fremaux, recorded in Birmingham Town Hall
- * "The Merry Widow"

==References==

{{Reflist}}

==External links==

- * [<http://www.gmmy.com/tenors/hughes.htm> Newspaper Cuttings]
- * [<http://www.davidhughestenor.co.uk/> Popular singer]
- * BBC Desert Island Discs <http://www.bbc.co.uk/programmes/p009y9cq>
- * BBC Desert Island Discs <http://www.bbc.co.uk/programmes/p009ndjy>

{{Authority control}}

{{DEFAULTSORT:Hughes, David}}
[[Category:1925 births]]
[[Category:1972 deaths]]
[[Category:English opera singers]]
[[Category:English tenors]]
[[Category:English people of Welsh descent]]
[[Category:English pop singers]]
[[Category:Alumni of the Royal Academy of Music]]
[[Category:20th-century English singers]]
[[Category:20th-century opera singers]]
[[Category:Royal Air Force airmen]]

Τμήμα κώδικα 2: Το άρθρο 1 μετά από αφαίρεση πινάκων και λιστών

[[Image:David Hughes.jpg|thumb|David Hughes (Geoffrey Paddison)]]

'''David Hughes''' (born '''Geoffrey Paddison'''; 11 October 1925&nbs;- 19 October 1972)<ref name="British Hit Singles & Albums"/> was an English-born [[popular music|popular]] [[singer]] of Welsh extraction who became an opera singer.

==The popular tenor==

Paddison was born in [[Bournbrook]], [[Birmingham]], [[England]] of [[Welsh people|Welsh]] parents. As a child he listened to [[gramophone record|records]] by [[Enrico Caruso|Caruso]]. He found work as a clerk, and was invited to sing "[[On the Road to Mandalay (song)|On the Road to Mandalay]]" at an office [[concert]]. This was so well received that he started taking professional singing lessons. In 1945 he joined the [[Royal Air Force|RAF]] and sang on ZBW in [[Kowloon]], the armed forces' [[radio station]]. At this time he was singing [[Bing Crosby]] [[song]]s. In 1947 he received a grant to study singing at Wigmore Hall.

[[Image:48 Alton Road.jpg|thumb|48 Alton Road, Bournbrook - David Hughes's birthplace]]

After that he studied at the Royal Academy of Music and Dramatic Art. He had an early break in 1948 appearing with [[Ginger Rogers]] and [[Lizbeth Webb]] in ''[[Carissima (musical)|Carissima]]'', a [[West End theatre|West End]] [[musical theatre|musical]]. In 1951 he appeared on [[Henry Hall (bandleader)|Henry Hall]]'s "Guest Night". He was introduced by Hall as "the young Welsh tenor". This prompted him to take his stage name "David Hughes", an archetypical Welsh name. He appeared often in the 1950s on [[television]] and [[radio]]. These shows included ''Presenting David Hughes'', ''[[Sunday Night at the London Palladium]]'' and ''Make Mine Music'' (1959). In 1954, while touring [[Australia]], he met and [[marriage|married]] Ann Sullivan. He appeared in the stage show ''Summer Song'' in 1956, a biographical musical about [[Antonín Dvořák]]'s visit to the [[United States]]. [[Sally Ann Howes]] was the female lead. In 1956 he had his only [[hit record|hit]] in the [[UK Singles Chart]], "By The Fountains of Rome".<ref name="British Hit Singles & Albums">{{cite book

==The opera singer==

Shortly after appearing in the musical ''Seagulls Over Sorrento'' (1962) he moved into [[opera]]. He sang at [[Glyndebourne]] and with [[Sadler's Wells]] and [[Welsh National]] operas. [[John Barbirolli|Sir John Barbirolli]] conducted Verdi's "Requiem" several times, with Hughes singing. He earned a reputation as a thorough professional, popular with colleagues. His most famous role was as Lieutenant Pinkerton in ''[[Madama Butterfly|Madam Butterfly]]''.

On 8 October 1972, he fell ill while singing the part of Pinkerton at the [[London Coliseum]]. He collapsed in the wings near the end but managed to complete the final scene. He died the following day, from [[heart failure]]. Just before the ambulance men took him out of the theatre he said "I didn't let them down, did I?"{{Citation needed|date=June 2008}}

==Namesake==

==Discography==

==Popular songs==

==Stage musicals==

```

====Classical music====

==References==

==External links==

[[Category:1925 births]]
[[Category:1972 deaths]]
[[Category:English opera singers]]
[[Category:English tenors]]
[[Category:Welsh opera singers]]
[[Category:Welsh pop singers]]
[[Category:Alumni of the Royal Academy of Music]]
[[Category:Welsh tenors]]
[[Category:20th-century English singers]]
[[Category:20th-century opera singers]]
[[Category:Royal Air Force airmen]]

```

Το κείμενο που προκύπτει μετά από αυτή την επεξεργασία είναι ως επί το πλείστον αδόμητο, οπότε σε αυτή τη φάση ο parser συλλέγει τους συνδέσμους που υπάρχουν σε αυτό χρησιμοποιώντας την κατάλληλη κανονική έκφραση (regular expression) του Wikitext. Από αυτούς τους συνδέσμους αγνοούνται όσοι αναφέρονται σε εικόνες και κατηγορίες της Wikipedia καθώς συνήθως βρίσκονται εκτός του κυρίως σώματος του κειμένου. Στο dataset μας βρέθηκαν συνολικά 43550 σύνδεσμοι, οι οποίοι αποτελούν και τις αναφορές του συστήματος.

3.1.2 Παραγωγή υποψήφιων οντοτήτων

Αφού προσδιορίστηκαν οι αναφορές, πρέπει να παραχθούν τα αντίστοιχα σύνολα υποψήφιων οντοτήτων. Αυτό έγινε εξάγοντας όλες τις πιθανές σελίδες οντοτήτων στις οποίες οδηγούν οι σύνδεσμοι της Wikipedia με anchor text τις αναφορές του dataset από ένα Wikipedia dump. Το dump βρίσκεται σε XML μορφή και περιέχει tags με πληροφορίες όπως το όνομα της σελίδας, namespace, αριθμό αναθέωρησης, κείμενο. Για διευκόλυνση της διαδικασίας αυτής, έγινε αρχικά προεπεξεργασία του dump με τη χρήση ενός XML parsing tool. Συγκεκριμένα, με την προεπεξεργασία αυτή εξάγονται features που αφορούν τους συνδέσμους των σελίδων της Wikipedia από άρθρα οντοτήτων και χρησιμεύει γιατί αφενός εξαιρεί τις σελίδες που είναι περιττές για την εργασία, δηλαδή όσες δεν έχουν namespace=0, και αφετέρου γιατί καθιστά εύκολα προσβάσιμη όλη την πληροφορία που χρειαζόμαστε σε αυτό το στάδιο σε ένα πιο συμπιεσμένο αρχείο. Στο αρχείο μορφής csv που προκύπτει, κάθε σειρά αντιστοιχεί σε ένα σύνδεσμο και περιέχει διαχωρισμένες

με ένα κόμμα πληροφορίες, σχετικές με αυτόν και την σελίδα στην οποία βρίσκεται. Οι πληροφορίες είναι οι εξής: id σελίδας, τίτλος σελίδας, id αναθεώρησης, id γονέα αναθεώρησης, timestamp αναθεώρησης, είδος χρήστη, όνομα χρήστη, id χρήστη, μικρή αναθεώρηση, οντότητα συνδέσμου, anchor text συνδέσμου, όνομα ενότητας συνδέσμου, επίπεδο ενότητας συνδέσμου.

Έχοντας αυτό το αρχείο, είναι εύκολο να βρεθούν όλες οι υποψήφιες οντότητες των αναφορών. Για κάθε γραμμή, αν το anchor text βρίσκεται στο σετ των αναφορών του dataset, η οντότητα στην οποία οδηγεί ο σύνδεσμος προστίθεται στο σύνολο των υποψήφιων οντοτήτων της αναφοράς. Σε αυτό το σημείο υπολογίζεται επίσης ο αριθμός εμφανίσεων κάθε οντότητας για κάποια συγκεκριμένη αναφορά, που θα χρησιμεύσει στο βήμα ταξινόμησης των υποψήφιων οντοτήτων.

Αλγόριθμος 1 Ψευδοκώδικας για τη δημιουργία των υποψήφιων οντοτήτων

Require: *mentions*, the set of mentions in the dataset

Require: *wikidump_links*, the text of the parsed wiki dump xml file

Require: *CsvTokens*, a way to iterate over the tokens of the csv line

```

1: function FindSenses(mentions, wikidump_links)
2:   for each line in wikidump_links do
3:     anchor_text  $\leftarrow$  CsvTokens(line)[10]
4:     if anchor_text  $\in$  mentions then
5:       entity  $\leftarrow$  CsvTokens(line)[9]
6:       if entity  $\notin$  mentions[anchor_text].senses then
7:         mentions[anchor_text].senses.add(entity)
8:         mentions[anchor_text].senses[entity].appearances  $\leftarrow$  1
9:       else
10:        mentions[anchor_text].senses[entity].appearances.incrementBy(1)
11:       end if
12:     end if
13:   end for
14:   return mentions
15: end function

```

3.1.3 Εξαγωγή χαρακτηριστικών

Σε αυτή την ενότητα αναλύονται τα χαρακτηριστικά που χρησιμοποιούνται από το σύστημα για την ταξινόμηση των οντοτήτων μέσα στο σύνολο υποψήφιων οντοτήτων E_m , ώστε να αντιστοιχιστεί η σωστή οντότητα στην αναφορά m . Τα χαρακτηριστικά είναι η εκ των προτέρων πιθανότητα της κάθε οντότητας, η σημασιολογική συνοχή της με τα συμφραζόμενα (relatedness) και η πιθανότητα μια αναφορά να αποτελεί σύνδεσμο (link probability).

3.1.3.1 Εκ των προτέρων πιθανότητα

Όπως περιγράφηκε στην υποενότητα 2.3.2.1, η εκ των προτέρων πιθανότητα ή δημοτικότητα μιας οντότητας e ως προς την αναφορά της m είναι η αναλογία του αριθμού που η m αναφέρεται στην οντότητα e συγκριτικά με τον αριθμό όλων το συνδέσμων της αναφοράς m . Έχοντας το χαρακτηριστικό αυτό, φάνηκε χρήσιμο να οριστεί ένα κατώφλι κάτω από το οποίο θα απορρίπτονται οι οντότητες. Ακολουθώντας τους Milne και Witten [29], αφαιρούνται όλες οι οντότητες για τις οποίες η εκ των προτέρων πιθανότητα είναι μικρότερη του 2%. Οι οντότητες που απορρίπτονται είναι αρκετά απίθανο να είναι είναι οι σωστές κι έτσι αυξάνεται το recall ενώ μειώνεται το precision από τα false positives των περιπτώσεων που είναι σωστές. Με αυτόν τον τρόπο γίνεται εξοικονόμηση χρόνου χωρίς να επηρεάζεται η επίδοση του συστήματος.

Αλγόριθμος 2 Ψευδοκώδικας τον υπολογισμό της εκ των προτέρων πιθανότητας των υποψήφιων οντοτήτων μιας αναφοράς

```

1: function CalcCommonness(mention)
2:   total  $\leftarrow$  0
3:   for each sense in mention.senses do
4:     total  $\leftarrow$  total + sense.appearances
5:   end for
6:   for each sense in mention.senses do
7:     sense.commonness  $\leftarrow$  sense.appearances/total
8:     if sense.commonness < 0.02 then
9:       del sense
10:    end if
11:   end for
12: end function
```

3.1.3.2 Πιθανότητα αναφοράς να αποτελεί σύνδεσμο

Στο μοντέλο αυτό, οι όροι των συμφραζομένων δεν θεωρούνται εξίσου χρήσιμοι. Η λέξη 'the' για παράδειγμα, υπάρχει σε εκατομμύρια άρθρα χωρίς να είναι σύνδεσμος και έχει μηδενική αξία για την αποσαφήνιση των αναφορών τους. Για τον εντοπισμών τέτοιων περιπτώσεων χρησιμοποιείται το χαρακτηριστικό της πιθανότητας μιας αναφοράς να αποτελεί anchor text συνδέσμου στη Wikipedia (link probability). Αν $freq(m)$ είναι ο αριθμός των άρθρων της Wikipedia στα οποία υπάρχει η αναφορά m και $link(m)$ ο αριθμός των άρθρων στα οποία η m είναι anchor text συνδέσμου τότε ορίζουμε:

$$link_probability(m) = \frac{link(m)}{freq(m)}$$

Η διαδικασία υπολογισμού είναι αρκετά χρονοβόρα καθώς πρέπει να αναζητήσουμε όλες τις αναφορές του dataset σε όλα τα άρθρα της Wikipedia. Για να διευκολυνθεί ο σκοπός αυτός, χρησιμοποιείται parser που εξάγει και καθαρίζει το κείμενο των άρθρων Wikipedia dump, διατηρώντας όμως τους συνδέσμους όπου υπάρχουν. Για κάθε άρθρο της εξόδου του parser, αν υπάρχει η αναφορά στους σύνδεσμος, αυξάνεται κατά ένα ο δείκτης που δηλώνει τον αριθμό των άρθρων που εμφανίζεται αλλά και ο δείκτης που δηλώνει ότι είναι σύνδεσμος σε άρθρο. Διαφορετικά, αν υπάρχει στο κείμενο του άρθρου, αυξάνεται μόνο ο δεύτερος δείκτης.

Αλγόριθμος 3 Ψευδοκώδικας για τον υπολογισμό της πιθανότητας μιας αναφοράς να αποτελεί σύνδεσμο

Require: *wikidump_text*, the parsed clean Wikipedia text with links included
Require: *GetLinks*, a method to get the links in the text

```

1: function LinkProb(mentions, wikidump_text)
2:   for each article in wikidump_text do
3:     links  $\leftarrow$  GetLinks(article)
4:     for each mention  $\in$  mentions do
5:       if mention in links then
6:         mention.as_link  $\leftarrow$  mention.as_link + 1
7:         mention.appearances  $\leftarrow$  mention.appearances + 1
8:       else if mention  $\in$  article then
9:         mention.appearances  $\leftarrow$  mention.appearances + 1
10:      end if
11:    end for
12:  end for
13: end function
```

3.1.3.3 Σημασιολογική συνοχή

Η σημασιολογική συνοχή υπολογίζεται χρησιμοποιώντας τη μέθοδο WLM όπως περιγράφηκε στην υποενότητα 2.3.2.1. Ο αλγόριθμος για τον υπολογισμό της πρέπει να κάνει ένα μεγάλο αριθμό συγκρίσεων, γι' αυτό ομοίως με τους Milne και Witten [29] λαμβάνονται υπόψιν μόνο οι εισερχόμενοι σύνδεσμοι προς κάθε άρθρο.

Για να βρεθούν οι εισερχόμενοι σύνδεσμοι χρησιμοποιείται το MediaWiki API με το οποίο λαμβάνονται όλα τα backlinks της σελίδας κάθε οντότητας. Επειδή το πλήθος των υποψήφιων οντοτήτων είναι μεγάλο, για εξοικονόμηση χρόνου αλλά και γενικότερη βελτίωση του dataset, βρίσκουμε αρχικά το page id που αντιστοιχεί σε κάθε οντότητα. Με τη διαδικασία αυτή παρατηρήθηκε ότι ορισμένες διαφορετικές υποψήφιες οντότητες αναφέρονταν στην ίδια σελίδα. Για παράδειγμα οι

οντότητες "mass production", "Mass Production" και "mass-production" είχαν το ίδιο page id και επομένως μπορούν να συμπτυχθούν σε μία μόνο υποψήφια οντότητα. Επίσης, άλλες οντότητες δεν έχουν pageid, κάτι που σημαίνει πως πρόκειται για κόκκινους συνδέσμους της Wikipedia, δηλαδή συνδέσμους που οδηγούν σε σελίδες που δεν έχουν δημιουργηθεί ακόμα ή που διαγράφτηκαν.

