Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Σημασιολογικές Αναπαραστάσεις Λέξεων με χρήση Θεματικής Μοντελοποίησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΕΛΕΥΘΕΡΙΑΣ ΜΠΡΙΑΚΟΥ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής

Εργαστηριο Ορασης Υπολογιστων, Επεξεργασιας Λογου και Επεξεργασιας Σηματων
Αθήνα, Ιούνιος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Σημασιολογικές Αναπαραστάσεις Λέξεων με χρήση Θεματικής Μοντελοποίησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΕΛΕΥΘΕΡΙΑΣ ΜΠΡΙΑΚΟΥ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής

Εγκρίθηκε από την επιτροπή στις 25 Ιουνίου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Ποταμιάνος	Γιώργος Στάμου	Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π	Ε.Μ.Π.

Αθήνα, Ιούνιος 2018

.....

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

Ευχαριστίες

Με τη διπλωματική αυτή εργασία σηματοδοτείται το τέλος των φοιτητικών μου χρόνων στη Σχολή των Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Η πραγματοποίηση αυτής της εργασίας δεν θα ήταν εφικτή δίχως την υποστήριξη και βοήθεια πολλών ατόμων. Θα ήθελα, λοιπόν, να εκφράσω τις ειλικρινείς μου ευχαριστίες προς όλα αυτά τα άτομα.

Πρώτα απ' όλα, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Αλέξανδρο Ποταμιάνο για την υποστήριξή του και την ενθάρρυνσή του κατά τη διάρκεια της διπλωματικής μου εργασίας. Σε πολλά στάδια της έρευνάς μου επωφελήθηκα από τις συμβουλές του, ιδιαίτερα κατά τη διερεύνηση νέων ιδεών. Θα ήθελα επίσης να τον ευχαριστήσω βαθύτατα που με εισήγαγε στους τομείς της Μηχανικής Μάθησης και της Επεξεργασίας Φυσικού Λόγου μέσα από τα μαθήματά του, καθώς η προσέγγισή του διαδραμάτισε σημαντικό ρόλο στην απόφασή μου να συνεχίσω τις σπουδές μου σε αυτούς τους τομείς.

Θα ήθελα να ευχαριστήσω τη Φένια για την υποστήριξη και τη συμβολή της σε ένα μεγάλο μέρος αυτής της διπλωματικής εργασίας, καθώς και για τις ουσιαστικές συζητήσεις και ιδέες που μοιραστήκαμε αυτό το χρόνο. Θα ήθελα να ευχαριστήσω ιδιαίτερα τη Μαλβίνα που αποτέλεσε σταθερή συνεργάτιδα και φίλη των φοιτητικών μου χρόνων, καθώς και όλους τους Πολιτικούς Μηχανικούς φίλους μου!

Τέλος, πρέπει να εκφράσω την βαθύτατη ευγνωμοσύνη μου προς την οικογένειά μου για τη μόνιμη υποστήριξη και τη συνεχή ενθάρρυνση που μου παρείχαν καθ'όλη τη διάρκεια των σπουδών μου καθώς επίσης και κατά τη διάρκεια της έρευνας και συγγραφής αυτής της εργασίας. Τίποτα από όσα έχω καταφέρει έως σήμερα δεν θα ήταν εφικτά χωρίς αυτούς. Ευχαριστώ.

Η διπλωματική αυτή εργασία αφιερώνεται στο θείο μου Γιώργο.

Ελευθερία Μπριάκου
25 Ιουνίου 2018

Περίληψη

Τα Κατανεμημένα Σημασιολογικά Μοντέλα (ΚΣΜ) αποτελούν μια δημοφιλή μέθοδο που κωδικοποιεί τις έννοιες των λέξεων μέσω της στατιστικής ανάλυσης των συμφραζόμενων πλαισίων τους. Οι προκύπτουσες διανυσματικές αναπαραστάσεις λέξεων έχουν χρησιμοποιηθεί επιτυχώς σε διάφορες εφαρμογές της Επεξεργασίας Φυσικού Λόγου (Natural Language Processing, NLP), ενώ χρησιμοποιούνται επίσης για τον υπολογισμό των σημασιολογικών ομοιοτήτων ανάμεσα σε ζεύγη λέξεων. Ωστόσο, μια σημαντική έλλειψη των παραδοσιακών ΚΣΜ είναι ότι οι πολλαπλές έννοιες μιας πολυσήμαντης λέξης συγχωνεύονται σε μια μοναδική διανυσματική αναπαράσταση.

Ο σκοπός αυτής της διπλωματικής εργασίας είναι η αντιμετώπιση του παραπάνω προβλήματος, μέσω της χρήσης δύο μοντέλων που αξιοποιούν θεματικές διανυσματικές αναπαραστάσεις λέξεων οι οποίες έχουν εξαχθεί από Θεματικά ΚΣΜ (ΘΚΣΜ). Αρχικά, βασιζόμενοι στην παρατήρηση ότι οι λέξεις εμφανίζονται συνήθως με μια συγκεκριμένη έννοια σε κάθε θεματική περιοχή, προτείνουμε ένα μείγμα σημασιολογικών μοντέλων που επιτρέπει τον συνδυασμό των ομοιοτήτων ζευγών λέξεων οι οποίες υπολογίζονται σε πολλαπλά ΘΚΣΜ. Στη συνέχεια, επεκτείνουμε αυτό το μοντέλο προκειμένου να αποκτήσουμε μια ενιαία αναπαράσταση των πολλαπλών θεματικών εννοιών των λέξεων σε έναν κοινό διανυσματικό χώρο. Προς αυτή την κατεύθυνση, κάθε ένα από τα ΘΚΣΜ ευθυγραμμίζεται ως προς έναν κοινό διανυσματικό χώρο μέσω γραμμικής απεικόνισης. Αυτή η μέθοδος οδηγεί σε ένα σύνολο διανυσματικών αναπαραστάσεων ανά λέξη, το πλήθος των οποίων ισούται με το πλήθος των θεμάτων. Έπειτα, το πλήθος των προκυπτόντων διανυσμάτων μειώνεται περαιτέρω μέσω συσσωρευτικής ταξιδόμησης.

Επιπλέον, έναν από τους κύριους στόχους αυτής της εργασίας αποτελεί η διερεύνηση των διαφορετικών τρόπων εκτέλεσης των σημασιολογικών απεικονίσεων ανάμεσα στους θεματικούς υποχώρους και στον ενοποιημένο σημασιολογικό χώρο. Συγκεκριμένα, υποθέτουμε ότι τα ΘΚΣΜ ενσωματώνουν σημαντικές διακυμάνσεις στη χρήση των πολυσήμαντων λέξεων, ενώ παράλληλα διατηρούν τις σχετικές σημασιολογικές αποστάσεις ανάμεσα στις μονοσήμαντες λέξεις. Αυτό, μας οδήγησε στο να αντιμετωπίσουμε τις μονοσήμαντες λέξεις ως *σημασιολογικές άγκυρες* που καθορίζουν τις αντιστοιχίσεις ανάμεσα στους σημασιολογικούς μας χώρους. Απ'όσο γνωρίζουμε, αυτή είναι η πρώτη φορά που απεικονίσεις μεταξύ σημασιολογικών χώρων εφαρμόζονται στο πρόβλημα της εκμάθησης πολλαπλών διανυσματικών αναπαραστάσεων για πολυσήμαντες λέξεις.

Τα προτεινόμενα μοντέλα μπορούν να αξιολογηθούν σε σύνολα δεδομένων τα οποία παρέχουν ζεύγη λέξεων παρουσία ή απουσία συμφραζόμενων πλαισίων, επιδεικνύοντας σημαντική βελτίωση της συσχέτισης με τις τιμές αλήθειας οι οποίες παρέχονται από ανθρώπινες εκτιμήσεις, σε σύγκριση με μια βασική προσέγγιση που δεν χρησιμοποιεί θεματικά μοντέλα.

Επιπλέον, τα μοντέλα μας σημειώνουν επιδόσεις συγκρίσιμες με τα καλύτερα προβλεπτικά συστήματα τα οποία προτείνονται στη βιβλιογραφία.

Λέξεις Κλειδιά

σημασιολογική ανάλυση, θεματική μοντελοποίηση, LDA, κατανεμημένα σημασιολογικά μοντέλα, θεματικά κατανεμημένα σημασιολογικά μοντέλα, Word2Vec, σημασιολογικές απεικονίσεις

Abstract

Distributional Semantic Models (DSMs) constitutes a popular method that estimates the meaning of words from the statistical analysis of their contexts. The extracted word representations have been successfully applied to various Natural Language Processing (NLP) applications, and they are typically utilized to compute pairwise semantic similarities of words. However, one major deficiency of traditional DSMs is that the multiple senses of a polysemous word are conflated into a single vector space representation.

The goal of this diploma thesis is to alleviate the above problem, via proposing two models that leverage topic representations of words extracted from Topic-based DSMs (TDSMs). Firstly, motivated by the fact that typically words appear with a specific sense in each topic, we discuss a semantic mixture model that enables the combination of word similarity scores estimated across multiple TDSMs. Afterwards, we extend this work in order to acquire a unified representation of the multiple topic-senses of words in a common space. In this direction, each of the TDSMs are aligned to a common vector space via linear mapping. This results in a set of embedding vectors per word with cardinality equal to the number of topics; the number of resulting vectors is further reduced via agglomerative clustering.

Furthermore, on of the main scopes of this thesis is to investigate different ways to perform the mappings from the topic sub-spaces to the unified semantic space. Specifically, we hypothesize that TDSMs capture meaningful variations in usage of polysemous words, while the relative semantic distance between monosemous words is preserved. This, motivated as to think of monosemous words as *semantic anchors* that determine the mappings between our semantic spaces. Up to our knowledge, this is the first time that mappings between semantic spaces are applied to the problem of learning multiple embeddings for polysemous words.

The proposed models can be evaluated on both contextual and in-isolation semantic similarity tasks, showing a significant improvement of correlation with human annotations, compared to a baseline approach that does not use topic models. Moreover, our models report performances comparable to the best predictive systems that are proposed in the literature.

Keywords

semantic analysis, topic modeling, LDA, distributional semantic models, Word2Vec, embeddings, topic embeddings, semantic mappings

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

BOW	Bag-of-words
LDA	Latent Dirichlet Allocation
PMI	Point-wise Mutual Information
TF-IDF	Term frequency-Inverse Topic Frequency
POS	Part-of-Speech
LSE	Least Squares Estimation
CBOW	Continuous Bag-of-words
W2V	Word2Vec
CCA	Canonical Correlation Analysis
SVD	Singular Value Decomposition
VSM	Vector Space Model
DSM	Distributional Semantic Model
HDP	Hierarchical Dirichlet Process
KL	Kullback-Leibler
TDSM	Topic-based Distributional Semantic Model
UTDSM	Unified multi-Topic Distributional Semantic Model
PCA	Principal Component Analysis
LR	Linear Regression
ΚΣΜ	Κατανεμημένο Σημασιολογικό Μοντέλο
ΘΚΣΜ	Θεματικά Κατανεμημένο Σημασιολογικό Μοντέλο

Κεφάλαιο 1

Εισαγωγή

1.1 Λεξική Σημασιολογία

Με τον όρο *σημασιολογία* αναφερόμαστε στη μελέτη της σημασίας στη γλώσσα. Ο ευρύς ορισμός της ενσωματώνει δύο διαφορετικές έννοιες: η μία είναι η *φιλοσοφική σημασιολογία*, ο στόχος της οποίας είναι να διατυπώσει μια γενική θεωρία της σημασίας και ο δεύτερος είναι η *λεξική σημασιολογία*, ο στόχος της οποίας είναι να καταγράψει τις έννοιες που έχουν λεξικοποιηθεί σε συγκεκριμένες γλώσσες [Miller and Charles, 1991]. Η παρούσα εργασία εμπίπτει σαφώς στη δεύτερη κατηγορία, η οποία αποτελεί επίσης γνωστό τομέα της Επεξεργασίας Φυσικού Λόγου (Natural Language Processing, NLP).

Εξετάζοντας τα παραπάνω από μια υπολογιστική σκοπιά, η λεξική σημασιολογία στοχεύει στο να διευκολύνει τους υπολογιστές να ανιχνεύουν πτυχές της σημασίας των λέξεων στη γλώσσα, καθώς και να κωδικοποιούν αυτές τις πληροφορίες με έναν φορμαλιστικό τρόπο που να επιτρέπει την ερμηνεία τους από τους υπολογιστές. Η βαρύτητα της κατανόησης της σημασίας των λεξικών μονάδων (λέξεων) είναι πρωταρχικής σημασίας για την κατανόηση των γλωσσών, καθώς αυτές αποτελούν τα βασικά συστατικά της ανθρώπινης γλώσσας. Για να κατανοήσουμε με ποίον τρόπο η μη ενσωμάτωση αυτής της γνώσης στα συστήματα ηλεκτρονικών υπολογιστών θα μπορούσε να οδηγήσει σε ανεπαρκή αποτελέσματα, ας εξετάσουμε ένα πραγματικό παράδειγμα. Ας υποθέσουμε ότι χρησιμοποιούμε μια μηχανή αναζήτησης και διαβιβάζουμε σε αυτή την ερώτηση "Πότε γεννήθηκε ο Λίνους Τόρβαλντς;". Η μηχανή απαντά: "ο Λίνους Τόρβαλντς γεννήθηκε το *", θέτοντας απλά ένα ερώτημα στη μηχανή αναζήτησης (π.χ., της Google) μετά τη μετατροπή της ερώτησης σε κατάφαση χρησιμοποιώντας συντακτικούς κανόνες. Αυτό είναι ένα εύκολο παράδειγμα που δεν απαιτεί καμία γνώση της λεξικής σημασιολογίας προκειμένου να απαντηθεί σωστά. Ας υποθέσουμε τώρα ότι προωθούμε στη μηχανή την ερώτηση "Ποιος είναι ο μεγαλύτερος Φυσικός στην Ελλάδα;" η μηχανή εξάγει "Η Ελλάδα είναι από τις χώρες της Ευρώπης με τη μεγαλύτερη ενεργειακή εξάρτηση σε πετρέλαιο και φυσικό αέριο", το οποίο δείχνει σαφώς ότι το σύστημα είναι ανίκανο να κατανοήσει την ερώτηση. Γιατί είναι λοιπόν δύσκολο για τη μηχανή να επιστρέψει τη σωστή έξοδο στο προηγούμενο παράδειγμα; Η απάντηση είναι απλή: η σημασιολογία των λέξεων δεν έχει ληφθεί υπόψη. Συνεχίζοντας με το δεύτερο παράδειγμα, έρχεται στην επιφάνεια ένα άλλο γλωσσικό φαινόμενο που διαδραματίζει σημαντικό ρόλο στην κατανόηση της γλώσσας. Συγκεκριμένα, η πολυσημία της λέξης *φυσικός* θα μπορούσε να επηρεάσει δραστικά την αξιοπιστία της εξόδου του συστήματος, καθώς η λέξη θα μπορούσε είτε να αντιστοιχίζεται στην έννοια "επιστήμο-

νας" είτε στην έννοια του "προερχόμενου από τη φύση", ανάλογα με τα συμφραζόμενα στα οποία συναντάται.

Επιπλέον, πειραματικές μελέτες φαίνεται να δείχνουν ότι η κατανόηση αυτών των σημασιολογικών στοιχείων θα μπορούσε να ενισχυθεί από τις θεμελιώδεις σημασιολογικές σχέσεις μεταξύ των λέξεων [Marmaridou, 2000]. Μια απλή συνέπεια αυτής της παρατήρησης είναι ότι η σημασία μιας λέξης εξαρτάται σε μεγάλο βαθμό από τις σημασιολογικές σχέσεις που έχει με άλλες λέξεις. Για τον λόγο αυτό, ο υπολογισμός σημασιολογικής ομοιότητας — ο οποίος αντιστοιχεί στον βαθμό ομοιότητας του νοήματος δύο λέξεων — αποτελεί έναν δημοφιλή τομέα της Επεξεργασίας Φυσικού Λόγου που προσπαθεί να αποκαλύψει την πραγματική σημασία των λέξεων. Ορισμένες από τις εφαρμογές της Επεξεργασίας Φυσικού Λόγου που ενσωματώνουν τις παραπάνω πληροφορίες στα συστήματά τους περιλαμβάνουν: την αυτόματη μετάφραση κειμένων, την ανάκτηση πληροφοριών, την αυτόματη περίληψη κειμένου, τη δημιουργία φυσικής γλώσσας, την απάντηση σε ερωτήσεις, τις μηχανές αναζήτησης και τη μετατροπή της προφορικής ομιλίας σε κείμενο.

1.2 Συνεισφορά Εργασίας

Οι αλγόριθμοι εκμάθησης λεξικών αναπαραστάσεων υιοθετούν την *υπόθεση της κατανεμημένης έννοιας των λέξεων* (distributional hypothesis) [Harris, 1954], υποθέτοντας ότι υπάρχει μια συσχέτιση μεταξύ της σχέσης κατανομής και της σημασιολογικής σχέσης των λέξεων. Συνήθως, τα μοντέλα που χρησιμοποιούν τους παραπάνω αλγορίθμους κωδικοποιούν την πληροφορία των λέξεων με βάση τα συμφραζόμενά τους σε διανύσματα χαρακτηριστικών — συχνά αναφερόμενα ως *embeddings* — ενός χώρου διάστασης k, δημιουργώντας ένα μοντέλο διανυσματικού χώρου (Vector Space Model, VSM). Τα παραπάνω μοντέλα έχουν αποδειχθεί χρήσιμα σε διάφορες εφαρμογές της Επεξεργασίας Φυσικού Λόγου, όπως η ανάκτηση πληροφοριών [Manning et al., 2008], η συναισθηματική ανάλυση κειμένου [Tai et al., 2015], η μηχανική μετάφραση [Sharaf et al., 2017] και άλλες.

Παρά τη δημοτικότητά τους, τα παραδοσιακά Κατανεμημένα Σημασιολογικά Μοντέλα (ΚΣΜ, Distributional Semantic Models) βασίζονται αποκλειστικά σε μοντέλα όπου κάθε λέξη αντιπροσωπεύεται μοναδικά από ένα σημείο στον χώρο. Από μια γλωσσολογική σκοπιά, τα μοντέλα αυτά δεν μπορούν να ενσωματώσουν με ακρίβεια την έννοια των πολυσήμαντων λέξεων (π.χ., *καρκίνος* ή *φυσικός*), με αποτέλεσμα να συνδυάζουν τις διαφορετικές σημασιολογικές έννοιες των λέξεων σε μια αναπαράσταση. Προς επίλυση αυτού του προβλήματος, μερικές μελέτες ενσωματώνουν πολλαπλές αναπαραστάσεις ανά λέξη στα αντίστοιχα VSMs, με βάση την κατηγοριοποίηση των τοπικών συμφραζόμενων των λέξεων [Reisinger and Mooney, 2010, Tian et al., 2014, Neelakantan et al., 2014]. Παράλληλα, γλωσσικές μελέτες δείχνουν ότι το ευρύτερο πλαίσιο των συμφραζόμενων μιας λέξης μπορεί επίσης να βοηθήσει στην κατανόηση της γλώσσας. Τα μοντέλα θεματικής μοντελοποίησης της γλώσσας εισήχθησαν ως ένας φυσικός τρόπος αναπαράστασης του ευρύτερου πλαισίου των λεξικών συμφραζόμενων από τους Liu et al. [2015b].

Ακολουθώντας την ίδια κατεύθυνση, προτείνουμε αρχικά έναν συνδυασμό σημασιολογικών μοντέλων κάνοντας χρήση τεχνικών θεματικής μοντελοποίησης. Αυτή η προσέγγιση περιγράφεται επίσης και στη δημοσιευμένη εργασία των Christopoulou et al. [2018]. Συγκεκριμένα, περιγράφουμε έναν συνδυασμό σημασιολογικών ομοιοτήτων που ορίζονται ανάμεσα

σε ζεύγη λέξεων των οποίων οι αναπαραστάσεις έχουν εξαχθεί από Θεματικά Κατανεμημένα Σημασιολογικά Μοντέλα (ΘΚΣΜ, Topic-based Distributional Semantic Models). Η κύρια υπόθεση του συγκεκριμένου μοντέλου είναι ότι οι έννοιες των λέξεων μεταβάλλονται ανάλογα με το θεματικό περιεχόμενο στο οποίο εντάσσονται. Αυτό είναι παρόμοιο με τη χρήση ενός συνδυασμού σημασιολογικών μοντέλων προκειμένου να κωδικοποιηθούν οι πολλαπλές αισθήσεις/έννοιες των λέξεων. Ωστόσο, μια από τις σημαντικότερες ελλείψεις της παραπάνω προσέγγισης είναι οτι δεν καταφέρνει να συλλάβει τις σχέσεις μεταξύ των λέξεων που δεν ανήκουν στον ίδιο θεματικό τομέα, καθώς η σύγκριση ενός ζεύγους λέξεων περιορίζεται σε επίπεδο θεματικής περιοχής. Προς επίλυση αυτού του προβλήματος, προτείνουμε επιπλέον μια πιο ευέλικτη προσέγγιση που χρησιμοποιεί Θεματικά Κατανεμημένα Σημασιολογικά Μοντέλα (ΘΚΣΜ), προκειμένου να δημιουργήσει ένα ενοποιημένο μοντέλο πολλαπλών θεματικών αναπαραστάσεων για κάθε λέξη. Για το σκοπό αυτό, τα ΘΚΣΜ θα πρέπει να ευθυγραμμιστούν ως προς ένα κοινό σύστημα συντεταγμένων που θα επιτρέπει τη σύγκριση των θεματικών αναπαραστάσεων λέξεων που εξάγονται από διαφορετικούς θεματικούς χώρους. Με βάση τις υπάρχουσες μεθόδους απεικόνισης, αναλύουμε τον ρόλο των μονοσήμαντων λέξεων στον καθορισμό μετασχηματισμών μεταξύ των ΘΚΣΜ. Απ' όσο γνωρίζουμε, αυτή είναι η πρώτη δουλειά που εφαρμόζει τεχνικές απεικόνισης μεταξύ σημασιολογικών χώρων της ίδιας γλώσσας, προκειμένου να περιγράψει την πολυσημία των λέξεων.

1.3 Διάρθρωση Εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται ως εξής:

- Το Κεφάλαιο 2 παρέχει μια περιγραφή της βιβλιογραφίας που χωρίζεται σε δύο βασικά τμήματα, σύμφωνα με τα πεδία που καλύπτει η εργασία. Η πρώτη ενότητα δίνει μια σύντομη επισκόπηση των Κατανεμημένων Σημασιολογικών Μοντέλων (ΚΣΜ). Συγκεκριμένα, μελετώνται οι βασικές κατηγορίες των παραδοσιακών ΚΣΜ (μοναδικές αναπαραστάσεις ανά λέξη) και των ΚΣΜ που χρησιμοποιούν πολλαπλές αναπαραστάσεις ανά λεξη. Στο δεύτερο τμήμα συζητάμε τις βασικές μεθόδους μετασχηματισμού μεταξύ σημασιολογικών χώρων μαζί με τις εφαρμογές τους.

- Στο Κεφάλαιο 3 παρουσιάζονται οι αλγόριθμοι: Latent Dirichlet Allocation (LDA) και Word2Vec που αποτελούν τα βασικά συστήματα που χρησιμοποιούν οι δύο προτεινόμενες προσεγγίσεις μας.

- Το Κεφάλαιο 4 περιγράφει το κύριο κίνητρο και την αρχιτεκτονική του συστήματος της πρώτης μας προσέγγισης, την οποία ονομάζουμε: Μείγμα Θεματικά Κατανεμημένων Σημασιολογικών Μοντέλων (Mixture of Topic-based Distributional Semantic Models).

- Το Κεφάλαιο 5 περιγράφει το κύριο κίνητρο και την αρχιτεκτονική του συστήματος της δεύτερής μας προσέγγισης, την οποία ονομάζουμε: Ενιαίο Κατανεμημένο Σημασιολογικό Μοντέλο πολλαπλών Θεματικών Αναπαραστάσεων (Unified multi-Topic Based Distributional Semantic Model).

- Το Κεφάλαιο 6 παρουσιάζει την πειραματική διαδικασία και τα αποτελέσματα για τα διάφορα πειράματα που πραγματοποιήθηκαν σε διαφορετικά σύνολα δεδομένων.

- Το κεφάλαιο 7 περιέχει τα συμπεράσματα της διπλωματικής εργασίας και προτείνει κατευθύνσεις για μελλοντικές μελέτες.

Κεφάλαιο 2

Κατανεμημένα Σημασιολογικά Μοντέλα

2.1 Υπόθεση της Κατανεμημένης έννοιας των λέξεων

Τα Κατανεμημένα Σημασιολογικά Μοντέλα (Distributional Semantic Models, DSMs) περιλαμβάνουν ένα ευρύ φάσμα μεθόδων που βασίζονται στην υπόθεση της κατανεμημένης έννοιας των λέξεων (distributional hypothesis) προσπαθώντας να συλλάβουν τις έννοιες των γλωσσικών οντοτήτων (λέξεων, φράσεων) από τη χρήση τους στη γλώσσα. Αυτή η υπόθεση συχνά περιγράφεται από το διάσημο απόφθεγμα "Μπορείτε να αναγνωρίζετε μια λέξη από τις λέξεις με τις οποίες συνυπάρχει" [Firth, 1957]. Άμεση συνέπεια αυτής της υπόθεσης είναι ότι δύο λέξεις που θεωρούμε ότι είναι σημασιολογικά παρόμοιες αναμένεται να εμφανίζονται σε παρόμοια *συμφραζόμενα πλαίσια*, και αντίστροφα. Προκειμένου να μπορέσουμε να χρησιμοποιήσουμε αυτή την υπόθεση στην Επεξεργασία Φυσικού Λόγου, απαιτείται ο φορμαλιστικός ορισμός του τι θεωρούμε ως *συμφραζόμενα* μιας λέξης. Σε αυτή τη διπλωματική εργασία, ακολουθούμε τον κοινώς χρησιμοποιούμενο ορισμό του *συμφραζόμενου πλαισίου* ως ένα σύνολο λέξεων που υπάρχουν μέσα σε ένα παράθυρο γύρω από κάθε εμφάνιση της λέξης ενδιαφέροντος.

The functional interplay of philosophy and	science	should, as a minimum, guarantee...
...and among works of dystopian	science	fiction...
The rapid advance in	science	today suggests...
...calculus, which are more popular in	science	-oriented schools.
But because	science	is based on mathematics...
...the value of opinions formed in	science	as well as in the religions...
...if	science	can discover the laws of human nature....
...is an art, not an exact	science	.
...factors shaping the future of our civilization:	science	and religion.
...certainty which every new discovery in	science	either replaces or reshapes.
...if the new technology of computer	science	is to grow significantly
He got a	science	scholarship to Yale.
...frightened by the powers of destruction	science	has given...
...but there is also specialization in	science	and technology...

Σχήμα 2.1: Παράδειγμα της υποθετικής σημασιολογικής κατανομής της λέξης *science*.

2.2 Μονές Αναπαραστάσεις Λέξεων

Τα μοντέλα που ακολουθούν την υπόθεση της κατανεμημένης έννοιας των λέξεων αναφέρονται συχνά ως Κατανεμημένα Σημασιολογικά Μοντέλα (ΚΣΜ). Συνήθως, αυτά τα μοντέλα κωδικοποιούν την πληροφορία των λέξεων με βάση τα συμφραζόμενα πλαίσιά τους σε διανύσματα χαρακτηριστικών τα οποία τοποθετούνται σε έναν χώρο διάστασης k, δημιουργώντας έναν διανυσματικό χώρο (Vector Space) που συχνά αναφέρεται ως σημασιολογικός χώρος (Semantic Space). Εξετάζοντας αυτά τα μοντέλα από γεωμετρική σκοπιά, κάθε λέξη αντιπροσωπεύεται ως ένα σημείο στον χώρο, ενώ σημασιολογικά παρόμοιες λέξεις τοποθετούνται πιο κοντά στον σημασιολογικό χώρο και οι ανόμοιες λέξεις τοποθετούνται η μία μακρυά από την άλλη. Οι Baroni et al. [2014] αναφέρουν ότι τα ΚΣΜ που χρησιμοποιούν μία αναπαράσταση για κάθε λέξη μπορούν να χωριστούν σε δύο ευρείες κατηγορίες: τα μοντέλα που βασίζονται σε μέτρησεις και τα μοντέλα πρόβλεψης.

2.2.1 Μοντέλα βασισμένα σε μετρήσεις

Στην απλούστερη περίπτωση των παραδοσιακών ΚΣΜ, κάθε διάσταση συμπεριλαμβάνει στατιστικές πληροφορίες των συμφραζόμενων στοιχείων που παρατηρούνται να συνυπάρχουν με τη λέξη ενδιαφέροντος σε απόσταση μικρότερη μιας σταθεράς c. Αυτή η απλή μέθοδος καταμέτρησης καταλήγει σε έναν πίνακα συνεμφάνισης, όπου οι συνιστώσες του κάθε διανύσματος μπορούν να ερμηνευθούν ως βάρη που υποδηλώνουν τη δύναμη της σημασιολογικής σχέσης ανάμεσα στη λέξη ενδιαφέροντος και το αντίστοιχο συμφραζόμενο στοιχείο. Ωστόσο, έχει παρατηρηθεί ότι οι μετρήσεις συν-εμφανίσεων δεν αποτελούν από μόνες τους μια αξιόπιστη πηγή πληροφοριών για την εξαγωγή σημασιολογικών σχέσεων, καθώς συχνά εμφανιζόμενες λέξεις (όπως η λέξη «a» στο παράδειγμά του Σχήματος 2.1) τείνουν να συνυπάρχουν με πολλές λέξεις με μεγάλη συχνότητα.

Προκειμένου να αμβλυνθεί αυτό το πρόβλημα, μπορούν να εφαρμοστούν μη γραμμικές πράξεις στον πίνακα συν-εμφάνισης για τον οποίο συζητήσαμε προηγουμένως. Συγκεκριμένα, η ενσωμάτωση μη γραμμικών πράξεων μπορεί να οδηγήσει στην 'υποβάθμιση' του ρόλου που διαδραματίζουν οι λέξεις με υψηλή συχνότητα εμφάνισης. Ο ευρύτερα χρησιμοποιούμενος μετασχηματισμός που ακολουθεί την παραπάνω κατεύθυνση είναι ο Positive Pointwise Mutual Information (PPMI) που ορίζεται από τους Church and Hanks [1989] ως:

$$\text{PPMI}(word_i, word_j) = \max(0, \text{PMI}(word_i, word_j)) \tag{2.1}$$

$$\text{PMI}(word_i, word_j) = \log_2 \frac{P(word_i) \cap P(word_j)}{P(word_i)P(word_j)} \tag{2.2}$$

Στην παραπάνω σχέση ο αριθμητής μας δίνει πληροφορία για το πόσο συχνά συν-εμφανίζονται οι δύο λέξεις, ενώ ο παρονομαστής μας λέει πόσο συχνά θα περίμενε κανείς ότι οι δύο λέξεις συνυπάρχουν, αν υποθέσουμε ότι οι εμφανίσεις τους είναι ανεξάρτητες, έτσι ώστε οι πιθανότητές τους να πολλαπλασιάζονται (βλ. παράδειγμα στον Πίνακα 2.1).

Δεδομένου ότι οι μέθοδοι που βασίζονται σε μετρήσεις υπολογίζουν τον πίνακα συν-εμφάνισης για όλες τις λέξεις, οδηγούν σε αραιές (sparse) αναπαραστάσεις υψηλής διάστασης —δηλαδή τα περισσότερα από τα στοιχεία των διανυσμάτων είναι μηδενικά— καθώς μια λέξη

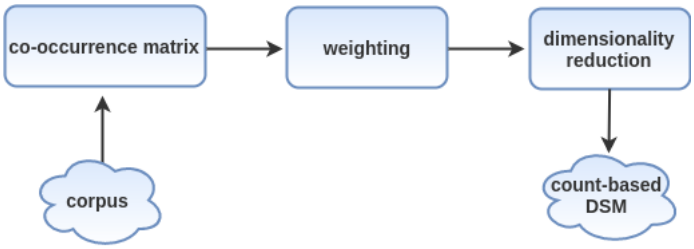
	player	field	court	Athenian	cart	a
baseball	546	350	5	1	35	975
basketball	485	10	410	1	45	1053
democracy	1	5	2	350	10	375
monarchy	2	1	4	7	276	330

↓

	player	field	court	Athenian	cart	a
baseball	0.38	0.97	0	0	0	0
basketball	0.21	0	0	0	0	0.01
democracy	0	0	0	1.93	0	0
monarchy	0	0	0	0	1.86	0.03

Πίνακας 2.1: Παράδειγμα ενός πίνακα συν-εμφανίσεων που εξάγεται χρησιμοποιώντας μετρήσεις πριν (άνω πίνακας) και μετά (κάτω πίνακας) από το μετασχηματισμό PPMI.

συνήθως σχετίζεται σημασιολογικά με ένα μικρό ποσοστό από το σύνολο των πιθανών συμφραζόμενων στοιχείων. Συνήθως, ως επόμενο βήμα μειώνεται η διάσταση του μεγάλου πίνακα συν-εμφάνισης (στην περίπτωση αυτή, ο πίνακας PPMI-ζυγίζεται) προκειμένου να μειωθεί ο θόρυβος και ο διανυσματικός χώρος να γίνει λιγότερο αραιός. Η βασική ιδέα είναι να δημιουργηθεί μια προσέγγιση μικρότερης διάστασης από τον αρχικό πίνακα, διατηρώντας παράλληλα τις σχέσεις μεταξύ των διανυσμάτων. Ο προκύπτων μικρότερος χώρος διαστάσεων αντιπροσωπεύεται από τις σημαντικότερες διαστάσεις του συνόλου δεδομένων, κατά μήκος των οποίων παρατηρείται η μεγαλύτερη μεταβλητότητα (variation). Η πιο δημοφιλής μέθοδος για τη δημιουργία τέτοιων προσεγγιστικών πινάκων είναι η Singular Value Decomposition (SVD) [Landauer and Dumais, 1997], η οποία βασίζεται στην εξαγωγή των ριζών των ιδιοτιμών (singular values) του αρχικού πίνακα. Το Σχήμα 2.2 συνοψίζει τα βήματα δημιουργίας ενός ΚΣΜ βασισμένου σε μετρήσεις.



Σχήμα 2.2: Αναπαράσταση της δημιουργίας ενός Κατανεμημένου Σημασιολογικού Μοντέλου που βασίζεται σε μετρήσεις.

2.2.2 Μοντέλα βασισμένα σε προβλέψεις

Τα μοντέλα που βασίζονται σε προβλέψεις ανήκουν στη νέα γενιά των ΚΣΜ, που προσεγγίζουν το πρόβλημα εκτίμησης διανυσμάτων ως ένα πρόβλημα χωρίς επίβλεψη. Οι Bengio et al. [2003] ήταν οι πρώτοι που εισήγαγαν τεχνητά νευρωνικά δίκτυα στον τομέα των σημασιολογικών αναπαραστάσεων λέξεων και λίγα χρόνια αργότερα οι Collobert and Weston

[2008] ήταν οι πρώτα που τα καθιέρωσαν ως ένα αποτελεσματικό εργαλείο σε εφαρμογές της Επεξεργασίας Φυσικού Λόγου. Αν και οι δύο κατηγορίες μοντέλων βασίζουν το θεωρητικό τους υπόβαθρο στην *υπόθεση της κατανεμημένης έννοιας των λέξεων*, διαφέρουν ως προς την υπολογιστική προσέγγιση που ακολουθούν για να μάθουν γεωμετρικές κωδικοποιήσεις λέξεων. Η νέα γενιά των ΚΣΜ αντί να συλλέγει τα στατιστικά των συμφραζόμενων λέξεων για να αποκαλύψει το σημασιολογικό τους νόημα, χρησιμοποιεί τεχνικές μηχανικής μάθησης και προσπαθεί να δημιουργήσει προγράμματα που μαθαίνουν να λαμβάνουν σωστές αποφάσεις ανάλογα με τα δεδομένα πάνω στα οποία εκπαιδεύονται και να βελτιώνονται με την εμπειρία.

Η γενική δομή των προγνωστικών μοντέλων βασίζεται σε ένα πιθανοτικό νευρωνικό δίκτυο με τροφοδοσία προς τα εμπρός, το οποίο παίρνει ως είσοδο λέξεις και τις ενσωματώνει σαν διανύσματα σε έναν χώρο μικρότερης διάστασης, ο οποίος στη συνέχεια ρυθμίζεται με λεπτομέρεια χρησιμοποιώντας back-propagation. Για τον λόγο αυτό, οι διανυσματικές αναπαραστάσεις που εξάγονται ως βάρη του πρώτου επιπέδου (layer) του νευρωνικού μοντέλου αναφέρονται συνήθως ως *embeddings* στη βιβλιογραφία. Μια συστηματική σύγκριση μεταξύ των *embeddings* και των διανυσματικών αναπαραστάσεων που λαμβάνονται μέσω μοντέλων που βασίζονται σε μετρήσεις έχει μελετηθεί διεξοδικά από τους Baroni et al. [2014]. Η ανωτερότητα των πυκνών αναπαραστάσεων (dense feature vectors) έχει επίσης αποδοθεί σε υπολογιστικούς λόγους από τους Goldberg [2017], καθώς η πλειοψηφία των εργαλειοθηκών (toolkits) δεν λειτουργεί αποδοτικά με πολύ μεγάλης διαστάσεως, αραιά διανύσματα.

Το Word2Vec των Mikolov et al. [2013b], αποτελεί το πιο προεξέχον μοντέλο που δημιουργεί διανυσματικές αναπαραστάσεις λέξεων οι οποίες αντικατοπτρίζουν τη σημασιολογική τους έννοια. Η ιδέα πίσω από το συγκεκριμένο μοντέλο βρίσκεται στη δημιουργία διανυσματικών αναπαραστάσεων για λέξεις-στόχους που επιτρέπουν την πρόβλεψη των πιο πιθανών συμφραζόμενων στοιχείων τους ή αντίθετα στη χρήση των συμφραζόμενων στοιχείων προκειμένου να προβλεφθούν οι αντίστοιχες λέξεις-στόχοι. Αφήνουμε την αναλυτική εξήγησή του μοντέλου για την Ενότητα 4.

2.3 Πολλαπλές αναπαραστάσεις λέξεων

Ένα μεγάλο ποσοστό πρόσφατων εργασιών σχετικά με τη σημασιολογική αναπαράσταση λέξεων βασίζεται αποκλειστικά σε μοντέλα όπου κάθε λέξη αντιπροσωπεύεται από ένα σημείο στον σημασιολογικό χώρο. Από γλωσσολογική σκοπιά, τα μοντέλα αυτά δεν μπορούν να αποδόσουν με ακρίβεια την έννοια μιας πολυσήμαντης λέξης, οδηγώντας σε αναπαραστάσεις οι οποίες συγχέουν τις διαφορετικές έννοιές τους. Η προβληματική φύση των μοντέλων που χρησιμοποιούν μονές αναπαραστάσεις λέξεων είναι προφανής στα ακόλουθα δύο παραδείγματα όπου γίνεται χρήση δύο διαφορετικών εννοιών της λέξης *python*.

- '...students find coding in ***python*** a satisfying experience...'
- '..***python*** uses its sharp, backward-curving teeth...'

Εδώ, οι υπονοούμενες έννοιες της λέξης python είναι τελείως διαφορετικές στα δύο πλαίσια συμφραζομένων (γλώσσα προγραμματισμού, φίδι). Προκειμένου να καταστούν οι διαφορές αυτές δυνατές στην Επεξεργασία Φυσικού Λόγου, πρέπει να λάβουμε υπόψη την πολυσημία στα μοντέλα μας και να μετατρέψουμε τις μονές αναπαραστάσεις λέξεων σε πολλαπλές. Στις επόμενες ενότητες συγκεντρώνουμε τις μεθόδους που αποδίδουν πολλαπλές αναπαραστάσεις

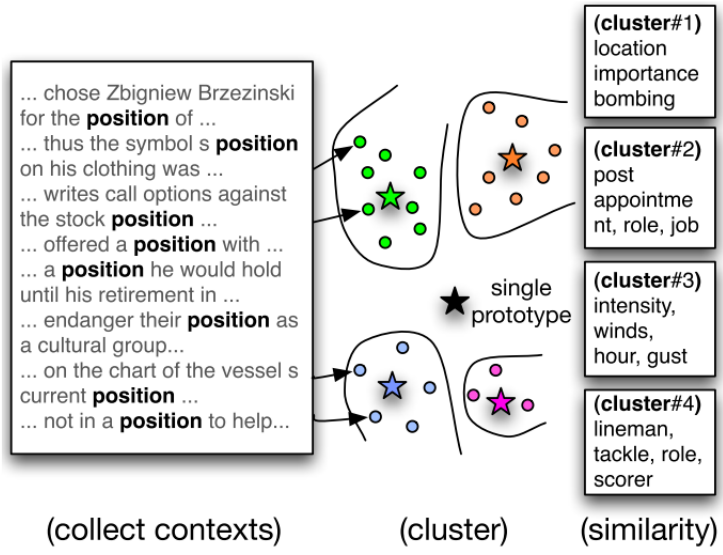
ανά λέξη, χωρίζοντάς τες σε δύο ευρείες κατηγορίες: η πρώτη κατηγορία περιλαμβάνει τις *Μεθόδους Χωρίς Επίβλεψη* οι οποίες δημιουργούν πολλαπλές αναπαραστάσεις χωρίς τη χρήση σημασιολογικών λεξικών πόρων, ενώ η δεύτερη κατηγορία συμπεριλαμβάνει τις *Μεθόδους Με Επίβλεψη* που βασίζονται σε κάποια πρότερη γνώση αναφορικά με τις διαφορετικές έννοιες των λέξεων.

2.3.1 Μέθοδοι χωρίς Επίβλεψη

Σταθερός πλήθος αναπαραστάσεων για κάθε λέξη

Οι Reisinger and Mooney [2010] ήταν οι πρώτοι που εισήγαγαν πολλαπλές αναπαραστάσεις λέξεων. Βασιζόμενοι στην *υπόθεση της κατανεμημένης έννοιας των λέξεων*, συγκέντρωσαν τοπικά συμφραζόμενα για κάθε λέξη ενδιαφέροντος τα οποία αναπαρίστανται ως διανύσματα που σχηματίστηκαν με τη συλλογή στατιστικών στοιχείων συχνότητας σε ένα σταθερό παράθυρο γύρω από αυτές. Έπειτα εφάρμοσαν ομαδοποίηση (clustering) στα προκύπτοντα διανύσματα, με το πλήθος των ομάδων να αποτελεί τη μοναδική παράμετρο του μοντέλου τους. Τα κεντροειδή των προκυπτουσών ομάδων χρησιμοποιήθηκαν για να δημιουργηθεί ένα σύνολο διανυσμάτων για κάθε λέξη ενδιαφέροντος (Σχήμα 2.3).

Ακολουθώντας την προσέγγιση της ομαδοποίησης, οι Huang et al. [2012] πρότειναν ένα αναδρομικό νευρωνικό δίκτυο που ενσωμάτωνε τόσο το καθολικό όσο και το τοπικό πλαίσιο των συμφραζόμενων λέξεων προκειμένου να μάθουν πολλαπλα διανύσματα χαμηλής διάστασης για κάθε λέξη. Και πάλι το πλήθος των πιθανών εννοιών που αντιστοιχούν σε κάθε λέξη συμπίπτει με το σταθερό πλήθος ομάδων.



Σχήμα 2.3: Επισκόπηση της μεθόδου πολλαπλών αναπαραστάσεων ανά λέξη που χρησιμοποιεί η προσέγγιση ομαδοποίησης συμφραζομένων.

Ένα πιθανοτικό πλαίσιο εισήχθη αργότερα από τους Tian et al. [2014] που επέκτειναν το μοντέλο Word2Vec μέσω της αντιπροσώπευσης της πιθανότητας μιας συμφραζόμενης λέξης, δεδομένης της λέξης ενδιαφέροντος, ως έναν πεπερασμένο συνδυασμό όλων των δυνατών αναπαραστάσεων της λέξης ενδιαφέροντος. Χρησιμοποιώντας αυτό το πλαίσιο, σχεδίασαν

έναν Expectation-Maximization αλγόριθμο για να μάθουν πολλαπλές αναπαραστάσεις, όπου το πλήθος των εννοιών που αποδίδονται σε κάθε λέξη αποτελεί μια προκαθορισμένη παράμετρο του συστήματος.

Παρά το γεγονός ότι τα μοντέλα με σταθερό πλήθος πρωτοτύπων ανά λέξη αποτέλεσαν τις πρώτες προσπάθειες ενσωμάτωσης της πολυσημίας σε μοντέλα σημασιολογικών αναπαραστάσεων, οι πιο πρόσφατες προσεγγίσεις παρέχουν πιο ευέλικτες λύσεις στο πρόβλημα. Η ευελιξία τους αποδίδεται στο γεγονός ότι το πραγματικό πλήθος των εννοιών για τις λέξεις διαφέρει ανάλογα: με τον βαθμό πολυσημίας τους (σημειώστε ότι μερικές λέξεις έχουν μόνο μία έννοια, γνωστές και ως μονοσήμαντες λέξεις), και τις αλλαγές που υφίστανται στο χρόνο καθώς η εξέλιξη της γλώσσας προκαλεί τη δημιουργία νέων εννοιών (π.χ. λέξη *python* ως γλώσσα προγραμματισμού).

Προσαρμοστικό Πλήθος Αναπαραστάσεων για κάθε λέξη

Οι πιο πρόσφατες προσεγγίσεις επικεντρώνονται κυρίως σε αρχιτεκτονικές νευρωνικών δικτύων που κωδικοποιούν τις πολλαπλές έννοιες των λέξεων. Οι Neelakantan et al. [2014] βασιζόμενοι στις προηγούμενες προσεγγίσεις ομαδοποίησης των συμφραζόμενων, ακολούθησαν μια online μέθοδο εκμάθησης skip-gram εννοιολογικών αναπαραστάσεων λέξεων, κατά την οποία υπολόγιζαν επίσης το πλήθος των ομάδων. Σε αντίθεση με τις προηγούμενες προσεγγίσεις, αμφότερα τα διάνυσμα των συμφραζόμενων καθώς και της λέξης ενδιαφέροντος μαθαίνονται ταυτόχρονα, αντί να μαθαίνονται οι πρώτες ως μέρος ενός βήματος προ-επεξεργασίας. Αργότερα, ένα δυναμικό Γκαουσιανό μοντέλο skip-gram εισήχθη από τους Chen et al. [2015] επιτρέποντας την ανίχνευση διαφορετικού πλήθους αισθήσεων για κάθε λέξη κατά τη διάρκεια της εκπαίδευσης. Σε αυτή τη δουλειά, κάθε λέξη παριστάνεται ως ένα Γκαουσιανό μείγμα αντί μιας διανυσματικής αναπαράστησης, όπου κάθε Γκαουσιανό στοιχείο αντιπροσωπεύει μια έννοια της λέξης. Σε μια πιο πρόσφατη εργασία, οι Amiri et al. [2016] έκαναν χρήση των αυτόματων κωδικοποιητών (autoencoders) για να αντιστοιχίσουν κάθε λέξη σε αναπαραστάσεις ανάλογα με τα συμφραζόμενά τους, ενώ οι Lee and Chen [2017], Guo et al. [2014] υλοποίησαν εννοιολογική αποσαφήνιση λέξεων μέσω της ενισχυτικής μάθησης.

Όλες οι παραπάνω μέθοδοι έκαναν χρήση των συμφραζόμενων για κάθε εμφάνιση μιας λέξης χωρίς να ληφθεί υπόψη η σχετική σειρά των λέξεων στο παράθυρο συμφραζόμενων. Οι Zheng et al. [2017] θεώρησαν ότι αυτή η παράλειψη υποβαθμίζει την ποιότητα των πολλαπλών λεξιλογικών αναπαραστάσεων που προκύπτουν από μεθόδους που βασίζονται σε ομάδες και σημείωσαν ότι η σειρά των λέξεων που εμφανίζονται στα συμφραζόμενα διαδραματίζει κάποιο ρόλο στην κατανόηση της έννοιας που αντιστοιχεί στη λέξη ενδιαφέροντος. Για να αντιμετωπίσουν αυτό το ζήτημα, χρησιμοποίησαν ένα νευρωνικο δίκτυο, το οποίο ονομάστηκε CSV (Context-Specific Vector), και παρήγαγε αναπαραστάσεις λέξεων και συμφραζόμενων. Η προτεινόμενη αρχιτεκτονική του νευρικού δικτύου περιλάμβανε ένα συνελικτικό επίπεδο (convolutional layer) που σχεδιάστηκε για να παράγει παραστάσεις συμφραζόμενων που αντικατοπτρίζουν τη σειρά των στοιχείων τους. Οι προκύπτουσες αναπαραστάσεις χρησιμοποιήθηκαν για τη δημιουργία πολλαπλών αναπαραστάσεων για κάθε λέξη.

Ένας άλλος ορισμός των *συμφραζόμενων* στοιχείων περιγράφηκε από τους Liu et al. [2015b], οι οποίοι αντιμετώπισαν τα συμφραζόμενα ως έναν θεματικό τομέα. Βασιζόμενοι στην παρατήρηση ότι οι πολυσήμαντες λέξεις αλλάζουν συνήθως το νόημά τους όταν συναντώνται σε διαφορετικές θεματικές περιοχές, χρησιμοποίησαν τεχνικές θεματικής μοντελοποίησης προ-

κειμένου να μάθουν πολλαπλές αναπαραστάσεις για κάθε λέξη. Συγκεκριμένα, ο αλγόριθμος Latent Dirichlet Allocation (LDA) εφαρμόστηκε στο μοντέλο skip-gram προκειμένου να εξαχθεί η κατανομή μίας λέξης πάνω σε πιθανές θεματικές περιοχές· εν συνεχεία χρησιμοποιήθηκε για την εξαγωγή θεματικών αναπαραστάσεων λέξεων (topical embeddings). Σε μια πιο πρόσφατη εργασία, ο LDA χρησιμοποιήθηκε για να επάγει τα βάρη κάθε θέματος. Τα βάρη αυτά χρησιμοποιήθηκαν για να ορίσουν ένα μείγμα διανυσμάτων για κάθε λέξη ενδιαφέροντος που προέβλεπε τις αντίστοιχες συμφραζόμενες λέξεις [Nguyen et al., 2017]. Επιπλέον, οι Wu and Giles [2015] χρησιμοποιώντας άρθρα εξαγόμενα από τη Wikipedia υπέθεσαν ότι οι εμφανίσεις των λέξεων που συνυπάρχουν στο ίδιο άρθρο αναφέρονται σε μία έννοια της λέξης. Συγκεκριμένα, οι εννοιολογικές αναπαραστάσεις των λέξεων εξήχθησαν αφού πρώτα οι σελίδες της Wikipedia ομαδοποιήθηκαν με βάση το καθολικό και το τοπικό πλαίσιο συμφραζομένων των λέξεων ενδιαφέροντος.

