Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Σημασιολογική Ομοιότητα Κειμένων - Notion Oriented Approach (NOA)** **Text Semantic Similarity - Notion Oriented Approach (NOA)**
Ονοματεπώνυμο Φοιτητή	**Θεόδωρος Παπασωτηρίου**
Πατρώνυμο	**Γεώργιος**
Αριθμός Μητρώου	**ΜΠΣΠ/ 11012**
Επιβλέπων	**Θεμιστοκλής Παναγιωτόπουλος, Καθηγητής**

Ημερομηνία Παράδοσης **Οκτώβριος 2013**

Περιεχόμενα

Θα ήθελα να ευχαριστήσω ιδιαίτερα την ερευνήτρια επεξεργασίας φυσικής γλώσσας του University of North Texas, **Rada Mihalcea** και το **ΕΚΕΦΕ Δημόκριτος**, για τις συμβουλές και την καθοδήγηση που μου προσέφεραν στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας και συγκεκριμένα τον τομέα της σημασιολογικής ομοιότητας κειμένων.

1. Περίληψη

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP) είναι ένας τομέας της τεχνητής νοημοσύνης που συντροφεύει την πληροφορική από τα πρώτα της χρόνια. Χωρίζεται σε διάφορες επιμέρους κατηγορίες με επικρατέστερες αυτές κυρίως της μετάφρασης, των αλγορίθμων διόρθωσης της ορθογραφίας που χρησιμοποιούν οι περισσότεροι πλέον κειμενογράφοι, και τους αλγορίθμους αναγνώρισης φωνής ή μετατροπής κειμένου σε φωνή. Βέβαια, αν και αυτές οι κατηγορίες είναι πιο γνωστές εμπορικά στο ευρύ κοινό, η κοινότητα της τεχνητής νοημοσύνης που έχει επιλέξει να ασχοληθεί με την επεξεργασία φυσικής γλώσσας, έχει να διαλέξει μέσα από μια πληθώρα επιπλέον πεδίων, όπως η αναγνώριση οντοτήτων μέσα από φυσικό κείμενο (Named Entity Recognition), η δημιουργία περιλήψεων, οι αυτόματες απαντήσεις σε μικρού μήκος και όχι μόνο ερωτήματα, η σημασιολογική ομοιότητα μεταξύ κειμένων, κ.α.. Η σημασιολογική ομοιότητα κειμένων είναι αυτή που διαπραγματεύεται η παρούσα εργασία και ίσως αφορά ένα από τα πιο πολύπλοκα προβλήματα που καλείται να λύσει η επιστημονική κοινότητα της NLP. Αν και τα πρώτα μοντέλα δημιουργήθηκαν πριν από αρκετά χρόνια, οι διάφορες προσεγγίσεις αλλάζουν ανά μεγάλα χρονικά διαστήματα, προσπαθώντας να βρουν την πιο αξιόπιστη και απτή λύση στο παρών πρόβλημα.

Η τρέχουσες προσεγγίσεις, έχουν κατευθυνθεί σε μοντέλα που στηρίζονται κυρίως σε μετρικές δύο μεγάλων κατηγοριών. Έχουμε λοιπόν μετρικές βασισμένες στη γνώση (Knowledge-Based) και μετρικές βασισμένες σε λεκτικό δείγμα (Corpus-Based). Αυτές οι

μετρικές χρησιμοποιούνται είτε μεμονωμένα, είτε συνδυαστικά στις εκάστοτε προσεγγίσεις για το πρόβλημα της σημασιολογικής ομοιότητας των κειμένων. Αν και προτείνονται συνεχώς νέα μοντέλα δεν έχει βρεθεί κάποιο, το οποίο να μπορεί να απαντήσει με απόλυτη αξιοπιστία και σιγουριά αν δύο κείμενα έχουν σημασιολογική ομοιότητα μεταξύ τους. Τα περισσότερα μοντέλα επιστρέφουν μία κανονικοποιημένη τιμή η οποία ορίζει το ποσοστό ομοιότητας μεταξύ των κειμένων και χρησιμοποιείται κυρίως για κατηγοριοποίηση κειμένων βάσει του θέματος που αναφέρονται. Τα βέλτιστα αποτελέσματα φτάνουν σε απόδοση κοντά στο 85% των κειμένων, ποσοστό αρκετά ικανοποιητικό, αλλά μη ικανό να δημιουργήσει εφαρμογές για παραγωγή αξιόπιστων μοντέλων βασισμένων σε φυσική γλώσσα.

Η παρούσα εργασία πραγματεύεται ακριβώς αυτό το πρόβλημα, προτείνοντας ένα νέο μοντέλο τύπου Knowledge-Based, το οποίο σε αντίθεση με τα καθιερωμένα μοντέλα που ως Βάση Γνώσης χρησιμοποιούν το WordNet ή τη Wikipedia, προτείνει μια νέα Βάση Γνώσης, η οποία καλύπτει στο μέγιστο τις ανάγκες του. Η Βάση Γνώσης αυτή έχει τις βάσεις της στη δουλειά των Collins & Quillian (Collins & Quillian, 1969), και αναπαριστά όλη τη γνώση των λέξεων και των εννοιών που συνδέονται ως ένα σημασιολογικό δίκτυο. Με βάση τα παραπάνω και μια εφαρμογή κανόνων που βασίζονται στα Hidden Markov Models (HMMs) το μοντέλο που προτείνεται μπορεί να εξάγει μία τιμή αληθείας για το αν δύο κείμενα είναι σημασιολογικά όμοια ή όχι με μεγαλύτερη αξιοπιστία, αφού σέβεται τους όποιους γραμματικούς και συντακτικούς κανόνες ορίζει η εκάστοτε γλώσσα για περαιτέρω αποσαφήνιση της σημασιολογίας των προτάσεών της.

Abstract

Natural Language Processing is a field of Artificial Intelligence that accompanies Computer Science through it's early steps. It bifurcates into various branches with most important these of translation, spelling errors correction that most modern text processors use and speech recognition or Text-To-Speech algorithms. Even though these are the most known to the public categories, the community of Artificial Intelligence that works on Natural Language Processing, has a huge variety of sub-domains to choose, which also includes Named Entity Recognition, Summary Extraction, Auto Question Answering, Text Semantic Similarity, etc. Text Semantic Similarity is the one that this dissertation deals with and probably is one of the most complex problems that researchers of NLP tries to solve. Even if the first models were created enough years ago, the various approaches change periodically, in a journey of seeking for the most robust and feasibly, optimized solution in the current problem.

The current approaches, have proposed models based on metrics of two basic categories, Knowledge-Based and Corpus-Based metrics. Those metrics have been used solely or combined in various approaches that deal with the Text Semantic Similarity problem. Up to now, none of the proposed models can answer with total reliability and confidence if two texts are semantically identical. Most models return a normalized value that defines the percentage of similarity between input texts and this techniques are used

mainly for text categorization or clustering, depending on their subject. Best systems evaluation results up to 85% of texts, which is a good percentage, but not adequate for applications that will produce reliable models based in natural language.

The current dissertation deals with this problem, proposing a new Knowledge-Based model, which, unlike the majority of other models that use WordNet or Wikipedia as a Knowledge Base, proposes a new Knowledge Base, which covers in depth its functionality. This Knowledge Base is based on the work of Collins & Quillian, and represents the words and the notions that are related to, as a semantic network. With the Knowledge Base described above and the application of rules, based on Hidden Markov Models (HMMs) the proposed model can result a boolean value that will define if the two input texts are semantically similar or not, with respect in grammatical and syntactical rules of the language that has been used for the examined texts, which play a huge role in further disambiguation of the semantic meaning of a sentence.

2. Εισαγωγή

Πολλά χρόνια τώρα το ανθρώπινο είδος προσπαθεί να οργανώσει όλη τη γνώση που έχει σε ένα σημείο. Οι εγκυκλοπαίδειες είναι αυτές που έχουν μια πιο γενική προσέγγιση. Τα πρώτα χρόνια της έρευνας της Τεχνητής Νοημοσύνης, οι Buchanan & Feigenbaum ορίσανε τη γνώση ως ισχυρή υπόθεση, που απαιτεί "Η δύναμη ενός έξυπνου προγράμματος να υλοποιεί διεργασίες εξαρτάται αρχικά στην ποσότητα και στην ποιότητα της γνώσης που έχει σχετικά με τη διεργασία αυτή" (Buchanan & Feigenbaum, 1982). Όταν οι υπολογιστές πρέπει να υλοποιήσουν διεργασίες που χρειάζονται ανθρώπινη νοημοσύνη, είναι φυσικό να χρησιμοποιούν μια εγκυκλοπαίδεια για να ανακαλέσουν τη διαθέσιμη γνώση. Ωστόσο υπάρχουν κάποια εμπόδια που χρήζουν αρκετά δύσκολο να υλοποιηθεί κάτι τέτοιο. Πρώτον, η γνώση αυτή είναι διαθέσιμη σε μορφή κειμένου και για να χρησιμοποιηθεί από τον υπολογιστή απαιτεί κατανόηση της φυσικής γλώσσας, ένα βασικό πρόβλημα από μόνο του. Επιπλέον, η κατανόηση της γλώσσας μπορεί να μην είναι αρκετή, αφού τις περισσότερες φορές τα κείμενα που γράφονται από ανθρώπους υποθέτουν ότι ο αναγνώστης έχει ένα μεγάλο ποσό από κοινώς δεδομένη γνώση, και επομένως την παραλείπουν.

Από τα παραπάνω είναι εμφανές ότι η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) είναι το κλειδί για το επόμενο βήμα της Τεχνητής

Νοημοσύνης και όχι μόνο. Τα τελευταία χρόνια το πεδίο αυτό κερδίζει όλο και περισσότερα βλέμματα και ερευνητές, σε μια προσπάθεια δόμησης της Φυσικής Γλώσσας, την οποία χρησιμοποιούν όλοι οι άνθρωποι στην καθημερινότητά τους. Η Επεξεργασία Φυσικής Γλώσσας ανήκει στο ευρύτερο πεδίο της Τεχνητής Νοημοσύνης (Artificial Intelligence - A.I.) και εστιάζει σε θέματα πληροφορικής όπου ο υπολογιστής αλληλεπιδρά με κάποια ανθρώπινη φυσική γλώσσα. Μερικά από τα θέματα με τα οποία ασχολείται η NLP είναι η αυτόματη δημιουργία περίληψης(Automatic Summarization), μηχανική μετάφραση (Machine Translation), αναγνώριση οντοτήτων εντός του κειμένου (Named-Entity Recognition), εξόρυξη πληροφορίας μέσα από απλό κείμενο (Information Extraction), οπτική αναγνώριση χαρακτήρων (Optical Character Recognition - OCR), ηχητική αναγνώριση κειμένου (Speech Recognition), αναγνώριση θέματος (Topic Recognition), ερωταπαντήσεις (Question Answering), κ.α.

Στην παρούσα εργασία, εστιάζουμε σε έναν ανοιχτό τομέα της NLP, ο οποίος έχει να κάνει με την σημασιολογική ομοιότητα των κειμένων (Text Semantic Similarity) και αν και υπάρχουν αρκετές προσπάθειες και προσεγγίσεις για τη λύση προβλημάτων τέτοιου τύπου, δεν υπάρχει κάποια προσέγγιση η οποία να είναι απόλυτα αξιόπιστη και να έχει γενική χρήση. Το πρόβλημα ανεύρεσης της σημασιολογικής ομοιότητας μεταξύ δύο κειμένων είναι από τα πιο πολύπλοκα προβλήματα στον τομέα της επεξεργασίας φυσικής γλώσσας και οι υπάρχουσες προσεγγίσεις χρησιμοποιούν μεγάλο πλήθος τεχνικών και διαδικασιών για να το επιλύσουν. Ετησίως, διοργανώνεται ένα συνέδριο, γνωστό ως SEMEVAL (Semantic Evaluation) *(SEMEVAL 2013, 2013)*, που παρουσιάζονται νέες προσεγγίσεις και τεχνικές που επιδιώκουν να λύσουν το παραπάνω πρόβλημα.

Η πλειοψηφία των περισσότερων προτάσεων που έχουν γίνει μέχρι σήμερα σχετικά με το πρόβλημα εστιάζουν στο να ορίσουν ένα ποσοστό ομοιότητας μεταξύ δύο κειμένων ή δύο προτάσεων και χρησιμοποιούνται με μεγάλη επιτυχία κυρίως στην κατηγοριοποίηση των κειμένων (clustering). Ωστόσο, αξιόπιστες λύσεις στο συγκεκριμένο πρόβλημα, θα μπορούσαν να ανοίξουν νέους δρόμους σε διάφορους τομείς με την μετατροπή της φυσικής γλώσσας, σε δομημένη, κατανοητή από τον υπολογιστή. Μερικά από αυτά θα μπορούσαμε να πούμε ότι είναι η αυτόματη και ταχύτατη συγκέντρωση της γνώσης και των ειδήσεων, η εξόρυξη χρήσιμης πληροφορίας μέσα από κείμενο, η δημιουργία νέων μοντέλων μέσα από κείμενο, κ.λπ.

2.1. Ιστορική Αναδρομή

Η ιστορία της ξεκινάει περίπου στο 1950, όπου ο Alan Turing εκδίδει ένα άρθρο με τίτλο "Computing Machinery and Intelligence" *(Tuning, 1950)* (Μηχανήματα Υπολογισμών και Ευφυΐας), το οποίο έμεινε γνωστό ως "Turing Test" και θεωρείται κριτήριο εύρεσης ευφυΐας σε μηχανές. Το συγκεκριμένο τεστ, δοκιμάζει την ικανότητα μιας μηχανής, να παρουσιάσει ευφυΐα, αντίστοιχη με της ανθρώπινης. Ένας άνθρωπος εμπλέκεται σε μια ομιλία φυσικής γλώσσας με μια μηχανή, η οποία έχει σχεδιαστεί ώστε να εξομοιώνει την ανθρώπινη συμπεριφορά. Ένας επιπλέον άνθρωπος, ορίζεται ως κριτής και καλείται να βρει με βεβαιότητα ποιος είναι ο άνθρωπος και ποια η μηχανή. Οι συμμετέχοντες είναι σε διαφορετικά σημεία ο ένας από τον άλλον και δεν έχουν οπτική επαφή. Φυσικά, δεν κρίνεται η ορθότητα των απαντήσεων, αλλά μόνο πόσο κοντά είναι στην ανθρώπινη συμπεριφορά. Αν ο κριτής δεν μπορεί να ξεχωρίσει με βεβαιότητα την μηχανή, τότε η μηχανή έχει περάσει επιτυχώς τη δοκιμασία. Από τότε ξεκίνησε ένας αγώνας των ερευνητών ώστε να καταφέρουν να περάσουν με επιτυχία το τεστ αυτό.

Παράλληλα, με την εξάπλωση των τηλεπικοινωνιών και του διαδικτύου, ο κόσμος καλωσόριζε την παγκοσμιοποίηση και νέες ανάγκες ήρθαν στο προσκήνιο. Το πείραμα στην Georgetown από το τοπικό πανεπιστήμιο και την IBM, το 1954, *(IBM Translator, 1954),* εμπεριέχει πλήρης αυτόματη μετάφραση από περισσότερες από 60 Ρώσικες προτάσεις, στα Αγγλικά. Οι συγγραφείς ισχυρίζονταν ότι σε 3 με 5 χρόνια η μηχανική

μετάφραση θα είναι λυμένο πρόβλημα. Ωστόσο, η πραγματική πρόοδος ήταν πολύ πιο αργή και μετά την αναφορά της ALPAC το 1966, όπου όριζε ότι η δεκαετής έρευνα απέτυχε να εκπληρώσει τις προσδοκίες, η χρηματοδότηση για την μηχανική μετάφραση (machine translation) μειώθηκε δραματικά. Λίγο παραπάνω έρευνα στο συγκεκριμένο τομέα, διεξάχθηκε μέχρι το τέλος του 1980, όπου το πρώτο σύστημα στατιστικής μηχανικής ανάλυσης δημιουργήθηκε.

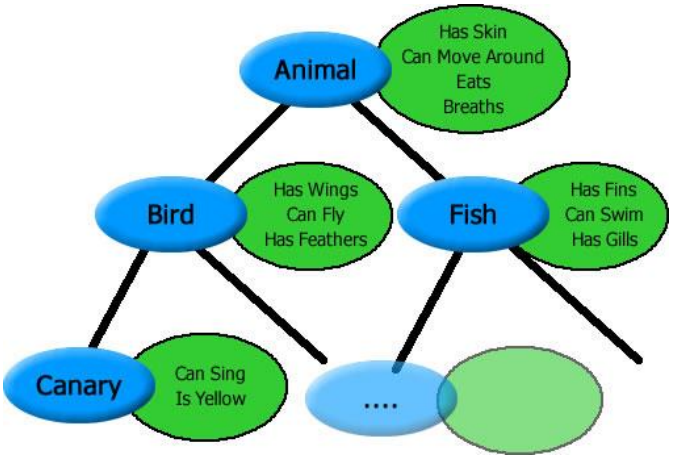
Μερικά ακόμη επιτυχημένα NLP συστήματα δημιουργήθηκαν το 1960-1970, όπως το SHRDLU *(SHRDLU)* και το ELIZA *(Joseph, 1966)*, που χρησιμοποιώντας σχεδόν καμία πληροφορία για την ανθρώπινη σκέψη ή συναίσθημα, παρείχαν μια μερικώς ανθρώπινη αλληλεπίδραση. Κατά το 1970 πολλοί προγραμματιστές άρχισαν να γράφουν εννοιολογικές οντολογίες (conceptual ontologies), όπου δόμησαν πληροφορία του πραγματικού κόσμου σε κατανοητά από τον υπολογιστή δεδομένα. Παραδείγματα είναι το MARGIE *(Schank, Goldman, Rieger, & Riesbeck, 1973)*, SAM *(Schank & Abelson, 1978)* και αρκετά άλλα. Εκείνη την περίοδο γράφτηκαν και πολλά chatterbots (προγράμματα εξομοίωσης ανθρώπινου συνομιλητή σε chat).

Μέχρι το 1980, τα περισσότερα συστήματα NLP βασιζόντουσαν σε πολύπλοκα σετ κανόνων, γραμμένων χειροκίνητα. Από το 1980 και μετά ξεκίνησε μια επανάσταση στην NLP με την εισαγωγή των αλγορίθμων μηχανικής μάθησης (machine learning) για γλωσσική επεξεργασία. Μερικοί από τους πρώτους αλγορίθμους μηχανικής μάθησης, όπως τα δέντρα αποφάσεων, παρήγαγαν συστήματα από if-then κανόνες, παρόμοιους με αυτούς που χρησιμοποιούνταν ως τώρα. Η περισσότερη όμως έρευνα στράφηκε προς στατιστικά μοντέλα, τα οποία έπαιρναν αποφάσεις βάσει πιθανοτήτων και τοποθετώντας βάρη στα χαρακτηριστικά των δεδομένων εισόδου.

Πολλές από τις πρώτες επιτυχίες έγιναν στο πεδίο της μηχανικής μετάφρασης, ειδικά στην IBM, όπου πιο πολύπλοκα στατιστικά μοντέλα δημιουργήθηκαν. Αυτά τα συστήματα μπορούσαν να πάρουν το πλεονέκτημα από μια υπάρχουσα, πολύγλωσση βάση κειμένου που είχε παραχθεί από τη βουλή του Καναδά και την Ευρωπαϊκή Ένωση, σαν αποτέλεσμα της μετάφρασης των νόμων σε όλες τις επίσημες γλώσσες.

Πιο πρόσφατες έρευνες έχουν εστιάσει κυρίως σε μη-επιβλεπόμενους (unsupervised) ή ημι-επιβλεπόμενους (semi-supervised) αλγορίθμους μάθησης. Τέτοιοι αλγόριθμοι μπορούν να μαθαίνουν από δεδομένα που δεν έχουν εξ αρχής τις επιθυμητές απαντήσεις, ή που χρησιμοποιούν έναν συνδυασμό σχολιασμένων και μη σχολιασμένων δεδομένων.

Ο κλάδος της Σημασιολογικής Ομοιότητας ξεκινάει περίπου στο 1968 *(Quillian, 1967)*, όπου ο M. Quillian προτείνει ένα μοντέλο για να καταχωρηθούν σημασιολογικές πληροφορίες στη μνήμη ενός υπολογιστή. Σε αυτό το μοντέλο, κάθε λέξη έχει αποθηκεύεται μαζί με μια ρύθμιση των δεικτών (pointers) σε άλλες λέξεις στη μνήμη. Αυτή η ρύθμιση αναπαριστά τη σημασία της τρέχουσας λέξης. Για παράδειγμα η λέξη "canary" (καναρίνι) θα έχει έναν δείκτη στη λέξη "bird" (πουλί), που θα είναι το όνομα της κατηγορίας ή το superset της λέξης "canary", καθώς θα έχει και 2 δείκτες στις 2 ιδιότητες: "yellow" (κίτρινο) και "can sing" (μπορεί να τραγουδήσει). Πληροφορίες που ισχύουν για όλα τα πουλιά δεν χρειάζεται να αποθηκευτούν σαν δείκτες στον κόμβο για κάθε είδος πουλιών χωριστά. Δηλαδή, η πληροφορία ότι "canary can fly" μπορεί να υπονοηθεί βρίσκοντας ότι το καναρίνι είναι πουλί και ότι τα πουλιά μπορούν να πετάξουν. Επομένως, στη λέξη "bird" χρειάζεται ένας δείκτης στο "can fly" (μπορεί να πετάξει), κ.ο.κ, όπως φαίνεται και στο παρακάτω σχήμα *(Collins & Quillian, 1969)*:



Αυτή είναι και η πρώτη αναπαράσταση της φυσικής γλώσσας σε δομημένη μορφή από υπολογιστή.

Μία άλλη σημαντική δημοσίευση στην ιστορία του text semantic similarity είναι αυτή των A. Collins και E. Loftus *(Collins & Loftus, A spreading-activation theory of semantic processing, 1975)*, η οποία βασίζεται στην παραπάνω θεωρία του Quillian για τη σημασιολογική μνήμη και δημιουργήθηκε για να αποσαφηνίσει παρερμηνείες που είχαν δημιουργηθεί από την ευρεία χρήση της, κυρίως από τη γνωστική οικονομία (cognitive economy). Ουσιαστικά, μπορεί να θεωρηθεί σαν μια επιπλέον δημοσίευση που εδραίωσε το μοντέλο του Quillian για αναπαράσταση της ανθρώπινης γνώσης σε υπολογιστή.

Αν και οι παραπάνω δημοσιεύσεις δεν συγκρίνουν την ομοιότητα της σημασιολογίας μεταξύ δύο κειμένων, κατάφεραν να ανοίξουν τον δρόμο σε περαιτέρω αναζητήσεις στον τομέα της σημασιολογικής ομοιότητας κειμένων.

3. Εργαλεία και Τεχνικές

Όπως προαναφέρθηκε, η σημασιολογική ομοιότητα μεταξύ δύο κειμένων είναι ένα αρκετά σύνθετο πρόβλημα το οποίο αναλύεται σε επιμέρους βήματα και κάνει χρήση αρκετών εργαλείων και τεχνικών. Πριν προχωρήσουμε λοιπόν στην επεξήγηση και κατανόηση ορισμένων προσεγγίσεων κρίνεται απαραίτητο ο αναγνώστης να έχει εφοδιαστεί με τη γνώση των βασικών εργαλείων και τεχνικών που θα χρησιμοποιηθούν.

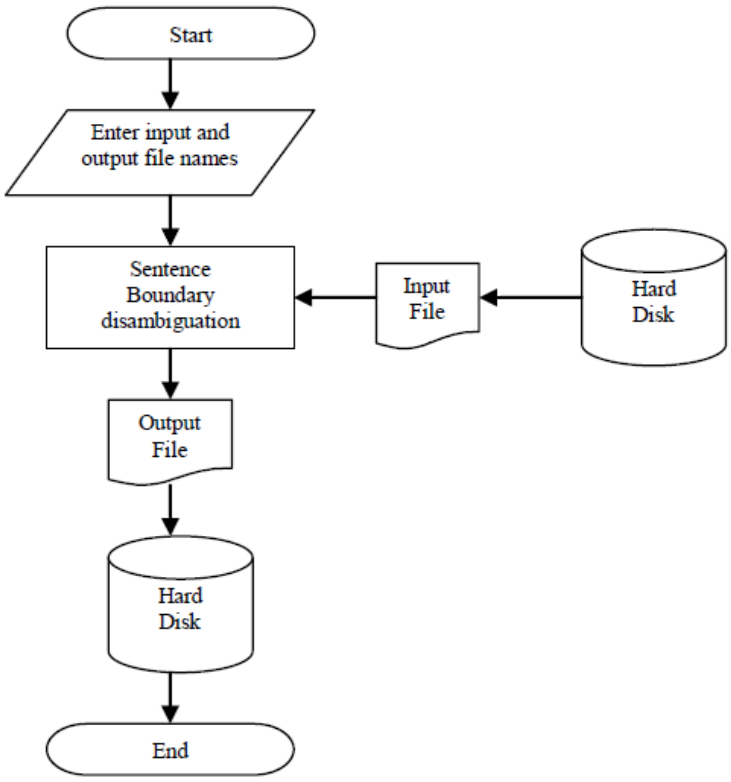
3.1. Sentence Boundary Disambiguation- S.B.D.

Γνωρίζουμε από τη γλωσσολογία ότι κάθε κείμενο αποτελείται από προτάσεις και φράσεις. Το τέλος μιας πρότασης σηματοδοτείται από κάποιο σημείο στίξης όπως: '.', '?' '!', κ.α. Στην πλειοψηφία τους οι προτάσεις χρησιμοποιούν το σημείο στίξης της τελείας για να δηλώσουν το τέλος τους. Ωστόσο, τα σημεία στίξης που αναφέραμε δεν σηματοδοτούν μόνο το τέλος μιας πρότασης, αλλά μπορεί να χρησιμοποιούνται και με διαφορετικούς τρόπους μέσα σε αυτή. Ένα παράδειγμα είναι οι συντομογραφίες, που χρησιμοποιούν την τελεία, χωρίς να

υποδηλώνουν ότι η πρόταση κλείνει. Επομένως δημιουργείται αμφιβολία για το πότε τα σημεία στίξης ορίζουν το τέλος μιας πρότασης ή χρησιμοποιούνται με διαφορετικό τρόπο. Αυτή την αμφιβολία καλούνται να αποσαφηνίσουν οι τεχνικές που είναι γνωστές ως Sentence Boundary Disambiguation (S.B.D.).

Για την αποσαφήνιση των ορίων της πρότασης έχουν προταθεί κατά διαστήματα διάφορες τεχνικές. Οι περισσότερες από αυτές εστιάζουν γύρω από την ίδια διαδικασία, όπου χρησιμοποιούνται κανόνες με τη μορφή regular expressions και διάφορες εξαιρέσεις για να οριοθετήσουν τις προτάσεις του κειμένου.

Μια ενδιαφέρουσα και απλή τεχνική είναι αυτή που προτείνεται από τους Pritam Singh Negi, M.M.S Rauthan και H.S Dhami (Negi, Rauthan, & Dhami, 2010). Η διαδικασία που ακολουθείται για την αποσαφήνιση των ορίων των προτάσεων του κειμένου προϋποθέτει το κείμενο να εισαγχθεί σαν είσοδος στο σύστημα. Το σύστημα πρώτα ελέγχει για συχνές λέξεις που περιέχουν σημεία στίξης (mr., mrs., κ.α.) Για να γίνει αυτό επεξεργάζεται το αρχείο κειμένου χαρακτήρα-χαρακτήρα μέχρι να εντοπίσει πλήρης ταίριασμα με τη ζητούμενη λέξη. Τότε, μετατρέπει τη λέξη αυτή σε μια καινούργια στην οποία θα λείπουν τα σημεία στίξης. Κατόπιν ελέγχει με τον ίδιο τρόπο για δεκαδικούς αριθμούς, οι οποίοι διαχωρίζονται με τελεία. Επόμενο βήμα είναι να ελεγχθεί αν υπάρχουν ταιριάσματα για ιδιάζουσες μορφές όπως ο υπερ-σύνδεσμος μιας ιστοσελίδας (www.google.com) ή συνεχόμενα σημεία στίξης (Ohh, My God!!!!). Αφού εντοπίσει τα σημεία όπου τα σημεία στίξης δεν υποδηλώνουν τέλος πρότασης, δημιουργεί ένα νέο κείμενο όπου τα έχει αφαιρέσει ή αντικαταστήσει όπως φαίνεται στο παρακάτω διάγραμμα ροής.



Η παραπάνω τεχνική παρουσιάζει κάποια προβλήματα, όπως στην περίπτωση όπου αρχικά που χρησιμοποιούν τελείες ως διαχωριστικά, βρίσκονται στο τέλος της πρότασης, ή υπάρχουν αντίστοιχα σημεία στίξης σε υπο-προτάσεις (π.χ. μέσα σε παρενθέσεις ή εισαγωγικά). Οι D. Palmer και M. Hearst προτείνουν μια ενδιαφέρουσα τεχνική για αποτελεσματική εφαρμογή Sentence Boundary Disambiguation σε κείμενα, η οποία δεν

προϋποθέτει καταγραφή όλων των πιθανών κανόνων, αλλά και εφαρμόζεται σχετικά γρήγορα, παρέχοντας αξιόπιστα αποτελέσματα (Palmer & Hearst, 1994).

Ο πυρήνας του αλγορίθμου που προτείνουν, βασίζεται στην παρακάτω λογική: Οι πιθανότητες του μέρους του λόγου (Part-of-Speech) που ορίζει τις λέξεις που περιβάλλουν ένα σημείο στίξης, χρησιμοποιούνται σαν είσοδος σε ένα νευρωνικό δίκτυο του οποίου τα αποτελέσματα ορίζουν την ετικέτα του σημείου στίξης. Επομένως, είναι χρήσιμο να καταγραφεί η πιθανότητα για κάθε λέξη να εμφανίζεται πριν ή μετά το τέλος μιας πρότασης. Επειδή όμως κάτι τέτοιο απαιτεί αρκετό χρόνο και μεγάλη Βάση Γνώσης, η διαδικασία μεταφέρεται αντίστοιχα στις πιθανότητες για το αντίστοιχο μέρος του λόγου (Part-of-Speech), να είναι πριν ή μετά το τέλος μιας πρότασης. Η προσέγγιση αυτή φαίνεται να λειτουργεί κυκλικά γιατί τα περισσότερα συστήματα POS tagging απαιτούν να έχει εφαρμοστεί Sentence Boundary Disambiguation τεχνική. Για το λόγο αυτό, ο αλγόριθμος χρησιμοποιεί τις πιθανότητες για όλα τα πιθανά POS tags μόνο των λέξεων που μας ενδιαφέρουν, χωρίς να προσπαθεί να ορίσει το σωστό, αφήνοντας αυτή τη διαδικασία για το Part-of-Speech Tagging που θα δούμε αναλυτικότερα παρακάτω.

3.2. Tokenization

Αφού οριστούν τα όρια της πρότασης είναι σύνηθες, πριν γίνει οποιαδήποτε ανάλυση και διαδικασία να χρειάζεται να εντοπιστούν ξεκάθαρα οι λέξεις του κειμένου. Η απλότητα της αναγνώρισης των αγγλικών λέξεων μέσα από κενά διαστήματα ως διαχωριστικά παραπλανεί ώστε να παραβλέπουμε την πολυπλοκότητα σε άλλες μονάδες των αγγλικών, όπως ιδιωματισμοί και συγκεκριμένες εκφράσεις. Αυτή η πολυπλοκότητα είναι ακόμη μεγαλύτερη όταν πρόκειται για άλλες γλώσσες όπως τα Κινέζικα, τα οποία δεν έχουν διαχωριστικού χαρακτήρες (όπως το κενό διάστημα) μεταξύ των λέξεων. Απαιτείται λοιπόν μια πιο σύνθετη διαδικασία, ώστε να εντοπιστούν όλες οι υπο-μονάδες του κειμένου (συνήθως είναι λέξεις), η οποία είναι γνωστή ως Tokenization.

Πιστεύεται ότι οι ιδιωματισμοί και κάποιες εκφράσεις, πρέπει να παίρνονται σαν ολόκληρες μονάδες λέξης και να μην αποσυντίθενται σε μικρότερες, όπως στις λέξεις που τα αποτελούν. Ωστόσο, σε κάποιες περιπτώσεις, όπως όταν οι ιδιωματισμοί δεν είναι συνεχόμενοι, θα ήταν χρήσιμο να τους αποσυνθέσουμε στις επιμέρους λέξεις και να παρθούν αυτές ως tokens.