Ως συμφραζόμενα θεωρούνται όλοι οι σύνδεσμοι του κειμένου που έχουν μία μόνο υποψήφια οντότητα και άρα δε χρήζουν αποσαφήνισης (unambiguous terms). Επίσης, επειδή κάθε οντότητα των συμφραζομένων δεν έχει την ίδια αξία, της ανατίθεται ένα βάρος που είναι ο μέσος όρος της πιθανότητας η αναφορά να αποτελεί σύνδεσμο, όπως περιγράφηκε παραπάνω, και της συνοχής που έχει με τις υπόλοιπες οντότητες των συμφραζομένων. Έτσι ακολουθεί ο υπολογισμός της συνοχής των εναπομεινάντων αναφορών προς αποσαφήνιση με τον τύπο WLM. Συγκεκριμένα, η συνοχή μιας υποψήφιας οντότητας είναι ο σταθμισμένος μέσος όρος της συνοχής της με κάθε άρθρο των συμφραζομένων, όπου το βάρος κάθε σύγκρισης είναι αυτό που ορίσαμε παραπάνω.

Αλγόριθμος 4 Υπολογισμός βάρους μίας οντότητας των συμφραζομένων

Require: relatedness(a,b): the function to get the relatedness between two sets of incoming links with WLM method

```

1: global variables
2:   mentions, the dictionary of mentions in the dataset
3: end global variables
4:
5: ▷ unamb_term: the unambiguous term for which we are calculating the weight
6: ▷ rest: the remaining unambiguous terms of the article
7: function UnambWeight(unamb_term, rest)
8:   total_rel ← 0
9:   num_of_terms ← length(rest)
10:  for each other_term in rest do
11:    total ← total + relatedness(unamb_term.incoming, other_term.incoming)
12:  end for
13:  sem_rel ← total/num_of_terms
14:  weight ← (sem_rel + mentions[unamb_term.mention].link_prob)/2
15:  return weight
16: end function

```

3.1.3.4 Ποιότητα συμφραζομένων

Η ποιότητα συμφραζομένων είναι ένα χαρακτηριστικό που χρησιμοποιείται για να εξισορροπθεί η εκ των προτέρων πιθανότητα και η σημασιολογική συνοχή, λαμβάνοντας υπόψιν πόσο χρήσιμα είναι τα συμφραζόμενα. Αν είναι άφθονα και ομοιογενή, τότε η συνοχή έχει μεγαλύτερη αξία. Αν δε το κείμενο έχει ασαφές και

μπερδεμένο περιεχόμενο, η εκ των προτέρων πιθανότητα είναι πιο σημαντική. Η ποιότητα των συμφραζομένων δίνεται από το άθροισμα των βαρών που ανατέθηκαν προηγουμένως στις οντότητες των συμφραζομένων. Αυτό λαμβάνει υπόψη τον αριθμό των όρων που εμπλέκονται, την έκταση που σχετίζονται μεταξύ τους και τον τρόπο με τον οποίο συχνά χρησιμοποιούνται ως σύνδεσμοι της Wikipedia.

Αλγόριθμος 5 Υπολογισμός ποιότητας των συμφραζομένων

```

1: function ContextQuality(article)
2:   context_quality  $\leftarrow 0$ 
3:   for each term in article.unamb_terms do
4:     rest  $\leftarrow$  article.unamb_terms.remove(term)
5:     context_quality  $\leftarrow$  article.context_quality + UnambWeight(term, rest)
6:   end for
7:   return context_quality
8: end function

```

Αλγόριθμος 6 Υπολογισμός συνοχής των υποψήφιων οντοτήτων μίας αναφοράς

Require: relatedness(*a,b*): the function to get the relatedness between two sets of incoming links with WLM method

```

1: global variables
2:   mentions, the dictionary of mentions in the dataset
3: end global variables
4:
5: function AmbiguousRelatedness(article)
6:   for each term in article.amb_terms do
7:     for each sense in mentions[term.mention].senses do
8:       total  $\leftarrow 0$ 
9:       for each unmab_term in article.unamb_terms do
10:        total  $\leftarrow$  total + unmab_term.weight *
11:          relatedness(term.incoming, unmab_term.incoming)
12:        end for
13:      end for
14:      term.sem_rel  $\leftarrow$  total/ContextQuality(article)
15:    end for
16: end function

```

3.1.4 Ταξινόμηση υποψήφιων οντοτήτων

Το τελευταίο τμήμα του συστήματος είναι ένας classifier. Κατά το στάδιο της εκπαίδευσης, παίρνοντας ένα training set 600 άρθρων με training samples που έχουν τρία χαρακτηριστικά, την εκ ων προτέρων πιθανότητα, τη συνοχή και την ποιότητα των συμφραζομένων του άρθρου και δύο κλάσεις, 0 για τις λάθος οντότητες και 1 για τη σωστή, ο classifier εκπαιδεύεται να αναγνωρίζει την οντότητα που αντιστοιχεί στην αναφορά. Το πρόβλημα των μη αποσαφηνίσιμων αναφορών

δεν λαμβάνεται υπόψιν καθώς όλες οι αναφορές προς αποσαφήνιση του dataset είναι σύνδεσμοι της Wikipedia που οδηγούν σε κάποια σελίδα και άρα δεν εισάγεται το NIL στις υποψήφιες οντότητες.

Στο στάδιο της αξιολόγησης χρησιμοποιούνται 100 άρθρα για τα οποία ο πλέον εκπαιδευμένος classifier ξεχωρίζει τις ορθές οντότητες από τις λανθασμένες. Στην πραγματικότητα δεν επιλέγει την καλύτερη οντότητα για κάθε αναφορά. Αντιθέτως, εξετάζει κάθε οντότητα ανεξάρτητα από τις υπόλοιπες, παράγει μια πιθανότητα να είναι έγκυρη και στη συνέχεια επιλέγουμε την οντότητα που έχει την υψηλότερη πιθανότητα.

Πειραματιστήκαμε με τρεις διαφορετικούς classifiers:

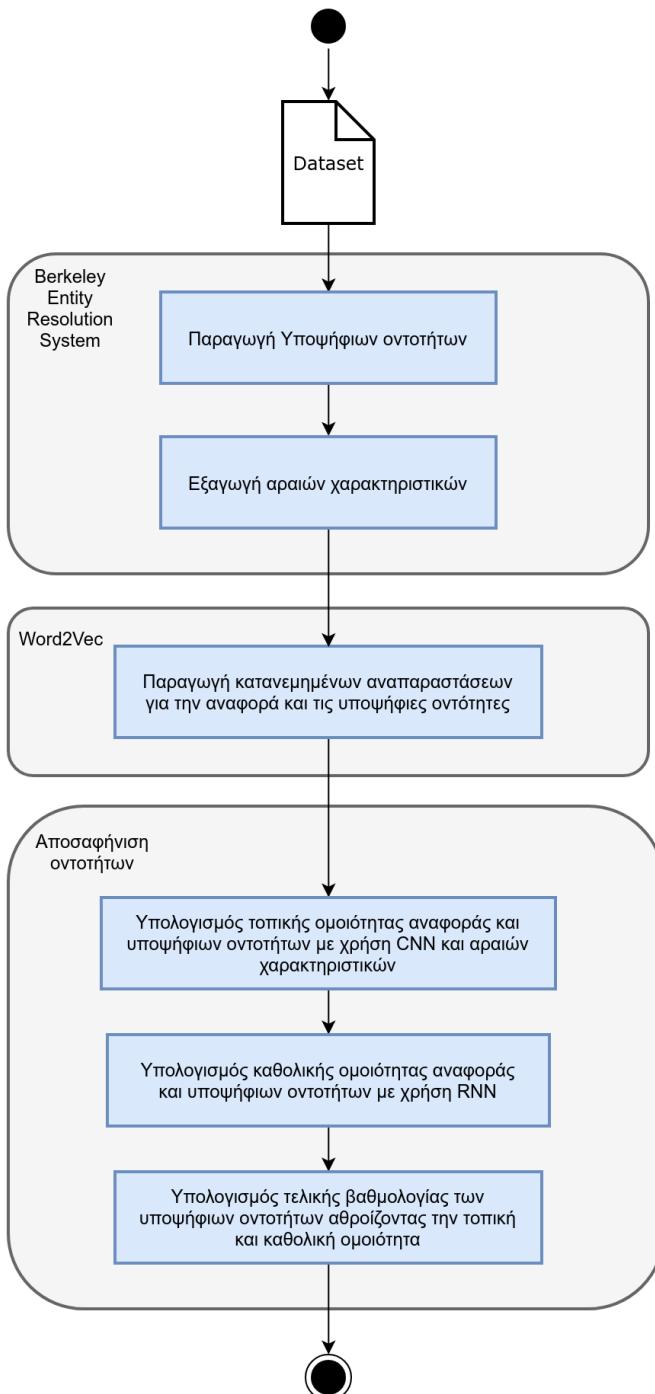
- Έναν ταξινομητή Naïve Bayes
- Έναν random forest ταξινομητή
- Μία Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine - SVM)

3.2 Συνδυαστικό Μοντέλο

Η πρώιμη προσέγγιση του προβλήματος αποσαφήνισης οντοτήτων όπως είδαμε χρησιμοποιεί για την ταξινόμηση των υποψήφιων οντοτήτων διάφορα διακριτά και "σχεδιασμένα με το χέρι" χαρακτηριστικά, τα οποία είναι συνήθως συγκεκριμένα για κάθε ζεύγος αναφοράς και υποψήφιας οντότητας. Στο δεύτερο σύστημα που υλοποιήθηκε με τη βοήθεια μοντέλων νευρωνικών δικτύων, οι λέξεις αντιπροσωπεύονται από συνεχείς αναπαραστάσεις και χρησιμοποιούνται χαρακτηριστικά των αναφορών και των υποψήφιων οντοτήτων, τόσο στο τοπικό επίπεδο μίας αναφοράς όσο και στο επίπεδο της καθολικής συγγένειας των αναφορών ολόκληρου του άρθρου, που μαθαίνονται αυτόματα από τα δεδομένα.

3.2.1 Παραγωγή υποψήφιων οντοτήτων

Το σύστημα αυτό, ομοίως με το προηγούμενο, δεν ασχολείται με το πρόβλημα εντοπισμού ονοματικών οντοτήτων και λαμβάνει ως είσοδο dataset στα οποία οι αναφορές προς αποσαφήνιση έχουν εντοπιστεί και επισημαίνονται μαζί με την οντότητα που τους αντιστοιχεί από τη Wikipedia.



Σχήμα 8: Σχεδιάγραμμα της αρχιτεκτονικής του συνδυαστικού μοντέλου.

Η παραγωγή υποψήφιων οντοτήτων, ακολουθώντας τους Durrett και Klein [33], γίνεται με τη βοήθεια του Berkeley Entity Resolution System. Με βάση αυτό, εισάγεται μια λανθάνουσα μεταβλητή q , η οποία καθορίζει πώς η αναφορά θα μετατραπεί σε ερώτημα (query) που θα επιστρέψει από τη Wikipedia το σύνολο υποψήφιων οντοτήτων. Η δημιουργία ερωτημάτων περιλαμβάνει την πιθανή αφαίρεση συνηθισμένων λέξεων, επιθεμάτων πληθυντικού, σημείων στίξεως και αρχικών

ή τελικών λέξεων. Αυτή η διαδικασία παράγει κατά μέσο όρο 9 ερωτήματα για κάθε αναφορά. Κάθε ερώτημα παράγει ένα σύνολο οντοτήτων με βάση τον αριθμό συνδέσμων που περιέχουν, και η ένωση των συνόλων οντοτήτων των ερωτημάτων δίνει στη συνέχεια το σύνολο υποψήφιων οντοτήτων της αναφοράς, συμπεριλαμβανομένου και του NIL. Για παράδειγμα, από την αναφορά "Deep Purple" προκύπτουν τα ερωτήματα "Deep Purple" και "Purple", από τα οποία παράγονται μεταξύ άλλων οι οντότητες της Wikipedia [Deep_Purple](#), [Deep_Purple_\(album\)](#) από το πρώτο ερώτημα, και [Purple](#) από το δεύτερο.

3.2.2 Ταξινόμηση υποψήφιων οντοτήτων

Το πρόβλημα της αποσαφήνισης οντοτήτων σε αυτό το σύστημα μπορεί να μοντελοποιηθεί ως εξής. Έστω ότι D είναι ένα άρθρο, $M = \{m_1, \dots, m_k\}$ οι αναφορές του και για κάθε αναφορά $m_i \in D$ το σύνολο των υποψήφιων οντοτήτων της είναι $P_i = \{p_{i1}, \dots, p_{in_i}\}$, όπου n_i ο αριθμός υποψήφιων οντοτήτων της αναφοράς m_i και $p_i^* \in P_i$ η σωστή οντότητα που αντιστοιχεί στη m_i . Αναπαριστούμε κάθε αναφορά με μία πλειάδα $m_i = (s_i, c_i, d_i)$, όπου s_i είναι η συμβολοσειρά (surface string) της m_i , c_i είναι τα συμφραζόμενα στα πλαίσια ενός παραθύρου γύρο από την m_i , και d_i είναι ολόκληρο το άρθρο που περιέχει την m_i . Για τις υποψήφιες οντότητες χρησιμοποιείται η πλειάδα $p_{ij} = (t_{ij}, b_{ij})$, όπου t_{ij} και b_{ij} είναι ο τίτλος και τα κείμενο των άρθρων τους αντίστοιχα. Για διευκόλυνση, σημαίνουμε την προηγούμενη πλειάδα που αντιστοιχεί στη σωστή οντότητα ως $p_i^* = (t_i^*, b_i^*)$.

Για να ταξινομηθούν οι οντότητες, ανατίθεται βαθμολογία σχετικότητας $\phi(m_i, p_{ij})$ σε κάθε υποψήφια οντότητα $p_{i,j}$ της m_i ως:

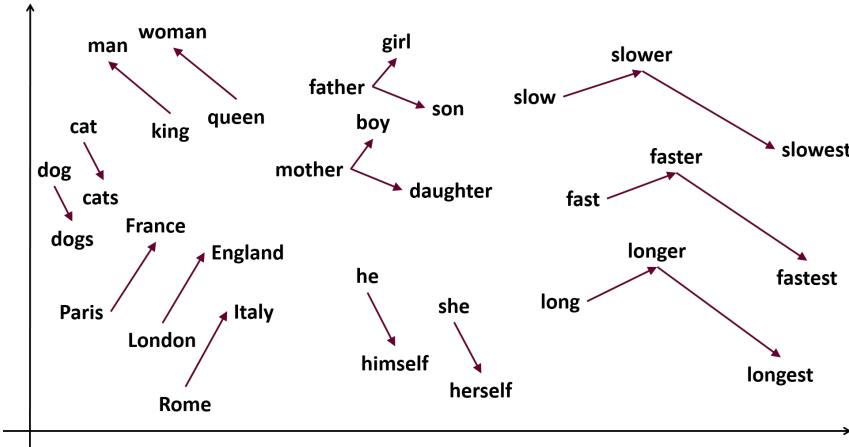
$$\phi(m_i, p_{ij}) = \phi_{local}(m_i, p_{ij}) + \phi_{global}(m_1, \dots, m_i, P_1, \dots, P_i)$$

Σε αυτή τη σχέση, το $\phi_{local}(m_i, p_{ij})$ αντιπροσωπεύει τις τοπικές ομοιότητες των m_i, p_{ij} , δηλαδή αυτές που εξάγονται χρησιμοποιώντας μόνο πληροφορίες που σχετίζονται με την αναφορά και την οντότητα. Το $\phi_{global}(m_1, \dots, m_i, P_1, \dots, P_i)$ από την άλλη υποδηλώνει τον υπολογισμό όλων των υποψήφιων οντοτήτων από όλες τις αναφορές του άρθρου, λαμβάνοντας υπόψη τη σειρά εμφάνισής τους.