Μια πιθανοτική προσέγγιση ακολουθήθηκε έπειτα από τους Li and Jurafsky [2015], οι οποίοι θεώρησαν ότι μια λέξη θα πρέπει να συνδέεται με μια νέα έννοια όταν τα συμφραζόμενά της υποδηλώνουν ότι διαφέρει αρκετά από τις αρχικές έννοιες που της έχουν αποδοθεί. Η υπόθεση αυτή τους οδήγησε στη χρήση της μεθόδου Chinese Restaurant Process, με βάση την οποία κάθε εμφάνιση μιας λέξης αντιστοιχεί σε έναν πελάτη, ενώ κάθε τραπέζι (table) αντιστοιχεί σε μια έννοια της λέξης. Υπό αυτούς τους όρους, μια νέα εμφάνιση λέξης θα μπορούσε είτε να καθίσει σε ένα κατειλημμένο τραπέζι (που έχει εκχωρηθεί σε μια υπάρχουσα έννοια της λέξης), είτε να επιλέξει ένα μη κατειλημμένο τραπέζι για να καθίσει (οπότε το τραπέζι εκχωρείται σε μια νέα έννοια της λέξης).

Τέλος, οι Guo et al. [2014] πρότειναν ένα διαφορετικό θεωρητικό πλαίσιο για να δημιουργήσουν πολλαπλές εννοιολογικές αναπαραστάσεις για κάθε λέξη, χρησιμοποιώντας ένα αναδρομικό νευρωνικό δίκτυο. Αντί να χρησιμοποιήσουν την πληροφορία που περιέχεται στα συμφραζόμενα των λέξεων ενδιαφέροντος ως ένδειξη των πιθανών εννοιών τους, χρησιμοποίησαν δίγλωσσους πόρους, υποθέτοντας ότι μια λέξη με πολλαπλές έννοιες μπορεί να έχει διαφορετική μετάφραση σε μια άλλη γλώσσα.

2.3.2 Μέθοδοι με Επίβλεψη

Οι μέθοδοι χωρίς επίβλεψη που έχουν εξεταστεί μέχρι στιγμής επιχειρούν να ανακαλύψουν την πολυσημική φύση των λέξεων μέσω της δημιουργίας πολλαπλών αναπαραστάσεων από ακατέργαστες πληροφορίες συμφραζόμενων που εξάγονται από μεγάλα κείμενα. Οι πιο πρόσφατες τεχνικές που επιτυγχάνουν state-of-the-art επιδόσεις βασίζονται σε επιβλεπόμενες μεθόδους. Γενικά, αυτές οι προσεγγίσεις χρησιμοποιούν μια ελλιπή βάση δεδομένων (knowledge base) μαζί με ένα μεγάλο σύνολο κειμένων και προσπαθούν να χρησιμοποιήσουν τη βάση αυτή ως πρότερη γνώση στο πρόβλημα. Ο ευρύτερα διαδεδομένος κατάλογος που περιλαμβάνει εννοιολογική πληροφορία λέξεων και χρησιμοποιείται ως βοηθητική γνώση για την εξαγωγή πολλαπλών αναπαραστάσεων είναι το WordNet. Τα ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα ομαδοποιούνται σε σύνολα γνωστικών συνωνύμων που ονομάζονται *synsets*, όπου το καθένα εκφράζει μια ξεχωριστή έννοια όπως περιγράφεται από τους Miller et al. [1990].

Οι Chen et al. [2014] χρησιμοποίησαν τους ορισμούς που παρέχονται για κάθε λέξη από το WordNet προκειμένου να δημιουργήσουν διανυσματικές αναπαραστάσεις για κάθε μία από τις έννοιες της. Χρησιμοποιώντας αυτές τις αναπαραστάσεις ως αρχικές εκτιμήσεις μαζί με μονές αναπαραστάσεις για κάθε λέξη, δημιούργησαν πιο ακριβείς εννοιολογικές διανυσματικές

αναπαραστάσεις λέξεων κάνοντας χρήση αλγορίθμων αποσαφήνισης λέξεων (word sense disambiguation algorithm). Γνωρίζοντας τις αποσαφηνισμένες λέξεις, τροποποίησαν το μοντέλο skip-gram προκειμένου να εκπαιδεύσουν διανυσματικές αναπαραστάσεις λέξεων και εννοιών. Αργότερα, οι Iacobacci et al. [2015] χρησιμοποίησαν τη βάση BabelNet η οποία αποτελεί μια εμπλουτισμένη βάση δεδομένων της WordNet. Χρησιμοποιώντας τη γνώση της βάσης αυτής, εξήγαγαν ένα επισημειωμένο σύνολο κειμένων (annotated corpus), χρησιμοποιώντας έναν αλγόριθμο αποσαφήνισης λέξεων. Οι εννοιολογικές διανυσματικές αναπαραστάσεις των λέξεων εξήχθησαν μέσω της εκπαίδευσης του μοντέλου skip-gram πάνω στο επισημειωμένο σύνολο κειμένων.

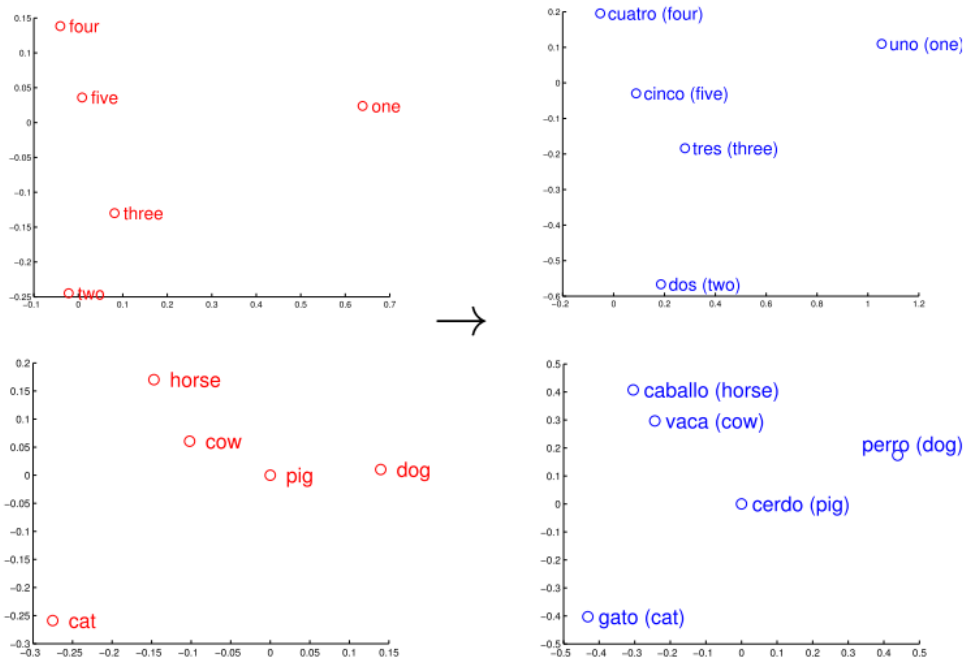
Μια άλλη προσέγγιση βασισμένη σε πρότερη γνώση των πιθανών εννοιών μιας λέξης εισήχθη από τους Rothe and Schütze [2015] οι οποίοι θεώρησαν ότι οι διανυσματικές αναπαραστάσεις των λέξεων προκύπτουν ως άθροισμα των αναπαραστάσεων που αντιστοιχούν στις λεξικές μονάδες των synsets τους. Η υλοποίηση της παραπάνω θεωρητικής υπόθεσης επιτυγχάνεται με χρήση αλγεβρικών πράξεων ανάμεσα στις διανυσματικές αναπαραστάσεις των αντίστοιχων λέξεων. Συγκεκριμένα, αναπαραστάσεις οι οποίες είχαν προεκπαιδευτεί (pre-trained embeddings) επεκτάθηκαν έτσι ώστε να αναπαραστήσουν lexemes και synsets, όπως αυτά ορίζονται στη βάση WordNet. Πρόσφατα, οι Pilehvar and Collier [2016] αποσαφήνισαν προεκπαιδευμένες διανυσματικές αναπαραστάσεις λέξεων χρησιμοποιώντας επίσης τη WordNet. Έπειτα από τη σύνδεση αυτών των προ-εκπαιδευμένων αναπαραστάσεων στη βάση αυτή, εξήγαγαν έναν κατάλογο σημασιολογικά biased λέξεων ως προς την πολυσήμαντη λέξη ενδιαφέροντος. Λαμβάνοντας υπόψη τις αναπαραστάσεις τόσο των biased λέξεων όσο και της λέξης ενδιαφέροντος, εξήγαγαν μια αναπαράσταση που αντιστοιχίζονταν σε μία συγκεκριμένη έννοια της λέξης ενδιαφέροντος αναζητώντας το biased διάνυσμα που είχε την ελάχιστη απόσταση από αυτή.

2.4 Μετασχηματισμοί σημασιολογικών χώρων

Όπως αναφέρθηκε προηγουμένως, οι αρχιτεκτονικές που χρησιμποποιούν νευρωνικα δίκτυα, όπως το Word2Vec, έχουν γίνει πολύ δημοφιλείς, καθώς έχει αποδειχθεί οτι υπερέχουν έναντι των παραδοσιακών μεθόδων που βασίζονται σε μετρήσεις. Πολλοί αποδίδουν αυτή την υπεροχή στη φυσική εξήγηση στην οποία βασίζεται η χρήση των νευρωνικών δικτύων, η οποία δεν συναντάται στις μεθόδους που χρησιμοποιούν ακατέργαστα χαρακτηριστικά συνεμφάνισης λέξεων. Ένα από τα κύρια χαρακτηριστικά των Κατανεμημένων Σημασιολογικών Μοντέλων που βασίζονται σε μεθόδους πρόβλεψης είναι ότι δημιουργούν σημασιολογικούς χώρους που δεν είναι ευθυγραμμισμένοι ως προς ένα σταθερό σύστημα συντεταγμένων, εξαιτίας της μη ντετερμινιστικής φύσης τους. Συγκεκριμένα, αυτό σημαίνει ότι αν τρέξουμε τον ίδιο αλγόριθμο δύο φορές για το ίδιο σύνολο δεδομένων, υπάρχει μεγάλη πιθανότητα οι σημασιολογικοί χώροι που προκύπτουν να έχουν διαφορετική καθολική γεωμετρική δομή. Για το λόγο αυτό, το πρόβλημα του μετασχηματισμού ανάμεσα σε σημασιολογικούς χώρους έχει προσελκύσει το ενδιαφέρον της επιστημονικής κοινότητας, καθώς επιτρέπει τη σύγκριση ανάμεσα σε κατανεμημένες αναπαραστάσεις λέξεων που έχουν εξαχθεί από διαφορετικά σύνολα δεδομένων.

Η πιο δημοφιλής εφαρμογή των μετασχηματισμών ανάμεσα σε σημασιολογικούς χώρους είναι η μηχανική μετάφραση, στόχος της οποίας είναι να αυτοματοποιήσει τη διαδικασία δημιουργίας μεγάλων λεξικών ξεκινώντας από λίγα δίγλωσσα δεδομένα. Οι Mikolov et al.

[2013a] ήταν οι πρώτοι που εισήγαγαν τέτοιους σημασιολογικούς μετασχηματισμούς προκειμένου να προβλέψουν τις μεταφράσεις μεταξύ αγγλικών και ισπανικών λέξεων. Μετά την εκμάθηση των διανυσματικών αναπαραστάσεων για τις λέξεις και των δύο γλωσσών χρησιμοποιώντας το μοντέλο Word2Vec, πρότειναν μια γραμμική αντιστοίχιση μεταξύ των δύο χώρων που αντιπροσωπεύουν τους σημασιολογικούς χώρους των γλωσσών. Στο ευθυγραμμισμένο σύστημα συντεταγμένων η σωστή μετάφραση μιας λέξης αναμένεται να βρίσκεται κοντά σε αυτή τη λέξη.



Σχήμα 2.4: Προβολές των διανυσματικών αναπαραστάσεων των λέξεων που αντιστοιχίζονται σε αριθμούς και ζώα, στα αγγλικά (αριστερά) και τα ισπανικά (δεξιά) χρησιμοποιώντας το Principal Component Analysis [Mikolov et al., 2013a].

Όπως σημείωσαν, το βασικό τους κίνητρο ήταν ότι όλες οι γλώσσες έχουν παρόμοια γεωμετρικά χαρακτηριστικά, καθώς μοιράζονται τις έννοιες που υπάρχουν στον πραγματικό κόσμο. Οι επόμενες εργασίες που αφορούσαν τη μηχανική μετάφραση επικεντρώθηκαν στις ιδιότητες του πίνακα μετασχηματισμού [Xing et al., 2015], καθώς και στις ιδιότητες των απεικονιζόμενων διανυσματικών αναπαραστάσεων. Συγκεκριμένα, οι Dinu and Baroni [2014] έδειξαν ότι η περιοχή των κοντινότερων γειτόνων (nearest neighbors) των απεικονιζόμενων αναπαραστάσεων είναι πολύ "μολυσμένη" από τα λεγόμενα *hubs*, που αντιστοιχούν σε διανύσματα τα οποία τείνουν να είναι δημοφιλείς κοντινότεροι γείτονες πολλών άλλων διανυσμάτων.

Μια άλλη εφαρμογή του μετασχηματισμού ανάμεσα σε σημασιολογικούς χώρους μελετήθηκε αργότερα από τους Tan et al. [2015], οι οποίοι προσπάθησαν να διερευνήσουν τις σημασιολογικές διαφορές λέξεων μεταξύ των ανεπίσημων αγγλικών που χρησιμοποιούνται στα μέσα κοινωνικής δικτύωσης (Twitter corpus) και τα επίσημα αγγλικά που εμφανίζονται σε καλώς οργανωμένα κείμενα (Wikipedia corpus). Προκειμένου να ευθυγραμμιστούν οι δύο σημασιολογικοί χώροι, υπέθεσαν ότι υπάρχει μια απεικόνιση μεταξύ των πιο συνηθισμένων λέξεων των δύο χώρων. Μετά την απεικόνιση των δύο γλωσσών σε έναν κοινό χώρο, χρη-

σιμοποίησαν μια ομαλοποίηση των αποστάσεων των λέξεων με βάση τη συχνότητα εμφάνισής τους στα αντίστοιχα corpora και τελικά χρησιμοποίησαν αυτές τις αποστάσεις για να βρουν λέξεις των οποίων η χρήση διαφοροποιείται αισθητά ανάμεσα στα δύο corpora.

Η σημασιολογική εξέλιξη της έννοιας των λέξεων μπορεί να συλληφθεί σε μεγάλης κλίμακας corpora που αναφέρονται σε διαφορετικές χρονικές περιόδους. Οι Hamilton et al. [2016] δημιούργησαν διαχρονικά embeddings, κατασκευάζοντας αρχικά embeddings για κάθε χρονική περίοδο. Στη συνέχεια μαθαίνοντας διαδοχικούς γραμμικούς περιστροφικούς πίνακες απεικόνισαν τους διανυσματικούς χώρους που αντιστοιχίζονταν σε διαδοχικές ιστορικές περιόδους έτσι ώστε να ανιχνεύσουν τη σημασιολογική εξέλιξη των λέξεων στο πέρασμα των χρόνων. Η σχετικά υψηλή διάσταση των διαχρονικών embeddings (20, 30, 200 ...) αποτελεί πρόκληση καθώς συνήθως τα διανύσματα δεν βρίσκονται στις 2 ή 3 διαστάσεις που μπορούν εύκολα να ερμηνευτούν από τον άνθρωπο. Για τον λόγο αυτό, πραγματοποιούνται συνήθως τεχνικές μείωσης διαστάσεων, προκειμένου να απεικονιστεί η τροχιά που ακολουθεί μια λέξη με την πάροδο του χρόνου σε έναν χώρο 2 διαστάσεων.

Το Σχήμα 2.5 απεικονίζει παραδείγματα τροχιών που αντιστοιχίζονται στη χρονική σημασιολογική εξέλιξη λέξεων. Συγκρίνοντας τη σχετική θέση των λέξεων με τους προσωρινούς πλησιέστερους γείτονές τους μπορούμε να παρακολουθήσουμε ενδιαφέρουσες σημασιολογικές μετατοπίσεις του νοήματος των λέξεων, οι οποίες θα μπορούσαν επίσης να αντανακλούν την πολιτισμική τους εξέλιξη. Για παράδειγμα, η λέξη "gay" μετατοπίστηκε από το να σημαίνει "χαρούμενος" ή "εύθραυστος" στο να αναφέρεται στην ομοφυλοφιλία. Στις αρχές του 20ου αιώνα η λέξη "broadcast" αναφερόταν στη "σπορά", ωστόσο, με την άνοδο της τηλεόρασης και του ραδιοφώνου η σημασία της μετατοπίστηκε στο να σημαίνει "σήμα μετάδοσης". Η λέξη "awful" μετατοπίστηκε από το νόημα "γεμάτος δέος" στο νόημα "'τρομερός" όπως αναφέρεται στους Hamilton et al. [2016].

Σχήμα 2.5: Δισδιάστατη απεικόνιση της σημασιολογικής αλλαγής τριών αγγλικών λέξεων.

Πρόσφατα οι Prokhorov et al. [2017] εφάρμοσαν μετασχηματισμούς ανάμεσα σε σημασιολογικούς χώρους σε μια προσπάθεια εμπλουτισμού ενός υπάρχοντος λεξιλογίου με σπάνιες λέξεις. Αυτό που είναι πολύ ενδιαφέρον στην προσέγγισή τους είναι ότι οι πληροφορίες που προέρχονται από συλλογές κειμένων θα μπορούσαν να χρησιμοποιηθούν για να συμπληρωθούν τα ελλείποντα τμήματα των γνωστικών βάσεων, και αντίστροφα. Προκειμένου να πετύχουν τα παραπάνω, δημιούργησαν μια απεικόνιση μεταξύ ενός κατανεμημένου σημασιολογικού χώρου και μιας λεξικής οντολογίας χρησιμοποιώντας *σημασιολογικές γέφυρες* (semantic bridges)

μονοσήμαντων λέξεων.

Μέθοδοι Απεικόνισης

Στο σημείο αυτό παραθέτουμε τη βασική ορολογία που απαιτείται για να εξηγήσουμε τις πιο δημοφιλείς μεθόδους μετασχηματισμών ανάμεσα σε σημασιολογικούς χώρους που αναφέρονται στη βιβλιογραφία. Αρχικά, θεωρούμε ότι οι πίνακες X και Y περιέχουν τις διανυσματικές αναπαραστάσεις των λέξεων στη γλώσσα προέλευσης (source language) και στη γλώσσα προορισμού (target language), αντίστοιχα. Η i-οστή στήλη του πίνακα X περιέχει την κατανεμημένη αναπαράσταση του διανύσματος $x_i \in \mathbb{R}^d$ της λέξης i στον χώρο προέλευσης, ενώ $y_i \in \mathbb{R}^d$ είναι η ισοδύναμη κατανεμημένη αναπαράσταση της λέξης στον χώρο προορισμού. Στόχος μας είναι να βρούμε έναν πίνακα μετασχηματισμού $W \in \mathbb{R}^{d \times d}$, ο οποίος απεικονίζει τη γλώσσα προέλευσης στη γλώσσα προορισμού, έτσι ώστε ο WX να είναι όσο το δυνατόν πιο κοντά στον Y. Όπως περιγράφεται από τους Artetxe et al. [2018], αυτός ο πίνακας μετασχηματισμού θα μπορούσε να υπολογιστεί μέσω γραμμικής παρεμβολής (Linear), κανονικών μεθόδων (Canonical Correlation Analysis) ή ορθογώνιων (Orthogonal) μεθόδων.

- Οι **Γραμμικές** μέθοδοι χρησιμοποιήθηκαν από τους Mikolov et al. [2013a] οι οποίοι υπολόγισαν μια γραμμική απεικόνιση, αποτελώντας την πρώτη προσπάθεια ευθυγράμμισης σημασιολογικών χώρων με εφαρμογή στην μηχανική μετάφραση. Συγκεκριμένα, χρησιμοποίησαν τη συνάρτηση απώλειας ελαχίστων τετραγώνων (least square objective function), η οποία ελαχιστοποιεί το άθροισμα των τετραγώνων των Ευκλείδειων αποστάσεων ανάμεσα στις διανυσματικές αναπαραστάσεις μεταφρασμένων ζευγών λέξεων που ανήκουν σε δύο γλώσσες ενδιαφέροντος, χωρίς να επιβάλλει κάποιον περιορισμό στον πίνακα μετασχηματισμού. Αυτό το πρόβλημα (γνωστό και ως Ordinary Least Squares) έχει μια λύση κλειστής μορφής όπως υποδεικνύεται στην εξίσωση 2.3.

$$W = \arg\min_W \|WX - Y\|_F = (X^t X)^{-1} X^t Y. \tag{2.3}$$

  Λίγα χρόνια αργότερα, οι Dinu and Baroni [2014] ενσωμάτωσαν έναν L2-regularization όρο στην παραπάνω εξίσωση.

- Οι **Ορθογώνιες** μέθοδοι προτάθηκαν αρχικά από τους Xing et al. [2015] οι οποίοι παρατήρησαν ότι τόσο τα διανύσματα του χώρου προέλευσης όσο και τα διανύσματα του χώρου προορισμού πρέπει να παραμείνουν ορθογώνια κατά τη φάση εκμάθησης του πίνακα μετασχηματισμού. Σημείωσαν επίσης ότι η κανονικοποίηση (normalization) είναι ένα κρίσιμο χαρακτηριστικό που πρέπει να διατηρούν οι ευθυγραμμισμένες διανυσματικές αναπαραστάσεις, καθώς εξασφαλίζει ότι το εσωτερικό γινόμενο (dot product) δύο διανυσμάτων αντιστοιχεί στην ομοιότητα συνημιτόνου (cosine similarity) τους, η οποία αποτελεί την πιο ευρέως χρησιμοποιούμενη μετρική σημασιολογικής ομοιότητας ανάμεσα σε embeddings. Για το λόγο αυτό, η απεικόνιση του χώρου προέλευσης στον χώρο προορισμού λαμβάνεται μέσω της επίλυσης του παρακάτω προβλήματος βελτιστοποίησης που ενσωματώνει έναν περιορισμό στον πίνακα W όπως περιγράφεται στην Εξίσωση 2.4.

$$W = \arg\min_W \|WX - Y\|_F, \quad \text{s.t.} \quad WW^T = \mathbb{I}. \tag{2.4}$$

Από μαθηματική σκοπιά, το παραπάνω πρόβλημα είναι γνωστό ως το Orthogonal Procrustes Problem και έχει μια λύση κλειστής μορφής. Ο βέλτιστος πίνακας μετασχηματισμού W ανακτάται από το UV^T, όπου οι U και V αποκτώνται μέσω της τεχνικής Singular Value Dicomposition (ίση με $(U\Sigma V^T)$) του YX^T . Για μια λεπτομερέστερη ανασκόπηση του προβλήματος παραπέμπουμε τον αναγνώστη στη μελέτη του Schönemann [1966].

- Οι **Canonical** μέθοδοι από την άλλη πλευρά, υπολογίζουν πρώτα δύο διακριτές γραμμικές απεικονίσεις M_1 και M_2, όπου ο στόχος είναι η μεγιστοποίηση της συσχέτισης μεταξύ των διαστάσεων των δύο προβαλλομένων πινάκων M_1X και M_2Y. Μετά τον υπολογισμό των δύο μεατασχηματισμών, ο πίνακας μετασχηματισμού W ανακτάται μέσω μιας απλής αλγεβρικής διαδικασίας όπως σημειώνεται στην εξίσωση 2.4. Οι Faruqui and Dyer [2014] ήταν οι πρώτοι που χρησιμοποίησαν την ανάλυση αυτή, γνωστή και ως Canonical Correlation Analysis, για να απεικονίσουν δύο σημασιολογικούς χώρους σε έναν κοινό διανυσματικό χώρο. Η μέθοδος αυτή οδηγεί σε παρόμοια αποτελέσματα με την ορθογώνια απεικόνιση.

$$W = M_1^{-1}M_2, \quad \text{where} \quad M_1, M_2 = \underset{M_1,M_2}{\operatorname{argmax}} \operatorname{cov}(M_1X, M_2Y) \tag{2.5}$$

Κεφάλαιο 3

Υπόβαθρο

Στις επόμενες υποενότητες εξηγούμε τους βασικούς αλγορίθμους που χρησιμοποιούνται από τα μοντέλα που προτείνουμε σε αυτή τη διπλωματική εργασία. Συγκεκριμένα, αναλύουμε: το μοντέλο Word2Vec, τον αλγόριθμο Latent Dirichlet Allocation (LDA) και την ιεραρχική μέθοδο ομαδοποίησης δεδομένων (hierarchical clustering).

3.1 Μοντέλο Word2Vec

Το μοντέλο Word2Vec εισήχθη από τους Mikolov et al. [2013b] και αποτελεί έναν από τους πιο ευρέως χρησιμοποιούμενους αλγορίθμους για τη δημιουργία διανυσματικών αναπαραστάσεων λέξεων υψηλής ποιότητας, οι οποίες είναι γνωστές ως *embeddings*. Εξετάζοντας το μοντέλο από μία γενική σκοπιά, το Word2Vec παίρνει ως είσοδο ένα σύνολο κειμένων, χρησιμοποιεί τα στατιστικά χαρακτηριστικά της εισόδου και ενσωματώνει (embed) κάθε λέξη σε έναν διανυσματικό χώρο με τέτοιο τρόπο ώστε να μπορούν να εξαχθούν σημαντικές σημασιολογικές σχέσεις μεταξύ τους, μέσω απλών μαθηματικών πράξεων ανάμεσα στα διανύσματα που τους έχουν αντιστοιχιστεί. Αν και οι πυκνές (dense) αναπαραστάσεις που παράγει αυτό το μοντέλο συχνά χρησιμοποιούνται ως κύριος πυρήνας συστημάτων Βαθιάς Μάθησης (Deep Learing) στην Επεξεργασία Φυσικής Γλώσσας, το Word2Vec εντάσσεται στην κατηγορία των ρηχών νευρωνικών δικτύων (shallow neural networks) καθώς αποτελείται από μόλις ένα κρυφό επίπεδο. Ως αποτέλεσμα, η απλή αυτή δομή καθιστά το Word2vVec ένα πολύ αποδοτικό μοντέλο από υπολογιστικής απόψεως.

Σε αυτό το σημείο, ορίζουμε μια πρότυπη πρόταση που θα χρησιμεύσει ως το εισαχθέν corpus μας για να περιγράψουμε τους δύο αλγόριθμους που μπορούν να χρησιμοποιηθούν από το Word2Vec για να μάθουμε διανυσματικές αναπαραστάσεις λέξεων: το μοντέλο Continuous Bag of Words και το μοντέλο Skip-gram. Ας θεωρήσουμε λοιπόν οτι το εισαχθέν corpus μας αποτελείται από την πρόταση D:

$$D = \{\textit{former president Obama speaks to the media in Washington about terrorism}\}$$

Προκειμένου να εκπαιδεύσουμε το νευρωνικό μας δίκτυο, θα πρέπει πρώτα να εξάγαγουμε ένα λεξικό (vocabulary) αποτελούμενο από λέξεις για τις οποίες θέλουμε να δημιουργήσουμε διανυσματικές αναπαραστάσεις. Στην περίπτωσή μας το λεξικό ορίζεται ως εξής:

$$V = \{\textit{former, president, Obama, speaks, to, the, media, in, Washington, about, terrorism}\}$$

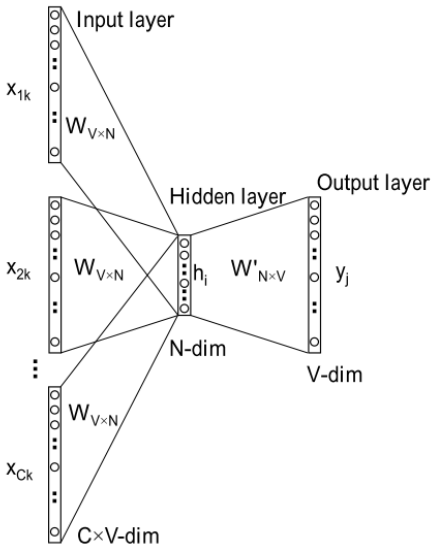
Με βάση αυτό το λεξικό δημιουργούμε μαθηματικές αναπαραστάσεις λέξεων, προκειμένου να τις τροφοδοτήσουμε στο επίπεδο εισόδου του νευρικού μας δικτύου. Για το λόγο αυτό, κατασκευάζουμε για κάθε λέξη μια one-hot αναπαράσταση, η οποία αντιστοιχεί σε ένα αραιό διάνυσμα με μέγεθος ίσο με το πλήθος των στοιχείων του λεξικού ($|V|$). Η αναπαράσταση αυτή έχει μηδενικά στοιχεία σε όλες τις συντεταγμένες της εκτός από ένα στοιχείο το οποίο τίθεται ίσο με τη μονάδα και τοποθετείται στη συντεταγμένη εκείνη που αντικατοπτρίζει τη σχετική θέση της λέξης στο σύνολο του λεξικού. Στο παράδειγμά μας, οι one-hot αναπαραστάσεις των λέξεων είναι οι εξής:

$[1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = \textit{former}$
$[0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = \textit{president}$
$[0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = \textit{Obama}$
$[0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = \textit{speaks}$
$[0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = \textit{to}$
$[0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0] = \textit{the}$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0] = \textit{media}$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0] = \textit{in}$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0] = \textit{Washington}$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0] = \textit{about}$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1] = \textit{terrorism}$

3.1.1 Το μοντέλο Continuous Bag of Words

Το μοντέλο Continuous Bag of Words (CBOW) προβλέπει μια λέξη ενδιαφέροντος χρησιμοποιώντας τις συμφραζόμενες λέξεις που βρίσκονται σε ένα σταθερό παράθυρο (context window) γύρω από αυτή. Η βασική ιδέα του αλγορίθμου πηγάζει από την *υπόθεση της κατανεμημένης έννοιας των λέξεων*, καθώς προσπαθεί να βρει στοιχεία για τη σημασία της λέξης κοιτώντας τα συμφραζόμενα της με σκοπό να μάθει μια πυκνή διανυσματική αναπαράσταση της. Στη συνέχεια ακολουθούμε την ορολογία και τα σύμβολα που χρησιμοποιεί ο Rong [2014] στη λεπτομερή ανασκόπηση του Word2Vec. Η δομή του μοντέλου CBOW παρουσιάζεται στο Σχήμα 3.1.

Το μοντέλο CBOW του παρακάτω σχήματος λαμβάνει ως είσοδο C one-hot διανύσματα, όπου το καθένα από αυτά έχει διάσταση $|V|$, και τα οποία αντιστοιχούν στις συμφραζόμενες αναπαραστάσεις της εκάστοτε λέξης ενδιαφέροντος. Το πλήθος των γειτονικών λέξεων που πρέπει να ληφθούν υπόψη αποτελεί σχεδιαστική απόφαση που καθορίζεται από τον δημιουργό του νευρωνικού δικτύου. Στο πρώτο στάδιο, πολλαπλασιάζουμε κάθε one-hot διάνυσμα με τον πίνακα $W \in \mathbb{R}^{V \times N}$, ο οποίος αντιπροσωπεύει τον πίνακα βαρών ανάμεσα στο επίπεδο εισόδου και το πρώτο κρυφό επίπεδο (hidden layer). Σημειώστε επίσης ότι η διάσταση του κρυφού επιπέδου είναι πολύ μικρότερη από αυτή των διανυσμάτων εισόδου και συμπίπτει με τη διάσταση των παραγόμενων embeddings. Δεδομένου ότι τα διανύσματα εισόδου είναι one-hot κωδικοποιήσεις, παρατηρούμε ότι η πραγματική λειτουργία του επιπέδου εισόδου για τη k-οστή συμφραζόμενη λέξη είναι η προώθηση της k-οστής γραμμής του πίνακα βαρών στο κρυφό επίπεδο. Στη συνέχεια, τα διανύσματα που συνδέονται με κάθε μια από τις συμφραζόμενες λέξεις προστίθενται στο κρυφό διάνυσμα (hidden vector). Έπειτα, το κρυφό διάνυσμα πολλα-



Σχήμα 3.1: Το Continuous Bag of Word μοντέλο όπως παρουσιάζεται από τον Rong [2014].

πλασιάζεται με τον πίνακα βαρών $W' \in \mathbb{R}^{N \times V}$ που αντιπροσωπεύει τη μετάβαση από το κρυφό επίπεδο προς το επίπεδο εξοδου και τελικά περνά από τη soft-max συνάρτηση προκειμένου να υπολογιστει το επίπεδο εξόδου του δικτύου. Η πρόβλεψη της λέξης ενδιαφέροντος πραγματοποιείται συγκρίνοντας τη θέση στην οποία εμφανίζεται η μέγιστη τιμή του διανύσματος εξόδου με την one-hot κωδικοποίηση του λεξικού στην είσοδο του δικτύου. Κατά τη διάρκεια της εκπαίδευσης του νευρωνικού μας δικτύου, χρησιμοποιούμε τη σωστή έξοδο (λέξη ενδιαφέροντος) έτσι ώστε να υπολογίσουμε το σφάλμα της πρόβλεψής μας και να το διαδώσουμε προς τα πίσω ενημερώνοντας τους πίνακες βάρων μέχρι να ικανοποιηθεί ένα κριτήριο τερματισμού.

Συνεχίζοντας με την προηγούμενη πρότυπη πρόταση, δίνουμε ένα αριθμητικό παράδειγμα της πρόβλεψης εξόδου. Ας υποθέσουμε ότι θέλουμε να προβλέψουμε τη λέξη *Washington*, λαμβάνοντας υπόψη τις συμφραζόμενες λέξεις *media*, *in*, *about* και *terrorism*. Θεωρήστε ότι ο πίνακας βαρών που συνδέει το επίπεδο εισόδου της λέξης ενδιαφέροντος *media* με το κρυφό επίπεδο, ισούται με:

$$W = \begin{pmatrix} 0.1 & 0.0 & 0.1 \\ 0.0 & 0.3 & 0.7 \\ 0.9 & 0.1 & 0.6 \\ 0.3 & 0.3 & 0.6 \\ 0.2 & 0.6 & 0.9 \\ 0.2 & 0.3 & 0.7 \\ 0.8 & 0.9 & 0.1 \\ 0.1 & 0.3 & 0.5 \\ 0.1 & 0.7 & 0.1 \\ 0.5 & 0.5 & 0.1 \end{pmatrix}$$

Η σκιασμένη σειρά αντιστοιχεί στην πυκνή αναπαράσταση της λέξης *media* η οποία έπειτα προωθείται στο κρυφό επίπεδο έτσι ώστε να αθροιστεί μαζί με τις πυκνές αναπαραστάσεις των άλλων συμφραζόμενων λέξεων (που εντάσσονται στο παράθυρο συμφραζόμενων). Στη

συνέχεια, υποθέτουμε ότι οι γραμμές του πίνακα βαρών της εισόδου, που αντιστοιχούν σε γειτονικές λέξεις, αθροίζονται στο κρυφό διάνυσμα:

$$h = \begin{bmatrix} 0.4 & 0.7 & 0.6 \end{bmatrix},$$

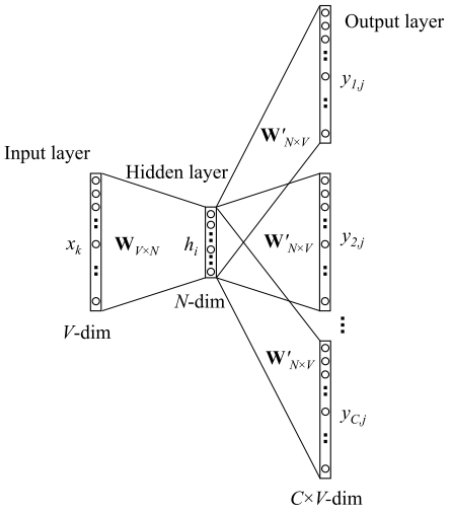
το οποίο στη συνέχεια μετατρέπεται σε ένα διάνυσμα V διαστάσεων στο στρώμα εξόδου, στο οποίο εφαρμόζεται η soft-max συνάρτηση:

$$y = \begin{bmatrix} 0.01 & 0.08 & 0.10 & 0.07 & \mathbf{0.13} & 0.10 & 0.09 & 0.09 & 0.07 & 0.11 \end{bmatrix}.$$

Η προβλεπόμενη λέξη στο παράδειγμά μας αντιστοιχεί στη λέξη 5 του λεξιλογίου μας (λέξη *to*).

3.1.2 Το μοντέλο Skip-gram

Το μοντέλο Skip-gram βασίζεται στην αντίθετη λογική από το μοντέλο CBOW. Σκοπός του είναι να προβλέψει τις συμφραζόμενες λέξεις οι οποίες βρίσκονται σε ένα σταθερό παράθυρο γύρω από μια λέξη ενδιαφέροντος, όταν η τελευταία λέξη είναι η μόνη παρεχόμενη πληροφορία. Όπως και το μοντέλο CBOW, εντάσσεται στην κατηγορία των ρηχών νευρωνικών δικτύων με ένα κρυφό επίπεδο όπως απεικονίζεται στο Σχήμα 3.2. Η είσοδός του είναι η one-hot διανυσματική αναπαράσταση της λέξης ενδιαφέροντος, η οποία προωθεί την αντίστοιχη πυκνή αναπαράσταση που δίνεται στον πίνακα βαρών της εισόδου στο κρυφό διάνυσμα.



Σχήμα 3.2: Το μοντέλο Skip-gram όπως παρουσιάζεται από τον Rong [2014].

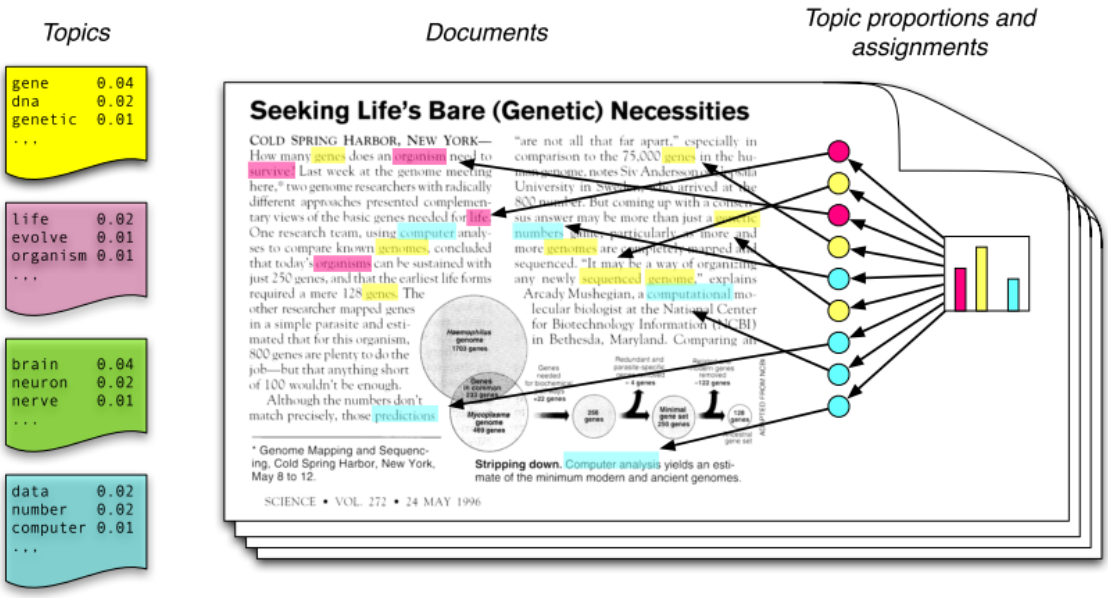
Έπειτα, το κρυφό διάνυσμα πολλαπλασιάζεται με κάθε πίνακα βαρών που αντιστοιχούν στο επίπεδο εξόδου $W' \in \mathbb{R}^{N\times V}$ και η soft-max συνάρτηση εφαρμόζεται για την παραγωγή των διανυσμάτων εξόδου όμοια με το μοντέλο CBOW. Η μόνη διαφορά είναι ότι αυτό το μοντέλο δημιουργεί C one-hot αναπαραστάσεις στην έξοδο $\{y_i\}_{i=1,2,..,C}$, κάθε μια από τις οποίες υποδηλώνει μια λέξη στο λεξικό ως προβλεπόμενο γείτονα της λέξης ενδιαφέροντος.

3.2 Latent Dirichlet Allocation

Ο αλγόριθμος Latent Dirichlet (LDA) που εισήχθη από τους Blei et al. [2003], είναι ένα πιθανοτικό αναπαραγωγικό (generative) μοντέλο ενός συνόλου κειμένων, το οποίο προσπαθεί να εντοπίσει τα κρυμμένα θέματα (topics) που βρίσκονται σε αυτά. Στη γλωσσολογία, η λέξη "θέμα" αναφέρεται σε ένα αφηρημένο σχήμα που δίνει μια γενική πληροφορία για το τι περιέχεται/συζητιέται σε ένα σύνολο λέξεων (πρόταση / έγγραφο). Τοποθετώντας αυτόν τον ορισμό σε ένα μαθηματικό πλαίσιο, θα μπορούσαμε να φανταστούμε ότι στις εφαρμογές της Επεξεργασίας Φυσικού Λόγου ένα "θέμα" περιγράφεται ως ένα σύνολο λέξεων οι οποίες συναντώνται συχνά μαζί ή χρησιμοποιώντας στατιστικούς όρους θα μπορούσε να αποδοθεί ως μια κατανομή πάνω στο λεξικό του συνόλου κειμένων. Σημειώστε επίσης ότι με δεδομένη την κατανομή, μπορούμε να αποκτήσουμε το σύνολο των πιο σχετικών λέξεων του λεξικού σε σχέση με ένα συγκεκριμένο θέμα μέσω της εφαρμογής μιας μεθόδου κατωφλίου στην κατανομή (διατήρηση λέξεων με μεγάλη πιθανότητα).

3.2.1 Βασική Ιδέα

Η βασική ιδέα του LDA είναι ότι τα έγγραφα (σύνολα φράσεων) εκπροσωπούνται ως μείγματα θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μια κατανομή πάνω σε ένα λεξικό. Αυτή η υπόθεση υπονοεί ότι ένα έγγραφο δεν θα μπορούσε να αναλύει μόνο ένα θέμα, το οποίο φαίνεται λογικό, καθώς τα έγγραφα είναι μεγάλες οντότητες κειμένου. Για να αποκτήσουμε μια βαθύτερη διαίσθηση της βασικής ιδέας του LDA, ας εξετάσουμε το άρθρο "Seeking Life's Bare (Genetic) Necessities', καθώς και την κατανομή πιθανών θεμάτων σε αυτό, όπως παρουσιάζεται στο Σχήμα 3.3.



Σχήμα 3.3: Βασική ιδέα του LDA. Το παρόν παράδειγμα αναφέρεται στους Blei [2012].

Όπως εξηγείται στους Blei [2012], το άρθρο αφορά την ανάλυση δεδομένων για τον προσ-

διορισμό του πλήθους των γονιδίων που χρειάζεται ένας οργανισμός για να επιβιώσει (από εξελικτικής σκοπιάς). Το άρθρο επισημάνθηκε με το χέρι για να δημιουργηθούν ομάδες λέξεων που θα μπορούσαν να αποδοθούν σε καθένα από τα θέματα για τα οποία γίνεται λόγος σε αυτό: τη γενετική, την ανάλυση δεδομένων και την εξελικτική βιολογία. Οι λέξεις σχετικά με την ανάλυση δεδομένων, όπως "computational" και "prediction" έχουν επισημανθεί με μπλε χρώμα, οι λέξεις που παραπέμπουν στην εξελικτική βιολογία όπως "survive" και "organism", έχουν επισημανθεί με ροζ χρώμα. Τέλος οι λέξεις που αναφέροντια στη γενετική, όπως "sequenced" και "genes" έχουν επισημανθεί με κίτρινο χρώμα. Ο LDA προσπαθεί να ενσωματώσει την παραπάνω ιδέα και να αυτοματοποιήσει τη διαδικασία της εκχώρησης θεμάτων σε έγγραφα και την εξαγωγή κατανομών γι΄ αυτά. Τα παραπάνω προϋποθέτουν ότι κάθε έγγραφο δημιουργείται ως εξής:

1. Επιλέγουμε τυχαία μια κατανομή πάνω στα θέματα (ιστόγραμμα στα δεξιά).
2. Για κάθε λέξη του εγγράφου:
   (α΄) Επιλέγουμε τυχαία ένα θέμα από την παραπάνω κατανομή (έγχρωμα κέρματα).
   (β΄) Επιλέγουμε τυχαία μια λέξη από την αντίστοιχη κατανομή τους προηγούμενου θέματος πάνω στο λεξικό.

3.2.2 Συμβολισμοί και Ορολογία

Καθώς πρόκειται να εισαγάγουμε την παραπάνω ιδέα σε ένα μαθηματικό πλαίσιο, θα πρέπει πρώτα να αναφερθούμε στη βασική ορολογία και τους συμβολισμούς που απαιτούνται για να περιγράψουμε γλωσσικούς όρους και έννοιες όπως "λέξεις", "έγγραφα", "σύνολο κειμένων", "θέμα", καθώς και τις κατανομές των εγγράφων πάνω στα θέματα (document-topic) και τις κατανομές των θεμάτων πάνω στο λεξικό (topic-word). Συγκεκριμένα, ακολουθώντας τους Blei et al. [2003] ορίζουμε:

- Η *λέξη* θεωρείται ως η βασική μονάδα των δεδομένων μας, και ορίζεται ως στοιχείο ενός λεξικού του οποίου τα στοιχεία έχουν θέσεις $\{1, ..., V\}$. Μαθηματικά, μια *λέξη* αντιπροσωπεύεται ως ένα διάνυσμα που έχει ένα στοιχείο ίσο με τη μονάδα ενώ όλα τα άλλα στοιχεία είναι ίσα με μηδέν. Για παράδειγμα, η αναπαράσταση της πρώτης λέξης του λεξικού αντιστοιχεί στο διάστασης V διάνυσμα $w_1 = [\ 1\ 0\ 0\ 0\ 0\ ...\ ]$.
- Ένα *έγγραφο* είναι μια ομάδα N λέξεων που συμβολίζεται ως $\boldsymbol{d} = (w_1, w_2, ..., w_N)$, όπου w_n είναι η n-οστή λέξη της ομάδας.
- Ένα *σύνολο κειμένων* είναι μια συλλογή M εγγράφων που συμβολίζονται ως $\boldsymbol{C} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, ..., \boldsymbol{d_M}\}$.
- Ένα *θέμα* αντιστοιχεί σε μια κατανομή πάνω στο λεξικό και συμβολίζεται ως β (το β_k υποδηλώνει την κατανομή του k-οστού θέματος, όπου $k \in K$ και το K αντιστοιχεί στο συνολικό πλήθος των θεμάτων).
- Η *document-topic* κατανομή για το έγγραφο $\boldsymbol{d}$ ορίζεται ως θ_d, ενώ το $\theta_{k,d}$ είναι η αναλογία του θέματος β_k στο έγγραφο $\boldsymbol{d}$.
- Η *topic-word* κατανομή για το έγγραφο $\boldsymbol{d}$ ορίζεται ως z_d, ενώ $z_{d,n}$ είναι η ανάθεση θεμάτων για τη λέξη w_n στο έγγραφο $\boldsymbol{d}$.

3.2.3 Αλγόριθμος

Γενικά, ο LDA θα μπορούσε να περιγραφεί ως ένα πιθανοτικό αναπαραγωγικό μοντέλο ενός συνόλου κειμένων, όπου οι παρατηρούμενες μεταβλητές είναι τα έγγραφα και οι κρυφές μεταβλητές (hidden variables) είναι τα θέματα που διαμένουν στο υπό εξέταση σύνολο. Όπως αναφέρθηκε παραπάνω, η βασική ιδέα του αλγορίθμου είναι ότι σε κάθε έγγραφο θα μπορούσε να ανατεθεί μια κατανομή πάνω σε θέματα, όπου κάθε θέμα είναι μια κατανομή πάνω σε λέξεις. Για να εξαγάγει αυτές τις κατανομές ο LDA ακολουθεί την παρακάτω αναπαραγωγική διαδικασία για κάθε έγγραφο $\boldsymbol{d}$ σε ένα σύνολο κειμένων $\boldsymbol{C}$, του οποίου η γραφική αναπαράσταση δίνεται στο Σχήμα 3.4.

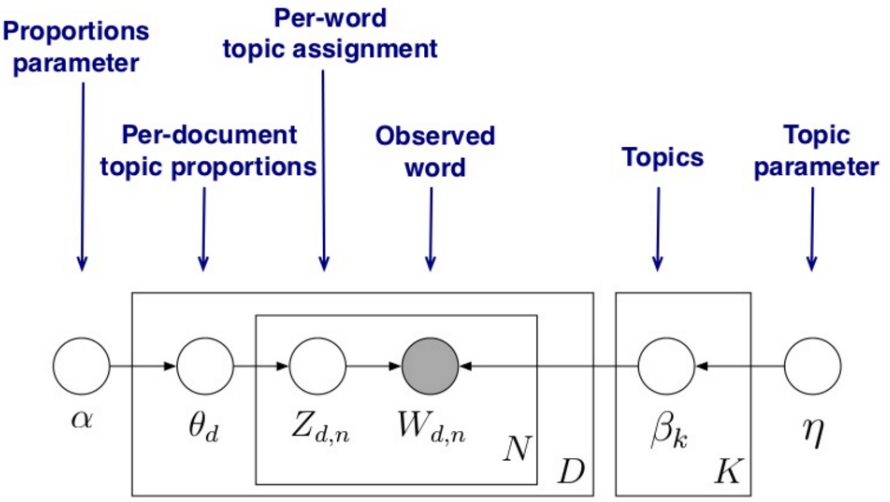
1. Επέλεξε $N \sim \text{Poisson}(\xi)$, όπου το N αντιστοιχεί στο πλήθος των λέξεων του $\boldsymbol{d}$.
2. Επέλεξε $\vartheta \sim \text{Dirichlet}(\alpha)$
3. Για κάθε μια από τις N λέξεις, w_n:
   (α′) Διάλεξε ένα θέμα $z_n \sim \text{Multinomial}(\vartheta)$
   (β′) Διάλεξε μια λέξη w_n από την $p(w_n|z_n, \beta)$, που έιναι μία multinomial δεσμευμένη πιθανότητα του θέματος z_n.