Η εύρεση των λέξεων μέσα στην πρόταση απαιτεί το ταίριασμα των γραμματοσειρών που της αναπαριστούν με αντίστοιχες σε μια λίστα λέξεων που έχουμε ως Βάση Γνώσης. όπως ένα λεξικό. Κάθε αυτόματη μέθοδος τμηματοποίησης τέτοιου τύπου ορίζεται από τρεις παράγοντες όπως φαίνεται στο μοντέλο Automatic Segmenting Method, ASM(d,a,m) όπου, $d \in \{+1, -1\}$ και υποδεικνύει την κατεύθυνση αναζήτησης για το ταίριασμα με το λεξικό, $a \in \{+1, -1\}$ και υποδεικνύει την προσθήκη ή παράληψη χαρακτήρα σε κάθε γύρω που ταιριάζει μια λέξη και $m \in \{+1, -1\}$ που υποδεικνύει τη χρήση το μέγιστου ή ελάχιστου ταιριάσματος αντίστοιχα.

Οι J. Webster και C. Kit (Webster & Kit, 1992), προτείνουν μια μέθοδο όπου η Βάση Γνώσης δημιουργείται μέσα από ένα εκπαιδευόμενο νευρωνικό δίκτυο το οποίο αναγνωρίζει τις επιμέρους λεκτικές μονάδες από τα συνοδευτικά τους και τις σχέσεις τους. Προτείνεται να εκπαιδευτεί ένα νευρωνικό δίκτυο για να αναγνωρίζει τα tokens βάσει των σχέσεων με τα συνοδευτικά τους. Αφού τελειώσει η εκπαίδευση, το νευρωνικό δίκτυο θα μπορεί να κάνει tokenization και αποσαφήνιση ταιριάζοντας τα δεδομένα εισόδου με τα πρότυπα που έμαθε. Κάθε οντότητα στο λεξικό είναι παράλληλα ένα "όλο" αλλά και ένα "στιγμιότυπο". Είναι δηλαδή, η αντίληψη της σύνθεσης της από τα μέρη της και παραγωγή κάποιου μοντέλου, αλλά και η αντιστοίχιση της σε μια λεκτική σταθερά.

Η απλούστερη προσέγγιση για την αναγνώριση των συστατικών λεκτικών μονάδων είναι το ταίριασμα table-look-up. Υποθέτει ότι υπάρχει ήδη μια λίστα από λεκτικές μονάδες ως δείγματα σε ικανοποιητικό αριθμό. Κατόπιν, αρχικά αποσυνθέτει κάθε λέξη, συνεχίζει τα ταιριάσματα για να βρει αν υπάρχουν και άλλες συστατικές λεκτικές μονάδες ανάμεσα σε αυτές τις λέξεις. Αυτή η μέθοδος μπορεί να αναγνωρίσει αγγλικούς ιδιωματισμούς και άλλα συστατικά λεκτικά τμήματα που είναι συνεχόμενα, αλλά δεν μπορεί να εφαρμοστεί σε μη-συνεχόμενα. Για αυτή την περίπτωση υπάρχει η Generalized Table-look-up μέθοδος .

Για παράδειγμα η έκφραση "keep in mind" είναι στάνταρ έκφραση και το "in mind" είναι συνήθως μέρος ενός μεγαλύτερου token. Μεταξύ αυτών των δύο μερών υπάρχει μια φράση ουσιαστικού (Noun Phrase - NP), η οποία συνήθως είναι μικρή (keep [NP] in mind). Πρέπει να καθορίσουμε λοιπόν την φράση ουσιαστικού (NP) . Διαδικασίες σαν αυτή πρέπει να χρησιμοποιήσουν συντακτική ανάλυση. Γι αυτό το λόγο χρειάζεται να ενσωματωθεί στο table-look-up μερικό parsing. Εδώ φαίνεται η σημασία της δομικής ανάλυσης στην αναγνώριση των συστατικών λεκτικών μονάδων. Εκτός από τη δομική ανάλυση, χρειάζεται γνώση για τις λεκτικές μονάδες αυτές, όπως που πρέπει να τοποθετηθεί η φράση ουσιαστικού (NP) στο μη συνεχόμενο "keep.... in mind".

Αφού γίνουν τα παραπάνω επιτυχώς έχουμε μια πλήρη αναγνώριση όλων των τμημάτων που αποτελούν τις προτάσεις είτε αυτά είναι λέξεις, είτε ιδιωματισμοί, είτε φράσεις. Το μεγαλύτερο όμως πλεονέκτημα της παραπάνω τεχνικής είναι η ικανότητά της να αντιληφθεί όλες τις λεκτικές μονάδες, ακόμη και στην περίπτωση που δεν είναι συνεχόμενες. Φυσικά, το κόστος είναι ότι κάποιος πρέπει να καταχωρήσει σε μια βάση γνώσης τον τρόπο που συντάσσονται.

3.3. Stemming

Η Βάση Γνώσης που χρησιμοποιεί η τεχνική του tokenization, στηρίζεται σε ένα μεγάλο λεξικό, το οποίο έχει γνώση όλων των λέξεων που θα συναντήσουμε στο κείμενο. Αυτό σημαίνει, ότι δύο λέξεις για παράδειγμα, οι οποίες έχουν την ίδια ρίζα, αλλά διαφοροποιούνται στις καταλήξεις, θα πρέπει να υπάρχουν και οι δύο στη Βάση Γνώσης για να γίνει η ταυτοποίηση. Αυτό σημαίνει αρκετός επιπλέον κόπος αν σκεφτούμε ότι κάθε ουσιαστικό έχει και πληθυντικό αριθμό, όπου διαφοροποιείται ελάχιστα. Αντίστοιχα κάθε ρήμα διαφοροποιεί την κατάληξή του ανάλογα με τον χρόνο που χρησιμοποιείται, και κάθε επίθετο διαφοροποιείται ανάλογα τον βαθμό που χρησιμοποιείται. Τα παραδείγματα όπου λέξεις με κοινή ρίζα, διαφοροποιούνται ελάχιστα μεταξύ τους είναι αρκετά, και γίνονται ακόμη περισσότερα αν λάβουμε υπόψη και τη δυνατότητα της σύνθεσης νέων λέξεων, παραγόμενες από υπάρχουσες, μέσα στο κείμενο (π.χ. googler).

Με τη διαδικασία του Stemming μπορούμε να έχουμε μόνο την κεντρική λέξη που χρησιμοποιείται ως ρίζα, στη Βάση Γνώσης. Όταν λοιπόν, στο κείμενο συναντήσουμε λέξεις οι οποίες παράγονται από μια άλλη λέξη (foxes - fox), τότε με επεξεργασία όπως, αφαίρεση των καταλήξεων και των προθεμάτων, μπορούμε να βρούμε αν η λέξη ρίζα υπάρχει στη Βάση Γνώσης που χρησιμοποιείται σαν σημείο αναφοράς, και να τις συσχετίσουμε. Ο πιο γνωστός αλγόριθμος που κάνει την παραπάνω διαδικασία είναι ο Porter Stemming Algorithm (Porter, 1997), ο οποίος μπορεί να οριστεί με τα παρακάτω βασική βήματα:

1. Χωρίζει τη λέξη σε πιθανά μορφήματα
2. Παίρνει την ενδιάμεση μορφή της
3. Ταιριάζει τις ρίζες σε κατηγορίες σχετικά με την έννοια
4. Βρίσκει τη σωστή φόρμα

Πιθανώς το καλύτερο κριτήριο για να αφαιρέσουμε τις καταλήξεις από δύο λέξεις W_1 και W_2 για να παράγουμε τον κορμό S, είναι να βρούμε την κοινή γραμματοσειρά των λέξεων. Αλλά αν οι λέξεις είναι το "Relate" και το "Relativity" τότε θα καταλήξουμε στη λέξη "relat" που δεν βγάζει κάποιο νόημα. Από την άλλη, αν οριοθετηθούν κανόνες του τύπου να

αφαιρεθεί η κατάληξη "-er" γιατί δηλώνει συγκριτικό βαθμό όπως στο "small" - "smaller", κινδυνεύουμε να την αφαιρέσουμε και από σημεία που δεν θα έπρεπε όπως το "wander", όπου θα καταλήξουμε σε τελείως διαφορετικό νόημα. Υπάρχει ένα στάδιο κατά τη δημιουργία ενός προγράμματος που αφαιρεί τις καταλήξεις, κατά το οποίο η προσθήκη κανόνων για να αυξηθεί η απόδοση σε μια περιοχή του λεξιλογίου καταλήγει να μειώνει την απόδοση ανάλογα σε μια άλλη περιοχή.

Ο αλγόριθμος του Porter για το Stemming προτείνει αρχικά να μετατρέψουμε τη λέξη σε μαθηματικό τύπο όπου αναπαριστούμε τα σύμφωνα με το γράμμα **c** (consonant) και τα φωνήεντα με το γράμμα **v** (vowel). Αν έχουμε περισσότερα από ένα σύμφωνα στη σειρά τα γράφουμε με **C** και πολλά φωνήεντα με **V**. Αντίστοιχα και τις ακολουθίες συμφώνων-φωνηέντων όταν επαναλαμβάνονται τις γράφουμε σε μορφή δύναμης.

$$[C]\,(VC)^m\,[V]$$

Οι αγκύλες ορίζουν την υποχρεωτική παρουσία του περιεχομένου

Οι κανόνες για αφαίρεση της κατάληξης θα δίνονται στη μορφή:

$$(\text{condition})\ S_1 \rightarrow S_2$$

Αυτό σημαίνει ότι αν μια λέξη τελειώνει με κατάληξη S_1 και το τμήμα της πριν το S_1 ικανοποιεί μια συνθήκη, τότε το S_1 αντικαθιστάται από το S_2. Στους κανόνες που επικαλύπτονται διαλέγουμε αυτόν που αφορά τη μεγαλύτερη ακολουθία χαρακτήρων όπως στο παράδειγμα CARESSES που θα πρέπει να βρει το CARESS με τους παρακάτω κανόνες:

SSES →SS

IES → I

SS → SS

S →

Επομένως, με τις τεχνικές stemming η Βάση Γνώσης μειώνεται σημαντικά. Ωστόσο αυτό δεν γίνεται χωρίς κάποιο τίμημα, το οποίο στη συγκεκριμένη περίπτωση είναι επιπλέον υπολογιστικός χρόνος για την εφαρμογή του αλγόριθμου.

3.4. Word Sense Disambiguation (WSD)

Μία λέξη μπορεί να έχει παραπάνω από μία χρήσεις και ερμηνείες. Η διαφορετική ερμηνεία της λέξης ονομάζεται word-sense και καθορίζεται από την πρόταση που βρίσκεται και τον τρόπο που χρησιμοποιείται. Για παράδειγμα, η λέξη fire, μπορεί να σημαίνει φωτιά, μπορεί να σημαίνει πυροβολώ, απολύω, κ.α. Η διαδικασία καθορισμού του σωστού word-sense ονομάζεται Word Sense Disambiguation.Υπάρχουν αρκετές μέθοδοι οι οποίες μπορούν να βοηθήσουν στην αποσαφήνιση του word-sense, οι περισσότερες εκ των οποίων βασίζονται σε στατιστική ανάλυση και εύρεση προτύπων μεταξύ των λέξεων και των αντίστοιχων σημασιών τους πάνω σε ένα μεγάλο λεκτικό δείγμα. Οι μέθοδοι με χρήση εύρεσης προτύπων χρειάζονται και ένα λεκτικό δοκιμαστικό δείγμα, γιατί τα πρότυπα συνήθως βρίσκονται με τη χρήση αλγορίθμων μηχανικής μάθησης.

3.5. Part-of-Speech Tagging (POS Tagging)

Οι λέξεις, ανάλογα με το word-sense που έχουν στην εκάστοτε πρόταση, όπως και το σημείο στο οποίο εμφανίζονται μέσα στην πρόταση, έχουν και έναν αντίστοιχο συντακτικό

προσδιορισμό. Θα πρέπει λοιπόν να δοθεί μια ετικέτα για τη σωστή χρήση της λέξης στην πρόταση. Το Part-of-Speech Tagging είναι μια τεχνική που χρησιμοποιείται για να δώσει την ετικέτα που ορίζει το μέρος του λόγου που αναπαριστά κάθε λέξη της πρότασης.

3.6. STOP-Words

Σε κάθε σωστή αγγλική πρόταση υπάρχουν πολλές λέξεις οι οποίες δεν συμμετέχουν στο να καθοριστεί η σημασία της πρότασης αυτής. Λέξεις όπως το "for", "of" και το "to" χρειάζονται μόνο για να βγάζει νόημα η πρόταση σαν σύνολο και ώστε να υπάρχει σωστή συντακτική δομή, αλλά δεν περιγράφουν κάποιο συγκεκριμένο νόημα. Δηλαδή, χρησιμοποιούνται μόνο για να βοηθήσουν στην κατανόηση της πρότασης μέσα από τους γνωστούς συντακτικούς κανόνες. Για παράδειγμα στην πρόταση "the car is fast", οι λέξεις "car" και "fast" θεωρούνται το βασικό μέρος της πρότασης, ενώ οι υπόλοιπες περιγράφουν την σύνδεση μεταξύ των δύο βασικών λέξεων. Οι λέξεις αυτές που δεν συνεισφέρουν στον ορισμό του νοήματος της πρότασης, ονομάζονται STOP-Words και αφαιρούνται από το κείμενο στις περισσότερες τεχνικές σημασιολογικής ομοιότητας κειμένων.

3.7. WordNet

Σε πολλές διαδικασίες της NLP καταλήγουμε να χρειαζόμαστε μια Βάση Γνώσης με τις λέξεις που υπάρχουν σε κάποιο λεξικό και μια σύνδεση μεταξύ τους. Σε αυτές τις διαδικασίες, η πλειοψηφία των τεχνικών που έχουν προταθεί ανά διαστήματα κάνουν χρήση του WordNet. Το WordNet είναι ένα λεξικό της Αγγλικής αρχικά γλώσσας που δημιουργήθηκε από το πανεπιστήμιο του Princeton για να ορίσει συνδέσεις μεταξύ τεσσάρων ειδών μερών του λόγου (ουσιαστικά, ρήματα, επίθετα, επιρρήματα). Περιέχει μία περιγραφή των λέξεων και διάφορες συσχετίσεις κάθε λέξης προς άλλες. Οι συσχετίσεις αυτές είναι τύπου "συνωνυμία" (synonymy), "αντωνυμία" (antonymy), "υπονυμία" (hyponymy), κ.α.

Υπονυμία είναι η κατηγορία και **Hyponym** ορίζεται το αντικείμενο που ανήκει σε αυτή. (π.χ. Hyponym: vehicle, Hypernym: car). Αυτή η σχέση ονομάζεται **IS-A** σχέση (car **IS-A** vehicle). Ένας άλλος τύπος σχέσης είναι το **PART-OF**, που χρησιμοποιεί τις κλάσεις **Meronym** και **Holonym** (π.χ.: Meronym: wheel, Holonym: car). Τα ρήματα σχετίζονται χρησιμοποιώντας **Troponyms**, όπου είναι μια συγκεκριμένη περίπτωση που σημαίνει άλλο ρήμα (π.χ.: το "duel" είναι troponym του "fight"). Επειδή το WordNet ορίζει μια ιεραρχική δομή συσχέτισης όλων των λέξεων, μπορεί να θεωρηθεί ως μία ταξονομία (taxonomy).

Κάθε λεξικό, οποιασδήποτε γλώσσας, μπορεί να οριστεί ως ένα σύνολο φορμών, που κάθε μία σχετίζεται με μία ή περισσότερες έννοιες (senses) οι οποίες ονομάζονται synsets. Κάθε synset έχει ένα gloss το οποίο περιγράφει την έννοια που αναπαριστά. Αν μια φόρμα έχει περισσότερες από μία έννοιες, είναι πολυσήμαντη (polysemous). Αντίθετα, αν δύο λέξεις μοιράζονται την ίδια έννοια (sense), τότε είναι συνώνυμες (synonymous). Στο WordNet οι συνώνυμες λέξεις τοποθετούνται μαζί σε ομάδες που ονομάζονται synsets.

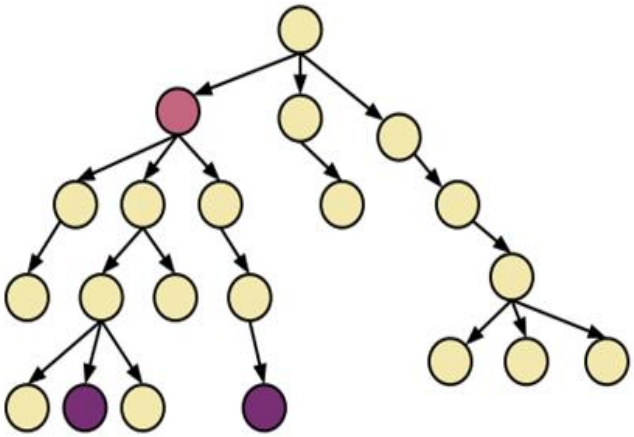
Όπως αναφέρθηκε και παραπάνω το WordNet έχει λέξεις από τέσσερις συντακτικές κλάσεις (ουσιαστικά, ρήματα, επίθετα, επιρρήματα). Εκτός από μια σύνδεση μεταξύ ουσιαστικών και επιθέτων, δεν υπάρχει σύνδεση μεταξύ των ξεχωριστών κλάσεων, και

επομένως η σύγκριση μεταξύ λέξεων διαφορετικών κλάσεων δεν είναι εφικτή μέσω του WordNet. Επίσης, το WordNet δεν έχει όλες τις μορφές ενός ρήματος (π.χ. run, runs, ran, running), αλλά μόνο τη ρίζα/βάση της λέξης (run). Αν κάποιο ρήμα χρησιμοποιείται σε διαφορετικό χρόνο, θα πρέπει να εφαρμοστεί τεχνική stemming για να γίνει η συσχέτιση.

Στην ταξονομία των ουσιαστικών, κάθε synset είναι συνδεδεμένο με τουλάχιστον ένα ακόμη synset και περιέχει μία ρίζα (root). Η ταξονομία είναι δομημένη σαν δένδρο. Το στοιχείο ρίζα του δένδρου ονομάζεται οντότητα (entity).

3.8. Lowest Common Ancestor (LCA)

Σε δενδρικές ταξονομίες όπως το WordNet μπορούμε να συγκρίνουμε την σημασιολογική ομοιότητα, ή πιο σωστά τη συνάφεια δύο λέξεων, αν μετακινηθούμε στην ταξονομία από τους 2 κόμβους που καλούμαστε να συγκρίνουμε προς τους κόμβους γονείς, έως ότου βρούμε κάποιον κοινό κόμβο. Αυτός ο κοινός κόμβος που θα βρεθεί ονομάζεται Lowest Common Ancestor (LCA) και όταν ο αριθμός των βημάτων που χρειάζονται για βρεθεί είναι μικρός, μπορούμε να αντιληφθούμε ότι και οι δύο λέξεις/κόμβοι που πυροδότησαν τον αλγόριθμο αναζήτησής του, είναι κοντά στην ταξονομία και επομένως συναφής.



Σε T δένδρο με n κόμβους, το LCA μεταξύ δύο κόμβων v και w ορίζεται ως ο κοντινότερος κόμβος στο T που έχει το v και w ως απογόνους. Ο LCA των v και w στο δένδρο T, είναι ο κοινός πρόγονός τους, που εντοπίζεται πιο μακριά από τη ρίζα του δένδρου. Πιο αναλυτικά, σε μια δομή δένδρου, όπου κάθε κόμβος οδηγεί στους γονείς του, ο LCA μπορεί να οριστεί εύκολα, βρίσκοντας την πρώτη διχοτόμηση των διαδρομών από τους κόμβους v και w αντίστοιχα, προς τη ρίζα του δένδρου (Aho, Hopcroft, & Ullman, 1973). Ο χρόνος υπολογισμού του αλγορίθμου αυτού είναι O(h), όπου h είναι το ύψος του δένδρου. Υπάρχουν βέβαια και περιπτώσεις, όπου ο αλγόριθμος μπορεί να τροποποιηθεί χάρη στη δομή του δένδρου, ώστε να βελτιωθεί η απόδοσή του.

3.9. Levenshtein Distance (Edit Distance) Algorithm

Η απλούστερη μετρική για να βρούμε την συνάφεια μεταξύ δύο λέξεων είναι ο αλγόριθμος του Levenshtein που είναι γνωστός και ως Edit Distance. Ο συγκεκριμένος αλγόριθμος ελέγχει πόσα βήματα γίνονται ώστε να μετατραπεί μια λέξη σε μια άλλη. Δηλαδή, μετράει πόσα γράμματα από την λέξη W_1 θα προστεθούν/ αφαιρεθούν/ αντικατασταθούν ώστε να καταλήξουμε στην λέξη W_2. Για παράδειγμα, η απόσταση Levenshtein μεταξύ των λέξεων "apple" και "maple" είναι 2. Το αποτέλεσμα του αλγορίθμου αυτού δεν μπορεί να είναι

μεγαλύτερο από το μέγεθος της μεγαλύτερης εκ των δύο λέξεων που συγκρίνονται. Ωστόσο, οι τιμές που μπορεί να επιστραφούν ποικίλουν και δυσκολεύουν τη χρήση τους. Επομένως, κρίνεται χρήσιμο να κανονικοποιηθούν, ώστε να επιστρέφουν τιμές μεταξύ 0 και 1, όπου 1 θα σημαίνει απόλυτη ομοιότητα μεταξύ των λέξεων και 0 ότι δεν υπάρχει καμία ομοιότητα. Η συνάρτηση που προτείνεται από τον Levenshtein (Levenshtein, 1966), είναι η παρακάτω:

$$\mathbf{sim}_{LD}(W_1,W_2) = 1 - \frac{\mathrm{edit}(W1,W2)}{\max(|W1|,|W2|)}$$

Ωστόσο, όταν έχουμε να συγκρίνουμε σημασιολογική ομοιότητα μεταξύ λέξεων, αυτός ο αλγόριθμος δεν μπορεί να βοηθήσει ιδιαίτερα γιατί χάνει την πληροφορία της σημασιολογικής συσχέτισης των λέξεων, και λαμβάνει υπόψη μόνο τις αλλαγές στα γράμματα που συμβολίζουν την εκάστοτε λέξη. Συνήθως χρησιμοποιείται σε προηγούμενα βήματα για τον εντοπισμό κακογραμμένων λέξεων, όπου υπάρχουν ορθογραφικά λάθη ή αναγραμματισμοί, και προτείνει στον χρήστη την αντικατάστασή τους. Ένα γνωστό σε όλους παράδειγμα που χρησιμοποιεί αυτόν τον αλγόριθμο είναι η ιστοσελίδα της Google, η οποία προτείνει κάποια άλλη λέξη ή φράση όταν γράφουμε κάτι στην μπάρα αναζήτησης. Αντίστοιχα λειτουργεί και το Auto-Correct στα κινητά τηλέφωνα, όταν συντάσσουμε κείμενο.

3.10. Pointwise Mutual Information

Η Pointwise Mutual Information (Turney, 2001) χρησιμοποιεί τα δεδομένα που συλλέχθηκαν με ανάκτηση πληροφορίας και είναι μη επιβλεπόμενο μέτρο για την αξιολόγηση της σημασιολογικής ομοιότητας των λέξεων. Βασίζεται στην επανεμφάνιση των λέξεων χρησιμοποιώντας μετρητές που ενημερώνονται μέσα από ένα μεγάλο λεκτικό δείγμα (π.χ. Web). Σε δύο λέξεις w_1 και w_2 το PMI-IR τους μετριέται ως:

$$\text{PMI-IR}(w_1, w_2) = log_2 \frac{p(w_1 \,\&\, w_2)}{p(w_1) * p(w_2)}$$

, το οποίο δείχνει τον βαθμό της στατιστικής εξάρτησης μεταξύ του w_1 και w_2, και μπορεί να χρησιμοποιηθεί σαν μέτρο της σημασιολογικής ομοιότητας των δύο αυτών λέξεων.

Με τη χρήση του NEAR query (επανεμφάνιση σε παράθυρο 10 λέξεων) αναπαριστάται μια ισορροπία μεταξύ ακρίβειας και αποδοτικότητας. Συγκεκριμένα το παρακάτω ερώτημα χρησιμοποιείται για να μετρήσει στο AltaVista:

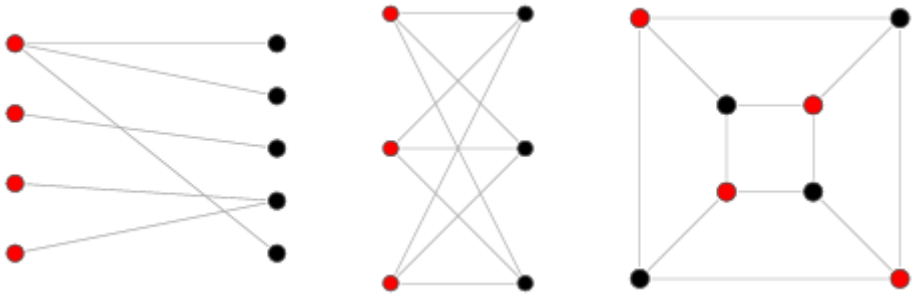
$$p_{NEAR}(w_1, w_2) \simeq \frac{hits(w_1 NEAR\ w_2)}{WebSize}$$

Προσεγγίζοντας το $p(w_1 \,\&\, w_2)$ από την πρώτη εξίσωση με $PNEAR(w_1 \,\&\, w_2)$ και υπολογίζοντας το $p(w_i)$ να είναι ίσο με το $hits(w_i)$/WebSize βρίσκουμε το παρακάτω PMI-IR:

$$\text{PMI-IR}(w_1, w_2) = log_2 \frac{hits(w_1 AND\ w_2) * WebSize}{hits(w_1) * hits(w_2)}$$

3.11. Bipartite Graph

Ο γράφος bipartite, γνωστός και ως bigraph, είναι ένα σετ κορυφών γράφου, που αποσυντίθενται σε δύο χωριστά μέρη U και V, ώστε δύο κορυφές των δύο γράφων εντός του ίδιου σετ να μην είναι γειτονικές. Ουσιαστικά κάθε ένωση εντός του γράφου, θα πρέπει να ενώνει μια κορυφή από το U μέρος, με μια κορυφή του V μέρους. Το U και το V είναι ανεξάρτητα σετ. Μπορούμε να σκεφτούμε το U και το V ως διαφορετικούς χρωματισμούς των κόμβων/κορυφών του γενικού γράφου. Επομένως, κάθε ένωση στην άλλη άκρη της θα έχει διαφορετικό χρώμα κορυφής από αυτό που είχε στο σημείο που ξεκινάει, όπως φαίνεται στα παρακάτω σχήματα:



Στη σύγκριση της σημασιολογικής ομοιότητας κειμένων, ο γράφος bipartite, θα δούμε ότι χρησιμοποιείται αρκετά συχνά, αφού ως το τμήμα U του γράφου μπορούμε να θεωρήσουμε το ένα κείμενο και αντίστοιχα το άλλο κείμενο ως το τμήμα V. Επομένως μπορούμε να δούμε το βαθμό των ενώσεων του γράφου και να καθορίσουμε ένα ποσό συσχέτισης μεταξύ των κειμένων.

3.12. Μήτρα Ομοιότητας (Similarity Matrix)

Πολλές φορές όταν χρειάζεται να υπολογιστεί το σκορ ομοιότητας μεταξύ δύο προτάσεων, τα σκορ ομοιότητας για κάθε ζευγάρι λέξεων αυτών των προτάσεων, θα πρέπει να συλλεχθούν σε μια δομή δεδομένων, ώστε να μπορούμε να υπολογίσουμε το συνολικό σκορ των ζευγαριών. Η λύση είναι να δημιουργηθεί μια μήτρα ομοιότητας N x M, όπου N είναι ο αριθμός των λέξεων της πρώτης πρότασης και M είναι ο αριθμός των λέξεων της δεύτερης πρότασης. Αυτή η μήτρα που δημιουργείται είναι γνωστή ως Μήτρα Ομοιότητας (Similarity Matrix). Παρακάτω απεικονίζεται ένα παράδειγμα Similarity Matrix για τις προτάσεις "Chocolate cake is unhealthy" και "Apple are important for making a pie"

	apple	**important**	**making**	**pie**
chocolate	0,31	0,11	0,00	0,00
cake	0,62	0,11	0,00	0,77
unhealthy	0,11	0,00	0,11	0,11

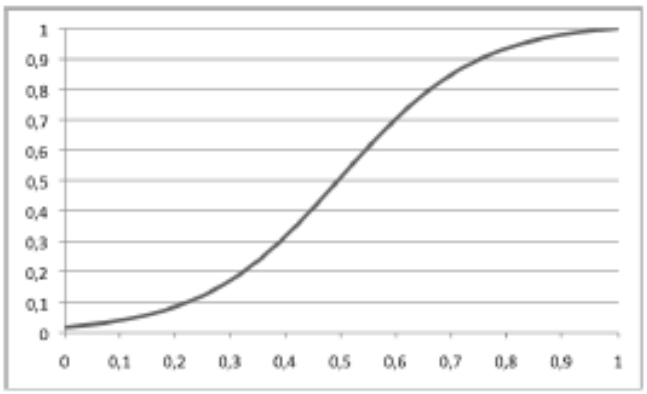
Για να βελτιωθεί το αποτέλεσμα της Μήτρας Ομοιότητας, οι STOP-Words μπορούν να αφαιρεθούν, ώστε να μειωθεί ο θόρυβος στα αποτελέσματα της μήτρας. Οι τιμές ομοιότητας των ζευγαριών των λέξεων θα χρησιμοποιηθούν για να υπολογιστεί η συνολική ομοιότητα των προτάσεων. Ένα τρόπος για να επιτευχθεί αυτό είναι με τη μεταφορά της μήτρας σε bipartite γράφο, όπου οι ενώσεις αντιστοιχούν σε κάθε ζευγάρι λέξεων της μήτρας και οι τιμές ομοιότητας των ζευγαριών ορίζουν τα βάρη που τοποθετούνται στις ενώσεις.

3.13. Σιγμοειδής Συνάρτηση (Sigmoid Function)

Η Sigmoid συνάρτηση είναι μια μαθηματική συνάρτηση η οποία έχει το σχήμα "S", από όπου προκύπτει το όνομά της, καθώς φέρει και την ονομασία σιγμοειδής καμπύλη. Χρησιμοποιείται για να μοντελοποιήσει συστήματα που κυμαίνονται σε μεγάλες τιμές συνεχόμενων ροών εξόδου. Η συνάρτηση κλιμακώνεται ώστε να καλύπτει τιμές από 0 έως 1.

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

Όπως φαίνεται και στο παρακάτω διάγραμμα, με τη συνάρτηση Sigmoid, επιτυγχάνεται μια μικρή τιμή να γίνει ακόμα μικρότερη, ενώ αντίθετα μια μεγάλη τιμή να γίνει ακόμα μεγαλύτερη.



Επομένως, είναι ιδανική για να ενισχύει τη διαφορά μεταξύ των αποτελεσμάτων.

3.14. Perceptron Algorithm

Αρκετές διαδικασίες της NLP, συμπεριλαμβανομένης της εύρεσης του Text Semantic Similarity, ορίζουν την εκπαίδευση του συστήματος μέσα από ένα λεκτικό δείγμα, ώστε να βγουν τα πρότυπα με τα οποία το σύστημα θα δουλέψει. Είναι λοιπόν επακόλουθο να χρησιμοποιηθούν νευρωνικά δίκτυα, στο σύστημα εκπαίδευσης.

Ο Perceptron (Rosenblatt, 1958) είναι ένα υποθετικό νευρωνικό σύστημα, ή μηχανή, που σχεδιάστηκε για να απεικονίσει μερικές από τις βασικές ιδιότητες των ευφυών συστημάτων που συναντιούνται στους ζωντανούς οργανισμούς, χωρίς να εμβαθύνει στις ειδικές και συχνά άγνωστες συνθήκες που επικρατούν για τους βιολογικούς οργανισμούς. Η λειτουργική ομοιότητα μεταξύ των νευρώνων και των απλών on-off μονάδων, βάσει των οποίων κατασκευάστηκαν οι υπολογιστές, είχε εντυπωσιάσει τους θεωρητικούς και έστρεψε τα βλέμματα προς μεθόδους ανάλυσης και απεικόνισης πολύπλοκων λογικών λειτουργιών σε τέτοια μορφή.

Πολλοί θεωρητικοί, όπως ο Ashby και ο Von Neumann προσέγγισαν προβλήματα για το πως ένα μη τέλειο νευρωνικό δίκτυο, που εμπεριέχει τυχαίες συνδέσεις, μπορεί να δημιουργηθεί για να εκτελεί αξιόπιστα τέτοιες συναρτήσεις που μπορεί να αναπαριστούνται

από εξιδανικευμένα συνδεδεμένα διαγράμματα. Δυστυχώς, η γλώσσα της συμβολική λογικής και η άλγεβρα Boole δεν είναι κατάλληλες για τέτοιες αναζητήσεις. Χρειάζεται μια πιο κατάλληλη γλώσσα για τη μαθηματική ανάλυση των γεγονότων. Για αυτό το λόγο δημιουργήθηκε το μοντέλο του Perceptron, με τη χρήση της θεωρίας των πιθανοτήτων αντί της συμβολικής λογικής. Είναι ένας δυαδικός ταξινομητής που αντιστοιχεί την είσοδο x σε μια τιμή εξόδου f(x)

$$f(x) = \begin{cases} 1 & if\ w * x + b > 0 \\ 0 & otherwise \end{cases}$$

,όπου w είναι ένα διάνυσμα βάρους πραγματικών τιμών και w*x είναι το εσωτερικό γινόμενο μεταξύ των διανυσμάτων w και x. Το b (bias) είναι ένας σταθερός όρος που δεν εξαρτάται από τιμή εισόδου. Η τιμή f(x) χρησιμοποιείται για να ταξινομήσει το x είτε ως θετικό είτε ως αρνητικό στιγμιότυπο. Το bias χρησιμοποιείται για την μετατόπιση της συνάρτησης ενεργοποίησης ή για να δώσει στον νευρώνα εξόδου ένα βασικό επίπεδο δραστηριότητας.