3.2.2.1 Κωδικοποίηση

Προκειμένου να αποκτηθεί η κατανεμημένη αναπαράσταση κάθε ακολουθίας λέξεων $x \in \{s_i, c_i, d_i\}_i, \cup \{t_{ij}, b_{ij}\}_{i,j}$, μετατρέπουμε κάθε λέξη $x_i \in x$ σε ένα διάνυσμα

πραγματικών αριθμών w_i αξιοποιώντας έναν πίνακα ενσωμάτωσης λέξεων (word embedding table) της μεθόδου Word2vec [36], οπότε προκύπτει μια ακολουθία διανυσμάτων $w = \{w_1, \dots, w_n\}$. Με τη μέθοδο αυτή, τα διανύσματα των λέξεων τοποθετούνται στο χώρο των διανυσμάτων με τέτοιο τρόπο ώστε οι λέξεις που είναι σημασιολογικά κοινές να βρίσκονται κοντά μεταξύ τους.



Σχήμα 9: Διανύσματα λέξεων Word2vec.

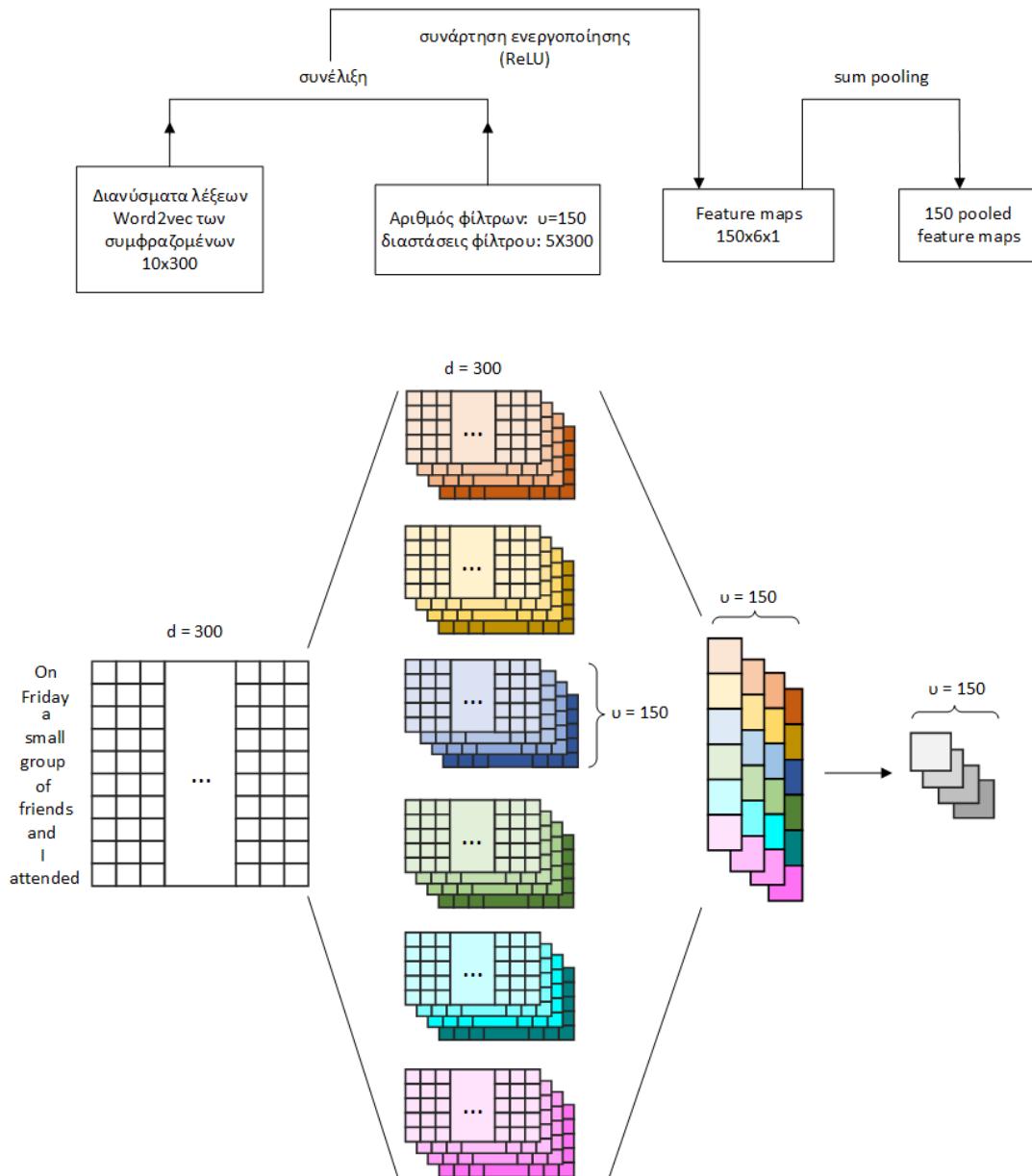
Στη συνέχεια αυτές οι λέξεις μετατρέπονται σε ένα διάνυσμα σταθερού μεγέθους μέσω ενός συνελικτικού δικτύου με πίνακα συνέλιξης $M \in \mathbb{R}^{v \times dL}$ ¹. Το αποτέλεσμα περνάει μέσα από ένα rectified linear unit (ReLU) και έπειτα τα αποτελέσματα συνδυάζονται με sum pooling δίνοντας την κατανεμημένη αναπαράσταση του x :

$$\bar{x} = \sum_{i=1}^{n-l+1} \max\{0, M_g w_{i:i+l-1}\}$$

όπου $w_{i:i+l}$ είναι το διάνυσμα συνένωσης των διανυσμάτων των λέξεων και M_g οι παράμετροι του φίλτρου που είναι διαφορετικές για κάθε είδος συνέλιξης (αναφορά, συμφραζόμενα, κείμενο). Από τα n-grams που οδηγούν στη μέγιστη ενεργοποίηση των διαφόρων φίλτρων παρατηρείται ότι πολλά φίλτρα μαθαίνουν να αναγνωρίζουν n-grams που είναι σχετικά με κάποιο συγκεκριμένο θέμα, ενώ άλλα φαίνεται να περιέχουν συνδυασμούς θεμάτων που δεν έχουν κάποια εμφανή συγγένεια.

Έστω $\bar{s}_i, \bar{c}_i, \bar{d}_i, \bar{t}_{ij}, \bar{b}_{ij}, \bar{t}_i^*, \bar{b}_i^*$ οι κατανεμημένες αναπαραστάσεις των $s_i, c_i, d_i, t_{ij}, b_{ij}, t_i^*, b_i^*$ αντίστοιχα που προκύπτουν από την παραπάνω συνελικτική διαδικασία. Αυτές θα τροφοδοτηθούν στα επόμενα τμήματα του συστήματος προκειμένου να υπολογιστούν τα χαρακτηριστικά για την αποσαφήνιση των οντοτήτων.

¹Θέσαμε αριθμό φίλτρων $v = 150$, διαστάσεις διανυσμάτων Word2vec $d = 300$, μέγεθος παραθύρου $l = 5$



Σχήμα 10: Γραφική απεικόνιση του υπολογισμού της κατανεμημένης αναπαράστασης μιας ακολουθίας 10 λέξεων. Τα Word2vec διανύσματα των λέξεων δίνονται ως είσοδος σε συνελικτικό δίκτυο με 150 φίλτρα πάνω σε 5 λέξεις κάθε φορά. Από τη διαδικασία συνέλιξης προκύπτουν 150 feature maps και μετά από sum pooling έχουμε το διάνυσμα που αποτελεί την κατανεμημένη αναπαράσταση της ακολουθίας των λέξεων.

3.2.2.2 Τοπική ομοιότητα

Οι τοπική ομοιότητα μια αναφοράς με μια υποψήφια οντότητα της εκφράζεται από τον παρακάτω τύπο:

$$\begin{aligned}\phi_{local}(m_i, p_{ij}) &= \phi_{sparse}(m_i, p_{ij}) + \phi_{CNN}(m_i, p_{ij}) = \\ &= W_{sparse}F_{sparse}(m_i, p_{ij}) + W_{CNN}F_{CNN}(m_i, p_{ij})\end{aligned}$$

όπου οι μεταβλητές W_{sparse} και W_{CNN} είναι τα βάρη των διανυσμάτων χαρακτηριστικών F_{sparse} και F_{CNN} αντίστοιχα.

Το διάνυσμα $F_{sparse}(m_i, p_{ij})$ αποτελεί ένα σύνολο αραιών χαρακτηριστικών (sparse features) τα οποία εντοπίστηκαν από προηγούμενες έρευνες του προβλήματος αποσαφήνισης οντοτήτων. Περιέχει διάφορες γλωσσικές ιδιότητες και στατιστικά όπως τον αριθμό εμφανίσεων της αναφοράς στη Wikipedia, τιμές μετρικών ομοιότητας της συμβολοσειράς με τους τίτλους των άρθρων των υποψήφιων οντοτήτων και ομοιότητα συνημιτόνου των tf-idf διανυσμάτων [33].

Το δε διάνυσμα F_{CNN} προκύπτει από τον υπολογισμό της ομοιότητας συνημιτόνου μεταξύ των διαφορετικών βαθμών ανάλυσης της αναφοράς και των υποψήφιων οντοτήτων, δηλαδή:

$$\begin{aligned}F_{CNN}(m_i, p_{ij}) &= [cosim(\bar{s}_i, \bar{t}_{ij}), cosim(\bar{c}_i, \bar{t}_{ij}), cosim(\bar{d}_i, \bar{t}_{ij}), \\ &\quad cosim(\bar{s}_i, \bar{b}_{ij}), cosim(\bar{c}_i, \bar{b}_{ij}), cosim(\bar{d}_i, \bar{b}_{ij})]\end{aligned}$$

3.2.2.3 Καθολική ομοιότητα

Όσον αφορά το κομμάτι της καθολικής ομοιότητας (global similarity), κρίθηκε ιδανική η χρήση αναδρομικού νευρωνικού δικτύου χάρη στην ικανότητά του να αποθηκεύει στη μνήμη του παρελθοντική πληροφορία. Στην περίπτωση του συστήματός μας, η πληροφορία αυτή αφορά το περιεχόμενο προηγούμενων οντοτήτων του άρθρου, με τη σειρά που εμφανίστηκαν.

Με σκοπό να εκφραστεί η συνοχή μεταξύ των αναφορών και των υποψήφιων οντοτήτων τους, τα αναδρομικά νευρωνικά δίκτυα δέχονται ως είσοδο την ακολουθία των διανυσμάτων αναπαράστασης του τίτλου ($\bar{t}_1^*, \bar{t}_2^*, \dots, \bar{t}_k^*$) και του κειμένου ($\bar{b}_1^*, \bar{b}_2^*, \dots, \bar{b}_k^*$) των άρθρων των υποψήφιων οντοτήτων. Οι ακολουθίες αυτές

αντιστοιχούν στις golden οντότητες κατά το στάδιο της εκπαίδευσης του συστήματος, ενώ κατά το στάδιο της αξιολόγησης λαμβάνονται από τις από τις οντότητες που προβλέπονται σύμφωνα με τα χαρακτηριστικά καθολικής ομοιότητας, όπως θα περιγραφούν στη συνέχεια. Για κάθε είδος ακολουθίας διανυσμάτων εκπαιδεύεται ένα διαφορετικό αναδρομικό νευρωνικό δίκτυο, άρα χρησιμοποιείται ένα για για τα $(\bar{t}_1^*, \bar{t}_2^*, \dots, \bar{t}_k^*)$ και ένα για τα $(\bar{b}_1^*, \bar{b}_2^*, \dots, \bar{b}_k^*)$.

Χρησιμοποιώντας ως παράδειγμα την ακολουθία διανυσμάτων για το κείμενο των άρθρων, $(\bar{b}_1^*, \bar{b}_2^*, \dots, \bar{b}_k^*)$, θα περιγράψουμε τη διαδικασία που ακολουθείται για τον υπολογισμό της καθολικής ομοιότητας. Το αναδρομικό νευρωνικό δίκτυο παίρνοντας αυτή την ακολουθία θα παράξει μια ακολουθία κρυφών καταστάσεων (hidden states) $h_1^b, h_2^b, \dots, h_k^b$ που προκύπτουν ως:

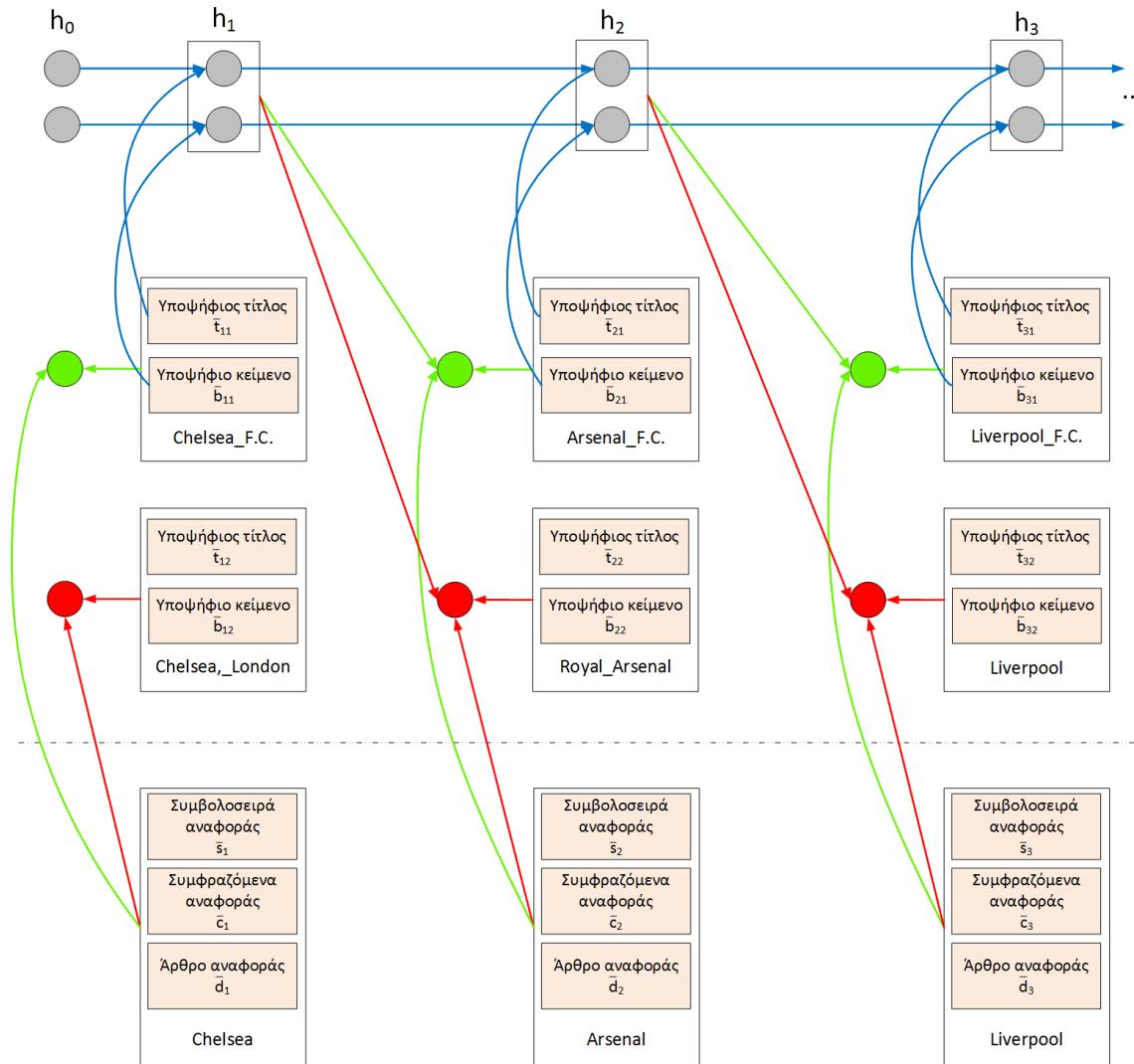
$$h_i^b = \Phi(h_{i-1}^b, \bar{b}_i^*)$$

όπου Φ η αναδρομική συνάρτηση ενός GRU (Gated Recurrent Unit) όπως προτάθηκε από τον Cho και τους συνεργάτες του [32]. Όπως αναφέρθηκε στην ενότητα 2.4.5.2, το GRU χρησιμοποιείται για την αντιμετώπιση του προβλήματος της εξαφανιζόμενης κλίσης (vanishing gradient problem) και είναι μια απλοποιημένη εκδοχή του αναδρομικού δικτύου Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory ή LSTM) που έχει δειχθεί πως επιτυγχάνει συγκρίσιμη επίδοση [37]. Λόγω, λοιπόν, της φύσης των αναδρομικών δικτύων, κάθε διάνυσμα h_i^b της ακολουθίας συνοψίζει την πληροφορία που αφορά το περιεχόμενο των προηγούμενων, πριν το i , υποψήφιων οντοτήτων στο άρθρο.