Απώτερος στόχος της παραπάνω διαδικασίας είναι να υπολογίσει τις κρυφές κατανομές $\theta_{1:D}$, $z_{1:D}$, $\beta_{1:K}$, δοθεισών των γνωστών μεταβλητών $w_{1:D}$. Ως αποτέλεσμα, το βασικό πρόβλημα που πρέπει να επιλύσουμε προκειμένου να χρησιμοποιήσουμε τον LDA είναι αυτό του υπολογισμού των posterior κατανομών των κρυφών μεταβλητών δεδομένων του συνόλου κειμένων όπως αναλύεται στην σχέση 3.1, χρησιμοποιώντας το Θεώρημα του Bayes.

$$p(\beta_{1:K}, \vartheta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{3.1}$$

Ο αριθμητής του παραπάνω κλάσματος μπορεί να υπολογιστεί ως η κοινή κατανομή όλων των τυχαίων μεταβλητών. Ωστόσο, για να υπολογίσουμε τον παρονομαστή του κλάσματος, πρέπει να ολοκληρώσουμε ως προς όλα τα πιθανά θέματα που ορίζονται από τις $\theta_{1:D}$, $z_{1:D}$ και $\beta_{1:K}$. Όταν συμβαίνει κάτι τέτοιο, δημιουργείται σύζευξη ανάμεσα στις $\theta_{1:D}$ και $\beta_{1:K}$ καθιστώντας αδύνατο τον διαχωρισμό τους στον υπολογισμό της συνάρτησης log likelihood. Έτσι, αφού ο ακριβής υπολογισμός του κλάσματος δεν είναι εφικτός, έχουν προταθεί διάφορες τεχνικές για την προσέγγιση της παραπάνω λύσης:

- **Variational Inference**. Η ιδέα που προτάθηκε από τον Reed [2012] ήταν η τροποποίηση του αρχικού γραφικού μοντέλου του Σχήματος 3.3 αφαιρώντας τις μεταβάσεις και τους κόμβους που είναι υπεύθυνοι για τη δημιουργία της ανεπιθύμητης ζεύξης που αναφέρθηκε παραπάνω. Ως αποτέλεσμα, χρησιμοποιήθηκε μια απλούστερη κατανομή για την προσέγγιση της πραγματικής.
- **Collapsed Gibbs Sampling**. Η προσέγγιση που προτάθηκε από τους Griffiths and Steyvers [2004] ήταν ότι μια κατανομή υψηλής διάστασης μπορεί να προσομοιωθεί μέσω δειγματοληψίας σε υποσύνολα μεταβλητών μικρότερης διάστασης όπου κάθε υποσύνολο εξαρτάται από τις τιμές των υπόλοιπων μεταβλητών. Η δειγματοληψία γίνεται διαδοχικά και συνεχίζεται έως ότου οι τιμές των δειγμάτων προσεγγίσουν την πραγματική κατανομή.



Σχήμα 3.4: Γραφική απεικόνιση του μοντέλου LDA από τους Blei et al. [2003].

- **Collapsed Variational Inference**. Οι Teh et al. [2006] έκαναν πιο χαλαρές παραδοχές παραγοντοποίησης από εκείνες που έγιναν από τον αλγόριθμο Variational Inference προκειμένου να προσεγγίσουν την πραγματική posterior πιθανότητα. Συγκεκριμένα, αντί να υποθέσουν ότι οι παράμετροι είναι ανεξάρτητες από τις κρυφές μεταβλητές, απομόνωσαν τις μεταβλητές θ και β από το ολοκλήρωμα.

- **Online Variational Inference**. Αργότερα, οι Hoffman et al. [2010] σημείωσαν ότι ο αλγόριθμος Varational Inference απαιτεί ένα πλήρες πέρασμα από το σύνολο κειμένων σε κάθε επανάληψη, καθιστώντας τη διαδικασία αργή για μεγάλα σύνολα δεδομένων. Προς αυτή την κατεύθυνση, πρότειναν έναν εναλλακτικό αλγόριθμο βασισμένο στη στοχαστική βελτιστοποίηση με natural gradient βήμα. Έδειξαν επίσης ότι ο αλγόριθμος παράγει καλές εκτιμήσεις παραμέτρων εντυπωσιακά ταχύτερα από τους batch αλγορίθμους για μεγάλα σύνολα δεδομένων.

3.3 Συσσωρευτική ταξιδόμηση

Σε αυτή την ενότητα περιγράφουμε συνοπτικά τη συσσωρευτική ταξιδόμηση (agglomerative clustering) που χρησιμοποιείται ως τεχνική εξομάλυνσης (smoothing) στην προτεινόμενη προσέγγισή μας. Η συσσωρευτική ταξιδόμηση αποτελεί μια μέθοδο ιεραρχικής ομαδοποίησης που επιχειρεί να δημιουργήσει μια ιεραρχία ομάδων, ακολουθώντας μια προσέγγιση "από τη βάση προς τα πάνω" χωρίζοντας ένα σύνολο αρχικών παρατηρήσεων σε διαφορετικές ομάδες με βάση κάποιο κριτήριο ομοιότητας. Συγκεκριμένα, δεδομένου ενός συνόλου K παρατηρήσεων προς ομαδοποίηση και ενός $K \times K$ πίνακα αποστάσεων (που περιέχει τις αποστάσεις μεταξύ παρατηρήσεων), η βασική διαδικασία της ιεραρχικής ομαδοποίησης, όπως περιγράφεται από τον Johnson [1967], αποτελείται από τα παρακάτω βήματα:

1. Ξεκινάμε αναθέτοντας κάθε στοιχείο στη δική του ομάδα, έτσι ώστε εάν έχουμε K στοιχεία, να τοποθετούνται σε K ομάδες η κάθε μια από τις οποίες περιέχει μόνο ένα

στοιχείο. Θεωρούμε ότι οι αποστάσεις μεταξύ των ομάδων ισούνται με τις αποστάσεις μεταξύ των στοιχείων που αυτές περιέχουν.

2. Βρίσκουμε το πλησιέστερο ζεύγος ομάδων (η εγγύτητα ορίζεται με βάση κάποια μετρική) και τις συγχωνεύουμε σε μια ενιαία ομάδα, έτσι ώστε να προκύψει ένα μικρότερο σύνολο ομάδων.

3. Υπολογίζουμε τις αποστάσεις μεταξύ της νέας ομάδας και κάθε παλαιάς ομάδας.

4. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι να συγκεντρωθούν όλα τα στοιχεία σε μια ομάδα μεγέθους K.

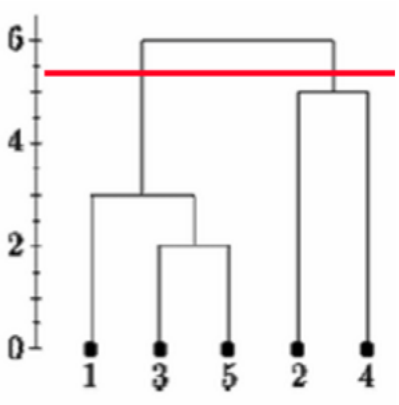
Προκειμένου να ακολουθήσουμε την παραπάνω διαδικασία, πρέπει να ορίσουμε τις αποστάσεις μεταξύ των παρατηρήσεων καθώς και τις αποστάσεις μεταξύ ομάδων αποτελούμενων από παρατηρήσεις, χρησιμοποιώντας οποιονδήποτε από τους ακόλουθους ορισμούς, όπως συνοψίζονται στους Πίνακες 3.1 και 3.2.

Distance	Equation
Euclidean	$\|a-b\|_2 = \sum \sqrt{(a_i - b_i)^2}$
Squared Euclidean	$\|a-b\|_2 = \sum (a_i - b_i)^2$
Manhattan	$\|a-b\|_1 = \sum \|a_i - b_i\|$
Maximum	$\|a-b\|_\infty = \max_i (a_i - b_i)^2$
Cosine	$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$
Pearson	$1 - \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$
Spearman	$1 - \frac{6\sum d_i^2}{n(n^2-1)}$

Πίνακας 3.1: Διαφορετικές μετρικές απόστασης μεταξύ ζευγών παρατηρήσεων.

Linkage Criterion	Equation
Average	$\frac{1}{\|A\|-\|B\|}\sum_{a\in A}\sum_{b\in B} d(a,b)$
Single	$\min\{d(a,b) : a \in A, b \in B\}$
Complete	$\max\{d(a,b) : a \in A, b \in B\}$

Πίνακας 3.2: Διαφορετικά κριτήρια σύνδεσης (linkage ctiteria) που καθορίζουν την απόσταση μεταξύ ομάδων που αποτελούνται από παρατηρήσεις.



Σχήμα 3.5: Δενδρόγραμμα που απεικονίζει τη συσσωρευτική ταξιδόμηση 5 παρατηρήσεων.

Η παραπάνω διαδικασία οδηγεί στη δημιουργία μιας ιεραρχίας παρατηρήσεων, όπως απεικονίζεται στο Σχήμα 3.5. Ένα από τα προβλήματα της ιεραρχικής ομαδοποίησης είναι ότι δεν υπάρχει κανένας αντικειμενικός τρόπος να ορίσουμε πόσες ομάδες παρατηρήσεων υπάρχουν. Για να αποκτήσουμε ομάδες παρατηρήσεων, πρέπει να "κόψουμε" το δέντρο ιεραρχίας (δενδρόγραμμα) σε κάποιο σημείο. Για παράδειγμα, η κόκκινη γραμμή που απεικονίζεται στο Σχήμα 3.5 δηλώνει τη δημιουργία δύο ομάδων.

Κεφάλαιο 4

Μείγμα Θεματικά Κατανεμημένων Σημασιολογικών Μοντέλων

4.1 Κίνητρο

Πρόσφατες προσεγγίσεις που δημιουργούν πολλαπλές κατανεμημένες αναπαραστάσεις για κάθε λέξη χρησιμοποιούν τεχνικές θεματικής μοντελοποίησης (topic modeling) όπως αναφέρεται στο Κεφάλαιο 2. Ένα θεματικό μοντέλο (topic model) καταλήγει σε μια απλή αναπαράσταση των θεματικών περιοχών που υπάρχουν στο σύνολο κειμένων που υπόκειται σε ανάλυση. Συνήθως, κάθε θέμα αντιπροσωπεύεται από μια κατανομή πάνω στο λεξικό, η οποία υποδηλώνει τις λέξεις εκείνες που είναι ιδιαίτερα εμφανείς για το συγκεκριμένο θέμα. Το κύριο κίνητρο πίσω από τη χρήση της θεματικής μοντελοποίησης στο πρόβλημα υπολογισμού της σημασιολογικής ομοιότητας λέξεων είναι η προσαρμογή των εκτιμήσεων ομοιότητας σε διάφορες θεματικές περιοχές. Αυτό είναι παρόμοιο με τη χρήση ενός συνδυασμού σημασιολογικών μοντέλων για την κωδικοποίηση των πολλαπλών αισθήσεων των λέξεων. Σε αυτό το κεφάλαιο, παρουσιάζουμε ένα μείγμα σημασιολογικών μοντέλων βασισμένο σε θεματικούς υπολογισμούς της σημασιολογικής ομοιότητας ανάμεσα σε ζεύγη λέξεων. Το μοντέλο αυτό βασίζεται σε προηγούμενες προσεγγίσεις που χρησιμοποιούν έναν συνδυασμό από σημασιολογικές ομοιότητες λέξεων που υπολογίζονται χρησιμοποιώντας Θεματικά Κατανεμημένα Σημασιολογικά Μοντέλα (ΘΚΣΜ - Topic-based Distributional Semantic Models, TDSMs).

4.2 Περιγραφή Αλγορίθμου

Το προτεινόμενο μοντέλο που εξετάζεται σε αυτό το κεφάλαιο ακολουθεί μια προσέγγιση δύο σταδίων για την εξαγωγή σημασιολογικών ομοιότητων μεταξύ ζευγών λέξεων που παρέχονται είτε σε συμφραζόμενα πλαίσια είτε απουσία συμφραζόμενων πλαισίων, όπως παρουσιάζεται στις επόμενες υποενότητες.

4.2.1 Θεματικά Κατανεμημένα Σημασιολογικά Μοντέλα

Θεματικά υποσύνολα κειμένων

Συνήθως, τα σύνολα κειμένων που χρησιμοποιούνται σε πολλές εφαρμογές της Επεξεργασίας Φυσικού Λόγου αποτελούνται από μεγάλες συλλογές αρχείων που εμπεριέχουν πληροφορία από διάφορους θεματικούς τομείς, ενσωματώνοντας έτσι συμφραζόμενα πλαίσια γενικών πληροφοριών. Παρόλα αυτά, τα θεματικά σύνολα κειμένων (topic-based corpora) είναι εξαιρετικά χρήσιμα για την εκμάθηση και την κατανόηση της φυσικής γλώσσας. Παρά τη χρησιμότητά τους, θεματικά σύνολα κειμένων δεν είναι ακόμη διαθέσιμα για πολλούς θεματικούς τομείς. Ένας τρόπος για να αντιμετωπιστεί αυτή η έλλειψη θεματικών δεδομένων είναι η χρήση των ήδη διαθέσιμων πόρων, όπως τα γενικά σύνολα κειμένων, προκειμένου να κατασκευαστούν αντίστοιχα σύνολα που περιέχουν θεματική πληροφορία. Όπως περφιγράφουμε λεπτομερώς παρακάτω, παρουσιάζουμε μια μέθοδο που δημιουργεί θεματικά υποσύνολα κειμένων χρησιμοποιώντας τον Latent Dirichlet Allocation αλγόριθμο.

Αρχικά, τα μόνο δεδομένα που απαιτούνται είναι ένα μεγάλο σύνολο κειμένων (global corpus). Μπορεί να είναι ένα γενικό σύνολο (που έχει ληφθεί από τον ιστό (π.χ. Wikipedia)) ή μια συλλογή από κριτικές (π.χ. κριτικές ταινιών). Θα χρειαστούμε δύο εκδόσεις αυτού του συνόλου κειμένων. Ένα που περιέχει έγγραφα και ένα άλλο ποθ περιέχει προτάσεις.

- Η πρώτη έκδοση που περιέχει έγγραφα (**document-level**) θα προωθηθεί ως είσοδος στον αλγοριθμό Latent Dirichlet Allocation, δεδομένου ότι στην προεπιλεγμένη μορφή του υποθέτει ότι κάθε έγγραφο περιέχει πολλαπλά θέματα.

- Η δεύτερη έκδοση που περιέχει προτάσεις (**sentence-level**) θα χρησιμοποιηθεί για την ομαδοποίηση των αρχικών δεδομένων σε συγκεκριμένες θεματικές κλάσεις, με βάση την υπόθεση ότι μικρότερα σύνολα κειμένων (προτάσεις) είναι πιθανόν να μιλούν για ένα μόνο θέμα, οπότε η ταξινόμηση μπορεί να είναι αυστηρή. Επιπλέον, η επιλογή αυτή ακολουθεί τις βασικές αρχές της θεματικής μοντελοποίησης, καθώς οι προτάσεις είναι θεματικά πλήρεις και συνεκτικές μονάδες.

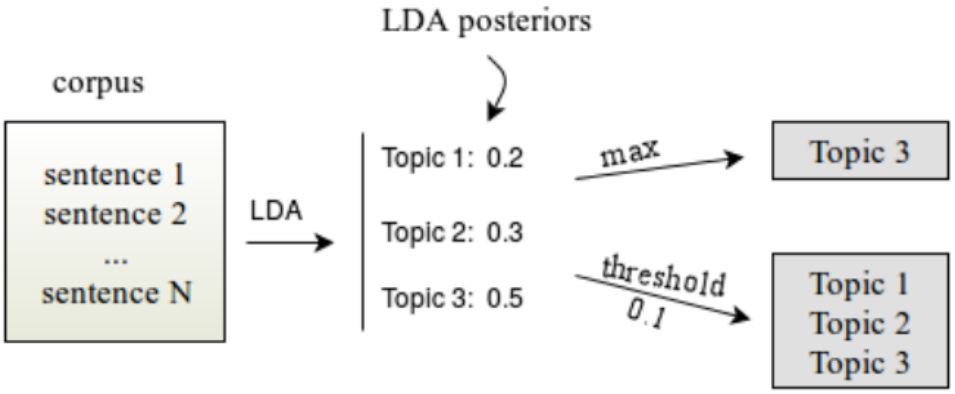
Η δημιουργία των θεματικών υποσυνόλων κειμένων (βλ. Σχήμα 4.1) περιγράφεται από τα τρία ακόλουθα βήματα:

1. Ξεκινώντας από ένα γενικό σύνολο εγγράφων, χρησιμοποιούμε τον αλγόριθμο Latent Dirichlet Allocation (LDA) για να εκπαιδεύσουμε ένα θεματικό μοντέλο, για ένα σταθερό πλήθος θεμάτων ίσο με K. Ο LDA συνδέει κάθε έγγραφο με αναλογίες θεμάτων βασισμένος στην ιδέα ότι μια ποικιλία θεμάτων αναφέρεται σε κάθε έγγραφο. Το εκπαιδευμένο θεματικό μοντέλο παράγει μια κατανομή πανω στις λέξεις του λεξικού για κάθε θέμα.

2. Στη συνέχεια, εφαρμόζουμε το εκπαιδευμένο μοντέλο LDA στην έκδοση του συνόλου κειμένων που αποτελείται από προτάσεις. Ως αποτέλεσμα, κάθε πρόταση συνδέεται με μια λίστα θεμάτων, που πιθανώς συζητήθηκαν στην πρόταση, σύμφωνα με το εκπαιδευμένο θεματικό μοντέλο.

3. Τέλος, δημιουργούμε ένα θεματικό υποσύνολο κειμένων για κάθε θέμα $k \in K$ ομαδοποιώντας τις προτάσεις των οποίων οι posterior πιθανότητες μεγιστοποιούνται για το

θέμα k. Αυτό το σχήμα ισχυρής ομαδοποίησης (hard clustering) μπορεί να οδηγήσει στη δημιουργία θεματικών υποσυνόλων περιορισμένου μεγέθους. Για να αποφευχθεί αυτό, υιοθετούμε ένα σύστημα ελαστικής ομαδοποίησης (soft clustering). Συγκεκριμένα, μια πρόταση αντιστοιχίζεται σε ένα θέμα όταν η posterior πιθανότητα γι'αυτό υπερβαίνει την τιμή ενός κατωφλίου h. Οι προτάσεις που παρουσιάζουν ίσες posterior πιθανότητες για όλα τα θέματα εξαιρούνται από αυτή τη διαδικασία, καθώς θεωρούνται υπερβολικά γενικές για την παροχή οποιασδήποτε θεματικής πληροφορίας.

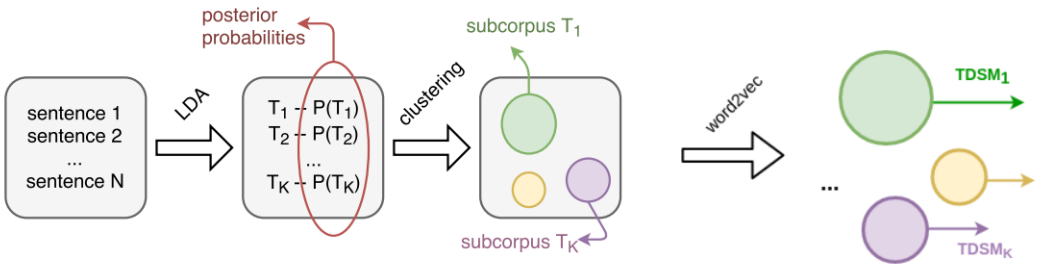


Σχήμα 4.1: Αφηρημένη απεικόνιση της δημιουργίας θεματικών υποσυνόλων κειμένων όπως παρουσίαζεται στην Christopoulou [2016].

Θεματικά Κατανεμημένα Σημασιολογικά Μοντέλα

Όπως αναφέρεται στην Ενότητα 2, τα Κατανεμημένα Σημασιολογικά Μοντέλα (ΚΣΜ) κωδικοποιούν τις συμφραζόμενες πληροφορίες των λέξεων που προέρχονται από τα σύνολα κειμένων, σε πυκνές αναπαραστάσεις χαρακτηριστικών. Ως αποτέλεσμα, τα αντίστοιχα ΚΣΜ που έχουν εκπαιδευτεί κάτω από αυτά οδηγούν στη δημιουργία γενικών διανυσματικών αναπαραστάσεων, καθώς δεν λαμβάνουν υπόψη τις εννοιολογικές διαφοροποιήσεις που εμφανίζει μια λέξη σε διαφορετικές θεματικές περιοχές. Η απομόνωση των διαφορετικών εννοιών των λέξεων θα μπορούσε να επιτευχθεί με τη συλλογή κειμένων που περιέχουν θεματολογική πληροφορία, χρησιμοποιώντας τεχνικές θεματικής μοντελοποίησης όπως περιγράφεται παραπάνω.

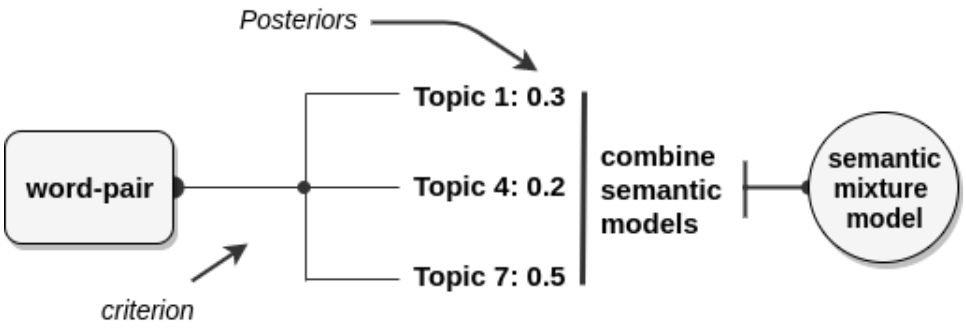
Χρησιμοποιούμε τον όρο Θεματικά Κατανεμημένα Σημασιολογικα Μοντέλα (ΘΚΣΜ) σε όλη την έκταση της παρούσας εργασίας, για να αναφερθούμε σε ένα ΚΣΜ το οποίο εκπαιδεύεται σε ένα θεματικό σύνολο κειμένων. Συγκεκριμένα, χρησιμοποιούμε τον αλγόριθμο Word2Vec για να δημιουργήσουμε θεματικές αναπαραστάσεις. Σημειώνουμε επίσης ότι τα ΘΚΣΜ θα μπορούσαν να εκπαιδευτούν πανω σε θεματικά σύνολα κειμένων, χρησιμοποιώντας οποιοδήποτε ΚΣΜ που κωδικοποιεί γλωσσικά χαρακτηριστικά κειμένων για να ενσωματώσουν λέξεις σε έναν σημασιολογικό χώρο. Το Σχήμα 4.2 συνοψίζει τη διαδικασία που ακολουθούμε για τη δημιουργία θεματικών αναπαραστάσεων λέξεων που ξεκινούν από ένα γενικό σύνολο δεδομένων.



Σχήμα 4.2: Αφηρημένη απεικόνιση της κατασκευής Θεματικά Κατανεμημένων Σημασιολογικών Μοντέλων.

4.2.2 Σημασιολογικά Μείγματα

Συνήθως, ο υπολογισμός της σημασιολογικής ομοιότητας μεταξύ ενός ζεύγους λέξεων ενσωματώνει όλες τις πιθανές έννοιες με τις οποίες οι λέξεις εμφανίζονται σε ένα σύνολο κειμένων. Για διάφορες σημασιολογικές εργασίες που σχετίζονται με τον υπολογισμό λεξικών ομοιοτήτων, τα παραπάνω μοντέλα επιτυγχάνουν πολύ καλές επιδόσεις, παρά την απόκλισή τους από την υπόθεση της μέγιστης ομοιότητας των εννοιών (maximum sense similarity), που υποδηλώνει ότι η σημασιολογική ομοιότητα μεταξύ δύο λέξεων μπορεί να εκτιμηθεί ως η ομοιότητα των δύο πλησιέστερων εννοιών που τις απαρτίζουν. Στο μοντέλο που προτείνουμε σε αυτό το κεφάλαιο, η προαναφερθείσα υπόθεση υιοθετείται μέσω της δημιουργίας θεματικών υποσυνόλων κειμένων για κάθε ζεύγος που αποτελείται από τις λέξεις $(word_i, word_j)$. Ο στόχος είναι οι λέξεις του ζεύγους να συνυπάρχουν σε κάθε θεματικό υποσύνολο κειμένων με τις πλησιέστερες έννοιές τους, που είναι συναφείς με τα αντίστοιχα θέματα. Αυτή η προσέγγιση είναι διαφορετική σε σύγκριση με την τυπική επαγωγή των εννοιών των λέξεων με βάση το σύνολα κειμένων (που επίσης αναφέρεται ως sense discovery), όπου η επαγωγή των εννοιών γίνεται ξεχωριστά για κάθε λέξη. Η ομοιότητα ανάμεσα στις λέξεις $word_i$ και $word_j$ υπολογίζεται ως ένα μείγμα από ομοιότητες οι οποίες έχουν εξαχθεί από θεματικά υποσύνολα κειμένων, όπως απεικονίζεται στο Σχήμα 4.3.



Σχήμα 4.3: Αφηρημένη παρουσίαση της μεθόδου που βασίζεται σε μείγμα σημασιολογικών μοντέλων.

Συγκεκριμένα, ακολουθούμε δύο διαφορετικές προσεγγίσεις για να υπολογίσουμε τις *posteriors* που απεικονίζονται παραπάνω χρησιμοποιώντας διάφορα κριτήρια, βασισμένα στο

πρόβλημα σημασιολογικής ομοιότητας που χρησιμοποιούμε για την αξιολόγηση του μοντέλου. Ορίζουμε ως L_K το σύνολο των K ΘΚΣΜ που προέρχονται από τον αλγόριθμο LDA, όπου λ_k είναι το ΚΣΜ που αντιστοιχεί στο k-οστό θέμα από το σύνολο των K θεμάτων.

Σημασιολογικοί Συνδυασμοί παρουσία συμφραζόμενων πλαισίων

Όταν παρέχονται συμφραζόμενα πλαίσια για ένα ζεύγος λέξεων (w, w') για τις οποίες θέλουμε να υπολογίσουμε τη σημασιολογική τους εγγύτητα, δημιουργούμε ένα κοινό πλαίσιο συμφραζόμενων $c'' = c \oplus c'$ το οποίο διαμορφώνεται συνενώνοντας τα επιμέρους πλαίσια κάθε λέξης c και c', αντίστοιχα . Το θεματικό μοντέλο τροφοδοτείται με το c'' και εξάγει μια λίστα υποψηφίων θεμάτων γι'αυτό, μαζί με τις αντίστοιχες posterior πιθανότητες. Αυτά τα θέματα χρησιμοποιούνται για την αναγνώριση των αντίστοιχων θεματικών υποσυνόλων κειμένων, τα οποία χρησιμοποιούνται για την εκπαίδευση συγκεκριμένων Θεματικών ΚΣΜ (ΘΚΣΜ). Προκειμένου να ληφθούν υπόψη οι συμφραζόμενες πληροφορίες των λέξεων, ορίζουμε δύο σημασιολογικούς συνδυασμούς:

$$S_{\text{AvgSimC}}(w, w'; L_K) = \frac{\sum_{k=1}^{T(c'')} p(k|c'') S_k(w, w'; \lambda_k)}{\sum_{k=1}^{T(c'')} p(k|c'')}, \tag{4.1}$$

$$S_{\text{MaxSimC}}(w, w'; L_K) = S_{\hat{k}}(w, w'; \lambda_{\hat{k}}) \quad \text{where} \quad \hat{k} = \arg\max_{k \in T(c'')} p(k|c''), \tag{4.2}$$

όπου $T(c'')$ είναι τα υποψήφια θέματα που επιστρέφονται από το θεματικό μοντέλο με posterior πιθανότητα μεγαλύτερη από 0.01 δοθείσας της εισόδου c'', η $p(k|c'')$ δηλώνει την posterior πιθανότητα του k-οστού θέματος για c'', ενώ $S_k(w, w'; \lambda_k)$ είναι η σημασιολογική ομοιότητα ανάμεσα στις λέξεις w και w' όπως προκύπτει από το ΚΣΜ που αντιστοιχεί στο k-οστό θέμα. Επειδή το πλήθος των υποψηφίων θεμάτων μπορεί να είναι μικρότερο ή ίσο με το συνολικό πλήθος θεμάτων ($T(c'') \leq K$) για τα οποία εκπαιδεύεται ο LDA οι posterior πιθανότητες κανονικοποιούνται ώστε να αθροίζουν στη μονάδα.

Δεδομένου ότι το c'' εισάγεται στο θεματικό μοντέλο, η σχέση 4.1 υπολογίζει έναν σταθμισμένο μέσο όρο των θεματικών σημασιολογικών ομοιοτήτων χρησιμοποιώντας τις posterior πιθανότητες των θεμάτων ως βάρη. Σημειώστε ότι για τα ζεύγη που αναφέρονται στην ίδια λέξη αλλά βρίσκονται σε διαφορετικά συμφραζόμενα πλαίσια, το μοντέλο τους αποδίδει πάντοτε την ίδια τιμή σημασιολογικής ομοιότητας η οποία ισούται με 1 (μέγιστη δυνατή τιμή ομοιότητας), καθώς οι αναπαραστάσεις τους εξάγονται από τα ίδια ΘΚΣΜ. Το μοντέλο ακολουθεί τη μέση οδό ανάμεσα στην υπόθεση της μέγιστης ομοιότητας των λέξεων και στα ΚΣΜ που επάγουν ξεχωριστές αναπαραστάσεις για τις έννοιες των πολυσήμαντων λέξεων. Η παραπάνω υπόθεση υιοθετείται μέσω της αναγνώρισης των θεματικών υποσυνόλων κειμένων στα οποία οι λέξεις w και w' εμφανίζονται με συναφείς έννοιες στον θεματικό τομέα των αντίστοιχων κειμένων. Η ενσωμάτωση των βαρών του σημασιολογικού μείγματος στον υπολογισμό της τελικής ομοιότητας χαλαρώνει την παραπάνω υπόθεση. Χρησιμοποιώντας τη σχέση 4.2 αποδίδεται σε ένα ζεύγος η σημασιολογική ομοιότητα που αντιστοιχεί στο θέμα με τη μέγιστη posterior πιθανότητα, το οποίο επίσης αποτελεί το κυρίαρχο θέμα του παρεχόμενου συμφραζόμενου πλαισίου.

Επιπλέον, εισάγουμε ένα μοντέλο σύντηξης (fusion) που συνδυάζει πληροφορίες από πολλαπλά θεματικά μοντέλα εκπαιδευμένα για διαφορετικά πλήθη θεμάτων. Συγκεκριμένα, για

ένα εκπαιδευμένο θεματικό μοντέλο με K θέματα, η σημασιολογική ομοιότητα ενός ζεύγους λέξεων υπολογίζεται χρησιμοποιώντας έναν από τους προαναφερθέντες συνδυασμούς, όπως ορίζεται από τις σχέσεις 4.1, 4.2. Μεταξύ των ομοιοτήτων που παράγονται από την εκπαίδευση πολλαπλών θεματικών μοντέλων για διάφορα πλήθη θεμάτων, επιλέγουμε τη μέγιστη ομοιότητα ζεύγους της θεματικής ομάδας G, που αποτελείται από σύνολα ΘΚΣΜ που ορίζονται ως L_K, όπως παρουσιάζεται παρακάτω:

$$S_{\text{Fuse}}(w, w') = \max_{L_K \in G} S_{*\text{Sim}}(w, w'; L_K), \tag{4.3}$$

όπου $S_{*\text{Sim}}(w, w; L_K)$ είναι η ομοιότητα για το ζεύγος (w, w') που υπολογίζεται χρησιμοποιώντας έναν από τους 4.1, 4.2 συνδυασμούς, K είναι το πλήθος των θεμάτων που περιέχονται σε μία συγκεκριμένη ομάδα και G είναι η ομάδα των ΚΣΜ στην οποία εφαρμόζεται η μέθοδος σύντηξης.

Σημασιολογικοί Συνδυασμοί απουσία συμφραζόμενων πλαισίων

Η σημασιολογική ομοιότητα μεταξύ δύο λέξεων w και w' για τις οποίες δεν μας παρέχεται συμφραζόμενη πληροφορία υπολογίζεται χρησιμοποιώντας διαφορετικούς συνδυασμούς ομοιότητας σε σύγκριση με την περίπτωση στην οποία γνωρίζουμε το πλαίσιο συμφραζόμενων στο όποιο εντάσσεται το ζεύγος ενδιαφέροντος. Συγκεκριμένα, ένα μοντέλο μίγματος θεματικών ομοιοτήτων ενσωματώνεται για να παραγάγει την τελική ομοιότητα $S(w, w')$ μεταξύ του ζεύγους λέξεων. Αυτό πραγματοποιείται ακολουθώντας τους Reisinger and Mooney [2010], και ορίζοντας τους δύο παρακάτω συνδυασμούς:

$$S_{\text{AvgSim}}(w, w'; L_K) = \frac{1}{K} \sum_{k=1}^{K} S_k(w, w; ; \lambda_k), \tag{4.4}$$

$$S_{\text{MaxSim}}(w, w'; L_K) = \max_{k \in K} \{S_k(w, w; ; \lambda_k)\}, \tag{4.5}$$

όπου $S_k(w, w'; \lambda_k)$ είναι η σημασιολογική ομοιότητα ανάμεσα στις w και w' που υπολογίζεται από το λ_k ΚΣΜ, το οποίο κατασκευάστηκε χρησιμοποιώντας το υποσύνολο κειμένων που αντιστοιχεί στο k-οστό θέμα. Στη σχέση 4.4 υπολογίζεται ο μη σταθμισμένος μέσος όρος όλων των θεματικών ομοιοτήτων για το υπό εξέταση ζεύγος λέξεων. Στην σχέση 4.5 έχει επιλεγεί μόνο η μέγιστη ομοιότητα ανά ζεύγος, μεταξύ των K πιθανών θεματικών περιοχών.

Τέλος, χρησιμοποιούμε ένα μοντέλο γραμμικής παρεμβολής (linear regression) για να συνδυάσουμε τα ζεύγη ομοιοτήτων που εξήχθησαν από το κάθε ΘΚΣΜ, τα οποία προκύπτουν από ένα θεματικό μοντέλο εκπαιδευμένο για K θέματα. Το μοντέλο στοχεύει στην ελαχιστοποίηση του Μέσου Τετραγωνικού Σφάλματος (Mean Squared Error, MSE) εκπαιδεύοντας ένα σύνολο από βάρη (β) σε ένα σύνολο από θεματικές ομοιότητες λέξεων. Το βασικό κίνητρο πίσω από αυτή την ιδέα είναι να μάθουμε πώς να συνδυάζουμε θεματικές ομοιότητες για ζεύγη λέξεων που παρέχονται μεμονωμένα (χωρίς συμφραζόμενη πληροφορία). Συγκεκριμένα, ο συνδυασμός που προτάθηκε στην παραπάνω υποενότητα (σχέση 4.1) απαιτεί την πρόσθετη εισαγωγή του συμφραζόμενου πλαισίου των λέξεων προκειμένου να υπολογίσει το πόσο θα σταθμιστεί κάθε θεματική ομοιότητα. Αντίθετα, όταν δεν υπάρχει κανένα συμφραζόμενο

πλαίσιο, αντί να υποθέσουμε ότι όλα τα θέματα συμβάλλουν εξίσου στην εκτίμηση της ομοιότητας ενός ζεύγους, όπως περιγράφεται στην σχέση 4.4, υποστηρίζουμε ότι ένας γραμμικός συνδυασμός θεματικών ομοιοτήτων είναι ικανός να παραγάγει μια ακριβέστερη εκτίμηση,

$$S_{\mathrm{LR}}(w, w'; L_K) = \beta_0 + \frac{1}{K} \sum_{k=1}^{K} \beta_k S_k(w, w'; \lambda_k), \tag{4.6}$$

όπου β_k είναι τα βάρη τα οποία μαθαίνουμε από το μοντέλο γραμμικής παρεμβολής για το k-οστό θέμα, $S_k(w, w'; \lambda_k)$ είναι η ομοιότητα του ζεύγους (w, w') εξαγόμενη από το ΚΣΜ που έχει εκπαιδευτεί στο k-οστό υποσύνολο κειμένων και β_k είναι ένα βάρος μεροληψίας. Τα βάρη β αθροίζουν στην μονάδα.

Κεφάλαιο 5

Θεματικά Κατανεμημένο Σημασιολογικό Μοντέλο πολλαπλών αναπαραστάσεων

5.1 Κίνητρο

Τα παραδοσιακά Κατανεμημένα Σημασιολογικά Μοντέλα (ΚΣΜ) αντιπροσωπεύουν σημασιολογικούς χώρους χαρακτηριστικών όπου οι πολλαπλές έννοιες των πολυσήμαντων λέξεων συγχωνεύονται σε μεμονωμένες αναπαραστάσεις. Σε αυτό το κεφάλαιο, περιγράφουμε τη δημιουργία ενός ενοποιημένου μοντέλου που αναθέτει πολλαπλές κατανεμημένες αναπαραστάσεις ανά λέξη, μέσω της ευθυγράμμισης των Θεματικά Κατανεμημένων Σημασιολογικών Μοντέλων (ΘΚΣΜ) σε έναν κοινό χώρο. Βασιζόμενοι στην παραδοχή ότι οι σημασιολογικές σχέσεις ανάμεσα στις μονοσήμαντες λέξεις δεν αλλάζουν σε διαφορετικές θεματικές περιοχές, υποθέτουμε ότι οι σχετικές αποστάσεις των αντίστοιχων διανυσματικών αναπαραστάσεων τους μένουν σταθερές σε όλα τα ΘΚΣΜ, ενεργώντας ως *σημασιολογικές άγκυρες* (semantic anchors) που καθορίζουν τις αντιστοιχίσεις μεταξύ τους.

Συγκεκριμένα προτείνουμε οτι το Θεματικά Κατανεμημένο Σημασιολογικό Μοντέλο πολλαπλών αναπαραστάσεων (Unified multi-topic Distributional Semantic Model, UTDSM) αποτελεί ένα πιο ευέλικτο μοντέλο σε σύγκριση με το Μείγμα των ΘΚΣΜ που περιγράφηκε στο Κεφάλαιο 4, για δύο κυρίως λόγους:

- Το μείγμα από ΘΚΣΜ αποτυγχάνει να αποδώσει τις διαφορετικές σημασιολογικές έννοιες για ένα ζεύγος που αποτελείται από την ίδια λέξη. Λεπτομερέστερα, η σύγκριση μεταξύ δύο λέξεων περιορίζεται πάντοτε σε επίπεδο θεματικής περιοχής. Ως αποτέλεσμα, οι ίδιες λέξεις —παρά τις έννοιες τους— αντιπροσωπεύονται πάντοτε από τις ίδιες αναπαραστάσεις και ως εκ τούτου αποδίδεται σε αυτές η μέγιστη δυνατή ομοιότητα που ορίζεται από τη μετρική απόσταση που χρησιμοποιείται.

- Ο προαναφερθείς περιορισμός οδηγεί επίσης σε μια ακόμα αδυναμία του προηγούμενου μοντέλου που συναντάται όταν συγκρίνονται δύο διαφορετικές λέξεις καθώς υποθέτει ότι η σημασιολογική σχέση δύο λέξεων ορίζεται σε μία *κοινή* θεματική περιοχή. Κατά συνέπεια, δύο λέξεις των οποίων οι έννοιες δεν μπορούν να συναντηθούν στην ίδια θεματική περιοχή δεν μπορούν να συγκριθούν με ακρίβεια.

5.2 Περιγραφή Αλγορίθμου

Το σύστημά μας ακολουθεί μια προσέγγιση τεσσάρων σταδίων που περιγράφεται συνοπτικά στις παρακάτω υπορουτίνες:

1. **Καθολικό Κατανεμημένο Σημασιολογικό Μοντέλο** (Global Distributional Semantic Model). Δεδομένης μιας μεγάλης συλλογής δεδομένων κειμένου, εκπαιδεύουμε ένα ΚΣΜ που κωδικοποιεί τη σημασιολογική πληροφορία κάθε λέξης σε μια ενιαία αναπαράσταση. Το μοντέλο αυτό αναφέρεται επίσης και ως καθολικό-ΚΣΜ (global-DSM).

2. **Θεματικά Κατανεμημένα Σημασιολογικά Μοντέλα.** Στη συνέχεια, ένα θεματικό μοντέλο εκπαιδεύεται χρησιμοποιώντας την ίδια συλλογή δεδομένων κειμένου. Το θεματικό μοντέλο χωρίζει την παραπάνω συλλογή σε K (πιθανώς αλληλεπικαλυπτόμενα) θεματικά υποσύνολα κειμένων SC_1, ..., SC_K (όπως περιγράφηκε στο Κεφάλαιο 4). Στη συνέχεια, ένα ΚΣΜ εκπαιδεύεται για κάθε θεματικό υποσύνολο με αποτέλεσμα τη δημιουργία K ΘΚΣΜ (TDSM), δηλαδή TDSM_1, ..., TDSM_K. Η τοπική προσαρμογή του σημασιολογικού χώρου λαμβάνει υπόψη τις εννοιολογικές διαφοροποιήσεις που παρουσιάζει μια λέξη σε διαφορετικούς θεματικούς τομείς και ως εκ τούτου οδηγεί στη δημιουργία ειδικών θεματικών διανυσμάτων (topic embeddings).

3. **Απεικονίσεις θεματικών διανυσμάτων.** Στη συνέχεια, απεικονίζουμε τον διανυσματικό χώρο κάθε ΘΚΣΜ στον κοινό χώρο του καθολικού-ΚΣΜ, χρησιμοποιώντας μια λίστα μονοσήμαντων λέξεων που σχετίζονται στατιστικά με το αντίστοιχο θέμα. Στον ενοποιημένο σημασιολογικό χώρο, κάθε λέξη αντιπροσωπεύεται από ένα σύνολο θεματικών αναπαραστάσεων που προηγουμένως ήταν απομονωμένες σε ξεχωριστούς διανυσματικούς χώρους, δημιουργώντας έτσι ένα ενοποιημένο θεματικό ΚΣΜ (Unified multi-topic DSM, UTDSM).

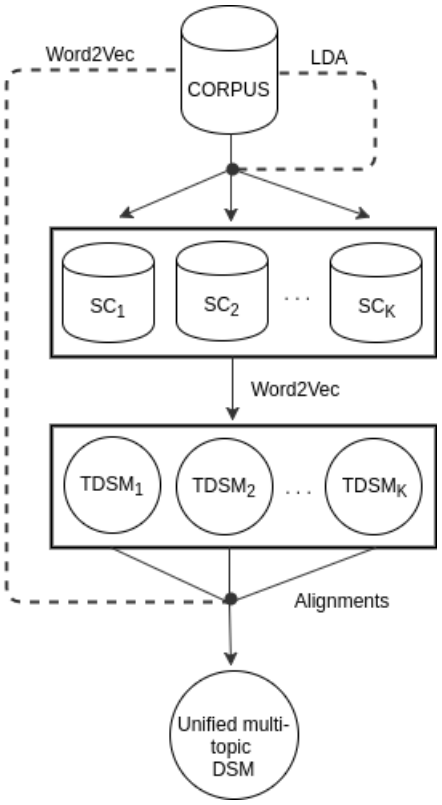
4. **Εξομάλυνση θεματικών διανυσμάτων.** Ως προαιρετικό βήμα, χρησιμοποιούμε μια προσέγγιση εξομάλυνσης βασισμένη στη χρήση της συσσωρευτικής ταξιδόμησης των θεματικών αναπαραστάσεων μιας λέξης σε N ομάδες. Θεωρούμε ότι το βήμα αυτό μειώνει το θόρυβο που εισάγεται στο σύστημά μας μέσω των σημασιολογικών απεικονίσεων και των αραιών δεδομένων εκπαίδευσης.

Η ευθυγράμμιση των ΘΚΣΜ σε έναν κοινό χώρο οδηγεί στη δημιουργία ενός ενοποιημένου μοντέλου πολλαπλών θεματικών κατανεμημένων αναπαραστάσεων, όπως απεικονίζεται στο Σχήμα 5.1.

5.2.1 Κατανεμημένα Σημασιολογικά Μοντέλα

Ξεκινώντας από ένα γενικό σύνολο κειμένων, εκπαιδεύουμε ένα καθολικό-ΚΣΜ καθώς και K ΘΚΣΜ όπως προτείναμε στο Κεφάλιαο 4, όπου το K αποτελεί σταθερή παράμετρο του συστήματός μας. Προβαίνουμε στις ακόλουθες παραδοχές για τα εξαγόμενα ΚΣΜ σχετικά με τα γλωσσικά χαρακτηριστικά που ενσωματώνουν στα αντίστοιχα διανύσματά τους.

- Το **Καθολικό-ΚΣΜ** παρέχει μία μόνο διανυσματική αναπαράσταση ανά λέξη. Καθώς αυτές οι αναπαραστάσεις εξάγονται από μια γενική περιοχή, οι έννοιες που περιέχονται



Σχήμα 5.1: Θεματικά Κατανεμημένο Σημασιολογικό Μοντέλο πολλαπλών αναπαραστάσεων

σε αυτές δεν εξαρτώνται από κάποιο θέμα. Για το λόγο αυτό, αναμένουμε ότι το διάνυσμα που αποδίδεται σε μια μονοσήμαντη λέξη (λέξη που έχει μόνο μία έννοια) θα είναι καλώς τοποθετημένο σε αυτόν τον χώρο. Από την άλλη πλευρά, αναμένουμε ότι το διάνυσμα που έχει εκχωρηθεί σε μια πολυσήμαντη λέξη (λέξη που έχει πολλαπλές έννοιες) θα τοποθετηθεί σε μια θέση που αντανακλά μια μέση αναπαράσταση των πραγματικών εννοιών της. Για παράδειγμα, η πολυσήμαντη λέξη *cancer* (καρκίνος) αναμένεται να έχει μια σχετικά κοντινή απόσταση από τις λέξεις *astrology* (αστρολογία) και *tumor* (όγκος).

- Τα **ΘΚΣΜ** παρέχουν επίσης μια ενιαία διανυσματική αναπαράσταση ανά λέξη. Η διαφορά τους από το καθολικό-ΚΣΜ είναι ότι ενσωματώνουν διαφοροποιήσεις στη χρήση της γλώσσας οι οποίες συναντώνται εντός ενός θεματικού τομέα, γεγονός που διευκολύνει την απομόνωση των πολλαπλών σημασιών των πολυσήμαντων λέξεων στις αντιπροσωπευτικές θεματικές αναπαραστάσεις τους. Ως αποτέλεσμα, η πολυσήμαντη λέξη *cancer* θα μετατοπιστεί προς τη λέξη *astrology* σε έναν ζωδιακό σημασιολογικό χώρο. Από την άλλη πλευρά, η ίδια λέξη θα τοποθετηθεί πιο κοντά στην αναπαράσταση της λέξης *tumor* σε έναν ιατρικό σημασιολογικό χώρο.

5.2.2 Απεικονίσεις θεματικών διανυσμάτων

Ο εγγενής μη-ντετερμινισμός του αλγορίθμου Word2Vec οδηγεί στη δημιουργία συνεχών διανυσματικών χώρων που δεν ευθυγραμμίζονται κατά φυσιολογικό τρόπο σε ένα αναφορικό σύστημα συντεταγμένων. Για το λόγο αυτό, πρέπει να ευθυγραμμίσουμε τα διανύσματα των

λέξεων που προέρχονται από διαφορετικά ΘΚΣΜ κάτω από κοινούς άξονες συντεταγμένων, προκειμένου να καταστεί εφικτή η σύγκρισή τους. Συγκεκριμένα, θεωρούμε ότι τα ΘΚΣΜ συλλαμβάνουν σημαντικές διαφοροποιήσεις στη χρήση των πολυσήμαντων λέξεων, ενώ παράλληλα διατηρούν τις σχετικές θέσεις μονοσήμαντων λέξεων, των οποίων η σημασία μπορεί να κωδικοποιηθεί σε μονές διανυσματικές αναπαραστάσεις. Αυτή η παρατήρηση μας ώθησε στο να δούμε τις μονοσήμαντες λέξεις ως σταθερά σημεία στους θεματικούς σημασιολογικούς μας χώρους, που θα μπορούσαν να χρησιμοποιηθούν ως *σημασιολογικές άγκυρες* για να προσδιορίσουν τις αντιστοιχίσεις μεταξύ τους.

Σε αυτή την εργασία, υποθέτουμε ότι υπάρχει ένας πίνακας μετασχηματισμού M_k ανάμεσα στις διανυσματικές αναπαραστάσεις των μονοσήμαντων λέξεων κάθε ΘΚΣΜ και τις αντίστοιχες αναπαραστάσεις του καθολικού-ΚΣΜ. Δεδομένου ότι απώτερος στόχος μας είναι να ευθυγραμμίσουμε όλα τα ΘΚΣΜ στο ενιαίο σύστημα συντεταγμένων, χρησιμοποιούμε ένα καθολικό-ΚΣΜ ως τον χώρο προορισμού, ενώ το ΘΚΣΜ που αντιπροσωπεύει το θέμα k χρησιμεύει ως χώρος πηγής. Προτείνουμε τη χρήση του καθολικού-ΚΣΜ ως χώρου προορισμού καθώς παρέχει καλές αναπαραστάσεις για τις μονοσήμαντες λέξεις.

Έστω ότι $X = [x_1, x_2, ..., x_n]$ και $Y = [y_1, y_2, ..., y_n]$ είναι οι πίνακες με τις διανυσματικές αναπαραστάσεις λέξεων που αντιστοιχούν στους χώρους πηγής και προορισμού, όπου x_i, y_i $\in \mathbb{R}^d$ και $X, Y \in \mathbb{R}^{d \times n}$. Απώτερος στόχος μας είναι να βρούμε ένα πίνακα μετασχηματισμού $M_k \in \mathbb{R}^{d \times d}$ για κάθε θέμα k, ο οποίος προσεγγίζει την σχέση 5.1:

$$M_k X = Y. \tag{5.1}$$

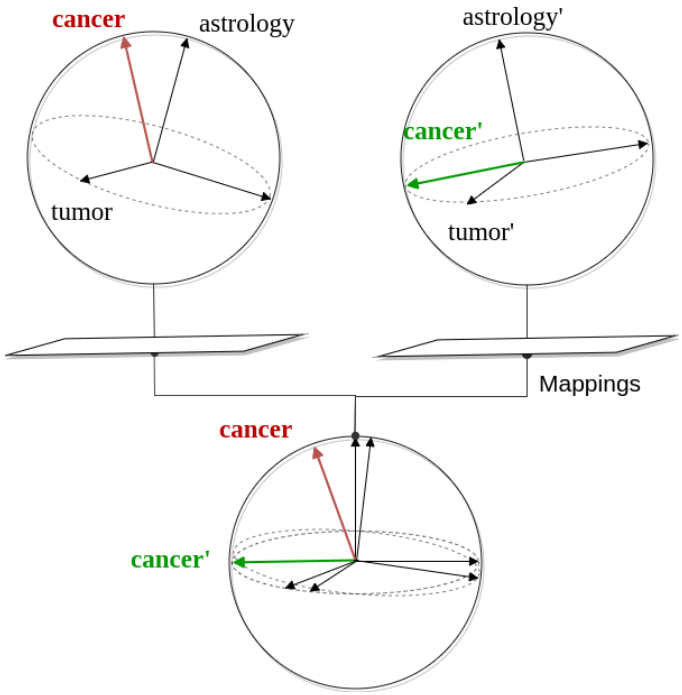
Για την επίλυση του παραπάνω προβλήματος πειραματιζόμαστε με τις γνωστές μεθόδους απεικόνισης που χρησιμοποιούνται στη βιβλιογραφία και έχουν ήδη περιγραφεί στο Κεφάλαιο 2.