4. Μετρικές Σημασιολογικής Ομοιότητας

Οι περισσότερες αν όχι όλες οι προσεγγίσεις της σύγκρισης σημασιολογικής ομοιότητας μεταξύ δύο κειμένων επιστρατεύουν μετρικές οι οποίες παράγουν μια τιμή που ορίζει την ομοιότητα των κειμένων που εισήχθησαν στο σύστημα. Αυτές οι μετρικές χωρίζονται σε δύο μεγάλες κατηγορίες. Είναι οι μετρικές που προϋποθέτουν ότι το σύστημα έχει κάποια προϋπάρχουσα γνώση, με την οποία συγκρίνει τα δεδομένα και παράγει νέα πληροφορία και ορίζονται ως Knowledge-Based μετρικές, καθώς είναι και οι μετρικές, που δεν απαιτούν προϋπάρχουσα γνώση του συστήματος σχετική με τα δεδομένα. Αυτές στηρίζονται στην επί τόπου εκπαίδευση του συστήματος βάσει ενός λεκτικού δείγματος, και την εξόρυξη προτύπων μέσα από αυτό. Η δεύτερη κατηγορία μετρικών, ονομάζεται Corpus-Based.

Αν και οι Corpus-based μετρικές, έχουν ένα πλεονέκτημα στο ότι δεν απαιτούν επίπονη και κοπιαστική δουλειά από τη μεριά του διαχειριστή του συστήματος, ώστε να εισάγει δεδομένα σε κάποια Βάση Γνώσης, καθώς και μπορούν να εφαρμοστούν και σε άγνωστες λέξεις ή ακόμα και σε διαφορετική γλώσσα από αυτή που προοριζόταν αρχικά το σύστημα, τείνουν να είναι λιγότερο αξιόπιστες και ως εκ τούτου και λιγότερο προτιμητέες από τις Knowledge-Based μετρικές στο πρόβλημα της σημασιολογικής ομοιότητας κειμένων. Επίσης δημιουργούν επιπλέον υπολογιστικό κόστος στο σύστημα. Ωστόσο, δεν είναι λίγες οι προσεγγίσεις που τις χρησιμοποιούν, είτε αυτόνομες, είτε ως επιπλέον έλεγχο σε συνδυασμό με τις Knowledge-Based, ώστε το σύστημα να παράγει ακόμη πιο αξιόπιστα αποτελέσματα.

4.1. Μετρικές Knowledge-Based

Παρακάτω θα εξετάσουμε αρχικά τις Knowledge-Based τεχνικές που χρησιμοποιούνται από τις περισσότερες προσεγγίσεις επίλυσης του προβλήματος, καθότι πιο διαδεδομένες. Οι Knowledge-Based προσεγγίσεις παράγουν ένα μέτρο συσχέτισης, χρησιμοποιώντας λεξικούς πόρους και οντολογίες όπως το WordNet για να μετρήσουν την επικάλυψη των ορισμών ή την απόσταση των λέξεων μέσα σε μια ταξονομία.

4.1.1. Lesk

Ο αλγόριθμος Lesk αρχικά προτάθηκε σαν λύση στο πρόβλημα του Word Sense Disambiguation (Lesk, 1986). Η διαδικασία που προτείνεται, χρησιμοποιεί τα διαθέσιμα λεξικά, ώστε να βρει κοινά χαρακτηριστικά μεταξύ της πληροφορίας που παρέχεται από το λεξικό και των κοντινών λέξεων από τη λέξη που εξετάζεται στο κείμενο. Ελέγχοντας λοιπόν τις λέξεις στα δεξιά και αριστερά από τη λέξη που εξετάζουμε, βλέπουμε στο λεξικό που διατίθεται (π.χ. WordNet) αν υπάρχει στην περιγραφή των πιθανών Word-Senses που αντιστοιχούν στην εξετάζουσα λέξη. Έτσι δημιουργείται ένας μετρητής για κάθε Word-Sense (synset) όπου αυξάνεται για κάθε ταυτοποίηση που βρίσκεται με την παραπάνω διαδικασία. Το word sense με τον μεγαλύτερο μετρητή υπερισχύει και κατά αυτό τον τρόπο γίνεται η αποσαφήνιση της χρήσης της λέξης.

Τα θετικά αυτής της μεθόδου, είναι ότι δεν βασίζεται στο συντακτικό και μπορεί να χρησιμοποιηθεί σαν καλό συμπλήρωμα στη συντακτική ανάλυση. Για παράδειγμα, στην πρόταση "I know a hawk from a handsaw" το πρόγραμμα δεν μπορεί να βρει ότι το "hawk" (strong, swift, keen-sighted bird of prey") υπερισχύει από το sense του "hawk" (offer goods for sale), αλλά η σύνταξη της πρότασης θα βοηθούσε άμεσα, εφόσον το ρήμα δεν μπορεί να εμφανίζεται αμέσως μετά το άρθρο.

Η εφαρμογή του παραπάνω αλγορίθμου ως μετρική σημασιολογικής ομοιότητας δύο λέξεων, με τη χρήση του WordNet εξετάζει τα δύο gloss (περιγραφές των word senses) που περιγράφουν την συγκεκριμένη λέξη για να βρει κοινά στοιχεία μεταξύ τους. Το βασικό ζητούμενο είναι να μετρήσει τον αριθμό των λέξεων που μοιράζονται μεταξύ των δύο glosses. Όσο πιο μεγάλος είναι αυτός ο αριθμός, τόσο πιο κοντά είναι οι δύο έννοιες που εξετάζονται. Βασικά, ο αλγόριθμος κάνει μια σύγκριση των ορισμών των λέξεων για κάθε word-sense κάθε λέξης στην πρόταση και υπολογίζει το σκορ. Για παράδειγμα, στην πρόταση "time flies like an arrow" ο αλγόριθμος συγκρίνει κάθε ερμηνεία(sense) της λέξης "time" με τις ερμηνείες των λέξεων "fly" και "arrow". Μετά, κάνει το ίδιο για τις ερμηνείες του "fly", κ.ο.κ. Για τα ζευγάρια ερμηνειών που έχουν συγκριθεί, δεν χρειάζεται να ξαναγίνεται ο υπολογισμός αυτός. Επίσης, παρατηρούμε, ότι πριν γίνει όποια σύγκριση, έχουν αφαιρεθεί τα STOP-Words, ώστε να διευκολυνθεί η απόδοση του αλγορίθμου, ενώ παράλληλα να μειωθεί και ο θόρυβος, καθώς τα STOP-Words θα αύξαναν το σκορ της σημασιολογικής συσχέτισης.

Παρακάτω παρατίθεται ο αλγόριθμος με τον οποίο λειτουργεί η μετρική του Lesk:

```
function SIMPLIFIED LESK(word,sentence) returns best sense of word
best-sense <- most frequent sense for word
max-overlap <- 0
context <- set of words in sentence
for each sense in senses of word do
signature <- set of words in the gloss and examples of sense
overlap <- COMPUTEOVERLAP (signature,context)
if overlap > max-overlap then
max-overlap <- overlap
best-sense <- sense

end return (best-sense)
```

Οι Banerjee και Pedersen χρησιμοποιώντας την βασική αρχή του αλγόριθμου του Lesk έφτιαξαν μια προέκταση, με σκοπό να παράγουν ακόμη πιο αξιόπιστα αποτελέσματα. Η προέκταση αυτή προτείνει τη χρήση των ορισμών των hypernyms και των hyponyms μιας λέξης. Δηλαδή, όπως ο απλός αλγόριθμος του Lesk, εξετάζει το gloss, μόνο για τη ζητούμενη λέξη, η προέκταση, εξετάζει τα gloss και για τα hypernyms και hyponyms της ζητούμενης λέξης (Banerjee & Pedersen, 2002). Φυσικά, αυτή η προσέγγιση έχει τόσο θετικά όσο και αρνητικά. Αν και μπορεί να οδηγήσει σε ακριβέστερα αποτελέσματα, αφού έχουμε περισσότερο υλικό για να συγκρίνουμε, ο χρόνος υπολογισμού αυξάνεται σημαντικά γιατί ο αλγόριθμος έχει να ψάξει μέσα σε μεγαλύτερο αριθμό λέξεων. Λόγω της χειρότερης απόδοσης τις περισσότερες φορές δεν προτιμάται από τον απλό αλγόριθμο του Lesk.

4.1.2. Wu και Palmer

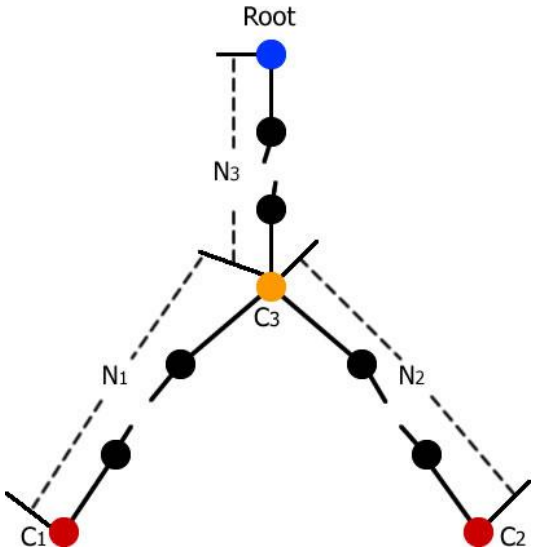
Μία άλλη μετρική της σημασιολογικής ομοιότητας, είναι αυτή που εισάγουν ο Wu και ο Palmer (Wu & Palmer, 1994), η οποία έχει τις βάσεις της στην μηχανική μετάφραση, όπου πρέπει να βρεθεί η λέξη-στόχος που είναι εννοιολογικά κοντινότερη στη λέξη-πηγή. Οι διαθέσιμες πληροφορίες που δίνονται για τη λήψη αυτής της απόφασης, είναι το κείμενο πηγή, τα λεξικά και οι Βάσεις Γνώσης του συστήματος.

Για μια σωστή μετάφραση θα πρέπει να ληφθεί υπόψη το περιεχόμενο γύρω από την λέξη που θέλουμε να μεταφράσουμε και η γνώση του νοήματος της πρότασης στην οποία βρίσκεται. Οπότε κρίνεται σκόπιμο να οριστούν εννοιολογικοί τομείς. Εντός ενός εννοιολογικού τομέα, δύο έννοιες ορίζονται από το πόσο κοντά βρίσκονται στη ιεραρχία.

Η σημασιολογική ομοιότητα μεταξύ των C_1 και C_2 είναι:

$$\text{ConSim}(C_1,C_2) = \frac{2*N_3}{N_1+N_2+2*N_3}$$

Το C_3 είναι η κοντινότερη κοινή υπερ-έννοια του C_1 και του C_2. Το N_1 είναι ο αριθμός των κόμβων της διαδρομής από το C_1 στο C_3,το N_1 είναι ο αριθμός των κόμβων από το C_2 στο C_3 και το N_3 είναι ο αριθμός των κόμβων από το C_3 στη ρίζα (root), όπως φαίνεται και στο παρακάτω σχήμα:



Αφού καθοριστεί το μέτρο ομοιότητας σε έναν τομέα, η ομοιότητα μεταξύ των σημασιών των ρημάτων μπορεί να οριστεί ως το άθροισμα των σταθμισμένων ομοιοτήτων μεταξύ ζευγαριών απλούστερων εννοιών για κάθε έναν από τους τομείς που εμπεριέχουν τα ρήματα.

$$\mathbf{WordSim}(V_1, V_2) = \Sigma_i \, wi * \mathbf{ConSim}(Ci_{,1}, Ci_{,2})$$

Η υλοποίηση τω Wu & Palmer ακούει στο όνομα UNICON όπου η είσοδος στο σύστημα είναι η το ρήμα-πηγή για να δημιουργηθεί η ιεραρχική δομή. Μετά την εφαρμογή τεχνικών Word Sense Disambiguation η εσωτερική αναπαράσταση της πρότασης μπορεί να σχηματιστεί. Το σύστημα κατόπιν, προσπαθεί να βρει την κατανόηση του ρήματος στόχου για εσωτερική αναπαράσταση. Αν οι έννοιες δεν έχουν κάποιο ρήμα, το σύστημα παίρνει τις κοντινότερες έννοιες ως υποψήφιους για να δει αν υπάρχει κατανόηση του ρήματος στόχου. Αν βρεθεί γίνεται ένα ταίριασμα μεταξύ της εσωτερικής αναπαράστασης και της σημασίας του ρήματος στόχου, μαζί με τους περιορισμούς που συσχετίζονται με το ρήμα στόχο. Κατόπιν το σύστημα πρέπει να μετρήσει τη σημασιολογική ομοιότητα της εσωτερικής αναπαράστασης και του ρήματος στόχου, καθώς και το βαθμό ικανοποίησης των περιορισμών που συσχετίζονται με το ρήμα στόχο.

4.1.3. Jiang και Conrath

Η μετρική των Jiang και Conrath (Jiang & Conrath, 1997), συνδυάζει μια λεκτική ταξονομία με στατιστική πληροφορία προερχόμενη από το λεκτικό δείγμα, ώστε η σημασιολογική απόσταση μεταξύ κόμβων στο σημασιολογικό χώρο που κατασκευάζεται από την ταξονομία να μπορεί καλύτερα να ποσοτικοποιηθεί με στοιχεία που προέρχονται από την ανάλυση της κατανομής των λεκτικών δεδομένων.

Σε μια corpus-based προσέγγιση, οι σχέσεις των λέξεων προέρχονται συνήθως από τον τρόπο που εμφανίζονται στο λεκτικό δείγμα.

Η χρήση μιας ψευδο-βάσης γνώσης μπορεί να συμπληρώσει αυτή τη στατιστική προσέγγιση, στην οποία η πραγματική κατανόηση του κειμένου είναι αδύνατη. Έτσι το στατιστικό μοντέλο μπορεί να έχει το πλεονέκτημα ενός εννοιολογικού χώρου που δομείται από μια ταξονομία που δημιουργήθηκε με το χέρι, ενώ παρέχει στοιχεία από την στατιστική ανάλυση του λεκτικού δείγματος.

Η εν λόγω μετρική, χρησιμοποιεί την τεχνική του περιεχομένου της πληροφορίας (Information-Content) που προτείνει ο Resnik (βλ. παρακάτω) για να ορίσει την σημασιολογική ομοιότητα. Για έναν πολυδιάστατο χώρο στον οποίο οι κόμβοι αναπαριστούν μια μοναδική έννοια που αποτελείται από συγκεκριμένη ποσότητα πληροφορίας και οι ενώσεις αναπαριστούν την σχέση μεταξύ δύο εννοιών, η ομοιότητα μεταξύ δύο εννοιών είναι η έκταση στην οποία μοιράζονται κοινές πληροφορίες. Σε μια ιεραρχική δομή, η κοινή πληροφορία είναι συγκεκριμένες έννοιες/κόμβοι που συμπεριλαμβάνουν και τους δύο κόμβους που αναπαριστούν τις λέξεις που συγκρίνονται. Η τιμή ομοιότητας ορίζεται ως η τιμή του περιεχομένου της πληροφορίας αυτού του υπερ-κόμβου. Η τιμή αυτή υπολογίζεται από την πιθανότητα εμφάνισης της έννοιας αυτού του κόμβου σε ένα μεγάλο λεκτικό δείγμα.

$$IC(c) = \log^{-1} P(c)$$

, όπου P(c) είναι η πιθανότητα να βρεθεί στιγμιότυπο της έννοιας c. Στην περίπτωση της ιεραρχικής δομής, όπου οι έννοιες στην ιεραρχία συμπεριλαμβάνουν και άλλες που βρίσκονται χαμηλότερα στην ιεραρχία αυτή, υπονοείται ότι το P(c) είναι μονοτονικό όσο κινείται κάποιος στην ιεραρχία. Όσο η πιθανότητα του κόμβου αυξάνεται, το περιεχόμενο της

πληροφορίας του μειώνεται. Αν είναι ένας μοναδικός κόμβος στην ιεραρχία, τότε η πιθανότητά του είναι 1 και το περιεχόμενο της πληροφορίας του 0.

Δεδομένου του μονοτονικού χαρακτηριστικού της τιμής του περιεχομένου της πληροφορίας, η ομοιότητα δύο εννοιών ορίζεται ως:

$$sim(c_1, c_2) = \max_{c \in Sup(c_1,c_2)} [IC(c)] = \max_{c \in Sup(c_1,c_2)} [-\log p(c)]$$

, όπου το $Sup(c_1, c_2)$ είναι το σετ των εννοιών που συμπεριλαμβάνουν το c_1 και το c_2. Πρακτικά βρίσκουμε το κοντινότερο ανώτατο όριο (Lowest Upper Bound) μεταξύ αυτών που συμπεριλαμβάνουν το c_1 και το c_2.

Στην περίπτωση πολλαπλής κληρονομικότητας, όπου οι λέξεις μπορεί να έχουν παραπάνω από ένα sense και επομένως πολλαπλές υπερ-κλάσεις, η λεκτική ομοιότητα μπορεί να οριστεί ως η καλύτερη τιμή ομοιότητα μεταξύ αυτών των υπερ-κλάσεων.

$$sim(w_1, w_2) = \max_{c_1 \in sen(w_1), c_2 \in sen(w_2)} [sim(c_1, c_2)]$$

, όπου sen(w) ορίζει το σετ των πιθανών senses για τη λέξη w.

Edge-based (Distance) Προσέγγιση

Ο πιο φυσικός και απευθείας τρόπος για να αποτιμηθεί η σημασιολογική ομοιότητα σε μια ταξονομία είναι μια προσέγγιση η οποία θα βασίζεται στην απόσταση της διαδρομής εντός της ταξονομίας. Υπολογίζει την απόσταση μεταξύ των κόμβων που αναπαριστούν την έννοια των λέξεων που συγκρίνονται. Η απόσταση αυτή μπορεί να μετρηθεί από τη γεωμετρική απόσταση των κόμβων αυτών. Όσο πιο μικρή είναι η διαδρομή από τον έναν κόμβο στον άλλον, τόσο περισσότερο όμοιοι είναι μεταξύ τους.

Η σύνδεση μεταξύ δύο κόμβων θα πρέπει να σταθμιστεί. Για να οριστούν αυτά τα βάρη αυτόματα, πρέπει να ληφθούν υπόψην διάφορες συνιστώσες όπως πυκνότητα του δικτύου, βάθος του κόμβου στην ιεραρχία, τύπος σύνδεσης και η ισχύς της ένωσης των κόμβων.

Ένας τρόπος υπολογισμού των παραπάνω δίνεται από τον Sussna (Sussna, 1993), όπου το βάρος δύο κόμβων c_1 και c_2 υπολογίζεται ως:

$$wt(c_1, c_2) = \frac{wt(c_1 \rightarrow_r c_2) + wt(c_2 \rightarrow_{r'} c_1)}{2d}$$

δεδομένου ότι:

$$wt(x \rightarrow_r y) = max_r - \frac{max_r - min_r}{n_r(x)}$$

όπου $\rightarrow_r$ είναι μια σχέση τύπου r, $\rightarrow_{r'}$ είναι το αντίστροφο, d είναι το βάθος του βαθύτερου από τα 2, max και min το μέγιστο και ελάχιστο πιθανό βάρος για μια συγκεκριμένη σχέση τύπου r και $n_r(x)$ είναι ο αριθμός των σχέσεων τύπου r που ξεκινάνε από τον κόμβο x.

Για να μετατρέψουμε την απόσταση σε μέτρο ομοιότητας, αρκεί να αφαιρέσουμε το μήκος της διαδρομής από το μέγιστο δυνατό μήκος διαδρομής

$$sim(w_1, w_2) = 2d_{max} - [\min_{c_1 \in sen(w_1),\ c_2 \in sen(w_2)} len(c_1, c_2)]$$

Μια συνδυαστική προσέγγιση

Στην edge-based προσέγγιση μπορούμε να προσθέσουμε το περιεχόμενο της πληροφορίας σαν παράγοντα απόφασης. Ιδιαίτερη προσοχή θα πρέπει να δοθεί στον καθορισμό της ισχύς του συνδέσμου μεταξύ ενός κόμβου-γονέα και κόμβου-παιδιού.

Η δύναμη ενός συνδέσμου-παιδιού είναι αναλογική στην πιθανότητα να βρεθεί ένα στιγμιότυπο της έννοιας-παιδί c_i για ένα στιγμιότυπο της έννοιας-γονέας p: $P(c_i|p)$

$$P(c_i|p) = \frac{P(c_i \cap p)}{P(p)} = \frac{P(c_i)}{P(p)}$$

Ακολουθώντας την κλασική παραδοχή της θεωρίας της πληροφορίας, ορίζουμε τη δύναμη του συνδέσμου (LS) παίρνοντας τον αρνητικό λογάριθμο της παραπάνω πιθανότητας:

$LS(c_i, p) = - \log(P(c_i | p) = IC(c_i)-IC(p)$

Αυτό ορίζει ότι το LS είναι απλά η διαφορά των τιμών πληροφοριακού περιεχομένου μεταξύ έννοιας παιδί - γονέα. Έχοντας υπόψην και άλλους παράγοντες όπως την πυκνότητα, το βάθος και τον τύπο, το συνολικό βάρος για έναν σύνδεσμο ορίζεται ως:

$$wt(c, p) = \left(\beta + (1-\beta)\frac{\bar{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha} [IC(c) - IC(p)] T (c,p)$$

,όπου d(p) είναι το βάθος ενός κόμβου p στην ιεραρχία, E(p) ο αριθμός των ενώσεων στους συνδέσμους-παιδιά, $\bar{E}$ είναι η μέση πυκνότητα σε όλη την ιεραρχία και T(c,p) είναι ο σύνδεσμος σχέση/τύπος. Οι παράμετροι α(α>=0) και β(0<=β<=1) ελέγχουν τον βαθμό του πόσο πολύ το βάθος και η πυκνότητα συμμετέχουν στον υπολογισμό του βάρους. Γίνονται πιο ασήμαντοι όσο το α πλησιάζει στο 0 και το β στο 1.

Η συνολική απόσταση μεταξύ δύο κόμβων θα είναι το άθροισμα των βαρών στο shortest path σύνδεσμο δύο κόμβων:

$$Dist(w_1, w_2) = \sum_{c \in \{\text{path}(c_1,c_2) - \text{LSuper}(c_1,c_2)\}} wt(c, parent(c))$$

,όπου c_1=sen(w_1), c_2=sen(w_2) και path(c_1, c_2) είναι το σετ που περιέχει όλους τους κόμβους στο shortest path από c_1 ως c_2. LSuper(c_1, c_2) ορίζει το ελάχιστο κοινό υπερκόμβο (lowest super-ordinate) του c_1 και c_2.Η συνάρτηση της απόστασης μπορεί να οριστεί και ως:

$$Dist(w_1, w_2) = IC(c_1)+IC(c_2)-2 \times IC(LSuper(c_1, c_2))$$

4.1.4. Leacock και Chodorow

Η αποσαφήνιση της χρήσης μια λέξης είναι η αντιστοίχιση μεταξύ λέξεων ενός κειμένου και τον κατάλληλων χρήσεών τους από ένα λεξικό. Οι corpus-based τεχνικές, εκπαιδεύουν έναν στατιστικό ταξινομητή ώστε να μπορεί να βρει χαρακτηριστικά αποσαφήνισης των λέξεων

βάσει άλλων λέξεων που συναντούνται στο δείγμα εκπαίδευσης και κοινών συσχετίσεων με το πρότυπο. Ένα βασικό πρόβλημα των corpus-based μεθόδων είναι ότι τα δεδομένα που δίνονται ως δείγμα εκπαίδευσης είναι διάσπαρτα. Αν θέλουμε να επιβεβαιώσουμε ότι το δείγμα εκπαίδευσης είναι αξιόπιστο θα πρέπει να ορίσουμε για κάθε πολυσήμαντη λέξη, τη χρήση της (word sense) με το χέρι. Αυτή είναι μια πολύ αργή και κοπιαστική διαδικασία.

Τα περιεχόμενα του κειμένου στο οποίο βρίσκεται η λέξη έχουν δείξει ότι είναι ένας πολύ καλός τρόπος για την αποσαφήνιση της χρήσης της στο κείμενο. Γι αυτό το λόγο πολλές στατιστικές προσεγγίσεις αυξάνουν το παράθυρο του περιεχομένου. Δηλαδή, τις γειτονικές λέξεις προς τη λέξη που χρήζει αποσαφήνισης. Έχει βρεθεί ότι ο Μπαϋεσιανός ταξινομητής (Bayesian classifier) δουλεύει καλύτερα με ένα παράθυρο που έχει πενήντα λέξεις αριστερά και άλλες τόσες δεξιά από την λέξη στόχο (Gale, Church, & Yarowksy, 1992). Ωστόσο, στην εν λόγω μετρική για να βελτιωθεί η απόδοση του συστήματος το παράθυρο ορίζεται σε μικρότερη κλίμακα.

Όπως ορίζουν οι Leacock και Chodorow (Leacock & Chodorow, 1998) για κάθε θέση p μέσα στο παράθυρο, δημιουργείται μια λίστα με όλα τα πιθανά POS tags που εμφανίζονται στη θέση p, βάσει όλων των πιθανών χρήσεων της συγκεκριμένης λέξης. Κατόπιν, για κάθε χρήση της λέξης (word-sense) ένα σκορ υπολογίζεται παίρνοντας το παράγωγο (product) του 1 και προσθέτοντας την πιθανότητα του συγκεκριμένου word-sense, δεδομένου ότι βρέθηκε το POS tag για κάθε θέση p στο παράθυρο.

$$pos\_score_i = \prod_{p=-2}^{2} \left(P(sense \mid pos_{tag_p}) \right) + 1$$

Ο παραπάνω τύπος μας δίνει τη δυνατότητα να συγκρίνουμε τα στοιχεία για ένα sense ακόμα και αν οι πιθανότητες είναι 0. Κατόπιν, επιλέγεται το sense με το μεγαλύτερο σκορ και συγκρίνεται με τις τιμές που είχαν δοθεί από τους ανθρώπινους κριτές.

Επόμενο βήμα είναι να ενισχύσουμε τον τοπικό ταξινομητή περιεχομένου με σημασιολογικές πληροφορίες. Σε αυτό το βήμα προσπαθούμε να καλύψουμε τα κενά σε ένα αραιό δείγμα εκπαίδευσης, βρίσκοντας μέτρα ομοιότητας βασισμένοι στο WordNet. Βάσει του τρόπου που είναι σχεδιασμένη η ταξονομία των ουσιαστικών στο WordNet αν μια λέξη έχει πολλαπλά senses θα εμφανίζεται και σε πολλαπλά synsets σε διάφορες τοποθεσίες της ταξονομίας.

Μια προφανής μέτρηση της σημασιολογικής ομοιότητας μεταξύ δύο λέξεων είναι να μετρηθεί η απόστασή τους εντός της ταξονομίας του WordNet. Αυτό μπορεί να γίνει βρίσκοντας όλες τις διαδρομές από τη μία λέξη προς την άλλη και μετά επιλέγοντας τη μικρότερη διαδρομή. Για να υπολογιστεί το μήκος της διαδρομής μεταξύ των λέξεων χρησιμοποιείται ο παρακάτω τύπος:

$$sim_{ab} = \max[-\log(Np / 2D)]$$

,όπου Np είναι ο αριθμός των κόμβων σε μια διαδρομή p από το a στο b και D είναι το μέγιστο βάθος της ταξονομίας. Σημειώνεται ότι το μήκος τις διαδρομής μετριέται σε κόμβους και όχι σε ενώσεις των κόμβων.

Όπως τονίζει ο Resnik τα μέτρα που βασίζονται στο μήκος της διαδρομής θα έχουν πρόβλημα από μεγάλες διαφορές στο βάθος των ταξινομιών του WordNet. Προτείνει να χρησιμοποιηθεί ένα μέτρο βασισμένο στην πληροφορία, την περισσότερο πληροφοριακή κλάση, όταν υπολογίζεται η ομοιότητα. Η κλάση αποτελείται από τα συνώνυμα που βρέθηκαν σε έναν κόμβο και τα συνώνυμα σε όλους τους κόμβος που βρίσκονται κάτω από αυτόν (hyponyms). Η συχνότητα της κλάσης είναι η συχνότητα όλων των λέξεων στην

κλάση, όπως ορίζεται από την στατιστική ανάλυση του λεκτικού δείγματος. Από αυτές τις συχνότητες μπορούμε να υπολογίσουμε την πιθανότητα της κλάσης. Επομένως, η ομοιότητα μεταξύ των λέξεων a και b ορίζεται από τη λιγότερο πιθανή κλάση που ανήκουν, όπως φαίνεται εδώ:

$$sim_{ab} = \max[-\log(\Pr(c))]$$

,όπου Pr(c) είναι η πιθανότητα μια κλάσης c που περιλαμβάνει το sense του a και το sense του b. Αυτό είναι το μέτρο ομοιότητας βάσει της περισσότερο πληροφοριακής κλάσης (most informative class). Πριν συγκριθεί ο τοπικός ταξινομητής περιεχομένου με τα μέτρα ομοιότητας, χρειάζεται να καθοριστεί πόσο γενίκευση παρέχεται.

4.1.5. Lin

Ένα πρόβλημα που έχουν διάφορες προσεγγίσεις μετρικών ομοιότητας είναι ότι είναι συνδεδεμένες σε μια συγκεκριμένη εφαρμογή ή υποθέτουν την ύπαρξη ενός μοντέλου σε συγκεκριμένο τομέα. Ένα άλλο πρόβλημα είναι ότι βασικές υποθέσεις δεν αναφέρονται ρητά.

Ο Lin (Lin, 1998) παρουσιάζει έναν ορισμό για την ομοιότητα που εμπεριέχει:

- Καθολικότητα (Universality) - Είναι εφαρμόσιμο όσο ο τομέας έχει πιθανολογικό μοντέλο. Εφόσον η θεωρία πιθανοτήτων μπορεί να εφαρμοστεί σε διάφορα είδη αναπαράστασης γνώσης ο ορισμός της ομοιότητας μπορεί να εφαρμοστεί σε πολλούς διαφορετικούς τομείς.
- Θεωρητική Αιτιολόγηση (theoretical justification) - Το μέτρο ομοιότητας δεν ορίζεται απευθείας από έναν τύπο. Προκύπτει από ένα σύνολο υποθέσεων σχετικά με την ομοιότητα.

Δύο αντικείμενα A και B ορίζονται όμοια όταν :

- Έχουν κοινά περιεχόμενα. Όσο πιο πολλά έχουν, τόσο πιο όμοια είναι.
- Δεν έχουν διαφορές. Όσο πιο λίγες έχουν τόσο πιο όμοια είναι.
- Η μέγιστη ομοιότητα μεταξύ τους επιτυγχάνεται όταν είναι πανομοιότυπα.

Το θεώρημα της ομοιότητας ορίζεται ως εξής:

Η ομοιότητα μεταξύ A και B μετράται από τη σχέση μεταξύ του ποσού της πληροφορίας που χρειάζεται για να δηλωθούν τα κοινά του A και του B και της πληροφορίας που χρειάζεται για να περιγραφθούν πλήρως το A και το B:

$$\text{sim}(A,B) = \frac{\log P\ (common(A,B))}{\log P\ (description\ (A,B))}$$

Τα χαρακτηριστικά διανύσματα είναι μια από τις απλούστερες και πιο κοινές φόρμες αναπαράστασης της γνώσης. Συνήθως ορίζονται βάρη στα χαρακτηριστικά για το γεγονός ότι η ανομοιότητα που προήλθε από πιο σημαντικά χαρακτηριστικά είναι μεγαλύτερη από αυτή που προήλθε από λιγότερο σημαντικά χαρακτηριστικά.

Οι δοκιμές που πραγματοποίησε ο Lin έγιναν με 3 μέτρα ομοιότητας:

1.

$$sim_{edit}(x, y) = \frac{1}{1 + editDist(x, y)}$$

, όπου το editDist(x, y) είναι ο ελάχιστος αριθμός εισαγωγής και διαγραφής χαρακτήρων που χρειάζονται για να μετατραπεί η λέξη x στην λέξη y.

2. Το δεύτερο βασίζεται στον αριθμό των διαφορετικών τριγραμμάτων στις δύο λέξεις.

$$sim_{tri}(x, y) = \frac{1}{1 + |tri(x)| + |tri(y)| - 2 * |tri(x) \cap tri(y)|}$$

,όπου tri(x) είναι το σετ των τριγραμμάτων στη x (π.χ.: tri (eloquent) = {elo, loq, oqu, que, ent}).