Εχοντας επομένως αυτή την ακολουθία, υπολογίζονται οι ομοιότητες συνημιτόνου μεταξύ του h_{i-1}^b και των διανυσμάτων αναπαράστασης κάθε υποψήφιας οντότητας της αναφοράς m_i , δηλαδή τα $\cos(h_{i-1}^b, \bar{t}_{ij}^*)$ και $\cos(h_{i-1}^b, \bar{b}_{ij}^*)$ αποτελούν και τα χαρακτηριστικά της καθολικής ομοιότητας. Ετσι μπορούμε να εξασφαλίσουμε πως η υποψήφια οντότητα της τρέχουσας αναφοράς m_i που θα επιλεχθεί ταιριάζει με το περιεχόμενο του κειμένου βάσει της πληροφορίας που περιέχεται στην κρυφή κατάσταση h_{i-1}^b .

Επαναλαμβάνοντας τη διαδικασία αυτή για για την ακολουθία διανυσμάτων των τίτλων $(\bar{t}_1^*, \bar{t}_2^*, \dots, \bar{t}_k^*)$ και συγκεντρώνοντας τις ομοιότητες συνημιτόνου σε ένα διάνυσμα F_{global} , υπολογίζεται ο βαθμός καθολικής ομοιότητας:

$$\phi_{global}(m_1, \dots, m_i, P_1, \dots, P_i) = W_{global} F_{global}(m_1, \dots, m_i, P_1, \dots, P_i)$$



Σχήμα 11: Γραφική απεικόνιση του υπολογισμού της καθολικής ομοιότητας. Οι γκρι κύκλοι είναι οι κρυφές καταστάσεις του RNN. Οι πράσινοι και κόκκινοι κύκλοι δηλώνουν τη βαθμολογία του μοντέλου για τις υποψήφιες οντότητες, όπου οι πράσινες είναι οι οντότητες που επιλέχθηκαν. Στη συνέχεια, οι κατανεμημένες αναπαραστάσεις τους χρησιμοποιούνται ως είσοδος για το επόμενο βήμα του RNN.

3.3 Σύγκριση των συστημάτων

Αφού έγινε η περιγραφή των δύο συστημάτων αξίζει να γίνει μια μεταξύ τους σύγκριση. Παρακάτω θα αναφερθούν ομοιότητες και διαφορές που αφορούν τη δομή τους και τη στρατηγική που ακολουθούν για την αποσαφήνιση οντοτήτων.

3.3.1 Ομοιότητες

Τόσο το πρώτο όσο και το δεύτερο σύστημα προσεγγίζουν το πρόβλημα της αποσαφήνισης οντοτήτων από δύο πλευρές. Η απόφαση για τη σωστή υποψήφια οντότητα των αναφορών λαμβάνεται αντιμετωπίζοντας από τη μια πλευρά την κάθε αναφορά ξεχωριστά και από την άλλη λαμβάνοντας υπόψιν τη σημασιολογική συνοχή των αναφορών στο κείμενο. Η τοπική προσέγγιση στο πρώτο σύστημα εκφράζεται με την την εκ των προτέρων πιθανότητα, ενώ στο δεύτερο με τα αραιά χαρακτηριστικά και την ομοιότητα συνημιτόνου των εξόδων των συνελικτικών νευρωνικών δικτύων. Η δε καθολική προσέγγιση στο πρώτο σύστημα εκφράζεται μέσω των χαρακτηριστικών της συνοχής και της ποιότητας συμφραζομένων, ενώ το δεύτερο σύστημα χρησιμοποιεί το αναδρομικό νευρωνικό δίκτυο για να αποσαφηνίσει ταυτόχρονα τις αναφορές ενός κειμένου και να αποκτήσει ένα σύνολο υποψήφιων οντοτήτων με σημασιολογική συνοχή.

Επίσης, και τα δύο συστήματα χρησιμοποιούν "χειροποίητα" χαρακτηριστικά. Το βασικό μοντέλο εξ' ολοκλήρου, ενώ το συνδυαστικό μοντέλο μέσω των αραιών χαρακτηριστικών του berkeley-entity συστήματος. Τέτοια χαρακτηριστικά που έχουν εντοπιστεί σε παλαιότερες έρευνες και έχουν εφαρμοστεί με επιτυχία αποτελούν καλή βάση για τα συστήματα αποσαφήνισης.

3.3.2 Διαφορές

Το πρώτο μοντέλο είναι ένα σύστημα αποσαφήνισης οντοτήτων απογυμνωμένο από όσο το δυνατό περισσότερες εξαρτήσεις από άλλα συμπληρωματικά συστήματα. Το πρόβλημα εντοπισμού ονομαστικών οντοτήτων παρακάμπτεται χρησιμοποιώντας άρθρα της Wikipedia, όπου οι ονοματικές οντότητες και τα συμφραζόμενα μπορούν να συγκεντρωθούν από τους συνδέσμους των σελίδων. Η παραγωγή υποψήφιων οντοτήτων γίνεται επίσης με επεξεργασία του περιεχομένου της Wikipedia και όχι με κάποιο συμπληρωματικό στατιστικό μοντέλο.

Βασική ακόμα διαφορά των δύο συστημάτων είναι ο τρόπος που αντιμετωπίζουν τις μη αποσαφηνίσιμες αναφορές. Το βασικό σύστημα δεν κάνει κάποια προσπάθεια να να τις αναγνωρίζει και απλώς τις αφαιρεί από το σύνολο των αναφορών προς αποσαφήνιση, εξετάζοντας μόνο τις αναφορές που είναι σύνδεσμοι προς κάποια σελίδα της Wikipedia. Το συνδυαστικό σύστημα προσθέτει την NIL οντότητα στο σύνολο υποψήφιων οντοτήτων και εκπαιδεύεται να της αναθέτει μια βαθμολογία ομοίως με τις υπόλοιπες υποψήφιες οντότητες.

Η εξάρτηση του βασικού συστήματος από χειροποίητα χαρακτηριστικά εισάγει κάποια μειονεκτήματα σε αυτό, τα οποία δεν εμφανίζονται στο συνδυαστικό μοντέλο. Καταρχάς, ο υπολογισμός τους είναι μια αρκετά χρονοβόρα διαδικασία καθώς απαιτεί επεξεργασία ολόκληρου του wikipedia dump αλλά και πολλά http requests προς το MediaWiki API. Θα ήταν χρήσιμο να υπήρχαν τα χαρακτηριστικά αυτά για κάθε οντότητα της Wikipedia διαθέσιμα σε κάποια βάση δεδομένων ώστε να εξάγονται γρήγορα και χωρίς σύνδεση στο διαδίκτυο. Εκτός αυτού, τα χειροποίητα χαρακτηριστικά είναι πιθανό να μην είναι εξίσου αποτελεσματικά με το πέρασμα του χρόνου, καθώς με τον αυξανόμενο αριθμό οντοτήτων στη Wikipedia, εισάγονται νέες σχέσεις και έτσι κάποια χαρακτηριστικά μπορεί να αποκτήσουν μεγαλύτερη βαρύτητα ενώ άλλα να καταστούν παρωχημένα.

Κεφάλαιο 4

Αξιολόγηση Συστημάτων

Στο κεφάλαιο αυτό θα παρουσιαστούν και θα σχολιαστούν τα αποτελέσματα από την εκτέλεση των δύο συστημάτων και θα αξιολογηθούν χρησιμοποιώντας τις μετρικές αξιολόγησης precision, recall και F1 score οι οποίες περιγράφηκαν στην ενότητα 2.5.

Σημειώνεται ότι τα δύο συστήματα αξιολογούνται σε διαφορετικά datasets. Η επιλογή αυτή βασίζεται στο ότι σύγκριση των συστημάτων ακόμα και σε κοινό dataset δεν θα είχε βάση για τους εξής λόγους:

- Το βασικό μοντέλο δεν κάνει αναγνώριση των μη αποσαφηνίσιμων αναφορών. Το πρόβλημα της αναγνώρισης μη αποσαφηνίσιμων αναφορών έχει τις περισσότερες φορές χαμηλότερη ακρίβεια από την αποσαφήνιση των οντοτήτων της βάσης γνώσης, γεγονός που θα οδηγούσε σε υψηλότερες τιμές στις μετρικές απόδοσης συγκριτικά με το αν λαμβανόταν υπόψιν, όπως γίνεται στο συνδυαστικό μοντέλο.
- Τα dataset που χρησιμοποιούνται για το πρόβλημα αποσαφήνισης οντοτήτων περιέχουν αναφορές που στην πλειοψηφία τους έχουν περισσότερες από μια υποψήφιες οντότητες. Το πρώτο σύστημα βασίζεται σε αναφορές που δεν είναι αμφίσημες για να δημιουργήσει τα συμφραζόμενα του κειμένου. Άρα στην περίπτωση των dataset που χρησιμοποιούνται στο δεύτερο σύστημα, θα έπρεπε να κάνει χρήση ενός συστήματος αναγνώρισης οντοτήτων ώστε να δημιουργήσει τα συμφραζόμενά του. Έτσι εισάγεται ένας εξωτερικός παράγοντας που επηρεάζει την απόδοση του μοντέλου. Από την άλλη, έχει πλεονέκτημα στα άρθρα της Wikipedia, όπου οι οντότητες που απαρτίζουν τα συμφραζόμενα είναι προσημειωμένες ως σύνδεσμοι και συνήθως σχετικές με το περιεχόμενο του άρθρου.

4.1 Βασικό μοντέλο

4.1.1 Στατιστικά του dataset

Όπως αναφέρθηκε στην ενότητα 3.1.1, ως dataset του συστήματος χρησιμοποιήθηκαν 700 τυχαία άρθρα της Wikipedia με πάνω από 50 συνδέσμους το καθένα. Αφού έγινε η εξαγωγή των έγκυρων αναφορών και των πιθανών οντοτήτων τους, υπολογίστηκαν τα παρακάτω στοιχεία για το dataset:

- Συνολικά περιέχει 58334 αναφορές. Από αυτές 38688 έχουν μία μόνο υποψήφια οντότητα και αποτελούν τα συμφραζόμενα. Οι υπόλοιπες 19646 είναι οι αναφορές προς αποσαφήνιση.
- Το μέσο μέγεθος του συνόλου υποψήφιων οντοτήτων μιας αναφοράς προς αποσαφήνιση είναι 3.6.
- 45777 ξεχωριστές οντότητες της Wikipedia.
- 83 σύνδεσμοι κατά μέσο όρο σε κάθε άρθρο.

4.1.2 Ρύθμιση υπερπαραμέτρων

Οι υπερπαράμετροι είναι παράμετροι των ταξινομητών που δεν ρυθμίζονται απευθείας από τη διαδικασία εκμάθησης, αλλά περνιούνται ως παράμετροι των constructor των κλάσεων τους. Επομένως, για να βρεθεί ο συνδυασμός των υπερπαραμέτρων κάθε μοντέλου που βελτιστοποιεί την απόδοσή του, πραγματοποιείται εξαντλητική αναζήτηση ενός υποσυνόλου του χώρου όλων των συνδυασμών υπερπαραμέτρων.

Για να γίνει αυτό, χρησιμοποιείται η τεχνική διασταυρωμένης επικύρωσης (cross-validation), κατά την οποία ο ταξινομητής εκπαιδεύεται πάνω σε ένα σύνολο εκπαίδευσης και αξιολογείται σε ένα σύνολο επικύρωσης (validation set), που όμως είναι διαφορετικό του συνόλου αξιολόγησης για να μην μειωθεί η δυνατότητα γενίκευσης του ταξινομητή. Πιο συγκεκριμένα, εκτελείται διασταυρωμένη επικύρωση σε 3 μέρη, κατά την οποία το αρχικό σύνολο εκπαίδευσης 600 άρθρων χωρίζεται σε τρία τυχαία υποσύνολα 200 άρθρων το καθένα. Από τα τρία αυτά υποσύνολα, ένα χρησιμοποιείται ως το σύνολο επικύρωσης για την αξιολόγηση του μοντέλου και τα υπόλοιπα ως σύνολα εκπαίδευσης. Η διαδικασία διασταυρωμένης επικύρωσης στη συνέχεια επαναλαμβάνεται τρεις φορές, με καθένα από τα τρία υποσύνολα να χρησιμοποιείται ακριβώς μία φορά ως το σύνολο

επικύρωσης. Έπειτα υπολογίζεται ο μέσος όρος των τριών αποτελεσμάτων από κάθε μέρος για να παραχθεί μια ενιαία εκτίμηση. Ακολουθεί ο πίνακας με τους διαφορετικούς συνδυασμούς υπερπαραμέτρων που εξετάστηκαν για τα τυχαίο δάσος και τη μηχανή διανυσμάτων υποστήριξης καθώς ο ταξινομητής Naive Bayes δεν περιέχει υπερπαραμέτρους. Ως μετρική αξιολόγησης χρησιμοποιήθηκε το F1-score.

min_samples_split	n_estimators	max_depth	f1_score
2	10	6	81.35
2	50	6	81.65
2	128	6	81.75
10	10	6	81.69
10	50	6	81.66
10	128	6	81.49
2	10	50	79.12
2	50	50	80.52
2	128	50	80.81
10	10	50	80.56
10	50	50	81.21
10	128	50	81.47
2	10	100	79.33
2	50	100	80.61
2	128	100	80.84
10	10	100	80.52
10	50	100	81.22
10	128	100	81.41
2	10	None	79.36
2	50	None	80.51
2	128	None	80.69
10	10	None	80.60
10	50	None	81.34
10	128	None	81.42

Πίνακας 2: Ρύθμιση υπερπαραμέτρων για το μοντέλο τυχαίου δάσους.

Έτσι, για την τελική αξιολόγηση του συστήματος επιλέχθηκε τυχαίο δάσος με n_estimators = 128, max_depth = 6, min_samples_split = 2 και μηχανή διανυσμάτων υποστήριξης με kernel = 'poly', C=100, gamma=0.1, degree=3 όπως φαίνεται και από τα αποτελέσματα των πινάκων 2 και 3 αντίστοιχα.