- Linear
- Orthogonal
- Canonical Correlation Analysis

Προκειμένου να αποκτήσουμε πολλαπλές θεματικές αναπαραστάσεις που βρίσκονται σε έναν ενοποιημένο διανυσματικό χώρο, ακολουθούμε τον παραπάνω μετασχηματισμό για κάθε ΘΚΣΜ, που σημαίνει ότι για μια ομάδα θεμάτων που αποτελείται από K θέματα, μαθαίνουμε ένα σύνολο πινάκων $\{M_k\}_{k=1}^{K}$. Το σύνολο θεματικών αναπαραστάσεων που αντιστοιχούν σε μια συγκεκριμένη λέξη αντιπροσωπεύεται ως $\{s_k\}_{k=1}^{K}$. Συγκεκριμένα, δεδομένης μιας λέξης και της κατανεμημένης αναπαράστασης της $u_k \in \mathbb{R}^d$, υπολογίζουμε την προβαλόμενη αναπαράστασή της $s_k \in \mathbb{R}^d$ ως εξής:

$$s_k = M_k u_k. \tag{5.2}$$

Επιστρέφοντας στο προηγούμενο παράδειγμα, η λέξη *cancer* θα έχει αντιστοιχιστεί με πολλαπλές κατανεμημένες αναπαραστάσεις που αντικατοπτρίζουν τις διαφορετικές έννοιές της στον κοινό διανυσματικό χώρο. Μια απλοποιημένη απεικόνιση αυτού του παραδείγματος παρουσιάζεται στο Σχήμα 5.2.



Σχήμα 5.2: Απλοποιημένη απεικόνιση που συνοψίζει την ιδέα πίσω από τη διαδικασία ευθυγράμμισης των θεματικών αναπαραστάσεων. Στο ενοποιημένο σύστημα, η πολυσήμαντη λέξη *cancer* αντιπροσωπεύεται από δύο θεματικά διανύσματα που καταγράφουν διαφορετικές σημασιολογικές ιδιότητες της λέξης σε ένα ζωδιακό και ένα ιατρικό θέμα. Οι λέξεις *astrology* και *tumor* είναι παραδείγματα *σημασιολογικών αγκυρών* που ορίζουν τις αντιστοιχίσεις και διατηρούν τις σχετικές θέσεις των μονοσήμαντων-μονοσήμαντων και μονοσήμαντων-πολυσήμαντων λέξεων.

5.2.3 Εξομάλυνση θεματικών διανυσμάτων

Ξεκινώντας από το σύνολο των ευθυγραμμισμένων θεματικών αναπαραστάσεων $\{s_k\}_{k=1}^{K}$ για κάθε λέξη, πραγματοποιούμε συσσωρευτική ταξιδόμηση του συνόλου σε N κλάσεις, όπου οι κοντινά τοποθετημένες θεματικές αναπαραστάσεις μιας λέξης εκχωρούνται στην ίδια ομάδα. Θεωρούμε ότι κάθε ομάδα σχηματίζει μια σημασιολογικά συνεκτική οντότητα που αντιστοιχεί σε στενά σχετιζόμενες σημασίες της λέξης ενδιαφέροντος. Στη συνέχεια, τα διανύσματα της κάθε ομάδας αθροίζονται προκειμένου να δημιουργήσουν ένα αντιπροσωπευτικό διάνυσμα της ομάδας, οδηγώντας σε ένα νέο σύνολο *εξομαλυμένων* (smoothed) θεματικών αναπαραστάσεων $\{s'_n\}_{n=1}^{N}$ για κάθε λέξη, όπου $s'_n \in \mathbb{R}^d$.

5.3 Σημασιολογική Ομοιότητα

Αυτή η ενότητα περιγράφει τον τρόπο με τον οποίο αξιοποιούμε τις θεματικές αναπαραστάσεις του ενοποιημένου ΘΚΣΜ για τον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ δύο λέξεων. Παραδοσιακά, αυτό το πρόβλημα μετρά τον βαθμό ομοιότητας ενός ζεύγους λέξεων, ελλείψει συμφραζόμενου πλαισίου. Ωστόσο, το πραγματικό πρόβλημα που προσπαθούν να επιλύσουν τα συστήματα που αποδίδουν πολλαπλές κατανεμημένες αναπαραστάσεις σε λέξεις

είναι εκείνο της πολυσημικής φύσης των λέξεων που εξαρτάται σε μεγάλο βαθμό από τα συμφραζόμενα της λέξης. Για το λόγο αυτό, συζητάμε επίσης την εκτίμηση της σημασιολογικής ομοιότητας μεταξύ λέξεων, όπου συγκεκριμένες έννοιες των πολυσήμαντων λέξεων ενεργοποιούνται από παρεχόμενες συμφραζόμενες πληροφορίες. Γενικά, για να υπολογίσουμε τη σημασιολογική ομοιότητα μεταξύ ενός ζεύγους λέξεων (είτε παρέχονται μεμονωμένα είτε σε συμφραζόμενα πλαίσια), ακολουθούμε τις γνωστές μετρικές που χρησιμοποιούνται στη βιβλιογραφία από συστήματα πολλαπλών κατανεμημένων αναπαραστάσεων λέξεων.

5.3.1 Μετρικές βασισμένες σε συμφραζόμενα πλαίσια

Όταν παρέχονται πληροφορίες συμφραζόμενων πλαισίων για ένα ζεύγος λέξεων, χρησιμοποιούμε τις μετρικές AvgSimC και MaxSimC, οι οποίες συζητήθηκαν αρχικά από τους Reisinger and Mooney [2010]. Με δεδομένο το ζεύγος λέξεων (w, w') και τα παρεχόμενα πλαίσια τους (c, c') ορίζουμε:

$$\text{AvgSimC}(w, w') = \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} p(j|w, c) p(k|w', c') d(u_j(w), u_k(w')), \tag{5.3}$$

$$\text{MaxSimC}(w, w') = d(\hat{u}(w), \hat{u}(w')), \tag{5.4}$$

όπου K είναι το πλήθος των θεμάτων που επιστρέφονται από ένα εκπαιδευμένο μοντέλο LDA, u_j είναι το εκπαιδευμένο ΘΚΣΜ στο θεματικό υποσύνολο κειμένων που αντιστοιχεί στο j-οστό θέμα αφού προβάλλεται στον ενοποιημένο διανυσματικό χώρο, η $p(j|w, c)$ υποδηλώνει την posterior πιθανότητα του θέματος j που επιστρέφεται από τον LDA στον οποίο έχει δοθεί ως είσοδος το πλαίσιο c της λέξης w, το d υποδηλώνει την ομοιότητα μεταξύ των δύο αναπαραστάσεων εισόδου και τέλος το $\hat{u}(w) = u_{\arg\max_{1 \le k \le K} p(k|w)}(w)$ είναι το διάνυσμα της λέξης w που αντιστοιχεί στο θέμα με τη μέγιστη posterior για c.

Συνεπώς, η μετρική AvgSimC αντιστοιχεί σε χαλαρή εκχώρηση θεματικών περιοχών, σταθμίζοντας κάθε όρο ομοιότητας με την πιθανότητα των λέξεων να εμφανίζονται στα αντίστοιχα θέματα τους, ενώ η μετρική MaxSimC αντιστοιχεί σε σκληρή αντιστοίχιση χρησιμοποιώντας μόνο την πιο πιθανή εκχώρηση θέματος.

5.3.2 Μετρικές που δεν βασίζονται σε συμφραζόμενα πλαίσια

Όταν δεν παρέχονται συμφραζόμενα πλαίσια, ορίζουμε την ομοιότητα μεταξύ του ζεύγους λέξεων (w, w') χρησιμοποιώντας τις ακόλουθες μετρικές:

$$\text{MaxSim}(w, w') = \max_{\substack{1 \le j \le K \\ 1 \le k \le K}} d(u_j(w), u_k(w')), \tag{5.5}$$

$$\text{AvgSim}(w, w') = \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} d(u_j(w), u_k(w')). \tag{5.6}$$

Η μετρική MaxSim ενσωματώνει τη δημοφιλή υπόθεση της μέγιστης ομοιότητας των λέξεων σύμφωνα με την οποία: *η ομοιότητα μεταξύ δύο λέξεων είναι η μέγιστη ομοιότητα μεταξύ των όλων των πιθανών εννοιών τους*, ενώ η μετρική AvgSim προϋποθέτει ότι όλες οι πιθανές

αναπαραστάσεις μιας λέξης συμβάλλουν εξίσου στον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ των w και w'. Σημειώστε ότι οι AvgSim και MaxSim μπορούν να θεωρηθούν ως ειδικές περιπτώσεις των AvgSimC και MaxSimC με ομοιόμορφο βάρος σε κάθε θέμα. Συνεπώς, οι AvgSimC και MaxSimC μπορούν να χρησιμοποιηθούν για να συγκρίνουν ζεύγη λέξεων που εμφανίζονται τόσο παρουσία όσο και απουσία συμφραζόμενων πλαισίων.

Οι δύο παραπάνω μετρικές 5.5, 5.6 έχουν χρησιμοποιηθεί ευρέως από τα συστήματα πολλαπλών κατανεμημένων αναπαραστάσεων κατά τη σύγκριση των λέξεων για τις οποίες δεν διατίθεται κάποια συμφραζόμενη πληροφορία. Ωστόσο, δεν λαμβάνουν υπόψη ούτε την ποιότητα των αντίστοιχων αναπαραστάσεων μιας λέξης ούτε την κυριαρχία της «έννοιας» που αντιπροσωπεύει κάθε embedding. Καθώς δεν παρέχονται πληροφορίες συμφραζόμενων πλαισίων που να καθοδηγούν την επιλογή των θεματικών αναπαραστάσεων της κάθε λέξης, θα μπορούσε κανείς να χρησιμοποιήσει τεχνικές μηχανικής μάθησης για να ανακαλύψει ένα μοτίβο συντήξεως των θεματικών ομοιοτήτων που αντιστοιχούν στα υπό εξέταση ζεύγη. Αυτή η μέθοδος εφαρμόστηκε στο Κεφάλαιο 4, όπου χρησιμοποιήσαμε γραμμική παρεμβολή για να μάθουμε τους συντελεστές βάρους του συνδυαστικού μοντέλου μας. Ωστόσο οφείλουμε να σημειώσουμε οτι η μέθοδος αυτή δεν κλιμακώνει σε μεγάλο πλήθος αναπαραστάσεων ανά λέξη (απαιτούνται πολλά δεδομένα για την εκπαίδευση).

Οι Iacobacci et al. [2015] πρότειναν ένα άλλο κλιμακωτό ζυγισμένο σχήμα συγχώνευσης που ενσωματώνει θεματικές ομοιότητες λέξεων χρησιμοποιώντας τη μηχανική μάθηση. Τόνισαν μια ανεπάρκεια της μετρικής MaxSim βασισμένοι σε ψυχογλωσσικές μελέτες. Συγκεκριμένα, σημείωσαν ότι λαμβάνοντας υπόψη μόνο την ομοιότητα των πλησιέστερων εννοιών δύο λέξεων αγνοείται το γεγονός ότι οι άλλες έννοιες τους μπορούν επίσης να συμβάλλουν στη διαδικασία διαμόρφωσης της ομοιότητας. Στην πραγματικότητα, ψυχολογικές μελέτες υποδηλώνουν ότι οι άνθρωποι, όταν κρίνουν τη σημασιολογική ομοιότητα ενός ζεύγους λέξεων, λαμβάνουν υπόψη πολλές διαφορετικές έννοιες των δύο λέξεων και όχι μόνο τις πλησιέστερες. Για το λόγο αυτό, χρησιμοποίησαν μια σταθμισμένη μετρική ομοιότητας στην οποία διαφορετικές έννοιες των δύο λέξεων συμβάλλουν στον υπολογισμό της ομοιότητάς τους, ενώ οι συνεισφορές τους κλιμακώνονται ανάλογα με το πόσο κυρίαρχες είναι. Ακολουθώντας την σταθμισμένη στρατηγική ομοιότητας, χρησιμοποιούμε τη μετρική AvgSimW που αποτελεί μια τροποποιημένη έκδοση της μετρικής AvgSimC όταν δεν παρέχονται συμφραζόμενες πληροφορίες. Συγκεκριμένα, η ομοιότητα μεταξύ του ζεύγους λέξεων (w, w') ορίζεται ως εξής:

$$\text{AvgSimW}(w, w') = \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} p(j|w) p(k|w') d(u_j(w), u_k(w'))^{\alpha}, \tag{5.7}$$

όπου η συνεισφορά της j-οστής θεματικής αναπαράστασης της λέξης w σταθμίζεται από τον όρο $p(j|w)$ που υποδηλώνει τη δεσμευμένη πιθανότητα του θέματος j δεδομένης της w. Επιπλέον, σύμφωνα με τους Iacobacci et al. [2015], το βάρος/κυριαρχία μιας συγκεκριμένης έννοιας μιας λέξης μπορεί επίσης να εξαρτάται από τη λέξη με την οποία συγκρίνεται. Αυτό προσομοιώνεται μέσω της επιβολής ενός bias στον υπολογισμό της ομοιότητας των κοντινότερων εννοιών των υπό εξέταση λέξεων, με την αύξηση της συνεισφοράς τους μέσω μιας συνάρτησης υψωμένης εις την σταθερή δύναμη που αντιστοιχίζεται στην παράμετρο α.

Επιπλέον, για να αξιολογήσουμε την απόδοση του μοντέλου μας στο συμπυκνωμένο σύνολο των *εξομαλυμένων* θεματικών αναπαραστάσεων, οι πιθανότητες των σχέσεων 5.3, 5.4 και 5.7 αθροίζονται για κάθε ένα από τα ομαδοποιημένα διανύσματα.

Κεφάλαιο 6

Πειραματική Διαδικασία & Αποτελέσματα

Σε αυτό το κεφάλαιο, περιγράφουμε την πειραματική διαδικασία και τα αποτελέσματα των μοντέλων μας. Αρχικά, παρουσιάζουμε τη δημιουργία του συνόλου κειμένων που χρησιμοποιούμε, τις τιμές παραμέτρων των μοντέλων, τα σύνολα δεδομένων που χρησιμοποιούνται για σκοπούς αξιολόγησης και το baseline σύστημα μας (βλ. υποενότητα 6.1). Έπειτα, συνεχίζουμε με την αξιολόγηση του πρώτου μας μοντέλου (Μείγμα από ΘΚΣΜ) σε προβλήματα σημασιολογικής ομοιότητας λέξεων (βλ. υποενότητα 6.2). Στη συνέχεια, παρουσιάζουμε τα πειράματα του δεύτερου μοντέλου μας (ΘΚΣΜ πολλαπλών αναπαραστάσεων) που περιλαμβάνουν την αξιολόγηση των σημασιολογικών απεικονίσεων στο πρόβλημα της σημασιολογικής ομοιότητας λέξεων δοθέντων των συμφραζόμενων πλαισίων τους (βλ. υποενότητα 6.3), τη διερεύνηση μιας τεχνικής εξομάλυνσης (βλ. υποενότητα 6.4) και την παρουσίαση των καλύτερων προβλέψεων του μοντέλου για τον υπολογισμό σημασιολογικής ομοιότητας λέξεων απουσία συμφραζόμενων πλαισίων (βλ. υποκεφάλαιο 6.5). Συγκρίνουμε επίσης τα δύο μοντέλα μας με μοντέλα state-of-the-art που υπάρχουν στη βιβλιογραφία (βλ. υποενότητα 6.6). Τέλος, παρουσιάζουμε παραδείγματα απεικόνισης πολυσήμαντων λέξεων σε έναν δισδιάστατο χώρο όπως προκύπτουν από το δεύτερο μοντέλο (βλ. υποενότητα 6.7).

6.1 Πειραματικές ρυθμίσεις

Σύνολο κειμένων. Το κύριο σύνολο κειμένων που χρησιμοποιείται για την κατασκευή των μοντέλων μας είναι ένα γενικό σύνολο, το οποίο έχει συλλεχθεί από το διαδίκτυο και δημιουργείται σύμφωνα με τους Iosif and Potamianos [2015]. Ως πρώτο βήμα, απαιτείται ο ορισμός ενός λεξικού σε μια συγκεκριμένη γλώσσα. Για να το κάνουμε αυτό, χρησιμοποιούμε έναν κατάλογο αποτελούμενο από 8752 αγγλικά ουσιαστικά που εξάγονται από το συνολο κειμένων SemCor3. Στη συνέχεια, για κάθε λέξη του λεξικού διαμορφώνεται ένα μεμονωμένο ερώτημα (ένα "κείμενο αναζήτησης" που αποστέλλεται σε μια μηχανή αναζήτησης) και περιέχει μία λέξη. Έπειτα, κάθε ερώτημα υποβάλλεται ξεχωριστά στη μηχανή αναζήτησης Yahoo! και στη συνέχεια συλλέγονται τα 1000 έγγραφα με την υψηλότερη κατάταξη. Ακολούθως, από κάθε έγγραφο, αποσπάται το απόσπασμά του (μικρή παράγραφος κάτω από τη διεύθυνση URL του αποτελέσματος που συνήθως περιγράφει κάθε έγγραφο) και τελικά όλα τα αποσπάσματα από όλα τα ερωτήματα συγκεντρώνονται, οδηγώντας σε ένα σύνολο 116 εκατομμυρίων προ-

τάσεων. Όπως περιγράφεται στην Ενότητα 3, το μοντέλο μας απαιτεί τόσο μια έκδοση σε μορφή προτάσεων όσο και μια έκδοση σε μορφή εγγράφων. Προκειμένου να δημιουργηθεί η τελευταία, οι προτάσεις του συνόλου κειμένων συνδέονται διαδοχικά σε ομάδες ανα 100, σχηματίζοντας μια έκδοση του συνόλου κειμένων που αποτελείται απο 900 χιλιάδες έγγραφα. Τέλος, οι παραδοσιακές τεχνικές προεπεξεργασίας εφαρμόζονται και στις δύο εκδοχές του συνόλου κειμένων μας, συμπεριλαμβανομένου του tokenization (διάσπαση των προτάσεων σε tokens), της αφαίρεσης των stop-words (λέξεις που δεν φέρουν σημαντικές πληροφορίες όπως 'and' και 'to'), της αφαίρεσης σημείων στίξης και επαναλαμβανόμενων γραμμών και της μετατροπής των κεφαλαίων γραμμάτων σε πεζά.

Παράμετροι. Για την κατασκευή θεματικών ΚΣΜ, χρησιμοποιούμε την Gensim υλοποίηση του θεματικού μοντέλου Latent Dirichlet Allocation [Řehůřek and Sojka, 2010] και εκπαιδεύουμε 7 μοντέλα με το πλήθος θεμάτων να κυμαίνεται από 5 έως 60. Το κατώφλι που χρησιμοποιείται για τη χαλαρή ταξινόμηση του γενικού συνόλου κειμένων σε θεματικά υποσύνολα κειμένων ισούται με 0.1. Η υλοποίηση της Google του Word2vec μοντέλου χρησιμοποιείται για την εκπαίδευση τόσο των καθολικών και των θεματικών ΚΣΜ. Η παράμετρος που ορίζει το μέγεθος του συμφραζόμενου παραθύρου για το Word2vec τίθεται ίση με 5, ενώ πειραματιζόμαστε με διαστάσεις 100 και 300 για τα ΚΣΜ μοντέλα. Οποιαδήποτε παράμετρος που δεν αναφέρεται ορίζεται ίση με τις προεπιλεγμένες τιμές των αντίστοιχων υλοποιήσεων που χρησιμοποιούνται (π.χ. Word2Vec, Gensim LDA). Για μια λεπτομερή επεξήγηση των παραπάνω παραμετρικών επιλογών παραπέμπουμε τον αναγνώστη στη διπλωματική εργασία της Christopoulou [2016]. Όσον αφορά τις αντιστοιχίσεις και την εξομάλυνση των θεματικών ενσωματώσεων, χρησιμοποιούμε διαφορετικό πλήθος μονοσήμαντων ή τυχαίων λέξεων που κυμαίνονται από 1.000 έως 6.000 και διαφορετικό πλήθος κλάσεων που είναι ανάλογος με το πλήθος των θεματικών αναπαραστάσεων κάθε λέξης. Συγκεκριμένα, δεδομένου ενός συνόλου θεματικών αναπαραστάσεων με μέγεθος ίσο με K, το τελικό σύνολο των εξομαλυμένων αναπαραστάσεων έχει μέγεθος pK όπου το $p \in (0, 1]$. Συνολικά, σε αυτή την εργασία διερευνάμε τον ρόλο των ακόλουθων τεσσάρων παραμέτρων των μοντέλων μας:

- Πλήθος διαστάσεων των ΚΣΜ.
- Πλήθος θεμάτων.
- Πλήθος των λέξεων που χρησιμοποιούνται για τις αντιστοιχήσεις θεματικών αναπαραστάσεων.
- Πλήθος κλάσεων που χρησιμοποιούνται για την εξομάλυνση των θεματικών αναπαραστάσεων.

Σύνολα δεδομένων. Για να αξιολογήσουμε την απόδοση των μοντέλων μας, χρησιμοποιούμε σύνολα δεδομένων που παρέχουν ανθρώπινες εκτιμήσεις για τη σημασιολογική ομοιότητα μεταξύ ζευγών λέξεων. Χρησιμοποιώντας αυτές τις ανθρώπινες εκτιμήσεις ως αναφορικές/πρότυπες τιμές εκτιμούμε την αξιοπιστία των προβλέψεων των μοντέλων μας. Ειδικότερα, ο συντελεστής συσχέτισης του Spearman επιλέγεται ως μετρική αξιολόγησης για να συγκρίνει τις εκτιμώμενες ομοιότητές μας με τις πρότυπες τιμές. Η κατασκευή των παραπάνω συνόλων δεδομένων ακολουθεί συνήθως την ακόλουθη διαδικασία: ένα ζεύγος λέξεων παρουσιάζεται στους ανθρώπους για εκτίμηση, όπου κάθε εκτίμηση μετρά πόσο παρόμοιες είναι οι δύο λέξεις

σε μια προκαθορισμένη κλίμακα, στη συνέχεια οι εκτιμήσεις συγκεντρώνονται για να ληφθεί ένας μέσος όρος της ομοιότητας μεταξύ των δύο λέξεων. Για παράδειγμα, το ζεύγος (*automobile*, *vehicle*) λαμβάνει μια μέση τιμή 45 σε μια κλίμακα $[0, 50]$, γεγονός που υποδηλώνει ότι οι παραπάνω λέξεις αξιολογούνται ως παρόμοιες. Το παραπάνω ζεύγος αποτελεί ένα παράδειγμα λέξεων που παρουσιάζονται απουσία συμφραζόμενων πλαισίων στους ανθρώπινους σχολιαστές, οδηγώντας έτσι στη δημιουργία ενός συνόλου δεδομένων που αποτελείται από μεμονωμένες λέξεις. Από την άλλη πλευρά, μπορούν να δημιουργηθούν σύνολα δεδομένων με συμφραζόμενα πλαίσια μέσω της παροχής ενός συνθηματικού πλαισίου για κάθε λέξη κατά τη φάση εκτίμησης από τους σχολιαστές. Για παράδειγμα, το ζεύγος (*tiger*, *tiger*) λαμβάνει μια μέση τιμή ίση με 2 σε μια κλίμακα $[0, 10]$, όταν οι λέξεις παρουσιάζονται στα περιβάλλοντα (*... tiger with as many as of hunts ending in a kill reproduction over the course of her life*, *... famous sport figures like tiger woods ...*), όπου οι λέξεις αξιολογούνται ως πολύ ανόμοιες δεδομένων των αντίστοιχων πλαισίων τους. Για τα πειράματά μας χρησιμοποιούμε 3 σύνολα δεδομένων χωρίς συμφραζόμενη πληροφορία, και το μόνο διαθέσιμο σύνολο δεδομένων στο οποίο οι λέξεις παρουσιάζονται εντός ενός συμφραζόμενου πλαισίου. Οι περιγραφές τους παρουσιάζονται παρακάτω:

- Το **WordSimilarity-353** (WS-353) είναι ένα ευρέως χρησιμοποιούμενο σύνολο δεδομένων για την αξιολόγηση της σημασιολογικής ομοιότητας λέξεων. Αποτελείται από ένα σύνολο 353 ζευγών λέξεων μαζί με εκτιμήσεις ομοιότητας που έχουν καθοριστεί από ανθρώπους σε μία κλίμακα $[0, 10]$. Το WS-353 είναι μια συλλογή ζευγών για τη μέτρηση τόσο της ομοιότητας λέξεων όσο και της συγγένειας (relatedness), οπότε έχει χωριστεί σε δύο υποσύνολα, ένα για την αξιολόγηση της ομοιότητας και το άλλο για την αξιολόγηση της συγγένειας [Finkelstein et al., 2002].

- Το **MEN** περιέχει 3.000 ζεύγη τυχαία επιλεγμένων λέξεων προκειμένου να εξασφαλιστεί ένα ισορροπημένο εύρος ομοιότητας στα επιλεγμένα ζεύγη λέξεων. Οι πρότυπες εκτιμήσεις αξιολογήθηκαν σε μια κλίμακα $[0, 50]$ με βάση τον πληθοπορισμό [Bruni et al., 2014].

- Το **RG** ή Rubenstein and Goodenough είναι ένα σύνολο 65 ζεύγη με εκτιμήσεις ομοιότητας υπολογισμένες από ανθρώπους σε μια κλίμακα $[0, 4]$ [Rubenstein and Goodenough, 1965].

- Το **SCWS** ή αλλιώς το Stanford Contextual Word Similarity σύνολο δεδομένων αποτελεί το μοναδικό σύνολο που παρέχει συμφραζόμενα πλαίσια λέξεων και δημοσιεύτηκε από τους Huang et al. [2012]. Αποτελείται από 2.003 ζεύγη λέξεων μαζί με φράσεις που περιέχουν αυτές τις λέξεις ενώ οι ομοιότητες έχουν βαθμολογηθεί σε μια κλίμακα $[0, 10]$.

Baseline. Η απόδοση του καθολικού ΚΣΜ μοντέλου χρησιμοποιείται στα ακόλουθα πειράματα ως το βασικό μας σύστημα (baseline). Θυμίζουμε ότι το καθολικό ΚΣΜ (global-DSM) παρέχει ένα αντιπροσωπευτικό διάνυσμα για κάθε λέξη και ως εκ τούτου δεν λαμβάνει υπόψη την πολυσημική φύση των λέξεων.

6.2 Σημασιολογικά Μείγματα

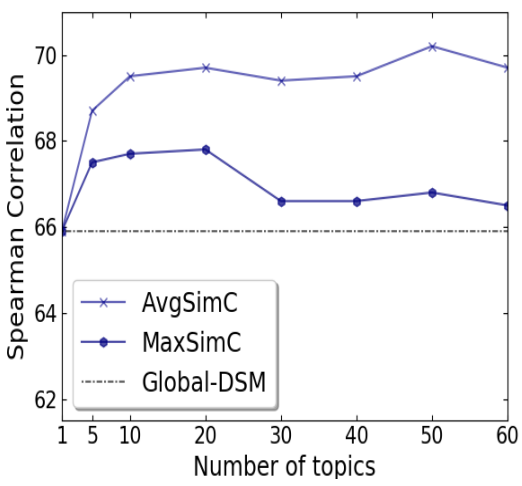
Σε αυτή την ενότητα παρουσιάζουμε αποτελέσματα αξιολόγησης που έχουν εξαχθεί χρησιμοποιώντας το Μείγμα Θεματικών Κατανεμημένων Σημασιολογικών Μοντέλων (ΘΚΣΜ), όπως περιγράφεται στο Κεφάλαιο 4. Η απόδοση του ΘΚΣΜ μοντέλου, για κάθε συνδυασμό σημασιολογικών ομοιοτήτων, απεικονίζεται στο Σχήμα 6.1 (α΄) ως συνάρτηση του πλήθους των θεμάτων για το σύνολο δεδομένων SCWS. Η κορυφαία απόδοση (70.2) επιτυγχάνεται με τον συνδυασμό AvgSimC όταν χρησιμοποιούνται 40 - 50 θέματα. Τα Σχήμα 6.1 (β΄) απεικονίζει την απόδοση του συνδυασμού γραμμικής παρεμβολής των ΘΚΣΜ, ως συνάρτηση του πλήθους των θεμάτων. Όσον αφορά τα σύνολα δεδομένων MEN και WS-353, η προτεινόμενη προσέγγιση φαίνεται να έχει καλύτερες επιδόσεις από το baseline μοντέλο για όλο τον αριθμό θεμάτων και για τα δύο σύνολα δεδομένων. Η κορυφαία τιμή συσχέτισης (83.8) επιτυγχάνεται για 40 θέματα στο σύνολο δεδομένων MEN. Για το σύνολο δεδομένων WS-353, ο ίδιος συνδυασμός θεμάτων παρέχει την κορυφαία απόδοση (72.7). Στη συνέχεια, συζητάμε πιο αναλυτικά τα αποτελέσματα που ελήφθησαν για κάθε σχήμα συνδυασμού θεματικών ομοιοτήτων όπως περιγράφεται στο Κεφάλαιο 4. Οι καλύτερες προβλέψεις για κάθε έναν από αυτούς παρατίθενται στον Πίνακα 6.1.

Η βελτίωση που επιτυγχάνεται με την προτεινόμενη προσέγγιση σε σχέση με την επίδοση του baseline μοντέλου για το πρόβλημα υπολογισμού της σημασιολογικής ομοιότητας, αποδεικνύεται μέσα από την χρήση τριών συνόλων δεδομένων. Παρατηρούμε ότι η κορυφαία απόδοση (72.7), που επιτυγχάνεται για το σύνολο δεδομένων WS-353, είναι χαμηλότερη σε σύγκριση με τη υψηλότερη απόδοση συσχετισης (83.8) που αποκτήθηκε για το MEN. Αυτό μπορεί να αποδοθεί στις διαφορετικές μορφές των δύο συνόλων δεδομένων, π.χ. τον τύπο της σημασιολογικής σχέσης, καθώς και τη διαδικασία που ακολουθείται για τη συλλογή ανθρώπινων αξιολογήσεων. Συνολικά, η αναφερθείσα βελτίωση για το σύνολο δεδομένων MEN είναι στατιστικά πιο σημαντική σε σύγκριση με την περίπτωση του WS-353 λόγω του μεγαλύτερου μεγέθους του συνόλου δεδομένων MEN (3000 έναντι 353 ζευγών).

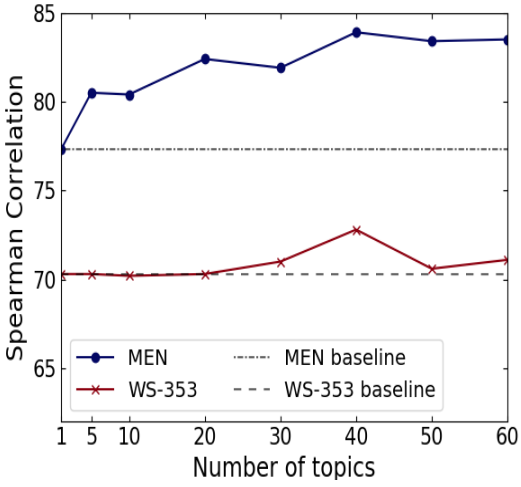
Το πλήθος των θεμάτων αποτελεί βασική παράμετρο της προτεινόμενης προσέγγισης. Τα

Mixture	In-context		Out-of-context	
	WS-353	MEN	SCWS	
			MaxSimC	AvgSimC
Global-DSM	70.3	77.3	65.9	65.9
TDSMs	-	-	67.8	70.2
TDSMs-Fuse	-	-	67.4	70.5
TDSMs-LR	72.2	83.8	-	-
TDSMs-MaxSim	72.2	80.0	-	-
TDSMs-AvgSim	71.3	79.6	-	-

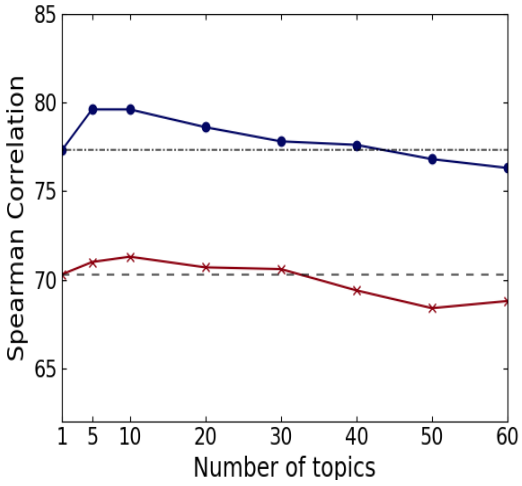
Πίνακας 6.1: Συγκριτική απόδοση μεταξύ των καλύτερων αποτελεσμάτων που λαμβάνονται για διαφορετικά σχήματα συνδυασμού θεματικών σημασιολογικών μοντέλων και διαφορετικά σύνολα δεδομένων, για τον υπολογισμό της σημασιολογικής ομοιότητας ζευγών λέξεων, χρησιμοποιώντας τη μετρική συσχέτισης Spearman ρ.



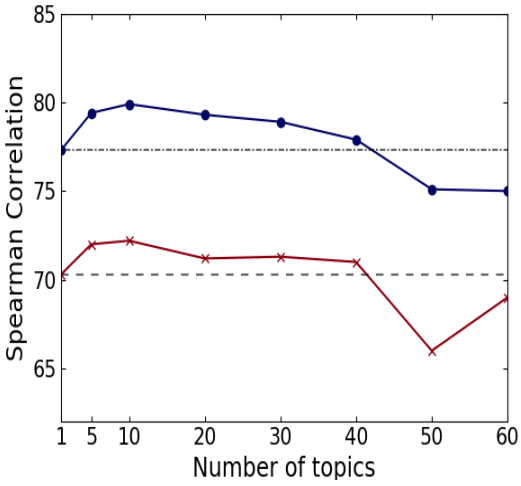
(α′) Contextual



(β′) Non-Contextual, LinearRegression



(γ′) Non-Contextual, AvgSim



(δ′) Non-Contextual, MaxSim

Σχήμα 6.1: Σύγκριση απόδοσης για διαφορετικούς αριθμούς θεμάτων και διαφορετικούς συνδυασμούς ΘΚΣΜ, χρησιμοποιώντας τη μετρική συσχέτισης Spearman. Το γράφημα (α′) απεικονίζει την απόδοση των συνδυασμών AvgSimC και MaxSimC στο σύνολο δεδομένων SCWS. Τα γραφήματα (β′), (γ′) και (δ′) απεικονίζουν τις επιδόσεις των συνδυασμών γραμμικής παρεμβολής, AvgSim και MaxSim, αντίστοιχα, στα σύνολα δεδομένων MEN και WS-353. Επίσης παρουσιάζονται οι αποδόσεις του baseline μοντέλου.

προσδιορισμένα θέματα χρησιμοποιούνται για το φιλτράρισμα των θεματικών υποσυνόλων κειμένων στα οποία βασίζεται η δημιουργία των ΘΚΣΜ. Σε αυτό το πλαίσιο, όταν υπολογίζουμε την ομοιότητα μεταξύ ενός ζεύγους λέξεων, υποστηρίζουμε ότι το χρησιμοποιούμενο υποσύνολο κειμένων έχει δύο ιδιότητες: i) το υποσύνολο είναι σημασιολογικά συνεπές, δηλαδή οι δύο λέξεις πρέπει να εμφανίζονται με τις πλησιέστερες έννοιες τους σε αυτό και ii) υπάρχουν επαρκή δεδομένα που επιτρέπουν τον υπολογισμό των ΚΣΜ. Συνήθως, ένας μεγαλύτερος

αριθμός θεμάτων βελτιώνει τη σημασιολογική συνοχή του αντίστοιχου υποσυνόλου κειμένων (αυξημένη ειδικότητα του θέματος), αλλά μπορεί να προκαλέσει τον κατακερματισμό των δεδομένων εκπαίδευσης, μειώνοντας την ποιότητα των σημασιολογικών μοντέλων. Αυτό το πρόβλημα είναι περισσότερο εμφανές στις περιπτώσεις όπου χρησιμοποιούνται οι συνδυασμοί MaxSim και AvgSim όπως φαίνεται στα Σχήματα 6.1 (γ′) και (δ′) για μεγάλες τιμές της παραμέτρου K.

Προκειμένου να ξεπεράσουμε το πρόβλημα αυτό, θεωρούμε οτι η προσέγγιση γραμμικής παρεμβολής είναι κατάλληλη καθώς οδηγεί στην επιλογή των καλύτερων ομοιοτήτων από τα αντίστοιχα θεματικά ΚΣΜ, για ζεύγη λέξεων που εμφανίζονται χωρίς συμφραζόμενο πλαίσιο όπως φαίνεται στο Σχήμα 6.1 (β′). Η μέθοδος ξεπερνά το baseline μοντέλο για πολύ μικρό αριθμό θεμάτων, αλλά φαίνεται να λειτουργεί καλύτερα για μεγαλύτερο αριθμό θεμάτων παρά τον κατακερματισμό των δεδομένων. Αυτή η συμπεριφορά μπορεί να εξηγηθεί θεωρώντας ότι χωρίς ένα δεδομένο συμφραζόμενο πλαίσιο, μια λέξη θα μπορούσε να έχει έναν αυθαίρετο αριθμό εννοιών. Ως αποτέλεσμα, ένας αυξανόμενος χώρος εννοιών επιτρέπει την ακριβή εκτίμηση της ομοιότητας ενός ζεύγους λέξεων μέσω του γραμμικού συνδυασμού των διαφορετικών θεματικών ομοιοτήτων του.

Τέλος, όπως φαίνεται στον Πίνακα 6.1, το μοντέλο σύντηξης (fusion) βελτιώνει την απόδοση της καλύτερης προγνωστικής διαμόρφωσης που αντιστοιχεί στον συνδυασμό θεματικών ομοιοτήτων που εξάγεται από μια ομάδα θεμάτων G. Συγκεκριμένα, το μοντέλο σύντηξης παρέχει τα καλύτερα αποτελέσματα όταν χρησιμοποιηθούν όλες οι ομάδες θεμάτων. Αυτό είναι αναμενόμενο καθώς η μέθοδος σύντηξης λειτουργεί σαν ένα ιεραρχικό θεματικό μοντέλο. Τα ιεραρχικά θεματικά μοντέλα χαλαρώνουν την υπόθεση μίας μοναδικής κατανομής θεμάτων πάνω στο αρχικό σύνολο κειμένων. Έτσι, επιλέγοντας τη μέγιστη ομοιότητα σε διάφορες πιθανές κατανομές, μπορεί να προσεγγιστεί το πραγματικό πλήθος των εννοιών που αντιστοιχούν σε κάθε λέξη.

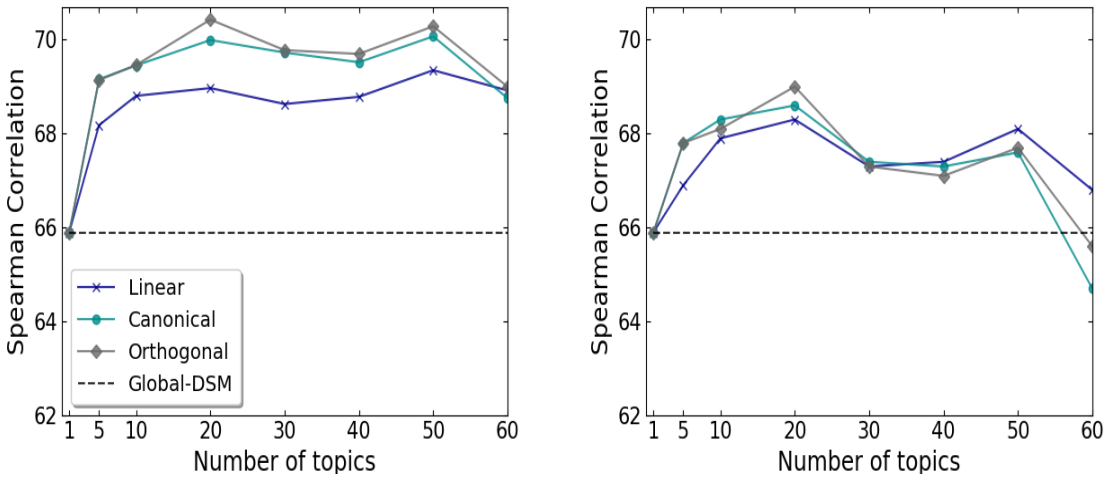
6.3 Σημασιολογικές απεικονίσεις

Τα παρακάτω πειράματα αντιστοιχούν στο δεύτερο μοντέλο που περιγράψαμε στο Κεφάλαιο 5. Συγκεκριμένα, εξετάζουμε τον ρόλο: (1) του αλγορίθμου που χρησιμοποιείται για τον καθορισμό των σημασιολογικών απεικονίσεων, (2) του αριθμού των λέξεων των οποίων οι κατανεμημένες αναπαραστάσεις χρησιμεύουν ως σημασιολογικες άγκυρες μεταξύ των χώρων πηγής και στόχου, (3) της πολυσημίας των σημασιολικών αγκυρών.

6.3.1 Μέθοδοι Απεικόνισης

Το Σχήμα 6.2 παρουσιάζει τα αποτελέσματα του πρώτου μας πειράματος. Στόχος μας είναι να αξιολογήσουμε την αξιοπιστία των σημασιολογικών μας απεικονίσεων χρησιμοποιώντας διαφορετικές τεχνικές που υιοθετούνται από τη βιβλιογραφία. Για το σκοπό αυτό, χρησιμοποιούμε πρότυπες λίστες σημασιολογικών αγκυρών αποτελούμενων από συχνές μονοσήμαντες λέξεις για κάθε θέμα και εφαρμόζουμε τις τρεις ακόλουθες τεχνικές: Linear, Orthogonal, Canonical Correlation Analysis.

Παρατηρούμε ότι γενικά η ορθογώνια μέθοδος σημειώνει ελαφρώς καλύτερη απόδοση σε σύγκριση με την κανονική μέθοδο, ενώ η γραμμική μέθοδος έχει ως αποτέλεσμα την χαμηλότερη απόδοση ως προς τη μετρική AvgSimC για όλα τα θέματα. Οι διαφορές απόδοσης μεταξύ



(α′) AvgSimC (β′) MaxSimC

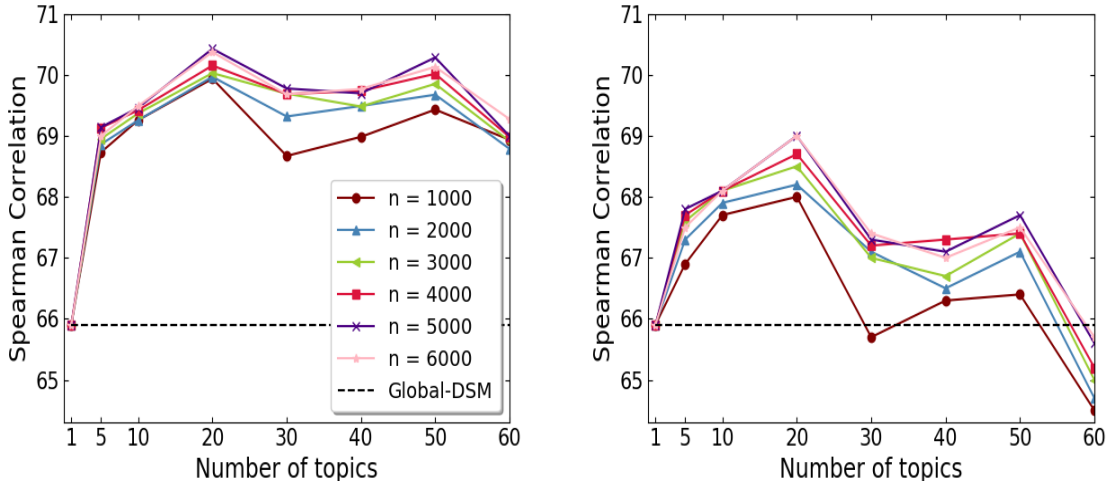
Σχήμα 6.2: Σύγκριση απόδοσης για διαφορετικούς αριθμούς θεμάτων και διαφορετικούς αλγορίθμους απεικόνισης, χρησιμοποιώντας τις μετρικές (α′) AvgSimC και (β′) MaxSimC και τη μετρική συσχέτισης Spearman, στο σύνολο δεδομένων SCWS. Η διάσταση των σημασιολογικών χώρων έχει τεθεί ίση με 300 ενώ λίστες 5000 συχνών μονοσήμαντων λέξεων έχουν χρησιμοποιηθεί ως *σημασιολογικές άγκυρες* για κάθε θέμα. Επιπλέον απεικονίζεται η απόδοση του baseline μοντέλου.

των μεθόδων απεικόνισης δεν είναι τόσο εμφανείς όσον αφορά τη μετρική MaxSimC. Επιπλέον, οι απεικονιζόμενες επιδόσεις έρχονται σε συμφωνία με τα αποτελέσματα της βιβλιογραφίας υποδεικνύοντας ότι οι ορθογώνιοι μετασχηματισμοί οδηγούν σε καλύτερες αντιστοιχίσεις μεταξύ σημασιολογικών χώρων μονόγλωσσων ή δίγλωσσων δεδομένων. Για μια λεπτομερή σύγκριση των μεθόδων απεικόνισης παραπέμπουμε τον αναγνώστη στη μελέτη των Artetxe et al. [2018]. Για την υπόλοιπη εργασία επιλέγουμε τους ορθογώνιους μετασχηματισμούς ως τεχνική σημασιολογικών απεικονίσεων.

6.3.2 Πλήθος μονοσήμαντων λέξεων

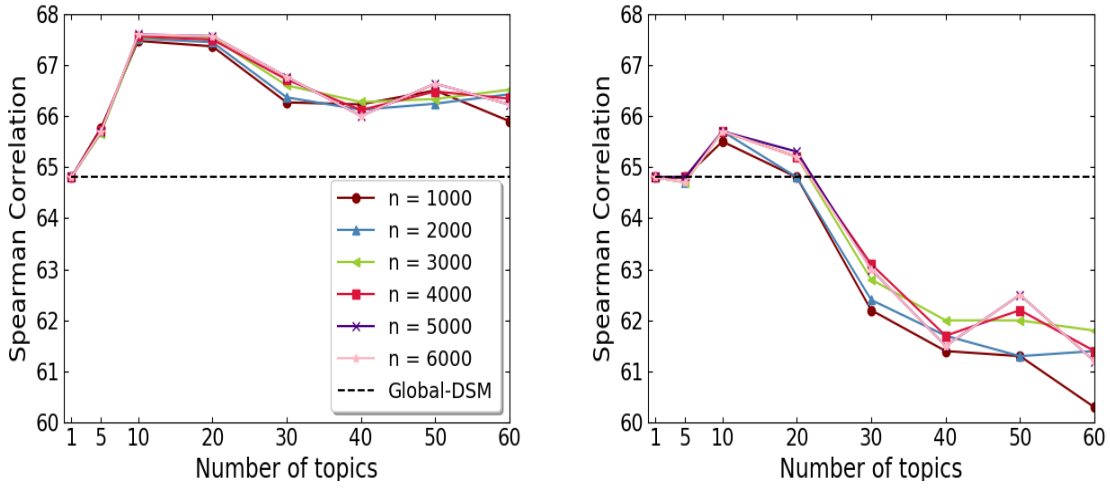
Στο Σχήμα 6.3 δείχνουμε την απόδοση του μοντέλου μας χρησιμοποιώντας ορθογώνιες απεικονίσεις και διαφορετικό πλήθος μονοσήμαντων λέξεων n, όταν οι διαστάσεις των χώρων προέλευσης και στόχου έχει τεθεί ίσες με 300. Αρχικά παρατηρούμε μια σαφή σχέση μεταξύ της συσχέτισης των προβλέψεών μας και του αριθμού των χρησιμοποιούμενων λέξεων: Μια μεγαλύτερη λίστα σημασιολογικών αγκυρών οδηγεί στην επαγωγή πιο αξιόπιστων σημασιολογικών απεικονίσεων επιτυγχάνοντας τις κορυφαίες επιδόσεις. Αυτή η συμπεριφορά είναι αναμενόμενη, δεδομένου του μεγάλου αριθμού των παραμέτρων που προσπαθούμε να μάθουμε κατά τους μετασχηματισμούς μας. Θυμίζουμε ότι κάθε πίνακας μετασχηματισμού $M_k \in \mathbb{R}^{d \times d}$, όπου d είναι η διάσταση των σημασιολογικών χώρων μας.

Έπειτα παρατηρούμε ότι η συσχέτιση του μοντέλου μας παρουσιάζει πολύ μικρές διαφορές απόδοσης όταν ο αριθμός των λέξεων αντιστοιχεί σε μεγάλες τιμές. Συγκεκριμένα, η απόδοση δεν βελτιώνεται όταν χρησιμοποιούμε περισσότερες από 5.000 μονοσήμαντες λέξεις ως σημα-



(α′) AvgSimC (β′) MaxSimC

Σχήμα 6.3: Σύγκριση απόδοσης για διαφορετικά πλήθη θεμάτων και διαφορετικούς αριθμούς μονοσήμαντων λέξεων n που χρησιμεύουν ως *σημασιολογικές άγκυρες* στις ορθογώνιες απεικονίσεις. Ως μετρικές έχουν χρησιμοποιηθεί η συσχέτιση Spearman και οι (α′) AvgSimC και (β′) MaxSimC στο SCWS σύνολο δεδομένων. Η **διάσταση** των σημασιολογικών χώρων ισούται με **300**. Επιπλέον απεικονίζεται η απόδοση του baseline μοντέλου.



(α′) AvgSimC (β′) MaxSimC

Σχήμα 6.4: Σύγκριση απόδοσης για διαφορετικά πλήθη θεμάτων και διαφορετικούς αριθμούς μονοσήμαντων λέξεων n που χρησιμεύουν ως *σημασιολογικές άγκυρες* στις ορθογώνιες απεικονίσεις. Ως μετρικές έχουν χρησιμοποιηθεί η συσχέτιση Spearman και οι (α′) AvgSimC και (β′) MaxSimC στο SCWS σύνολο δεδομένων. Η **διάσταση** των σημασιολογικών χώρων ισούται με **100**. Επιπλέον απεικονίζεται η απόδοση του baseline μοντέλου.

σιολογικές άγκυρες. Αυτό θα μπορούσε να αποδοθεί στο γεγονός ότι εισάγονται στο μοντέλο μας πιο σπάνιες μονοσήμαντες λέξεις καθώς αυξάνουμε το μέγεθος της λίστας των σημασιολογικών αγκυρών. Ως αποτέλεσμα, παρόλο που προσαρμόζουμε περισσότερα δεδομένα στο μοντέλο μας, η αξιοπιστία τους είναι μειωμένη και δεν οδηγεί σε μεγαλύτερη ακρίβεια των προβλέψεών μας.