3. Το τρίτο βασίζεται στον ορισμό που δώσαμε για την ομοιότητα υπό την υπόθεση ότι η πιθανότητα ενός τριγράμματος ορίζεται σε μια λέξη ανεξάρτητα από τα άλλα τριγράμματα της λέξης:

$$\text{sim(x,y)} = \frac{2 * \sum_{t \in \text{tri(x)} \cap \text{tri(y)}} \log P(t)}{\sum_{t \in \text{tri(x)}} \log P(t) + \sum_{t \in \text{tri(y)}} \log P(t)}$$

Στην αποτίμηση τα αποτελέσματα έδειξαν αρκετά καλύτερα με τη χρήση της τρίτης μεθόδου. Χρησιμοποιείται ένας parser για να εξάγει τις τριάδες συσχέτισης από το κείμενο που δίνεται σαν είσοδος. Η τριάδα αποτελείται από το κεφάλι (head), τύπος σχέσης (dependency type) και τροποποιητή (modifier) (π.χ.: "I have a brown dog" -> (have subj I), (have obj dog), (dog adj-mod brown), (dog det a))

Ορίζεται λοιπόν ένας πίνακας F(w), όπως ο πίνακας παρακάτω για τις λέξεις "duty" και "sanction":

Feature	duty	sanction	I(f_i)
f_1: $subj-of$ ($include$)	X	X	3.15
f_2: $obj-of$ ($assume$)	X		5.43
f_3: $obj-of$ ($avert$)	X	X	5.88
f_4: $obj-of$ ($ease$)		X	4.99
f_5: $obj-of$ ($impose$)	X	X	4.97
f_6: $adj-mod$ ($fiduciary$)	X		7.76

f_7: $adj-mod$ $(punitive)$	X	X	7.10
f_8: $adj-mod$ $(economic)$		X	3.70

Με όλα τα χαρακτηριστικά μιας λέξης και μπορούμε να ορίσουμε τα κοινά μεταξύ των λέξεων 1 και 2 ως:

$$\text{sim}(w_1, w_2) = \frac{2 * I\,(F(w_1) \cap F(w_2))}{I\,(F(w_1)) + I(F(w_2))}$$

, όπου I(S) είναι το ποσό της πληροφορίας που εμπεριέχεται στο σύνολο των χαρακτηριστικών S.

Υποθέτοντας ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο,

$$\text{I(S)} = -\sum_{f \in S} \log P(f)$$

, όπου P(f) είναι η πιθανότητα του χαρακτηριστικού f. Η μέγιστη ομοιότητα παίρνει την τιμή 1. Η 4η στήλη στον πίνακα δείχνει το ποσό της πληροφορίας που εμπεριέχεται σε κάθε χαρακτηριστικό.

Σημασιολογική Ομοιότητα σε μια Ταξονομία

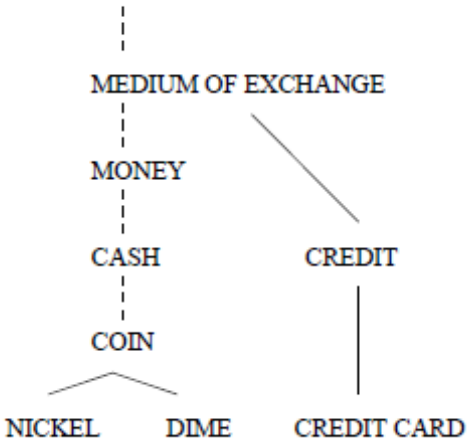
Η σημασιολογική ομοιότητα σε δύο κλάσεις C και C' δεν αφορά τις ίδιες τις κλάσεις, αλλά τις γενικεύσεις τους. Επομένως ορίζουμε το sim(C, C') ως η ομοιότητα μεταξύ του x και x' αν όλα όσα ξέρουμε για το x και x' είναι ότι x ∈ C και x' ∈ C'. Υποθέτοντας ότι η ταξονομία είναι δενδρική, αν x1 ∈ C_0 ∧ w_2∈ C0, όπου C0 είναι η πιο συγκεκριμένη κλάση για το C_1 και C_2, τότε:

$$\text{sim}(x_1, x_2) = \frac{2 * \log P(C_0)}{\log P(C_1) + logP(C_2)}$$

4.1.6. Resnik

Η μέθοδος του Resnik (Resnik, 1999), κάνει και αυτή χρήση του κλασσικού τρόπου υπολογισμού της σημασιολογικής ομοιότητας σε μια ταξονομία λαμβάνοντας υπόψη την απόσταση μεταξύ των κόμβων που αντιστοιχούν στις λέξεις που συγκρίνονται. Όσο πιο μικρή η διαδρομή, τόσο πιο όμοια είναι. Ωστόσο, αυτό βασίζεται στην λογική ότι οι σύνδεσμοι μεταξύ των κόμβων αναπαριστούν ομοιόμορφες αποστάσεις. Δυστυχώς αυτό είναι δύσκολο να οριστεί και πιο δύσκολο να χειριστεί.

Έστω ότι C είναι το σύνολο των εννοιών σε μια IS-A ταξονομία, που επιτρέπει πολλαπλή κληρονομικότητα. Το κλειδί για την ομοιότητα δύο εννοιών είναι η προέκταση στην οποία μοιράζονται πληροφορία. Η μέθοδος μέτρησης των πλευρών (edge-counting) το βρίσκει αυτό έμμεσα, αφού αν η μικρότερη διαδρομή των IS-A συνδέσμων μεταξύ των κόμβων είναι μεγάλη, αυτό σημαίνει ότι έχει προχωρήσει πολύ στην ταξονομία, σε αφηρημένες έννοιες, ώστε να βρει το κοινό τους σημείο.



Συσχετίζοντας πιθανότητες με έννοιες στην ταξονομία, είναι πιθανό να βρούμε την ίδια λογική, αποφεύγοντας την αναξιοπιστία των αποστάσεων. Έστω ότι η ταξονομία ενισχύεται με μια συνάρτηση p: C→ [0,1] ώστε για κάθε c ∈ C, p(c) είναι η πιθανότητα να βρεθεί ένα στιγμιότυπο της έννοιας c. Αυτό υπονοεί ότι το p είναι μονοτονικά μη φθίνων, όσο κάποιος προχωράει στην ταξονομία: Αν c_1 IS-A c_2, τότε $p(c_1) \leq p(c_2)$. Επιπλέον αν η ταξονομία έχει έναν μοναδικό Ανώτατο κόμβο τότε η πιθανότητα είναι 1. Ακολουθώντας την επιχειρηματολογία της θεωρίας της πληροφορίας (Ross, 1976), το περιεχόμενο της πληροφορίας μιας έννοιας c μπορεί να ποσοτικοποιηθεί ως αρνητικός λογάριθμος ομοιότητας, -log p(c).

Αυτός ο ποσοτικός χαρακτηρισμός της πληροφορίας παρέχει έναν νέο τρόπο μέτρησης της σημασιολογικής ομοιότητας. Όσο περισσότερη πληροφορία δύο έννοιες μοιράζονται, τόσο πιο όμοια είναι και η πληροφορία που μοιράζεται από τις δύο έννοιες υποδεικνύεται από το περιεχόμενο της πληροφορίας των δύο εννοιών που υπάγονται στην ταξονομία

$$\text{sim}(c_1, c_2) = \max_{c \in S(c_1,c_2)} [-\log p(c)]$$

, όπου $S(c_1, c_2)$ είναι το σετ των εννοιών που περικλείουν το c_1και c_2. Μια κλάση που επιτυγχάνει τη μέγιστη τιμή στην παραπάνω συνάρτηση ορίζεται ως "Most Informative Subsumer". Συνήθως είναι ένας μοναδικός Most Informative Subsumer, αν και δεν ισχύει πάντα.

Παρόλο που η ομοιότητα υπολογίζεται λαμβάνοντας υπόψη τα ανώτατα όρια των δύο εννοιών, το μέτρο της πληροφορίας μπορεί να αναγνωρίσει τα ελάχιστα ανώτατα όρια, αφού καμία κλάση δεν παρέχει λιγότερη πληροφορία από τις υποκλάσεις της. Αυτό μπορεί να κάνει τη διαφορά στις περιπτώσεις κληρονομικότητας. Δύο διαφορετικοί πρόγονοι-κόμβοι μπορούν και οι δύο να είναι το ελάχιστο ανώτατο όριο, αν μετρηθούν χρησιμοποιώντας την απόσταση στον γράφο, αλλά θα έχουν διαφορετικές τιμές χρησιμοποιώντας το περιεχόμενο.

Η υλοποίηση της μεθόδου γίνεται με τη χρήση της ταξονομίας εννοιών του WordNet. Οι συχνότητες των εννοιών στην ταξονομία υπολογίστηκαν χρησιμοποιώντας τις συχνότητες των ουσιαστικών από το Brown Corpus of American English, μια μεγάλη συλλογή κειμένου. Κάθε ουσιαστικό μετρήθηκε σαν μία ύπαρξη από κάθε κλάση της ταξονομίας που το περιέχει. Για παράδειγμα το ουσιαστικό "dime" μετρήθηκε για τη συχνότητα του "dime, coin, cash, κλπ"

$$\text{freq(c)} = \sum_{n \in \text{words(c)}} count(n)$$

, όπου words(c) είναι το σετ των λέξεων που υπάρχουν στην έννοια c. Οι πιθανότητες της έννοιας υπολογίστηκαν ως σχετική συχνότητα:

$$\hat{p}(c) = \frac{freq(c)}{N}$$

, όπου N ο συνολικός αριθμός των ουσιαστικών που μετρήθηκαν.

Για να οριστούν ταξονομικές πιθανότητες με σκοπό να μετρηθεί η σημασιολογική ομοιότητα, το παρών μοντέλο σχετίζει μια χωριστεί διωνυμική διανεμημένη τυχαία μεταβλητή για κάθε έννοια. Αυτό είναι, από τη μεριά της κάθε έννοιας c, ένα ουσιαστικό που είναι ή όχι στιγμιότυπο αυτής της έννοιας, με πιθανότητας p(c) και 1-p(c), αντίστοιχα. Αντίθετα με ένα μοντέλο το οποίο είναι μία απλή πολυωνυμική μεταβλητή με τιμές ανάμεσα στο σύνολο των εννοιών, αυτός ο σχεδιασμός ορίζει πιθανότητα 1 στην κορυφαία έννοια της ταξονομίας, οδηγώντας στις επιθυμητές συνέπειες που είναι ότι το περιεχόμενο της πληροφορίας είναι μηδενικό.

4.2. Μετρικές Corpus-Based

Οι Corpus-Based μετρικές, χρησιμοποιούν πιθανολογικές προσεγγίσεις για να αποσαφηνίσουν τη σημασία των λέξεων. Αποτελούνται από μη επιτηρούμενες μεθόδους, που χρησιμοποιούν την εμπεριεχόμενη πληροφορία και τα πρότυπα που παρατηρούνται στο κείμενο για να φτιάξουν σημασιολογικό προφίλ των λέξεων. Αντίθετα με τις Knowledge-Based μεθόδους, που έχουν περιορισμό στην εφαρμογή, οι Corpus-Based μπορούν να εντοπίσουν μια ομοιότητα ανάμεσα σε οποιεσδήποτε δύο λέξεις, υπό την προϋπόθεση ότι εμφανίζονται σε ένα πολύ μεγάλο κείμενο που χρησιμοποιείται για την εκπαίδευση του συστήματος.

4.2.1. Latent Semantic Analysis (LSA)

Η Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, An introduction to Latent Semantic Analysis, 1998), είναι μια θεωρία και μέθοδος για την εξαγωγή και αναπαράσταση της σημασίας των λέξεων με στατιστικούς υπολογισμούς που εφαρμόζονται σε ένα μεγάλο λεκτικό δείγμα (Landauer & Dumais, 1997). Αφού επεξεργαστεί ένα μεγάλο δείγμα κειμένου, τροποποιημένο ώστε να μπορεί να διαβαστεί από τον υπολογιστή, η LSA αναπαριστά τις λέξεις που χρησιμοποιήθηκαν σε αυτό, καθώς και κάθε σετ των λέξεων αυτών, σαν δείκτες σε ένα πολύ υψηλής διάστασης "σημασιολογικό χώρο". Η LSA βασίζεται σε τεχνική απλής ανάλυσης τιμής μιας μαθηματικής μήτρας.

Το νόημα των κειμένων και των λέξεων που παράγονται από την LSA, μπορούν να αναπαραστήσουν μια μεγάλη ποικιλία ανθρώπινων φαινομένων κατανόησης, από την δημιουργία της αναγνώρισης του λεξικού, μέχρι κατηγοριοποίηση των λέξεων, κατανόηση του νοήματός τους, κ.α.

Η LSA μπορεί να ερμηνευτεί με δύο τρόπους. Πρώτον, ως πρακτική μέθοδος για να βρούμε τις κατά προσέγγιση εκτιμήσεις της χρήσης του περιεχομένου μιας λέξης σε μεγαλύτερα τμήματα κειμένου και τις σημασιολογικές ομοιότητες μεταξύ των λέξεων και των

τμημάτων κειμένων και δεύτερον, σαν ένα μοντέλο υπολογισμών και αναπαραστάσεων σημαντικών τμημάτων της απόκτησης και αξιοποίησης της γνώσης.

Ως πρακτική μέθοδος για τον ορισμό της σημασίας της λέξης, η LSA παράγει λέξη προς λέξη (word-to-word), λέξη προς κείμενο (word-to-passage) και κείμενο προς κείμενο (passage-to-passage) μέτρα σημασιολογικών και όχι μόνο συσχετίσεων. Η αναπαράσταση της σημασίας μιας λέξης βρίσκεται μόνο από την ανάλυση του κειμένου και χωρίς καθόλου Βάσεις Γνώσης. Παρόλο που η γνώση που προέρχεται από την LSA δεν είναι τέλεια, θεωρείται ότι μπορεί να προσφέρει μια πολύ κοντινή προσέγγιση ως προς τον τρόπο που αποκτάται η γνώση από τον ίδιο τον άνθρωπο. Οι περιορισμοί που προκύπτουν από την συγκεκριμένη μέθοδο είναι ότι δεν κάνει χρήση της σειράς που εμφανίζονται οι λέξεις, και επομένως ούτε της συντακτικής ανάλυσής, της λογικής ή της μορφολογίας τους. Παρόλα αυτά τα αποτελέσματα είναι αρκετά ικανοποιητικά.

Η LSA διαφοροποιείται από μερικές στατιστικές προσεγγίσεις γιατί οι συσχετίσεις των δεδομένων εισόδου αποτελούνται απλά από λέξεις και εκφράσεις που δεν έχουν κανένα νόημα αντί για συνεχόμενες λέξεις. Θα μπορούσαμε να πούμε ότι η LSA αναπαριστά το νόημα μιας λέξης σαν ένα είδος του μέσου νοήματος όλων των κειμένων που εμφανίζεται η λέξη και το νόημα ενός κειμένου σαν τον μέσο όρο του νοήματος όλων των λέξεων που περιέχει.

Η LSA ως απόκτηση και αναπαράσταση γνώσης, θυμίζει τον τρόπο με τον οποίο οι άνθρωποι μαθαίνουν και αποκτούν περισσότερη γνώση μέσα από τις εμπειρίες τους. Ο μηχανισμός της LSA αποτελείται απλά από τον χειρισμό ενός πολύ μεγάλου αριθμού επανεμφανιζόμενων σχέσεων στον χώρο μιας συγκεκριμένη διάστασης. Οπότε πρέπει αρχικά να οριστεί ο χώρος/τομέας του κειμένου. Η LSA το κάνει αυτό με ανάλυση σε μια μήτρα, λαμβάνοντας υπόψη πληροφορία που εμπεριέχεται μέσα στις σταθερές, δομικές συσχετίσεις και στην αμοιβαία συνεπαγωγή των τοπικών παρατηρήσεων που διατίθενται μέσα από την εμπειρία.

Το πρώτο βήμα είναι η αναπαράσταση του κειμένου σαν μια μήτρα στην οποία κάθε λέξη αναπαριστάται σε μία γραμμή και κάθε στήλη είναι ένα κείμενο. Κάθε κελί εμπεριέχει τη συχνότητα στην οποία η λέξη εμφανίζεται στο αντίστοιχο κείμενο, όπως καθορίζει η στήλη.

WORD	Sentence 1	Sentence 2	Sentence 3
human	1	0	0
computer	1	1	0
user	0	1	1
system	0	1	2

Κατόπιν, κάθε κελί ορίζει βάρος που εκφράζει τη σημασία της λέξης στο συγκεκριμένο κείμενο, υπολογίζοντας τη σημαντικότητα, με σκοπό να μιμηθεί καλύτερα την αντίστοιχη ανθρώπινη διαδικασία. Για εξομοίωση της γλώσσας, η καλύτερη απόδοση παρατηρείται όταν οι συχνότητες συσσωρεύονται με έναν γραμμικό τρόπο σε κελιά

$$\log(freq_{ij} + 1)$$

,όπου $freq_{ij}$ είναι η συχνότητα της λέξης i στο κείμενο j, και αντιστρόφως με τη συνολική εμφάνιση της λέξης.

Ύστερα, η LSA εφαρμόζει στη μήτρα Αποσύνθεση Ιδιάζουσων Τιμών (Singular Value Decomposition - SVD). Η ορθογώνια μήτρα αναλύεται σε τρεις άλλες μήτρες. Η μία περιγράφει την γραμμής ως διανύσματα, η άλλη περιγράφει την στήλη με τον ίδιο τρόπο και η τρίτη είναι μια διαγώνια μήτρα που εμπεριέχει μια κλιμάκωση τιμών τέτοια ώστε όταν τα 3 συστατικά πολλαπλασιάζονται να ξαναδημιουργείται η αρχική μήτρα. Με αυτό το βήμα, όταν ανακατασκευάζουμε τη μήτρα των λέξεων, μερικές λέξεις εμφανίζονται με μεγαλύτερη ή μικρότερη συχνότητα και επίσης εμφανίζονται λέξεις που δεν υπήρχαν στην αρχική μήτρα. Ο λόγος που γίνεται αυτό είναι γιατί λαμβάνονται υπόψη οι έμμεσες σχέσεις των λέξεων μεταξύ τους. Οι k μεγαλύτερες ιδιάζουσες τιμές διατηρούνται και οι υπόλοιπες ορίζονται σε ο. Η αναπαράσταση του αποτελέσματος είναι η καλύτερη k-διαστάσεων προσέγγιση, στην αρχική μήτρα, με τη λιγότερο τετραγωνική έννοια. Κάθε κείμενο και όρος αναπαρίστανται σαν k-διαστάσεων διάνυσμα σε έναν χώρο που προέρχεται από το SVD. Στις περισσότερες εφαρμογές, οι k-διαστάσεις είναι πολύ μικρότερες από τον αριθμό των όρων στην αρχική μήτρα του κειμένου. Εικάζεται ότι σε πολλές περιπτώσεις, όπως η εξομοίωση γλώσσας, η βέλτιστη απόδοση διαστάσεων είναι σχετική με το θέμα που εξομοιώνεται και ορίζεται εμπειρικά. Το βήμα της μείωσης των διαστάσεων, ορίζει μια διαφορετική τιμή για την ομοιότητα κάθε λέξης προς μια άλλη, είτε έχουν εμφανιστεί σε κοινό περιεχόμενο, είτε όχι.

Ο αλγόριθμος της Latent Semantic Analysis έχει τα παρακάτω βήματα:

1. Ορίζουμε μια ορθογώνια μήτρα X, όρων t και κειμένων p (t * p).
2. Την αποσυνθέτουμε σε 3 άλλες μήτρες, ώστε X=T*S*P^T, όπου T είναι μια t*r μήτρα, P είναι μια p*r μήτρα και S είναι μια r*r διαγώνια μήτρα με τις εγγραφές να ταξινομούνται με φθίνουσα σειρά. Οι εγγραφές της S είναι ιδιάζουσες τιμές και οι T και P μήτρες είναι τα αριστερά και δεξιά ιδιάζουσα διανύσματα, που ανταποκρίνονται στα διανύσματα των λέξεων(t) και των κειμένων(p).
3. Η LSA χρησιμοποιώντας SVD κρατάει μόνο τις k μεγαλύτερες ιδιάζουσες τιμές και τα σχετικά διανύσματα, ώστε X=T_k*S_k*$P_k{}^T$ να είναι το μειωμένων διαστάσεων αποτέλεσμα.

4.2.2. Explicit Semantic Analysis (ESA)

Ο υπολογισμός της σημασιολογικής συσχέτισης κειμένων φυσικής γλώσσας απαιτεί πρόσβαση σε μεγάλες ποσότητες κοινής και καθετοποιημένης γνώσης. Η μέθοδος το ESA (Gabrilovich & Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, 2007), αναπαριστά το νόημα των κειμένων σε ένα high-dimensional space εννοιών που προέρχονται από τη Wikipedia. Χρησιμοποιούνται μηχανικές τεχνικής μάθησης για την αναπαράσταση του νοήματος κάθε κειμένου ως ένα σταθμισμένο διάνυσμα από έννοιες βασισμένες στη Wikipedia.

Ένα σημαντικό πλεονέκτημα της προσέγγισης αυτής, είναι η τεράστια ποσότητα άρθρων που έχουν οργανωθεί σωστά από ανθρώπους μέσα από τη δομή της Wikipedia. Επίσης, η Wikipedia εξακολουθεί να είναι ανοιχτό και δυναμικό project, οπότε το βάθος και το πλάτος της συνεχώς αυξάνονται με την πάροδο του χρόνου. Έχει διαπιστωθεί ότι η ποιότητα της Wikipedia είναι αντίστοιχη αυτής της Britannica (Giles, 2005).

Αρχικά, γίνεται χρήση τεχνικών μηχανικής μάθησης για να χτιστεί ένας σημασιολογικός διερμηνέας που αντιστοιχίζει τα τμήματα του κειμένου φυσικής γλώσσας σε σταθμισμένες ακολουθίες εννοιών της Wikipedia ταξινομημένες από τη σχετικότητά τους ως προς το

κείμενο. Κατά αυτό τον τρόπο τα κείμενα εισόδου, παρουσιάζονται σαν σταθμισμένα διανύσματα εννοιών, που ονομάζονται διερμηνευμένα διανύσματα (interpreted vectors). Ο υπολογισμός της σημασιολογικής συσχέτισης των κειμένων μετράει στο να υπολογιστούν τα διανύσματά τους στο χώρο που ορίζεται από τις έννοιες, για παράδειγμα να χρησιμοποιηθεί η μετρική συνημίτονου (cosine metric) (Zobel & Moffat, 1998). Η παρούσα σημασιολογική ανάλυση είναι ρητή υπό την έννοια ότι χειρίζεται τις έννοιες που βασίζονται στην ανθρώπινη γνώση, αντί για κρυμμένες έννοιες (latent concepts) που χειρίζεται η LSA.

Τα κείμενα εισόδου δίνονται στην ίδια φόρμα (απλό κείμενο) όπως τα άρθρα της wikipedia. Επομένως, μπορούν να χρησιμοποιηθούν συμβατικοί αλγόριθμοι ταξινόμησης για να κατατάξουν τις έννοιες που αναπαριστούνται από αυτά τα άρθρα βάσει τις σχετικότητάς τους ως προς το τμήμα του κειμένου που δίνεται σαν είσοδος. Αυτό είναι το κλειδί που μας επιτρέπει να χρησιμοποιούμε την εγκυκλοπαίδεια απευθείας, χωρίς να χρειάζεται βαθιά κατανόηση της γλώσσας ή προ-κατηγοριοποίηση της γενικής γνώσης.

Κάθε έννοια της Wikipedia αναπαριστάται σαν χαρακτηριστικό διάνυσμα των λέξεων που βρίσκονται στο αντίστοιχο άρθρο. Οι είσοδοι αυτών των διανυσμάτων ορίζουν βάρη χρησιμοποιώντας το TFIDF σχήμα (Salton & McGill, An introduction to modern information retrieval, 1983). Αυτά τα βάρη ορίζουν τη ποιότητα των συσχετίσεων μεταξύ των λέξεων και των εννοιών. Για να επιταχύνουμε τη σημασιολογική διερμηνεία, χτίζεται ένα αντεστραμμένο ευρετήριο, το οποίο ταιριάζει κάθε λέξη σε μια λίστα εννοιών που εμφανίζεται. Επίσης, χρησιμοποιείται αυτό το ευρετήριο για να απορριφθούν μη μεγάλης σημασίας συσχετίσεις μεταξύ λέξεων και εννοιών, αφαιρώντας τις έννοιες στις οποίες τα βάρη τους για τη λέξη που δίνεται είναι πολύ χαμηλά.

Πρώτα για το τμήμα κειμένου που δίνεται, αναπαρίσταται σαν διάνυσμα με τη χρήση του TFIDF σχήματος. Ο σημασιολογικός διερμηνευτής περνάει από όλες τις λέξεις του κειμένου, και ανακτά τις αντίστοιχες εγγραφές από το αντεστραμμένο ευρετήριο και τις ενώνει με ένα σταθμισμένο διάνυσμα εννοιών που αναπαριστούν τη ζητούμενη λέξη. Μετά το διάνυσμα του σημασιολογικού διερμηνευτή V για το κείμενο T είναι ένα διάνυσμα μήκους N, στο οποίο το βάρος κάθε έννοιας c_j ορίζεται ως:

$$\sum_{w_i \in \mathrm{T}} v_i * k_j$$

Οι εγγραφές αυτού του διανύσματος αντανακλούν τη σχετικότητα των αντίστοιχων εννοιών στο κείμενο T. Για να υπολογιστεί η σημασιολογική συσχέτιση ενός ζευγαριού τμημάτων κειμένου, υπολογίζουμε τα διανύσματά τους χρησιμοποιώντας μετρική συνημίτονου (cosine metric). Η μεθοδολογία αυτή μπορεί κάλλιστα να χρησιμοποιηθεί και για Word Sense Disambiguation.

4.2.3. Salient Semantic Analysis (SSA)

Ένα σύστημα για να επιτύχει ικανοποιητική εννοιολογική κατανόηση, δεν πρέπει μόνο να έχει ένα μεγάλο υπόβαθρο γνώσης, αλλά θα πρέπει να είναι ικανό να δημιουργεί γενικεύσεις και αφαιρετικές έννοιες. Στην Salient Semantic Analysis εισάγεται μια νέα και αξιόπιστη ερμηνεία των εννοιών που περιέχονται στο κείμενο που εξετάζεται, με το να αποδίδονται σε αυτές καλά ορισμένες και σαφής έννοιες με τη χρήση των εγκυκλοπαιδικών ορισμών. Για την υλοποίηση χρησιμοποιήθηκαν εγκυκλοπαιδικές πηγές όπως η Wikipedia, όπου έχουν οριστεί οι προερχόμενες έννοιες κάθε άρθρου. Το σχόλιο μπορεί να επεκταθεί, υλοποιώντας Word Sense Disambiguation heuristics. Περαιτέρω ανάλυση των παραγόμενων σημασιολογικών διανυσμάτων αναπαριστά το νόημα των λέξεων σε ένα χώρο εννοιών. Αυτή

η μεταφορά από έναν απλό χώρο λέξεων σε έναν πλουσιότερο χώρο εννοιών αποκρίνεται καλύτερα σε αυτό που ορίζεται ως αρχή γνώσης (Lenat & Feigenbaum, 1991).

Το συντακτικό έχει να κάνει με τους κανόνες που ορίζουν την φυσική γλώσσα. Ουσιαστικά είναι αυτό που καθορίζει τη σειρά με την οποία οι λέξεις παρουσιάζονται στο κείμενο. Η σημασιολογία βασίζεται στην ανάλυση των σχέσεων των λέξεων, φράσεων και προτάσεων ώστε να βρει το φανερό ή υπονοούμενο νόημα και ερμηνεία. Έτσι, επιτρέπει στην φυσική γλώσσα να είναι ένα μέσο ανταλλαγής γνώσης και πληροφορίας. Η πραγματολογία αφορά μη γλωσσολογικά στοιχεία που μπορούν να βρεθούν από το κείμενο, όπως το ύφος του κειμένου, η πρόθεση του συγγραφέα, κ.α. Αυτά τα επιπλέον στοιχεία επιτρέπουν περαιτέρω αποσαφήνιση του μηνύματος και συναφή σημασιολογική διερμηνεία.

Η έννοια της παράφρασης αφορά όταν βγαίνει μια λανθασμένη ή μη έγκυρη συμπερασματολογία επειδή δεν έγινε σαφές και ξεκάθαρο το νόημα του αρχικού κειμένου που δόθηκε ως είσοδος.

Μια από τις πρώτες εφαρμογές της συσχέτισης κειμένων είναι πιθανώς το διανυσματικό μοντέλο στην ανάκτηση πληροφορίας, όπου το πιο σχετικό έγγραφο σε ένα ερώτημα εισόδου, ορίζεται κατατάσσοντας τα έγγραφα σε φθίνουσα σειρά βάσει του ταιριάσματός τους ως προς τα δεδομένα εισόδου (Salton & Lesk, 1968). Η συσχέτιση κειμένων έχει επίσης χρησιμοποιηθεί για αντίστοιχη χρήση, κατηγοριοποίηση των κειμένων, WSD, κ.α..

Με μερικές εξαιρέσεις η τυπική προσέγγιση για να βρεθεί η συσχέτιση μεταξύ δύο τμημάτων είναι να χρησιμοποιηθεί μια απλή μέθοδος λεξικολογικού ταιριάσματος και να παραχθεί ένα σκορ βασισμένο στον αριθμό των λεκτικών μονάδων που βρίσκονται στα δύο τμήματα εισόδου. Βελτιώσεις σε αυτή την απλή μέθοδο είναι το stemming, αφαίρεση των stop-words, POS tagging, longest subsequence matching, όπως και διάφοροι τρόποι στάθμισης και κανονικοποίησης. Αν και επιτυχημένες μέχρι κάποιο βαθμό, αυτές οι μέθοδοι δεν μπορούν πάντα να αναγνωρίσουν τη σημασιολογική συσχέτιση των κειμένων. Χαρακτηριστικό παράδειγμα είναι οι φράσεις "we own a pet" και "I love animals" όπου οι περισσότερες μετρικές θα αποτύχουν να αναγνωρίσουν κάποιο τύπο σύνδεσης του μεταξύ τους λεξιλογίου.

Στα μοντέλα που βασίζονται στα χαρακτηριστικά (Feature-Based), κάθε λέξη επεκτείνεται χρησιμοποιώντας κάποιο χώρο χαρακτηριστικών που παράγει το διάνυσμα που την αναπαριστά. Τέτοια παραδείγματα είναι η LSA και η ESA.

Η πλειοψηφία των ψυχο-γλωσσολόγων χρησιμοποιούν το μοντέλο των Dijk & Kintsch (Dijk & Kintsch, 1983). με τα επίπεδα κατανόησης της ομιλίας (Graesser, Millis, & Zwaan, 1997), τα οποία είναι:

- Επιφανειακός κώδικας (Surface code): Έχει να κάνει με την αναπαράσταση του κειμένου και τη συντακτική δομή. Είναι συνήθως μια μικρής-ζωής μνήμη των λέξεων και των φράσεων στο κείμενο.
- Βάση κειμένου (Textbase): Αναπαριστά τις λογικές θέσεις που περικλείονται στο κείμενο, οι οποίες ενσαρκώνουν όλα τα πιθανά νοήματα των τμημάτων κειμένου.
- Μοντέλο καταστάσεων (Situational model): Δείχνει τον μικρόκοσμο που φανερώνει το κείμενο, ο οποίος ενσωματώνει τα συμπεράσματα που βγάζει ο αναγνώστης βάσει των γνώσεών του. Σε αυτό το σημείο τα τμήματα κειμένου χάνουν τη μοναδικότητά τους και ενοποιούνται στη γνώση του αναγνώστη (υποκειμενική).
- Επίπεδο επικοινωνίας (Communication level): Αναπαριστά την κατανόηση της πρόθεσης του συγγραφέα και την αναγνώριση της πραγματολογίας.
- Ύφος κειμένου (Text genre): Αναπαριστά το θεματικό σκοπό του κειμένου.

Εφόσον, η σημασιολογική συσχέτιση αναφέρεται στη χρήση της γνώσης του αναγνώστη, είναι πιο κατάλληλο, βάσει των ευρημάτων από την ψυχογλωσσολογία, να βασιστούμε σε συλλογές που εξηγούν (όπως το Wikipedia) ή σε λεκτικά δείγματα με

επιχειρηματολογία για να δημιουργήσουμε και να αναπαραστήσουμε τη Βάση Γνώσης για τον υπολογιστή.

Αν και υπάρχουν πολλά μοντέλα που εξηγούν και εξομοιώνουν την κατανόηση του λόγου, ένα μοντέλο υπερτερεί ως το πιο ακριβές μέχρι σήμερα: Το Construction-Integration του Kinstch (Kintsch, 1988). Σε αυτό η γνώση αναπαριστάται σαν δίκτυο συσχετίσεων, όπου οι κόμβοι αναπαριστούν έννοιες ή προτάσεις και οι ενώσεις την δύναμη της συσχέτισης των κόμβων.