4.1.3 Απόδοση του συστήματος

Ακολουθεί η αξιολόγηση του συστήματος στο σύνολο αξιολόγησης 100 άρθρων του dataset. Όπως αναφέρθηκε, στο στάδιο της αξιολόγησης, ο κάθε ταξινομητής

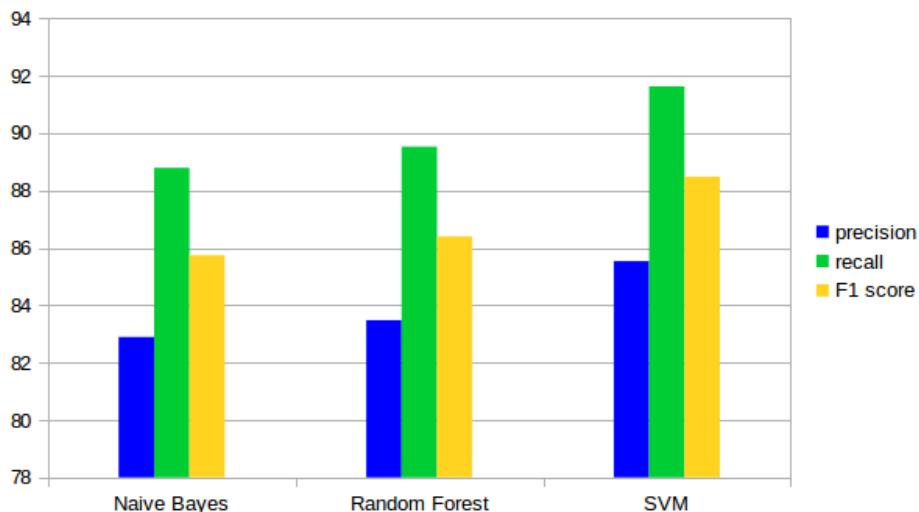
kernel	C	gamma	degree	f1_score
rbf	1	0.1	-	80.70
rbf	1	0.01	-	79.34
rbf	10	0.1	-	80.01
rbf	10	0.01	-	80.28
rbf	100	0.1	-	80.03
rbf	100	0.01	-	80.81
linear	1	-	-	80.32
linear	10	-	-	80.31
linear	100	-	-	80.32
poly	1	0.1	2	76.41
poly	1	0.01	2	0.00
poly	1	0.1	3	66.38
poly	1	0.01	3	0.00
poly	10	0.1	2	80.83
poly	10	0.01	2	70.86
poly	10	0.1	3	71.98
poly	10	0.01	3	0.00
poly	100	0.1	2	81.04
poly	100	0.01	2	76.41
poly	100	0.1	3	81.05
poly	100	0.01	3	0.00

Πίνακας 3: Ρύθμιση υπερπαραμέτρων για το μοντέλο μηχανής διανυσμάτων υποστήριξης.

model	precision	recall	F1 score
Naive Bayes	82.89	88.79	85.74
Random Forest	83.48	89.53	86.4
SVM	85.54	91.63	88.48

Πίνακας 4: Απόδοση των μοντέλων στο σύνολο αξιολόγησης.

δεν επιλέγει απευθείας για κάθε υποψήφια οντότητα την καλύτερη κλάση που της αντιστοιχεί, δηλαδή 0 για οντότητα που δεν αντιστοιχεί στη συγκεκριμένη αναφορά και 1 για να δηλώσει ότι είναι σωστή. Αντί αυτού, η έξοδος για κάθε υποψήφια οντότητα είναι μια πιθανότητα να είναι έγκυρη και από αυτές επιλέγεται μία μόνο οντότητα, αυτή που έχει την υψηλότερη πιθανότητα. Τα αποτελέσματα της πειραματικής αξιολόγησης των τριών μοντέλων με τις παραπάνω παραμετροποιήσεις φαίνονται φαίνονται στον πίνακα 4 και παρουσιάζονται γραφικά στην εικόνα 12.



Σχήμα 12: Γραφική απεικόνιση της απόδοση των μοντέλων του πρώτου συστήματος στο σύνολο αξιολόγησης.

4.1.4 Σχολιασμός αποτελεσμάτων

Όπως παρατηρούμε, ο ταξινομητής με τη χαμηλότερη απόδοση είναι ο Naive Bayes. Αυτό είναι αναμενόμενο διότι λειτουργεί με την υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, ενώ στην πραγματικότητα υπάρχουν εξαρτήσεις μεταξύ των χαρακτηριστικών όπως η μεγαλύτερη βαρύτητα της σημασιολογικής συνοχής σε κείμενα με πιο επικεντρωμένο περιεχόμενο. Τη δε μεγαλύτερη απόδοση επιτυγχάνει το SVM, κάτι που πιθανώς οφείλεται στη φύση του προβλήματος που είναι δυαδική ταξινόμηση.

Οι τιμές του precision είναι αρκετά χαμηλότερες από τις τιμές του recall σε όλους τους ταξινομητές. Αυτό οφείλεται στον αριθμό των αναφορών για τις οποίες η σωστή οντότητα δε βρίσκεται στο σύνολο υποψήφιων οντοτήτων, που σημαίνει ότι όλες οι υποψήφιες οντότητές τους ανήκουν στην κλάση 0 και όταν το σύστημα επιλέγει κάποια και την ταξινομεί στην κλάση 1, δημιουργείται false positive.

Το σύστημα αποδίδει ικανοποιητικά παρ' ότι είναι βασισμένο σε μοντέλο που χρησιμοποιήθηκε πρώτη φορά το 2008. Ωστόσο, συγκριτικά με τότε, ο αριθμός των άρθρων της Wikipedia είναι σχεδόν 2.5 φορές μεγαλύτερος τη στιγμή εγγραφής αυτής της εργασίας. Άρα η αποτελεσματικότητα των χαρακτηριστικών αυτών μπορεί να έχει επηρεαστεί.

4.2 Συνδυαστικό μοντέλο

4.2.1 Datasets

Για την αξιολόγηση του δεύτερου συστήματος χρησιμοποιήθηκαν δύο διαφορετικά dataset αποσαφήνισης ονοματικών οντοτήτων, το ACE 2005 και το CoNLL-YAGO. Σημειώνεται ότι αυτά τα dataset έχουν δεχθεί επεξεργασία και βρίσκονται σε μορφή CoNLL όπως απαιτεί το σύστημα Berkeley-Entity ώστε να παράξει το σύνολο υποψήφιων οντοτήτων και να εξάγει τα αραιά χαρακτηριστικά.

4.2.1.1 ACE

Το ACE 2005 dataset επικεντρώνεται σε 5 βασικά προβλήματα: την αναγνώριση οντοτήτων, αξιών, χρονικών εκφράσεων, σχέσεων και συμβάντων. Τα δεδομένα προέρχονται από διάφορες πηγές που αφορούν έξι διαφορετικούς τομείς: ραδιοτηλεοπτικές ειδήσεις, ραδιοτηλεοπτικές συζητήσεις, newswire, ιστολόγια, usenet και τηλεφωνικές συνομιλίες. Τα δεδομένα του είναι διαθέσιμα σε Αραβικά, Κινέζικα και Αγγλικά. Στην εργασία αυτή χρησιμοποιήθηκε μόνο η αγγλική γλώσσα. Περιέχει 599 άρθρα εκ των οποίων 117 αποτελούν το σύνολο αξιολόγησής του. Λόγω της χρονικής διαφοράς από τη δημιουργία του, ορισμένες οντότητες του δεν υπάρχουν πλέον στη Wikipedia. Μετά από φίλτραρισμα, από όλες τις οντότητές του dataset έμεινε προς αποσαφήνιση το 98.53%, ενώ από τις οντότητες του συνόλου αξιολόγησης το 95.23%.

4.2.1.2 CoNLL-YAGO

Το CoNLL-YAGO είναι βασισμένο στο CoNLL-2003 dataset. Περιέχει αναθέσεις οντοτήτων σε αναφορές ονοματικών οντοτήτων που προσημειώθηκαν για το αρχικό CoNLL-2003 πρόβλημα. Οι οντότητες αναγνωρίζονται από όνομα οντότητας YAGO2, Wikipedia url, ή Freebase mid. Απαρτίζεται από 1392 άρθρα εκ των οποίων 231 αποτελούν το σύνολο αξιολόγησής του και περιέχει κάποιες σπανιότερες οντότητες. Ομοίως με πριν, φίλτραρονται οι πλέον μη έγκυρες οντότητες και έτσι παρέμεινε συνολικά το 98.99% των οντοτήτων και συγκεκριμένα από το σύνολο αξιολόγησης το 94.91%.

4.2.2 Απόδοση του συστήματος

Αρχικά αξιολογείται ξεχωριστά η απόδοση της καθολικής ομοιότητας που επιτυγχάνεται από την πλευρά των RNN. Οι πρώτες αναφορές κάθε κειμένου αγνοούνται αφού η καθολική ομοιότητα προϋποθέτει την ύπαρξη προηγούμενων αναφορών για τον υπολογισμό της σημασιολογικής συνοχής του κειμένου.

	ACE	CoNLL
Global Similarity F1	68.3	71.4
Global Similarity F1-NIL	42.7	59.2

Πίνακας 5: Η απόδοση της καθολικής ομοιότητας. Στην πρώτη σειρά φαίνεται το F1 των αποσαφηνίσημων αναφορών ενώ στη δεύτερη φαίνεται το F1 των μη αποσαφηνίσημων (NIL) αναφορών.

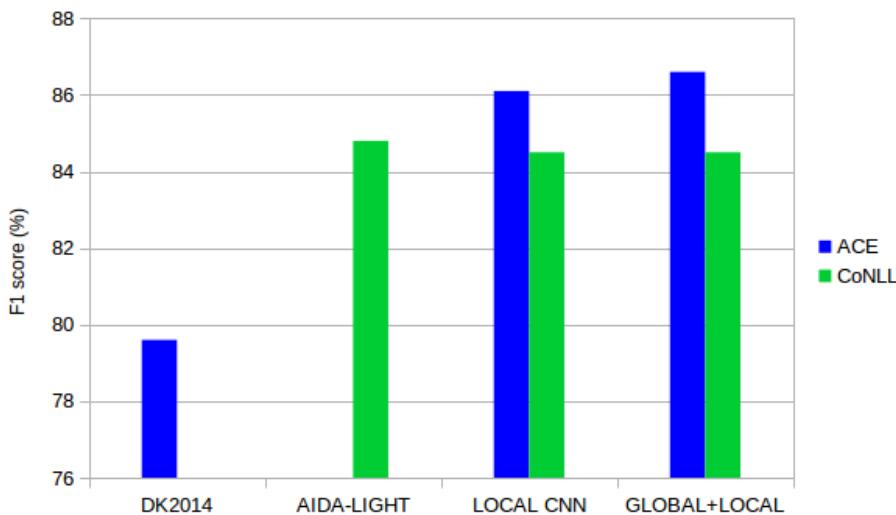
Έπειτα υπολογίζεται η απόδοση του συστήματος χρησιμοποιώντας το συνδυασμό τοπικής και καθολικής ομοιότητας. Τα αποτελέσματα φαίνονται στον πίνακα 6 και παρουσιάζονται γραφικά στην εικόνα 13. Συγκρίνονται με τα αποτελέσματα state-of-the-art μοντέλων, που είναι το μοντέλο που περιέχει μόνο τοπική ομοιότητα του Francis-Landau και των συνεργατών του [35], το μοντέλο ανάλυσης οντοτήτων των Durrett και Klein [33] και το μοντέλο AIDA-light του Nguen και των συνεργατών του [34] όπως περιγράφηκαν στην ενότητα 2.6. Τονίζεται ότι αυτά τα αποτελέσματα δεν έχουν αναπαραχθεί στο πλαίσιο αυτής της εργασίας, και η σύγκρισή με άλλες μεθόδους αποσαφήνισης έγινε βάσει βιβλιογραφίας.

	ACE	CoNLL
Προηγούμενες μέθοδοι		
DK2014	79.6*	-
AIDA-LIGHT	-	84.8*
LOCAL CNN	86.1	84.5
Μέθοδος της εργασίας		
GLOBAL + LOCAL	86.6	84.5

Πίνακας 6: Η απόδοση των διαφορετικών συστημάτων. Τα κελιά που έχουν σημειωθεί με * εξετάστηκαν στο Wikipedia dump του Δεκεμβρίου 2014, το οποίο είναι παλαιότερο από το δικό μας.

4.2.3 Σχολιασμός αποτελεσμάτων

Οπως φαίνεται η καθολική ομοιότητα είναι σχετικά χαμηλή και από μόνη της δεν επιτυγχάνει state-of-the-art αποτελέσματα. Αυτό οφείλεται σε μεγάλο βαθμό



Σχήμα 13: Γραφική απεικόνιση της απόδοση των διαφορετικών συστημάτων.

στην παρουσία των μη αποσαφηνίσημων αναφορών που πρέπει να αντιστοιχιστούν στην οντότητα NIL, καθώς αφαιρώντας τες από το CoNLL dataset, το F1-score ανέβηκε σε 77.6%. Η αξία της καθολικής ομοιότητας φαίνεται όταν λειτουργεί συμπληρωματικά και συνδυάζεται με την τοπική ομοιότητα ώστε το σύστημα να αποκτήσει μια πιο ολοκληρωμένη αντίληψη του κειμένου και των σχέσεων που διέπουν τις οντότητές του.

Το σύστημά μας έχει παρόμοια επίδοση με το Local CNN, το οποίο επεκτείνει, στο CoNLL dataset και ελαφρώς καλύτερη στο ACE dataset. Σημειώνεται ότι οι τα αποτελέσματα των DK2014 και AIDA-light αναμένεται να είναι χαμηλότερα στο Wikipedia dump του Ιουνίου 2016 που χρησιμοποιήσαμε καθώς η εισαγωγή νέων σελίδων (4.5 εκατομμύρια σελίδας στη Wikipedia το 2014, έναντι 5 εκατομμύρια το 2016) καθιστά τις αναφορές οντοτήτων περισσότερο αμφίσημες. Συγκριτικά, το local CNN μοντέλο είχε 89.9% F1 score στο ACE dataset και 85.5% στο CoNLL με το Wikipedia dump του Δεκεμβρίου 2014.

Τα αποτελέσματα του συστήματος δείχνουν ότι η τοπική και καθολική ομοιότητα που εξάγεται από τα συνελικτικά και αναδρομικά νευρωνικά δίκτυα αντίστοιχα, βελτιώνουν την απόδοση των αραιών χαρακτηριστικών που χρησιμοποιούνται και στο σύστημα DK2014. Πράγματι, αν ελέγξει κανείς τα βάρη των αραιών χαρακτηριστικών παρατηρείται ότι μεγαλύτερο βάρος έχουν συνήθως τα αυτά που δηλώνουν τον αριθμό της συχνότητας με την οποίο κάποια σελίδα αποτελεί προορισμό συνδέσμων και εκείνα που υποδεικνύουν συγκεκριμένα λεξιλογικά στοιχεία για την επιλογή της λανθάνουσας μεταβλητής q , κάτι που δηλώνει πως ενώ έχουν τη δυνατότητα να επιλέξουν τη σωστή πτυχή προς αποσαφήνιση μιας αναφοράς, στη συνέχεια επιλέγουν γενικά το άρθρο της Wikipedia

στο οποίο οδηγούν οι περισσότεροι σύνδεσμοι. Το συνδυαστικό σύστημα έχει καλύτερη δυνατότητα να επιλέγει λιγότερο συχνές οντότητες εάν τα θέματα που εξάγονται από τα CNN και η συνοχή από το RNN δείχνουν πως είναι ορθότερες επιλογές.

Κεφάλαιο 5

Επίλογος και Μελλοντικές Επεκτάσεις

Σε αυτή την εργασία είδαμε δύο προσεγγίσεις του προβλήματος αποσαφήνισης οντοτήτων. Η πρώτη, είναι βασισμένη σε ένα προγενέστερο σύστημα και είχε ως στόχο την εξοικείωση και κατανόηση του προβλήματος και την ανάδειξη μιας θεμελιώδης βάσης πάνω στην οποία στηρίζονται τα συστήματα αποσαφήνισης. Η δεύτερη, που μπορεί να θεωρηθεί επέκταση της πρώτης, χρησιμοποιεί συνδυασμό συνελικτικών νευρωνικών δικτύων για να εντοπίσει τοπικές ομοιότητες και αναδρομικών νευρωνικών δικτύων για να εισάγει τον παράγοντα της συνοχής σε επίπεδο κειμένου και επιτυγχάνει state-of-the-art απόδοση.