Για να παρακάμψουμε αυτό το πρόβλημα, μειώνουμε τον αριθμό των παραμέτρων που πρέπει να μάθουμε, δηλαδή τον αριθμό στοιχείων για κάθε πίνακα μετασχηματισμού M_k. Στο Σχήμα 6.4 παρουσιάζουμε τα αποτελέσματα των πειραμάτων μας χρησιμοποιώντας διαφορετικό αριθμό μονοσήμαντων λέξεων n, όταν η διασταση των σημασιολογικών μας χώρων έχει οριστεί ίση με 100. Γενικά, η απόδοση του μοντέλου μας παρουσιάζει μια πιο συνεπή συμπεριφορά ως προς την μετρική AvgSimC, σε σύγκριση με τη μετρική MaxSimC. Όπως αναμενόταν, οι διαφορές μεταξύ των επιδόσεων του μοντέλου μας δεν είναι τόσο εμφανείς όσο ο αριθμός των λέξεων αυξάνεται, πράγμα που σημαίνει ότι ο αριθμός των σημασιολογικών αγκυρών δεν αποτελεί βασική παράμετρο του προβλήματός μας όταν χρησιμοποιούμε κατανεμημένες αναπαραστάσεις χαμηλότερων διαστάσεων. Ωστόσο, αν και το προηγούμενο πρόβλημα μετριάζεται με τη μείωση των διαστάσεων των διανυσμάτων, ένα άλλο πρόβλημα εμφανίζεται: η γενική απόδοση του μοντέλου μας είναι μειωμένη και η διαφορά του από την απόδοση του baseline μοντέλου μειώνεται για όλες τις επιλογές θεμάτων. Αυτή η παρατήρηση δείχνει ότι η πολυπλοκότητα των σημασιολογικών μας χώρων περιγράφεται καλύτερα χρησιμοποιώντας 300 διαστάσεις αντί για 100.

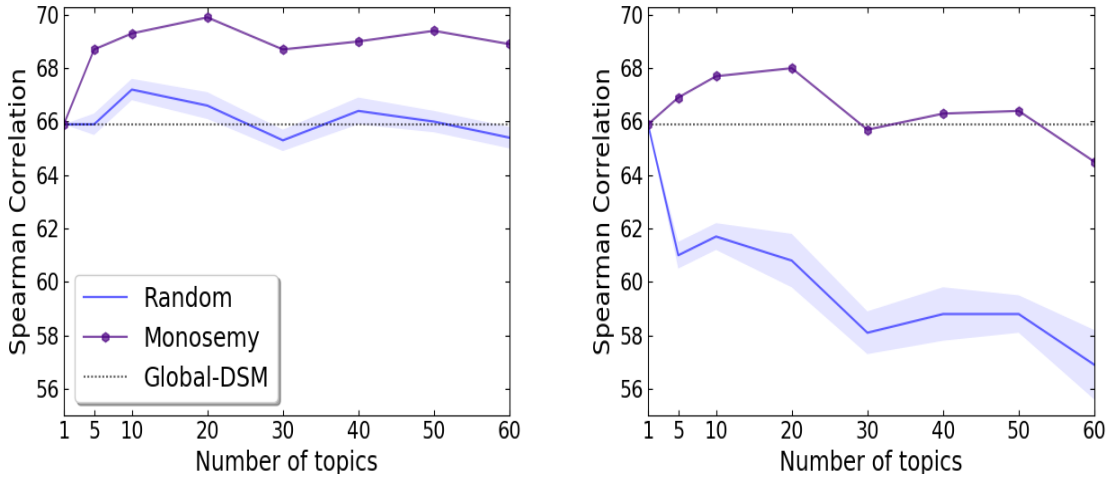
6.3.3 Σημασιολογικές Άγκυρες

Για να διερευνήσουμε τον ρόλο των *σημασιολογικών αγκυρών* στον προσδιορισμό των αντιστοιχιών μεταξύ των σημασιολογικών χώρων προέλευσης και στόχου διεξάγουμε δύο πειράματα. Μετά τις καλύτερες διαμορφώσεις που προέκυψαν από τα προηγούμενα πειράματα, χρησιμοποιούμε τη μέθοδο ορθογώνιας απεικόνισης και μια λίστα V_k που αποτελείται από n λέξεις για την ευθυγράμμιση του k-οστού θεματικού χώρου. Τα δύο πειράματα περιγράφονται ως εξής:

- Για το πρώτο πείραμα, $|V_k|$ σημασιολογικές άγκυρες εξάγονται για κάθε θέμα από τη λίστα των πιο συχνών μονοσήμαντων λέξεων (από το υποσύνολο κειμένων του συγκεκριμένου θέματος). Η πρωτότυπη (κύρια) λίστα μονοσήμαντων λέξεων εξάγεται από το WordNet [Fellbaum, 1998]. Για να ορίσουμε τη μονοσημία ακολουθούμε την ορολογία που παρέχεται από το WordNet, σύμφωνα με την οποία: *μια λέξη είναι μονοσήμαντη όταν έχει μία μόνο αίσθηση/έννοια σε κάθε συντακτική κατηγορία.*

- Για το δεύτερο πείραμα, χρησιμοποιώντας το κοινό σύνολο λέξεων που αναπαριστάται τόσο στον τόπο προέλευσης όσο και στον χώρο στόχου επιλέγουμε τυχαία $|V_k|$ λέξεις ως σημασιολογικές άγκυρες. Επαναλαμβάνουμε αυτό το πείραμα 10 φορές, λαμβάνοντας κάθε φορά δείγματα από μια διαφορετική λίστα από το κοινό σύνολο λέξεων και συλλέγουμε στατιστικά αποτελέσματα της απόδοσης που αντιστοιχούν στη μέση τιμή και την τυπική απόκλιση των μεμονωμένων πειραμάτων.

Στα Σχήματα 6.5 και 6.6 απεικονίζουμε την απόδοση του μοντέλου μας όταν χρησιμοποιούμε λίστες μονοσήμαντων και λίστες τυχαίων λέξεων για να καθορίσουμε τις απεικονίσεις



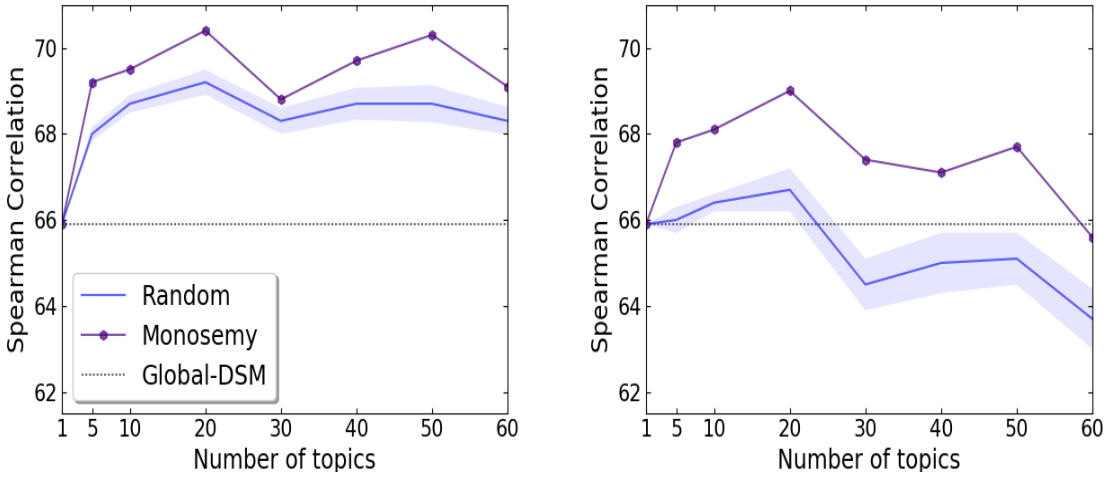
(α′) AvgSimC (β′) MaxSimC

Σχήμα 6.5: Σύγκριση απόδοσης για διαφορετικά πλήθη θεμάτων και απεικονίσεων που αποκτήθηκαν χρησιμοποιώντας λίστες μονοσήμαντων και λίστες τυχαίων λέξεων που χρησιμεύουν ως *σημασιολογικές άγκυρες*. Ως μετρικές έχουν χρησιμοποιηθεί η συσχέτιση Spearman και οι (α′) AvgSimC και (β′) MaxSimC στο SCWS σύνολο δεδομένων. Το **πλήθος** των σημασιολογικών αγκυρών που χρησιμοποιούνται ισούται με **1000**. Επιπλέον απεικονίζεται η απόδοση του baseline μοντέλου.

μεταξύ του χώρου πηγής και του χώρου προέλευσης. Συγκεκριμένα, στο Σχήμα 6.5 παρουσιάζουμε τα αποτελέσματα των πειραμάτων όταν ο αριθμός των λέξεων της κάθε λίστας περιορίζεται σε 1.000, ενώ το Σχήμα 6.6 παρουσιάζει τα αποτελέσματα του ίδιου πειράματος χρησιμοποιώντας 5.000 λέξεις. Επιπλέον, οι οριζόντιες γραμμές ($K = 1$) δείχνουν την απόδοση του καθολικού ΚΣΜ, το οποίο χρησιμοποιείται ως το *baseline* σύστημά μας και οι σκιασμένες περιοχές αντιπροσωπεύουν την απόκλιση της απόδοσης χρησιμοποιώντας διαφορετικές λίστες τυχαίων λέξεων.

Για ένα σταθερό πλήθος θεμάτων K, το μοντέλο μας επιτυγχάνει σταθερά υψηλότερες επιδόσεις όταν οι μονοσήμαντες λέξεις χρησιμοποιούνται ως *σημασιολογικές άγκυρες* έναντι τυχαία επιλεγμένων λέξεων, όπως απεικονίζεται και στα δύο Σχήματα 6.5 και 6.6. Στα Σχήματα 6.5, 6.6 (β′), παρατηρούμε ότι η απόδοση της μετρικής MaxSimC πέφτει κάτω από το *baseline* σύστημα για $K > 1$ και $K > 20$ αντίστοιχα, όταν τυχαία επιλεγμένες λέξεις χρησιμεύουν ως *σημασιολογικές άγκυρες*, ενώ αυτό δεν συμβαίνει όταν χρησιμοποιούνται μονοσήμαντες λέξεις ως άγκυρες για τις περισσότερες απεικονίσεις. Αυτό το αποτέλεσμα επικυρώνει την υπόθεσή μας ότι οι κατανεμημένες αναπαραστάσεις μονοσήμαντων λέξεων αποτελούν *σημασιολογικές άγκυρες* που καθορίζουν τις απεικονίσεις μεταξύ των σημασιολογικών διανυσματικών χώρων.

Επιπλέον, η παραπάνω παρατήρηση είναι πιο εμφανής όταν χρησιμοποιούμε λιγότερες λέξεις για να μάθουμε τις απεικονίσεις, όπως φαίνεται στο Σχήμα 6.5. Σημειώνουμε επίσης ότι η απόδοση του μοντέλου μας δεν μπορεί να ξεπεράσει την απόδοση του baseline μοντέλου όταν χρησιμοποιούνται $n = 1.000$ τυχαίες λέξεις για σχεδόν όλες τις επιλογές αριθμού θεμάτων. Από την άλλη πλευρά, οι επιδόσεις του μοντέλου μας ξεπερνούν σαφώς την απόδοση του



(α′) AvgSimC (β′) MaxSimC

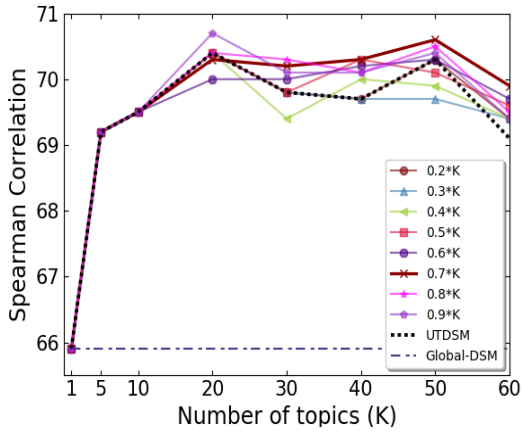
Σχήμα 6.6: Σύγκριση απόδοσης για διαφορετικά πλήθη θεμάτων και απεικονίσεων που αποκτήθηκαν χρησιμοποιώντας λίστες μονοσήμαντων και λίστες τυχαίων λέξεων που χρησιμεύουν ως *σημασιολογικές άγκυρες*. Ως μετρικές έχουν χρησιμοποιηθεί η συσχέτιση Spearman και οι (α′) AvgSimC και (β′) MaxSimC στο SCWS σύνολο δεδομένων. Το **πλήθος** των σημασιολογικών αγκυρών που χρησιμοποιούνται ισούται με **5000**. Επιπλέον απεικονίζεται η απόδοση του baseline μοντέλου.

baseline μοντέλου, όταν χρησιμοποιούνται $n = 1.000$ μονόσημες λέξεις. Αυτή η παρατήρηση δείχνει ότι η χρήση μονοσήμαντων σημασιολογικών αγκυρών μπορεί να επιτύχει ενδιαφέροντα αποτελέσματα ακόμη και αν δεν είναι διαθέσιμη μεγάλη ποσότητα δεδομένων.

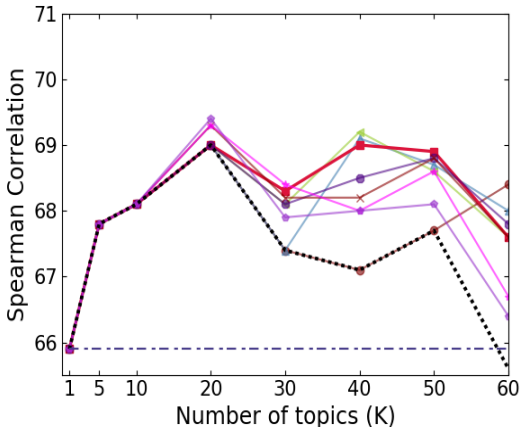
6.4 Εξομάλυνση θεματικών αναπαραστάσεων

Ως ένα επιπλέον βήμα πειραματιζόμαστε με μια τεχνική εξομάλυνσης χρησιμοποιώντας ιεραρχική ομαδοποίηση του συνόλου αναπαραστάσεων που αντιστοιχούν σε μια λέξη στο ενοποιημένο ΘΚΣΜ. Για να εκτελέσουμε την ιεραρχική ομαδοποίηση των θεματικών αναπαραστάσεων, χρησιμοποιούμε την υλοποίηση Scikit-learn της συσσωρευτικής ταξιδομησης [Pedregosa et al., 2011]. Το μέγεθος του συνόλου που αποτελείται από τις *εξομαλυμένες* θεματικές αναπαραστάσεις ορίζεται ως ποσοστό του αριθμού των θεμάτων K. Συγκεκριμένα, πειραματιζόμαστε με $n = \max\{10, pK\}$ κλάσεις, όπου $p \in (0, 1]$ (σημειώνουμε ότι όταν δεν εφαρμόζεται η τεχνική εξομάλυνσης $p = 1$). Οι τιμές του N επιλέγονται έτσι ώστε να είναι μεγαλύτερες της τιμής 10, καθώς υποθέτουμε ότι η τεχνική εξομάλυνσης θα πρέπει να εφαρμόζεται σε θορυβώδεις αναπαραστάσεις που εισάγονται στο μοντέλο μας όταν το K είναι μεγάλο. Επιπλέον, πειραματιζόμαστε με διαφορετικά κριτήρια σύνδεσης (linkage criteria) τα οποία χρησιμοποιούνται για τον ορισμό της απόστασης μεταξύ των κλάσεων, ενώ η απόσταση συνημιτόνου[1] χρησιμοποιείται για τον καθορισμό των αποστάσεων μεταξύ των διανυσμάτων.

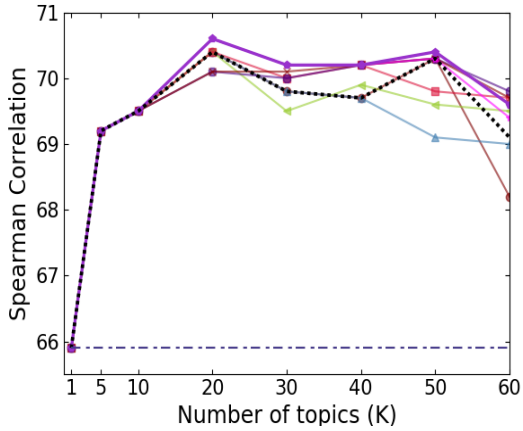
[1] Χρησιμοποιούμε αυτή τη μετρική απόστασης για να συγκρίνουμε τις διανυσματικές αναπαραστάσεις λέξεων σε όλη τη εργασία.



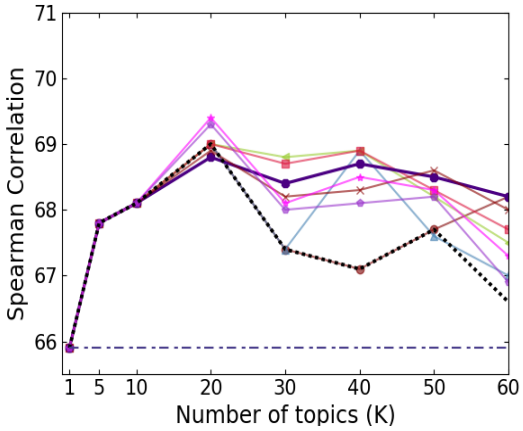
(α′) Complete linkage-AvgSimC



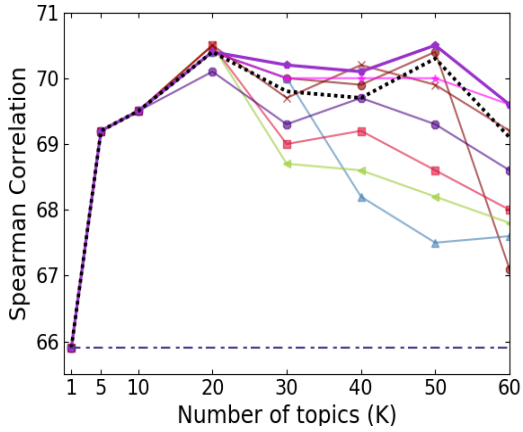
(β′) Complete linkage-MaxSimC



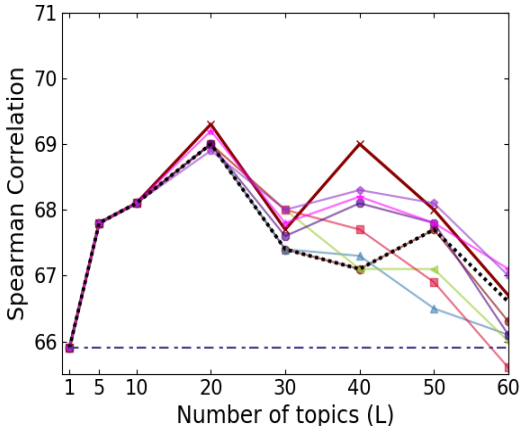
(γ′) Average linkage-AvgSimC



(δ′) Average linkage-MaxSimC



(ε′) Single linkage-AvgSimC



(ϛ′) Single linkage-MaxSimC

Σχήμα 6.7: Σύγκριση απόδοσης για διαφορετικές παραμέτρους εξομάλυνσης και κριτήρια σύνδεσης σε συνάρτηση με το πλήθος των θεμάτων. Ως μετρικές έχουν χρησιμοποιηθεί η συσχέτιση Spearman και οι AvgSimC και MaxSimC στο SCWS σύνολο δεδομένων. Ο αριθμός των μονοσήμαντων σημασιολογικών αγκυρών ισούται με 5.000 ενώ η διάσταση των σημασιολογικών χώρων ισούται με 300. Επιπλέον απεικονίζεται η απόδοση του baseline μοντέλου.

Στο Σχήμα 6.7 παρουσιάζουμε το πώς επηρεάζει η τεχνική εξομάλυνσης την απόδοση του μοντέλου χρησιμοποιώντας διαφορετικά κριτήρια σύνδεσης και διαφορετικό αριθμό κλάσεων ως συνάρτηση του αριθμού των θεμάτων. Παρατηρούμε πειραματικά ότι όταν χρησιμοποιούμε είτε το complete είτε το avergage κριτήριο σύνδεσης, τα αποτελέσματά μας δεν διαφέρουν σημαντικά. Ωστόσο, η χρήση του complete κριτηρίου σύνδεσης επιτυγχάνει ελαφρώς καλύτερες επιδόσεις. Αντιθέτως, παρατηρούμε ότι όταν χρησιμοποιούμε το single κριτήριο σύνδεσης, η απόδοσή μας παρουσιάζει μεγαλύτερες παραλλαγές ανάλογα με τον αριθμό των ομάδων N. Αυτή η παρατήρηση μπορεί να αποδοθεί στο γεγονός ότι η μικρότερη απόσταση μεταξύ των διανυσμάτων είναι η μοναδική τιμή που λαμβάνεται υπόψη. Ως αποτέλεσμα, πολλές από τις κλάσεις συνδέονται μεταξύ τους απλώς επειδή ένα από τα στοιχεία τους βρίσκεται σε άμεση γειτνίαση με ένα διάνυσμα που ανήκει σε κάποια απομακρυσμένη κλάση, επηρεάζοντας αρνητικά τη λύση του αλγορίθμου ομαδοποίησης.

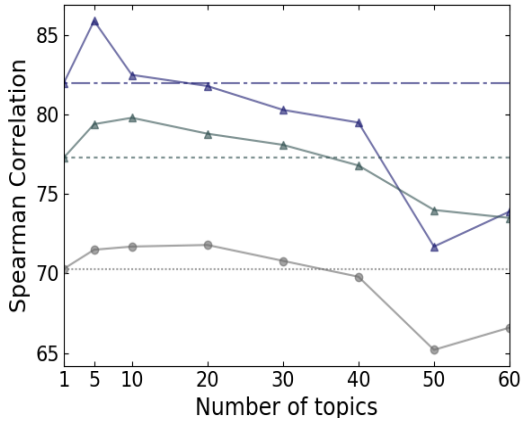
Επιπλέον, οι κορυφαίες επιδόσεις επιτυγχάνονται χρησιμοποιώντας μεγάλες τιμές της παραμέτρου p ως προς τις AvgSimC και MaxSimC μετρικές, όπως παρατηρείται στα δεξιά και αριστερά διαγράμματα του Σχήματος 6.7. Αυτό είναι ένα αναμενόμενο αποτέλεσμα, δεδομένου ότι οι μικρές τιμές του p οδηγούν στη δημιουργία μικρότερου αριθμού τελικών εξομαλυμένων διανυσμάτων για κάθε λέξη. Γενικά, η καλύτερη τιμή του p εξαρτάται τόσο από το κριτήριο σύνδεσης όσο και από τη βάση δεδομένων που χρησιμοποιούμε. Ωστόσο, οι διαφορές μεταξύ των επιδόσεων του μοντέλου μας για δύο διαδοχικές επιλογές της παραμέτρου p δεν είναι τόσο σημαντικές.

Τέλος, παρατηρούμε ότι χωρίς τη χρήση της τεχνικής εξομάλυνσης η απόδοση της προτεινόμενης προσέγγισης εξαρτάται σε μεγάλο βαθμό από τον αριθμό των θεμάτων K. Αποδίδουμε αυτή τη συμπεριφορά τόσο στις θορυβώδεις θεματικές αναπαραστάσεις που μπορεί να εισαχθούν στο μοντέλο μας καθώς ο αριθμός των θεμάτων αυξάνεται όσο και στα αραιά δεδομένα εκπαίδευσης (λιγότερα δεδομένα αποδίδονται σε κάθε ΘΚΣΜ όταν το K είναι μεγάλο). Ωστόσο, η επίδραση του θορύβου μειώνεται και η ανθεκτικότητα του μοντέλου μας ανακτάται όταν χρησιμοποιούμε την τεχνική εξομάλυνσης.

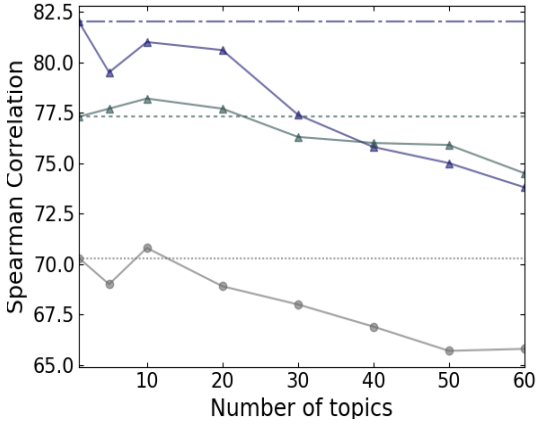
6.5 Αποτελέσματα σημασιολογικής ομοιότητας

Σε αυτή την ενότητα αξιολογούμε την απόδοση του δεύτερου μοντέλου μας σε μη συμφραζόμενα σύνολα δεδομένων, χρησιμοποιώντας τρεις μετρικές που υιοθετήθηκαν ιδιαίτερα στη βιβλιογραφία. Το Σχήμα 6.8 παρουσιάζει την απόδοση της προτεινόμενης προσέγγισής ως συνάρτηση του αριθμού θεμάτων K. Επίσης, οι αντίστοιχες αποδόσεις του *baseline* μοντέλου ($K = 1$) απεικονίζονται για κάθε σύνολο δεδομένων. Από τα αποτελέσματα που λαμβάνουμε για όλες τις μετρικές και για τα τρία σύνολα δεδομένων, παρατηρούμε μια σαφή σχέση μεταξύ του αριθμού των θεμάτων και της απόδοσης. Ο μικρότερος αριθμός θεμάτων επιτυγχάνει τις καλύτερες αποδόσεις, ενώ μεγαλύτερος είναι ο αριθμός των θεμάτων φαίνεται να οδηγεί στην υποβάθμιση της ποιότητας των προβλέψεών μας. Αποδίδουμε αυτή τη συμπεριφορά στην αυξημένη ειδικότητα των θεματικών αναπαραστάσεων που παρουσιάζονται στο σύστημα μας καθώς ο αριθμός των θεμάτων αυξάνεται (πιο σπάνια νοήματα λέξεων κωδικοποιούνται σε αυτά). Αναλυτικά, για κάθε μετρική παρατηρούμε ότι:

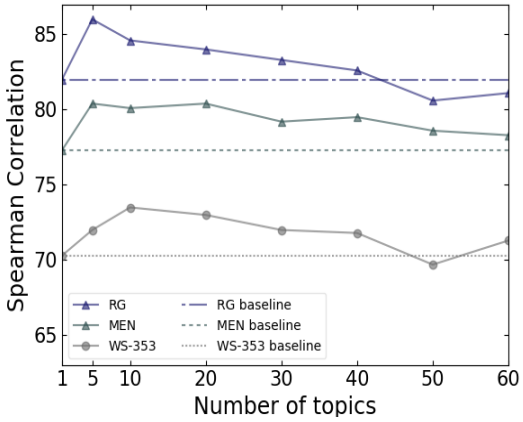
- MaxSim: Οι καλύτερες επιδόσεις επιτυγχάνονται χρησιμοποιώντας 5, 5 και 10 θέματα για τα σύνολα δεδομένων RG, WS-353 και MEN, αντίστοιχα. Η απόδοση του μοντέλου



(α′) MaxSim



(β′) AvgSim



(γ′) AvgSimW

Σχήμα 6.8: Η απόδοση του δεύτερου μοντέλου ως συνάρτηση του πλήθους των θεμάτων για μη συμφραζόμενα σύνολα δεδομένων, χρησιμοποιώντας τις μετρικές (α) MaxSim, (β) AvgSim και (γ) AvgSimW . Οι τρεις διακεκομμένες γραμμές αντιπροσωπεύουν τις αντίστοιχες αποδόσεις του baseline μοντέλου για κάθε σύνολο δεδομένων.

στα δύο τελευταία σύνολα δεδομένων παρουσιάζει σχεδόν την ίδια συμπεριφορά, καθώς δεν καταφέρνει να ξεπεράσει την αντίστοιχη απόδοση του baseline μοντέλου για $K > 30$, κάτι που δεν ισχύει για το RG.

- AvgSim: Το μοντέλο μας δεν καταφέρνει να ξεπεράσει την απόδοση του baseline μοντέλου για σχεδόν όλες τις επιλογές θεμάτων και στα τρία σύνολα δεδομένων. Αυτό θα μπορούσε να αποδοθεί στο γεγονός ότι τόσο οι θορυβώδεις όσο και οι καλά τοποθετημένες διανυσματικές αναπαραστάσεις μιας λέξης συμβάλλουν εξίσου στον σχηματισμό των σημασιολογικών ομοιοτήτων.

- AvgSimW: Και για τα τρία σύνολα δεδομένων το μοντέλο μας ξεπερνά τις αντίστοιχες επιδόσεις του baseline μοντέλου κατά μέσο όρο $1,5\%$ για κάθε $K > 1$. Οι καλύτερες επιδόσεις του μοντέλου μας επιτυγχάνονται με τη χρήση 5, 5 και 10 θεμάτων για τα

σύνολα δεδομένων RG, WS-353 και MEN αντίστοιχα.

Γενικά, το σύνολο δεδομένων MEN εμφανίζει την πιο ισχυρή απόδοση. Αυτό έχει μεγάλη σημασία δεδομένου ότι αποτελεί το πιο αξιόπιστο σύνολο δεδομένων μεταξύ των τριών συνόλων δεδομένων που χρησιμοποιούμε. Βασίζουμε αυτή την παραδοχή στα πειραματικά αποτελέσματα των Batchkarov et al. [2016] οι οποίοι σημειώνουν ότι το μικρό μέγεθος των RG και WS-353 καθιστά δύσκολη την αξιόπιστη διαφοροποίηση μεταξύ μοντέλων.

6.6 Σύγκριση με τη βιβλιογραφία

Στον Πίνακα 6.2 συγκρίνουμε τα δύο μοντέλα μας με state-of-the-art μοντέλα που χρησιμοποιούν πολλαπλές διανυσματικές αναπαραστάσεις για κάθε λέξη. Οι αναφερόμενες επιδόσεις αντιστοιχούν στις καλύτερες προβλέψεις για κάθε σύνολο δεδομένων. Συγκεκριμένα, οι επιδόσεις που αναφέρονται για το TDSMs σύστημα σε μη συμφραζόμενα σύνολα δεδομένων αντιστοιχούν στη μετρική MaxSim, ενώ η απόδοση του UTDSM στα ίδια σύνολα δεδομένων χρησιμοποιούν τη μετρική AvgSimW.[2] Παρατηρούμε ότι οι προσεγγίσεις που χρησιμοποιούν επίβλεψη (εξωτερικές βάσεις δεδομένων που παρέχουν κάποια πρότερη γνώση αναφορικά με τη σημασιολογία των λέξεων) επιτυγχάνουν υψηλότερες επιδόσεις τόσο σε συμφραζόμενα όσο και σε μη-συμφραζόμενα σύνολα δεδομένων.

Η πρώτη προσέγγισή μας (Μείγμα από ΘΚΣΜ) επιτυγχάνει την αρκετά υψηλή απόδοση $(68, 3)$ για τη μετρική MaxSimC, όσον αφορά το σύνολο δεδομένων SCWS. Ως προς τη μετρική AvgSimC, η προτεινόμενη προσέγγιση επιτυγχάνει απόδοση $70, 2$ που προσεγγίζει τα συστήματα με την υψηλότερη απόδοση $(70, 8$ και $71, 5$, αντίστοιχα). Το μοντέλο σύντηξης βελτιώνεται περαιτέρω την απόδοση του μοντέλου για τη μετρική AvgSimC $(70, 5)$. Όσον αφορά τα σύνολα δεδομένων για τα οποία δεν παρέχεται συμφραζόμενη πληροφορία, το μοντέλο γραμμικής παρεμβολής (TDSMs-LR) επιτυγχάνει κορυφαίες επιδόσεις $(83, 8)$ για το σύνολο δεδομένων MEN που υπερβαίνει τις επιδόσεις όλων των μοντέλων που προτείνονται στη βιβλιογραφία. Η ίδια προσέγγιση κατατάσσεται τέταρτη $(72, 7)$ σε σύγκριση με τα μοντέλα με τις καλύτερες επιδόσεις, όσον αφορά το WS-353.

Το δεύτερο μοντέλο (Ενιαίο ΘΚΣΜ) βελτιώνει τα αναφερόμενα αποτελέσματα της πρώτης μας προσέγγισης (όταν συγκρίνεται με τα σχήματα που δεν απαιτούν επίβλεψη), υποδεικνύοντας ότι η συμβολή των σημασιολογικών σχέσεων των λέξεων που διαμένουν σε διαφορετικούς θεματικούς τομείς διαδραματίζουν κάποιο ρόλο στην κρίση ομοιότητας. Επιπλέον, το μοντέλο μας ξεπερνά τις προηγούμενες προσεγγίσεις που δεν χρησιμοποιούν επίβλεψη σε όλα σχεδόν τα σύνολα δεδομένων και γενικά επιτυγχάνει αποτελέσματα συγκρίσιμα με τα καλύτερα συστήματα πρόβλεψης. Όσον αφορά το πρόβλημα της σημασιολογικής ομοιότητας λέξεων παρουσία συμφραζόμενων πλαισίων, το μοντέλο σημειώνε νέα state-of-the-art απόδοση ίση με 69.2, ως προς τη μετρική MaxSimC.

[2] Η απόδοση των μοντέλων για τη μετρική AvgSim δεν αναφέρονται στον Πίνακα 6.2.

	Approach	SCWS		MEN	WS-353	RG
		MaxSimC	AvgSimC			
Supervised	Chen et al. [2014]	-	68.9	-	-	-
	Iacobacci et al. [2015]	58.9	-	80.5	**77.9**	89.4
	Rothe and Schütze [2015]	-	69.8	-	-	-
	Pilehvar and Collier [2016]	-	**71.5**	78.6	-	**89.6**
Unsupervised	Huang et al. [2012]	-	65.7	-	71.3	-
	Neelakantan et al. [2014]	-	69.3	-	70.9	-
	Tian et al. [2014]	63.6	65.4	-	-	-
	Chen et al. [2015]	53.6	-	-	67.8	-
	Liu et al. [2015a]	67.9	69.5	-	-	-
	Liu et al. [2015b]	67.3	68.1	-	-	-
	Li and Jurafsky [2015]	-	69.7	-	-	-
	Amiri et al. [2016]	-	70.9		-	-
	Zheng et al. [2017]	-	69.9	-	-	-
	Nguyen et al. [2017]	66.9	66.7	76.4	72.4	-
	Lee and Chen [2017]	67.9	68.7	-	-	-
	Guo et al. [2018]	68.2	69.3	-	-	-
	Global-DSM	65.9	65.9	77.3	70.3	82.0
	TDSMs	67.8	70.2	80.0	72.2	-
	TDSMs-LR	-	-	**83.8**	72.2	-
	TDSMs-Fuse	67.4	70.5	-	-	-
	UTDSM	69.0	70.4	80.5	73.5	86.0
	UTDSM+smoothing	**69.2**	70.6	80.4	73.5	86.0

Πίνακας 6.2: Σύγκριση απόδοσης μεταξύ διαφορετικών state-of-the-art προσεγγίσεων σε διάφορα σύνολα δεδομένων για τον υπολογισμό της σημασιολογικής ομοιότητας, όσον αφορά τη συσχέτιση Spearman. Τα αποτελέσματα που παρουσιάζονται για τις δύο προτεινόμενες προσεγγίσεις της εργασίας αντιστοιχούν στις καλύτερες προβλέψεις μας, ενώ το Global-DSM αντιστοιχεί στο baseline σύστημα.

6.7 Οπτικοποιήσεις & Παραδείγματα

Σε αυτή την ενότητα πραγματοποιούμε μια σημασιολογική ανάλυση μεταξύ διαφορετικών θεματικών περιοχών (cross-domain analysis) για να ανιχνεύσουμε τις παραλλαγές της έννοιας μιας λέξης σε διαφορετικούς τομείς θεμάτων. Για το σκοπό αυτό, χρησιμοποιούμε μια λίστα γνωστών πολυσήμαντων λέξεων και μετράμε τη σημασιολογική ομοιότητα μεταξύ των διαφόρων θεματικών αναπαραστάσεων που αντιστοιχούν στην ίδια διφορούμενη λέξη. Απώτερος στόχος αυτής της ανάλυσης είναι να εξεταστεί αν το μοντέλο μας είναι ικανό να συλλάβει γνωστές θεματικές παραλλαγές στη σημασιολογία των πολυσήμαντων λέξεων.

Ο Πίνακας 6.3 περιλαμβάνει παραδείγματα της ανάλυσης μας. Οι πιο πιθανές λέξεις των θεμάτων (δεύτερη στήλη) δίνουν μια διαισθητική εικόνα των κύριων εννοιών που περιέχουν. Παρατηρούμε, για παράδειγμα, ότι η λέξη *python* μετατοπίζεται από το νόημα "φίδι" εντός ενός

Word	Topic Words	Meaning	Similarity
python	garden, plant, fish, bird, animal	snake	0.27
	software, forum, download, windows, web	programming language	
page	definition, dictionary, english, meaning, encyclopedia	sheet	0.65
	software, forum, download, windows, web	computing	
bank	loan, tax, cash, bank, insurance	financial institution	0.47
	boat, marine, ship, sailing, yacht	slope	
drug	health, medical, cancer, treatment, disease	medicine	0.61
	drug, health, marijuana, alcohol, effects	illegal substance	
power	news, nuclear, japan, energy, power	energy	0.50
	math, mathematics, theory, university, analysis	math operation	
apple	software, forum, download, windows, web	IT	0.30
	recipe, food, cooking, chicken, wine	fruit	
mouse	garden, plant, fish, bird, animal	rodent	0.48
	software, forum, download, windows, web	device	
window	forum, download, software, windows, web	computers	0.43
	car, parts, sale, auto, equipment	glass	
nursery	garden, plant, tree, flower, gardening	plants	0.46
	university, school, college, education, program	preschool	
history	university, school, college, education, program	course	0.68
	war, history, news, american, military	past	
act	law, court, legal, tax, state	law	0.39
	music, guitar, piano, dance, theatre	performance	
rock	mountain, river, park, road, trail	stone	0.43
	music, guitar, piano, dance, theatre	music	
house	mountain, river, park, road, trail	dwelling	0.57
	music, guitar, piano, dance, theatre	music	

Πίνακας 6.3: Παραδείγματα πολυσήμαντων λέξεων που υφίστανται αλλαγή στην έννοιά τους όταν συναντώνται σε δύο διαφορετικές θεματικές περιοχές. Στην πρώτη στήλη παρατίθενται οι λέξεις υπό εξέταση. Η δεύτερη στήλη περιλαμβάνει τις πιο πιθανές λέξεις των θεματικών τομέων στους οποίους συναντώνται αυτές οι λέξεις. Κάθε σειρά αντιστοιχεί σε διαφορετικό θεματικό τομέα. Η τρίτη στήλη συνάγει τη συγκεκριμένη σημασία της λέξης ενδιαφέροντος στον αντίστοιχο τομέα. Η τελευταία στήλη αντιστοιχεί στην ομοιότητα συνημιτόνου μεταξύ των δύο θεματικών αναπαραστάσεων της λέξης ενδιαφέροντος.

θέματος σχετικά με τα ζώα και τη φύση, στο να αναφέρεται σε μια "γλώσσα προγραμματισμού" σε ένα θέμα σχετικά με τους υπολογιστές. Η λέξη *drug* σχετίζεται κυρίως με την έννοια ενός "φαρμάκου" σε έναν ευρύ ιατρικό τομέα, υφίσταται όμως μια μικρή μετατόπιση από αυτό το νόημα όταν συναντάται σε ένα θέμα που αναφέρεται σε "παράνομες ουσίες". Η σημασία της εξαιρετικά πολυσήμαντης λέξης *act* μετατοπίζεται από το νόημα "θέσπισμα/νόμος" στο νόημα "παράσταση" σε ένα νομικό και σε ένα θέμα για την τέχνη αντίστοιχα. Σε ένα θεματικό πεδίο για τη μουσική, η λέξη *rock* αναφέρεται σε ένα "είδος μουσικής" ενώ σε ένα ευρύτερο θεματικό

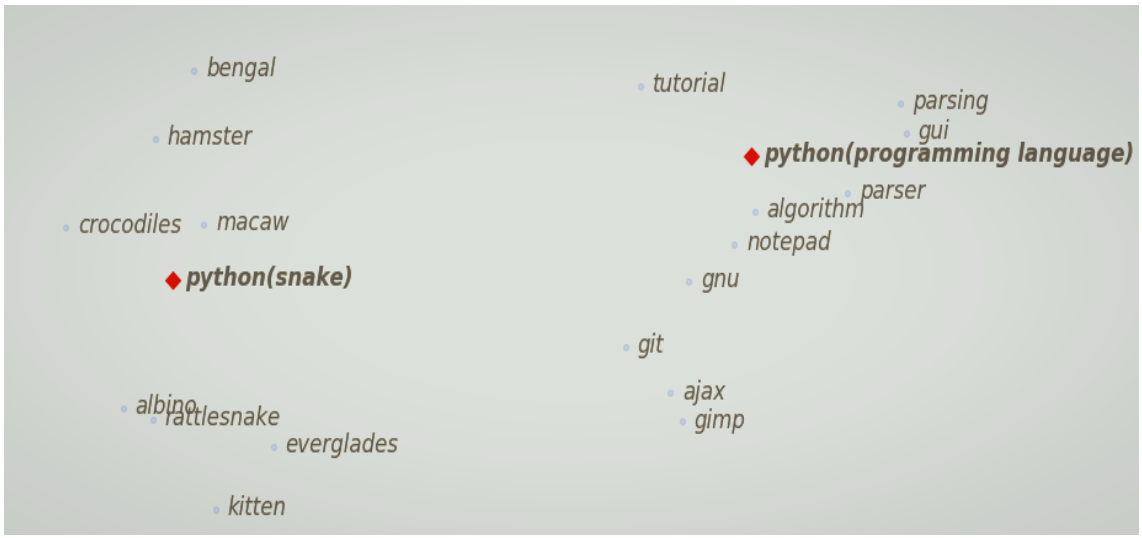
πλαίσιο που αναφέρεται στη φύση σημαίνει "βράχος". Τέλος, η λέξη *nursery* αντιστοιχεί σε μια "μονάδα παιδικής μέριμνας" σε ένα θέμα σχετικά με την εκπαίδευση, ενώ η σημασία της αλλάζει σε "φυτώριο" σε ένα θέμα σχετικά με τα φυτά.

Επιπλέον, στην Εικόνα 6.9 απεικονίζουμε τους κρυφούς σημασιολογικούς χώρους κάποιων μονοσήμαντων γειτονικών στοιχείων που αντιστοιχούν σε δύο διακριτές έννοιες των λέξεων *python*, *nursery*, *drug*, *page*, *apple* και *act* χρησιμοποιώντας την ανάλυση κύριων συνιστωσών (Principal Component Analysis, PCA). Συγκεκριμένα, για την απεικόνιση της σημασιολογικής αλλαγής μιας γνωστής πολυσήμαντης λέξης μεταξύ δύο διαφορετικών θεματικών περιοχών χρησιμοποιήσαμε την ακόλουθη διαδικασία, που βασίζεται στην ανάλυση κύριων συνιστωσών [Pedregosa et al., 2011] ως υπορουτίνα:

- Βρίσκουμε τις k κοντινότερες μονοσήμαντες λέξεις της υπό ανάλυση λέξης στους δύο θεματικούς χώρους που εξετάζουμε.
- Υπολογίζουμε τις PCA αναπαραστάσεις αυτών των λέξεων στον ενοποιημένο διανυσματικό χώρο.
- Οπτικοποιούμε τις δύο βασικές συνιστώσες των διανυσμάτων που ανήκουν στην ένωση των δύο μονοσήμαντων γειτονιών της λέξης ενδιαφέροντος μαζί με τις αντίστοιχες θεματικές αναπαραστάσεις της λέξης αυτής.

Σημειώστε ότι περιορίζουμε την ανάλυσή μας στην απεικόνιση των δύο μονοσήμαντων γειτονιών της λέξης ενδιαφέροντος, καθώς υποθέτουμε ότι οι μονοσήμαντες λέξεις διατηρούν σταθερές σχέσεις και στους δύο υπό εξέταση χώρους. Τονίζουμε ότι ο απώτερος σκοπός αυτής της διαδικασίας δεν η έρευση της απόλυτης σειράς των πλησιέστερων γειτόνων των δύο αναπαραστάσεων της λέξης ενδιαφέροντος (δεδομένου ότι πολυσήμαντες λέξεις θα μπορούσαν επίσης να συμπεριληφθούν στα γειτονικά σύνολα), αλλά στόχος μας είναι να ερευνήσουμε αν οι δύο θεματικές έννοιες της λέξης θα μπορούσαν να διαφοροποιηθούν με αξιοπιστία από το μοντέλο μας.

Τέλος, εξετάζοντας τις τοπικές μονοσήμαντες γειτονιές των λέξεων που υποβάλλονται σε ανάλυση, δείχνουμε ότι το μοντέλο μας παράγει όντως λογικά αποτελέσματα που αντικατοπτρίζουν την αναμενόμενη θεματική σημασία των λέξεων.


bengal
hamster
crocodiles
macaw
python(snake)
albino
rattlesnake
everglades
kitten
tutorial
parsing
gui
python(programming language)
parser
algorithm
notepad
gnu
git
ajax
gimp


evergreens
shrubs
succulents
perennials
camellias
ornamental
growers
grower
greenhouse
greenhouses
nursery(plants)
zoo
enrolment
website
district
kindergarten
infant
grammar
preschool
montessori
prep
nursery(preschool)
karate
gymnastics


amphetamines
narcotics
hallucinogens
insulin
therapies
drug(medicine)
mdma
heroin
vaccines
hormone
pharmaceuticals
vaccine
drug(illegal substance)
meth
cocaine
methamphetamine

Σχήμα 6.9: Παραδείγματα 2-διάστατων προβολών των κρυφών σημασιολογικών χώρων που κωδικοποιούνται στον ενοποιημένο διανυσματικό χώρο του μοντέλου μας, απεικονίζοντας τις μονοσήμαντες γειτονιές δύο θεματικών αναπαραστάσεων των λέξεων *python*, *nursery*, *drug*, *page*, *apple* και *act* που εξάγονται από διαφορετικούς θεματικούς τομείς.

Κεφάλαιο 7

Συμπέρασμα

7.1 Συμπεράσματα

Σε αυτή τη διπλωματική εργασία, προτείναμε αρχικά έναν συνδυασμό Θεματικά Κατανεμημένων Σημασιολογικών Μοντέλων για τον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ ζευγών λέξεων. Αποδείχθηκε οτι η προσέγγισή μας βελτίωσε την απόδοση του baseline μοντέλου (καθολικό ΚΣΜ). Η καλή απόδοση του μοντέλου αυτού θα μπορούσε να αποδοθεί στη δημιουργία θεματικών υποσυνόλων κειμένων όπου οι λέξεις ενδιαφέροντος εμφανίζονται με συναφείς θεματικές έννοιες. Στη συνέχεια, παρουσιάσαμε μια πιο ευέλικτη προσέγγιση που επεκτείνει την πρώτη μέθοδο, μέσω της απεικόνισης των Θεματικών ΚΣΜ σε έναν ενοποιημένο διανυσματικό χώρο. Το προκύπτον μοντέλο αποτελείται από πολλαπλές κατανεμημένες αναπαραστάσεις για κάθε λέξη, αντικατοπτρίζοντας τις θεματικές έννοιές τους.

Συνολικά, ως πρώτο βήμα, πραγματοποιήσαμε θεματικές προσαρμογές του σημασιολογικου χώρου μέσω της εκπαίδευσης Θεματικών ΚΣΜ που βασίζονται σε τεχνικές θεματικής μοντελοποίησης. Θεωρήσαμε ότι αυτές οι προσαρμογές οδηγούν στην απομόνωση των πολλαπλών θεματικών αισθήσεων που αντιστοιχούν σε πολυσήμαντες λέξεις, βασιζόμενοι στην παραδοχή ότι οι πολυσήμαντες λέξεις αλλάζουν τις έννοιές τους σύμφωνα με τις θεματικές περιοχές στις οποίες συναντώνται. Σε αυτό το σημείο, ένας συνδυασμός Θεματικών ΚΣΜ θα μπορούσε να χρησιμοποιηθεί για να συγκρίνει τις κοινές αισθήσεις δύο λέξεων κάτω από ένα συγκεκριμένο θέμα. Ωστόσο, η παραπάνω σύγκριση περιορίζεται σε επίπεδο θεματικών περιοχών. Για να ξεπεραστεί αυτός ο περιορισμός —αντί να χρησιμοποιήσουμε έναν συνδυσμό Θεματικών ΚΣΜ— προτείναμε ότι οι διανυσματικοί χώροι που ορίζουν τα Θεματικά ΚΣΜ πρέπει να ευθυγραμμιστούν ως προς έναν σταθερό/αναφορικό διανυσματικό χώρο, επιτρέποντας τη σύγκριση ανάμεσα στις θεματικές διανυσματικές αναπαραστάσεις λέξεων. Έπειτα, προτείναμε μια τεχνική εξομάλυνσης που μειώνει το θόρυβο που έχει εισαχθεί στο μοντέλο κατά τα διάφορα στάδια κατασκευής του και ανακτά την αξιοπιστία της απόδοσής του σε προβλήματα σημασιολογικής ομοιότητας λέξεων. Απ'όσο γνωρίζουμε, η δουλειά που περιγραφεται σε αυτή τη διπλωματική εργασία αποτελεί την πρώτη προσέγγιση κατά την οποία οι μέθοδοι απεικόνισης μεταξύ σημασιολογικών χώρων εφαρμόζονται στο πρόβλημα της εκμάθησης πολλαπλών αναπαραστάσεων πολυσήμαντων λέξεων.

Αναλυτικότερα, κύριος στόχος αυτής της εργασίας ήταν να διερευνήσει τις απεικονίσεις μεταξύ σημασιολογικών χώρων όταν χρησιμοποιούνται μονογλωσσικά δεδομένα. Ξεκινώντας από μια λίστα από *σημασιολογικές άγκυρες* που καθορίζουν τις παραπάνω απεικονίσεις, εξε-

τάσαμε τον ρόλο: (α) των αλγορίθμων απεικόνισης, (β) του πλήθους των σημασιολογικών αγκυρών, (γ) του βαθμού πολυσημίας των λέξεων οι αναπαραστάσεις των οποίων χρησιμοποιούνται ως σημασιολογικές άγκυρες. Συγκεκριμένα, το κύριο κίνητρό μας ήταν ότι οι πολυσήμαντες λέξεις αλλάζουν τις έννοιές τους σε διαφορετικούς θεματικούς τομείς και ως εκ τούτου οι σχετικές θέσεις τους σε διαφορετικούς θεματικούς χώρους θεμάτων παρουσιάζουν αντίστοιχες παραλλαγές. Ωστόσο, υποθέσαμε ότι κάτι τέτοιο δεν συμβαίνει όταν εξετάζονται μονοσήμαντες λέξεις. Συνεπώς, υποθέσαμε ότι οι κατανεμημένες αναπαραστάσεις μονοσήμαντων λέξεων αποτελούν *σημασιολογικές άγκυρες* που καθορίζουν τις αντιστοιχίσεις μεταξύ σημασιολογικών διανυσματικών χώρων, δεδομένου ότι διατηρούν σταθερές σημασιολογικές σχέσεις σε όλες τις θεματικές περιοχές. Η παραπάνω υπόθεση επικυρώθηκε μέσω των πειραματικών μας αποτελεσμάτων.

Επιπλέον, αξιολογήσαμε την απόδοση των μοντέλων μας σε προβλήματα σημασιολογικής ομοιότητας λέξεων οι οποίες εμφανίζονται τόσο παρουσία όσο και απουσία συμφραζόμενων πλαισίων. Επίσης έχουμε δείξει ότι μέσω της χρήσης γραμμικών ορθογώνιων απεικονίσεων το Ενιαίο Θεματικά ΚΣΜ πολλαπλών αναπαραστάσεων επιτυγχάνει σταθερά καλύτερη επίδοση όταν συγκρίνεται με το baseline σύστημα, υποδεικνύοντας την υπεροχή της χρήσης πολλαπλών αναπαραστάσεων ανά λέξη έναντι της χρήσεως μεμονωμένων αναπαραστάσεων λέξεων. Τέλος, εξετάζοντας τις τοπικές γειτονιές γνωστών πολυσήμαντων λέξεων που βρίσκονται σε διαφορετικούς θεματικούς τομείς, επιβεβαιώσαμε ότι το μοντέλο μας συλλαμβάνει λογικές θεματικές διαφοροποίησεις της ποικίλης σημασίας τους.

Εν κατακλείδι, οι απεικονίσεις μεταξύ σημασιολογικών χώρων φαίνεται να είναι πολύ ελπιδοφόρες στη σημασιολογική ανάλυση πολυσήμαντων λέξεων, καθώς αποτελούν μια ευέλικτη και κλιμακούμενη προσέγγιση που θα μπορούσε εύκολα να επεκταθεί σε ένα μοντέλο που λειτουργεί χωρίς καμία επίβλεψη.

7.2 Κατευθύνσεις για μελλοντική εργασία

Τα δύο μοντέλα που εξετάζονται σε αυτή την εργασία (Mixture of TDSMs, Unified multi-topic DSM) ακολουθούν προσεγγίσεις πολλαπλών σταδίων που περιέχουν διαφορετικές πρότυπες υλοποιήσεις και παραμέτρους που πρέπει να συντονιστούν για να βελτιώσουν την απόδοσή τους. Προτείνουμε τις ακόλουθες ιδέες για κάθε ένα από τα βασικά στάδια του μοντέλου μας ως κατευθύνσεις για μελλοντική εργασία:

Θεματική Μοντελοποίηση

- Ο αλγόριθμος Latent Dirichlet Allocation χρησιμοποιείται για τη δημιουργία Θεματικών ΚΣΜ. Ωστόσο, αυτό το θεματικό μοντέλο δεν είναι σε θέση να μοντελοποιήσει τις συσχετίσεις μεταξύ των θεμάτων, επειδή αποδίδει μια ενιαία κατανομή πάνω στα θέματα για κάθε έγγραφο. Αντί αυτού, θα μπορούσε να χρησιμοποιηθεί ένα ιεραρχικό θεματικό μοντέλο για να χαλαρώσει τον παραπάνω περιορισμό. Σύμφωνα με αυτό το σχήμα, ένα μείγμα ιεραρχικών ΘΚΣΜ θα μπορούσε να κατασκευαστεί σε μια προσπάθεια να συλλάβει τις ακριβείς σχέσεις ανάμεσα σε ζεύγη λέξεων καθορίζοντας μια διαδρομή εξάρτησης μεταξύ των διαφορετικών επιπέδων των θεματικών ΚΣΜ.

Σημασιολογικές απεικονίσεις

- Οι σημασιολογικές απεικονίσεις μεταξύ των χώρων προέλευσης και προορισμού χρησιμοποιούν μια λίστα μονοσήμαντων λέξεων που εξάγονται από τον κατάλογο εννοιών του WordNet. Η προτεινόμενη προσέγγιση θα μπορούσε εύκολα να μετατραπεί σε ένα μοντέλο χωρίς επίβλεψη, εάν οι μονοσήμαντες λέξεις αναγνωριστούν αυτόματα από το μοντέλο χωρίς τη χρήση του WordNet. Για να γίνει αυτό, θα μπορούσε να εξαχθεί ένας πίνακας αποστάσεων ανάμεσα σε ζεύγη λέξεων για κάθε θέμα. Στη συνέχεια, τα ζεύγη μονοσήμαντων λέξεων θα μπορούσαν να αναγνωριστούν ως εκείνα που έχουν μικρές αποκλίσεις απόστασης σε διαφορετικούς τομείς θεμάτων.

- Οι τεχνικές απεικόνισης που μελετώνται σε αυτή την εργασία υποθέτουν ότι υπάρχει ένας γραμμικός μετασχηματισμός μεταξύ των χώρων προέλευσης και στόχου. Ωστόσο, θα μπορούσαν επίσης να εξεταστούν μη γραμμικές απεικονίσεις μεταξύ των σημασιολογικών χώρων χρησιμοποιώντας αρχιτεκτονικές βαθιών νευρωνικών δικτύων.

Τεχνική εξομάλυσνης

- Όσον αφορά την τεχνική εξομάλυνσης, θα μπορούσαν να χρησιμοποιηθούν διάφοροι άλλοι αλγόριθμοι ταξιδόμησης για τη δημιουργία ομάδων θεματικών διανυσμάτων (π.χ. ομαδοποίηση k-μέσων). Επιπλέον, το πλήθος των κλάσεων θα μπορούσε να βελτιστοποιηθεί μέσω του αλγρίθμου ομαδοποίησης, προκειμένου να μειωθεί το πλήθος των παραμέτρων του μοντέλου μας. Αυτό θα μπορούσε να εφαρμοστεί μέσω της βελτιστοποίησης ενός κριτηρίου, όπως το εσωτερικό άθροισμα τετραγώνων των κλάσεων.

- Τέλος, θα μπορούσαν να χρησιμοποιηθούν πιο προηγμένες τεχνικές εξομάλυνσης των διαφορετικών θεματικών διανυσμάτων κάθε λέξης, με τη χρήση Γκαουσιανών διανυσμάτων. Η ιδέα αυτή βασίζεται στην αντίστοιχη δουλειά των Chen et al. [2015] οι οποίοι υποστηρίζουν ότι η αναπαράσταση των λέξεων ως σημείων σε ένα διανυσματικό χώρο δεν μπορεί να αντικατοπτρίσει τις σύνθετες σημασιολογικές σχέσεις μεταξύ των λέξεων.

Βιβλιογραφία

Hadi Amiri, Philip Resnik, Jordan Boyd-Graber, and Hal Daumé III. Learning text pair similarity with context-sensitive autoencoders. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1882–1892, 2016.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 238–247, 2014.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David J. Weir. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 7–12, 2016.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, pages 1137–1155, 2003.

David M. Blei. Introduction to Probabilistic Topic Modeling. *Communications of the ACM*, pages 77–84, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Resources (JAIR)*, 49:1–47, 2014.

Xinchi Chen, Xipeng Qiu, Jingxiang Jiang, and Xuanjing Huang. Gaussian mixture embeddings for multiple word prototypes. *arXiv preprint arXiv:1511.06246, 2015*, 2015.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, 2014.

Efstathia Christopoulou. Sentence-level sentiment analysis using topic modeling. 2016.

Fenia Christopoulou, Eleftheria Briakou, Elias Iosif, and Alexandros Potamianos. Mixture of topic-based distributional semantic and affective models. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 203–210, 2018.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, 16:76–83, 1989.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing. *Proceedings of the 25th International conference on Machine learning (ICML)*, pages 160–167, 2008.

Georgiana Dinu and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2014.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 426–471, 2014.

Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, pages 116–131, 2002.

J. R. Firth. A synopsis of linguistic theory. *Studies in Linguistic Analysis*, pages 1–32, 1957.

Yoav Goldberg. *Neural Network Methods in Natural Language Processing*, volume 37 of *Synthesis Lectures on Human Language Technologies*. 2017.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

Fenfei Guo, Mohit Iyyer, and Jordan Boyd-Graber. Inducing and embedding senses with scaled gumbel softmax. *arXiv preprint arXiv:1804.08077*, 2018.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proc. International Conference on Computational Linguistics (COLING)*, pages 497–507, 2014.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, 2016.

Zellig S. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954.

Matthew D. Hoffman, David M. Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, volume 1, pages 856–864, 2010.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882, 2012.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 95–105, 2015.

Elias Iosif and Alexandros Potamianos. Similarity computation using semantic networks created from web-harvested data. 21:49—79, 2015.

Stephen C. Johnson. *Hierarchical Clustering Schemes*. Psychometrika, 1967.

Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, pages 211–240, 1997.

Guang-He Lee and Yun-Nung Chen. Muse: Modularizing unsupervised sense embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMLP)*, pages 327–337, 2017.

Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? In *Proceedigns of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1732, 2015.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1284–1290. AAAI Press, 2015a.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Topical word embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, pages 2418–2424, 2015b.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Sophia Marmaridou. *Pragmatic Meaning and Cognition*. John Benjamins Publishing, 2000.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168*, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–751, 2013b.

George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, page 148, 1991.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, pages 235–244, 1990.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, 2014.

Dai Quoc Nguyen, Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. A mixture model for learning multi-sense word embeddings. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 121–127, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1680–1690, 2016.

Victor Prokhorov, Mohammad Taher Pilehvar, Dimitri Kartsaklis, Pietro Liò, and Nigel Collier. Learning rare word representations using semantic bridging. *CoRR*, abs/1707.07554, 2017.

Colorado Reed. Latent dirichlet allocation: Towards a deeper understanding, 2012.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

Joseph Reisinger and Raymond Mooney. Mixture Model with Sharing for Lexical Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1182, 2010.

Xin Rong. word2vec parameter learning explained. *arXiv:1411.2738*, 2014.

Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1793–1803, 2015.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, pages 627–633, 1965.

Peter H. Schönemann. *A generalized solution of the orthogonal procrustes problem.* 1966.

Amr Sharaf, Shi Feng, Khanh Nguyen, Kiante Brantley, and Hal Daumé III. The umd neural machine translation systems at wmt17 bandit learning task. In *Proceedings of the Second Conference on Machine Translation*, pages 667–673, 2017.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1556–1566, 2015.

Luchen Tan, Haotian Zhang, Charles L A Clarke, and Mark D Smucker. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 657–661, 2015.

Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 1353–1360, 2006.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings International Conference on Computational Linguistics (COLING)*, pages 151–160, 2014.

Zhaohui Wu and C Lee Giles. Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2188–2194, 2015.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 1006–1011, 2015.

Xiaoqing Zheng, Jiangtao Feng, Yi Chen, Haoyuan Peng, and Wenqing Zhang. Learning context-specific word/character embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3393–3399, 2017.

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics

Topic-based word embeddings

Diploma Thesis

of

ELEFTHERIA BRIAKOU

Supervisor: Alexandros Potamianos
Associate Professor

Computer Vision, Speech Communication and Signal Processing Laboratory
Athens, June 2018

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics

Topic-based word embeddings

Diploma Thesis

of

ELEFTHERIA BRIAKOU

Supervisor: Alexandros Potamianos
Associate Professor

Approved by the committee on June 15th 2018.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Alexandros Potamianos	Giorgos Stamou	Konstantinos Tzafestas
Associate Professor	Associate Professor	Associate Professor
N.T.U.A	N.T.U.A	N.T.U.A.

Athens, June 2018

.....

Eleftheria Briakou
Electrical and Computer Engineer, N.T.U.A.

Acknowledgments

This Diploma thesis signals the end of my studies at the National Technical University of Athens and it becomes a reality with the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

First and foremost, I would like to thank my supervisor Alexandros Potamianos for his support and encouragement during this thesis. At many stages in the course of my research I benefited from his advice, particularly so when exploring new ideas. I would also like to deeply thank him for introducing me to the fields of Machine Learning and Natural Language Processing, as his approach played a major role in my decision to continue my studies in this area.

I should thank Fenia for her support and contribution to a large part of this Diploma Thesis, as well as for the fruitful discussions and ideas we shared. Special thanks go also to Malvina —my stable co-worker and friend all these six years— as well as to all of my civil engineer friends!

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. None of my accomplishments would have been possible without them. Thank you.

This Diploma thesis is dedicated to my uncle Giorgos.

Eleftheria Briakou
June 25th, 2018

Abstract

Distributional Semantic Models (DSMs) constitutes a popular method that estimates the meaning of words from the statistical analysis of their contexts. The extracted word representations have been successfully applied to various Natural Language Processing (NLP) applications, and they are typically utilized to compute pairwise semantic similarities of words. However, one major deficiency of traditional DSMs is that the multiple senses of a polysemous word are conflated into a single vector space representation.

The goal of this Diploma Thesis is to alleviate the above problem, via proposing two models that leverage topic representations of words extracted from Topic-based DSMs (TDSMs). Firstly, motivated by the fact that typically words appear with a specific sense in each topic, we propose a semantic mixture model that enables the combination of word similarity scores estimated across multiple TDSMs. Afterwards, we extend this work in order to acquire a unified representation of the multiple topic-senses of words in a common space. In this direction, each of the TDSMs are aligned to a common vector space via linear mapping. This results in a set of embedding vectors per word with cardinality equal to the number of topics; the number of resulting vectors is further reduced via agglomerative clustering.

Furthermore, on of the main scopes of this thesis is to investigate different ways to perform the mappings from the topic sub-spaces to the unified semantic space. Specifically, we hypothesize that TDSMs capture meaningful variations in usage of polysemous words, while the relative semantic distance between monosemous words is preserved. This, motivated as to think of monosemous words as *semantic anchors* that determine the mappings between our semantic spaces. Up to our knowledge, this is the first time that mappings between semantic spaces are applied to the problem of learning multiple embeddings for polysemous words.

The proposed models can be evaluated on both contextual and in-isolation semantic similarity tasks, showing a significant improvement of correlation with human annotations, compared to a baseline approach that does not use topic models. Moreover, our models report performances comparable to the best predictive systems that are proposed in the literature.

Keywords

semantic analysis, topic modeling, LDA, distributional semantic models, Word2Vec, embeddings, topic embeddings, semantic mappings

Contents

List of Figures

List of Tables

Abbreviations

BOW	Bag-of-words
LDA	Latent Dirichlet Allocation
PMI	Point-wise Mutual Information
TF-IDF	Term frequency-Inverse Topic Frequency
POS	Part-of-Speech
LSE	Least Squares Estimation
CBOW	Continuous Bag-of-words
W2V	Word2Vec
CCA	Canonical Correlation Analysis
SVD	Singular Value Decomposition
VSM	Vector Space Model
DSM	Distributional Semantic Model
HDP	Hierarchical Dirichlet Process
KL	Kullback-Leibler
TDSM	Topic-based Distributional Semantic Model
UTDSM	Unified multi-Topic Distributional Semantic Model
PCA	Principal Component Analysis

Chapter 1

Introduction

1.1 Lexical semantics

By the term *semantics* we refer to the study of meaning in language. Its broad definition incorporates two different intellectual enterprises: one is *philosophical semantics* dignified and inscrutable; its goal is to formulate a general theory of meaning and the second is *lexical semantics*, grungy and laborious; its goal is to record meanings that have been lexicalized in particular languages [Miller and Charles, 1991]. The work presented in this Diploma Thesis clearly belongs to the second category, which also constitutes a well-known domain of Natural Language Processing (NLP).

From a computational perspective, lexical semantics aim to facilitate computers to detect aspects of meaning in language, as well as to encode this information in a formalistic way that enables their interpretability by computers. The importance of understanding the semantics of lexical units is paramount to language comprehension and acquisition, as they constitute the basic components of human language. To realize how the non-integration of this knowledge to computer systems could lead to undesirable results, let us examine a real example. Suppose that we use a question-answering machine and we forward to it the question "When was Linus Torvalds born?". The machine answers "Linus Torvalds was born in *", by just posing a Google query after converting the question into a statement using syntactic rules. This is an easy example that does not require any knowledge of lexical semantics to be answered correctly. Now suppose that we forward the question "What plants are native to Scotland?"; the machine outputs "A new chemical plant was opened in Scotland", which clearly indicates that the system is incapable of understanding the question. Why is it hard for the machine to infer a correct output in the previous example? The answer is straightforward: word semantics were not taken into account. Continuing with the latter example, another linguistic phenomenon is illustrated that plays a major role in language comprehension. Specifically, the polysemic nature of the word *plant* could drastically affect the reliability of the output answer as the word could be either reffered to a *small organism* or a *factory*, depending on the context it belongs to.

Cognitive experiments seem to indicate that the understanding of these semantics could be aided by the fundamental cognitive relationships between words [Marmaridou, 2000]. One straightforward implication of this observation is that a word's meaning highly

depends on the semantic relationships it shares with other words. For this reason the task of semantic similarity computation —which corresponds to the degree of likeness of meaning between two terms— is a popular domain in NLP that attempts to reveal the pragmatic semantics of words. Some of the applications of NLP that integrate the above information in their systems include: automatic translation of texts, information retrieval, automatically summarizing text, natural language generation, question answering, search engines and converting spoken speech into text.

1.2 Thesis Scope

Word-level representation learning algorithms adopt the *distributional hypothesis* [Harris, 1954], presuming a correlation between the distributional relationship and the semantic relationship of words. Typically, these models encode the contextual information of words into dense feature vectors —often referred to as *embeddings*— of a k-dimensional space, creating a vector space model (VSM) of meaning. These embeddings have been proved useful in various NLP applications, such as information retrieval [Manning et al., 2008], sentiment analysis [Tai et al., 2015], machine translation [Sharaf et al., 2017] and others.

Despite their popularity, traditional DSMs rely solely on models where each word is uniquely represented by one point in the vector space. From a linguistic perspective, these models could not accurately capture the meaning of polysemous words (e.g., *bank* or *cancer*), resulting in conflated word representations of their diverse contextual semantics. To alleviate this problem, some works incorporate multiple representations per word in their corresponding VSMs, based on clustering local contexts of individual words [Reisinger and Mooney, 2010, Tian et al., 2014, Neelakantan et al., 2014]. In parallel, psycholinguistic evidence seems to indicate that global context can also help language comprehension. Latent topic models are introduced as a natural way to represent global context of words in Liu et al. [2015b].

Following the same direction, we firstly propose a topic-based semantic mixture model that utilizes a combination of similarities extracted from Topic-based Distributional Semantic Models (TDSMs). This work is also described in the published conference paper of Christopoulou et al. [2018]. The main motivation behind the use of topic modeling, for the task of word semantic similarity computation, is to adapt the similarity estimates provided from various topics. This is similar to using semantic mixture models to encode multiple senses in words. Afterwards, we recognize that one of the major deficiencies of the above approach is that it fails to capture the relationships between words that do not reside in the same thematic domain, as the comparison of a pair of words is restricted to a topic-level. To overcome this problem, we additionally propose a more flexible framework that utilizes TDSMs, in order to create a unified model of multiple topic-based distributional semantic representations. To that end, TDSMs should be aligned to a reference coordinate system enabling the comparison of topic embeddings extracted from different topic semantic spaces. Based on existing mapping methods, we analyze the role of monosemous words in defining robust transformations between TDSMs. To our knowledge, this is the first work to apply mapping embedding techniques between semantic spaces of the same language in order to account for polysemy in word semantics.

1.3 Thesis outline

The thesis is organized as follows:

- Chapter 2 gives a description of the bibliography which is divided in two main sections, according to the fields that the thesis addresses. The first section gives a brief review of Distributional Semantic Models (DSMs). Specifically, the basic categories of traditional (single-prototype) and multi- prototype DSMs are studied. In the second section we discuss basic transformation methods between semantic spaces along with their applications.
- Chapter 3 presents the Latent Dirichlet Allocation (LDA) and Word2Vec algorithms that constitute the two main core systems that our two proposed approaches utilize.
- Chapter 4 describes the main motivation and the system architecture of our first approach; that is the Mixture of Topic-based Distributional Semantic Models (TDSMs), along with the mixture schemes we experiment with.
- Chapter 5 presents the main motivation and the system architecture of our second approach (extension of TDSMs); that is the Unified multi-Topic Distributional Semantic Model (UTDSM), along with the semantic metrics that we utilize for evaluation purposes.
- Chapter 6 displays the experimental procedure and the results for the different experiments that were conducted, on different datasets.
- Chapter 7 concludes the thesis and proposes directions for future work.

Chapter 2

Distributional Semantics

2.1 Distributional Hypothesis

Distributional Semantics embraces a wide range of approaches based on the distributional hypothesis, in an attempt to capture meanings of linguistic entities (words, phrases) from their usage in language. This hypothesis is often described by the famous quote "You shall know a word by the company it keeps" [Firth, 1957], which presumes a correlation between distributional similarity and meaning similarity. The direct implication of this hypothesis is that two words that are considered to be semantically similar are expected to occur in similar *contexts*, and vice-versa. The conceptualization of this hypothesis, requires a definition of what constitutes a *context* of a target word defined in a mathematical framework. In this work, we follow the commonly used definition of a context as the set of words existing within a window around each occurrence of the target word.

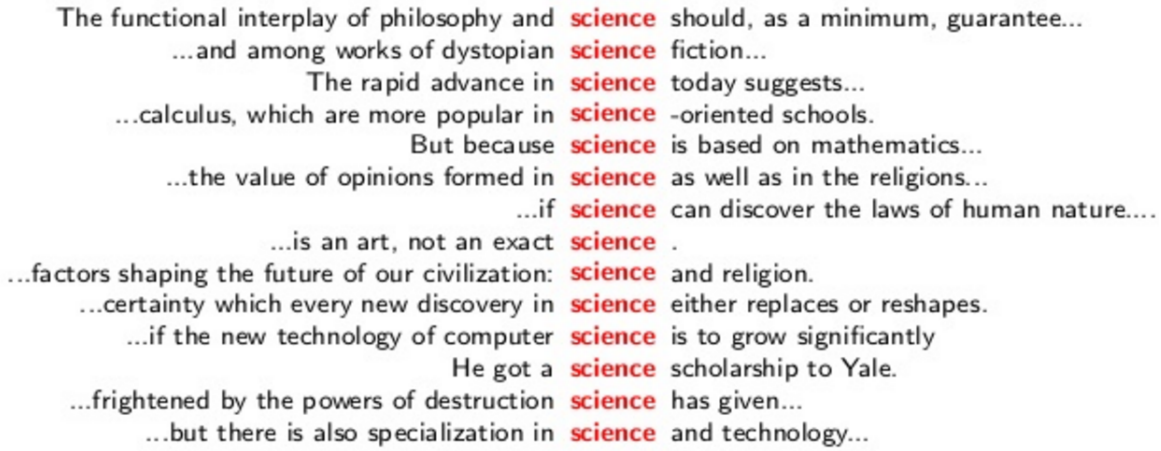


Figure 2.1: Example of the distributional hypothesis of meaning for the word *science*.

2.2 Single-prototype representations

The models that follow the distributional hypothesis are often referred to as Distributional Semantic Models (DSMs). Typically, these models encode the contextual information of words into feature-vectors placed in a k-dimensional space, thus creating a vector space model (VSM) of meaning. From a geometrical perspective, each word is represented as a single point in the VSM, while semantically similar words are placed closer in the semantic space and dissimilar words are being put far apart from each other. Baroni et al. [2014] noted that DSMs of single-prototype representations of meaning can be divided into two broad categories: the count-based and the predictive models.

2.2.1 Count-based Models

In the simplest case of traditional DSMs, each dimension captures statistical information for context items observed to co-occur no further than a fixed distance c from the target's instance. This simple counting method results in a co-occurrence matrix, where the components of each vector can be interpreted as weights denoting the strength of the relationship between the target and the respective context word. It can be observed though, that raw co-occurrences are not a reliable source of information for revealing meaning correlation, as frequent yet uninformative context words (such as the word "a" in the example of Figure 2.1) tend to co-occur with most of the target words at a high rate.

In order to mitigate this phenomenon, non-linear operations can be applied on the co-occurrence matrix in an attempt to downplay the role of highly frequent words. Typically, the most widely used transformation is the Positive Pointwise Mutual Information (PPMI) defined by Church and Hanks [1989] as:

$$\text{PPMI}(word_i, word_j) = \max(0, \text{PMI}(word_i, word_j)) \tag{2.1}$$

$$\text{PMI}(word_i, word_j) = \log_2 \frac{P(word_i) \cap P(word_j)}{P(word_i)P(word_j)}. \tag{2.2}$$

In the above relation the numerator gives us information about how often the two words occur together, while the denominator tells us how often we would expect the two words to co-occur assuming they occurred independently, so their probabilities could just be multiplied (see example in Table 2.1).

Since count-based methods calculate the co-occurrence matrix for all words, they result in sparse high-dimensional representations —that is, most of the components of the vectors are zero— as a word is often semantically related to a small percentage of context instances. Commonly, dimensionality reduction is applied to the large matrix (in this case, the PPMI-weigthed co-occurrence matrix) in order to lessen the noise and reduce the sparsity of the vector space. The basic idea is to generate a lower-rank approximation of the original matrix, while in parallel retain the relations between the vectors. The resulted lower dimensional space is represented by the most important dimensions of the data set, along which most variation happens. The most popular method to generate matrix approximations of any given rank k is Singular Value Decomposition (SVD) [Landauer and

	player	field	court	Athenian	cart	a
baseball	546	350	5	1	35	975
basketball	485	10	410	1	45	1053
democracy	1	5	2	350	10	375
monarchy	2	1	4	7	276	330

↓

	player	field	court	Athenian	cart	a
baseball	0.38	0.97	0	0	0	0
basketball	0.21	0	0	0	0	0.01
democracy	0	0	0	1.93	0	0
monarchy	0	0	0	0	1.86	0.03

Table 2.1: Example of co-occurence matrix, extracted using raw counts (upper table), and after PPMI transformation (lower table)

Dumais, 1997], based on extracting the singular values of the initial matrix. An abstract scheme that summarizes the steps of creating a count-based DSM in depicted in 2.2

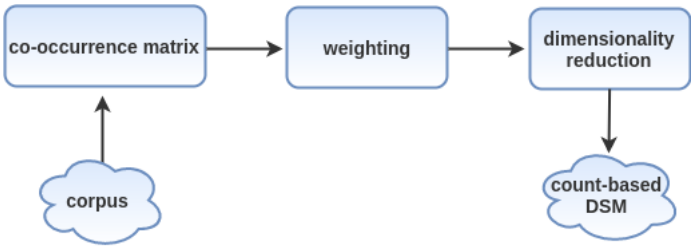


Figure 2.2: Abstract representation of count-based Distributional Semantic Models construction.

2.2.2 Predictive Models

Predictive models belong to the new generation of DSMs that frame the vector estimation problem directly as a unsupervised task. Bengio et al. [2003] were the first to introduce artificial neural networks in the field of word semantic representations, and few years later Collobert and Weston [2008] were the first to establish them as a highly efficient tool in NLP tasks. Although both count and predictive models base their theoretical background on the *distributional hypothesis*, they differ in the computational approach they follow in order to learn geometrical encodings of words. Instead of collecting contextual statistics of words to reveal semantic meaning, the new generation of DSMs turns the problem to a machine learning task and attempts to build programs which learn to make correct decisions on training data and improve with experience.

The general structure of predictive models is based on a probabilistic feed-forward neural network that takes words as input, and embeds them as vectors into a lower dimensional space, which it then fine-tunes through back-propagation. For this reason, the vector representations that are extracted as the weights of the first neural layer of the

model are usually referred to as *embeddings* in the literature. A systematic comparison between *embeddings* and vector representations obtained through count-based models has been excessively studied in Baroni et al. [2014] in terms of downstream tasks. The superiority of dense representations has been also attributed to computational reasons in Goldberg [2017], as the majority of neural network toolkits do not play well with very high-dimensional, sparse vectors.

Word2Vec released by Mikolov et al. [2013b] constitutes the most prominent neural language model that builds distributional representations of meaning. Its underlying idea lies in storing into each word vector the information that allows it to predict its most relevant contexts or vice-versa; predict the contexts from the target word. We leave its analytical explanation for Chapter 4.

2.3 Multiple-prototype representations

A great percentage of recent work on distributional semantics relies solely on models where each word is uniquely represented by one point in the semantic space. From a linguistic perspective, these models could not accurately capture the meaning of a polysemous word, resulting in a conflated representation of its diverse contexts. The problematic nature of single-prototype models could be better understood in the following two examples, which present two different contextual occurrences of the word *python*.

- "...students find coding in ***python*** a satisfying experience..."
- "...***python*** uses its sharp, backward-curving teeth..."

Here, the inferred meanings of the word python are totally different in the two contexts (programming language, snake). In order to make these distinctions possible in NLP, we need to account for polysemy in our models and turn the single-prototype representations to multiple-prototype representations. In the following sections we group the methods that assign multiple representations per word into two broad categories: *unsupervised* models induce multiple representations without leveraging semantic lexical resources, while *supervised* models constitute knowledge-based approaches.

2.3.1 Unsupervised Models

Fixed number of prototypes per word

Reisinger and Mooney [2010] were the first to introduce multiple-prototype representations for lexical semantics. Motivated by the *distributional hypothesis*, they collected local contexts for each target word (as a vector formed by collecting frequency statistics in a fixed window around it) and applied clustering on them, with the number of clusters being the single parameter of the model. The centroids of the created clusters were used in order to create a set of "sense-specific" vectors for each target word (Figure 2.3).

Following the clustering approach, Huang et al. [2012] proposed a recurrent neural network that incorporated both global and local context to learn multiple dense, low-dimensional embeddings. Again the number of possible senses corresponding to each word coincided with the fixed number of clusters.

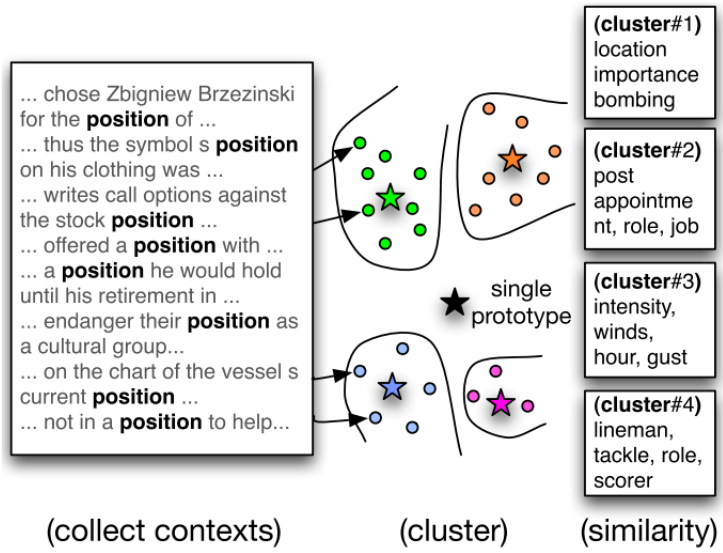


Figure 2.3: Overview of the multi-prototype approach using contextual clustering.

A probabilistic framework was later introduced by Tian et al. [2014] who extended the Word2Vec model via representing the probability of a context word given the target word as a finite mixture of the prototypes of the target word. Using this framework, they designed an Expectation-Maximization algorithm to learn multiple embeddings, where the number of senses attributed to each word constituted a predetermined design decision.

Despite the fact that models with a fixed number of prototypes per word established the first attempts to provide vector representations that integrated the polysemic nature of words, more recent approaches provide more flexible solutions to the problem. Their flexibility is attributed to the fact that the real number of senses for words differs according to their polysemy degree (note that some words have only one sense, a.k.a. monosemous words) and changes through time as the evolution of language causes the creation of new senses (e.g., word *python* as a programming language).

Adaptive number of prototypes per word

More recent approaches mostly relied on neural network architectures that encode multi-sense information. Neelakantan et al. [2014] motivated by the clustering approaches of previous models, followed an online method of learning skip-gram sense embeddings during which they also estimated the number of clusters. Contrary to previous work, both context and word vectors were learned simultaneously, instead of learning context representations as part of a pre-processing step. Later, a dynamic Gaussian skip-gram mixture model was introduced by Chen et al. [2015] enabling the detection of different number of senses for each word during training. In that work, each word was represented as a Gaussian mixture instead of a point vector in the embedding space, where each Gaussian component represented a sense of the word. In a more recent work, Amiri et al. [2016] made use of autoencoders to map each word to a context-specific representation, while Lee and Chen [2017], Guo et al. [2014] implemented discrete sense selection through reinforcement learning.

All of the above methods utilized the context information of each word occurrence without taking into account the relative order of words in the context window. Zheng et al. [2017] suggested that this omission impairs the quality of multi-prototype representations derived by clustering-based methods, and noted that the order of context words matters to the meaning of the target word. To tackle this issue, they used a neural network model, called CSV (Context-Specific Vector), which can generate both word and context representations. Their proposed neural network architecture contained a convolutional layer that was designed to produce context representations reflecting the order of their constituents. After the refined generated context representations were extracted, they were used to learn context-specific multi-prototype word embeddings.

Another definition of context was also described in Liu et al. [2015b], who treated context as a topic domain. Motivated by the observation that polysemous words usually change their meaning when they reside in different topic domains, they were the first to utilize topic modeling to learn multiple-prototype representations. Specifically, the Latent Dirichlet Allocation (LDA) algorithm was employed into the skip-gram model to get the distribution of a word over topics, which was further utilized to extract topic-word embeddings. In a more recent work, LDA was utilized to infer the weights of each topic. The weights were further used to define a mixture vector representation for each target word that predicted its corresponding context words [Nguyen et al., 2017]. Moreover, Wu and Giles [2015] exploited Wikipedia articles and assumed that words co-occurring in articles under the same subject share the same sense. The sense-aware prototypes were produced via clustering the Wikipedia pages based on the global and local contextual information of the target word.

A probabilistic approach was followed by Li and Jurafsky [2015], who proposed that a word should be associated with a new sense when there is evidence in its context suggesting that it sufficiently differs from its early senses. They also noted that such a theoretical scheme naturally points to Chinese Restaurant Process. According to this probabilistic framework, each word occurrence corresponds to a costumer while each table corresponds to a sense of a word. In these terms, a new word occurrence could either sit in an occupied table (assigned to an existing word sense), or choose an unoccupied table to sit (assigned to a new word sense).

Guo et al. [2014] proposed a different theoretical framework to induce multiple sense-specific embeddings for each ambiguous word, using a recurrent neural network. Instead of using the contextual information of words as evidence of their possible meanings, they utilized bilingual resources motivated by the fact that a word with multiple senses could have a different translation in another language.

2.3.2 Supervised Models

The unsupervised methods reviewed so far attempt to conceptualize the polysemic nature of words via creating multiple-prototype representations from raw contextual information extracted from massive text corpora. More recent techniques that achieve state-of-the-art performance in contextual semantic similarity tasks, involve knowledge-based approaches. In general, these knowledge-based approaches utilize an incomplete knowledge base along with a large corpus of text and try to use the first as a prior knowledge

to the problem. The most widely known sense inventory used as an auxiliary knowledge for multiple-prototype representations extraction is WordNet. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called *synsets*, each expressing a distinct concept as described in Miller et al. [1990].

Chen et al. [2014] used the definitions provided for each word by WordNet, in order to assign vector representation to senses. Using these sense vectors as initial estimations along with single-prototype word vectors, they refined them through word sense disambiguation algorithms. Given the disambiguated words, they finally modified the skip-gram model in order to jointly train words and sense vectors. Later, Iacobacci et al. [2015] used BabelNet as their underlying sense inventory, which constitutes an enriched database of WordNet. By leveraging the knowledge of the inventory they automatically generated sense-annotated corpora, using a word sense disambiguation algorithm. Sense-agnostic representations were extracted via employing the skip-gram model over the annotated corpus.

Another knowledge based approach introduced by Rothe and Schütze [2015] thought of words as sums of their lexemes (units), and synsets as sums of their lexemes. The interpretation of this theoretical foundation naturally establishes algebraic operations between word vectors in a mathematical algorithm. More specifically, pre-trained word embeddings were extended to embeddings of lexemes and synsets, with the help of WordNet. Recently, Pilehvar and Collier [2016] de-conflated pre-trained word representations based on the deep knowledge derived from WordNet. After linking these pre-trained representations to WordNet, they extracted a list of semantically biased words towards the ambiguous word. Given the biased words and the target word's lemma representations, they extracted a sense-aware representation for the target word via searching for the vector with the minimum distance from it.

2.4 Transformations in semantic spaces

As mentioned previously, neural network models —such as Word2Vec— have become very popular recently, as it has been proved that they significantly and continuously outperform the traditional count-based models. Many scientists attributed this superiority to the natural edge of neural networks over methods that solely relied on word co-occurrence counts. One of the main characteristics of predictive Distributional Semantic Models is that they create semantic spaces which are not aligned to a fixed coordinate system, due to their non-deterministic nature. Basically, this means that if the algorithm runs under the same dataset twice, it can be noted that the resulted semantic spaces have drastically different global structures. For this reason, the problem of defining transformations between semantic spaces has attracted a lot of attention recently, as it enables the comparison of the distributed representations that belong to different datasets.

The most popular application of semantic spaces transformation is machine translation, where the ultimate goal is to automate the process of generating large dictionaries starting from few bilingual data. Mikolov et al. [2013a] were the first to introduce such mappings in order to predict translations of words between English and Spanish. After learning word representations for the two languages using the Word2Vec model, they proposed a

linear mapping between the two spaces representing the language semantic spaces. In the aligned coordinate system the correct translation of a target word is expected to lie near the target word.

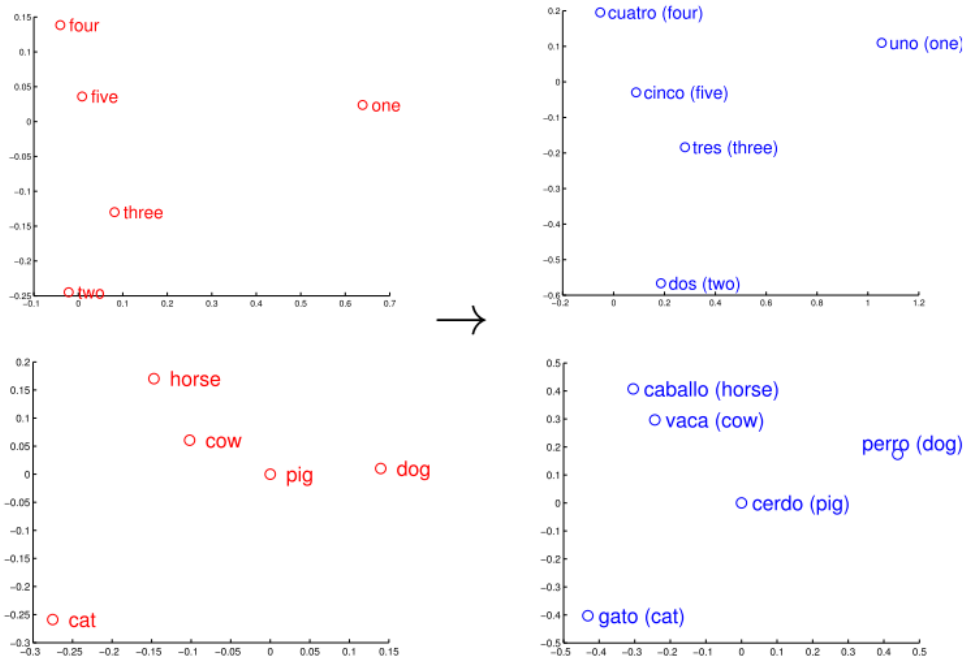


Figure 2.4: Projections of distributed word vector representations of numbers and animals in English (left) and Spanish (right) using PCA as presented in Mikolov et al. [2013a].

As they noted, their core motivation was that all common languages have similar geometric arrangements, as they share concepts that are grounded in the real world. Later work on machine translation focused on the properties of the transformation matrices between languages [Xing et al., 2015], as well as on the properties of the embeddings being mapped to the shared space. Specifically, Dinu and Baroni [2014] showed that the neighborhoods of the mapped embeddings are highly polluted by *hubs*, which are defined as vectors that tend to be popular nearest neighbors of many items.

Another application of semantic spaces transformation was later studied by Tan et al. [2015], who attempted to explore the semantic differences of words between the informal English of social media (Twitter corpus), and the formal English of well organized texts (Wikipedia corpus). In order to align the two semantic spaces, they assumed that a mapping existed between the most frequent words of the two corpora. After mapping the two languages to a common space, they employed a normalization of word distances based on term-frequency, and finally used these distances to find words whose usage is discriminative between the two corpora.

The semantic evolution of words' meaning can be captured in large-scale corpora that refer to different periods of time. Hamilton et al. [2016] created diachronic embeddings, by first constructing embeddings in each time-period and then learn consecutive linear rotational matrices that mapped the vector spaces of historic corpora that corresponded to different time intervals, in order to track the semantic drifts of words within-years. The

relative high dimensionality of diachronical embeddings (20, 30, 200 ...) poses a challenge, as they are typically not embedded in 2 or 3 dimensions that can be easily interpreted by humans. For this reason, dimensionality reduction techniques usually take place in order to visualize the trajectory a word follows over time in a 2 dimensional space.

Figure 2.5 illustrates an example of words' trajectories that reveal semantic evolution of words through time. By comparing the relative position of the words with their temporal nearest neighbor we could track interesting semantic shifts in their meaning that could also reflect cultural evolution. For instance, the word gay shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. In the early 20th century broadcast referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". Awful underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" as reported in Hamilton et al. [2016].

Figure 2.5: Two-dimensional visualization of semantic change of three English words.

Recently Prokhorov et al. [2017] applied semantic space transformations in an attempt to enrich the coverage of an existing vocabulary with rare or unseen words. The interesting property of their approach is that distributional information derived from text corpora, could be used in order to complete the missing parts of knowledge bases and vice-versa. To do so, they created a mapping between a distributional semantic space and a lexical ontology using *semantic bridges* of monosemous words.

Mapping Methods

We start by defining basic terminology in order to explain the most popular methods of alignment between semantic spaces that can be found in the literature. Let X and Y be the word embedding matrices of the source and target language, respectively. The i-th column of matrix X is the distributed vector representation $x_i \in \mathbb{R}^d$ of word i, while $y_i \in \mathbb{R}^d$ is its equivalent distributed vector representation in the target language. We aim to find a transformation matrix $W \in \mathbb{R}^{d \times d}$ that maps the source language to the target language, such that WX is as close as possible to Y. As summarized in Artetxe et al. [2018], this transformation matrix could be computed through linear, orthogonal of canonical methods.

- **Linear** methods in this area were introduced by Mikolov et al. [2013a] who used a

linear mapping as the first attempt to align semantic spaces for machine translation. They used a least squares objective function that minimizes the sum of squared Euclidean distances between the translated pairs vectors of two languages, without imposing a restriction to the matrix. This problem (also known as Ordinary Least Square) has a closed-form solution as indicated in Equation 2.3.

$$W = \arg\min_{W} \|WX - Y\|_F = (X^t X)^{-1} X^t Y. \tag{2.3}$$

Few years later, Dinu and Baroni [2014], incorporated an L2-regularization term to the objective function.

- **Orthogonal** methods were firstly proposed by Xing et al. [2015] who noticed that both the source and the target vectors should remain normalized to unit length during the learning phase of the mapping algorithm. They also noted that normalization is a crucial characteristic that the aligned representations should hold, as it ensures that the dot product between two vectors falls back to their cosine similarity, the most widely used distance measurement between word embeddings. For this reason, they mapped the source space to the target via solving the constraint optimization problem of Equation 2.4.

$$W = \arg\min_{W} \|WX - Y\|_F, \quad \text{subject to} \quad WW^T = \mathbb{I}. \tag{2.4}$$

  From a mathematical perspective, the above problem is known as the orthogonal Procrustes problem and it has a closed form solution. The optimal W is recovered by UV^T, where U and V, are obtained through the Singular Value Decomposition (equal to $(U\Sigma V^T)$) of YX^T. For a more detailed review of the problem we refer the reader to Schönemann [1966].

- **Canonical** methods on the other side, compute two distinct linear mappings M_1 and M_2 first, where the objective is to maximize the correlation between the dimensions of the projected matrices M_1X ans M_2Y. After computing the two mappings, the transformation matrix W is recovered through a simple algebraic operation as noted in Equation 2.4. Faruqui and Dyer [2014] were the first to use Canonical Correlation Analysis in order to map two semantic spaces, which was later proved to give similar results to the orthogonal mapping.

$$W = M_1^{-1} M_2, \quad \text{where} \quad M_1, M_2 = \underset{M_1, M_2}{\text{argmax}} \ \text{cov}(M_1 X, M_2 Y). \tag{2.5}$$

Chapter 3

Background

In the following subsections we are going to explain the Word2Vec and Latent Dirichlet Allocation (LDA) algorithms, the two main core models used in this thesis.

3.1 Word2Vec

The Word2Vec model, introduced by Mikolov et al. [2013b], constitutes one of the most widely used algorithms to create high quality vector representations of words, which are typically known as *embeddings*. From a generic point of view, Word2Vec takes as input a bunch of text, computes some statistics on that text and embeds each word in a vector space in such a way that important semantic relationships between words can be revealed through simple mathematical operations on their vectors. Despite the fact that the dense representations that this model produces are often used as the main core of applying Deep learning in Natural Language Processing, Word2Vec is itself a shallow neural network consisting of one hidden layer. This simple structure makes Word2Vec a very computationally-efficient model.

At this point, we are going to define an example sentence that will serve as our input corpus in order to describe the two algorithms that can be used to create word embeddings: the Continuous Bag of Words and the Skip-gram model. Let our input corpus consists of the sentence D:

$$D = \{\textit{former president Obama speaks to the media in Washington about terrorism}\}$$

In order to train our neural network we should firstly extract a vocabulary consisting of words we are interested in learning representations for. In this case the vocabulary is defined as follows:

$$V = \{\textit{former, president, Obama, speaks, to, the, media, in, Washington, about, terrorism}\}$$

Based on that vocabulary we create naive mathematical word representations, in order to feed the input layer of our neural network. For this reason, we construct a one-hot vector representation for each word, which corresponds to a sparse vector with size equal to the cardinality of the vocabulary ($|V|$), and zero elements in all indices except for a 1 value in a unique index reflecting its relative position in the vocabulary set. In our example the

one-hot vectors are:

$[1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = former$
$[0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = president$
$[0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = Obama$
$[0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = speaks$
$[0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0] = to$
$[0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0] = the$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0] = media$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0] = in$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0] = Washington$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0] = about$
$[0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1] = terrorism$

3.1.1 Continuous Bag of Words Model

The Continuous Bag of Words (CBOW) model aims to predict the target word by its nearby words that are included in a fixed window around it. The motivation behind this algorithm stems from the *distributional hypothesis*, as it tries to find evidence in the context a word lives to learn its dense representation. We are going to follow the terminology and notation introduced by Rong [2014] in its detailed review of Word2Vec parameters learning. The structure of the CBOW model is presented in Figure 3.1.

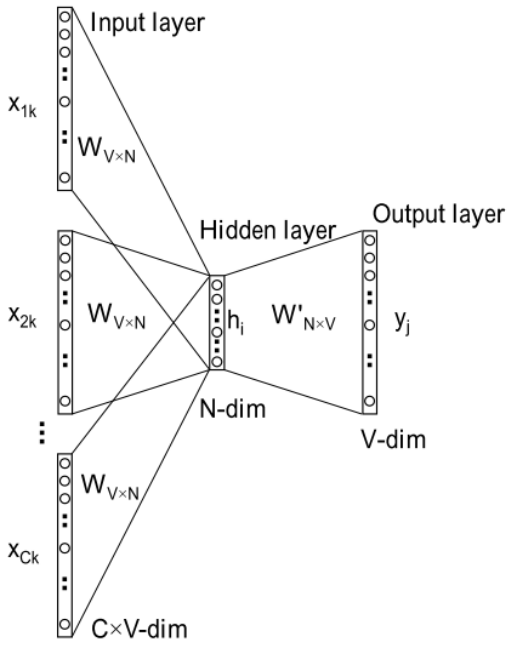


Figure 3.1: Continuous Bag of Word Model as presented in Rong [2014].

The CBOW model of the above figure takes as input C one-hot vectors of $|V|$ dimensions each, that correspond to the context representations of the target word. The

number of neighbors to be taken into account constitutes a design decision determined by the network creator. At the first stage, we multiply each one-hot vector with the matrix $W \in \mathbb{R}^{V \times N}$, which represents the weight matrix between the input and the hidden layer. Note that the dimensionality of the hidden layer is much smaller than that of the input vectors, and it coincides with the dimensionality of the produced embeddings. Given that the input vectors are one-hot encodings, we can notice that the true functionality of the input layer for the k-th context word is to forward the k-th row of the weight matrix to the hidden layer. After that, the vectors associated to each of the context words are added to the hidden vector. Consequently, the hidden vector is multiplied with the hidden to output weight matrix $W' \in \mathbb{R}^{N \times V}$ and finally it is passed through the soft-max function to compute the network's output layer. By taking the maximum value of the output vector and comparing its index with 1-hot-encoded words, we predict our target word. During the training of our neural network, we have the actual output of the target word so we can compute the error of our prediction, and back-propagate it to update the weight matrices until we meet a termination criterion.

To give a numerical example of the output prediction we continue on our previous paradigm. Suppose that we want to predict the word *Washington*, given the context words *media*, *in*, *about* and *terrorism*. Let the weight matrix that connects the input layer of the context word *media* to the hidden layer be equal to:

$$W = \begin{pmatrix} 0.1 & 0.0 & 0.1 \\ 0.0 & 0.3 & 0.7 \\ 0.9 & 0.1 & 0.6 \\ 0.3 & 0.3 & 0.6 \\ 0.2 & 0.6 & 0.9 \\ 0.2 & 0.3 & 0.7 \\ 0.8 & 0.9 & 0.1 \\ 0.1 & 0.3 & 0.5 \\ 0.1 & 0.7 & 0.1 \\ 0.5 & 0.5 & 0.1 \end{pmatrix}$$

The shaded row corresponds to the dense representation of the word *media* that will be passed for addition to the hidden layer, along with the dense representations of the other context words. Now, assume that the rows of the input weight matrix, which correspond to neighbor words, sum up to the hidden vector:

$$h = \begin{bmatrix} 0.4 & 0.7 & 0.6 \end{bmatrix},$$

which is subsequently transformed into a V-dimensional vector at the output layer, upon which the soft-max function is applied:

$$y = \begin{bmatrix} 0.01 & 0.08 & 0.10 & 0.07 & \mathbf{0.13} & 0.10 & 0.09 & 0.09 & 0.07 & 0.11 \end{bmatrix}.$$

The predicted word in our example corresponds to the 5-th word of our vocabulary (word *to*).

3.1.2 Skip-gram Model

The Skip-gram model is the opposite of the CBOW model. Its target is to predict the context words of a fixed window around a target word, when the only provided information is the latter. Like the CBOW model, it constitutes a shallow neural network with one hidden layer as depicted in Figure 3.2. Its input is the one-hot vector representation of the target word, which forwards the corresponding dense representation given in the input weight matrix to the hidden vector.

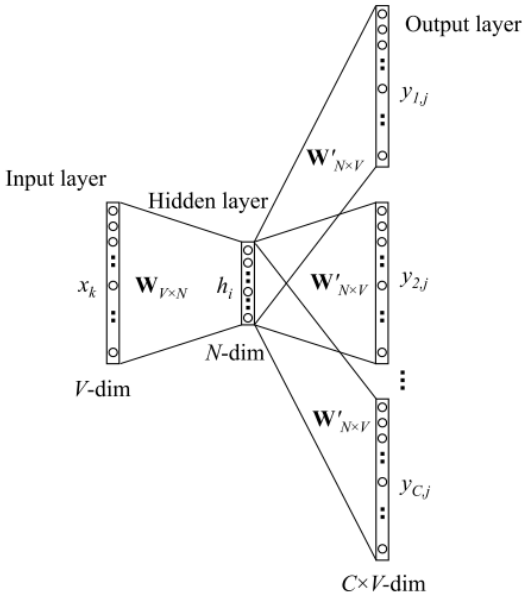


Figure 3.2: Skip-gram Model presented in Rong [2014].