Το μοντέλο Salient Semantic Analysis (SSA) δημιουργήθηκε βασισμένο στην υπόθεση ότι η αναπαράσταση του μέσου αναγνώστη για κάθε κείμενο εμπεριέχει ένα νοητικό πλαίσιο το οποίο κρατάει και ταιριάζει τις αμφίβολες έννοιες που παρατηρούνται στο κείμενο με έννοιες από τη Βάση Γνώσης που ήδη έχει (γνώσεις και εμπειρίες). Τα σημασιολογικά προφίλ βασίζονται στην Wikipedia χρησιμοποιώντας τους συνδέσμους μεταξύ των άρθρων για τις συσχετίσεις. Αυτά τα προφίλ βοηθάνε στην αποσαφήνιση των αμφίβολων χρήσης εννοιών. Στην ερμηνεία μια λέξη ορίζεται ως ένα σετ εννοιών που μοιράζονται το περιεχόμενό τους και σταθμίζονται από το pointwise mutual information τους.

Το σύστημα που χρησιμοποιείται για αποσαφήνιση είναι όμοιο με αυτό που χρησιμοποιείται στο Wikify, το οποίο εκχωρεί τα άρθρα της Wikipedia σε λέξεις ή φράσεις που έχουν υψηλά νούμερα υπερσυνδέσμων. Δηλαδή, πρώτα ορίζει τις φράσεις που έχουν μεγάλη πιθανότητα (>= 0.5) να οριστούν ως φράσεις-κλειδιά και να ανταποκρίνονται σε μια έννοια. Αυτή η πιθανότητα υπολογίζεται ως τον αριθμό των φορών που η λέξη ή φράση εμφανίζεται μέσα στο χειροκίνητα παραγόμενο σύνδεσμο, διαιρούμενο από το συνολικό αριθμό των φορών που η λέξη ή φράση εμφανίζεται στη Wikipedia. Για να υπολογιστεί η σημασιολογική συσχέτιση για ένα ζευγάρι λέξεων, η επικάλυψη μεταξύ των σημασιολογικών προφίλ των λέξεων συναθροίζεται για να παράγει ένα σκορ συσχέτισης:

$$score_{cos}(A,B) = \frac{\sum_{y=1}^{N}(P_{iy} * P_{jy})^{\gamma}}{\sqrt{\sum_{y=1}^{N} P_{iy}^{2\gamma} * \sum_{y=1}^{N} P_{jy}^{2\gamma}}}$$

Αφού το συνημίτονο είναι ένας κανονικοποιημένος αριθμός που ορίζει 1 για τους ίδιους όρους, επηρεάζεται αρνητικά από έναν αραιό χώρο και παράγει μικρά σκορ για τα κοντινά συνώνυμα. Αυτό δημιουργεί μεγάλα σημασιολογικά κενά μεταξύ των όρων που ταιριάζουν και των απόλυτα συσχετισμένων όρων. Για να μειωθεί αυτό το κενό και τα σκορ να έχουν μεγαλύτερη ουσία, δημιουργείται και μια κανονικοποίηση παράγοντα "λ":

$$sim(A,B) = \begin{cases} 1 & score_{cos}(A,B) > \lambda \\ \dfrac{score_{cos}(A,B)}{\lambda} & score_{cos}(A,B) \leq \lambda \end{cases}$$

Για να υπολογίσουμε τη σημασιολογική συσχέτιση μεταξύ δύο τμημάτων κειμένου, χρησιμοποιούμε τα ίδιο προφίλ λέξεων που δημιουργήθηκαν από τις προεξέχων εγκυκλοπαιδικές έννοιες, σε συνδυασμό με μια απλοποιημένη έκδοση της τεχνικής ταιριάσματος bipartite-graph.

Το T_a και T_b είναι δύο τμήματα κειμένου μεγέθους a και b. Αφαιρώντας όλα τα stop-words, καθορίζεται ο αριθμός των κοινών όρων (w) μεταξύ T_a και T_b. Μετά υπολογίζεται η σημασιολογική συσχέτιση όλων των πιθανών ταιριασμάτων μεταξύ των μη-κοινών όρων στο T_a και T_b, χρησιμοποιώντας κανονικοποίηση βάσει του συνημίτονου. Αυτοί οι πιθανοί συνδυασμοί φιλτράρονται περαιτέρω δημιουργώντας μια λίστα φ, που κρατάει τα ισχυρότερα

σημασιολογικά ταιριάσματα μεταξύ των όρων, με τέτοιο τρόπο ώστε κάθε όρος να ανήκει σε μόνο ένα ταίριασμα:

$$\text{sim}(T_a, T_b) = \frac{\left(\omega + \sum_{i=1}^{|\varphi|} \varphi_i\right) * (2ab)}{a+b}$$

, όπου w είναι ο αριθμός των κοινών όρων μεταξύ των τμημάτων κειμένου και φ_i είναι το σκορ ομοιότητας για το i ταίριασμα.

5. Διάφορες προσεγγίσεις για την εύρεση της σημασιολογικής ομοιότητας κειμένων

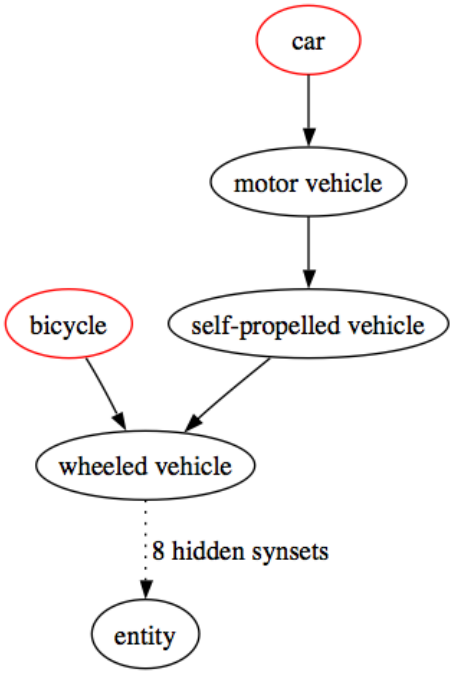
Η εύρεση σημασιολογικής ομοιότητας μεταξύ κειμένων, όπως θα πρέπει να έχει γίνει ήδη αντιληπτό, είναι ένας βασικός κλάδος της επεξεργασίας φυσικής γλώσσας. Χρησιμοποιείται σε πολλά προβλήματα όπως στο να παράγει μη ξεκάθαρο νόημα, στην εξόρυξη πληροφορίας, στην κατηγοριοποίηση κειμένων, στην δημιουργία περιλήψεων, κ.α., αλλά χρησιμοποιείται επίσης και σαν αυτόνομη διαδικασία σε προβλήματα που απλά πρέπει να οριστεί η σημασιολογική ομοιότητα δύο κειμένων. Γι αυτό το λόγο υπάρχουν αρκετές προσεγγίσεις στο πρόβλημα, η κάθε μία από τη δική της οπτική γωνία και εξυπηρετώντας τους δικούς της σκοπούς. Η πλειοψηφία των προσεγγίσεων αυτών κάνει χρήση των παραπάνω τεχνικών που είδαμε και περιστρέφεται κυρίως στο να αποδώσει ένα ποσοστό ομοιότητας μεταξύ δύο κειμένων σε κανονικοποιημένη τιμή. Παρακάτω αναφέρονται αναλυτικότερα οι κυριότερες από τις υπάρχουσες προσεγγίσεις και τεχνικές.

5.1. Comparing Similarity - Andreas Jensen & Niklas Boss

Αυτή η προσέγγιση προτείνει δύο διαφορετικές λύσεις για τη μέτρηση της σημασιολογικής ομοιότητας των κειμένων. Η μία λύση είναι χρησιμοποιώντας το WordNet και η άλλη χρησιμοποιώντας τη γνωστή τεχνική "edit distance" μεταξύ γραμματοσειρών.

Η κεντρική ιδέα είναι να αναπαρασταθεί η σημασιολογική ομοιότητα μεταξύ των λέξεων χρησιμοποιώντας μια λεξική Βάση Δεδομένων όπως το WordNet. Βρίσκοντας την τιμή ομοιότητας για κάθε συνδυασμό των λέξεων σε δύο προτάσεις, μπορεί να προκύψει μια τιμή σημασιολογικής ομοιότητας για τις προτάσεις αυτές. Με αντίστοιχο τρόπο, μπορεί να οριστεί και ομοιότητα μεταξύ κειμένων.

Στην πρώτη λύση με μεθόδους υπολογισμού ομοιότητας μεταξύ δύο synsets του δέντρου μιας ταξονομίας, βρίσκει πόσο κοντά είναι οι δύο λέξεις που συγκρίνονται μεταξύ τους. Οπότε, όσο πιο κοντά βρίσκονται, τόσο μεγαλύτερο βαθμό ομοιότητας έχουν. Κατόπιν, η απόστασή τους πρέπει να λαμβάνεται υπόψη. Γι αυτό το λόγω μπορεί να βρεθεί ο Lowest Common Ancestor (LCA) μεταξύ των δύο synset, όπως φαίνεται παρακάτω:



Αυτό θα οδηγήσει σε ένα αποτέλεσμα όπου το μεγάλο σκορ θα σημαίνει μεγάλη ομοιότητα, κάτι το οποίο δεν είναι το ζητούμενο. Αυτό που μας ενδιαφέρει είναι να δημιουργηθεί μια τιμή μεταξύ του 0 και του 1 όπου όσο πιο κοντά είναι στο 1, τόσο μεγαλύτερη ομοιότητα θα ορίζει. Για να γίνει αυτό υπάρχουν πολλές μετρικές που μπορούν να χρησιμοποιηθούν, όπως του Wu και Palmer, του Philip Resnik, των Leacock και Chodorow , κ.α.

Η μέθοδος των Leacock και Chodorow αν και δείχνει η καλύτερη που μπορεί να χρησιμοποιηθεί για τα ουσιαστικά, δεν μπορεί εύκολα να χρησιμοποιηθεί στα ρήματα, λόγω της διαφορετικότητας της ταξονομίας των ρημάτων στο WordNet. Για τη σύγκριση των ρημάτων προτείνεται η μετρική των Wu και Palmer.

Επομένως, θα πρέπει να οριστεί το Part-Of-Speech Tagging των λέξεων γιατί λόγω της ταξονομίας του WordNet δεν μπορούμε να βρούμε ομοιότητα μεταξύ ρημάτων και ουσιαστικών, αλλά μόνο μεταξύ λέξεων που ανήκουν στο ίδιο μέρος του λόγου. Για τη χρήση του POS Tagging χρησιμοποιείται ένας έτοιμος POS Tagger που έχει δημιουργηθεί στην Java. Σε αυτό το σημείο, κρίνεται χρήσιμο να οριστεί και ο μηχανισμός Word Sense Disambiguation που θα αποσαφηνίσει το word sense των λέξεων που θα συγκρίνουμε. Ένας από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους για Word Sense Disambiguation είναι ο Lesk Algorithm ο οποίος χρησιμοποιείται και σε αυτή την περίπτωση.

Αφού λοιπόν γίνει το POS Tagging, πρέπει να βρεθεί η ομοιότητα μεταξύ των λέξεων με τη χρήση της απόστασης Levenshtein. Προτείνεται να αφαιρεθούν όλες οι STOP-Words πριν την σύγκριση των προτάσεων, αφού μπορούν να επηρεάσουν σημαντικά τον βαθμό ομοιότητάς τους.

Κατόπιν, δημιουργείται μια μήτρα ομοιότητας (similarity matrix), η οποία για να αποδώσει σωστή τιμή μετατρέπεται σε bipartite graph. Οι ευρέως χρησιμοποιούμενες λέξεις αφαιρούνται για να μην επηρεάσουν τα αποτελέσματα ομοιότητας. Ωστόσο υπάρχει πιθανότητα οι λιγότερο κοινές λέξεις που υπάρχουν και στις δύο προτάσεις να μη συμμετέχουν στο νόημα των κειμένων. Αυτό σημαίνει ότι θα επηρεαστεί θετικά και το ποσό ομοιότητας που υπολογίζεται. Η πρόταση που γίνεται είναι να δημιουργηθεί μια συνάρτηση όπου οι υψηλές τιμές θα ανταμείβονται και οι χαμηλές θα δέχονται κάποια ποινή . Αυτό σημαίνει ότι αν και κάποιες λέξεις είναι όμοιες, αν το σύνολο είναι πολύ διαφορετικό, η τιμή

θα είναι αντίστοιχα χαμηλή. Γι αυτή τη χρήση προτείνεται η σιγμοειδής συνάρτηση (Sigmoid Function).

Αφού οριστεί η τιμή της ομοιότητας των προτάσεων με αντίστοιχη διαδικασία του bipartite graph μπορεί να βρεθεί και η ομοιότητα των κειμένων. Όπως μια πρόταση αποτελείται από λέξεις, αντίστοιχα ένα κείμενο αποτελείται από προτάσεις. Το σκορ ομοιότητας υπολογίζεται για κάθε ζευγάρι προτάσεων στα κείμενα που μπαίνουν στη μήτρα ομοιότητας. Η μήτρα μπορεί να μετατραπεί σε bipartite graph και χρησιμοποιώντας το δίκτυο ροής μπορεί να βρεθεί το μέγιστο bipartite ταίριασμα. Το τελικό σκορ ομοιότητας το βρίσκουμε χρησιμοποιώντας και πάλι τη σιγμοειδής συνάρτηση (Jensen & Boss, 2008).

5.2. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity - R. Mihalcea, C. Corley & C. Strapparava

(Mihalcea, Corley, & Strapparava, 2006) Εκτός από την ομοιότητα των λέξεων λαμβάνουν υπόψην και την ειδικότητα (specificity) των λέξεων, ώστε να μπορεί να δοθεί μεγαλύτερο βάρος σε ένα σημασιολογικό ταίριασμα μεταξύ δύο συγκεκριμένων λέξεων και δίνεται λιγότερη σημασία στην ομοιότητα που μετρήθηκε στις γενικές έννοιες. Η ειδικότητα αν και έχει μετρηθεί από το βάθος των λέξεων στη σημασιολογική ιεραρχία, ενισχύεται με βασισμένα στο λεκτικό δείγμα (corpus-based) μέτρα για την ειδικότητα της λέξη, που έχουν να κάνουν με τη διαμοιρασμένη πληροφορία που μαθεύτηκε από ένα μεγάλο λεκτικό δείγμα. Η ειδικότητα των λέξεων μετριέται με τη χρήση της Αντίστροφης Συχνότητας του Εγγράφου (Inverse Document Frequency - IDF), που ορίζει τον συνολικό αριθμό των εγγράφων στο λεκτικό δείγμα διαιρούμενο από τον συνολικό αριθμό των δεδομένων που εμπεριέχουν την λέξη (Jones, 1972) .

Δεδομένης της word-to-word ομοιότητας και της ειδικότητας, ορίζεται η σημασιολογική ομοιότητα μεταξύ δύο τμημάτων T_1 και T_2, χρησιμοποιώντας μια μετρική που συνδυάζει σημασιολογικές ομοιότητες του κάθε τμήματος του κειμένου. Πρώτα για κάθε λέξη w στο τμήμα T_1 προσπαθούμε να αναγνωρίσουμε τη λέξη στο τμήμα T_2 που έχει τη μεγαλύτερη σημασιολογική ομοιότητα, σύμφωνα με τα μέτρα word-to-word. Μετά, η ίδια διαδικασία εφαρμόζεται για να οριστεί η πιο όμοια λέξη στο T_1 ξεκινώντας από τις λέξεις του T_2. Στις ομοιότητες που θα βρεθούν, προστίθενται βάρη βάσει της αντίστοιχης ειδικότητας (specificity), αθροίζονται και κανονικοποιούνται με το μήκος κάθε τμήματος κειμένου. Τελικά, τα επιστρεφόμενα σκορ ομοιότητας συνδυάζονται με έναν απλό μέσο όρο.

Η παρακάτω συνάρτηση υπολογίζει το σκορ ομοιότητας μεταξύ των τμημάτων T_1 και T_2:

$$\text{sim}(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{T_1\}}(maxSim(w,T_2)*idf(w))}{\sum_{w\in\{T_1\}} idf(w)} + \frac{\sum_{w\in\{T_2\}}(maxSim(w,T_1)*idf(w))}{\sum_{w\in\{T_2\}} idf(w)}\right)$$

Αυτό το σκορ έχει τιμές μεταξύ 0 και 1, όπου το 1 σημαίνει ότι τα τμήματα είναι ίδια και 0 ότι δεν έχουν καμία ομοιότητα.

Οι μετρικές corpus-based word-to-word που χρησιμοποιούνται είναι οι Pointwise Mutual Information και η Latent Semantic Analysis. Στην Pointwise Mutual Information , με τη χρήση του NEAR query (επανεμφάνιση των λέξεων μέσα σε ένα παράθυρο δέκα λέξεων) και του AltaVista πέτυχε 72.5% σκορ.

Η LSA αντιμετωπίζεται σαν ένας τρόπος, ώστε να ξεπεραστούν τα πισωγυρίσματα που προκύπτουν από το κλασσικό διανυσματικό μοντέλο (vector space model). Για την ακρίβεια, η ομοιότητα LSA, υπολογίζεται σε χαμηλότερης διάστασης διάστημα, στο οποίο χρησιμοποιούνται και οι σχέσεις δευτέρου βαθμού μεταξύ των όρων και των κειμένων. Η ομοιότητα στον επιστρεφόμενο χώρο διανυσμάτων (vector space) μετράται με την κλασσική ομοιότητα συνημίτονου (cosine similarity).

Για τις μετρικές που βασίζονται στη γνώση (knowledge-based) υπάρχει ένας μεγάλος αριθμός από αυτές που χρησιμοποιούνται. Εδώ προτείνονται μερικές που δουλεύουν καλά με την ταξονομία του WordNet. Όλες αυτές οι μέθοδοι προϋποθέτουν την είσοδο ενός ζευγαριού εννοιών και επιστρέφουν μια τιμή που ορίζει τη σημασιολογική τους συσχέτιση. Οι μέθοδοι που χρησιμοποιήθηκαν είναι των Leacock & Chodorow, Lesk, Wu & Palmer, Resnik, Lin και τέλος των Jiang & Conrath. Ας σημειωθεί ότι όλες αυτές οι μέθοδοι ορίζονται μεταξύ εννοιών αντί για λέξεων, αλλά μπορούν εύκολα να μετατραπούν σε μετρικές ομοιότητας word-to-word , επιλέγοντας οποιοδήποτε ζευγάρι λέξεων που η σημασία τους να οδηγεί σε concept-to-concept ομοιότητα. Φυσικά σαν Βάση Γνώσης, χρησιμοποιείται το WordNet.

Οι τιμές που επιστρέφονται από τις παραπάνω μετρικές είναι κανονικοποιημένες. Η κανονικοποίηση γίνεται διαιρώντας το σκορ ομοιότητας για κάθε μέθοδο με το μέγιστο πιθανό σκορ που μπορεί να επιστρέψει η μέθοδος αυτή. Τέλος, για τα οκτώ σκορ που επιστράφηκαν από τις οκτώ παραπάνω μεθόδους, corpus-based και knowledge-based, υπολογίζεται ο μέσος όρος, που θα είναι και η τελική τιμή ομοιότητας των δύο κειμένων.

5.3. Text-To-Text Semantic Similarity for Automatic Short Answer Grading - M. Mohler & R. Mihalcea

Η συγκεκριμένη προσέγγιση (Mohler & Mihalcea, 2009), δημιουργήθηκε για να λύσει το πρόβλημα της βαθμολόγησης ενός μικρού κειμένου που έχει δοθεί ως απάντηση σε ένα ερώτημα. Το σύστημα λειτουργεί με σύγκριση της απάντησης σε σχέση με μία ή περισσότερες, πιστοποιημένα σωστές απαντήσεις. Οπότε το πρόβλημα αντιμετωπίζεται ως σημασιολογική σύγκριση των κειμένων και μετρικές σημασιολογικής ομοιότητας τύπου text-to-text.

Χρησιμοποιήθηκαν οχτώ Knowledge-based μετρικές (Shortest Path, Wu & Palmer, Resnik, Leacock & Chodorow, Lesk, Lin, Jiang & Conrath, Hirst & St. Onge) και δύο corpus-based (LSA, ESA).

Για τις Knowledge-based μεθόδους χρησιμοποιείται text-to-text μετρική ομοιότητας όπως προτείνεται και σε προηγούμενη δημοσίευση των συγγραφέων. Δηλαδή, για κάθε λέξη του ενός κειμένου χρησιμοποιείται η μέγιστη σημασιολογική ομοιότητα που μπορεί να επιτευχθεί αν την συγκρίνουμε με τις λέξεις του δευτέρου κειμένου. Όλα τα word-to-word σκορ ομοιότητας που βρίσκονται, προστίθενται και κανονικοποιούνται με το μήκος των δύο κειμένων. Για τη χρήση των Knowledge-Based μετρικών χρησιμοποιείται η ταξονομία του WordNet.

Τα corpus-based μέτρα, που δεν χρειάζονται βάση γνώσης, είναι ιδιαίτερα χρήσιμα στις περιπτώσεις που δεν υπάρχει διαθέσιμη κάποια ταξονομία όπως το WordNet. Για την LSA

χρησιμοποιείται το πακέτο InfoMap και για την ESA χρησιμοποιείται μια υλοποίηση του ESA αλγόριθμου (Gabrilovich & Markovitch, 2006). Και εδώ όλες οι τιμές που επιστρέφονται κανονικοποιούνται μεταξύ των τιμών του 0 και 1. Η κανονικοποίηση γίνεται μέσω της διαίρεσης του σκορ ομοιότητας που επιστρέφεται με το μέγιστο πιθανό σκορ.

Παρακάτω παρέχονται τα αποτελέσματα για κάθε μετρική που εφαρμόστηκε στο ίδιο σετ δεδομένων:

Μετρική	Αποτέλεσμα
Shortest path	0.4413
Leacock & Chodorow	0.2231
Lesk	0.3630
Wu & Palmer	0.3366
Resnik	0.2520
Lin	0.3916
Jiang & Conrath	0.4499
Hirst & St-Onge	0.1961
LSA BNC	0.4071
LSA Wikipedia	0.4286
ESA Wikipedia	0.4681

Μία σημαντική παρατήρηση που έγινε στις Corpus-Based μεθόδους είναι ότι στην LSA όταν το δείγμα ήταν από συγκεκριμένο domain είχαν μεγαλύτερη ακρίβεια σε σχέση με το δείγμα ανοιχτού θέματος, παρά το γεγονός ότι είναι μικρότερο σε μέγεθος. Το ακριβώς αντίθετο έδειχνε να συμβαίνει στην ESA. Αν οι υπόλοιπες συνθήκες είναι ίδιες, τότε τα μεγαλύτερα λεκτικά δείγματα τείνουν να δίνουν μεγαλύτερη ακρίβεια, όπως ήταν αναμενόμενο. Από τα παραπάνω αποτελέσματα προκύπτει ο μέσος όρος τους, ο οποίος ορίζει την τελική τιμή ομοιότητας με μια σωστή απάντηση.

5.4. UNT: A Supervised Synergistic Approach to Semantic Text Similarity - C. Banea, S. Hassan, M. Mohler & R. Mihalcea

Η εργασία των C. Banea, S. Hassan, M. Mohler & R. Mihalcea (Banea, Hassan, Mohler, & Mihalcea), κάνει αναφορά σε τρεις διαφορετικές προσεγγίσεις σημασιολογικής ομοιότητας κειμένων από την ομάδα του Πανεπιστημίου του Βόρειου Τέξας, που παρουσιάστηκαν στο SEMEVAL 2012. Το παρόν σύστημα χρησιμοποιεί τόσο knowledge-based μεθόδους, όσο και corpus-based. Οι προβλέψεις αυτών των ανεξάρτητων συστημάτων, ταιριασμένες με επιπλέον χαρακτηριστικά, χειρίζονται από ένα μετα-σύστημα το οποίο χρησιμοποιεί μηχανική μάθηση (machine learning). Σαν Βάσεις Γνώσης, χρησιμοποιήθηκαν η Wikipedia, η οποία κρίθηκε ιδανική λόγω της δομή της για τις μεθόδους ESA και SSA και το WordNet

Στις Knowledge-based μετρικές, βασισμένοι σε προηγούμενη δουλειά (Mihalcea, Corley, & Strapparava, 2006), (Mohler & Mihalcea, 2009), οι συγγραφείς, χρησιμοποιούν

μετρικές βασισμένες στο WordNet για την εύρεση της ομοιότητας των προτάσεων. Για κάθε λέξη στο ένα από τα 2 κείμενα που εισάγονται υπολογίζεται η μέγιστη σημασιολογική ομοιότητα σε σχέση με κάθε λέξη στο δεύτερο κείμενο. Όλα τα σκορ σημασιολογικής ομοιότητας που επιστρέφονται, προστίθενται και κανονικοποιούνται.

Αν και οι περισσότερες corpus-based μέθοδοι δημιουργούν σημασιολογικά προφίλ λέξεων βάσει της προ-εμφάνισης των λέξεων στο λεκτικό δείγμα, οι LSA, ESA και SSA διαφέρουν, αφού βασίζονται στην αναπαράσταση της έννοιας και όχι της λέξης. Σε αυτές τις μεθόδους, το σημασιολογικό προφίλ της λέξης εκφράζεται με όρους όπως: "υπονοείται"(LSA), "σαφές" (ESA) και "προερχόμενο" (SSA). Η μεταφορά από τον αραιό χώρο λέξεων (word-space) σε έναν πιο πυκνό, πλουσιότερο και με μικρότερη αμφιβολία χώρο εννοιών λύνει ένα από τα βασικά προβλήματα της σημασιολογικής συσχέτισης, που αφορά την αντιστοίχηση των λεξιλογίων.

Αφού εφαρμοστούν οι παραπάνω μέθοδοι, προσπαθούμε να βρούμε το σκορ συσχέτισης σε δύο στάδια.

Στο πρώτο στάδιο παρέχουμε στο σύστημα τους γράφους εξάρτησης για κάθε ζευγάρι προτάσεων. Για κάθε κόμβο σε έναν γράφο εξάρτησης υπολογίζουμε το σκορ ομοιότητας με κάθε κόμβο στον άλλο γράφου εξάρτησης, βασισμένοι σε ένα σετ λεκτικών, σημασιολογικών και συντακτικών χαρακτηριστικών που εφαρμόζονται στο ζευγάρι των κόμβων και στους αντίστοιχους υπο-γράφους. Η συνάρτηση υπολογισμού του σκορ έχει εκπαιδευτεί σε ένα μικρό σετ από χειροκίνητα ευθυγραμμισμένους γράφους, χρησιμοποιώντας τον μέσο perceptron αλγόριθμο. Ορίστηκαν 64 χαρακτηριστικά για το δείγμα εκπαίδευσης του συστήματος μηχανικής μάθησης, από τα οποία τα μισά(32) βασίζονται στην bag-of-words σημασιολογική ομοιότητα του υπο-γράφου και τα υπόλοιπα είναι λεκτικο-συντακτικά χαρακτηριστικά που συσχετίζονται με τους γονείς κόμβους του υπογράφου. Κατόπιν υπολογίζονται με τη χρήση μιας άλλης έκδοσης του αλγόριθμου perceptron (Rosenblatt, 1958), (Freund & Schapire, 1999)τα βάρη που σχετίζονται με τα 64 αυτά χαρακτηριστικά. Αφού οριστούν τα βάρη, ένα σκορ ομοιότητας για κάθε ζευγάρι κόμβων μπορεί να υπολογιστεί.

Στο δεύτερο στάδιο, το σκορ ομοιότητας των κόμβων που υπολογίζεται στο προηγούμενο βήμα, χρησιμοποιείται για να βρεθεί η βέλτιστη ευθυγράμμιση για τα ζευγάρια των συσχετισμένων γράφων. Ξεκινάμε με έναν bipartite γράφο, όπου κάθε κόμβος του ενός γράφου αναπαριστάται από έναν κόμβο στην αριστερή μεριά του bipartite γράφου και κάθε κόμβος του άλλου γράφου, αναπαριστάται στην δεξιά πλευρά. Το βάρος που συσχετίζεται με κάθε ένωση είναι το σκορ που υπολογίστηκε στο προηγούμενο στάδιο. Κατόπιν, ο bipartite γράφος ενισχύεται προσθέτοντας ψεύτικους κόμβους στις δύο μεριές όπου επιτρέπεται να ταιριάξουν με κάθε κόμβο με μηδενικό σκορ. Κατόπιν, η βέλτιστη ευθυγράμμιση μεταξύ των δύο γράφων υπολογίζεται αποδοτικά με τη χρήση του Hungarian algorithm (Hungarian Method). Αφού βρεθεί το βέλτιστο ταίριασμα, παράγονται τέσσερα σκορ βασισμένα στην ευθυγράμμιση, που επιλεκτικά κανονικοποιούνται με τον αριθμό των κόμβων και/ή το βάρος το κόμβο-ευθυγραμμίσεων βάσει του idf σκορ των λέξεων. Αυτό το σκορ καταγράφεται ως $graph_{none}$, $graph_{norm}$, $graph_{idf}$, $graph_{idfnorm}$.

Όλα τα συστήματα που περιγράφτηκαν παραπάνω χρησιμοποιούνται για να παράγουν ένα σκορ για κάθε δείγμα εκπαίδευσης και δοκιμής. Ύστερα, τα σκορ αυτά συναθροίζονται ανά δείγμα και χρησιμοποιούνται σε ένα πλαίσιο επιβλεπόμενης μάθησης. Αποφασίσαμε να χρησιμοποιήσουμε οπισθοδρομικό μοντέλο, αντί ταξινόμηση, αφού οι απαιτήσεις της εργασίας είναι να παραχθεί ένα σκορ μεταξύ 0 και 5. Συγκεκριμένα χρησιμοποιήθηκε support vector regression (Smola & Schoelkopf, 1998).Εξαιτίας της διαφορετικής μεθοδολογίας εκμάθησης και αφού ταιριάζει για να προβλέπει συνεχόμενες κλάσεις, το δεύτερο σύστημα χρησιμοποιεί ένα M5P αλγόριθμο δέντρου απόφασης (Wang & Witten, 1997) , (Landauer & Dumais, 1997).

5.5. Measuring Similarity between Sentences - T. Dao, T. Simpson

Οι T. Dao και T. Simpson (Dao & Simpson, 2005), με τη βοήθεια του WordNet, προσεγγίζουν το πρόβλημα της σημασιολογικής ομοιότητας των προτάσεων σε πέντε βασικά βήματα. Αρχικά χωρίζουν κάθε πρόταση σε λίστες από λεκτικά τμήματα (tokens). Σε αυτό το σημείο αφαιρούν τις STOP-Words ώστε να μειώσουν τον θόρυβο στα μετέπειτα αποτελέσματα όσο το δυνατόν νωρίτερα. Επομένως, κερδίζουν στην απόδοση του αλγορίθμου, αφού δεν χρειάζεται να εφαρμοστούν τα παρακάτω βήματα για λέξεις οι οποίες δεν θα καθορίσουν την ομοιότητα.

Στο δεύτερο βήμα εφαρμόζουν POS tagging για κάθε λέξη,στο τρίτο βήμα εφαρμόζουν τεχνικές stemming με τον αλγόριθμο του Porter και στο τέταρτο βήμα εφαρμόζουν Word Sense Disambiguation με χρήση της επέκτασης του αλγόριθμου του Lesk.

Τέλος, υπολογίζουν την ομοιότητα των προτάσεων βασισμένοι στην ομοιότητα των ζευγαριών των λέξεων μεταξύ των κειμένων. Για να βρεθεί η ομοιότητα δύο προτάσεων, πρέπει να βρεθεί η ομοιότητα μεταξύ των ζευγαριών των διάφορων word-senses που αντιστοιχούν στις δύο αυτές προτάσεις. Για τον ορισμό της ομοιότητας αυτής, χρησιμοποιείται η ομοιότητα της διαδρομής εντός της ταξονομίας του WordNet. Σε αυτό το σημείο μπορούν να εφαρμοστούν πολλές μετρικές όπως των Leacock & Chodorow, του Resnik, του Wu & Palmer, κ.α.

Αφού εφαρμοστούν όλα τα παραπάνω βήματα, για να μετρηθεί η σημασιολογική ομοιότητα μεταξύ δύο προτάσεων X και Y, με m το μήκος της X και n το μήκος της Y, θα πρέπει να δημιουργηθεί μια σχετική μήτρα σημασιολογικής συσχέτισης R[m, n] για κάθε ζευγάρι των word-senses. Αυτή η μήτρα κατόπιν θα μετατραπεί σε bipartite γράφου, όπου X και Y είναι τα δύο σετ των μη ενωμένων κόμβων. Αυτό θα εφαρμοστεί σε κάθε πρόταση και κατόπιν θα πρέπει να βρεθεί η συνολική ομοιότητα μεταξύ όλων των προτάσεων του πρώτου κειμένου, με τις προτάσεις του δευτέρου. Για αυτό το σκοπό προτείνονται δύο μέθοδοι:

Η πρώτη μέθοδος που προτείνεται είναι το ταίριασμα των μέσων όρων (Matching Average):

$$\frac{2 * Match(X, Y)}{|X| + |Y|}$$

, όπου match(X,Y) είναι τα λεκτικά τμήματα (tokens) που ταίριαξαν μεταξύ των X και Y. Αυτή η ομοιότητα υπολογίζεται διαιρώντας το άθροισμα των τιμών ομοιότητας όλων των υποψηφίων που ταίριαξαν από τις δύο προτάσεις με το συνολικό αριθμό των tokens.