Ενώ το συνδυαστικό αυτό σύστημα αποδίδει ικανοποιητικά, αξίζει να αναφερθούν σημεία βελτίωσης και επεκτάσεις. Κάποιες από τις τροποποιήσεις που θα μπορούσαν να πραγματοποιηθούν είναι οι εξής:

- *Υποστήριξη περισσότερων γλωσσών.* Αυτή τη στιγμή το σύστημα υποστηρίζει μόνο την αγγλική γλώσσα. Η λειτουργία του όμως μπορεί να επεκταθεί σε άλλες γλώσσες χρησιμοποιώντας pre-trained Word2Vec διανύσματα, ή εκπαιδεύοντας νέα από το Wikipedia dump της αντίστοιχης γλώσσας. Περιορισμό αποτελούν η παραγωγή υποψήφιων οντοτήτων και εξαγωγή των αραιών χαρακτηριστικών του Berkeley-Entity συστήματος, που θα πρέπει να γίνει με διαφορετικό τρόπο. Μια απλή προσέγγιση θα ήταν να γίνει με τρόπο όμοιο του πρώτου συστήματος, δηλαδή να ληφθούν ως υποψήφιες οντότητες όλες οι οντότητες της βάσης γνώσης στις οποίες οδηγεί η αναφορά και ως χαρακτηριστικό να χρησιμοποιηθεί η εκ των προτέρων πιθανότητα.

- *Πειραματισμός με διαφορετικά νευρωνικά μοντέλα.* Το συνδυαστικό σύστημα χρησιμοποιεί ίδιες παραμέτρους και αρχιτεκτονική των νευρωνικών δικτύων με τη βιβλιογραφία. Μπορεί να γίνει πειραματισμός με εναλλακτικές τιμές παραμέτρων, όπως το μέγεθος των Word2Vec διανυσμάτων, το παράθυρο και ο αριθμός των φίλτρων του CNN, καθώς και με διαφορετικά μοντέλα νευρωνικών δικτύων, όπως αναδρομικό δίκτυο Jordan αντί για Elman και αμφίδρομα αναδρομικά δίκτυα.
- *Διατομεακή προσαρμογή.* Τα συστήματα αποσαφήνισης που εκπαιδεύτηκαν σε κάποιο τομέα μπορεί να χάσουν ακρίβεια αν στη συνέχεια δοκιμαστούν σε διαφορετικό τομέα, λόγω διαφορών μεταξύ του τομέα εκπαίδευσης και του τομέα στόχου όπως λεξιλόγιο, κατανομές δεδομένων, στυλ κειμένου. Αυτό μπορεί να αντιμετωπιστεί με τεχνικές διατομεακής προσαρμογής. Η σημασιολογική συνοχή από τα RNN βιοηθάει σε αυτό το σημείο, αλλά υπάρχουν περιθώρια βελτίωσης όπως εύρεση χαρακτηριστικών που βιοθούν στη μετάδοση γνώσης από τον ένα τομέα στον άλλο [38].
- *Επέκταση σε μοντέλο NERD.* Στην παρούσα μορφή του, το σύστημα ασχολείται μόνο με το πρόβλημα αποσαφήνισης οντοτήτων. Η παραγωγή υποψήφιων οντοτήτων γίνεται από το σύστημα Berkeley-Entity και η αναγνώριση ονοματικών οντοτήτων, ενώ δεν χρειάστηκε στα dataset που χρησιμοποιήθηκαν, θα πρέπει επίσης να γίνει με εξωτερικό σύστημα. Επομένως, το σύστημα θα μπορούσε να αναλάβει επίσης τα παραπάνω προβλήματα και να επεκταθεί σε πλήρες σύστημα αναγνώρισης και αποσαφήνισης ονοματικών οντοτήτων (NERD).

Αναφορές

- [1] *List of Wikipedias - Meta.* https://meta.wikimedia.org/wiki/List_of_Wikipedias.
- [2] *Facts & Figures / DBpedia.* <http://wiki.dbpedia.org/about/facts-figures>.
- [3] Ben Hachey et al. «Evaluating entity linking with Wikipedia». In: *Artificial intelligence* 194 (2013), pp. 130–150.
- [4] Stephen Guo, Ming-Wei Chang, and Emre Kiciman. «To link or not to link? a study on end-to-end tweet entity linking». In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013, pp. 1020–1030.
- [5] Abhishek Gattani et al. «Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach». In: *Proceedings of the VLDB Endowment* 6.11 (2013), pp. 1126–1137.
- [6] Razvan Bunescu and Marius Pașca. «Using encyclopedic knowledge for named entity disambiguation». In: *11th conference of the European Chapter of the Association for Computational Linguistics*. 2006.
- [7] Silviu Cucerzan. «Large-scale named entity disambiguation based on Wikipedia data». In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 2007.
- [8] Xianpei Han and Jun Zhao. «NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking.» In: *TAC*. Citeseer. 2009.
- [9] Sayali Kulkarni et al. «Collective annotation of Wikipedia entities in web text». In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 457–466.
- [10] Vasudeva Varma et al. «IIIT Hyderabad at TAC 2009.» In: *TAC* 2009.

- [11] Vasudeva Varma et al. «IIIT Hyderabad in Guided Summarization and Knowledge Base Population.» In: *TAC*. 2010.
- [12] Wei Zhang et al. «Entity linking leveraging: automatically generated annotation». In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 1290–1298.
- [13] Wei Zhang et al. «NUS-I2R: Learning a Combined System for Entity Linking.» In: *TAC*. 2010.
- [14] Zheng Chen et al. «CUNY-BLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description.» In: *TAC*. 2010.
- [15] Zhicheng Zheng et al. «Learning to link entities with knowledge base». In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 483–491.
- [16] John Lehmann et al. «LCC Approaches to Knowledge Base Population at TAC 2010.» In: *TAC*. 2010.
- [17] Wei Zhang et al. «I2R-NUS-MSRA at TAC 2011: Entity Linking.» In: *TAC*. 2011.
- [18] Silviu Cucerzan. «TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation.» In: *TAC*. 2011.
- [19] Wei Zhang et al. «Entity linking with effective acronym expansion, instance selection, and topic modeling». In: *IJCAI*. Vol. 2011. 2011, pp. 1909–1914.
- [20] Sean Monahan et al. «Cross-Lingual Cross-Document Coreference with Entity Linking.» In: *TAC*. 2011.
- [21] Xianpei Han and Le Sun. «A generative entity-mention model for linking entities with knowledge base». In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, pp. 945–954.
- [22] Xianpei Han, Le Sun, and Jun Zhao. «Collective entity linking in web text: a graph-based method». In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 765–774.
- [23] Lev Ratinov et al. «Local and global algorithms for disambiguation to wikipedia». In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, pp. 1375–1384.

- [24] Swapna Gottipati and Jing Jiang. «Linking entities to a knowledge base with query expansion». In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 804–813.
- [25] Wei Shen et al. «Linden: linking named entities with knowledge base via semantic knowledge». In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 449–458.
- [26] Wei Shen et al. «Linking named entities in tweets with knowledge base via user interest modeling». In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 68–76.
- [27] Mark Dredze et al. «Entity disambiguation for knowledge base population». In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 277–285.
- [28] Johannes Hoffart et al. «Robust disambiguation of named entities in text». In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 782–792.
- [29] David Milne and Ian H Witten. «Learning to link with wikipedia». In: *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, pp. 509–518.
- [30] Rudi L Cilibrasi and Paul MB Vitanyi. «The google similarity distance». In: *IEEE Transactions on knowledge and data engineering* 19.3 (2007).
- [31] Paul McNamee. «HLTCOE Efforts in Entity Linking at TAC KBP 2010.» In: *TAC*. 2010.
- [32] Kyunghyun Cho et al. «Learning phrase representations using RNN encoder-decoder for statistical machine translation». In: *arXiv preprint arXiv:1406.1078*(2014).
- [33] Greg Durrett and Dan Klein. «A joint model for entity analysis: Coreference, typing, and linking». In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 477–490.
- [34] Dat Ba Nguyen et al. «AIDA-light: High-Throughput Named-Entity Disambiguation.» In: *LDOW1184* (2014).
- [35] Matthew Francis-Landau, Greg Durrett, and Dan Klein. «Capturing semantic similarity for entity linking with convolutional neural networks». In: *arXiv preprint arXiv:1604.00734*(2016).

- [36] Tomas Mikolov et al. «Distributed representations of words and phrases and their compositionality». In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.
- [37] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. «An empirical exploration of recurrent network architectures». In: *International Conference on Machine Learning*. 2015, pp. 2342–2350.
- [38] Thien Huu Nguyen and Ralph Grishman. «Employing word representations and regularization for domain adaptation of relation extraction». In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2014, pp. 68–74.
- [39] David Nadeau and Satoshi Sekine. «A survey of named entity recognition and classification». In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.
- [40] Wei Shen, Jianyong Wang, and Jiawei Han. «Entity linking with a knowledge base: Issues, techniques, and solutions». In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 443–460.
- [41] Omer Levy and Yoav Goldberg. «Dependency-based word embeddings». In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2014, pp. 302–308.
- [42] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. «On the difficulty of training recurrent neural networks». In: *International Conference on Machine Learning*. 2013, pp. 1310–1318.
- [43] Thien Huu Nguyen et al. «Joint learning of local and global features for entity linking via neural networks». In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 2310–2320.

Παράρτημα Α

Άρθρα της Wikipedia

Σε αυτό το παράρτημα παρατίθενται τα 700 άρθρα της Wikipedia που χρησιμοποιήθηκαν ως το dataset του βασικού μοντέλου.

1. (10115) 1992 SK
2. 109th United States Congress
3. 1892 vote of no confidence in the Salisbury ministry
4. 18th International Adana Golden Bell Film Festival
5. 1932 Cuba hurricane
6. 1945
7. 1989–90 Philadelphia Flyers season
8. 1991–92 Copa México
9. 1995 in motoring
10. 1st Armoured Division (Poland)
11. 2000 Labatt Brier
12. 2004–05 Indiana Pacers season
13. 2004 Wales rugby union tour of Argentina and South Africa
14. 2005 West of Scotland Cup Final
15. 2008 Friends Provident Trophy
16. 2008 Rafael Nadal tennis season
17. 2009 UEFA European Under-17 Championship
18. 2011 24 Hours of Le Mans
19. 2011 Canadian Soccer League season
20. 2011 Missouri River Flood
21. 2012 Assam violence
22. 2012 Major League Soccer season
23. 2012 NFL Draft
24. 2014–15 Wichita B-52s season
25. 2014 Carolina Panthers season
26. 2014 Giro del Trentino
27. 2015–16 BYU Cougars women's basketball team
28. 2015 Andy Murray tennis season
29. 582d Helicopter Group
30. Abbasid Revolution
31. A Christian Reflection on the New Age
32. Adam Schechter
33. Adrianople Vilayet
34. Adriatic sturgeon
35. Afghan Armed Forces
36. African-American president of the United States in popular culture
37. AfroBasket 2009
38. Ahmed el-Tayeb
39. Air India Express Flight 812
40. Alan Kirby
41. Albert I of Käfelnburg
42. Alexander Macfarlane
43. Alfred Rolfe (director)
44. Ålgård Line
45. All Singing, All Dancing
46. Almond
47. Alphonse, Count of Poitiers
48. American Negro Labor Congress
49. Anagnorisis
50. Anderson, South Carolina
51. Andrew McCarthy
52. Angel Robinson Garcia
53. Annora Brown
54. Ansett-ANA Flight 325
55. Anti-genre

56. Antonio Beccadelli (poet)
57. Antonio Cervantes
58. Archibald Murray
59. Area code 262
60. Arijan Komazec
61. Art belongs to the people (Leningrad, 1977)
62. Arthur Golding
63. Arthur Rackham
64. ASK Riga
65. Assembly language
66. Astartea
67. Atlee Ayres
68. Auguste Doriot
69. Avraham Kalmanowitz
70. Avro 720
71. Babe Carey
72. Bangil
73. Barbra Streisand...and Other Musical Instruments
74. Barry Town United F.C.
75. Barry Wood (American football)
76. Battle of Khajwa
77. Battle of Sedan (1940)
78. Battle of Taierzhuang
79. Bay Ronald
80. Becharaji
81. Beiarn
82. Bengali cuisine
83. Ben Horowitz
84. Beretta ARX160
85. Bernard Jenkin
86. Beroun District
87. Betcha Bottom Dollar
88. Betty Carter
89. Betty Comden
90. Betty Lennox
91. Big Hoops (Bigger the Better)
92. Bill Bentley (record producer)
93. Black Death in medieval culture
94. Blake Comeau
95. Bloody Bloody Andrew Jackson
96. Bloody Sunday (1938)
97. Boeing B-52 Stratofortress
98. Boston Harbor
99. Brenda Taylor (athlete)
100. Brian Schaefering
101. Brivanib alaninate
102. Broadway Hollywood Building
103. Bruce Ford (tenor)
104. Bruce Graham
105. Bryan D. O'Connor
106. Brygos Painter
107. Bülent Uygun
108. California Division of Juvenile Justice
109. Cane Belt Railroad
110. Canton of La Vallée des Gaves
111. Captain (ice hockey)
112. Carlos Huerta
113. Carol Zhao
114. Carpetbagger
115. Carrier Strike Group 15
116. Carteret County, North Carolina
117. Cassie Turner
118. Catatumbo lightning
119. Catholic Church in France
120. Cedars-Sinai Medical Center
121. Celebes crested macaque
122. Chaim Ozer Grodzinski
123. Chariots of Fire
124. Charles Djou
125. Charles McArthur
126. Charles Scribner's Sons
127. Chemically modified electrode
128. Cheshunt F.C.
129. Child labour in India
130. Chi (letter)
131. Chimes of Freedom (horse)
132. China Dragon
133. Chinese passport
134. Chitra Naik
135. Christie (band)
136. Christopher McCulloch
137. CIA activities in the Democratic Republic of the Congo
138. Cinco canciones populares argentinas
139. Cinema of the United Kingdom
140. Citizen Ruth
141. Classical conditioning
142. Clovis point
143. CMLL Super Viernes (July 2010)
144. Coat of arms of Ukraine
145. Cohen crime family
146. Connie Marrero
147. Contra (series)
148. Convention on Fishing and Conservation of the Living Resources of the High Seas
149. Cork City F.C.
150. Count Luitpold of Castell-Castell
151. Craig Farrell (footballer)
152. Craig Tatum
153. Creativity and mental illness
154. Cricket in South Africa
155. Criticism of Judaism
156. Curse of dimensionality
157. Dakota Territory