The hidden vector is then multiplied with each output weight matrix $W' \in \mathbb{R}^{N\times V}$, and the soft-max function is applied to produce the output vectors in complete accordance to the CBOW model. The only difference is that this model generates C one-hot output vectors $\{y_i\}_{i=1,2,..,C}$, each of them indicating a word in the vocabulary as the predicted neighbor of the target word.

3.2 Latent Dirichlet Allocation

The Latent Dirichlet Algorithm (LDA), introduced by Blei et al. [2003], is a generative probabilistic model of a corpus, that attempts to identify the hidden topics lying behind it. In linguistics, the word "topic" refers to an abstract scheme that gives us information about what is talked about in a set of words (sentence/document). Putting this definition into a mathematical framework, we could imagine that a "topic" in NLP applications is described as a set of words that frequently occur together, or in statistical terms it could be presented as a distribution over the vocabulary of a corpus. Note that given the distribution, we can obtain the set of most related words of the vocabulary with respect to a topic via applying a threshold to its distribution (retain words with high probability).

3.2.1 Intuition

The basic idea of LDA is that documents (set of sentences) are represented as mixtures over topics, where each topic is characterized by a distribution over words. This assumption implies that a document could not exhibit only a single topic, which seems to be logical as documents are large entities of text. To gain insight into this assumption let's examine the article "Seeking Life's Bare (Genetic) Necessities", as well as the distribution of possible topics on it, as presented in Figure 3.3.

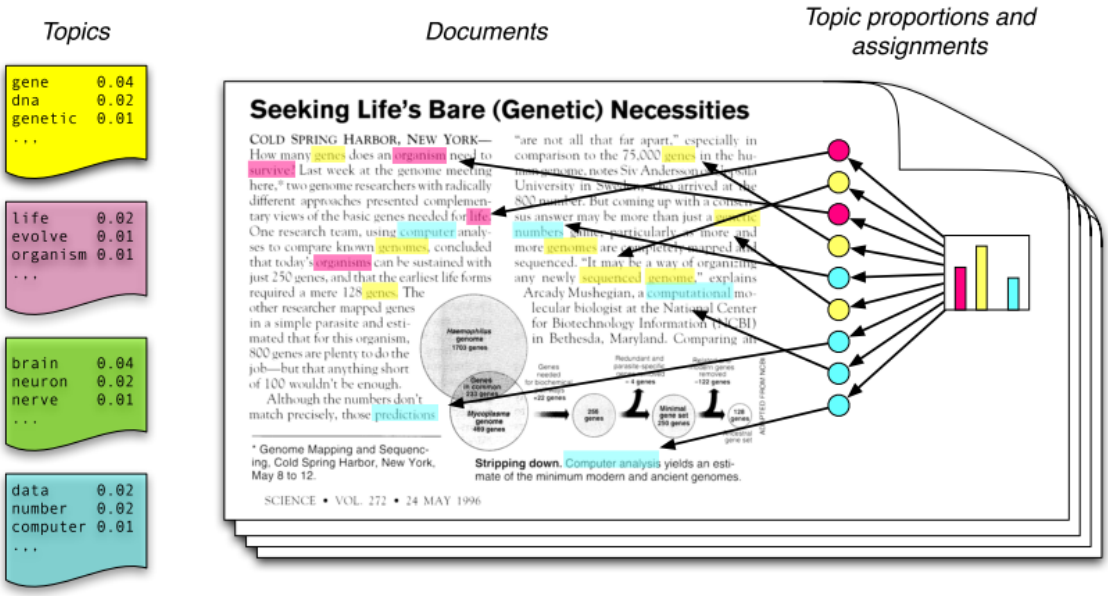


Figure 3.3: Intuition of LDA, an example presented in Blei [2012].

As explained in Blei [2012], the article is about using data analysis to determine the number of genes that an organism needs to survive (in an evolutionary sense). The article has been highlighted manually in order to create clusters of words that could be attributed to each of the topics residing in it: genetics, data analysis and evolutionary biology. The words about data analysis, such as "computational" and "prediction" have been highlighted in blue; words about evolutionary biology, such as "survive" and "organism", have been highlighted in pink; words about genetics, such as "sequenced" and "genes," have been highlighted in yellow. LDA tries to capture the above intuition, and automate the procedure of assigning topics to documents, and word distributions to topics. To do so, it supposes that each document is generated as follows:

1. Randomly choose a distribution over topics (histogram on the right).
2. For each word in the document:
   (a) Randomly choose a topic from the distribution over topics (colored coins).
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

3.2.2 Notation and Terminology

As we are going to put the above intuition into a mathematical framework, we should firstly introduce the basic notation and terminology needed to describe linguistic terms and concepts such as "words", "documents","corpora", "topic", as well as the document-topic and the topic-word distributions. Following Blei et al. [2003] we define:

- A *word* is considered the basic unit of our data, defined to be an item from a vocabulary indexed by $\{1, ..., V\}$. Mathematically, it is represented as a vector that has a single component equal to one and all other components equal to zero. For example, the representation of the first word of the vocabulary corresponds to a V-dimensional vector $w_1 = [\ 1\ 0\ 0\ 0\ 0\ ...\ ]$.
- A *document* is a group of N words denoted by $\boldsymbol{d} = (w_1, w_2, ..., w_N)$, where w_n is the n-th word in the group.
- A *corpus* is a collection of M documents denoted by $\boldsymbol{C} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, ..., \boldsymbol{d_M}\}$.
- A *topic* is a distribution over the vocabulary noted as β (β_k denotes the topic distribution of the k-th topic, where $k \in K$ and K corresponds to the total number of topics).
- The *document-topic* distribution for document $\boldsymbol{d}$ is defined as θ_d, while $\theta_{k,d}$ is the topic proportion of topic β_k in document $\boldsymbol{d}$.
- The *topic-word* distribution for document $\boldsymbol{d}$ is defined as z_d, while $z_{d,n}$ is the topic assignment for word w_n in document $\boldsymbol{d}$.

3.2.3 Algorithm

Generally, LDA could be described as a generative probabilistic model of a corpus, where the observed variables are documents and the latent variables are the topics residing in the corpus. As mentioned above, the basic idea of the algorithm is that each document could be assigned to a distribution over topics, where each topic is a distribution over words. In order to infer these distributions LDA assumes the following generative process for each document $\boldsymbol{d}$ in a corpus $\boldsymbol{C}$, whose graphical representation is given if Figure 3.4:

1. Choose $N \sim$ Poisson(ξ), where N corresponds to the number of words for $\boldsymbol{d}$.
2. Choose $\vartheta \sim$ Dirichlet(α)
3. For each of N words, w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(ϑ)
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

The ultimate goal of the above process is to estimate the hidden distributions $\theta_{1:D}$, $z_{1:D}$, $\beta_{1:K}$, given the observed variables $w_{1:D}$. As a result, the key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given the corpus as analyzed in Equation 3.1, using the Bayes' Theorem.

$$p(\beta_{1:K}, \vartheta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{3.1}$$

The numerator of the above fraction can be computed as the joint distribution of all random variables. However, in order to compute the denominator of the fraction we have to marginalize over all possible topic structures defined by $\theta_{1:D}$, $z_{1:D}$ and $\beta_{1:K}$. When doing so a coupling between $\theta_{1:D}$ and $\beta_{1:K}$ arises making the separation of them in the computation of the log likelihood function impossible.

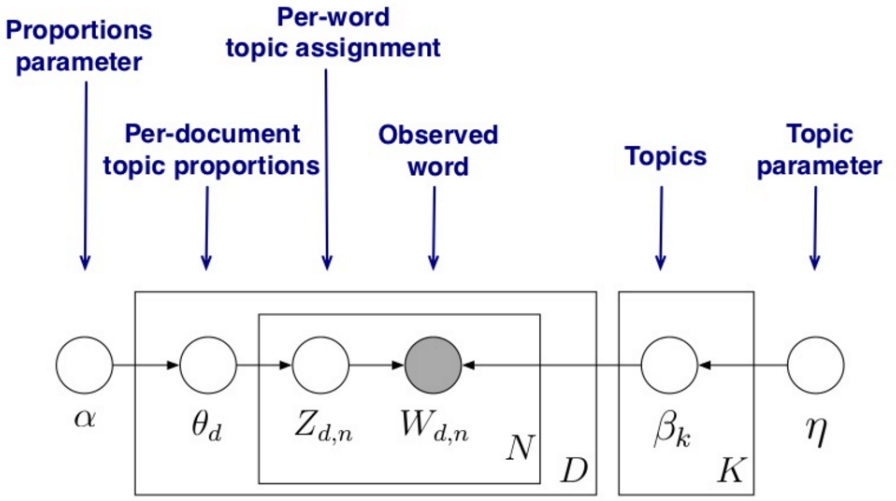


Figure 3.4: Graphical model representation of LDA as presented in Blei et al. [2003].

So while exact inference is not tractable, various inference techniques have been proposed in order to approximate the above solution:

- **Variational Inference**. The idea proposed by Reed [2012] was to modify the original graphical model of Figure 3.3 by removing the edges and nodes which are responsible for creating the undesirable coupling mentioned above. As a result a simpler distribution is used in order to approximate the real.

- **Collapsed Gibbs Sampling**. The approximation introduced by Griffiths and Steyvers [2004] was that a high-dimensional distribution is simulated by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others. The sampling is done sequentially and proceeds until the sampled values approximate the target distribution.

- **Collapsed Variational Inference**. Teh et al. [2006] made weaker factorization assumptions than those made by the Variational Inference algorithm in order to approximate the true posterior. Specifically, instead of assuming the parameters

to be independent from latent variables they treat their dependence on the topic variables, in an exact fashion marginalizing out the θ and β variables.

- **Online Variational Inference**. Later, Hoffman et al. [2010] noted that the Variational Inference algorithm requires a full pass through the entire corpus each iteration, making the whole procedure slow for large datasets. In this direction they proposed an online variational inference algorithm based on stochastic optimization with a natural gradient step. They also showed that the algorithm produces good parameter estimates on large datasets dramatically faster than batch algorithms.

3.3 Agglomerative Clustering

In this section we briefly describe agglomerative clustering that is utilized as a smoothing technique in our proposed approach. Agglomerative clustering constitutes a hierarchical clustering method which attempts to build a hierarchy of clusters, following a "bottom up" approach that separates a set of initial observations into distinct groups based on some measure of similarity. Specifically, given a set of K observations to be clustered and a $K \times K$ distance matrix (containing distances between observations), the basic process of hierarchical clustering, as described in Johnson [1967], is:

1. Start by assigning each item to its own cluster, so that if you had K items, you now have K clusters, each containing just one item. Let the distances between the clusters equal the distances between the items they contain.

2. Find the closest pair of clusters and merge them into a single cluster, so that now you have one less cluster.

3. Compute distances between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size K.

In order to iteratively follow the above procedure we could define distances between observations as well as distances between clusters of observations, via using any of the following definitions, as summarized in Tables 3.1 and 3.2.

Distance	Equation
Euclidean	$\|a-b\|_2 = \sum \sqrt{(a_i - b_i)^2}$
Squared Euclidean	$\|a-b\|_2 = \sum (a_i - b_i)^2$
Manhattan	$\|a-b\|_1 = \sum \|a_i - b_i\|$
Maximum	$\|a-b\|_\infty = \max_i (a_i - b_i)^2$
Cosine	$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$
Pearson	$1 - \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$
Spearman	$1 - \frac{6\sum d_i^2}{n(n^2-1)}$

Table 3.1: Different measures of distance between pairs of observations.

Linkage Criterion	Equation
Average	$\frac{1}{\|A\|-\|B\|}\sum_{a\in A}\sum_{b\in B} d(a,b)$
Single	$\min\{d(a,b) : a \in A, b \in B\}$
Complete	$\max\{d(a,b) : a \in A, b \in B\}$

Table 3.2: Different linkage criteria specifying the distance between clusters of observations.

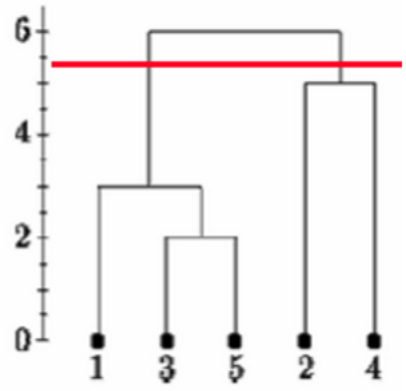


Figure 3.5: Dendogram depicting the hierarchical clustering of 5 observations.

The above procedure leads to the creation of a hierarchy of observations, as depicted in Figure 3.5. One of the problems with hierarchical clustering is that there is no objective way to say how many clusters there are. In order to acquire clusters of observations, we have to "cut" the hierarchy tree (a.k.a. dendogram) at some point. For example, the red line depicted in Figure 3.5 denotes the creation of two clusters.

Chapter 4

Mixture of Topic-based Distributional Semantic Models

4.1 Motivation

Recent approaches that produce multiple distributed representations per word make use of topic modeling techniques as discussed in Chapter 2. A topic model results into a parsimonious representation of the topics (thematic domains) that exist in the corpus under analysis. Typically, each topic is represented as a distribution of words being salient for the respective topic. The main motivation behind the use of topic models, for the task of word semantic similarity computation, is to adapt the similarity estimates provided from various topics. This is similar to using semantic mixture models to encode multiple senses in words. In this chapter, a topic-based semantic mixture model is discussed for the computation of semantic similarity between words. This is also motivated by previous approaches that utilize a combination of similarities computed via Topic-based Distributional Semantic Models (TDSMs).

4.2 Algorithm Description

The proposed model [Christopoulou et al., 2018] that is discussed in this chapter follows a two-step approach to extract semantic similarities between word-pairs provided either in-context or in-isolation, as presented in the following subsections.

4.2.1 Topic-based Distributional Semantic Models

Topic-based Sub-Corpora

Typically, the corpora used in many NLP applications consist of massive collections of text retrieved from various thematic domains, thus incorporating global contexts of generic information. However, topic-based (a.k.a. in-domain) corpora are extremely useful for language learning and comprehension. Despite their usefulness, substantial corpora are not yet available for many topic domains. One way to alleviate this lack of thematic data is to exploit available and diverse resources, like generic corpora, to induce their

construction. As detailed below, we present a method that creates topic sub-corpora via utilizing the Latent Dirichlet Allocation algorithm:

First, the only data required is a large corpus. It can be a generic one (downloaded from the web (e.g., Wikipedia)) or a collection formed from reviews (e.g., Movie reviews). We will need two versions of this corpus. One that contains documents, and another the contains sentences.

- The **document-level** version will be forwarded as input to the Latent Dirichlet Allocation algorithm, as in its default form assumes that each document contains multiple topics.
- The **sentence-level** version will be used in order to cluster the initial data to domain specific clusters, based on the assumption that smaller pieces of document (sentences) will possibly talk about one topic, so the classification can be strict. Moreover, this choice adheres to the basic principles of topic modeling, since sentences are topically complete and coherent units.

The creation of sub-corpora (see Figure 4.1) follows a three-step procedure:

1. Starting from a generic corpus of documents we use the Latent Dirichlet Allocation (LDA) algorithm to train a topic model, for a number of topics K. LDA associates each document with topic proportions motivated by the idea that a variety of different topics resides in each of them. The trained topic model produces a distribution of words for each topic, that are semantically related under the corresponding topic.
2. Afterwards, we apply the trained LDA model to the sentence-version of the corpus. As a result, each sentence is probabilistically associated with a list of topics, discussed in the sentence, according to the trained topic model.
3. Finally, a sub-corpus is created for each topic $k \in K$ by aggregating the sentences, the posterior probabilities of which are maximized for k. This hard-clustering scheme may result in sub-corpora of limited size. In order to relax this limitation, a soft-clustering scheme is adopted. Specifically, a sentence is allowed to be included in a topic-specific corpus when the posterior probability for the corresponding topic exceeds a threshold h. Sentences exhibiting equal posterior probabilities across all topics are excluded from this process, as they are considered too generic to provide any topic-related information.

Topic-based Distributional Semantic Models

As mentioned in Chapter 2, DSMs encode the contextual information of words derived from corpora, into dense feature representations. As a result, the corresponding DSMs that are trained under them lead to the creation of generic word representations, as the contextual variations a word exhibits into different topic-domains are not taken into account. The isolation of different word senses could be achieved by collecting topic-related snippets into separate bodies of text, using topic modeling techniques as described above.

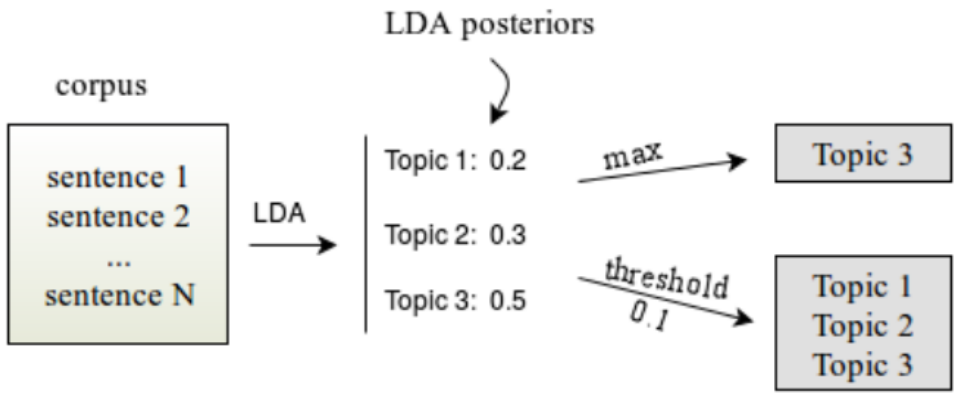


Figure 4.1: Abstract Depiction of Topic-based Sub-Corpora Creation [Christopoulou, 2016].

We use the term Topic-based Distributional Semantic Model (TDSM) throughout the thesis, to refer to a DSM that is constructed over a topic specific sub-corpus. Specifically, we run the Word2Vec algorithm in order to create topic-specific representations of words. Note that TDSMs could also be created over the extracted corpora, using any DSM that encodes linguistic features from text in order to embed words in a semantic space. Figure 4.2 summarizes the procedure we follow to induce the creation of topic-specific representations of words starting from a generic corpus.

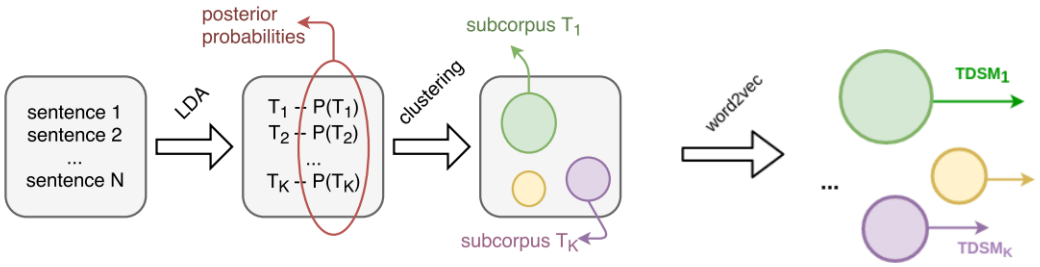


Figure 4.2: Abstract Depiction of Topic-based DSMs Construction.

4.2.2 Semantic Mixtures

Typically, the computation of semantic similarity between a pair of words is performed across all of their senses that appear in a corpus. For various semantic tasks related to similarity computation such models were found to achieve very good performance despite their divergence from the maximum sense similarity assumption, which suggests that the semantic similarity between two words can be estimated as the similarity of their two closest senses. In the discussed model, the aforementioned assumption is adopted via the creation of topic-based sub-corpora with respect to any pair of words, $word_i$ and $word_j$, subjected to similarity computation. The goal is the words of the pair to co-occur in each sub-corpus with their closest senses, pertaining to the relevance with the respective topics.

This approach is different compared to the typical corpus-based word sense induction (also referred to as sense discovery), where the discovery is performed individually for each word. The similarity between $word_i$ and $word_j$ is computed by a mixture model that combines similarity scores computed over multiple topic-based sub-corpora, as depicted in Figure 4.3.

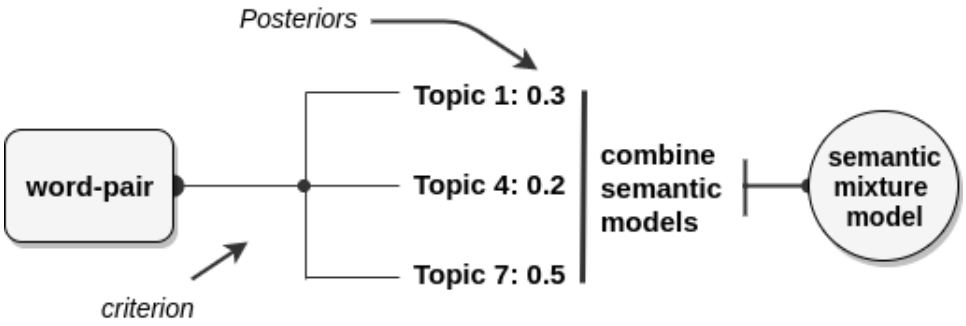


Figure 4.3: Abstract Depiction of the semantic mixture model.

Specifically, we follow two different approaches to compute the *posteriors* depicted above utilizing different *criteria*, based on the semantic similarity task we use for our evaluation. We define L_K as the set of K topic-specific DSMs (TDSMs) derived from the LDA algorithm, where λ_k is the DSM trained on topic k out of the K topics in total.

Contextual Semantic Mixture

When contextual information is provided for a word-pair (w, w'), a shared context $c'' = c \oplus c'$ is formulated by concatenating the contexts of each word c and c', respectively. The topic model is fed with c'' and outputs a list of candidate topics for c'', along with the corresponding posterior probabilities. These topics are utilized for identifying the respective sub-corpora, which are used to train topic-specific DSMs (TDSMs). In order to consider context information, we define two similarity metrics:

$$S_{\text{AvgSimC}}(w, w'; L_K) = \frac{\sum_{k=1}^{T(c'')} p(k|c'') S_k(w, w'; \lambda_k)}{\sum_{k=1}^{T(c'')} p(k|c'')}, \tag{4.1}$$

$$S_{\text{MaxSimC}}(w, w'; L_K) = S_{\hat{k}}(w, w'; \lambda_{\hat{k}}) \quad \text{where} \quad \hat{k} = \arg\max_{k \in T(c'')} p(k|c''), \tag{4.2}$$

where $T(c'')$ are the candidate topics returned by the topic model with a posterior probability larger than 0.01, when given as input a shared context c'', $p(k|c'')$ denotes the posterior probability of topic k for c'', while $S_k(w, w'; \lambda_k)$ is the cosine similarity of w and w' from the DSM that corresponds to topic k. Because the number of candidate topics can be less or equal to the total number of topics ($T(c'') \leq K$), for which LDA is trained, the posterior probabilities are normalized to sum to unity.

Given c'' as input to the topic model, Equation 4.1 computes a weighted average of topic-based semantic similarities using the topics posterior probabilities as weights. Note that for pairs that share the same word, but are found in different contexts, the

model always assigns to them a similarity score equal to one, as their representations are extracted from the same topic-based DSM. The described mixture scheme takes the middle road between the maximum sense similarity hypothesis and the sense-agnostic DSMs. This hypothesis is adopted for the identification of sub-corpora in which w and w' appear with related senses under the thematic domain of the corresponding topic. The incorporation of the mixture weights in the computation of the final similarity relaxes the hypothesis. Using Equation 4.2 a pair is assigned the semantic similarity of the topic with the maximum posterior probability, hence the dominant topic in the provided context.

Additionally, we introduce a fusion model that combines information from multiple topic models trained for different number of topics. In more detail, for a topic model trained on K topics, the semantic similarity of a word pair is calculated using one of the aforementioned metrics, as defined in Equations 4.1, 4.2. Among the similarities produced by training the topic model for various number of topics, we select the maximum pair similarity over a group G, of topic DSM sets L_K , generated by different topic models.

$$S_{\text{Fuse}}(w, w') = \max_{L_K \in G} S_{*\text{Sim}}(w, w'; L_K), \tag{4.3}$$

where $S_{*\text{Sim}}(w, w; L_K)$ is the (w, w') pair similarity computed with 4.1, 4.2, using a topic model trained on K topics and G is the group of DSM sets that will be fused.

Non-contextual Semantic Mixture

The semantic similarity between two words w and w' is computed using different similarity metrics with respect to the presence of context for each pair. A mixture model of topic-based semantic similarities is incorporated to produce the final similarity $S(w, w')$ between a word pair. In accordance with Reisinger and Mooney [2010], we define two non-contextual metrics:

$$S_{\text{AvgSim}}(w, w'; L_K) = \frac{1}{K} \sum_{k=1}^{K} S_k(w, w; ; \lambda_k), \tag{4.4}$$

$$S_{\text{MaxSim}}(w, w'; L_K) = \max_{k \in K} \{S_k(w, w; ; \lambda_k)\}, \tag{4.5}$$

where $S_k(w, w'; \lambda_k)$ is the cosine similarity of w and w' computed by the λ_k DSM, which was built using the sub-corpus that corresponds to topic k. In Equation 4.4, the unweighted average of all topic-based pairwise semantic similarities is computed. In Equation 4.5 only the maximum pairwise similarity, among K topics, is selected.

Finally, we employ a linear regression model to combine pairwise similarities between topic-specific DSMs, resulted from a topic model trained on K topics. The model aims to minimize the Mean Squared Error (MSE) by training a set of β weights on a group of similarities between words. The motivation behind this idea is to learn how to combine topic-specific similarities for isolated words. The context-dependent similarity metric (Equation 4.1) requires additional input (context) to estimate how much each topic-similarity will be weighted. In contrast, when no context is present, instead of assuming that all topics contribute equally to the estimation of a pairwise similarity, as described in Equation

4.4, we argue that a linear combination of topic-similarities will produce a more precise estimation,

$$S_{\mathrm{LR}}(w, w'; L_K) = \beta_0 + \frac{1}{K} \sum_{k=1}^{K} \beta_k S_k(w, w'; \lambda_k), \tag{4.6}$$

where β_k are learned weights by the regression model for the corresponding topic k, $S_k(w, w'; \lambda_k)$ is the similarity of a pair (w, w') computed from the DSM trained on the sub-corpus of topic k and β_k is a bias weight. The β weights sum to unity.

Chapter 5

Unified multi-Topic Distributional Semantic Model

5.1 Motivation

Traditional Distributional Semantic Models (DSMs) represent feature spaces of meaning, where the multiple senses of polysemous words are conflated in single representations. In this chapter, we discuss the creation of a unified model that assigns multiple distributed representations per word, via aligning Topic-based DSMs (TDSMs) to a shared space. Motivated by the assumption that semantic relationships between monosemous words do not change in different topic domains, we assume that the relative distances of their corresponding representations are preserved across TDSMs, acting as *semantic anchors* that determine the mappings between them.

We propose that the Unified multi-topic DSM (UTDSM) system constitutes a more flexible model when compared to mixture of TDSMs (described in Chapter 4), for two main reasons:

- TDSMs fail to capture the different senses for a pair consisting of the same words. In more detail, the comparison between two words is always restricted to be computed under the same topic domain. As a result, identical words —despite their senses—are always represented by the same embeddings and therefore are assigned the maximum possible similarity defined by the distance metric that is used.
- The described restriction leads also to a weakness of the previous work when two different words are compared, as it assumes that the semantic relationship of two words is defined in a *shared* topic domain. As a consequence, two words that do not share any sense in the same topic could not be accurately compared.

5.2 Algorithm description

Our system follows a four-step approach briefly described in the following sub-routines:

1. **Global Distributional Semantic Model.** Given a large collection of text data we train a DSM that encodes the contextual semantics of each word into a single representation, also referred to as global-DSM.

2. **Topic-based Distributional Semantic Models.** Next, a topic model is trained using the same corpus. The topic model splits the corpus into K (possibly overlapping) sub-corpora SC_1, ... , SC_K similar to Chapter 4. A DSM is then trained from each sub-corpus resulting in K topic-based DSMs, i.e., $TDSM_1$, ... ,$TDSM_K$. The topical adaptation of the semantic space takes into account the contextual variations a word exhibits under different thematic domains and therefore leads to the creation of "topic-specific" vectors (topic embeddings).

3. **Mappings of topic embeddings.** Next, we map the vector space of each topic-based DSM to the shared space of the global-DSM, using a list of monosemous words that is statistically related to the corresponding topic. In the unified semantic space each word is represented by a set of topic embeddings that were previously isolated in distinct vector spaces, thus creating a Unified multi-Topic DSM (UTDSM).

4. **Smoothing of topic embeddings.** As an optional step, we utilize a smoothing approach based on the use of agglomerative clustering of a word's topic embeddings into N groups. We suggest that this step reduces the noise introduced to our system through the semantic mappings and sparse training data.

The alignment of TDSMs under a common space, leads to the creation of a unified model of multiple topic-based distributional semantic representations, as depicted in the abstract scheme of Figure 5.1.

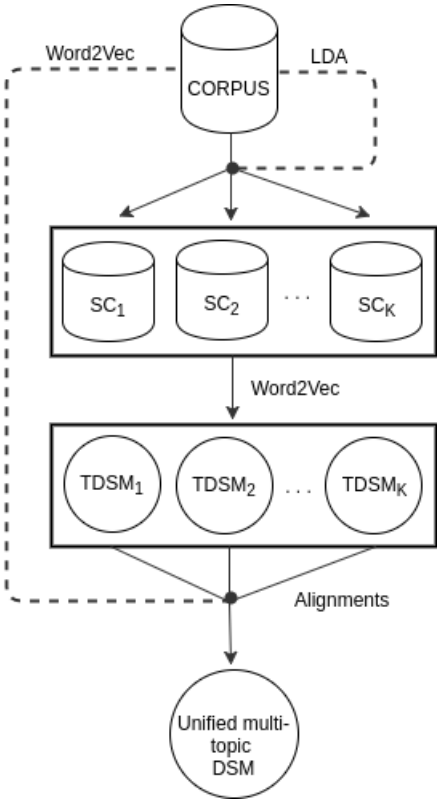


Figure 5.1: Unified multi-topic DSM.

5.2.1 Distributional Semantic Models

Starting from a generic corpus we train a global-DSM as well as K TDSMs as proposed in Chapter 4, where K constitutes a fixed parameter of our system. We make the following assumptions for the extracted DSMs regarding the linguistic features they capture in their corresponding vectors.

- **Global-DSM** provides a single feature representation per word. Given that these representations are extracted from a generic domain, we assume that the meanings encapsulated in them are not biased towards a topic direction. For this reason, we expect that the vector assigned to a monosemous word (a.k.a. word having only one sense) will be well positioned in this space. On the other hand, we expect that the vector assigned to a polysemous word (a.k.a. word having multiple senses) will be placed in a position that reflects an average representation of its actual senses. For instance, the polysemous word *cancer* is expected to have a close relative distance to the words *astrology* and *tumor*.

- **TDSMs** also provide a single feature representation per word. The difference from the global-DSM is that they encapsulate in-domain variations in usage of language, which facilitates the de-conflation of the multiple meanings of polysemous words in topic distributional representations. As a result, the polysemous word *cancer* will be shifted towards the word *astrology* in a zodiacal semantic space. On the other hand, the same word will be positioned closer to the representation of the word *tumor* in a medical semantic space.

5.2.2 Mapping of topic embeddings

The intrinsic non-determinism of the Word2Vec algorithm leads to the creation of continuous vector spaces that are not naturally aligned to a reference coordinate system. For this reason, we need to align word vectors from different TDSMs under common coordinate axes, in order to enable their comparison. In particular, we suggest that TDSMs capture meaningful variations in usage of polysemous words, while in parallel preserve the relative positions of monosemous words, whose meaning can be encoded in single vector representations. This observation motivated us to think of monosemous words as fixed points in our topic semantic spaces, that could be used as *semantic anchors* to determine the mappings between them.

In this work, we assume that there exists a transformation matrix M_k between the vector representations for the monosemous words of each TDSM and the corresponding representations of the global-DSM. As our ultimate goal is to align all TDSMs to the unified coordinate system, we use a global-DSM as the target space, whereas the TDSM representing topic k serves as the source space. We suggest the usage of the global-DSM as the target space as it provides robust representations for the monosemous words.

Let $X = [x_1, x_2, ..., x_n]$ and $Y = [y_1, y_2, ..., y_n]$ be the matrices of word embeddings corresponding to the source and target spaces, where $x_i,\ y_i \in \mathbb{R}^d$ and $X, Y \in \mathbb{R}^{d \times n}$. Our ultimate goal is to find a transformation matrix $M_k \in \mathbb{R}^{d \times d}$ for each topic k, that approximates the Equation 5.1:

$$M_k X = Y. \tag{5.1}$$

To solve the above problem we experiment with the well-known mapping methods that are used in the literature and have already been discussed in Chapter 2.

- Linear
- Orthogonal
- Canonical Correlation Analysis

In order to acquire multiple topic embeddings that reside in a unified vector space, we follow the above transformation for each TDSM, meaning that for a topic group consisting of K topics, we learn a set of matrices $\{M_k\}_{k=1}^{K}$. The set of topic embeddings that correspond to a specific word is represented as $\{s_k\}_{k=1}^{K}$. Specifically, given a word and its k-th topic distributed representation $u_k \in \mathbb{R}^d$, we compute its projected representation $s_k \in \mathbb{R}^d$ as follows:

$$s_k = M_k u_k. \tag{5.2}$$

Returning to the previous example, the word *cancer* will be assigned multiple distributed representations that reflect its different in-domain senses in the Unified Topic-based DSM. A simplified depiction of this example is presented in Figure 5.2.

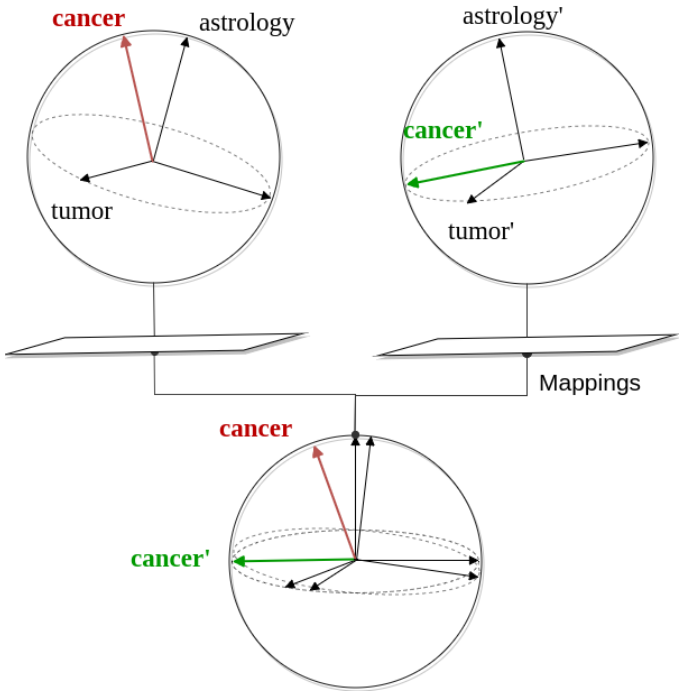


Figure 5.2: Simplified depiction summarizing the intuition behind the alignment process of topic embeddings. In the unified system, the polysemous word *cancer* is represented by two topic vectors that capture different semantic properties of the word under a zodiacal and a medical topic. Words *astrology* and *tumor* are examples of *semantic links* that define the mappings and preserve the relative positions of monosemous-monosemous and monosemous-polysemous words.

5.2.3 Smoothing of topic embeddings

Starting from the set of mapped topic embeddings $\{s_k\}_{k=1}^{K}$ for each word, we employ an agglomerative clustering of the set into N clusters, where closely positioned topic embeddings are assigned to the same group. We suggest that each cluster forms a semantic coherent unit that corresponds to closely related semantics of the target word. Subsequently, the vectors of each cluster are averaged to create a representative vector of the group, leading to a new set of *smoothed* topic embeddings $\{s'_n\}_{n=1}^{N}$ for each word, where $s'_n \in \mathbb{R}^d$.

5.3 Semantic Similarity

This subsection describes how we leverage the topic embeddings of the Unified TDSM for the computation of word similarity between two words. This task traditionally measures the degree of likeness of a word-pair in the absence of sentential context. However, the real problem that multiple distributed representation systems attempt to solve is that of the polysemic nature of words which highly depends on the words' contexts. For this reason, we also discuss the estimation of contextual semantic similarity between words where the specific meanings of polysemous words are triggered by the provided sentential information. Overall, to compute the semantic similarity between a pair of words (either provided in-isolation or in context), we follow the known metrics used in the literature by systems of multiple distributed word representations.

5.3.1 Contextual Metrics

When context information for a pair of words is provided, we employ the AvgSimC and MaxSimC contextual metrics, firstly discussed in Reisinger and Mooney [2010]. Given the word-pair (w, w') and their provided contexts (c, c') we define:

$$\text{AvgSimC}(w, w') = \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} p(j|w, c) p(k|w', c') d(u_j(w), u_k(w')), \tag{5.3}$$

$$\text{MaxSimC}(w, w') = d(\hat{u}(w), \hat{u}(w')), \tag{5.4}$$

where K is the number of topics returned by a trained LDA model, u_j is the TDSM trained on the sub-corpus corresponding to the j-th topic after being projected to the unified vector space, $p(j|w, c)$ denotes the posterior probability of topic j returned by LDA given as input the context c of word w, d denotes the cosine similarity between the two input representations and finally $\hat{u}(w) = u_{\arg\max_{1 \le k \le K} p(k|w,c)}(w)$ is the vector representation of word w that corresponds to the topic with the maximum posterior for c.

Thus, AvgSimC corresponds to soft cluster assignment, weighting each similarity term by the likelihood of the word contexts appearing in their respective topics, while MaxSimC corresponds to hard assignment, using only the most probable topic assignment.

5.3.2 Non-contextual Metrics

When context information is not provided, we define the similarity between the word pair (w, w') using the following metrics:

$$\text{MaxSim}(w, w') = \max_{\substack{1 \le j \le K \\ 1 \le k \le K}} d(u_j(w), u_k(w')), \tag{5.5}$$

$$\text{AvgSim}(w, w') = \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} d(u_j(w), u_k(w')). \tag{5.6}$$

The MaxSim metric incorporates the popular maximum sense hypothesis according to which: *the similarity between two words is the maximum similarity between their senses*, while the AvgSim metric assumes that all possible representations of a word should contribute equally to the computation of the semantic similarity between w and w'. Note that AvgSim and MaxSim can be thought of as special cases of AvgSimC and MaxSimC with uniform weight to each topic; hence AvgSimC and MaxSimC can be used to compare words in context and isolated words as well.

The two above non-contextual metrics have been widely used by the systems of multiple distributed representations when comparing words in-isolation. However, they do not take into account neither the quality of the corresponding representations of a word nor the dominance of the "sense" represented by each embedding. As context information is not provided to guide our choice of embeddings in this task, machine learning techniques could be used in the non-contextual metrics to help us discover a pattern of fusing the pair similarities computed between multiple representations. This method has been applied in Chapter 4, were we used linear regression to learn the weight coefficients of our mixture model. However, one drawback of the linear regression method is that it is not scalable to large number of representations per word (a lot of data are required).

Iacobacci et al. [2015] proposed another scalable weighted fusion scheme that incorporated similarities using machine learning. They highlighted a deficiency of the MaxSim metric based on psycholinguistic studies. Specifically, they noted that taking the similarity of the closest senses of two words as their overall similarity ignores the fact that the other senses can also contribute to the process of similarity judgement. In fact, psychological studies suggest that humans, while judging semantic similarity of a pair of words, consider different meanings of the two words and not only the closest ones. For this reason, they used a weighted similarity measurement in which different senses of the two words contribute to their similarity computation, but the contributions were scaled according to their dominance. Following their weighted similarity measurement strategy we use the AvgSimW metric that is a modified version of the AvgSimC metric when context information is not provided. Specifically, we define the similarity between the word pair (w, w') as follows:

$$\text{AvgSimW}(w, w') = \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} p(j|w) p(k|w') d(u_j(w), u_k(w'))^{\alpha}, \tag{5.7}$$

where the contribution of the j-th topic representation of word w is weighted by the term $p(j|w)$ that denotes the conditional probability of the topic j given w. Furthermore,

according to Iacobacci et al. [2015] the importance of a specific sense of a word can also be triggered by the word it is being compared with. This is modeled by biasing the similarity computation towards closer senses, by increasing the contribution of closer senses through a power function with parameter α.

Moreover, to evaluate the performance of our model on the condensed set of *smoothed* topic embeddings the probabilities of Equations 5.3, 5.4 and 5.7 are summed for each of the aggregated vectors.

Chapter 6

Evaluation & Results

In this chapter, we describe the experimental procedure and the evaluation results. First, we present the corpus creation, the parameter values of our models that are investigated in this thesis, the datasets that are utilized for evaluation purposes and our baseline system (see subsection 6.1). Then, we continue with the evaluation of our first model (Mixture of TDSMs) on semantic similarity tasks (see subsection 6.2). Afterwards we present the experiments of our second model (UTDSM) that includes the evaluation of our semantic mappings on the contextual semantic similarity task (see subection 6.3), the investigation of a smoothing technique (see subsection 6.4) and the presentation of our best predictive models for in-isolation semantic similarity computation (see subsection 6.5). We also compare our two models with state-of-the-art systems found in the literature (see subsection 6.6). Finally, we present visualization examples of our second model in a two dimensional space (see subsection 6.7).

6.1 Experimental Settings

Corpus. The main corpus that is used for the construction of our models is a generic, web-harvested corpus that is created according to Iosif and Potamianos [2015]. As a first step, the definition of a vocabulary in a specific language is required. To do so, we use a list consisting of $8,752$ English nouns extracted from the SemCor3 corpus. Then, for each word of the vocabulary an individual query (a "search text" that is sent to a search engine) is formulated that contains only the word itself. Afterwards, each query is posed separately to the Yahoo! search engine and then only the $1,000$ top-ranked documents are collected. Thereafter, from each document, its snippet (small paragraph under the URL of the result that usually describes each document) is extracted, and finally all snippets from all queries are aggregated, resulting in a corpus of 116 million sentences. As described in Chapter 3 our model requires both a sentence-version and a document-version of our corpus. In order to create the latter, the sentences of the corpus are concatenated sequentially in groups of 100, forming a document-level version of 900 thousand documents. Finally, traditional preprocessing techniques are applied to both versions of our corpus including tokenization (splitting sentences to tokens), removal of stop-words (words that do not carry important information like 'and', 'to', 'the' etc.), removal of punctuation and duplicate lines, and capitalization (reduce all letters to lower case).

Parameters. For the construction of Topic-based DSMs we use the Gensim implementation of the Latent Dirichlet Allocation topic model [Řehůřek and Sojka, 2010] and train 7 models with the number of topics spanning from 5 to 60. The threshold used for the soft-clustering of the generic corpus to topic sub-corpora is set to 0.1. Google's implementation of Word2Vec and Continuous Bag-of-Words method is employed to train both the global-DSM and the Topic-based DSMs. The context window parameter of Word2Vec is set to 5, while we experiment with 100 and 300 dimensions. Any parameter not mentioned is set to default values of the corresponding implementations (e.g., Word2Vec, Gensim LDA). For a thorough explanation of the above setting options we refer the reader to Christopoulou [2016]. Concerning the mappings and smoothing of topic embeddings we use different number of monosemous or random words ranging from $1,000$ to $6,000$, and different number of clusters which is analogous to the number of topic embeddings of each word. Specifically, given a set of topic embeddings with cardinality equal to K, the final set of smoothed embeddings has a cardinality of pK with $p \in (0, 1]$. Overall, in this thesis we investigate the role of the following four parameters of our models:

- Dimensionality of DSMs.
- Number of topics.
- Number of words used for the semantic mappings of topic embeddings.
- Number of clusters used for the smoothing of topic embeddings.

Datasets. To evaluate the performance of our models we use datasets that provide human judgments on semantic similarity between pairs of words. By using these human annotations as gold standard values we estimate the reliability of our models' predictions. More specifically, the Spearman's ρ correlation coefficient is selected as evaluation metric to compare our estimated similarities against the ground truth. The construction of the above datasets typically follows the following procedure: a pair of words is presented to humans for annotation, where every annotation measures how similar the two words are as perceived by a human on a predetermined scale; afterwards annotations are aggregated to obtain an average measure of similarity between the two words. For example, the pair (*automobile*, *vehicle*) receives an average value of 45 in a $[0, 50]$ scale, which indicates that the above words are evaluates as highly similar. The above pair constitutes an example of words presented in-isolation to the human annotators, thus leading to the creation of a non-contextual dataset. On the other hand, contextual datasets are created via providing a sentential context for each word during the annotation phase. For example, the pair (*tiger*, *tiger*) receives an average value of 2 in a $[0, 10]$ scale, when the words are presented in the contexts (*... tiger with as many as of hunts ending in a kill reproduction over the course of her life, ... famous sport figures like tiger woods ...*), which indicates that the words are evaluated as highly dissimilar given their corresponding contexts. For our experiments we use 3 non-contextual datasets, and the only available contextual dataset. Their descriptions are presented below:

- **WordSimilarity-353** (WS-353) is a widely used dataset for evaluation of semantic similarity; it constitutes a set of 353 English word pairs along with human-assigned

similarity judgments in a $[0, 10]$ scale. WS-353 is a collection of pairs for measuring both word similarity and relatedness, so it has been split into two subsets, one for evaluating similarity, and the other for evaluating relatedness [Finkelstein et al., 2002].

- **MEN** contains $3,000$ pairs of randomly selected words to ensure a balanced range of similarity levels in the selected word pairs. The ground truth judgments were evaluated on a $[0, 50]$ scale, based on crowd-sourcing [Bruni et al., 2014].
- **RG** or Rubenstein and Goodenough is a set of 65 noun pairs with human similarity ratings in a $[0, 4]$ scale [Rubenstein and Goodenough, 1965].
- **SCWS** or Stanford Contextual Word Similarity dataset constitutes the only contextual dataset, published by Huang et al. [2012]. It consists of $2,003$ pairs of words along with sentences containing these words and human labeled word similarity scores in a $[0, 10]$ scale.

Baseline. The performance of the global-DSM model is utilized in the following experiments as our baseline system. Recall that the global-DSM provides one vector representation per word and as a result it does not account for the polysemic nature of words.

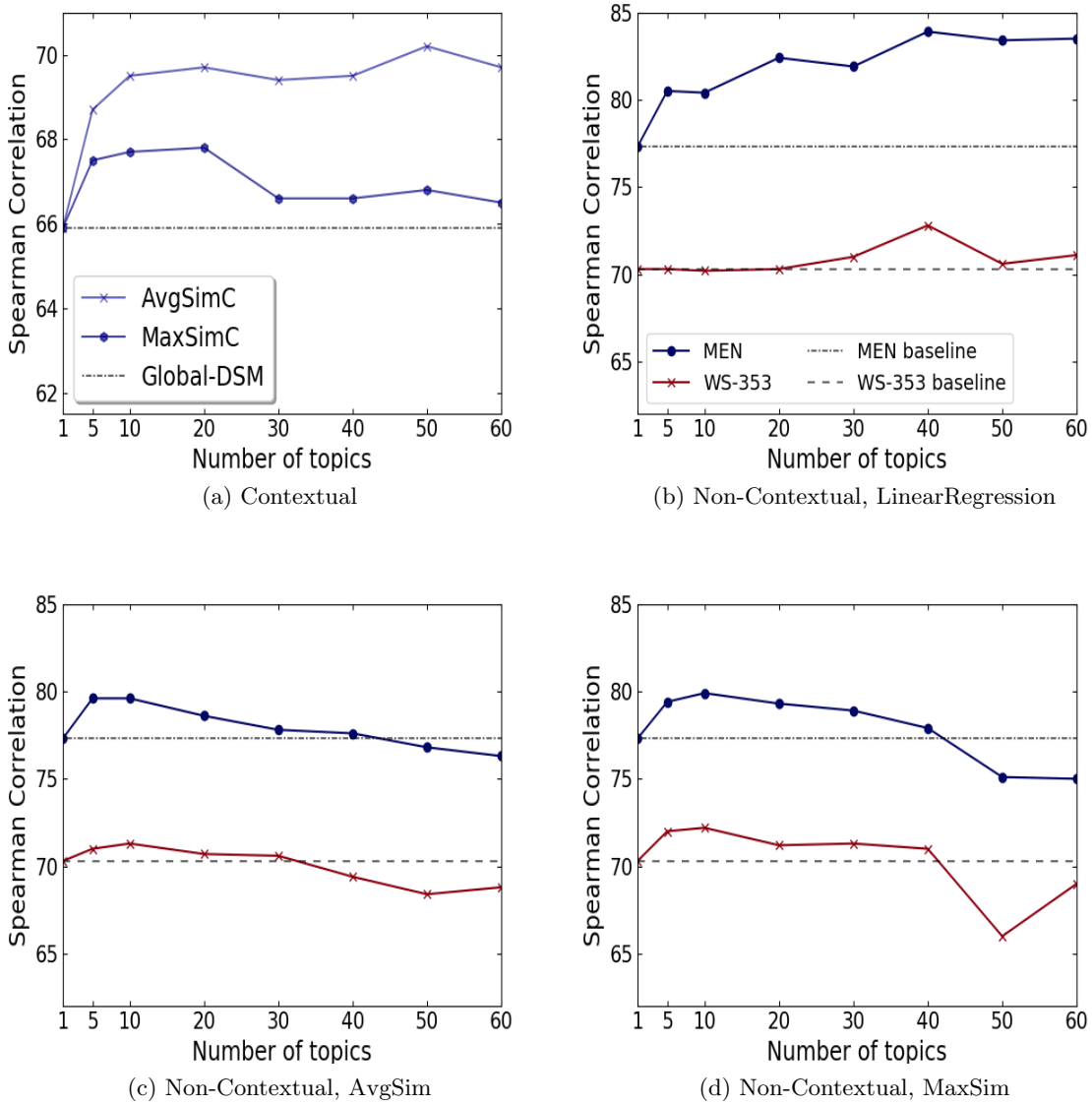
6.2 Semantic Mixtures

In this subsection we present the evaluation results obtained using Mixture of Topic-based Distributional Semantic Models (TDSMs), as described in Chapter 4. The performance of the TDSMs model, for each contextual semantic mixture scheme, is depicted in Figure 6.1 (a) as a function of the number of topics for the SCWS dataset. The top performance (70.2) is achieved by the AvgSimC metric when utilizing 40-50 topics. Figure 6.1 (b) illustrates the performance of the TDSMs-LR model, as a function of the number of topics. Regarding the MEN and WS-353 datasets, the proposed approach is shown

Mixture	In-context		Out-of-context	
	WS-353	MEN	SCWS	
			MaxSimC	AvgSimC
Global-DSM	70.3	77.3	65.9	65.9
TDSMs	-	-	67.8	70.2
TDSMs-Fuse	-	-	67.4	70.5
TDSMs-LR	72.2	83.8	-	-
TDSMs-MaxSim	72.2	80.0	-	-
TDSMs-AvgSim	71.3	79.6	-	-

Table 6.1: Performance comparison between best results obtained for different semantic mixture schemes and different datasets for semantic similarity computation in terms of Spearman's ρ correlation.

to outperform the baseline for all number of topics. The top correlation score (83.8) is achieved for 40 for the MEN dataset. For the WS-353 dataset, the same combination of topics and corpus provides the top performance (72.7). Consequently, we discuss in more detail the results obtained for each mixture scheme as described in Chapter 4. The best predictive performances for each of them are reported in Table 6.1.