Η δεύτερη μέθοδος είναι ο Συντελεστής Ζαριών (Dice Co-Efficient):

$$\frac{2 * |X \cap Y|}{|X| + |Y|}$$

, όπου επιστρέφει τη σχέση του αριθμού των tokens που ταίριαξαν ως προς το σύνολο των tokens.

5.6. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures - D. Bar, C. Blemann, I. Gurevych, T. Zesch

Κατά την πρώτη φάση, το σύστημα συγκρίνει το μήκος των μεγαλύτερων συνεχόμενων ακολουθιών χαρακτήρων ώστε να εντοπίσει ομοιότητες μεταξύ τους δοκιμάζοντας να εισάγει

ή να διαγράψει λέξεις. Όταν απαιτούνται λίγες αλλαγές για να προκύψει το ένα κείμενο μέσα από το άλλο, τα κείμενα τείνουν να είναι όμοια. Κατόπιν, συγκρίνονται τα n-grams των χαρακτήρων που ακολουθούν την υλοποίηση του Barron-Cedeno, γενικεύοντας το αυθεντικό trigram σε n = 2,3,....,15. Επίσης, συγκρίνονται τα n-grams των λέξεων, χρησιμοποιώντας τον συντελεστή Jaccard. Όταν στα n-grams το n τείνει να είναι μεγάλος αριθμός, οδηγεί σε αστάθειες του ταξινομητή, εξ αιτίας της μεταξύ συσχέτισης των λέξεων. Επομένως, μόνο τα n=1,2,3,4 χρησιμοποιήθηκαν.

Τα μέτρα για υπολογισμό της ομοιότητας των λέξεων σε σημασιολογικό επίπεδο, ενεργούν σε μια αναπαράσταση λέξεων εν μέσω γράφου και οι σημασιολογικές συσχετίσεις μεταξύ τους σε μια λεκτική-σημασιολογική πηγή. Για αυτό στο σύστημα χρησιμοποιήθηκαν οι αλγόριθμοι των Jiang & Conrath, Lin και του Resnik με τη χρήση του WordNet, ώστε να δημιουργηθούν οι γράφοι και οι συσχετίσεις. Για να κλιμακωθούν τα αποτελέσματα, εφαρμόστηκε η αθροιστική στρατηγική που προτείνει η Mihalcea (Mihalcea, Corley, & Strapparava, 2006) . Το άθροισμα των idf σταθμισμένων σκορ ομοιότητας για κάθε λέξη με το καλύτερο ταίριασμα που αντιστοιχεί στο άλλο κείμενο υπολογίζεται στις δύο κατευθύνσεις, και βγαίνει ο μέσος όρος. Στα πειράματα η μετρική του Resnik αποδείχθηκε ανώτερη σε σχέση με τις υπόλοιπες και χρησιμοποιήθηκε σε όλες τις ρυθμίσεις ομοιότητας λέξεων. Επίσης, χρησιμοποιήθηκε η μετρική Explicit Semantic Analysis (ESA), η οποία σαν πηγή λεκτικού δείγματος, εκτός από το WordNet χρησιμοποίησε το Wikipedia και το Wiktionary, ώστε να εμπλουτίσει το λεκτικό δείγμα κατά το μέγιστο δυνατό.

Επιπλέον, εφαρμόστηκε σύστημα λεκτικής αντικατάστασης βασισμένο σε επιβλεπόμενη τεχνική Word Sense Disambiguation (WSD) (Biemann, 2012). Το σύστημα αυτό παρέχει αυτόματα αντικαταστάσεις για ένα σετ περίπου 1000 συχνών αγγλικών ουσιαστικών με υψηλή ακρίβεια. Για κάθε ουσιαστικό προστέθηκαν οι αντικαταστάτες στο κείμενο και υπολογίστηκε το pairwise word similarity για τα κείμενα που περιγράφονται παραπάνω. Με αυτόν τον τρόπο εξομαλύνθηκε το λεκτικό κενό (αν υπήρχε) του υποσυνόλου των λέξεων. Ακόμη, χρησιμοποιήθηκε το Moses SMT System (Koehn, Hoang, Birch, Callison-Burch, & Federico, 2007) για να μεταφράσει τα αυθεντικά αγγλικά κείμενα μέσα από μια γέφυρα τριών γλωσσών (Ολλανδικά, Γερμανικά, Ισπανικά) ξανά πίσω στα Αγγλικά. Η ιδέα ήταν ότι στη διαδικασία μετάφρασης από τη μία γλώσσα στην άλλη εισάγονται επιπλέον έννοιες συσχετιζόμενες στην αρχική γραμματοσειρά που καλείται να μεταφραστεί, που εξομαλύνουν με τη σειρά τους, πιθανά λεκτικά κενά. Το σύστημα εκπαιδεύτηκε με τη χρήση του Europarl (Koehn, 2005), που έγινε διαθέσιμο από τον Koehn, χρησιμοποιώντας τις παρακάτω ρυθμίσεις οι οποίες δεν βελτιστοποιήθηκαν για αυτή τη διαδικασία: WMT11 baseline χωρίς tuning, με MGIZA στοίχιση.

Εκτός από τα παραπάνω κοινά μέτρα σημασιολογικής ομοιότητας, δόθηκε βάση και στον υπολογισμό της δομικής ομοιότητας των κειμένων. Μια τεχνική εντοπισμού της δομικής ομοιότητας μεταξύ κειμένων γίνεται υπολογίζοντας n-grams για τα stopwords (Stamatatos, 2011). Όλες οι λέξεις που φέρουν περιεχόμενο αφαιρούνται ενώ οι stopword λέξεις διατηρούνται. Αυτό στηρίζεται στη λογική ότι η συντακτική δομή κειμένων που αναφέρουν το ίδιο πράγμα είναι αντίστοιχη και αλλάζουν οι υπόλοιπες λέξεις με συνώνυμες. Η μέθοδος αυτή προτάθηκε για να εντοπίζει τη λογοκλοπή. Τα stopword n-grams των δύο κειμένων συγκρίνονται χρησιμοποιώντας τη μετρική περιεχομένου του Border (Border, 1997), η οποία ορίζει τιμή αναλόγως του ποσοστού που το ένα κείμενο εμπεριέχεται στο άλλο. Αντίστοιχα με τα stopword n-grams υπολογίζονται και Part-of-speech n grams για διάφορα POS tags τα οποία κατόπιν συγκρίνονται χρησιμοποιώντας μέτρα περιεχομένου και τον συντελεστή Jaccard. Ακόμη, χρησιμοποιήθηκαν δύο μέτρα ομοιότητας μεταξύ ζευγαριών λέξεων (Hatzivassiloglou, 1999). Το Word pair order που δείχνει αν δύο λέξεις είναι στην ίδια σειρά στα δύο κείμενα (με οποιοδήποτε αριθμό λέξεων ανάμεσα) και το word pair distance που μετράει τον αριθμό των λέξεων που βρίσκονται ανάμεσα στο ζευγάρι των λέξεων που εξετάζονται. Για να συγκριθούν τα κείμενα σε σχέση με τη διάσταση του στυλ τους, χρησιμοποιήθηκε ακόμη μια συνάρτηση συχνοτήτων των λέξεων (Dinu & Popescu, 2009), η οποία ενεργεί σε ένα σετ 70 λειτουργικών λέξεων που ορίστηκαν από τον Mosteller και

Wallace (Mosteller & Wallace, 1964). Τα διανύσματα συχνοτήτων των λειτουργικών λέξεων υπολογίζονται και συγκρίνονται από τη συσχέτιση Pearson.

Επομένως, αφού εφαρμόστηκαν αρκετές μετρικές σημασιολογικής και μορφολογικής ομοιότητας μεταξύ των δύο κειμένων, επιστράφηκαν τιμές για κάθε μία από αυτές τις μετρικές, οι οποίες κανονικοποιούνται ώστε να φέρουν μια τελική τιμή που θα κρίνει την ομοιότητα μεταξύ των δύο κειμένων. Αυτή η προσέγγιση είχε τα καλύτερα αποτελέσματα στην SEMEVAL 2012, κάτι που είναι αναμενόμενο, αφού κάνει χρήση μεγάλης ποικιλίας μεθόδων σύγκρισης της ομοιότητας των δύο κειμένων.

6. Προτεινόμενη Προσέγγιση - Notions Oriented Approach (N.O.A.)

6.1. Εισαγωγή και περιγραφή

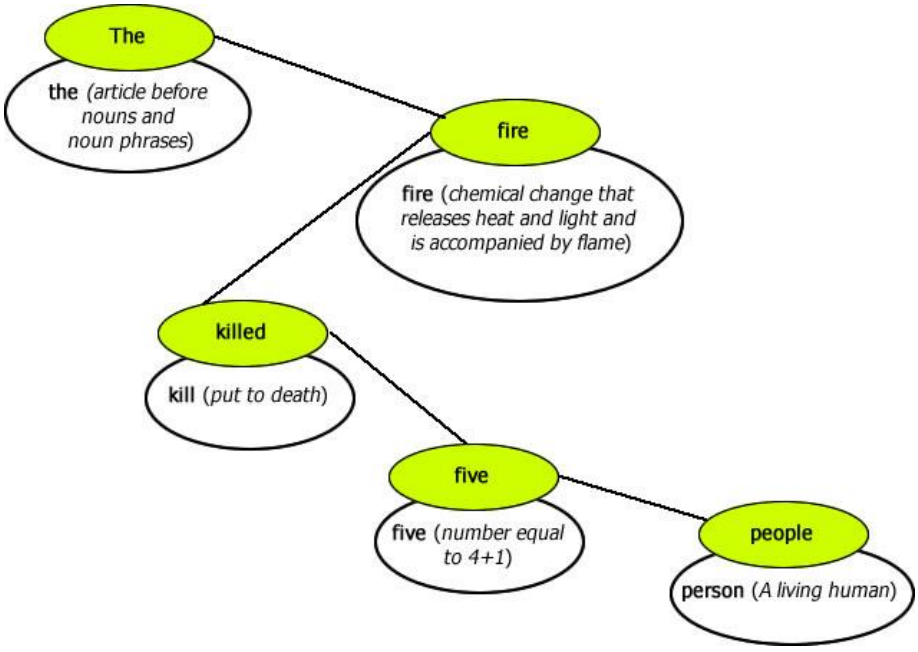
Όπως είναι εμφανές, στις περισσότερες προσεγγίσεις για την εύρεση της σημασιολογικής ομοιότητας, ακολουθείται μια παρόμοια διαδικασία, όπου εξάγεται μία τιμή που ορίζει τον βαθμό ομοιότητας των προτάσεων, βασισμένη στην ομοιότητα των λέξεων και αυτή η διαδικασία επεκτείνεται ώστε να βρεθεί και η ομοιότητα των κειμένων αντίστοιχα. Μια τέτοια προσέγγιση θα μπορούσε να φανεί χρήσιμη σε μια πρόχειρη κατηγοριοποίηση των κειμένων, όπως σε μια διαδικτυακή εφημερίδα. Ωστόσο, αν σκεφτούμε ένα παράδειγμα όπου ένας χρηματιστής, ο οποίος θέλει να ενημερωθεί γρήγορα για τις ειδήσεις που αφορούν τις μετοχές που παρακολουθεί, οι παραπάνω προσεγγίσεις μπορούν μόνο να του κερδίσουν χρόνο δημιουργώντας μια ομαδοποίηση των ειδήσεων, η οποία πιθανώς θα έχει εξαιρέσει ή προσθέσει κάποιες ειδήσεις οι οποίες δεν έχουν απόλυτη συνάφεια. Η αιτία που συμβαίνει αυτό με την πλειοψηφία των υπαρχόντων προσεγγίσεων είναι ότι αν και λαμβάνουν υπόψη την ομοιότητα των λέξεων μεταξύ δύο προτάσεων, δεν λαμβάνουν υπόψη τη συντακτική δομή των προτάσεων, η οποία μπορεί να παίξει καθοριστικό ρόλο στη σημασία του κειμένου. Κάποιες Corpus-Based μετρικές, προσπαθούν με στατιστική ανάλυση ή/και μηχανική μάθηση να δημιουργήσουν πρότυπα για τις ακολουθίες των λέξεων σε μια γλώσσα και τις συσχετίσεις μεταξύ τους. Αυτά τα πρότυπα όμως δεν καλύπτουν την ανάγκη του εξεταζόμενου παραδείγματος. Ο χρηματιστής χρειάζεται μια τεχνική που να κάνει απαλοιφή των επαναλαμβανόμενων ειδήσεων με ακρίβεια, χωρίς να απορρίπτει πιθανή χρήσιμη πληροφορία. Κατά αυτό τον τρόπο, ο χρηματιστής θα μπορεί να διαβάσει το σύνολο των ειδήσεων μόνο μια φορά, χωρίς να σπαταλάει χρόνο για να διαβάζει τα ίδια από διαφορετικές πηγές. Μάλιστα, θα μπορούσε η επαναλαμβανόμενη πληροφορία να χαρακτηριστεί ως πιο έγκυρη.

Μια τέτοια τεχνική, η οποία να διαγράφει την επαναλαμβανόμενη πληροφορία και να συντάσσει ένα νέο κείμενο που θα εμπεριέχει το σύνολο της πληροφορίας από τα κείμενα

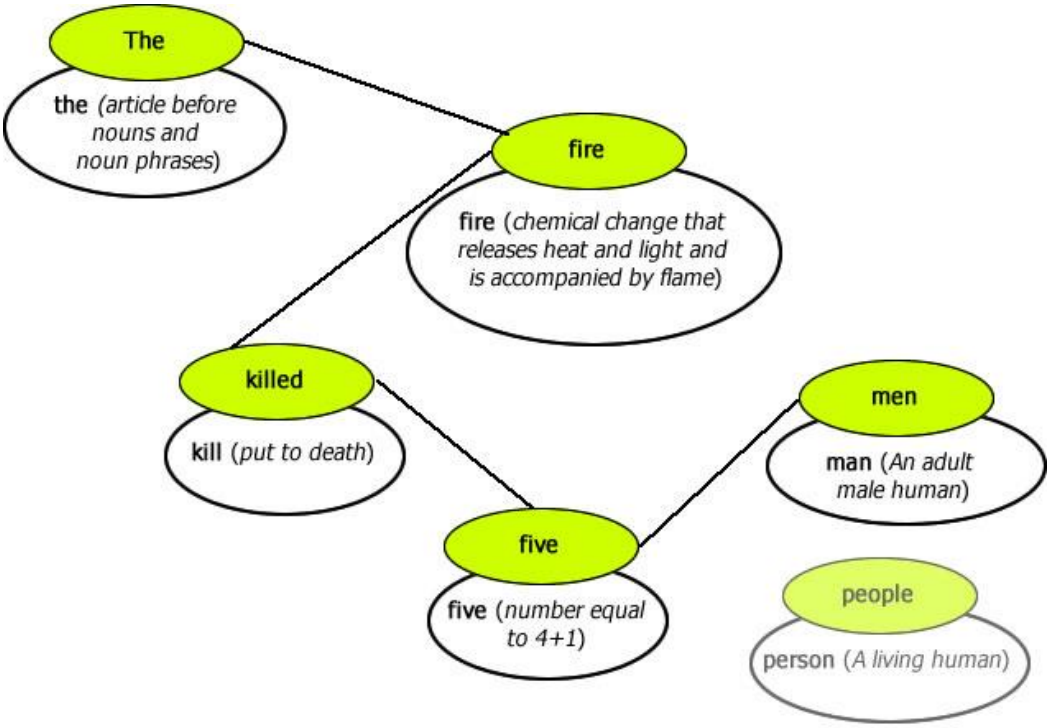
που δόθηκαν σαν είσοδος, θα μπορούσε να έχει πολλές και χρήσιμες εφαρμογές. Το μοντέλο που προτείνεται παρακάτω στρέφει την προσοχή του προς αυτή την κατεύθυνση για την σημασιολογική ομοιότητα των κειμένων, βάσει της επεξεργασίας φυσικής γλώσσας.

Τα θεμέλια του εν λόγω μοντέλου στηρίζονται στην παραδοχή ότι ο άνθρωπος έχει ορίσει την γλώσσα σαν σύμβαση επικοινωνίας και όχι σαν την πραγματική αναπαράσταση της γνώσης. Μια λέξη συμβολίζει ένα βαθύτερο νόημα, το οποίο μπορεί να συμβολίζει και μια άλλη λέξη, είτε στην ίδια γλώσσα (συνώνυμα), είτε σε άλλη. Το νόημα όμως που εμπεριέχει αυτή η λέξη, ή η ακολουθία λέξεων, ορίζεται κοινό σε έναν αντικειμενικό ανθρώπινο εγκέφαλο. Βέβαια, η αντικειμενικότητα παίζει καθοριστικό ρόλο ώστε να αντιληφθούν δύο ή παραπάνω άνθρωποι την έννοια που κρύβει αυτή η λέξη, με τον ίδιο τρόπο, αλλά αυτό είναι μια συζήτηση η οποία ξεφεύγει από τα όρια που θέτουμε για τις ανάγκες της εν λόγω προσέγγισης. Αντίστοιχα, μια έννοια μπορεί να αναπαρασταθεί από μια εικόνα, πιθανώς κάποιον ήχο, μια μυρωδιά, κ.ο.κ. Οπότε, καταλήγουμε στο συμπέρασμα που έκανε και ο Quillian (Collins & Quillian, 1969), και σκεφτόμαστε την αναπαράσταση της γνώσης ως ένα σημασιολογικό δίκτυο.

Μια πρόταση λοιπόν θα μπορούσε να θεωρηθεί ως μια διαδρομή αυτού του σημασιολογικού δικτύου, που διασχίζει όλες τις έννοιες που εμπεριέχει με μια συγκεκριμένη σειρά. Όπως φαίνεται στην παρακάτω εικόνα το σημασιολογικό δίκτυο για την πρόταση "The fire killed five people" διαμορφώνεται ως εξής:

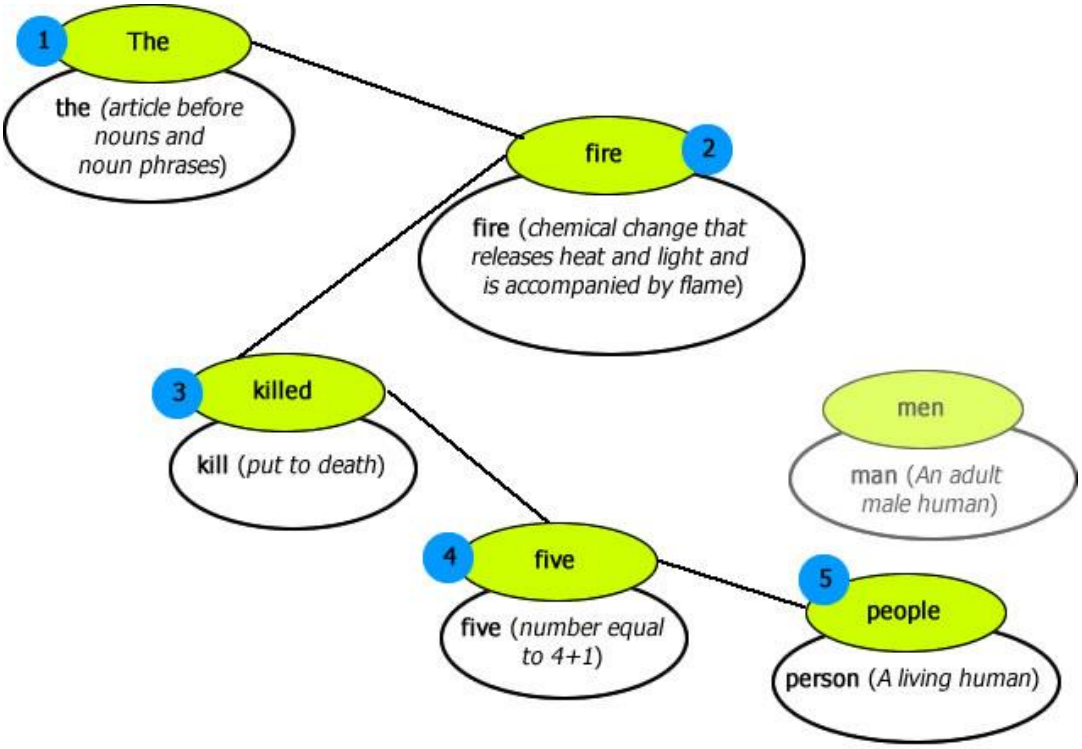


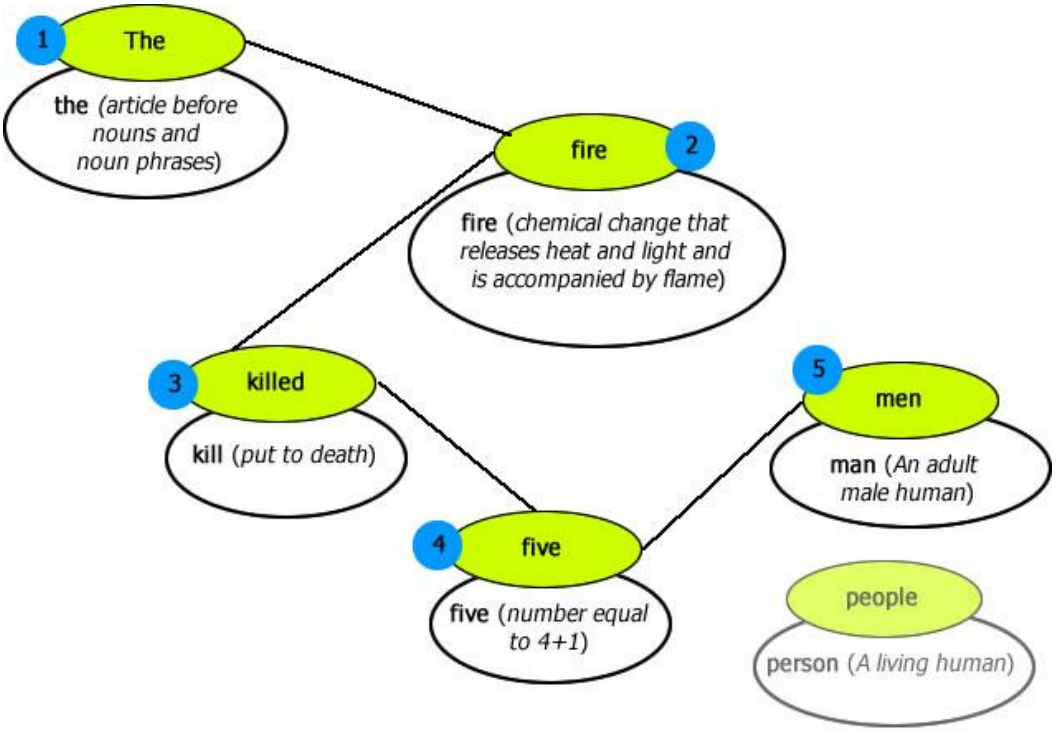
, όπου με πράσινο χρώμα φαίνονται οι λέξεις και στους λευκούς κύκλους αναπαρίσταται η έννοια με λεκτική απεικόνιση για την καλύτερη κατανόηση του παραδείγματος. Όπως φαίνεται, για να γίνει η συσχέτιση της λέξης με την έννοια, πιθανώς να χρειαστούν τεχνικές stemming, με σκοπό να μειωθεί σημαντικά το μέγεθος της Βάσης Γνώσης. Αν λοιπόν αντί της πρότασης "The fire killed five people" είχαμε την πρόταση "The fire killed five men" βλέπουμε ότι η διαδρομή στο σημασιολογικό δίκτυο θα είναι η ίδια με εξαίρεση στην τελευταία έννοια:



Πρακτικά, η δεύτερη πρόταση αν και είναι πανομοιότυπη της πρώτης, προσδίδει περισσότερη πληροφορία, αφού ενημερώνει και για το φύλο των ανθρώπων που πέθαναν από τη φωτιά. Οπότε χρήζει την πρώτη πρόταση περιττή. Οι περισσότερες μετρικές, φυσικά θα μπορούσαν να βρουν μεγάλο βαθμό ομοιότητας μεταξύ των δύο προτάσεων, αλλά θα παρουσιαζόταν έντονο πρόβλημα αν οι προτάσεις ήταν οι "I am a programmer" και "I am not a programmer", όπου με την προσθήκη μίας και μόνο έννοιας αποκτούν αντικρουόμενη σημασία, ενώ όλες οι υπόλοιπες λέξεις εξακολουθούν να χρησιμοποιούν το ίδιο word-sense. Το πρόβλημα αυτό μπορεί να γίνει αρκετά πιο πολύπλοκο, αναλόγως με την πολυπλοκότητα της συντακτικής δομής της πρότασης, της προσθήκης αναφορών σε άλλες προηγούμενες προτάσεις και όταν το κείμενο υπονοεί κάτι χωρίς να το αναφέρει. Αυτά εξακολουθούν να είναι σημαντικά προβλήματα στην επεξεργασία φυσικής γλώσσας.

Θα χρειαστεί λοιπόν μια μέθοδος, η οποία θα σέβεται την ύπαρξη των STOP-Words, καθότι όπως είδαμε μπορούν να αποτελέσουν σημαντικό παράγοντα για το νόημα μιας πρότασης, αλλά κυρίως θα σέβεται τη σειρά με την οποία εμφανίζονται οι λέξεις στην πρόταση. Δηλαδή, δεν αρκεί να δημιουργήσουμε δύο σημασιολογικά δίκτυα (είτε bipartite γράφους, όπως είδαμε να χρησιμοποιούνται στις περισσότερες προσεγγίσεις), αλλά να τηρηθεί και η σειρά από την οποία περνάει η διαδρομή από τους διάφορους κόμβους. Έτσι, στο παράδειγμα των προτάσεων "The fire killed five people" και "The fire killed five men" θα έχουμε τα παρακάτω σημασιολογικά δίκτυα με τις αντίστοιχες διαδρομές:

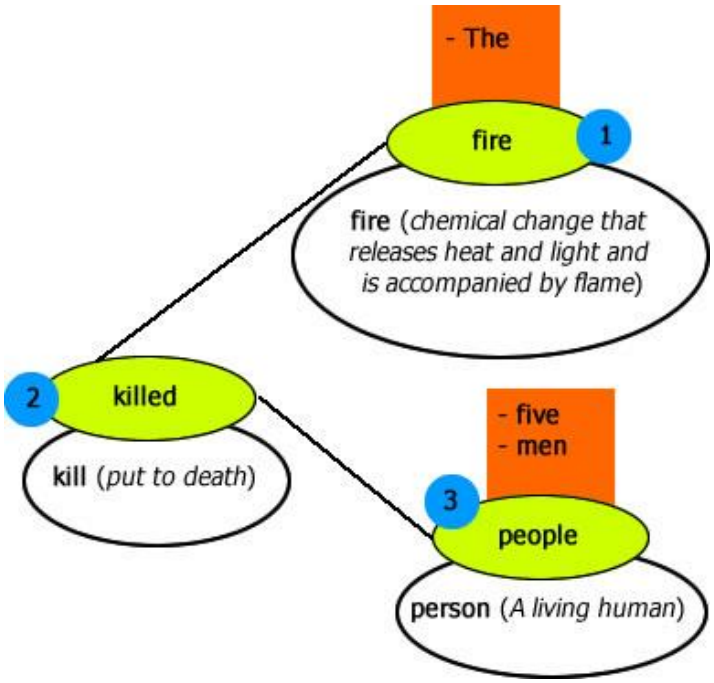




Είναι εμφανές ότι και στις δύο περιπτώσεις ακολουθείται η ίδια διαδρομή με εξαίρεση στην τελευταία έννοια. Στην πρώτη περίπτωση χρησιμοποιείται η έννοια των ανθρώπων και στη δεύτερη των ανδρών. Γνωρίζουμε φυσικά ότι αυτές οι δύο λέξεις σχετίζονται μεταξύ τους. Στην ταξονομία του WordNet το "man" είναι υπονυμία (hyponym) του "person".

Επομένως με μια γενικότερη γνώση από κάποια Βάση Γνώσης, θα μπορούσαμε να φτάσουμε στο συμπέρασμα ότι οι δύο έννοιες που διαφέρουν στο τέλος της διαδρομής έχουν έναν τύπο σύνδεσης μεταξύ τους. Το γεγονός ότι η έννοια "man" ανήκει στην ευρύτερη έννοια "person", και συγκεκριμένα γίνεται χρήση αυτής της ιδιότητας στην εν λόγω πρόταση, θα μπορούσε να δημιουργήσει μια διαδρομή εντός του σημασιολογικού δικτύου, η οποία να είναι ίδια με αυτή που αναπαρίσταται στην πρώτη πρόταση ("The fire killed five people"), όπου το "people" να έχει σαν ιδιότητα το "men". Αυτή η διαδρομή θα μπορούσε να αναπαραστήσει όλη τη χρήσιμη πληροφορία που προκύπτει από τις δύο προτάσεις.

Αντίστοιχα, η έννοια "five" μπορεί να προστεθεί σαν χαρακτηριστικό του "people" στη συγκεκριμένη πρόταση, αφού από το "people" με τους κανόνες Porter Stemming, μπορούμε να προσδιορίσουμε ότι είναι πληθυντικός για το "person", και το "five" προσδιορίζει των αριθμό. Επομένως, με την παραπάνω λογική καταλήγουμε σε μια διαδρομή, η οποία διαπερνά έννοιες κλειδιά εντός του σημασιολογικού δικτύου, οι οποίες με τη σειρά τους περικλείουν χαρακτηριστικά από άλλες έννοιες, μετατρεπόμενες κατά αυτό τον τρόπο σε στιγμιότυπα που χρησιμοποιούνται στην εξεταζόμενη πρόταση. Δημιουργείται επομένως για κάθε κόμβο της διαδρομής εντός του σημασιολογικού δικτύου, ένα υπο-δίκτυο με στιγμιότυπα εννοιών που θυμίζει ένα bag-of-words, όπως φαίνεται παρακάτω:



Η παραπάνω απεικόνιση είναι κοινή και για τις δύο προτάσεις και εμπεριέχει όλη τη χρήσιμη πληροφορία. Επομένως, αν δύο προτάσεις καταλήγουν σε ένα κοινό σημασιολογικό δίκτυο, αντίστοιχο του παραπάνω θα μπορούμε να ορίσουμε ότι οι δύο προτάσεις είναι σημασιολογικά όμοιες, και να κρατήσουμε τα στιγμιότυπα των εννοιών με τις περισσότερες ιδιότητες, ώστε να παραχθεί μια τρίτη πρόταση που θα εμπεριέχει όλο το νόημα. Στην παρούσα προσέγγιση, το ζητούμενο είναι να υλοποιηθεί μια βηματολογία/μέθοδος για την δημιουργία αυτού του σημασιολογικού δικτύου αναπαράστασης των προτάσεων.

Αρχικά θα πρέπει να εφαρμοστούν τεχνικές Sentence Boundary Disambiguation, για να εξορυχτούν οι προτάσεις που εμπεριέχονται στο κείμενο, όπως γίνεται σε κάθε προσέγγιση εύρεση της σημασιολογικής ομοιότητας των κειμένων. Αφού βρεθούν οι προτάσεις, κάθε μία από αυτές θα πρέπει να μετατραπεί σε μια διαδρομή εντός του σημασιολογικού δικτύου των εννοιών. Επομένως, το δεύτερο βήμα είναι η διαδικασία tokenization ώστε να βρεθούν οι λεκτικές ακολουθίες που απαρτίζουν την πρόταση.

Τα tokens που βρέθηκαν στο προηγούμενο βήμα μπορεί να αναπαριστούν, ή να συμμετέχουν στην αναπαράσταση μίας ή περισσότερων εννοιών. Θα πρέπει λοιπόν για τα λεκτικά τμήματα που συνδέονται στη Βάση Γνώσης με περισσότερες από μία έννοιες να αποσαφηνιστεί ποια έννοια αναπαριστούν στην παρούσα πρόταση.

Εδώ, πρέπει να δοθεί ιδιαίτερη σημασία στο ότι αναφερθήκαμε σε έννοιες και όχι σε word-senses, όπως οι συμβατές τεχνικές. Αυτή είναι μια βασική διαφοροποίηση της παρούσας προσέγγισης. Η έννοια έχει κάποιες βασικές διαφορές από το word-sense, όπως το ότι είναι πιο αφηρημένη και μπορεί να είναι η αντιστοίχηση όχι μόνο μιας λέξης, αλλά και μιας ακολουθίας λέξεων, μιας εικόνας, ενός ήχου, κ.ο.κ.. Ένα ακόμη προτέρημα της χρήσης της έννοιας, έναντι του word-sense είναι ότι λέξεις διαφορετικών γλωσσών μπορεί να καταλήγουν στην ίδια έννοια. Αντίθετα τα word-sense είναι αλληλένδετα με τη γλώσσα που χρησιμοποιεί η ταξονομία και για άλλη γλώσσα θα χρειαστεί άλλη Βάση Γνώσης. Επομένως, με τη χρήση των εννοιών η Βάση Γνώσης γίνεται πιο ευέλικτη.