158. Đakovo
159. Dan Campbell
160. Dangerous Woman (song)
161. Daniel Prodan
162. Danny Barnes (musician)
163. Dark Angel (TV series)
164. Datsakorn Thonglao
165. Dave Bush
166. Dave Manders
167. David Brown (footballer, born 1887)
168. David Goffin
169. David Hughes (tenor)
170. David Lieber
171. David Shaltiel
172. David Stevenson (admiral)
173. D. C. Boonzaier
174. Dead Sea Scrolls
175. Deansgate
176. Derek Raymond
177. Destroy Destroy Destroy
178. Diahann Carroll
179. Diana Vreeland
180. Dino Radja
181. Dissociative identity disorder
182. Dizzy Trout
183. Dominic Treadwell-Collins
184. Done In 60 Seconds Award
185. Don Ho
186. Donie Bush
187. Dønna
188. Doping at the 2007 Tour de France
189. Douglas J. McCarron
190. Drought in Australia
191. Dutch elm disease
192. Early life of Mirza Ghulam Ahmad
193. Echuca
194. Economic policy of the Nicolás Maduro government
195. EDAG
196. Eddie Cochran
197. Edmund Payne
198. El Chapulín Colorado
199. Eliot Engel
200. Elizabeth Craven
201. Elliott Sadler
202. Elverum
203. Epcot International Food & Wine Festival
204. Epes Randolph
205. Erich Leinsdorf
206. Erskine Hawkins
207. Esmond Knight
208. Evel Knievel
209. E. W. Hornung
210. Exit Through the Kwik-E-Mart
211. Fährinsel
212. Fălticeni
213. Fannie Farmer
214. Fargo (TV series)
215. Fashion blog
216. Fayetteville Regional Airport
217. Federal architecture
218. Félix González-Torres
219. Fernando Giudicelli
220. Fick's laws of diffusion
221. Fifth Avenue Hotel
222. File Explorer
223. Fill Your Head with Rock
224. First Great Awakening
225. Fishing industry in the United States
226. Flea (musician)
227. Floyd Thompson (lawyer)
228. Folk metal
229. Foot (unit)
230. Forrest Gump (soundtrack)
231. Francisco Coloma y Maceda
232. Francis Dashwood, 11th Baron le Despencer
233. Franco Di Santo
234. Frank Mount Pleasant
235. Frederick Brewing Company
236. Frederick I, Margrave of Meissen
237. Frigolet Abbey
238. Frogmore
239. Funtime (Iggy Pop song)
240. Fusui County
241. Gaetano Fraschini
242. Galois connection
243. Gary Bennett (footballer, born 1961)
244. Gayle Hunnicutt
245. Genzyme
246. George C. Butte
247. George Curtis (footballer, born 1919)
248. George Gabriel
249. George Porter
250. Georges Baklanoff
251. Georg Friedrich, Margrave of Baden-Durlach
252. Gerald Drucker
253. Germany–Japan relations
254. Gerry Ryan
255. Gmina Puck
256. Gnaeus Julius Agricola
257. Good Will Hunting
258. Gordon Messenger
259. Gordon Muortat Mayen
260. Gore (band)

261. Graphical timeline of prehistoric life
262. Greater Boston
263. Gregory Lee Johnson
264. Gros Morne National Park
265. Gullah
266. GWR 4073 Class 4079 Pendennis Castle
267. Hagai Shaham
268. Hail to the King (Avenged Sevenfold album)
269. Hakkō ichiu
270. Halden Prison
271. Hank Anderson
272. Hanover
273. Hans Ulrich Gumbrecht
274. Harpe brothers
275. Harts (musician)
276. Headquarters (album)
277. Health threat from cosmic rays
278. Heart
279. Heather Raffo
280. Helmet-mounted display
281. Henry Cohen, 1st Baron Cohen of Birkenhead
282. Henry Home, Lord Kames
283. Herbert Wehner
284. Hierarchy problem
285. Highway systems by country
286. Hikaru Shida
287. Hilltop, Minnesota
288. History of energy
289. History of Hartford City, Indiana
290. History of Palestine
291. Holographic interference microscopy
292. Holomorphic functional calculus
293. Holywell
294. HSPA8
295. Hückel method
296. Hudson Jet
297. Hunan Avetisyan
298. Hundred of Hoo Railway
299. Hydrocynus vittatus
300. Hydrogen-cooled turbo generator
301. Ian McCall (footballer)
302. Icahn School of Medicine at Mount Sinai
303. Iconostasis
304. Iligan
305. Incense Route
306. Industrial Workers of the World
307. International business
308. International Convention for the Regulation of Whaling
309. International Criminal Tribunal for Rwanda
310. Irvine Welsh's Ecstasy
311. Iry LeJeune
312. Isaac Hawkins Browne (coal owner)
313. Isa Boletini
314. Jack Off Jill
315. Jack Taylor (Arizona politician)
316. James A. Kowalski
317. James K. Johnson
318. Jefferson Pepper
319. Jerame Tuman
320. Jie people
321. Jihlava (river)
322. Jim Chen
323. Joe Thunder
324. John Beresford (statesman)
325. John Darnton
326. John Fieldhouse (rugby league)
327. John M. Jones
328. John Parkin Taylor
329. John Petrucci
330. John William Waterhouse
331. John Y. Mason
332. Jonah Kapena
333. Joseph Tiefenthaler
334. Josh Judy
335. Joyce Grable
336. JPEG
337. Juan Antonio Sotillo
338. Juan Sánchez Moreno
339. Julio Aguilera
340. Junction City, Kansas
341. Jundallah (Iran)
342. Justin Labonte
343. Kamioka Observatory
344. Kansas City Mavericks
345. Kappa Alpha Order
346. Karim Alrawi
347. Karl Schwarzschild
348. Karma in Jainism
349. Katerini
350. Kathy Long
351. KDLD
352. Kendall Langford
353. Ken Sprague (cartoonist)
354. Kernavė
355. K. Hariharan (director)
356. Kid Acne
357. Kim Kaphwan
358. Kingdom of Redonda
359. King World Productions
360. KLKN
361. Knattspyrnufélag Reykjavíkur
362. Kõrgessaare Parish

363. Kosača noble family
364. KOSC
365. Koyukuk River
366. Kremsmünster Abbey
367. Krishna Janmashtami
368. Kris Radlinski
369. KRLD-FM
370. Lactarius indigo
371. Lamborghini Gallardo
372. Languages of Bhutan
373. Larry Butler (darts player)
374. Late years of Pope Pius XII
375. Laura Boersma
376. Lawyer
377. Leeland (band)
378. Le Monde
379. Leone Cattani
380. Liberty, Missouri
381. Li Bing
382. Lie algebra representation
383. Linda Duncan
384. Linda Gilroy
385. Lo Mejor de Tu Vida
386. London and Continental Railways
387. London Borough of Sutton
388. Long March 5
389. Lorenzo Esposito Fornasari
390. Lotte Lehmann
391. Louisa Wall
392. Louisiana District Courts
393. Louisville, Colorado
394. Love Story (Andy Williams album)
395. Low-rise (fashion)
396. Lucian Croitoru
397. Ludovic Quistin
398. Luke James
399. Luke McAlister
400. Luwian language
401. Macron Stadium
402. Mako (actor)
403. Ma Liang (Three Kingdoms)
404. Mama Said Knock You Out (song)
405. Marco Delvecchio
406. Marjorie Main
407. Mary Pickford (politician)
408. Mary Printz
409. Master of the Mornauer Portrait
410. Material (band)
411. Mathern Palace
412. Matt Gutierrez
413. Matt Targett
414. Maya Lin
415. McCormick Tribune Plaza & Ice Rink
416. McMillan Plan
417. Media ownership in Canada
418. Mega Man III (Game Boy)
419. Megapode
420. Melodifestivalen 2007
421. Memphis, Egypt
422. Mercury Prize
423. Methotrexate
424. Michael Brennan (photographer)
425. Michael Polenske
426. Microfilament
427. Middlesbrough South and East Cleveland (UK Parliament constituency)
428. Mie goreng
429. Mihailo Petrović
430. Mike Douglas
431. Mikhail Rudy
432. Milan Ristić
433. Million Dollar Legs (1932 film)
434. Missouri State University
435. Moneybomb
436. Monsour del Rosario
437. Mortal Kombat 4
438. Mount Tambora
439. Mount Tremper
440. MTS system architecture
441. Muhammad Abu Khubza
442. MV Cemfjord
443. Mycena galopus
444. My Chief and My Regiment
445. Nailsea and Backwell railway station
446. Nathuram Mirdha
447. National Cycle Route 44
448. National flag
449. Nation Party (Turkey, 1948)
450. Neil Harris (footballer, born 1977)
451. Newcomb's snail
452. New South Wales 48 class locomotive
453. Nick Bacon
454. Nick Waplington
455. Nicole Cooke
456. Nikolai Noskov
457. Nirad C. Chaudhuri
458. Noble savage
459. Nobody's Fool (1994 film)
460. Norman Kirkman
461. Norteños
462. Northrop YF-23
463. North Yorkshire Moors Railway
464. Novica Veličković
465. Nuclei Armati Rivoluzionari
466. Nuts (1987 film)
467. OBike (Taiwan)

468. Occitan cross
469. O. G. S. Crawford
470. Ole von Beust
471. Olivier (comics)
472. Omega Psi Phi
473. Order of battle for Operation Barbarossa
474. Orfeo ed Euridice
475. Organizational culture
476. Orr (surname)
477. OS-9
478. Outer space
479. Owain Jones
480. Oxandrolone
481. På minuten
482. Paparoa National Park
483. Pasquale Cannone
484. Pat Coombs
485. Patricia Hornsby-Smith, Baroness Hornsby-Smith
486. Patrick Hillery
487. Paul Dacre
488. Paul Pierce (American football)
489. Payment card industry
490. Pentecostal Foreign Mission of Norway
491. Peter Arguindegui
492. Peter Weir
493. PFC Cherno More Varna
494. Phaerimm
495. Philip Bailey
496. Philippine passport
497. Philip Seymour Hoffman
498. Phoroneus
499. Pierre Gaveaux
500. Pituitary apoplexy
501. Plain tobacco packaging
502. Playhouse 90
503. Politics Can Be Different
504. Politics of Flanders
505. Polvo
506. Porroglossum
507. Post-industrial society
508. Presidential \$1 Coin Program
509. Prince Georg of Hanover
510. Princess Anastasia (1986)
511. Princess Elisabeth of Saxe-Altenburg (1865–1927)
512. Progressive music
513. Protein phosphorylation
514. Provisional Constitutional Order
515. Pseudocapacitance
516. Punnagai Mannan
517. Pure Gold (various artists compilation album)
518. Races of StarCraft
519. Racing Extinction
520. Rapastinel
521. Ray Burke (Irish politician)
522. Ray Jacobs
523. ReCAPTCHA
524. Redline (2007 film)
525. Reginald Judson
526. Rey de Reyes (2017)
527. Ricardo Ortega Fernández
528. Ricardo Villa
529. Richard de Bury
530. Richard Fleischer
531. Richard Glücks
532. Rickon Stark
533. Rise of Nations: Rise of Legends
534. RMS Tahiti
535. Robert Geathers
536. Robert Gordon University - Garthdee campus
537. Robert R. McCormick
538. Robert Whitehead
539. Rob Moore (politician)
540. Rockingham County, Virginia
541. Rock N Roll McDonald's
542. Rolls-Royce BR700
543. Roman Catholic Archdiocese of Canberra and Goulburn
544. Roseanna Vitro
545. Ross Muir
546. Route 417 (Israel)
547. Route nationale 7
548. Roy Innis
549. Roy Williams (safety)
550. RS-232
551. Rubicon Global
552. Rüdiger Fikentscher
553. Ruth Brown
554. Sachtleben Chemie
555. Sahib Singh Sokhey
556. Samuel D. Ingham
557. Samuel May Williams House
558. Samuel S. Cox
559. Samuel Sutton
560. San Angelo Army Air Field
561. San Luis Reservoir
562. Sarah Borwell
563. Sara Tancredi
564. Satellite
565. Seaport Music Festival
566. Seneca, South Carolina
567. Sergei Shamba
568. Shakespeare Theatre of New Jersey

569. Shirlington, Arlington, Virginia
570. Shkodër
571. Shoma Uno
572. Shoot 'Em Up (film)
573. Shooting of Collin Rose
574. Siege of Fort Erie
575. Signorini problem
576. Simon Harris
577. Sir Gabriel Goldney, 1st Baronet
578. Sir James Heath, 1st Baronet
579. Skaro
580. Smile
581. SMIM23
582. Soletsky District
583. Solomon Adeniyi Babalola
584. Sonic the Hedgehog (1991 video game)
585. Soule
586. Southside Railroad (Virginia)
587. Spider-Man (1967 TV series)
588. Stages (Jimi Hendrix album)
589. Standard Motor Company
590. State highways in Virginia
591. St Columba's Convent, Dalby
592. Steep, Hampshire
593. Stephen Chow
594. Steve Cotterill
595. St. Vincent (film)
596. Summer Lake Wildlife Area
597. Sunghursh
598. Supercharger
599. Superheater
600. Supervixen
601. Syed Murad Ali Shah
602. Syed Sheh Hassan Barakbah
603. Symphony No. 1 (Walton)
604. Syncron International AB
605. System of a Down
606. Tagiades japeretus
607. Taguchi methods
608. Tamasha (soundtrack)
609. Taoyuan, Taiwan
610. Ted Barrett
611. Teenage Mutant Ninja Turtles (2003 TV series) (season 2)
612. Telecommunications in Kenya
613. Terpenoid
614. Terry Dodson
615. Terry W. Virts
616. The Baxters
617. The Best Classics... Ever! vol. 2
618. The Cape (2011 TV series)
619. The Devil in Miss Jones
620. The Heavy (band)
621. The History Boys (film)
622. The Incredible World of James Bond
623. Theology of John Calvin
624. The Political Cesspool
625. The PTA Disbands
626. Therosaurus
627. The Taste of Money
628. The Ten Commandments (1923 film)
629. The Voice UK (series 3)
630. The Wombles (band)
631. The Woolpack
632. Thomas Cranmer
633. Three lookouts
634. Tiergarten (park)
635. Timeline of the Deepwater Horizon oil spill (May 2010)
636. Timo Liekoski
637. Titanium Man
638. Toe
639. Tokyo 6th district
640. Tolomeo (horse)
641. Tom Cora
642. Tommy McLaren
643. Tony McWalter
644. Tony Mendez
645. Tornrak
646. Totnes railway station
647. Transporter 2
648. Transport Integration Act 2010
649. Trap Lord
650. Travis Mays
651. Treaty of Mutual Cooperation and Security between the United States and Japan
652. T. Ryder Smith
653. Tuborg GreenFest
654. Tupolev Tu-124
655. Two of Diamonds (album)
656. Tyrant
657. UCC Philosophical Society
658. Unforgettable (Philippine TV series)
659. University of Detroit Mercy
660. Uranium mining in the United States
661. USS Mahan (DD-364)
662. USS Norfolk (DL-1)
663. USS Pegasus (AK-48)
664. USS Reedbird (AMS-51)
665. USS Satterlee (DD-626)
666. Utah Motorsports Campus
667. Uveal melanoma
668. Van Alen Building
669. Velimir Perasović
670. Verliebt in Berlin

671. Vincent J. Donehue
672. Visual arts education
673. Vito Rizzuto
674. Wang Meng (Former Qin)
675. Wayne Williams
676. Wei Zhaodu
677. Wes Fletcher
678. West Vancouver Memorial Library
679. WFHM-FM
680. Whitehorse Ranch
681. Wieland Förster
682. Wielbark culture
683. William B. Ide
684. William Boyd Carpenter
685. William Torrey Harris
686. Winchester Model 1897
687. Win Win (film)
688. Wisley
689. WNBJ-LD
690. World Expo 88
691. World of Ghost in the Shell
692. Wulfric Spot
693. Yavne-Yam
694. Yolanda Marculescu
695. Yoshi's Woolly World
696. Youth Code
697. Yo Yogi!
698. Zhangjiakou–Hohhot High-Speed Railway
699. Zone of the Enders: The 2nd Runner
700. Zoran Milinković (footballer)

Παράρτημα Β

Εργαλεία

Σε αυτό το παράρτημα παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση των δύο συστημάτων. Και τα δύο συστήματα γράφτηκαν στη γλώσσα Python, η οποία έχει ένα μεγάλο αριθμό από βιβλιοθήκες που αφαιρούν φόρτο από τον προγραμματιστή, όπως βιβλιοθήκες επεξεργασίας φυσικής γλώσσας, επικοινωνίας με API και κατασκευής machine learning συστημάτων.

B.1 Βασικό μοντέλο

B.1.1 Mediawiki API

Το MediaWiki API ειναι μια διαδικτυακη υπηρεσια που προσφερει προσβαση στα δεδομένα της Wikipedia. Η χρήση του API γίνεται απλά στέλνοντας μία ερώτηση HTTP για κάποιο URL, μέσω του web browser ή κάποιας γλώσσας προγραμματισμού. Παρακάτω δίνεται ένα χαρακτηριστικό παράδειγμα της μορφής του URL, το οποίο επιστρέφει ένα έγγραφο μορφής JSON που περιλαμβάνει το wikitext για τη σελίδα με τίτλο "Main Page":

```
https://en.wikipedia.org/w/api.php?action=query&titles>Main  
%20Page&prop=revisions&rvprop=content&format=json
```

Συγκεκριμένα για την Python, αντί να κατασκευάζονται τα αντίστοιχα HTTP requests, είναι ευκολότερο και πιο αποδοτικό να χρησιμοποιηθεί ένα από τα υπάρχοντα interfaces που υπάρχουν για το MediaWiki API. Επιλέχθηκε το mwclient που περιείχε όλες τις λειτουργίες που χρειάστηκαν για την εργασία. Συγκεκριμένα, η κλάση site=mwclient.Site(site_name) παρέχει λειτουργίες για ένα mediawiki site με hostname το site_name, στην

περίπτωσή μας η αγγλική Wikipedia 'en.wikipedia.org' και με την κλάση `page = mwclient.page.Page(site, title)` λαμβάνουμε πληροφορίες της σελίδας με τίτλο 'title', για την ιστοσελίδα 'site' που ορίστηκε προηγουμένως. Χρησιμοποιούμε τις παρακάτω μεθόδους αυτών των κλάσεων:

- `site.random()`: Ένας generator που επιστρέφει τυχαίες σελίδες της Wikipedia
- `page.text()`: Επιστρέφει το wikitext της τρέχουσας σελίδας
- `page.redirects_to()`: Αν η σελίδα είναι σελίδα ανακατεύθυνσης, επιστρέφει την σελίδα στην οποία οδηγεί
- `page.backlinks(namespace=0)`: Επιστρέφει μια λίστα με τα άρθρα της Wikipedia που έχουν συνδέσμους προς την τρέχουσα σελίδα.

B.1.2 Wikipedia parsers

Κάποιες πληροφορίες που χρειαζόμαστε από τη Wikipedia είτε δεν προσφέρονται με εύκολο τρόπο από το MediaWiki API είτε οδηγούν σε προβλήματα απόδοσης λόγω της εξάρτησης από πολλά HTTP requests. Σε αυτές τις περιπτώσεις είναι προτιμότερο να βρεθούν οι πληροφορίες από ένα Wikipedia dump. Για το πρώτο σύστημα χρησιμοποιήθηκε το dump από τον Απρίλιο του 2017 και δύο tools.

- **wikidump**: Το `wikidump` είναι ένα εργαλείο για την εξαγωγή πληροφοριών από το XML dump της Wikipedia. Εκτελώντας το με την εντολή `python -m wikidump FILE [FILE ...] OUTPUT_DIR` επεξεργάζεται το Wikipedia dump αρχείο που ορίζεται από την παράμετρο `FILE` και εξάγει χαρακτηριστικά στον φάκελο που ορίζεται από την παράμετρο `OUTPUT_DIR` σε ένα αρχείο `FILE.features.xml`. Διευκολύνει την παραγωγή υποψήφιων οντοτήτων όπως περιγράφτηκε στην ενότητα 3.1.2.

Τμήμα κώδικα 3: Οι πρώτες σειρές του parsed XML dump της Wikipedia από το `wikidump`. Στην πρώτη σειρά φαίνεται το πεδίο που αντιστοιχεί σε κάθε comma-separated value. Κάθε επόμενη σειρά αντιστοιχεί σε ένα σύνδεσμο και τις τιμές του για τα πεδία που ορίστηκαν στην πρώτη γραμμή.

```

1 page_id,page_title,revision_id,revision_parent_id,revision_timestamp,user_type, [
  ↵ user_username,user_id,revision_minor,wikilink.link,wikilink.anchor, [
    ↵ wikilink.section_name,wikilink.section_level,wikilink.section_number
2 10,AccessibleComputing,767284433,631144794,2017-02-25T00:30:28Z,registered, [
    ↵ Godsy,23257138,0,Computer accessibility,Computer accessibility,-----
    ↵ incipit -----,0,0
3 12,Anarchism,772891928,772621616,2017-03-29T23:51:28Z,registered,Sk4rpHed1n, [
    ↵ 13952908,0,political philosophy,political philosophy,----- incipit
    ↵ -----,0,0

```

```

4   12,Anarchism,772891928,772621616,2017-03-29T23:51:28Z,registered,Sk4rpHedln,J
    ↳ 13952908,0,self-governance,self-governed,----- incipit
    ↳ -----,0,0
5   12,Anarchism,772891928,772621616,2017-03-29T23:51:28Z,registered,Sk4rpHedln,J
    ↳ 13952908,0,stateless society,stateless societies,----- incipit
    ↳ -----,0,0

```

- **wikiextractor:** Το wikiextractor είναι ένα script που καθαρίζει το κείμενο ενός Wikipedia database dump. Εκτελείται με παράμετρο ένα αρχείο Wikipedia dump και αποθηκεύει την έξοδό του σε πολλά μικρότερα αρχεία στο directory που προσδιορίζεται. Στην εργασία χρησιμοποιήθηκε για να εξάγουμε το κείμενο των άρθρων της wikipedia διατηρώντας τους συνδέσμους με σκοπό να υπολογιστεί η πιθανότητα κάθε αναφοράς του dataset να αποτελεί σύνδεσμο όπως περιγράφηκε στην ενότητα 3.1.3.2. Αυτό έγινε με την η εντολή:

```
python WikiExtractor.py --output ./TEXT --links  
--no-templates
```

Όπου η σημασία των παραμέτρων είναι η εξής:

- **--output ./TEXT:** Ορίζεται το ”./TEXT” ως directory των αρχείων εξόδου.
- **--links:** Διατηρούνται οι σύνδεσμοι στο κείμενο.
- **--no-templates:** Δεν επεκτείνονται τα MediaWiki template, επιταχύνοντας σημαντικά το parsing.

B.2 Συνδυαστικό Μοντέλο

B.2.1 Berkeley Entity Resolution System

Το Berkeley Entity Resolution System [33] είναι ένα συνδυαστικό σύστημα αναγνώρισης οντοτήτων, επίλυσης συναναφοράς (coreference resolution) και αποσαφήνισης οντοτήτων με ένα πλούσιο σε χαρακτηριστικά μοντέλο.

Τμήματα του Berkeley Entity Resolution System χρησιμοποιήθηκαν στην εργασία ώστε να παραχθεί το σύνολο υποψήφιων οντοτήτων και να εξαχθούν τα αραιά χαρακτηριστικά, τα οποία ενσωματώνονται στη συνέχεια στα τοπικά χαρακτηριστικά του συστήματός μας. Αρχικά δίνεται ως είσοδος το Wikipedia dump και το dataset, και παράγεται μια μικρότερη βάση

δεδομένων που είναι σχετική με το συγκεκριμένο dataset. Αυτό γίνεται χρησιμοποιώντας την κλάση WikipediaInterface και παράγεται το αρχείο wikipedia-interface.ser.gz που χρειάζεται στο επόμενο βήμα.

Έχοντας το wikipedia interface και τις σελίδες της Wikipedia που αντιστοιχούν στις οντότητες του dataset, χρησιμοποιείται η κλάση JointQueryDenotation Chooser η οποία δημιουργεί το αρχείο dev-set-queries.json. Το αρχείο αυτό αποτελεί ένα dictionary, στο οποίο αναφερόμαστε ως queries, με κλειδιά τμήματα κειμένου και τιμές dictionaries για κάθε αναφορά του, με κλειδί τα συμφραζόμενα γύρω από αυτή και τιμές πληροφορίες όπως υποψήφιες οντότητες, χαρακτηριστικά της αναφοράς και των υποψήφιων οντοτήτων της, αν χρησιμοποιείται στην εκπαίδευση ή αξιολόγηση του συστήματος και την ή τις ονομασίες της σωστής οντότητας. Τα χαρακτηριστικά των ερωτημάτων και των υποψήφιων οντοτήτων είναι αριθμοί που δηλώνουν τον δείκτη του αραιού χαρακτηριστικού που ικανοποιούν από το dictionary featureNames. Τμήμα του dictionary queries φαίνεται στο τμήμα κώδικα 4 παρακάτω.

Τμήμα κώδικα 4: Παράδειγμα ζεύγους key-value από το dictionary των queries

```
"AFTER THE BELL - After hours slows in light volume . NEW YORK 1996-08-27 Traders said
on Tuesday after-hours activity was light . Both WorldCom Inc and MFS Communications
Co Inc were trading but they moved in line with their close . WorldCom , which said it
will buy MFS , shed 1-3/4 to close at 21 while MFS lost 3-8/16 to close at 41-5/16 .
The New York Stock Exchange said its session one volume was 5,700 shares compared to
53,400 shares Monday . Session two volume was 4,153,800 shares compared to no volume
Monday . The American Stock Exchange said there was no after-hours activity .": {
    "WorldCom , which said it will buy [MFS] , shed 1-3/4 to close at 21 while MFS
lost": {
        "vals": {
            "Tivo's Media File System": [0, [[36],[36],[36],[36],[36]], 0, [],[], 0],
            "Octaviar": [
                0, [[20, 22, 27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44, 45, 49, 50, 228,
                229, 230, 231, 232, 240],[20, 22, 27, 28, 29, 30, 31, 32, 33, 34, 35, 42,
                43, 44, 45, 49, 50, 228, 229, 230, 231, 232, 240],[20, 22, 27, 28, 29, 30,
                31, 32, 33, 34, 35, 42, 43, 44, 45, 49, 50, 228, 229, 230, 231, 232,
                240],[20, 22, 27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44, 45, 49, 50,
                228, 229, 230, 231, 232, 240],[36]], 0, [],[], 0
            ],
            "-NIL-": [0, [[15],[15],[15],[15],[52]], 0, [],[], 0],
            "XXNILXX": [0, [[36],[36],[36],[36],[36]], 0, [],[], 0],
            "Major facilitator family": [0, [[36],[36],[36],[36],[36]], 0, [],[], 0],
            "Miletich Fighting Systems": [
                0, [[27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44, 45, 47, 48, 49, 50, 228,
                229, 230, 231, 232, 243],[27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44,
                45, 47, 48, 49, 50, 228, 229, 230, 231, 232, 243],[27, 28, 29, 30, 31, 32,
                33, 34, 35, 42, 43, 44, 45, 47, 48, 49, 50, 228, 229, 230, 231, 232,
                243],[27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44, 45, 47, 48, 49, 50,
                228, 229, 230, 231, 232, 243],[36]], 0, [],[], 0
            ],
            "": []
        }
    }
}
```

```

"Syriac Military Council": [
    0, [[27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44, 45, 47, 48, 49, 50, 228,
        229, 230, 231, 232, 234],[27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44,
        45, 47, 48, 49, 50, 228, 229, 230, 231, 232, 234],[27, 28, 29, 30, 31, 32,
        33, 34, 35, 42, 43, 44, 45, 47, 48, 49, 50, 228, 229, 230, 231, 232,
        234],[27, 28, 29, 30, 31, 32, 33, 34, 35, 42, 43, 44, 45, 47, 48, 49, 50,
        228, 229, 230, 231, 232, 234],[36]], 0, [],[], 0
],
"MFS": [0, [[36],[36],[36],[36],[36]], 0, [],[], 0],
"MFS (label)": [...],
"TiVo Media File System": [...],
"Major facilitator superfamily": [...],
"Metropolitan Fiber Systems": [...],
"MFS Limited": [...],
"Macintosh File System": [...],
"MFS Investment Management": [...]
},
"query_vals": [
    [0, 1, 222, 9683, 13259, 5, 6, 7],
    [8, 1, 222, 9683, 13259, 5, 6, 9],
    [8, 1, 222, 9683, 13259, 5, 10, 7],
    [8, 1, 222, 9683, 13259, 5, 10, 9],
    [8, 11, 225, 558, 13260]
],
"training": "True",
"gold": ["Metropolitan Fiber Systems"]
},
"Both WorldCom Inc and [MFS Communications Co Inc] were trading but they moved in
line with their close": [...],
"[WorldCom] , which said it will buy MFS , shed 1-3/4": [...],
"buy MFS , shed 1-3/4 to close at 21 while [MFS] lost 3-8/16 to close at 41-5/16
.": [...],
"[NEW YORK] 1996-08-27": [...],
"The [American Stock Exchange] said there was no after-hours activity .": [...],
"The [New York Stock Exchange] said its session one volume was 5,700 shares
compared to": [...],
"Both [WorldCom Inc] and MFS Communications Co Inc were trading but they moved":
[...]
}

```

B.2.2 Word2vec

Όπως αναφέρθηκε, για την αναπαράσταση των λέξεων χρησιμοποιείται το μοντέλο Word2vec. Η εκπαίδευση και εκτέλεση του συστήματος που είναι υπεύθυνο για την δημιουργία των Word2vec απαιτεί ως είσοδο μια συλλογή κειμένων βάσει της οποίας παράγονται τα διανύσματα των λέξεων ανάλογα με τις σημασιολογικές σχέσεις που εντοπίζονται μεταξύ τους. Εκτός αυτού, δέχεται διάφορες άλλες παραμέτρους που επηρεάζουν τον τρόπο εκπαίδευσης και τη μορφή του αρχείου εξόδου. Για την εργασία αυτή επιλέχθηκαν ως εξής:

- **-train:** Το αρχείο εισόδου για την εκπαίδευση, που ορίστηκε να είναι το wikipedia dump.
- **-binary:** Καθορίζει αν το αρχείο εξόδου είναι σε binary μορφή. Δόθηκε 1, δηλαδή αληθές.
- **-window:** Το μέγεθος του παραθύρου των συμφραζομένων κατά την εκπαίδευση. Μεγαλύτερα παράθυρα παράγουν πιο σημασιολογικά εστιασμένα διανύσματα [41]. Ορίστηκε να είναι ίσο με 21.
- **-negative:** Το word2vec για να εξοικονομήσει χρόνο στην εκπαίδευση χρησιμοποιεί την τεχνική του negative sampling. Με βάση αυτή την τεχνική, αντί να ανανεώνονται τα βάρη όλων των αρνητικών λέξεων ενός παραδείγματος, επιλέγεται μόνο ένας τυχαίος αριθμός αυτών. Για παράδειγμα έχοντας την φράση "cat food" για τη λέξη "cat", η σωστή έξοδος του δίκτυου του word2vec είναι ένα 1-hot διάνυσμα στο οποίο ο νευρώνας που αντιστοιχεί στη λέξη "food" πρέπει να είναι 1 και όλες οι υπόλοιπες να είναι 0. Σε κάθε παράδειγμα εκπαίδευσης θα πρέπει να ενημερώνεται ένας τεράστιος αριθμός βαρών και το πρόβλημα αυτό λύνεται ενημερώνοντας εκτός από τη σωστή λέξη, ένα μικρό μόνο τμήμα των "αρνητικών" λέξεων που αποτελείται συνήθως από 5-20 λέξεις. Για το σύστημά μας ορίστηκε να είναι 10.
- **-min-count:** Ορίζει τον αριθμό των εμφανίσεων μιας λέξης κάτω από τον οποίο οι λέξεις θα απορρίπτονται. Επιλέχθηκε να είναι 10.
- **-size:** Οι διαστάσεις των word2vec διανυσμάτων που παράγονται. Ορίστηκε να είναι 300.