(a) Contextual

(b) Non-Contextual, LinearRegression

(c) Non-Contextual, AvgSim

(d) Non-Contextual, MaxSim

Figure 6.1: Performance comparison between different number of topics and different mixture schemes of TDSMs, in terms of Spearman's correlation on different datasets. Plot (a) depicts the performance of AvgSimC and MaxSimC mixtures on SCWS dataset. Plots (b), (c) and (d) depict the performance of linear regression, AvgSim and MaxSim mixture schemes on MEN and WS-353 datasets. Baseline performances are also depicted.

The improvement over the baseline performance, achieved by the proposed approach for the semantic similarity computation task, is clearly demonstrated through the use of

three datasets. We observe that the top performance (72.7), achieved for the WS-353 dataset, is lower compared to the highest correlation score (83.8) obtained for MEN. This can be attributed to the different dataset designs, e.g., the type of semantic relationship, as well as the procedure followed for the collection of human ratings. Overall, the reported improvement for the MEN dataset is statistically more significant compared to the case of WS-353 due to the larger size of the MEN dataset (i.e., 3000 vs. 353 pairs).

The number of topics constitutes a key parameter of the proposed approach. The identified topics are used for corpus filtering (i.e., creation of sub-corpora) upon which the creation of DSMs is based. In this framework, when computing the similarity between a word pair, we argue that the exploited sub-corpus exhibits two properties: i) The sub-corpus is semantically coherent, i.e., the two words should appear with their closest word senses, and ii) adequate data exist enabling the computation of DSMs. Typically, a larger number of topics improves the semantic coherence of the respective sub-corpus (increased topic specificity), but it may cause the fragmentation of the training data lowering the quality of the semantic models. This problem is more obvious in the cases when the MaxSim and AvgSim mixtures are used as illustrated in Figure 6.1 (c) and (d) where peak performances are obtained for large values of K.

In order to overcome this issue, the linear regression approach is suitable for selecting the best similarities from the respective topic-based DSMs, for pairs without context as shown in Figure 6.1 (b). The method surpasses the baseline for very small number of topics but seems to work better for a larger number of topics despite the data fragmentation. This behavior can be explained by considering that without a given context, a word could have an arbitrary number of senses. An augmented sense-space enables the accurate estimation of a word pair similarity as a linear combination of different sense-related similarities.

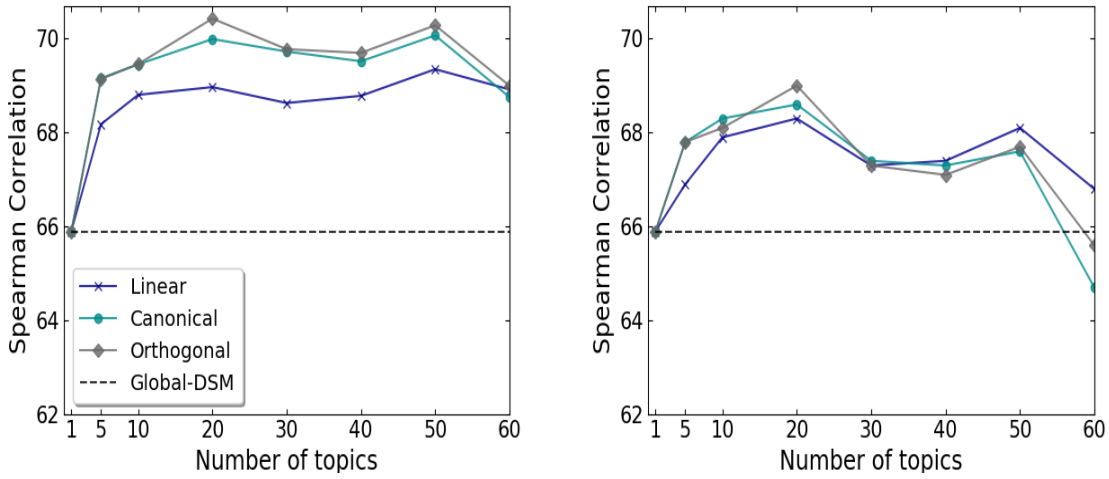
Finally, as shown in Table 6.1 the fusion model improves on the performance of the best predictive configuration that corresponds to the mixture of topic similarities extracted from a single topic groups G. Specifically, the fusion model provides the best results when all topic groups were used. This is expected as it resembles the functionality of a hierarchical topic model. Hierarchical topic models relax the hypothesis of a single distribution over a corpus. By selecting the maximum similarity over different possible distributions, the actual number of senses assigned to each word can be approached.

6.3 Semantic Mappings

The following experiments correspond to the second model that we described in Chapter 5. Specifically, we examine the role of: (1) the algorithm used to define the mappings, (2) the number of words whose distributed representations serve as anchor points between our source and target spaces, (3) the polysemic degree of words used as anchor points.

6.3.1 Mapping methods

Figure 6.2 presents the results of our first experiment. Our goal is to evaluate the reliability of our semantic mappings using different techniques that are highly adopted by the literature. To that end, we use standard lists of semantic anchors consisting of $5,000$ frequent monosemous words for each topic and apply the three following mapping



(a) AvgSimC (b) MaxSimC

Figure 6.2: Performance comparison between different number of topics and different mapping algorithms, in terms of Spearman's correlation, using the (a) AvgSimC and (b) MaxSimC contextual metrics, for SCWS dataset. The dimensionality of semantic vector spaces is set to 300, and lists of 5,000 frequent monosemous words serve as *semantic anchors* for each topic. Baseline performance is also depicted.

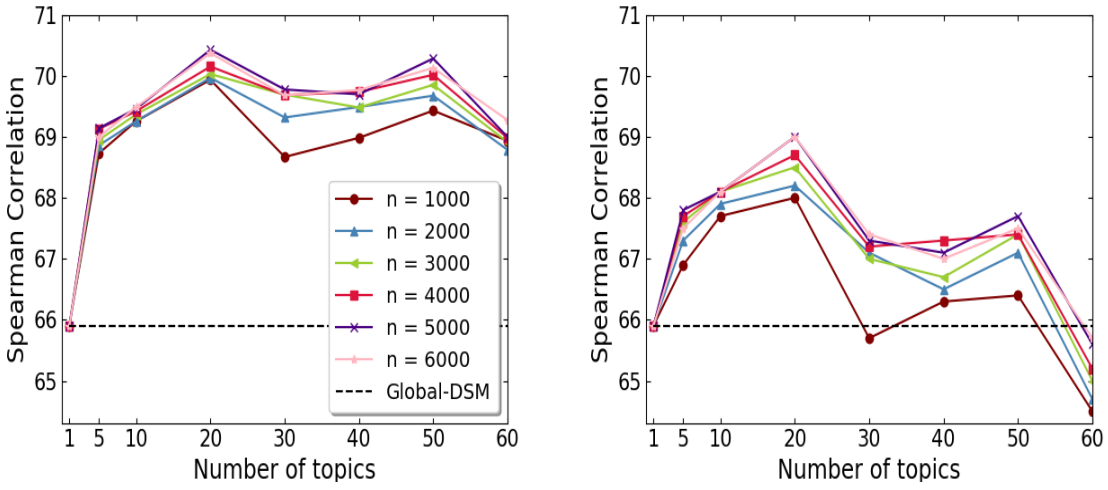
techniques: Linear transformation, Orthogonal Transformation, Canonical Correlation Analysis.

We observe that generally the orthogonal method yields slightly better performance when compared to the canonical method, while the linear method results in the poorest performance in terms of AvgSimC for all topic configurations. The differences in performance between the mapping methods is not so obvious in terms of MaxSimC. Furthermore, the depicted performances come to an agreement with literature results indicating that the orthogonal transformations lead to better mappings between semantic spaces of monolingual or bilingual data. For a thorough comparison of mapping methods we refer the reader to Artetxe et al. [2018]. For the rest of this thesis we opt for orthogonal transformations as our mapping technique.

6.3.2 Number of monosemous words

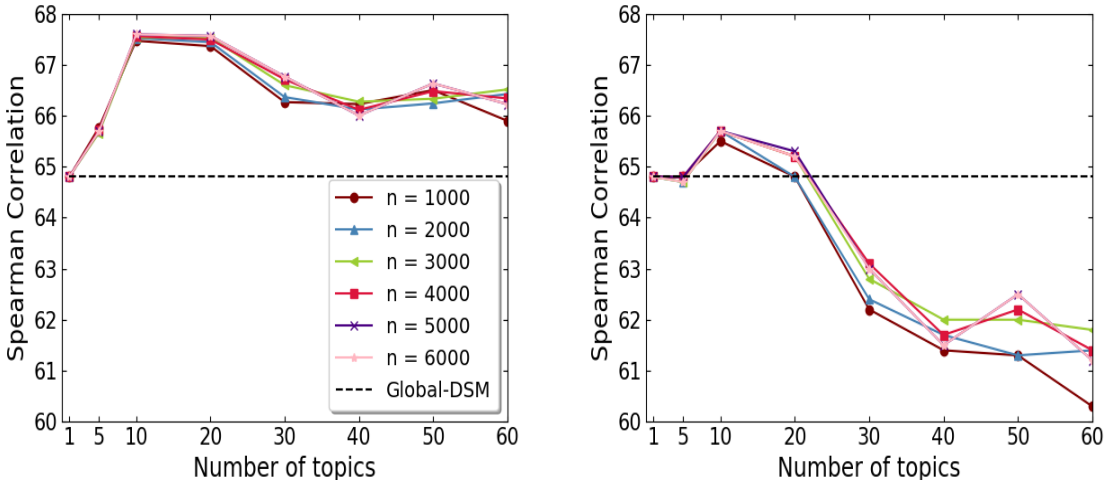
In Figure 6.3 we show the performance of our model using orthogonal mappings and different number of monosemous words n, when the dimensionality of the source and target spaces is set to 300. At first place we observe a clear relationship between the correlation of our predictions and the number of words used: A larger list of semantic anchors leads to the induction of more reliable semantic mappings that achieve peak performances. This behavior is expected, given the large number of parameters we attempt to learn. Recall that each transformation matrix $M_k \in R^{d \times d}$, where d is the dimensionality of our semantic spaces.

At second place we observe that the correlation of our model exhibits very slight



(a) AvgSimC (b) MaxSimC

Figure 6.3: Performance comparison between number of topics and different number of monosemous words n that serve as *semantic anchors* for the orthogonal mappings. Spearman's correlation is reported in terms of (a) AvgSimC and (b) MaxSimC contextual metrics, for SCWS dataset. The **dimensionality** of semantic vector spaces is set to **300**. Baseline performance is also depicted.



(a) AvgSimC (b) MaxSimC

Figure 6.4: Performance comparison between number of topics and different number of monosemous words n that serve as *semantic anchors* for the orthogonal mappings. Spearman's correlation is reported in terms of (a) AvgSimC and (b) MaxSimC contextual metrics, for SCWS dataset. The **dimensionality** of semantic vector spaces is set to **100**. Baseline performance is also depicted.

performance differences when the number of words is set to large values. Specifically, the performance is not improved when we use more than $5,000$ monosemous words as semantic anchors. This could be attributed to the fact that more rare monosemous words are introduced to our model as we increase the cardinality of our list of semantic anchors. As a result, even though we fit more data to our model, their reliability is impaired and do not result in higher accuracy of our predictions.

To circumvent this problem, we decrease the number of parameters that we have to learn, that is the number of components for each matrix M_k. In Figure 6.4 we present the results of our experiments using different number of monosemous words n, when the dimensionality of our semantic spaces is set to 100. Generally the performance of our model exhibits a more consistent behavior in terms of AvgSimC, when compared to the MaxSimC metric. As expected the differences between the performances of our model are not so obvious as the number of words increases, meaning that the cardinality of our anchor lists do not constitute a key parameter to our problem when we use distributed representations of lower dimensionality. However, although the previous problem is mitigated by decreasing the vectors' dimensions, another phenomenon arises: the general performance of our model is impaired and its difference from the baseline performance decreases for all topic configurations. This observation indicates that the complexity of our semantic spaces is better described using 300 dimensions instead of 100.

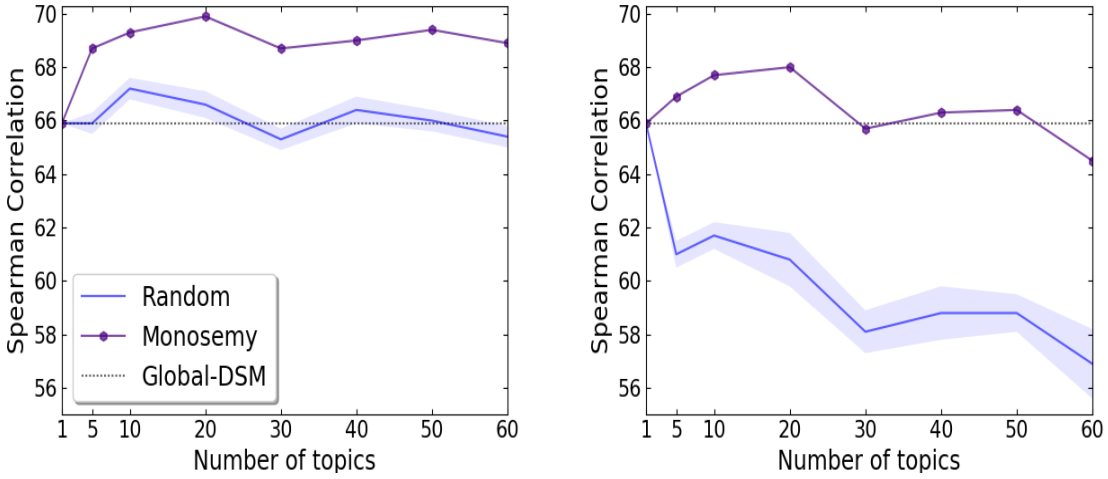
6.3.3 Semantic Anchors

To investigate the role of *semantic anchors* in determining the mappings between our source and target spaces we conduct two experiments. Following the best configurations obtained by the previous experiments we use the orthogonal mapping method and a list V_k consisting of n words for the alignemnt of the k-th topic. The two experiments are described as follows:

- For the first experiment, semantic anchors for each topic are extracted from the list of $|V_k|$ most frequent monosemous words (for that topic sub-corpus). The original (master) list of monosemous words is extracted from WordNet [Fellbaum, 1998]. In order to define monosemy we follow the terminology provided by WordNet, according to which: *a word is called monosemous when it has only one sense in every syntactic category.*

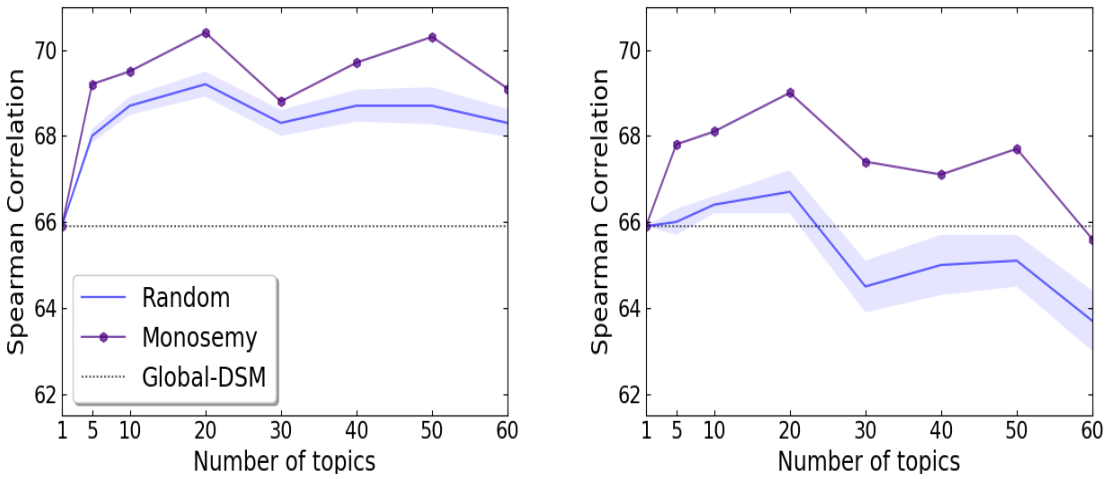
- For the second experiment, using the common set of words that are represented in both the source and the target spaces, we randomly sample $|V_k|$ of them as semantic anchors. We repeat this experiment 10 times, every time sampling a different list from the common set, and report average performance results, that correspond to the mean value and standard deviation of the individual experiments.

In Figure 6.5 and 6.6 we depict the performance of our model when we use lists of monosemous and lists of random words to determine the mappings between the source and target spaces. Specifically, in Figure 6.5 we show the results of the experiments when the cardinality of our lists is restricted to $1,000$ words, while Figure 6.6 presents the results of the same experiment using $5,000$ words. Furthermore, the horizontal lines ($K = 1$) show



(a) AvgSimC (b) MaxSimC

Figure 6.5: Performance comparison between different number of topics and mappings obtained using lists of monosemous and lists of random words to serve as *semantic anchors*. Spearman's correlation is reported in terms of (a) AvgSimC and (b) MaxSimC contextual metrics, for SCWS dataset. The **number** of anchor points for each topic is set to $\mathbf{1,000}$. Baseline performance is also depicted.



(a) AvgSimC (b) MaxSimC

Figure 6.6: Performance comparison between different number of topics and mappings obtained using lists of monosemous and lists of random words to serve as *semantic anchors*. Spearman's correlation is reported in terms of (a) AvgSimC and (b) MaxSimC contextual metrics, for SCWS dataset. The **number** of anchor points for each topic is set to $\mathbf{5,000}$. Baseline performance is also depicted.

the performance of the global-DSM, which is used as our *baseline* system and the shaded regions represent the deviation of the performance using different lists of random words.

For a fixed number of topics K, our model consistently yields higher performance when monosemous words are utilized as *semantic anchors* instead of randomly selected words, as depicted in both Figures 6.5 and 6.6. In Figures 6.5, 6.6 (b), we observe that the performance of MaxSimC falls bellow the *baseline* for $K > 1$ and $K > 20$, respectively, when randomly selected words serve as *semantic anchors*; while this is not the case when monosemous words are used as anchors for the most configurations. This result validates our hypothesis that the distributed representations of monosemous words constitute informative *semantic anchors* that determine the mappings between semantic vector spaces.

Furthermore, the above observation is more obvious when fewer words are used to learn the mappings, as illustrated in Figure 6.5. We note that the performance of our model fails to overcome the baseline performance when $n = 1,000$ random words are used for almost all topic configurations and metrics. On the other hand, the performance of our model clearly outperforms the baseline performance when $n = 1,000$ monosemous words are used for most of the configurations. This observation indicates that the usage of monosemous semantic anchors can achieve interesting results even if large amount of data is not available.

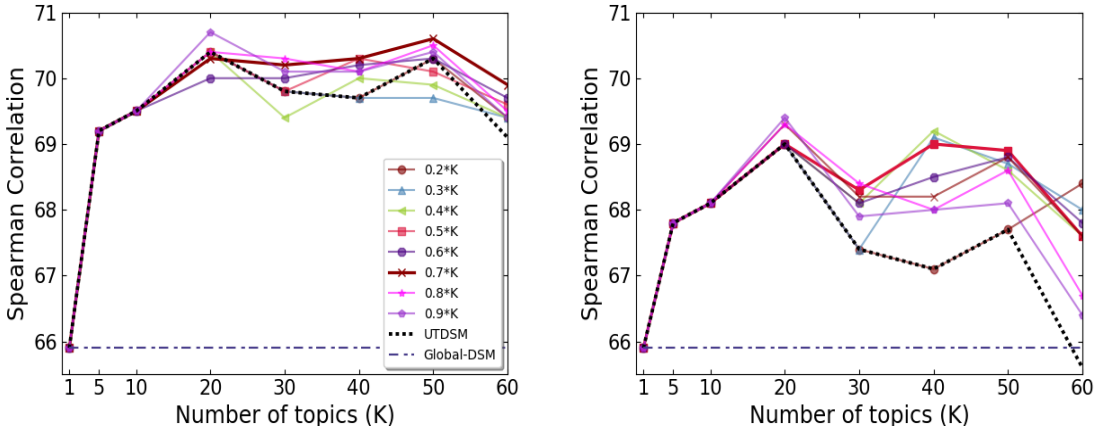
6.4 Smoothing of embeddings

As a further step we experiment with a smoothing technique using hierarchical clustering on the set of embeddings that correspond to a word in our unified multi-topic DSM. To perform hierarchical clustering of topic embeddings we use the Scikit-learn implementation of agglomerative clustering [Pedregosa et al., 2011]. The cardinality of the set consisting of *smoothed* topic embeddings is defined as a percentage of the number of topics K. Specifically, we experiment with $N = \max\{10, pK\}$ clusters, with $p \in (0, 1]$ (note that when $p = 1$ the smoothing technique is not applied). The possible values of N is restricted to be larger than 10, as we assume that the smoothing technique should be applied to noisy representations that are introduced to our model when K is large. Furthermore, we experiment with different linkage criteria that are used to define the distance between clusters, while the cosine distance[1] is used to define distances between vectors.

In Figure 6.7 we show the impact of smoothing using different linkage criteria and different number of clusters as a function of the number of topics. We experimentally note that when we use either the complete linkage or the average linkage criterion our results do not differ substantially. However, the usage of complete linkage achieves slightly better performance. On the other hand, we note that when using the single linkage criterion our performance exhibits higher variations depending on the number of clusters N. This observation could be attributed to the fact that the smallest distance between vectors is the only value taken into account. As a result, several clusters are joined together simply because one of their elements is within close proximity of a vector from a separate cluster, and thus the cluster solution is negatively impacted.

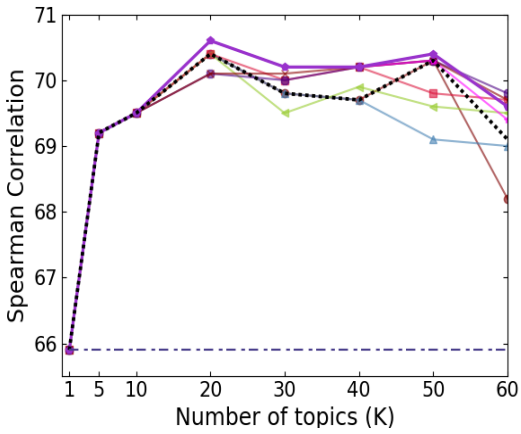
Furthermore, peak performances are achieved using large values of the parameter p in

[1] We use this distance metric to compare vector representations of words throughout the thesis.

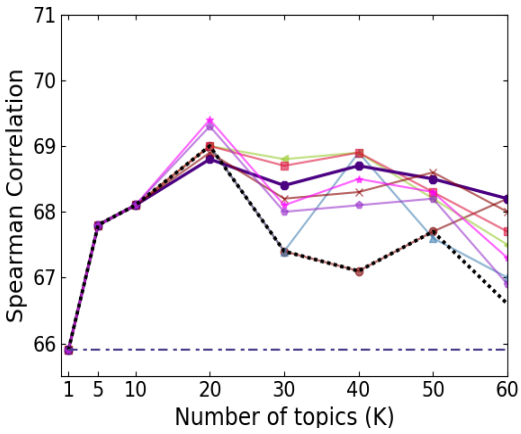


(a) Complete linkage-AvgSimC

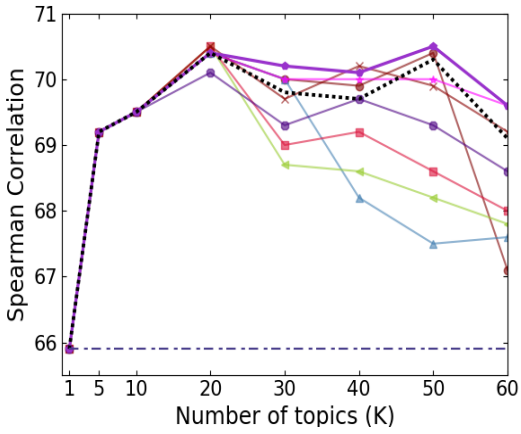
(b) Complete linkage-MaxSimC



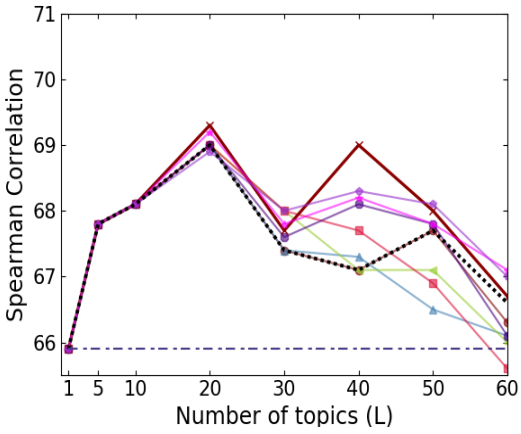
(c) Average linkage-AvgSimC



(d) Average linkage-MaxSimC



(e) Single linkage-AvgSimC



(f) Single linkage-MaxSimC

Figure 6.7: Performance comparison for different smoothing parameters and linkage criteria as a function of number of topics. Spearman's correlation is reported in terms of AvgSimC and MaxSimC metrics, for SCWS dataset. Lists of $5,000$ monosemous semantic anchors are used for the mappings. The dimensionality of semantic vector spaces is set to 300. Baseline performance is also depicted.

terms of AvgSimC and MaxSimC, as observed on the right and left plots of Figure 6.7. This is an expected outcome, given that small values of p lead to the creation of fewer vectors on the final set of smoothed embeddings. Generally, the best value of p depends on both the linkage criterion and the contextual metric we use. However, the differences between our model performances for two consecutive choices of the parameter p are not so significant.

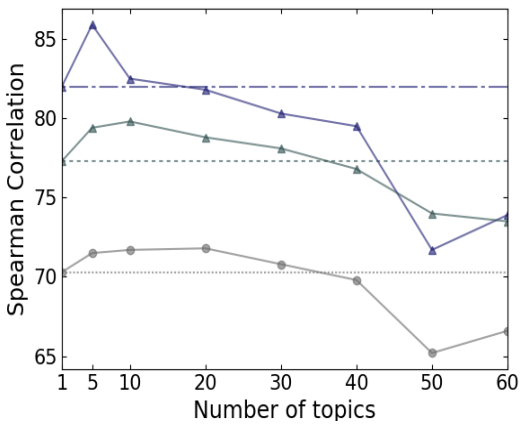
Generally, we observe that without utilizing the smoothing technique the performance of the proposed approach highly depends on the number of topics K. We attribute this behavior to the noisy topic representations that might be introduced to our model as the number of topics increases as well as to sparse training data (fewer data are assigned to each TDSM when K is large). However, the effect of this noise is reduced and the robustness of our model is recovered, when we make use of the smoothing technique.

6.5 Semanitc Similarity Results

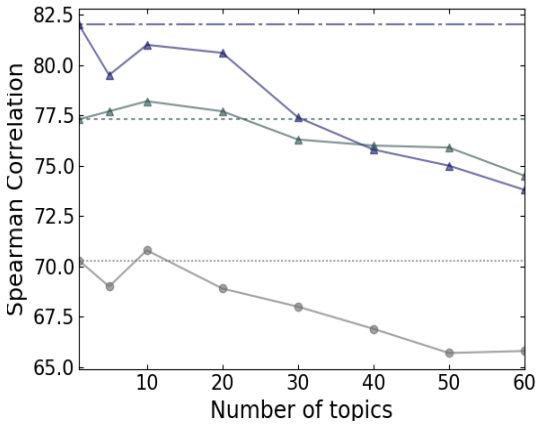
In this section we evaluate the performance of our second model on non-contextual datasets, using three metrics highly adopted in the literature. Figure 6.8 plots the performance of our proposed approach as a function of the number of topics K. The corresponding *baseline* performances ($K = 1$) are also plotted for each dataset. For all three datasets and metrics, we notice a clear relationship between the number of topics and performance. Smaller numbers of topics achieve peak performance, while larger numbers of topics seem to impair the quality of our predictions. We attribute this behavior to the increased topic-specificity of the "topic-specific" representations introduced to our system as the number of topics increases (more rare meanings of words are encoded in them). Analytically, for each metric we observe that:

- MaxSim: The best performances are achieved using 5, 5 and 10 topics for RG, WS-353 and MEN datasets, respectively. Correlation on the two latter datasets exhibits nearly the same behavior, as it fails to overcome the corresponding baselines performance for $K > 30$, which is not the case for RG.
- AvgSim: Our model fails to overcome the baseline performance for almost all topic configurations and datasets. This could be attributed to the fact that both noisy and well-positioned vector representations of a word contribute equally to the formation of the semantic similarities.
- AvgSimW: For all three datasets our model outperforms the corresponding *baseline* performances by an average of 1.5% for every $K > 1$. The best performances of our model are achieved using 5, 5 and 10 topics for the RG, WS-353 and MEN datasets, respectively.

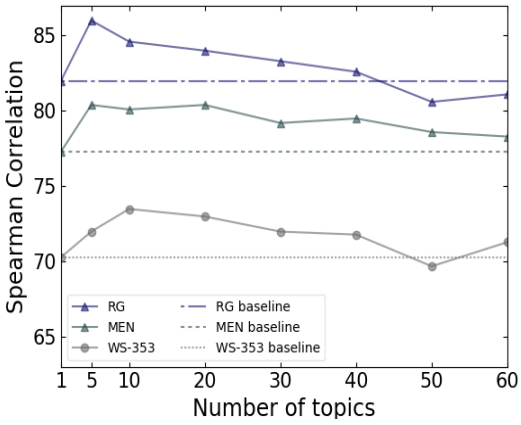
Generally, the MEN dataset exhibits the most robust performance; this is of significant importance given that it constitutes the most reliable dataset among the three non-contextual benchmarks that we use. We base this assumption on the experimental outcomes of Batchkarov et al. [2016] who note that the small size of the RG and WS-353 makes it difficult to reliably differentiate between models.



(a) MaxSim



(b) AvgSim



(c) AvgSimW

Figure 6.8: Performance as a function of the number of topics for the non-contextual datasets, using the (a) MaxSim, (b) AvgSim and (c) AvgSimW metrics. The three dot lines represent the corresponding baselines for each dataset.

6.6 Literature Comparison

In Table 6.2 we compare our two models with state-of-the-art systems that utilize multiple embeddings per word. The reported performances correspond to the best metrics and predictive configurations for each dataset. Specifically the performance reported for the TDSMs system on non-contextual datasets correspond to the MaxSim metric, while the UTDSM performance on the same dataset uses the AvgSimW metric.[2] Typically, supervised approaches that utilize external sense inventories and knowledge bases achieve higher performances in both contextual and non-contextual datasets.

Our first proposed approach (Mixture of topic-based models (TDSMs)) achieves for the MaxSimC scheme state-of-the-art performance (68.3), regarding the SCWS dataset. For the AvgSimC scheme, the proposed approach achieves 70.2 correlation being close to the

[2]The performance of our systems for the AvgSim metric is not reported in Table 6.2.

top performing systems (70.8 and 71.5, respectively). The fusion model further improves the performance of the TDSMs model for the AvgSimC metric (70.5). Concerning the out-of-context datasets, the Linear Regression model (TDSMs-LR) achieves state-of-the-art performance (83.8) for MEN dataset exceeding all models proposed in the literature. The same approach ranks fourth (72.7) compared to the top performing models, regarding WS-353.

Our second model (Unified multi-Topic Distributional Semantic Model (UTDSM)) improves on the reported results of the our first approach (when it is compared to the unsupervised mixture schemes), indicating that the contribution of the semantic relationships of words that reside in distinct domains plays a role in the similarity judgment. Furthermore, our model outperforms previous supervised approaches in almost all datasets, and generally achieves results comparable to the best predictive systems. Concerning the contextual semantic similarity task, we report new state-of-the-art performance in terms of MaxSimC (equal to 69.2).

	Approach	SCWS		MEN	WS-353	RG
		MaxSimC	AvgSimC			
Supervised	Chen et al. [2014]	-	68.9	-	-	-
	Iacobacci et al. [2015]	58.9	-	80.5	**77.9**	89.4
	Rothe and Schütze [2015]	-	69.8	-	-	-
	Pilehvar and Collier [2016]	-	**71.5**	78.6	-	**89.6**
Unsupervised	Huang et al. [2012]	-	65.7	-	71.3	-
	Neelakantan et al. [2014]	-	69.3	-	70.9	-
	Tian et al. [2014]	63.6	65.4	-	-	-
	Chen et al. [2015]	53.6	-	-	67.8	-
	Liu et al. [2015a]	67.9	69.5	-	-	-
	Liu et al. [2015b]	67.3	68.1	-	-	-
	Li and Jurafsky [2015]	-	69.7	-	-	-
	Amiri et al. [2016]	-	70.9		-	-
	Zheng et al. [2017]	-	69.9	-	-	-
	Nguyen et al. [2017]	66.9	66.7	76.4	72.4	-
	Lee and Chen [2017]	67.9	68.7	-	-	-
	Guo et al. [2018]	68.2	69.3	-	-	-
	Global-DSM	65.9	65.9	77.3	70.3	82.0
	TDSMs	67.8	70.2	80.0	72.2	-
	TDSMs-LR	-	-	**83.8**	72.2	-
	TDSMs-Fuse	67.4	70.5	-	-	-
	UTDSM	69.0	70.4	80.5	73.5	86.0
	UTDSM+smoothing	**69.2**	70.6	80.4	73.5	86.0

Table 6.2: Performance comparison between different state-of-the-art approaches for contextual and non-contextual datasets, in terms of Spearman's correlation. The results presented for two proposed approaches correspond to our best predictive configurations, while Global-DSM corresponds to our baseline system.

6.7 Visualizations & Examples

In this section we carry out a cross-domain semantic analysis to detect the variations of a word's meaning in different topic domains. To that end, we use a list of known polysemous words and measure the semantic similarity between different topic representations of the same ambiguous word. The ultimate goal of this analysis is to validate that our model captures known thematic variations in semantics of polysemous words.

Word	Topic Words	Meaning	Similarity
python	garden, plant, fish, bird, animal software, forum, download, windows, web	snake programming language	0.27
page	definition, dictionary, english, meaning, encyclopedia software, forum, download, windows, web	sheet computing	0.65
bank	loan, tax, cash, bank, insurance boat, marine, ship, sailing, yacht	financial institution slope	0.47
drug	health, medical, cancer, treatment, disease drug, health, marijuana, alcohol, effects	medicine illegal substance	0.61
power	news, nuclear, japan, energy, power math, mathematics, theory, university, analysis	energy math operation	0.50
apple	software, forum, download, windows, web recipe, food, cooking, chicken, wine	IT fruit	0.30
mouse	garden, plant, fish, bird, animal software, forum, download, windows, web	rodent device	0.48
window	forum, download, software, windows, web car, parts, sale, auto, equipment	computers glass	0.43
nursery	garden, plant, tree, flower, gardening university, school, college, education, program	plants preschool	0.46
history	university, school, college, education, program war, history, news, american, military	course past	0.68
act	law, court, legal, tax, state music, guitar, piano, dance, theatre	law performance	0.39
rock	mountain, river, park, road, trail music, guitar, piano, dance, theatre	stone music	0.43
house	mountain, river, park, road, trail music, guitar, piano, dance, theatre	dwelling music	0.57

Table 6.3: Examples of polysemous words and the change of meaning between different topic domains. First column lists the example target words. Second column includes the most probable words of the topic domains these words are assigned to. Each row corresponds to a different topic domain. Third column infers the specific meaning of the target word in the corresponding topic domain. The last column corresponds to the cosine similarity between the two topic representations of the target word.

Table 6.3 includes examples of our analysis. The most probable words of the topics (second column) give an intuitive sense of their major contexts. For example, we observe that the word *python* shifts from meaning "snake" in a topic about animals and nature, to referring to a "programming language" under a topic about computers. Word *drug* is

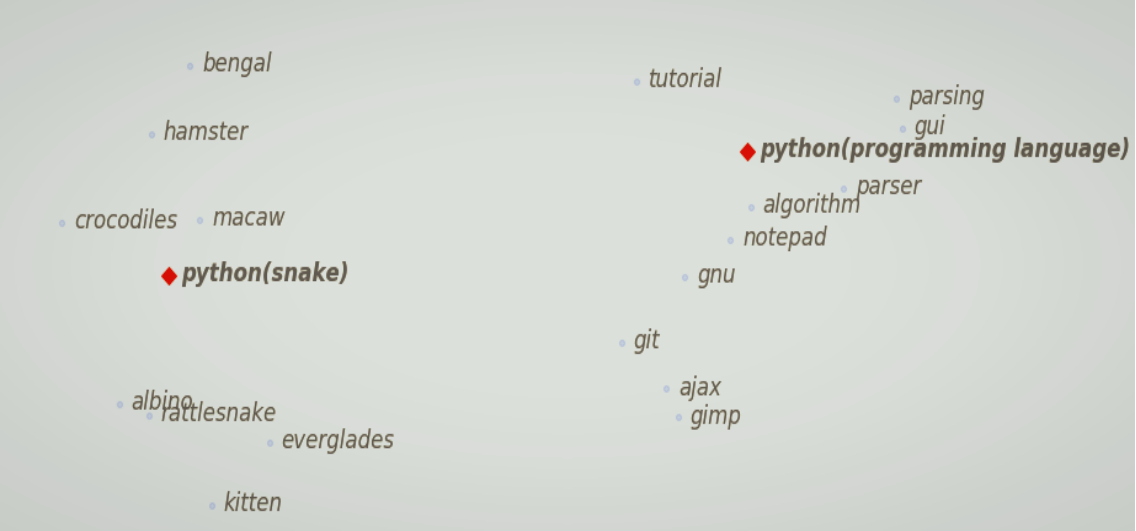
mostly related to "medication" in a broad medical domain; it experiences though a slight shift from this meaning when it resides in a topic about "illegal substances". The highly polysemous word *act* shifts from meaning "statute" to meaning "performance" under the corresponding law and art topics. In a thematic domain about music the word *rock* refers to a "music style" while in a more broad context about nature it refers to "stone". Finally, the word *nursery* corresponds to a "childcare facility" in a topic about education, whereas its meaning changes to "seedbed" in a topic about plants.

Moreover, in Figure 6.9 we visualize the latent semantic space of the monosemic neighborhoods for the two discriminative senses of words *python*, *nursery*, *drug*, *page*, *apple* and *act* using principal component analysis. Specifically, to visualize semantic change for a known polysemous word between two different topic domains we employed the following procedure, which relies on the Principal Component Analysis (PCA) [Pedregosa et al., 2011] as a subroutine:

- Find the word's k nearest monosemous neighbors over the two targeted topic domains.
- Compute the PCA embeddings of these words on the unified vector space.
- Visualize the two principal components for the union of the two monosemic neighborhoods of the target word along with its corresponding topic representations.

Note that we are limited to visualizing only the two monosemic neighborhoods of the target word, as we assume that monosemous words share consistent relationships in both spaces. The ultimate goal of this procedure is to investigate whether the two topic senses could be reliably differentiated by our model and not to reveal the actual set of their nearest neighbors (given that polysemous words could be also included in the sets).

By examining the local neighborhoods of the words subjected to analysis, we show that our model produces meaningful results that reflect the expected topic semantics of words.




evergreens
enrolment
kindergarten
shrubs
infant
succulents
grammar
preschool
perennials
nursery(plants)
montessori
camellias
prep
nursery(preschool)
website
ornamental
zoo
district
growers
greenhouse
grower
greenhouses
karate
gymnastics


amphetamines
narcotics
hallucinogens
insulitherapies
drug(medicine)
mdma
heroin
vaccines
hormone
drug(illegal substance)
pharmaceuticals
meth
cocaine
vaccine
methamphetamine

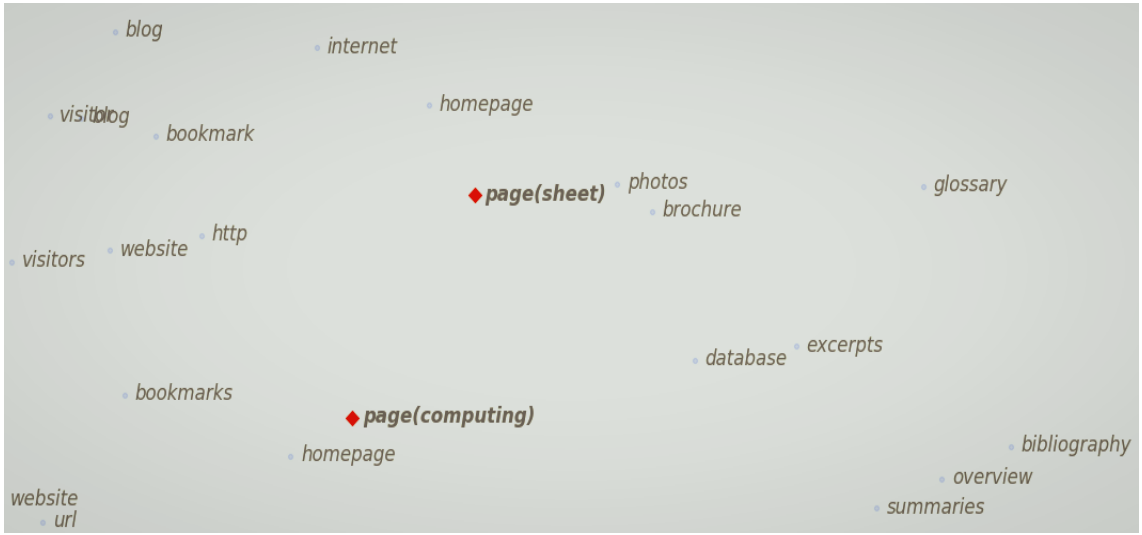

blog
internet
homepage
bookmark
page(sheet)
photos
glossary
brochure
http
website
visitors
excerpts
database
bookmarks
page(computing)
bibliography
homepage
overview
website
summaries
url

Figure 6.9: Examples of 2-dimensional projections of latent semantic spaces encoded in our unified vector space model, depicting the monosemic neighborhoods for two representations of the words *python*, *nursery*, *drug*, *page*, *apple* and *act* extracted form different thematic domains.

Chapter 7

Conclusion

7.1 Conclusions

In this thesis, a mixture model of topic-based DSMs was firstly proposed for the computation of semantic similarity between words. It was shown to outperform the baseline model (global-DSM). The good performance of the mixture model could be attributed to the creation of sub-corpora where the words of interest appear with topic-related senses. Afterwards, a more flexible approach was presented that extends the first approach, via mapping topic-based DSMs to a unified vector space. The resulting model consists of multiple distributed representations per word, reflecting their topic senses.

Overall, as a first step, we took advantage of the topic adaptation of semantic spaces via training topic-based DSMs. We assumed that these adaptations de-conflate the multiple senses of polysemous words into their principal topic senses, motivated by the assumption that polysemous words change their meanings according to the topic domains they reside in. At this point, a Mixture of topic-based DSMs could be utilized to compare the shared senses of two words under a specific topic domain. However, the above comparison is restricted to a topic-level. To overcome this restriction —instead of using a mixture of topic-based DSMs— we proposed that topic-based DSMs should be aligned to a reference vector space, enabling the comparison between word representations belonging to different topics. After that, we proposed a smoothing technique that smooths out noise from our model and recovers the robustness of our model's performance on semantic similarity tasks. To our knowledge this is the first time that mappings between semantic spaces are applied to the problem of learning multiple embeddings for polysemous words.

In more detail, the main goal of this thesis was to investigate the mappings between semantic spaces when monolingual data are used. Starting from a list of *semantic anchors* that determine the above mappings we examined the role of: (a) the mapping algorithms, (b) the number of semantic anchors, (c) the polysemic degree of words whose representations are used as semantic anchors. Specifically, our main motivation was that polysemous words change their meanings in different topic domains, and as a result their relative positions in different topic semantic spaces exhibit corresponding variations. However, we assumed that this is not the case when monosemous words are examined. As a consequence, we hypothesized that distributed representations of monosemous words constitute informative *semantic anchors* that determine the mappings between semantic vector

spaces, given that they share consistent semantic relationships in all topic domains. Our experimental results validated the above hypothesis.

Furthermore, we evaluated the performance of our models on contextual and in-isolation semantic similarity tasks. We showed that via using the best predictive configurations for the semantic mappings the unified multi-topic DSM outperformed the baseline system, indicating the superiority of using multiple representations per word instead of single representations. Finally, via examining the local neighborhoods of known polysemous that reside in different topic domains, we validated that our model captures meaningful variations of their diverse contextual semantics.

In summary, the mappings between semantic spaces seem to be very promising in the semantic analysis of polysemous words, as they establish a flexible and scalable approach that could be easily extended to a completely unsupervised model.

7.2 Future Work

The two models discussed in this thesis (Mixture of TDSMs, Unified multi-topic DSM) follow multi-step approaches that contain different standard implementations and parameters that need to be tuned in order to improve their performance. We propose the following ideas for each of the basic steps of our models as directions for future work:

Topic Modeling

- The Latent Dirichlet Allocation algorithm is utilized to induce the creation of topic-based DSMs. However, this topic model isn't able to model the correlations among topics because it attributes a single distribution over topics to each document. Therefore, a hierarchical topic model could be utilized to relax this restriction. Under this scheme, a mixture of hierarchical TDSMs could be constructed in an attempt to capture finer word-pair relationships by defining a dependency path across different levels of topic-based DSMs.

Semantic Mappings

- The semantic mappings between our source and target spaces utilize a list of monosemous words extracted from the sense inventory of WordNet. The proposed approach could be easily turned into a completely unsupervised model, if monosemous words were automatically identified from the model without using WordNet. In order to do so, a distance matrix that contains pairwise distances between the words of each topic could be extracted. Subsequently, pairs of monosemous words could be identified as those having small distance variances across different topic domains.

- The mapping techniques that are investigated in this thesis assume that there is a linear transformation between our source and target spaces. However, non-linear mappings between semantic spaces could be examined using deep neural network architectures.

Smoothing technique

- Concerning the smoothing technique a variety of other clustering algorithms could be used in order to create groups of topic embeddings (e.g., k-means). Furthermore, the number of clusters could be optimized by the clustering algorithm in order to reduce the number of parameters of our model. This could be implemented via optimizing a criterion, such as the within cluster sums of squares or the average silhouette.
- More advanced techniques could be also used for aggregating word embedding from different topics, using Gaussian mixture embeddings. This is motivated by the work of Chen et al. [2015] who argued that by representing words as points in an embedded space the rich relations between words cannot be reflected.

Bibliography

Hadi Amiri, Philip Resnik, Jordan Boyd-Graber, and Hal Daumé III. Learning text pair similarity with context-sensitive autoencoders. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1882–1892, 2016.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 238–247, 2014.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David J. Weir. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 7–12, 2016.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, pages 1137–1155, 2003.

David M. Blei. Introduction to Probabilistic Topic Modeling. *Communications of the ACM*, pages 77–84, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Resources (JAIR)*, 49:1–47, 2014.

Xinchi Chen, Xipeng Qiu, Jingxiang Jiang, and Xuanjing Huang. Gaussian mixture embeddings for multiple word prototypes. *arXiv preprint arXiv:1511.06246, 2015*, 2015.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, 2014.

Efstathia Christopoulou. Sentence-level sentiment analysis using topic modeling. 2016.

Fenia Christopoulou, Eleftheria Briakou, Elias Iosif, and Alexandros Potamianos. Mixture of topic-based distributional semantic and affective models. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 203–210, 2018.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, 16:76–83, 1989.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing. *Proceedings of the 25th International conference on Machine learning (ICML)*, pages 160–167, 2008.

Georgiana Dinu and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2014.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 426–471, 2014.

Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, pages 116–131, 2002.

J. R. Firth. A synopsis of linguistic theory. *Studies in Linguistic Analysis*, pages 1–32, 1957.

Yoav Goldberg. *Neural Network Methods in Natural Language Processing*, volume 37 of *Synthesis Lectures on Human Language Technologies*. 2017.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

Fenfei Guo, Mohit Iyyer, and Jordan Boyd-Graber. Inducing and embedding senses with scaled gumbel softmax. *arXiv preprint arXiv:1804.08077*, 2018.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proc. International Conference on Computational Linguistics (COLING)*, pages 497–507, 2014.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, 2016.

Zellig S. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954.

Matthew D. Hoffman, David M. Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, volume 1, pages 856–864, 2010.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882, 2012.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 95–105, 2015.

Elias Iosif and Alexandros Potamianos. Similarity computation using semantic networks created from web-harvested data. 21:49—79, 2015.

Stephen C. Johnson. *Hierarchical Clustering Schemes*. Psychometrika, 1967.

Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, pages 211–240, 1997.

Guang-He Lee and Yun-Nung Chen. Muse: Modularizing unsupervised sense embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMLP)*, pages 327–337, 2017.

Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? In *Proceedigns of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1732, 2015.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1284–1290. AAAI Press, 2015a.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Topical word embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, pages 2418–2424, 2015b.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Sophia Marmaridou. *Pragmatic Meaning and Cognition*. John Benjamins Publishing, 2000.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168*, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–751, 2013b.

George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, page 148, 1991.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, pages 235–244, 1990.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, 2014.

Dai Quoc Nguyen, Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. A mixture model for learning multi-sense word embeddings. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 121–127, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1680–1690, 2016.

Victor Prokhorov, Mohammad Taher Pilehvar, Dimitri Kartsaklis, Pietro Liò, and Nigel Collier. Learning rare word representations using semantic bridging. *CoRR*, abs/1707.07554, 2017.

Colorado Reed. Latent dirichlet allocation: Towards a deeper understanding, 2012.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

Joseph Reisinger and Raymond Mooney. Mixture Model with Sharing for Lexical Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1182, 2010.

Xin Rong. word2vec parameter learning explained. *arXiv:1411.2738*, 2014.

Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1793–1803, 2015.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, pages 627–633, 1965.

Peter H. Schönemann. *A generalized solution of the orthogonal procrustes problem.* 1966.

Amr Sharaf, Shi Feng, Khanh Nguyen, Kiante Brantley, and Hal Daumé III. The umd neural machine translation systems at wmt17 bandit learning task. In *Proceedings of the Second Conference on Machine Translation*, pages 667–673, 2017.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1556–1566, 2015.

Luchen Tan, Haotian Zhang, Charles L A Clarke, and Mark D Smucker. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 657–661, 2015.

Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 1353–1360, 2006.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings International Conference on Computational Linguistics (COLING)*, pages 151–160, 2014.

Zhaohui Wu and C Lee Giles. Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2188–2194, 2015.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 1006–1011, 2015.

Xiaoqing Zheng, Jiangtao Feng, Yi Chen, Haoyuan Peng, and Wenqing Zhang. Learning context-specific word/character embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3393–3399, 2017.