Για να μπορέσουμε λοιπόν να κάνουμε χρήση εννοιών, κατασκευάστηκε μια ολοκαίνουργια Βάση Γνώσης αντί να χρησιμοποιηθεί το WordNet, όπως σχεδόν έχει καθιερωθεί στις αντίστοιχες τεχνικές. Εκτός από την ανάγκη της χρήσης των εννοιών στην παρούσα τεχνική, υπήρξαν και άλλοι περιορισμοί στο WordNet οι οποίοι δημιουργούσαν προβλήματα στην υλοποίηση της νέας προσέγγισης. Ένας βασικός περιορισμός είναι ότι το WordNet δεν έχει "πολλά-προς-πολλά" σχέσεις μεταξύ των λέξεων και των word-sense. Αυτό δημιουργεί επιπλέον επεξεργαστικό κόστος στον υπολογιστή και ακόμη και προβλήματα, όταν καλούμαστε να δημιουργήσουμε τη διαδρομή στο γράφο των εννοιών, που απαιτείται γι αυτή την προσέγγιση. Τέλος, η νέα Βάση Γνώσης δεν περιορίζει τον αριθμό των σχέσεων μεταξύ των εννοιών. Το WordNet έχει ένα πεπερασμένο αριθμό σχέσεων όπως hyponym, hypernym, meronym, κ.ο.κ.. Η νέα Βάση Γνώσης, δημιουργήθηκε με σκοπό να υπερπηδήσει αυτό το πρόβλημα και να αφήσει ανοιχτό τον αριθμό των τύπων των σχέσεων, ώστε βάσει αυτών να μπορούν να δημιουργηθούν νέοι γραμματικοί και συντακτικοί κανόνες.

Για την αποσαφήνιση λοιπόν των εννοιών (notions), στα tokens που έχουν παραπάνω από μία χρησιμοποιήθηκαν δύο κατηγορίες μεθόδων. Η πρώτη αναζητά ακολουθίες tokens και προσπαθεί να τις ταιριάξει με ακολουθίες λέξεων που αναπαριστούν μια έννοια στη Βάση Γνώσης, με ιδιαίτερη προσοχή στη σειρά που εμφανίζονται στο κείμενο και να μην έχουν άλλες λέξεις ανάμεσά τους. Αν το ταίριασμα γίνει τότε το σύστημα αντιλαμβάνεται ότι εντόπισε ακολουθία και βρίσκει την αντίστοιχη έννοια για τα tokens που την αναπαριστούν. Αν για παράδειγμα έχω την ακολουθία "capital city of Greece" θα την αντιστοιχίσει με την έννοια "Athens". Ο αλγόριθμος είναι βασισμένος στα Hidden Markov Models και σε n-grams (Hidden Markov Models-HMM), τα οποία εξετάζουν τις πιθανότητες να ακολουθεί μια τιμή μετά από τις n προηγούμενες.

Αφού εφαρμόσει τις τεχνικές αναζήτησης ακολουθιών λέξεων, εφαρμόζει τεχνικές γραμματικών και συντακτικών κανόνων με σκοπό να αποσαφηνίσει ακόμα περισσότερο. Οι κανόνες αυτοί είναι καταγεγραμμένοι στη Βάση Γνώσης και αφορούν όχι μόνο τις έννοιες που καλούνται να ελέγξουν, αλλά και τις έννοιες με τις οποίες συνδέονται εντός της Βάσης Γνώσης. Δημιουργείται ένα Bag-Of-Words ή για την ακρίβεια ένα Bag-Of-Notions, στο οποίο οι σχέσεις κάθε έννοιας με την βασική ορίζονται σαφώς. Για να γίνει πιο κατανοητό ας υποθέσουμε ότι εξετάζουμε τη λέξη "fire" που έχει δύο έννοιες: "φωτιά" και "πυροβολώ". Ο κανόνας μπορεί να λέει ότι αν βρεθεί άρθρο πριν από τη λέξη που εξετάζουμε είναι ουσιαστικό ή επίθετο. Επομένως, αν στο κείμενο είχαμε τη φράση "The fire... " θα δημιουργήσει ένα Bag-Of-Notions για το "The", όπου μέσα σε αυτά θα βρίσκεται η έννοια του άρθρου με μια σχέση IS-A. Αυτό θα βοηθήσει να ελεγχθεί ο παραπάνω κανόνας και να αποκλειστεί το ενδεχόμενο του ρήματος, το οποίο έχει οριστεί στο Bag-Of-Notions της λέξης "fire" που εξετάζεται, επίσης με σχέση IS-A.

Οι παραπάνω κανόνες δεν είναι οι μόνοι που μπορούν να εφαρμοστούν για αποσαφήνιση της έννοιας των tokens που αντιστοιχούν σε παραπάνω από μία. Μπορούν να

οριστούν επιπλέον τύποι κανόνων. Κάτι τέτοιο όμως, απαιτεί αντικείμενο μελέτης για γλωσσολόγους και ξεφεύγει από τα όρια ορισμού της εναλλακτικής προσέγγισης. Στην παρούσα εργασία κρίθηκαν αρκετοί οι δύο παραπάνω τύποι. Αφού εφαρμοστούν οι δύο παραπάνω τύποι κανόνων, αν επέφεραν αλλαγές και αποσαφήνισαν την έννοια για κάποιο token, επανεκτελούνται. Ο λόγος είναι ότι εφόσον υπάρχουν αλλαγές, υπάρχει και περισσότερη πληροφορία για να μπορέσουν να εφαρμοστούν κι άλλοι κανόνες από το πρώτο σετ κανόνων και κατόπιν, πιθανώς και από το δεύτερο.

Το σύστημα είναι σε πιλοτική έκδοση και οι κανόνες είναι αρκετά περιορισμένοι. Οπότε υπάρχει ενδεχόμενο όταν τελειώσει με την εφαρμογή τους, να εξακολουθούν να υπάρχουν λέξεις tokens με περισσότερες από μία πιθανές έννοιες. Το σύστημα επομένως, ρωτάει το χρήστη να αποσαφηνίσει αυτές τις περιπτώσεις. Όσο πιο πολύ εξελίσσεται το σύστημα θα προστίθενται περισσότεροι κανόνες και η περίπτωση το σύστημα να μην μπορεί να αποσαφηνίσει την έννοια θα είναι όλο και πιο σπάνια.

Στο επόμενο βήμα το σύστημα έχει δημιουργήσει μια ακολουθία εννοιών, όπου αναπαριστά την κάθε πρόταση και στα δύο κείμενα που εξετάζονται. Το πιθανότερο είναι ότι οι δύο προτάσεις, ακόμη και αν είναι όμοιες, να μην έχουν την ίδια ακολουθία. Για παράδειγμα οι προτάσεις "The fire killed five people" και "Five people died by the fire" ενώ λένε το ίδιο ακριβώς πράγμα, έχουν εντελώς διαφορετική ακολουθία εννοιών. Σε αυτό το σημείο με χρήση κανόνων από τη Βάση Γνώσης, εφαρμόζουμε τεχνικές κανονικοποίησης των προτάσεων/φράσεων. Για παράδειγμα, ένας κανόνας που θα εφαρμοστεί στην παραπάνω πρόταση είναι ο παρακάτω: "Ουσιαστικό Α-Ρήμα σε Αόριστο-by-Ουσιαστικό Β" → "Ουσιαστικό Β -Ρήμα σε Αόριστο- Ουσιαστικό Α". Βέβαια, πριν εφαρμοστεί, θα πρέπει να έχουν εφαρμοστεί τεχνικές σύνδεσης λέξεων, όπως τα άρθρο με τα ουσιαστικά που ακολουθούν ή ο αριθμός με ουσιαστικό που ακολουθεί αμέσως μετά ως προσδιοριστικό, κ.ο.κ.. Αυτές οι τεχνικές βασίζονται επίσης στους κανόνες κανονικοποίησης. Αντίστοιχα, πρέπει να οριστούν κανόνες για την κανονικοποίηση περισσότερων περιπτώσεων φράσεων. Αυτό φυσικά είναι μια επίπονη και αρκετά δύσκολη δουλειά που πρέπει να γίνει από ειδικούς επιστήμονες της γλωσσολογίας.

Αν, λοιπόν δύο προτάσεις είναι όμοιες, με τους κανόνες κανονικοποήσης που εφαρμόστηκαν, θα καταλήξουν στην ίδια ακολουθία εννοιών. Επομένως, κατά αυτό τον τρόπο μπορούμε να δούμε τις φράσεις που επαναλαμβάνονται σε δύο κείμενα και να αφαιρέσουμε την επαναλαμβανόμενη πληροφορία από αυτά.

6.2. Notions-Based Βάση Γνώσης αντί WordNet

Αν και στην πλειοψηφία των knowledge-based προσεγγίσεων της εύρεσης της σημασιολογικής ομοιότητας δύο κειμένων χρησιμοποιείται είτε η ταξονομία του WordNet, είτε σε μερικές περιπτώσεις η Wikipedia, στην παρούσα προσέγγιση προτιμήθηκε να δημιουργηθεί ένα νέο σημασιολογικό δίκτυο, το οποίο θα ταιριάζει ακριβώς στις ανάγκες της προσέγγισης αυτής.

Το σημασιολογικό δίκτυο που δημιουργείται πάει ένα βήμα παραπέρα από την αντιστοίχιση λέξεων με word-senses που χρησιμοποιεί το WordNet και εισάγει την έννοια των "notions". Όπου "notion", ορίζεται η έννοια που αναπαριστά μια λέξη, είτε αυτούσια, είτε η λέξη αυτή συμμετέχει στην αναπαράσταση μιας έννοιας σε συνδυασμό με άλλες λέξεις. Επομένως, ενώ στο Word-Net έχουμε μια σχέση ένα προς πολλά, αφού μια λέξη μπορεί να έχει πολλά word-senses, ενώ ένα word-sense δεν μπορεί να αντιστοιχεί σε πολλές λέξεις, η παρούσα Βάση Γνώσης δημιουργεί σχέσεις πολλά προς πολλά μεταξύ των λέξεων και των εννοιών (notions).

Ένα παράδειγμα για να γίνει πιο κατανοητό το παραπάνω είναι οι λέξεις "beautiful" και "pretty" οι οποίες στην πραγματικότητα έχουν την ίδια έννοια. Ωστόσο, στο WordNet καταλήγουν σε διαφορετικά synsets (word-senses) τα οποία έχουν μεταξύ τους μια σχέση

similar (300218842(beautiful) - 300221022(pretty)). Γίνεται λοιπόν σαφής η σκοπιμότητα και η χρηστικότητα να εισαχθεί η έννοια της λέξης (notion) στη Βάση Γνώσης και να συσχετιστεί με τις αντίστοιχες λέξεις, καθώς το word-sense κρίνεται περιορισμένο για τις ανάγκες της εν λόγω προσέγγισης.

Όπως αναφέρθηκε και παραπάνω, η έννοια μπορεί να έχει αντιστοίχιση όχι μόνο με λέξεις από το λεξικό μίας και μόνο γλώσσας, αλλά με λέξεις από οποιοδήποτε λεξικό. Αυτό δίνει τη δυνατότητα στην τεχνική να χρησιμοποιεί κοινή Βάση Γνώσης ανεξάρτητα από τη γλώσσα στην οποία είναι γραμμένο το κείμενο που εξετάζεται. Επίσης, δίνεται ευελιξία, στην περίπτωση που ενώ το σύνολο του κειμένου είναι σε μία γλώσσα, υπάρχουν εντός του κειμένου λέξεις από άλλη γλώσσα, κάτι που είναι ιδιαίτερα σύνηθες τη σύγχρονη εποχή.

Επιπλέον, η δόμηση της Βάσης Γνώσης, επιτρέπει στο σύστημα την εισαγωγή και αντιστοίχηση με τις έννοιες, όχι μόνο λέξεων, αλλά και οποιουδήποτε στοιχείου, όπως εικόνα, ήχο, κ.α. Αυτό ανοίγει δρόμους στη χρήση του συστήματος σε μελλοντικές επεκτάσεις.

Επίσης, η νέα Βάση Γνώσης χειρίζεται τις ιδιότητες των εννοιών ως ανεξάρτητες έννοιες και κάνει αντιστοίχηση με έναν δυναμικό τρόπο σχέσεων, όπου ο διαχειριστής του συστήματος δεν περιορίζεται. Για παράδειγμα, η λέξη "car", αντιστοιχεί στην έννοια "car - an automobile", η οποία με τη σειρά της έχει σχέσεις τύπου IS-A με τις έννοιες "noun - grammar class", "singular - grammar used notion to define a singular number", κ.ο.κ.. Φυσικά, οι σχέσεις των εννοιών δεν είναι μόνο IS-A, αλλά δυναμικά οριζόμενες.

Η Βάση Γνώσης που δημιουργήθηκε, φεύγει ένα βήμα παραπέρα από την απλή αναγνώριση εννοιών και δίνει τη δυνατότητα ορισμού συντακτικών και γραμματικών κανόνων με μορφή n-grams (Hidden Markov Models-HMM), που χρησιμοποιούνται κατά τη διαδικασία αποσαφήνισης των εννοιών. Ο διαχειριστής μπορεί να ορίσει πιθανότητες για κάποια έννοια βασισμένος στην ακολουθία των εννοιών που βρίσκονται στα αριστερά ή στα δεξιά της εξεταζόμενης λέξης. Επομένως, αν εντοπιστεί εντός του κειμένου μια ακολουθία εννοιών που επιβεβαιώνει κάποιον κανόνα στη Βάση Γνώσης, στην περίπτωση που κάποια λέξη εντός της ακολουθίας χρήζει αποσαφήνισης ως προς την αντίστοιχη έννοια, η ακολουθία εννοιών που ταιριάζει με τον κανόνα αυτό μπορεί να βοηθήσει στην αποσαφήνιση αυτή. Στο παρόν μοντέλο αυτό ισχύει μόνο για την περίπτωση όπου μόνο μία λέξη της ακολουθίας δεν έχει αντιστοιχιστεί σε έννοια (notion), αλλά σε μελλοντικές επεκτάσεις του συστήματος, μπορούν να δημιουργηθούν τύποι κανόνων για ακόμα πιο πολύπλοκες περιπτώσεις.

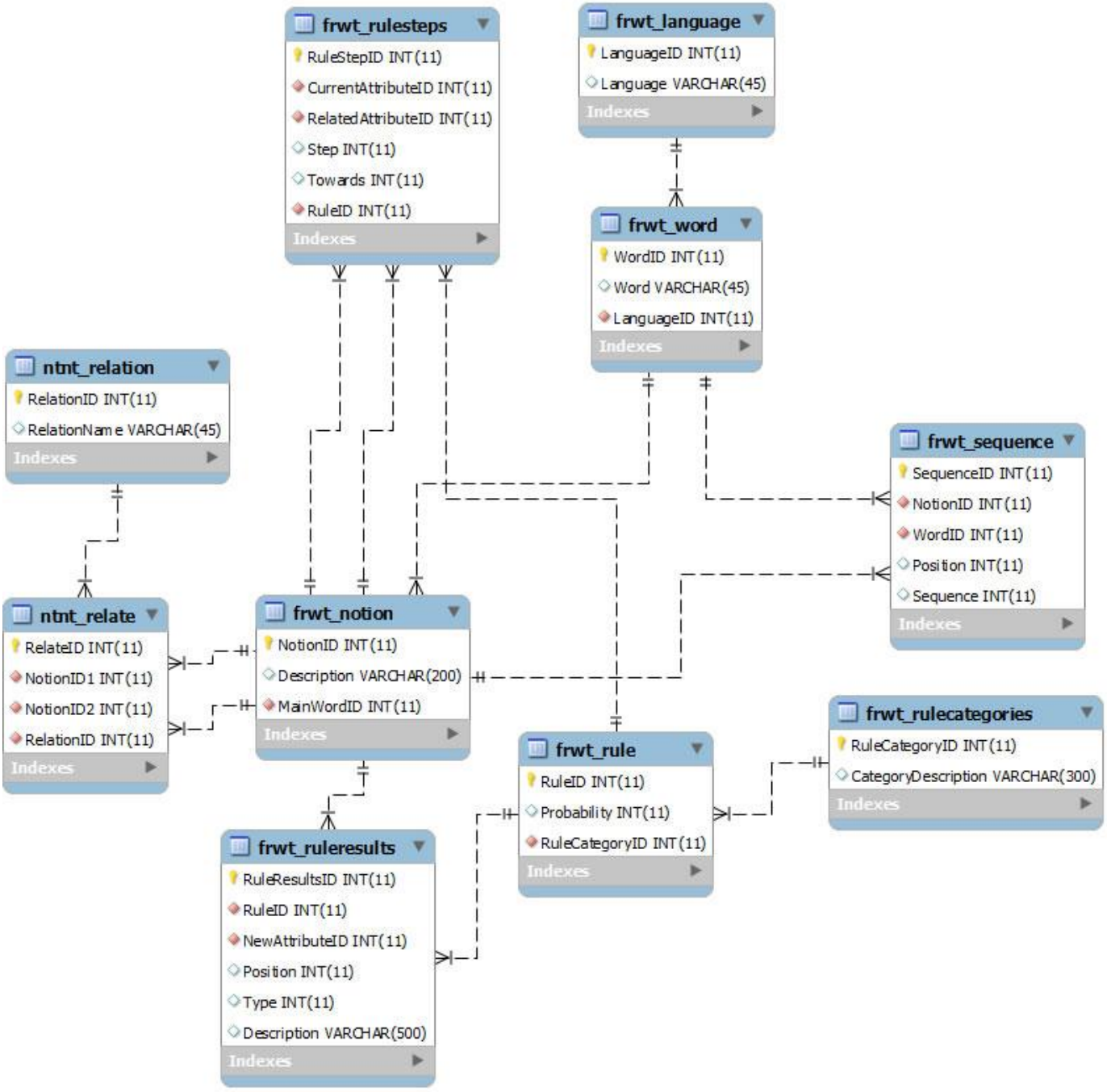
Τέλος, η Βάση Γνώσης έχει κανόνες για κανονικοποίηση των ακολουθιών που θα βρεθούν στα παραπάνω βήματα με αποτέλεσμα να καταλήξουν σε μια τελική κοινή μορφή οι ακολουθίες δύο διαφορετικών εξεταζόμενων ως προς την ομοιότητα κειμένων, στην περίπτωση που είναι όντως όμοια ως προς τη σημασιολογία τους. Αυτοί οι κανόνες αποτελούνται από δύο βασικά βήματα. Στο πρώτο βήμα εντοπίζονται αντίστοιχα με την τεχνική του Notion Disambiguation n-gram models από ακολουθίες εννοιών ή ιδιοτήτων τους. Στο δεύτερο βήμα γίνεται αντικατάσταση των ακολουθιών που εντοπίστηκαν από το πρώτο με την κανονικοποιημένη τους μορφή. Για παράδειγμα στην πρόταση "...people died by the fire" η ακολουθία "...the fire" θα αντικατασταθεί από το "fire" στο οποίο θα γίνει bind το "the" καθότι εντοπίστηκε κανόνας που ορίζει ότι όταν βρεθεί ακολουθία άρθρου και ουσιαστικού, το άρθρο θα πρέπει να γίνει bind στο ουσιαστικό. Επομένως η πρόταση διαμορφώνεται ως "... people died by $fire_{(the)}$". Κατόπιν, επανεξετάζεται το κείμενο για περαιτέρω εφαρμογή κανόνων κανονικοποίησης και εντοπίζεται ο κανόνας που ορίζει ότι η ακολουθία "ουσιαστικό Α - died - by - ουσιαστικό Β" μετατρέπεται σε "ουσιαστικό Β - killed - ουσιαστικό Α". Επομένως το κείμενο διαμορφώνεται ως "... $fire_{(the)}$ killed people".

Οι κανόνες αυτοί είναι σημαντικό να δομηθούν ώστε να μην επικαλύπτουν ο ένας τον άλλον, αλλά παράλληλα να είναι σαφώς ορισμένοι. Γι αυτό το λόγο κρίνεται σκόπιμο, η δόμηση των κανόνων ως n-gram Hidden Markov Models να δημιουργηθεί από επιστήμονες

της γλωσσολογίας. Οι κανόνες που έχουν ενταχθεί στο παρόν σύστημα είναι πολύ απλοϊκοί και εξυπηρετούν στο να κατανοήσει ο αναγνώστης τη λειτουργία της προτεινόμενης τεχνικής.

6.3. Σχήμα Βάσης Δεδομένων

Στο παρόν σημείο κρίνεται σκόπιμο να δοθεί μια περιληπτική επεξήγηση του σχήματος της Βάσης Δεδομένων που δημιουργήθηκε για να εξυπηρετήσει το παρόν σύστημα.



Οι πίνακες καταγράφονται με ένα πρόθεμα το οποία δηλώνει το τμήμα της εφαρμογής που εξυπηρετούν (FRW - Framework, NTN - Notion), ενώ ακολουθεί ένα "t" που υποδηλώνει ότι είναι πίνακας. Αντίστοιχα στις stored procedures και στις ρουτίνες υπάρχει ένα "f" μετά το πρόθεμα.

Ο πίνακας **FRWt_Word**, περιέχει όλες τις λέξεις που υπάρχουν στη Βάση Γνώσης του συστήματος, καθώς και ορίζει σε ποια γλώσσα αντιστοιχούν. Ο πίνακας **FRWt_Notion**

περιέχει όλες τις έννοιες που υπάρχουν στη Βάση Γνώσης. Η αντιστοιχία μεταξύ των εννοιών και των λέξεων γίνεται μέσω του πίνακα **FRWt_Sequence**. Όταν λοιπόν μια ακολουθία λέξεων αντιστοιχεί σε μία έννοια, θα υπάρχουν δύο η παραπάνω εγγραφές (ανάλογα με τον αριθμό των λέξεων της ακολουθίας) στον πίνακα **FRWt_Sequence**, όπου το πεδίο *Sequence* θα τις ομαδοποιεί στην ίδια ακολουθία και το πεδίο *Position* θα υποδεικνύει τη σειρά με την οποία εμφανίζονται στην ακολουθία. Οι σχέσεις μεταξύ των εννοιών δηλώνονται στον πίνακα **NTNt_Relate**, όπου ο τύπος σχέσης στο πεδίο *RelationID* βλέπει στον πίνακα **NTNt_Relation** που περιέχει όλους τους πιθανούς τύπους σχέσεων που μπορεί να συναντήσουμε. Η δομή της Βάσης Δεδομένων επιτρέπει στον διαχειριστή να προσθέσει εύκολα επιπλέον τύπους σχέσεων μεταξύ των εννοιών.

Οι κανόνες βρίσκονται τόσο για το Notion Disambiguation όσο και για την κανονικοποίηση, βρίσκονται στον πίνακα **FRWt_Rule**, όπου ο τύπος τους (*RuleCategoryID*) ορίζει σε ποιο βήμα θα χρησιμοποιηθούν. Οι ακολουθίες των εννοιών για να εντοπιστούν οι κανόνες βρίσκονται σε χωριστό πίνακα, στον **FRWt_RuleSteps**, όπου το πεδίο *RuleID* δείχνει σε ποιον κανόνα αναφέρεται το κάθε βήμα, η στήλη *Step* ορίζει τη σειρά εμφάνισης στην ακολουθία και το πεδίο *Towards* τη φορά (αριστερή ή δεξιά). Όταν πρόκειται για κανόνες κανονικοποιήσης, όπου θα γίνει αντικατάσταση της ακολουθίας με κάτι άλλο, για την αντικατάσταση αυτή, γίνεται αναζήτηση στον πίνακα **FRWt_RuleResults**, όπου και πάλι γίνεται αναφορά στον κανόνα που απευθύνονται από το πεδίο *RuleID*, το πεδίο *Position* δηλώνει την έννοια που εξετάζεται, ενώ το πεδίο *NewAttributeID* ορίζει είτε το νέο notion που θα μπει στη θέση της εξεταζόμενης έννοιας, είτε το σημείο που θα βρεθεί. Επειδή υπάρχουν και περιπτώσεις όπου η νέα έννοια δεν θα μπει 'καρφωτά' αλλά δυναμικά, βάσει αλλαγής της σειράς των ήδη εμφανιζόμενων εννοιών, το πεδίο *Type* ορίζει αν συμβαίνει κάτι τέτοιο. Στην περίπτωση που έχει την τιμή 1 το πεδίο *NewAttributeID*, παραπέμπει στο ID του notion που θα πάρει τη θέση του τρέχων, ενώ αν έχει την τιμή 2, το πεδίο *NewAttributeID* αναφέρεται στη θέση της ακολουθίας που υπάρχει το notion που θα κάνει την αντικατάσταση.

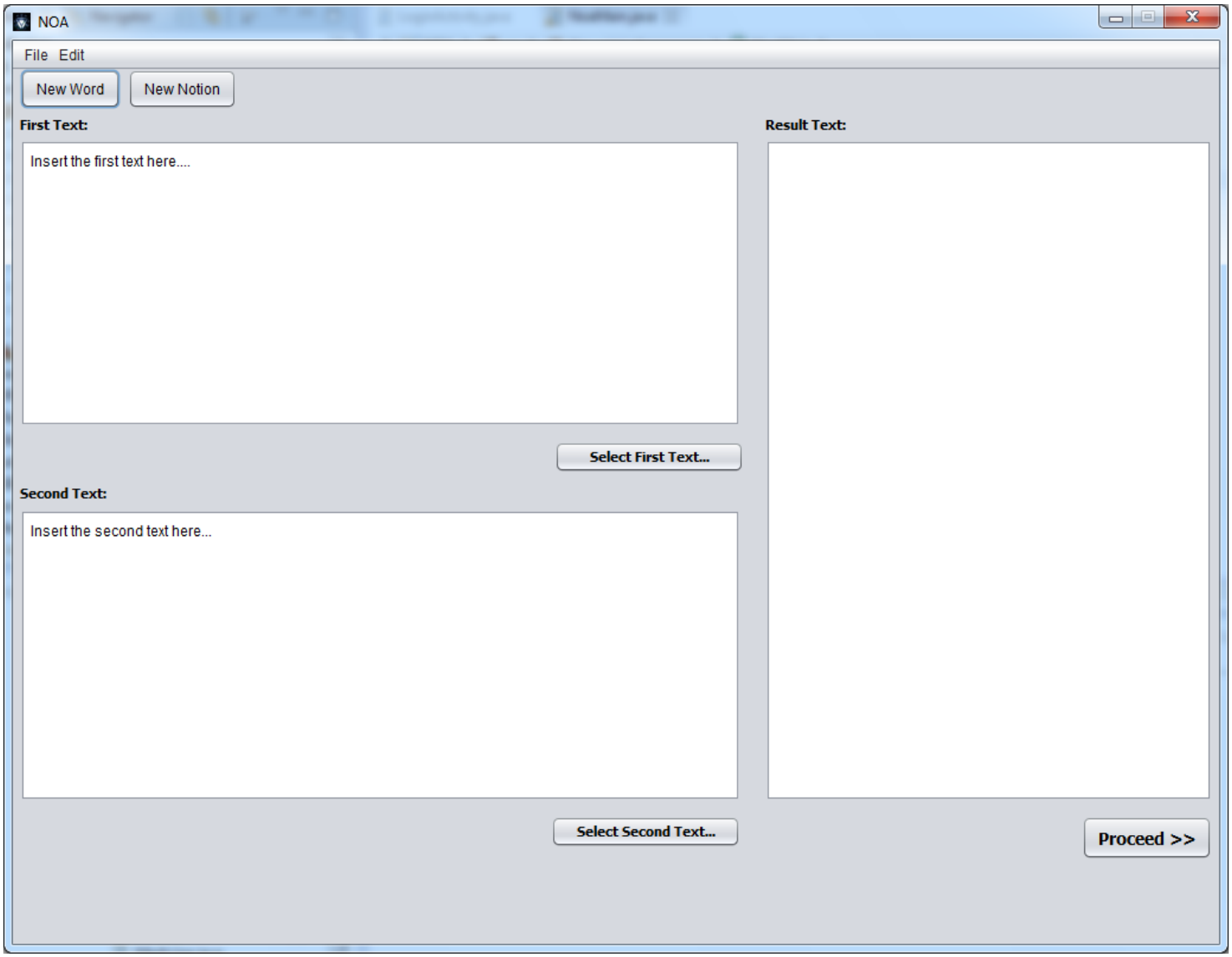
6.4. Υλοποίηση

Με σκοπό να δοκιμαστεί το παραπάνω μοντέλο έγινε μια υλοποίηση του NOA με τη χρήση της Java S.E. 7 και της MySQL για τη Βάση Γνώσης. Η υλοποίηση πραγματεύεται μια εφαρμογή που έγινε για σκοπούς δοκιμών επιτυχίας του μοντέλου και δέχεται δύο κείμενα ως είσοδο, τα επεξεργάζεται και εξάγει την τελική, κανονικοποιημένη ακολουθία εννοιών για κάθε κείμενο. Επομένως, αν οι ακολουθίες που εξάγονται από τα δύο κείμενα είναι ίδιες, συμπεραίνεται ότι και τα δύο κείμενα είναι σημασιολογικά ίδια. Η παρούσα υλοποίηση έχει δημιουργηθεί με σκοπό να εφαρμόσει το μοντέλο NOA σε απλές προτάσεις. Ωστόσο, η σχεδίαση υποστηρίζει τη δυνατότητα ανάπτυξης της εφαρμογής ώστε να υποστηρίξει ακόμα πιο μεγάλα και πολύπλοκα κείμενα σε επόμενες εκδόσεις.

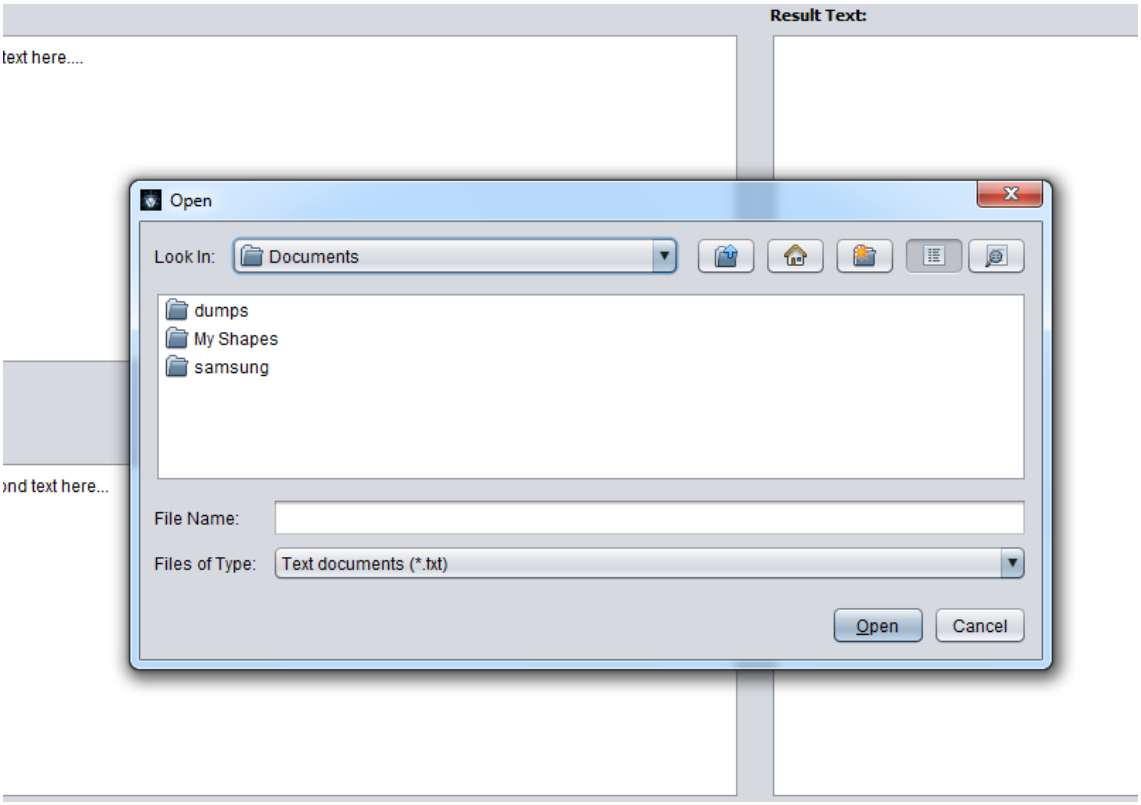
Η Βάση Δεδομένων της εφαρμογής έχει εγγραφές που απαιτούνται μόνο για τα τρία ζεύγη λεκτικών προτάσεων που χρησιμοποιήθηκαν ως δοκιμή της εφαρμογής. Μια πλήρης Βάση Γνώσης, για να μπορεί να τρέξει το μοντέλο σε οποιαδήποτε πρόταση, θα πρέπει να έχει όλες τις λέξεις για τη γλώσσα επί της οποίας θα τρέξει το μοντέλο και τις αντίστοιχες έννοιες. Επίσης, θα πρέπει να έχει όλους τους γραμματικούς και συντακτικούς κανόνες που θα χρειαστούν τόσο για το Notion Disambiguation, όσο και για την κανονικοποίηση των προτάσεων. Η παραπάνω δουλειά ανέρχεται σε αρκετά μεγάλο χρονικό διάστημα data entry, το οποίο θα πρέπει να γίνει σε συνεργασία με ειδικούς γλωσσολόγους για να μην υπάρχουν προβλήματα επικαλύψεων, κυρίως όσον αφορά τους γραμματικούς και συντακτικούς κανόνες. Επομένως, οι εγγραφές στη Βάση Γνώσης έγιναν με σκοπό να δοκιμαστούν κάποια τυχαία δείγματα κειμένων που τείνουν να είναι σημασιολογικά ίδια, αλλά διαφοροποιούνται συντακτικά ή γραμματικά, κυρίως για να φανεί στην πράξη η λειτουργικότητα του μοντέλου. Η γλώσσα που είναι γραμμένες οι προτάσεις είναι τα αγγλικά και για τη σύνδεση των λέξεων που χρησιμοποιούν οι προτάσεις αυτές με έννοιες χρησιμοποιήθηκε το διαδικτυακό λεξικό

της αγγλικής γλώσσας "the free dictionary" (http://www.thefreedictionary.com) που διατίθεται δωρεάν μέσω internet.

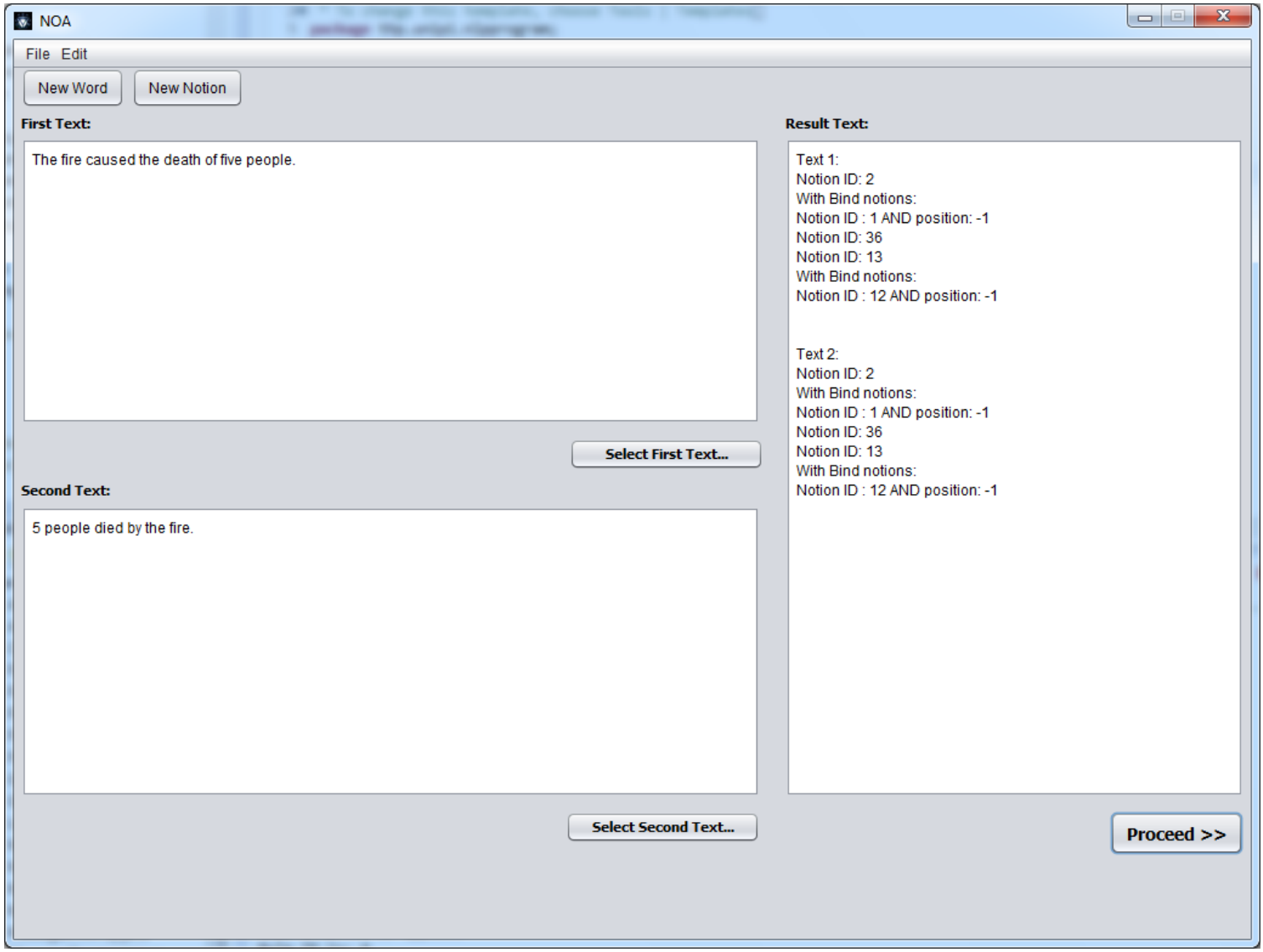
Ξεκινώντας την εφαρμογή ανοίγει ένα Graphics User Interface το οποίο δημιουργήθηκε με τη χρήση της Swing βιβλιοθήκης της Java, μέσα από το οποίο ο χρήστης του συστήματος μπορεί να επιλέξει εύκολα τα δύο κείμενα τα οποία θέλει να συγκρίνει.



Ο χρήστης πατώντας το κουμπί "Select First Text..." μπορεί να περιηγηθεί στο σύστημα αρχείων του λειτουργικού του συστήματος και να επιλέξει το κείμενο το οποίο θέλει. Αντίστοιχα μπορεί να κάνει και για το δεύτερο κείμενο. Η εφαρμογή έχει δημιουργηθεί για να μπορούν να εισαχθούν μόνο αρχεία τύπου .txt.



Αφού ο χρήστης εισάγει τα κείμενα τα οποία θέλει να συγκρίνει πατάει το πλήκτρο "Proceed>>" για να ξεκινήσει η εφαρμογή να τα συγκρίνει ως προς τη σημασιολογική τους ομοιότητα.



Στο παραπάνω παράδειγμα παρατηρείται στο "Result Text" ότι οι ακολουθίες εννοιών των δύο κειμένων που επιστρέφονται είναι ίδιες. Επομένως έχοντας δύο διαφορετικά κείμενα "The fire caused the death of five people" και "5 people died by the fire" τα οποία χρησιμοποιούν διαφορετική σύνταξη, καταλήγουμε στην ίδια ακολουθία εννοιών, που αποδεικνύει ότι τα δύο κείμενα είναι ίδια.

Στην παρούσα έκδοση της εφαρμογής δεν είναι πολύ κατανοητό τι αναπαριστούν τα ID των εννοιών που επιστρέφονται, αφού είναι αριθμοί μη κατανοητοί στον χρήστη. Ωστόσο, δεν έχει καμία ιδιαίτερη σημασία, αφού είναι αντιληπτό ότι οι ακολουθίες είναι ίδιες. Οι αριθμοί αυτοί είναι τα ID των εννοιών που πραγματεύονται οι προτάσεις και των συσχετιζόμενων με αυτά άλλων εννοιών, όπως έχουν οριστεί στη Βάση Γνώσης, στο FRWt_NOTION πίνακα.

Η εφαρμογή θα μπορούσε να επιστρέφει μια τιμή αληθείας (ναι ή όχι) για το αν τα κείμενα κρίθηκαν σημασιολογικά όμοια, όμως επειδή δημιουργήθηκε για να αποτιμηθεί η αξία του προτεινόμενου μοντέλου, κρίνεται χρησιμότερο το αποτέλεσμα που επιστρέφεται, εφόσον μπορεί να χρησιμοποιηθεί για να αιτιολογήσει τον τρόπο που η εφαρμογή έκρινε ότι τα δύο κείμενα είναι σημασιολογικά όμοια ή όχι. Η διαδικασία είναι ιδιαίτερα απλή για τον τελικό χρήστη, όμως καλούνται αρκετές τεχνικές και αντικείμενα τα οποία δεν είναι ορατά στο GUI της εφαρμογής.

Η διαδικασία που ακολουθείται, από τη στιγμή που ο χρήστης θα πατήσει το "Proceed" μέχρι την επιστροφή των αποτελεσμάτων περιγράφεται παρακάτω:

1. Αρχικά η εφαρμογή ελέγχει αν ο χρήστης έχει διαλέξει κείμενα και αν τα κείμενα αυτά παρουσιάζουν διαφορές μεταξύ τους. Αν ο χρήστης έχει διαλέξει κατά λάθος το ίδιο κείμενο δύο φορές η εφαρμογή θα τον ενημερώσει σχετικά και δεν θα προχωρήσει σε περαιτέρω επεξεργασία, αφού κρίνεται περιττή.
2. Δημιουργείται ένα αντικείμενο τύπου "SentenceDetector" το οποίο χρησιμοποιεί τον BreakIterator της Java για να μπορέσει να χωρίσει το κείμενο σε προτάσεις. Ο BreakIterator παρέχεται μαζί με τη βιβλιοθήκη του Collections και χρησιμοποιείται γι αυτόν ακριβώς το λόγο. Θα μπορούσε φυσικά να δημιουργηθεί ένα πιο αποδοτικό εργαλείο για το Sentence Boundary Detection (SBD), αλλά δεν αποτελεί αντικείμενο εστίασης της παρούσας εργασίας και κρίθηκε ότι ο BreakIterator καλύπτει τις ανάγκες. Η μέθοδος splitText που χρησιμοποιείται για να κάνει αυτή τη δουλειά, επιστρέφει το κείμενο σε ένα String Array.
3. Κατόπιν, κάθε εγγραφή στο Array που επιστράφηκε στο παραπάνω βήμα, πρέπει να σπάσει σε επιμέρους λέξεις (tokens). Γι αυτό το λόγο δημιουργήθηκε ένα object τύπου "Tokenizer", το οποίο χρησιμοποιεί τον StringTokenizer της Java util. Η μέθοδος splitSentence του εν λόγω αντικειμένου επιστρέφει ένα ArrayList με τις προτάσεις που σε κάθε θέση περιέχει ένα ArrayList με τις λέξεις της κάθε πρότασης.

```
public static String[] splitSentence(String curSentence)
{
	ArrayList<String> tokenizedSentenceList = new
	ArrayList<String>();
	StringTokenizer st = new StringTokenizer(curSentence);
	while (st.hasMoreTokens())
	{
		tokenizedSentenceList.add(st.nextToken());
	}
	String[] TempArray = new
	String[tokenizedSentenceList.size()];
	TempArray = tokenizedSentenceList.toArray(TempArray);
	return TempArray;
}
```

4. Σε αυτό το βήμα δημιουργείται ένα αντικείμενο τύπου ArrayList<ArrayList<HashMap<Notion, Integer>>> το οποίο θα φιλοξενήσει έναν πίνακα με όλες τις πιθανές έννοιες (notions) που αντιστοιχούν για κάθε token, από αυτά που βρέθηκαν παραπάνω. Για να γίνει αυτό χρησιμοποιείται η κλάση "SemanticDetector", η οποία ελέγχει για κάθε notion αν υπάρχει στη Βάση Γνώσης και με ποιες έννοιες συνδέεται.
5. Κατόπιν με την κλάση "SemanticDisambiguator" το σύστημα αναζητά όλα τα tokens που αντιστοίχησαν σε παραπάνω από μία έννοια στο προηγούμενο βήμα και προσπαθεί να τα αποσαφηνίσει με τη χρήση διάφορων κανόνων από

τη Βάση Γνώσης. Στην παρούσα υλοποίηση χρησιμοποιούνται δύο τύποι κανόνων για το Notion Disambiguation, οι κανόνες που ελέγχουν για ακολουθίες λέξεων και γραμματικοί κανόνες, οι οποίοι εκτελούνται και με τη σειρά που αναφέρθηκαν. Αν κάποιος κανόνας αποσαφηνίσει μια λέξη ως το προς το ποια έννοια είναι, μπορεί να δώσει περαιτέρω πληροφορία στο σύστημα για την αποσαφήνιση και των υπόλοιπων λέξεων. Οπότε έχει φτιαχτεί ένα σύστημα παλινδρομικής χρήσης των κανόνων αυτών εφόσον αλλάξει κάτι από την τελευταία φορά που χρησιμοποιήθηκαν. Το σύστημα μετά την εφαρμογή των κανόνων βλέπει πως ορίστηκαν οι πιθανότητες για τις υποψήφιες έννοιες ανά token και αφαιρεί αυτές που είναι κάτω από το threshold. Στην παρούσα υλοποίηση το threshold έχει οριστεί το 30%, αλλά θα ήταν ορθότερο να οριστεί με τη βοήθεια κάποιου γλωσσολόγου και περαιτέρω μελέτης.

6. Αφού η εφαρμογή τελειώσει με το πρώτο κείμενο ως προς την αποσαφήνιση των εννοιών που χρησιμοποιεί, προχωράει στην κανονικοποίησή του. Αρχικά λοιπόν, δημιουργεί ένα αντικείμενο "NotionsBinder" το οποίο στόχο έχει να εντοπίσει τις βοηθητικές έννοιες και να τις "κρεμάσει" σε άλλες έννοιες. Για παράδειγμα το άρθρο θεωρείται βοηθητική έννοια στο ουσιαστικό. Επομένως, ο "NotionBinder" θα αφαιρέσει το άρθρο από την ακολουθία των εννοιών και θα το κάνει bind στο σχετικό ουσιαστικό. Οι κανόνες για την εφαρμογή των "δεσιμάτων" αυτών μεταξύ των εννοιών, ορίζονται στη Βάση Γνώσης ως n-gram HMMs.
7. Επόμενο βήμα είναι στην υπόλοιπη ακολουθία εννοιών, όπως έχει διαμορφωθεί έως τώρα να εφαρμόσει τους λοιπούς γραμματικούς και συντακτικούς κανόνες από τη Βάση Γνώσης ώστε να κανονικοποιηθεί όσο το δυνατόν καλύτερα γίνεται. Αυτό το αναλαμβάνει το αντικείμενο "SyntacticRules", όπου βρίσκει κανόνες που συνδέονται με τις έννοιες τις ακολουθίας, ή με τις σχέσεις των εννοιών αυτών με άλλες έννοιες. Για παράδειγμα, αν η ακολουθία έχει την έννοια "car" θα αναζητήσει κανόνες όχι μόνο για αυτή την έννοια, αλλά και για την έννοια "noun", "singular", κ.ο.κ.. Σημειωτέον ότι με αντίστοιχο τρόπο λειτουργεί και ο "NotionBinder". Αυτό είναι επιτρεπτό χάρη στη δημιουργία της νέας Βάσης Γνώσης και της μη χρήσης του WordNet.
8. Όταν τελειώσουν όλα τα παραπάνω το πρώτο κείμενο αποθηκεύεται στη μνήμη ως μα κανονικοποιημένη ακολουθία εννοιών και η εφαρμογή προχωράει αντίστοιχα στο δεύτερο. Εφόσον, ολοκληρωθεί η διαδικασία και για τα δύο κείμενα εκτυπώνουν στο GUI τις ακολουθίες ως έχουν για να δει ο τελικός χρήστης αν είναι ίδιες ή όχι.

Για να γίνει λίγο πιο κατανοητός ο τρόπος με τον οποίο η εφαρμογή ελέγχει τους κανόνες παρατίθεται σχετικός κώδικας από τη μέθοδο εφαρμογής κανόνων κανονικοποίησης που χρησιμοποιείται κατά το βήμα 7:

```
public boolean applyRules(ArrayList<ArrayList<Notion>> finalNotionList,
Connector connectorArg) throws SQLException
{
 Connector curConnector = connectorArg;
 boolean rulesApplied = false;
 for(int i=0;i<finalNotionList.size();i++)
  {
   for (int j=0;j<finalNotionList.get(i).size();j++)
   {
    Notion tempNotion = (Notion) finalNotionList.get(i).get(j);
```

```
    ArrayList<Integer> tempList = tempNotion.getNotionAttributes();
    tempList.add(tempNotion.getNotionID());
    for(int k=0;k<tempList.size();k++)
    {
     Connection conn = curConnector.returnConnection();
     Statement st = conn.createStatement();
     st.execute("CALL FRWf_CurrentAttributeRules('" +
     (int) tempList.get(k)+"')");
     ResultSet res = st.getResultSet();
    while (res!=null && res.next())
    {
     boolean exists = false;
     int ruleID = res.getInt(1);
     Statement st2 = conn.createStatement();
     st2.execute("CALL FRWf_ReturnStepsOfRule('"+ruleID+"')");
     ResultSet res2 = st2.getResultSet();
     res2.last();
     int rows = res2.getRow();
     res2.beforeFirst();
     int counter = 0;
     while (res2!=null && res2.next())
     {
      int CurrentAttributeID = res2.getInt(1);
      int RelatedAttributeID = res2.getInt(2);
      int step = res2.getInt(3);
      int towards = res2.getInt(4);
//If towards = 1, then MainAttribute is in previous position than
RelatedAttribute
      if (towards == 1)
      {
       int position = j+step;
       if (position < finalNotionList.get(i).size())
       {
         Notion tempNotion2 = finalNotionList.get(i).get(position);
        ArrayList<Integer> tempList2 = tempNotion2.getNotionAttributes();
        tempList2.add(tempNotion2.getNotionID());
       if (tempList2.contains(RelatedAttributeID))
       {
         exists = true;
         counter++;
       }
       else
       {
         exists = false;
       }
      }
      }
      else //This means that towards = 2
      {
       // .........
      }
     }
```

```
//This means that all the steps of the rule were satisfied
     if(exists && counter == rows)
     {
      rulesApplied = true;
      Statement st3 = conn.createStatement();
      st3.execute("CALL FRWf_RuleEffects('"+ruleID+"')");
      ResultSet res3 = st3.getResultSet();
      HashMap<Notion, Integer> tempMap = new HashMap<>();
      while (res3!=null && res3.next())
      {
       if(res3.getInt(3) == 1) //type = 1
       {
        //Create the notion for the current value
        Notion shadowNotion = new Notion();
        Statement st4 = conn.createStatement();
        st4.execute("CALL FRWf_RelatedAttributes('"+res3.getInt(1)+"')");
        ResultSet res4 = st4.getResultSet();
        while (res4!=null && res4.next())
        {
         shadowNotion.addElement(res4.getInt(1));
        }
        shadowNotion.setNotionID(res3.getInt(1));
//add this element to the tempMap with the position that is need to be place
        tempMap.put(shadowNotion, ((j+res3.getInt(2))-1));
       }
       else if (res3.getInt(3) == 2) //type = 2
       {
        tempMap.put(finalNotionList.get(i).get((j+res3.getInt(1))-
        1),((j+res3.getInt(2))-1));
       }
     }
     if(tempMap.size()>0)
     {
      Iterator it2 = tempMap.entrySet().iterator();
      while(it2.hasNext())
      {
       Map.Entry pairs2 = (Map.Entry)it2.next();
       Notion test = (Notion)pairs2.getKey();
       int position = (int)pairs2.getValue();
       if(position < finalNotionList.get(i).size())
       {
        finalNotionList.get(i).set(position, test);
       }
       else
       {
        finalNotionList.get(i).add(test);
       }
      }
     }
    }
    else
    {
```

```
      System.out.println("Rule "+ruleID+" cannot be applied");
    }
   }
  }
 }
}
```

Αξίζει να αναφερθεί επίσης, ότι η εφαρμογή κάνει σύνδεση με τη Βάση Γνώσης μέσω του αντικειμένου Connector, το οποίο χρησιμοποιεί τον JDBC για να συνδεθεί στην MySQL και να τρέξει ερωτήματα.

Η παρούσα υλοποίηση έχει δοκιμαστεί με τρία ζευγάρια όμοιων προτάσεων διαφορετικού τύπου και συντακτικού και δείχνει να λειτουργεί σωστά. Ωστόσο, το δείγμα είναι αρκετά μικρό και οι προτάσεις δεν είναι ιδιαίτερα πολύπλοκες. Η δημιουργία κανόνων, οι οποίοι να μην δημιουργούν σύγχυση και να είναι ορθοί, είναι μια αρκετά δύσκολη διαδικασία που απαιτεί την ανάμιξη ειδικών επιστημόνων της γλωσσολογίας. Αντίστοιχα δύσκολη και κοπιαστική είναι η αναπαράσταση του λεξιλογίου ως σημασιολογικό δίκτυο μέσα στη Βάση Γνώσης. Η παρούσα εργασία σκοπό είχε να ανοίξει ένα δρόμο ως προς το να προτείνει μια εναλλακτική προσέγγιση στο πρόβλημα της σημασιολογικής ομοιότητας, αλλά ο δρόμος για περαιτέρω ανάπτυξη της εφαρμογής με σκοπό τη δημιουργία ενός ολοκληρωμένου συστήματος που μπορεί να ανταπεξέλθει σε μεγάλο αριθμό κειμένων και πιο πολύπλοκα δομημένων, απαιτεί τη σύσταση ειδικής ομάδας και δεν ήταν το ζήτημα της υλοποίησης. Η υλοποίηση απλά πιστοποίησε την δυνατότητα υλοποίησης του προτεινόμενου μοντέλου NOA.

6.5. Προβλήματα

Είναι επόμενο, εφόσον προτείνεται μια ριζοσπαστική αλλαγή στις τεχνικές αντιμετώπισης της αναγνώρισης σημασιολογικής ομοιότητας μεταξύ κειμένων να ακολουθήσουν και αρκετά προβλήματα, τα οποία καλούνται να λυθούν σε μεταγενέστερο χρόνο. Κάποια από αυτά που εντοπίζονται από τα πρώτα βήματα αναφέρονται παρακάτω.

Η εφαρμογή στηρίζεται εξ ολοκλήρου στην ύπαρξη μιας μεγάλης Βάσης Γνώσης, η οποία είναι σωστά ορισμένη και χρειάζεται ιδιαίτερη προσοχή, ιδιαίτερα στον ορισμό των κανόνων, ώστε να είναι έγκυροι και να μην επικαλύπτουν ο ένας τον άλλον. Αυτό προϋποθέτει ότι η Βάση Γνώσης, χρειάζεται ιδιαίτερη προσπάθεια από ειδικούς επιστήμονες της γλωσσολογίας, οι οποίοι θα πρέπει να την τροφοδοτήσουν με ιδιαίτερη προσοχή, μια αρκετά επίπονη και χρονοβόρα διαδικασία. Πέραν τούτου, η άρρηκτη σχέση του συστήματος με μια ενημερωμένη Βάση Γνώσης, συνεπάγεται συχνή αναζήτηση σε αυτή, κάτι που έχει σαν επακόλουθο τη μείωση της απόδοσης του συστήματος, ιδιαίτερα στην περίπτωση που λόγω του μεγάλου μεγέθους της, η Βάση Γνώσης θα βρίσκεται στον σκληρό δίσκο του υπολογιστή που φιλοξενεί το σύστημα.

Επίσης, το παρόν σύστημα βρίσκεται σε πρωταρχικό στάδιο και έχει αρκετές ατέλειες, όπως περιορισμένο αριθμό τύπου κανόνων, μη εμβάθυνση σε επιμέρους τεχνικές όπως το Tokenization ή η Sentence Boundary Disambiguation, κ.λπ. Αυτά τα συστήματα παρέμειναν σε πρωταρχικό στάδιο καθότι δεν εξυπηρετούν τους σκοπούς της παρούσας εργασίας, η οποία στοχεύει να προτείνει έναν εναλλακτικό δρόμο για να δημιουργηθούν πιο αξιόπιστα συστήματα ανεύρεσης της σημασιολογικής ομοιότητας των κειμένων. Ωστόσο, το σύστημα είναι ανοιχτό σε μελλοντικές επεκτάσεις και βελτιώσεις, ανάλογα με τις ανάγκες που θα προκύψουν.

6.6. Εφαρμογές και Μελλοντικές Επεκτάσεις

Η παραπάνω τεχνική, εφόσον έχει σωστά ορισμένους, όλους τους συντακτικούς και γραμματικούς κανόνες που χρειάζεται, ανοίγει δρόμους σε ένα μεγάλο πλήθος εφαρμογών, που αναζητούν αξιοπιστία στην σημασιολογική ομοιότητα κειμένων. Κάποια από τα προφανή παραδείγματα όπου θα μπορούσε να γίνει χρήση της παραπάνω τεχνικής αναφέρονται παρακάτω:

- Στη δημιουργία εταιρικών Βάσεων Γνώσης, όπου οι διαχειριστές θα μπορούν να εισάγουν πολλά κείμενα που αναφέρονται στο ίδιο θέμα και το σύστημα να τα εξάγει σε ένα κείμενο που εμπεριέχει όλη τη χρήσιμη πληροφορία. Κατά αυτόν τον τρόπο οι εργαζόμενοι, όταν θέλουν να επιμορφωθούν σε κάποιον τομέα, δεν θα χάνουν πολύτιμο χρόνο διαβάζοντας βιβλία και κείμενα τα οποία αναφέρουν τα ίδια πράγματα με άλλα βιβλία.
- Στη δημοσιογραφία και στην ενημέρωση, όπου οι ειδήσεις από πολλές πηγές θα μπορούν να συνοψίζονται σε ένα κείμενο. Επομένως, δεν θα χάνει χρόνο διαβάζοντας από διαφορετικές ιστοσελίδες τις ίδιες ειδήσεις, αλλά θα μπορεί να εξάγει τη συνολική πληροφορία που παρέχεται από τις διάφορες πηγές σε ένα μόνο κείμενο και να το διαβάσει χωρίς να χάνει λεπτομέρειες που μπορεί να αναφέρονται μόνο σε μία πηγή. Επίσης, σε αυτή την περίπτωση θα μπορούσε να οριστεί και βαθμός αξιοπιστίας στα κείμενα, βάσει του πόσες φορές επαναλαμβάνονται στις διάφορες πηγές. Οπότε μια πληροφορία που παρέχεται από πολλές πηγές τείνει να είναι πιο αξιόπιστη από μια πληροφορία που παρέχεται μόνο από μία πηγή.
- Σε επενδύσεις και κυρίως στη χρηματιστηριακή αγορά, όπου οι επενδυτές χρειάζεται να διαβάσουν σε μικρό χρονικό διάστημα πληροφορίες για μια επένδυση που θέλουν να κάνουν. Επομένως, θέλουν να συγκεντρώσουν όλη τη χρήσιμη πληροφορία σε ένα μόνο κείμενο.
- Στην εκπαίδευση, όπου οι μαθητές καλούνται να διαβάσουν διαφορετικά βιβλία, αναφερόμενα στο ίδιο θέμα, καταλήγοντας πολλές φορές να διαβάζουν παραγράφους με γνώση που έχουν αποκτήσει από προηγούμενα βιβλία που διάβασαν. Επομένως, όχι μόνο χάνουν χρόνο, αλλά επίσης χάνουν το ενδιαφέρον τους για το αντικείμενο που μελετάνε, αφού δεν προχωράνε με σταθερά βήματα παρακάτω.

Πέρα από τις παραπάνω εφαρμογές, το μοντέλο που προτείνεται στην παραπάνω τεχνική, δίνει τη δυνατότητα για περαιτέρω επεκτάσεις πέρα από τα όρια της επεξεργασίας φυσικής γλώσσας. Αρκεί το σημασιολογικό δίκτυο να επεκταθεί, περικλείοντας εικόνες και ήχους. Κατόπιν, οι δυνατότητες ενός τέτοιου συστήματος αυξάνονται εκθετικά, καθώς και η γνώση του συστήματος, πλησιάζει περισσότερο αυτή του ανθρώπινου εγκεφάλου.

Η αρχιτεκτονική του συστήματος δίνει τη δυνατότητα της σταδιακής δόμησης και επέκτασής του. Επίσης, στόχος είναι να δημιουργηθούν υποσυστήματα, τα οποία θα δίνουν στο σύστημα τη δυνατότητα όλο και περισσότερο εξελιγμένης αυτόματης μάθησης, με χρήση αλγορίθμων μηχανικής μάθησης και μεγάλες ροές δεδομένων σαν είσοδο, ώστε να αναπαρίσταται όλο και περισσότερη και καλύτερης ποιότητας πληροφορία από τον πραγματικό κόσμο.

7. Βιβλιογραφία

Aho, A., Hopcroft, J., & Ullman, J. (1973). On finding Lowest Common Ancestor.

Banea, C., Hassan, S., Mohler, M., & Mihalcea, R. (n.d.). UNT: A supervised synergetic approach to semantic text similarity.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for Word Sense Disambiguation using WordNet.

Biemann, C. (2012). Creating a system for lexical substitutions from scratch using crowdsourcing.

Border, A. (1997). On the resemblance and containment of documents.

Buchanan, B., & Feigenbaum, E. (1982). Knowledge-based systems in Artificial Intelligence.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory.

Dao, T., & Simpson, T. (2005). Measuring similarity between sentences.

Dijk, T., & Kintsch, W. (1983). Strategies of discourse comprehension.

Dinu, L., & Popescu, M. (2009). Ordinal measures in authorship identification.

Freund, Y., & Schapire, R. (1999). Large margin classification using Perceptron algorithm.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia-Enhancing text categorization with encyclopedic knowledge.

Gale, W., Church, K., & Yarowksy, D. (1992). Work on statistical methods for Word Sense Disambiguation.

Giles, J. (2005). Internet encyclopedias go head to head.

Graesser, A., Millis, K., & Zwaan, R. (1997). Discourse comprehension.

Hatzivassiloglou, V. (1999). Detecting text similarity over short passages: Exploring linguistics feature combinations via machine learning.

Hidden Markov Models-HMM. (n.d.). Ανάκτηση από http://en.wikipedia.org/wiki/Hidden_Markov_model

Hungarian Method. (n.d.). Ανάκτηση από http://en.wikipedia.org/wiki/Hungarian_algorithm

IBM Translator. (1954). Ανάκτηση από http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

Jensen, A., & Boss, N. (2008). Textual Similarity: Comparing texts in order to discover how closely they discuss the same topics.

Jiang, J., & Conrath, D. (1997). Semantic Similarity based on corpus statistics and lexical taxonomy.

Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval.

Joseph, W. (1966). ELIZA-A computer program for the study of natural language communication between man and machine.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model.

Koehn, P. (2005). Europal: A parallel corpus for statistical machine translation.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., & Federico, M. (2007). MOSES: Open source toolkit for statistical machine translation.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction and representation of knowledge.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to Latent Semantic Analysis.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification.

Lenat, D., & Feigenbaum, E. (1991). On the thresholds of knowledge.

Lesk, M. (1986). Automatic Sense Disambiguation using machine readable dictionaries-How to tell a pine cone from an ice cream cone.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals.

Lin, D. (1998). An information-theoretic definition of similarity.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and Knowledge-based measures of Text Semantic Similarity.

Mohler, M., & Mihalcea, R. (2009). Text-to-Text semantic similarity for automatic short answer grading.

Mosteller, F., & Wallace, D. (1964). Inference and disputed authorship.

Negi, P., Rauthan, M., & Dhami, H. (2010). Sentence Boundary Disambiguation-A user friendly approach.

Palmer, D., & Hearst, M. (1994). Adaptive Sentence Boundary Disambiguation.

Porter, M. (1997). An algorithm for suffix stripping.

Quillian, M. (1967). Word concepts: A theory and simulation of some basic semantic capabilities.

Resnik, P. (1999). Semantic similarity in a taxonomy-An information-based measure and its application to problems of ambiguity in natural language.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain.

Ross, S. (1976). A first course in probability.

Salton, G., & Lesk, M. (1968). Computer evaluation of indexing and text processing.

Salton, G., & McGill, M. (1983). An introduction to modern information retrieval.

Schank, R., & Abelson, R. (1978). Scripts, plans and knowledge.

Schank, R., Goldman, N., Rieger, C., & Riesbeck, C. (1973). MARGIE-Memory Analysis Response Generation and Inference on English.

SEMEVAL 2013. (2013). Ανάκτηση από http://www.cs.york.ac.uk/semeval-2013/

SHRDLU. (n.d.). Ανάκτηση από http://hci.stanford.edu/~winograd/shrdlu/

Smola, A., & Schoelkopf, B. (1998). A tutorial on support vector regression.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams.

Sussna, M. (1993). Word Sense Disambiguation for free-text indexing using a massive semantic network.

Tuning, A. (1950). Computing machinery and intelligence.

Turney, P. (2001). Mining the web for synonyms-PMI-IR versus LSA on TOEFL.

Wang, Y., & Witten, H. (1997). Induction of model trees for predicting continuous classes.

Webster, J., & Kit, C. (1992). Tokenization as the initial phase in NLP.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection.

Zobel, J., & Moffat, A. (1998). Exploring the similarity space.

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα