Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογιας Πληροφορικης και Υπολογιστων

Μελέτη Μεθόδων Υπολογισμού Σημασιολογικής Ομοιότητας Κειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρυσάνθη Σ. Γιαννούλη

Επιβλέπων : Γιώργος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογιας Πληροφορικης και Υπολογιστων

Μελέτη Μεθόδων Υπολογισμού Σημασιολογικής Ομοιότητας Κειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρυσάνθη Σ. Γιαννούλη

Επιβλέπων : Γιώργος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Μαρτίου 2021.

.....	.....	.....
Γιώργος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.	Ανδρέας-Γιώργος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2021

(*υπογραφή*)

.....

Χρυσάνθη Σ. Γιάννουλη

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

Περίληψη

Ο υπολογισμός της ομοιότητας μεταξύ κειμένων είναι μία σημαντική μέθοδος της ανάλυσης δεδομένων, η οποία μπορεί να χρησιμοποιηθεί περαιτέρω σε πολλές και διαφορετικές εφαρμογές της ΕΦΓ όπως είναι η ανάκτηση πληροφορίας, η ανάλυση συναισθημάτων, η μηχανική μετάφραση κτλ. Η παρούσα εργασία μελετά διάφορες μεθόδους για τον υπολογισμό της σημασιολογικής ομοιότητας κειμένων. Βασικό χαρακτηριστικό των μεθόδων αυτών, είναι η αναπαράσταση της φυσικής γλώσσας ενός κειμένου σε αριθμητική μορφή, με τρόπο που να συλλαμβάνεται πληροφορία για την σημασία του (embedding). Οι μέθοδοι βασίζονται στην μηχανική μάθηση για την δημιουργία των embeddings των λεκτικών όρων και εξετάζονται με κριτήριο την ικανότητα τους να εκτιμούν την ανθρώπινη κρίση για το νόημα του κειμένου.

Συγκεκριμένα, στην εργασία μελετώνται κωδικοποιητές για την δημιουργία embeddings λέξεων (word embeddings) και πως μπορούν να συνδυαστούν για να συλλάβουν το νόημα μιας πρότασης, καθώς και προ-εκπαιδευμένοι κωδικοποιητές για την δημιουργία embeddings προτάσεων (sentence embeddings). Για την πειραματική αξιολόγηση χρησιμοποιήθηκαν μικρά ειδησεογραφικά κείμενα, αντιπροσωπευτικά της αγγλικής γλώσσας, και ένα σετ ανθρώπινες μετρήσεις για την ομοιότητα των κειμένων. Τα αποτελέσματα των μετρήσεων δείχνουν πως η γενική επίδοση των μοντέλων να εκτιμούν την ανθρώπινη αντίληψη είναι καλή, χωρίς ιδιαίτερα κακές επιδόσεις. Αν και κάποια μοντέλα, πέτυχαν πολύ υψηλή απόδοση, η καλύτερη επίδοση επιτεύχθηκε στην περίπτωση που λάβαμε υπόψη το dataset μας, κάνοντας fine-tuning ένα μοντέλο σε αυτό.

Λέξεις Κλειδιά : ομοιότητα κειμένων, σημασιολογική ομοιότητα, ΕΦΓ, word embeddings, sentence embeddings , προ-εκπαιδευμένοι κωδικοποιητές , μέτρο ομοιότητας

Abstract

Computation of similarity between texts has been an important method of data analysis which can be further used in different NLP applications like information retrieval, sentiment analysis, machine translation. Generally, similarity between texts can be lexical or semantic. This thesis presents different approaches of modeling the semantic similarity between texts. The basic characteristic of these methods is the arithmetic representation of texts as vectors, in a way that captures information about text's meaning (embedding). The methods are based on machine learning for the creation of the embeddings and are assessed in terms of their ability to evaluate human judgments of similarity.

More specifically, in this thesis were studied models for creating word embeddings and different methods to combine these embeddings in order to capture the meaning of a sentence. Also, pre-trained sentence encoders were studied for creating sentence embeddings. The presented methods are evaluated experimentally using a small dataset of news and a dataset of the human ratings of the similarity of every pair of the texts. The evaluation results suggest that the proposed methods have generally good performance, without any model underperforming. However, the best performance measured was when we fine-tuned one of the models in the used dataset.

Keywords : document/text similarity, semantic similarity, NLP, word embeddings, sentence embeddings , pre-trained encoders , distributional semantics, similarity measure

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων

Κατάλογος Τμημάτων Κώδικα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Αυτό το κεφάλαιο περιλαμβάνει μια σύντομη εισαγωγή στα ζητήματα που πραγματεύεται η εργασία. Στην ενότητα 1.1 δίνονται κάποιοι ορισμοί που θα φανούν χρήσιμοι σε όλη την έκταση του κειμένου. Στην ενότητα 1.2 παρουσιάζεται η δυσκολία του προβλήματος της μελέτης της ομοιότητας κειμένων με ένα απλό παράδειγμα και αποσαφηνίζεται ο σκοπός της εργασίας. Στις ενότητες 1.3 και 1.4 παρουσιάζεται το εύρος και οι συνεισφορές της παρούσας εργασίας, αντίστοιχα. Τέλος, στην ενότητα 1.5 περιγράφεται η δομή του υπόλοιπου κειμένου.

1.1 Χρήσιμοι Ορισμοί

Σε αυτή την ενότητα καταγράφονται μερικοί ορισμοί εννοιών, οι οποίες χρησιμοποιούνται στο κείμενο και μπορεί να μην είναι οικείες στον αναγνώστη.

1. *Επεξεργασίας Φυσικής Γλώσσας / Natural Language Processing (ΕΦΓ / NLP)* είναι ένας διεπιστημονικός κλάδος της τεχνητής νοημοσύνης, της υπολογιστικής γλωσσολογίας και της επιστήμης της πληροφορικής, που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστων και των ανθρώπινων (φυσικών) γλωσσών. Πιο συγκεκριμένα, ασχολείται με το πως να προγραμματίσει κάποιος υπολογιστές για την επεξεργασία και ανάλυση μεγάλων ποσοτήτων φυσικής γλώσσας.

2. *Corpus* είναι μία συλλογή από αυθεντικά κείμενα οργανωμένα σε datasets. ‘Αυθεντικό’ στην περίπτωση αυτή, δηλώνει κείμενο γραμμένο από κάποιον στην μητρική του γλώσσα ή διάλεκτο. Ένα corpus μπορεί να αποτελείται από εφημερίδες, μυθιστορήματα μέχρι συνταγές και tweets.

3. *Μηχανική Μάθηση / Machine Learning* είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

4. *Νευρωνικό δίκτυο* ονομάζεται ένα κύκλωμα διασυνδεδεμένων μονάδων επεξεργασίας που ονομάζουμε Νευρώνες. Στους υπολογιστές είναι ένα υπολογιστικό μοντέλο που χρησιμοποιείται για την επίλυση κάποιου υπολογιστικών προβλημάτων.

5. *Embedding*, μαθηματικά ορίζεται ως η απεικόνιση ενός σετ σε ένα άλλο σετ. Στα πλαίσια αυτής της εργασίας, embedding είναι η απεικόνιση του νοήματος ενός λεκτικού όρου, σε ένα διάνυσμα πραγματικών αριθμών.

6. *Προ-εκπαιδευμένα Embeddings* είναι τα embeddings που μαθαίνονται σε ένα task και χρησιμοποιούνται για να λύσουν ένα άλλο παρόμοιο task.

7. *Εκπαίδευση Μοντέλου* είναι η διαδικασία καθορισμού των ιδανικών παραμέτρων ενός μοντέλου.

8. *Transfer Learning* είναι η ικανότητα του υπολογιστικού συστήματος να χρησιμοποιεί κάποια από τα concepts που μαθαίνει σε ένα task σε ένα άλλο διαφορετικό.

9. *Ground Truth* είναι πληροφορία που προέρχεται από άμεση παρατήρηση (π.χ. εμπειρικά στοιχεία) και όχι ως αποτέλεσμα συμπεράσματος.

10. *Fine tuning* είναι η διαδικασία κατά την οποία οι παράμετροι ενός μοντέλου προσαρμόζονται με ακρίβεια με σκοπό να ταιριάζουν με ορισμένες παρατηρήσεις.

11. *Baseline* είναι ένα απλό μοντέλο που παρέχει ικανοποιητικά αποτελέσματα για ένα πρόβλημα και δεν χρειάζεται ειδικές γνώσεις και χρόνο για να σχεδιαστεί. Χρησιμοποιείται ως σημείο αναφοράς για την σύγκριση του πόσο καλά ένα άλλο μοντέλο (συνήθως πιο περίπλοκο) αποδίδει.

1.2 Δυσκολία Μελέτης Ομοιότητας Κειμένων και Αποσαφήνιση Σκοπού Εργασίας

"*She had changed a lot since the last time we 'd seen her.* "

"*The legislature was instrumental in effecting changes to the benefit program.*"

Δοθέντων δύο μικρών κειμένων ή φράσεων, όπως παραπάνω, θέτουμε το ερώτημα εάν είναι όμοια. Για έναν άνθρωπο, που εκ φύσεως επεξεργάζεται εύκολα την φυσική γλώσσα και βγάζει συμπεράσματα γι' αυτή, η απάντηση στο ερώτημα είναι πολύ απλή. Το ίδιο όμως δεν ισχύει και για ένα υπολογιστικό σύστημα, και αυτό γιατί η φυσική γλώσσα στην πράξη είναι πολύ δύσκολη. *"...Είναι εξαιρετικά διφορούμενη... Επίσης, αλλάζει και εξελίσσεται συνεχώς. .... Συγχρόνως, ενώ οι άνθρωποι κάνουμε εξαιρετική*

χρήση της γλώσσας, είμαστε ... πολύ κακοί στην επίσημη κατανόηση και περιγραφή των κανόνων που κυβερνούν την γλώσσα."

— Σελίδα 1, Neural Network Methods in Natural Language Processing, 2017

Κάθε συμπέρασμα πάνω στην φυσική γλώσσα από ένα υπολογιστικό σύστημα είναι δύσκολο. Μάλιστα, η ομοιότητα κειμένων είναι ένα από τα ανοιχτά μεγάλα προβλήματα της τεχνητής νοημοσύνης τα τελευταία 50 χρόνια.

Όταν όμως μιλάμε για ομοιότητα, σε τι αναφερόμαστε; Γενικά, δύο κείμενα θεωρούνται όμοια εάν μοιάζουν 1. νοηματικά/σημασιολογικά ή 2. λεκτικά (surface closeness). Το πρώτο είδος ομοιότητας ονομάζεται σημασιολογική ομοιότητα (semantic similarity), και το δεύτερο λεκτική (lexical similarity). Π.χ. στις φράσεις "the cat ate the mouse" και "the mouse ate the cat food" , μπορούμε εύκολα να συμπεράνουμε πως μοιάζουν, καθώς έχουν 3 κοινές λέξεις, ωστόσο δεν φέρουν το ίδιο νόημα. Και αυτό είναι σημαντικό, γιατί αν και η λεκτική ομοιότητα δεν είναι τελείως αδιάφορη ως προς την πληροφορία που μας προσφέρει, η σημασιολογική ομοιότητα είναι αυτή που προσεγγίζει τον ανθρώπινο τρόπο αντίληψης της ομοιότητας.

Σκοπός της παρούσας εργασίας, είναι να μελετήσει τις σύγχρονες μεθόδους με τις οποίες ένα υπολογιστικό σύστημα μπορεί να εκτιμήσει την σημασιολογική ομοιότητα μεταξύ κειμένων και να τις αξιολογήσει με κριτήριο την αντίστοιχη ανθρώπινη εκτίμηση.

1.3 Εύρος της Εργασίας

Αυτή η εργασία επικεντρώνεται στην μελέτη μεθόδων για τον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ μικρών κειμένων. Μελετώνται τρόποι με τους οποίους μπορεί να αναγνωριστεί και αναπαρασταθεί το νόημα ενός κειμένου σε αριθμητική μορφή. Για τον σκοπό αυτό, μελετώνται δημοφιλή μοντέλα κωδικοποιητών για την δημιουργία embeddings λέξεων και προτάσεων. Κάθε μοντέλο αξιοποιείται και προσαρμόζεται ώστε να μπορεί να λειτουργεί για το μέγεθος του κειμένου που έχουμε. Τέλος, χρησιμοποιείται ένα μικρό dataset κειμένων, ώστε να δοκιμαστούν τα μοντέλα στην πράξη.

1.4 Συνεισφορές της Εργασίας

Η εργασία προσπαθεί να προσφέρει μία εμπεριστατωμένη ανάλυση των σύγχρονων μεθόδων που χρησιμοποιούνται για την μελέτη της σημασιολογικής ομοιότητας κειμένων, και παραθέτοντας πειραματικά τα αποτελέσματα των πιο απλών αλλά και πιο περίπλοκων μεθόδων προσφέρει μία γενική εικόνα για την απόδοση τους. Επίσης, στην εργασία παρουσιάζεται το συνολικό θεωρητικό υπόβαθρο της σημασιολογικής ομοιότητας.

1.5 Διάρθρωση Εργασίας

Στο Κεφάλαιο 2, χτίζεται το θεωρητικό υπόβαθρο που είναι αναγκαίο για την μελέτη μας. Είναι χωρισμένο σε τρία μέρη: 2.1 Αλγόριθμοι Σημασιολογικής Ομοιότητας 2.2 Word Embeddings 2.3 Sentence Embeddings και 2.4 Μέτρο Ομοιότητας. Στην συνέχεια, στο Κεφάλαιο 3, παρουσιάζονται οι μέθοδοι που θα μελετηθούν στην εργασία. Στο Κεφάλαιο 4, γίνεται η πειραματική αξιολόγηση των μεθόδων και ο σχολιασμός των πειραματικών αποτελεσμάτων. Επιπλέον, γίνεται fine-tuning μία εκ των μεθόδων και αξιολογούνται τα αποτελέσματα. Τέλος, στο Κεφάλαιο 5, καταγράφονται συμπεράσματα που προέκυψαν από την εκπόνηση της παρούσας εργασίας, καθώς και δυνατές επεκτάσεις έρευνας που μπορούν να πραγματοποιηθούν στο μέλλον.

ΚΕΦΑΛΑΙΟ 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό αναλύεται το θεωρητικό υπόβαθρο της ομοιότητας κειμένων. Αρχικά, στην ενότητα 2.1 παρουσιάζονται οι δύο βασικές κατηγορίες αλγορίθμων σημασιολογικής ομοιότητας, οι knowledge-based και corpus-based αλγόριθμοι. Στην ενότητα 2.2 παρουσιάζονται τα word embeddings και τρία βασικά μοντέλα για την δημιουργία τους. Στην ενότητα 2.3, αναφέρεται η βασική θεωρία για τα sentence embeddings. Τέλος, στην ενότητα 2.4 εξετάζονται τα μέτρα ομοιότητας.

2.1 Αλγόριθμοι Σημασιολογικής Ομοιότητας

Αρχικά, για την μελέτη της σημασιολογικής ομοιότητας χρειάζεται να εξετάσουμε την σημασία της ίδιας της έννοιας.

Σε πρώτο επίπεδο, η *'ομοιότητα'* ανάμεσα σε δύο αντικείμενα είναι ένα αριθμητικό μέτρο του πόσο μοιάζουν και μπορεί να λειτουργεί ως μια οργανωτική αρχή με την οποία (τα άτομα) ταξινομούν αντικείμενα, σχηματίζουν έννοιες, κάνουν γενικεύσεις. Σε δεύτερο επίπεδο, '*σημασιολογία*' (που προέρχεται από την αρχαία Ελληνική λέξη : "*σημαντικός"*) είναι η μελέτη της σημασίας. μιας λέξης.

Συνολικά, Σημασιολογική Ομοιότητα (Σ.Ο.) είναι το μέτρο της εννοιολογικής απόστασης μεταξύ δύο λεκτικών όρων (λέξεις, φράσεις, κείμενα), με βάση την αντιστοιχία της σημασίας τους. Διαφορετικά, η Σ.Ο. απαντάει στο πόσο όμοιοι είναι δύο λεκτικοί όροι, ως κάποια συνάρτηση απόστασης, όπου η ιδέα της απόστασης ανάμεσα στους όρους βασίζεται στην ομοιότητα της σημασίας τους και γίνεται χρήση κάποιας πηγής πληροφοριών. Χαρακτηριστικά παραδείγματα πηγών πληροφοριών που χρησιμοποιούνται, είναι τα σημασιολογικά δίκτυα, οι μηχανές αναζήτησεις, και η Wikipedia.

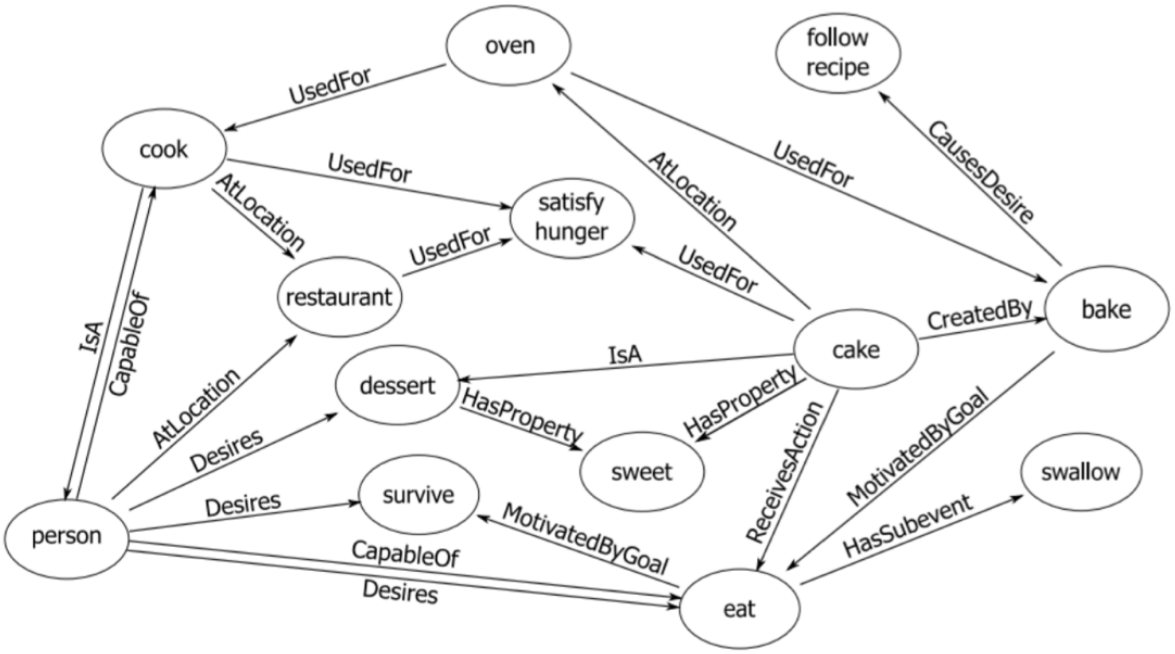
Ο καλός ορισμός της σημασιολογικής ομοιότητας απασχολεί τους ερευνητές εδώ και αρκετές δεκαετίες. Στην βιβλιογραφία, συναντάμε έναν πολύ μεγάλο αριθμό από προτεινόμενες μεθόδους. Ανάμεσα τους, μπορούμε να διακρίνουμε δύο βασικές κατηγορίες προσέγγισης της Σ.Ο. με βάση το είδος της πηγής που χρησιμοποιείται : τις corpus-based προσεγγίσεις και τις knowledge-based.

Αυτή η ενότητα αναλύει εν συντομία και τις δύο προσεγγίσεις.

2.1.1. Knowledge-based Αλγόριθμοι

Οι Knowledge-based αλγόριθμοι ορίζουν την σημασιολογική ομοιότητα χρησιμοποιώντας πληροφορίες από σημασιολογικά δίκτυα ή γενικά βάσεις γνώσης.

Ένα σημασιολογικό δίκτυο είναι μία βάση γνώσης που αναπαριστά τις σημασιολογικές σχέσεις μεταξύ εννοιών σε μορφή δικτύου. Τα σημασιολογικά δίκτυα συνήθως έχουν κόμβους που αντιπροσωπεύουν έννοιες και συνδέσεις που αντιπροσωπεύουν τις σημασιολογικές σχέσεις μεταξύ αυτών των εννοιών.



Σχήμα 1. Παράδειγμα Σημασιολογικού Δικτύου

Ενδεικτικά, στο WordNet, ένα από τα πιο δημοφιλή σημασιολογικά δίκτυα, οι λέξεις ενώνονται με τις σημασιολογικές σχέσεις : συνώνυμα, υποωνύμια, και μερωνύμια. Τα ουσιαστικά, ρήματα, επιρρήματα και επίθετα ομαδοποιούνται σε σετ γνωστικών συνωνύμων (cognitive synonyms - synset), που το καθένα εκφράζει μία διαφορετική έννοια.

Γενικά, τα σημασιολογικά δίκτυα καθώς προσφέρουν μία δομημένα αναπαράσταση των εννοιών, τα σημασιολογικά μέτρα που βασίζονται σε αυτά λαμβάνουν το πραγματικό νόημα των όρων, και άρα είναι χωρίς αμφισημία.Ωστόσο, καθώς “οι κανόνες της γλώσσας βασίζονται στην χρήση της και αλλαγές συμβαίνουν κάθε φορά που οι χρήστες την χρησιμοποιούν για να επικοινωνήσουν μεταξύ τους”, τα σημασιολογικά δίκτυα χρειάζεται να ενημερώνονται συχνά προκειμένου να διαμορφώνουν την δομή και την πληροφορία τους. Μια τέτοια διαδικασία, κοστίζει τόσο σε πόρους, όσο και σε χρόνο.

Υπάρχει πληθώρα μεθόδων που έχουν αναπτυχθεί για το προσδιορισμό της Σ.Ο. δύο λέξεων με την χρήση ενός σημασιολογικού δικτύου. Όλα αυτά τα μέτρα θεωρούν ως είσοδο ένα ζεύγος από έννοιες, και επιστρέφουν μία τιμή που δείχνει την σημασιολογική σχέση τους.Διαισθητικά, ένα είδος μέτρου της σημασιολογικής ομοιότητας σε ένα σημασιολογικό δίκτυο, είναι η απόσταση μεταξύ των εννοιών· δύο λέξεις θεωρούνται πιο όμοιες εάν βρίσκονται πιο κοντά στο δεδομένο δίκτυο. Πιο αναλυτικά, στην βιβλιογραφία μπορούμε να βρούμε τρεις βασικές διαφορετικές προσεγγίσεις για τον προσδιορισμό της Σ.Ο. σε ένα σημασιολογικό δίκτυο, οι οποίες είναι οι εξής :

- '*path-based*' : εκτιμούν την σημασιολογική ομοιότητα μετρώντας τον αριθμό των κόμβων/ακμών που χωρίζουν δύο έννοιες σε ένα δεδομένο σημασιολογικό δίκτυο

- '*information theoretic*' : αξιοποιούν την έννοια του '*information content*'. Information content, IC, ή αλλιώς 'πληροφορία περιεχομένου', είναι ένα μέτρο του πόση πληροφορία δίνει μία έννοια. Ένας τυπικός ορισμός του IC είναι ως : IC(a) = -log p(a) , όπου a ένας όρος σε ένα δεδομένο σημασιολογικό δίκτυο και p(a) η πιθανότητα να συναντήσουμε τον όρο 'a' σε ένα δεδομένο corpus.

  Στις 'information theoretic' προσεγγίσεις, η ομοιότητα εκτιμάται ως μία συνάρτηση του I.C. που δύο έννοιες έχουν κοινό, σε ένα δεδομένο σημασιολογικό δίκτυο.

- '*feature-based*' : Εδώ, οι έννοιες αναπαρίστανται ως σετ χαρακτηριστικών. Τα χαρακτηριστικά μιας έννοιας περιγράφονται από λέξεις, και συνήθως αφορούν το σύνολο των εννοιών που υπάγονται σε αυτήν. Η ομοιότητα επομένως, μπορεί να οριστεί ως μία συνάρτηση των κοινών και ξεχωριστών χαρακτηριστικών δύο εννοιών. Διαισθητικά, όσο πιο πολλές λέξεις αλληλοεπικαλύπτονται τόσο πιο όμοιες είναι οι δύο έννοιες.

  Αυτός ο τρόπος προσέγγισης κατάγεται από το 'feature-model' που προτάθηκε από τον Tversky το 1977. Μάλιστα, ο Tversky ήταν ο πρώτος που με το 'feature-model' διατύπωσε ένα πλαίσιο ορισμού της σημασιολογικής ομοιότητας.

Οι παραπάνω knowledge-based μέθοδοι μπορούν να επεκταθούν και για το μέτρο της ομοιότητας φράσεων ορίζοντας κανόνες συνάθροισης.

2.1.2 Corpus-based Αλγόριθμοι

Ο δεύτερος βασικός τρόπος προσέγγισης της Σ.Ο. είναι τα corpus-based semantics, ή αλλιώς statistical ή distributional semantics. Οι corpus-based αλγόριθμοι βασίζονται σε συλλογές κειμένων (corpus) τις οποίες επεξεργάζονται, παλιότερα με στατιστικά μέσα και σήμερα με μοντέλα νευρωνικών δικτύων, για να εξάγουν πληροφορία που θα αφορά το νόημα. Το νόημα μπορεί να αναφέρεται σε λέξεις, αλλά ακόμη και σε προτάσεις.

"....η κατανόηση της φυσικής γλώσσας απαιτεί μεγάλες ποσότητες γνώσης για την μορφολογία, την σύνταξη, την σημασιολογία και την πραγματολογία, καθώς και γενική γνώση για τον κόσμο. Η απόκτηση και αποκωδικοποίηση όλης αυτής την γνώσης είναι ένα από τα θεμελιώδη εμπόδια για την την ανάπτυξη αποτελεσματικών και ισχυρών γλωσσικών συστημάτων. Όπως τα στατιστικά μέσα ... η μηχανική μάθηση μπορεί να υποσχεθεί την αυτόματη απόκτηση αυτής της γνώσης από annotated ή unannotated συλλογές κειμένων "

—Σελίδα 377, The Oxford Handbook of Computational Linguistics, 2005.

Για την εξαγωγή της πληροφορίας που θα αφορά το νόημα, χρησιμοποιείται συνήθως το συγκείμενο πλαίσιο (context). Τα distributional semantics βασίζονται στην θεωρία του γλωσσολόγου J.R. Firth : "*You shall know a word by the company it keeps*" , (1957) και της διανεμητικής υπόθεσης (distributional hypothesis) : δύο λέξεις έχουν παρόμοιο νόημα, εάν χρησιμοποιούνται σε παρόμοια συγκείμενα πλαίσια, ή με άλλα λόγια, εάν έχουν όμοια ή παρόμοια διανομή (*distribution*) (1954). Π.χ. Τόσο το "σκυλί" όσο και η "γάτα" εμφανίζονται συχνά κοντά στις λέξεις "κτηνίατρος" , "κατοικίδιο", "ζωοτροφή", και αυτή η συνεμφάνιση αποδεικνύεται πως αποτελεί ένα επαρκές στοιχείο για την ομοιότητα τους.

Ο συνήθης τρόπος, που αναπαρίσταται η αναχθείσα πληροφορία για το νόημα μιας λέξης/ενός κειμένου, είναι αριθμητικά μέσω διανυσμάτων. Με ένα μέτρο ομοιότητας έτσι, μπορούμε να υπολογίσουμε την ομοιότητα των παραγόμενων διανυσμάτων, και κατ' έκταση, να έχουμε και την ομοιότητα των αντίστοιχων λέξεων/κειμένων.

Στην παρούσα εργασία, εξετάζονται μοντέλα που βασίζονται σε νευρωνικά δίκτυα για τις διανυσματικές αναπαραστάσεις του νοήματος λέξεων (word embeddings) και προτάσεων (sentence embeddings).

Δεδομένου ότι η απόδοση των αλγορίθμων εξαρτάται σε μεγάλο βαθμό, από το corpus που χρησιμοποιείται, η δημιουργία ενός αποδοτικού corpus είναι υψίστης σημασίας. Ωστόσο, ένα "ιδανικό corpus" ,στο βαθμό της γνώσης που αποκτήθηκε στα πλαίσια αυτής της εργασίας, δεν έχει ακόμη οριστεί από τους ερευνητές.

2.2 Word Embeddings

Ως 'Word Embeddings' ορίζουμε τις διανυσματικές αναπαραστάσεις λέξεων πραγματικών αριθμών, που κωδικοποιούν το νόημα των λέξεων, με τρόπο που λέξεις που βρίσκονται πιο κοντά στο διανυσματικό χώρο αναμένεται να μοιάζουν στην σημασία (meaningful space).

'man' → [0.33 , 0.46 , 0.71, 0.15 , 0.78]

Παράδειγμα ενδεικτικής αναπαράσταση της λέξης 'man' στο 5-διάστατο χώρο.
Δοθέντος ότι η λέξη 'woman', είναι νοηματικά συγγενική με την λέξη 'man', η αναπαράσταση της θα πρέπει να είναι τέτοια ώστε να βρίσκεται κοντά με την λέξη 'man' στο διανυσματικό χώρο.

Οι αναπαραστάσεις λαμβάνονται συνήθως με την εκπαίδευση ενός νευρωνικού δικτύου σε ένα μεγάλο corpus. Το μοντέλο εκπαίδευτε να κάνει προβλέψεις με βάση τις λέξεις και το συγκείμενο τους πλαίσιο (context). Τα βάρη που μαθαίνονται κατά την εκπαίδευση, χρησιμοποιούνται για την αναπαράσταση μιας λέξης.

Στην βιβλιογραφία, ο όρος 'word embeddings' χρησιμοποιείται ισοδύναμα με τον ορο 'distributed vector representations'.

Τα βασικά πλεονεκτήματα των word embeddings είναι τα εξής:

- Τα βάρη των διανυσμάτων έχουν συγκριτική τιμή, δηλαδή παρόμοιες λέξεις έχουν παρόμοια βάρη, και άρα και κοντινή θέση στον διανυσματικό χώρο.
- Μπορούμε να αναπαραστήσουμε μια λέξη με σχετικά μικρό αριθμό παραμέτρων.
- Παρέχουν πλούσιες αναπαραστάσεις για τις λέξεις, πολύ πιο ισχυρές από τη χρήση των ίδιων των λέξεων, καθώς μπορούν να φανερώσουν χαρακτηριστικά των λέξεων που δεν είναι εμφανή στα αρχικά δεδομένα
- Έχουν την ικανότητα γενίκευσης λόγω των κοινών χαρακτηριστικών μεταξύ των εννοιών, και αυτό τα κάνει χρήσιμα σε πολλές εφαρμογές της ΕΦΓ. Συνήθως, χρησιμοποιούνται ως γενικού σκοπού χαρακτηριστικά για λέξεις σε διάφορα προβλήματα της ΕΦΓ.
- Δεν χρειαζόμαστε ειδική γνώση τομέα για την δημιουργία τους.
- Δεν χρειάζονται συλλογές κειμένων με πλούσιο σχολιασμό, αλλά μπορούν να αντληθούν από μεγάλες ασχολιαστες πηγές που είναι ήδη διαθέσιμες.
- Τέλος, καθώς πρόκειται για διανύσματα, δίνουν την δυνατότητα αξιοποίησης μαθηματικών εργαλείων για την εξαγωγή χρήσιμων συμπερασμάτων.

Όσον αφορά τον στόχο της εργασίας, που είναι ο υπολογισμός της σημασιολογικής ομοιότητας δυο κειμένων, τα word embeddings δεν προσφέρουν μία άμεση λύση, αλλά

αποτελούν ένα εργαλείο που μπορούμε να χρησιμοποιήσουμε. Οι τρόποι με τους οποίους μπορούμε να τα αξιοποιήσουμε για να εκτιμήσουμε το νόημα ενός κειμένου, παρουσιάζονται στην ενότητα 'Μελέτη Μεθόδων'.

Κοινή πρακτική, είναι η χρησιμοποίηση προ-εκπαιδευμένων word embeddings για μία εργασία, δηλαδή η χρήση embeddings από μοντέλα νευρωνικών δικτύων που σχεδιάστηκαν για να να κωδικοποιούν γενικές σημασιολογικές σχέσεις και εκπαιδεύτηκαν σε μεγάλες συλλογές κειμένων. Αυτό οφείλεται στο γεγονός πως η μάθηση word embeddings από το μηδέν δεν είναι εύκολη.

Μία από τις μεγαλύτερες προκλήσεις στην ΕΦΓ είναι η έλλειψη δεδομένων εκπαίδευσης. Καθώς η ΕΦΓ είναι ένα εκτεταμένο πεδίο με πολλές διαφορετικές εφαρμογές, τα περισσότερα σύνολα δεδομένων για συγκεκριμένες εφαρμογές περιέχουν μερικές χιλιάδες ή μερικές εκατοντάδες χιλιάδες παραδείγματα εκπαίδευσης ταξινομημένα από ανθρώπους. Ωστόσο, τα σύγχρονα μοντέλα βαθιάς μάθησης της ΕΦΓ βλέπουν οφέλη από μεγαλύτερες ποσότητες δεδομένων, και βελτιώνονται όταν εκπαιδεύονται σε εκατομμύρια, ή δισεκατομμύρια, σχολιασμένα παραδείγματα εκπαίδευσης.

Για να μπορέσει να αντιμετωπιστεί αυτό το πρόβλημα, οι ερευνητές ανέπτυξαν μία πληθώρα από τεχνικές για την εκπαίδευση γενικού σκοπού μοντέλων, χρησιμοποιώντας την τεράστια ποσότητα από κείμενα στο διαδίκτυο (η οποία διαδικασία είναι γνωστή ως *pre-training*). Το προ-εκπαιδευμένο μοντέλο μπορεί στην συνέχεια να προσαρμοστεί με ακρίβεια (*fine-tuned*) σε εφαρμογές της ΕΦΓ που αποτελούνται από μικρότερο όγκο δεδομένων. Έτσι πετυχαίνεται από την μία πλευρά, βελτίωση στην ακρίβεια και από την άλλη, εξοικονόμηση χρόνου .

Οι πρώτοι που έδειξαν την χρησιμότητα των προ-εκπαιδευμένων word embeddings ήταν οι Collobert και Weston το 2008. Η εργασίας τους ορόσημο 'A unified architecture for natural language processing' όχι μόνο καθιέρωσε τα word embeddings ως ένα χρήσιμο εργαλείο για διάφορες εφαρμογές της ΕΦΓ, αλλά επίσης εισήγαγε την αρχιτεκτονική ενός νευρωνικού δικτύου που έθεσε την βάση για πολλές τρέχουσες προσεγγίσεις. Η τελική διάδοση των word embeddings μπορεί να αποδοθεί στους Mikolov κ.α., που το 2103 δημιουργησαν το word2vec, ένα μοντέλο που επιτρέπει την εκπαίδευση και την χρήση προ-εκπαιδευμένων embeddings.

Σήμερα, υπάρχουν αρκετά νευρωνικά μοντέλα που έχουν σχεδιαστεί για να μαθαίνουν word embeddings. Στην παρούσα εργασία θα μελετήσουμε τα μοντέλα:
Word2Vec (by Google - 2013) , GloVe (by Stanford - 2014) , FastText (by Facebook - 2015),
και θα χρησιμοποιήσουμε προ-εκπαιδευμένα embeddings τους.

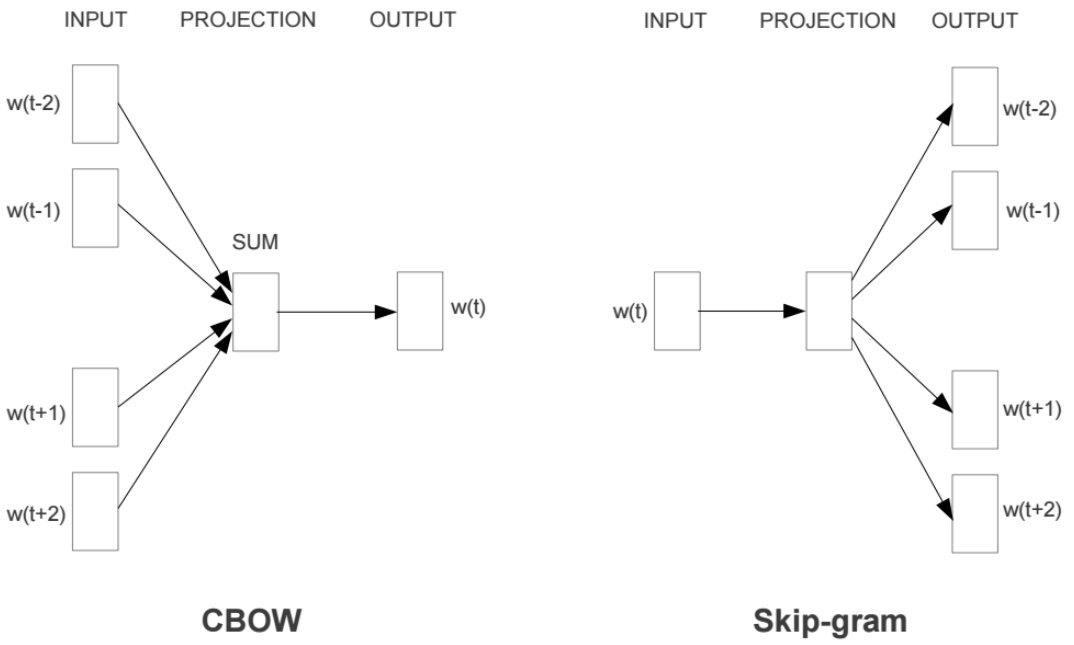
Παρακάτω παρουσιάζονται συνοπτικά, τα 3 αυτά μοντέλα.

2.2.1 Word2Vec

Η πρώτη και βασική μεθοδολογία για την εκμάθηση Embeddings Λέξεων, όπως αναφέρθηκε ήδη, είναι το μοντέλο Word2Vec (Mikolov et al., 2013).

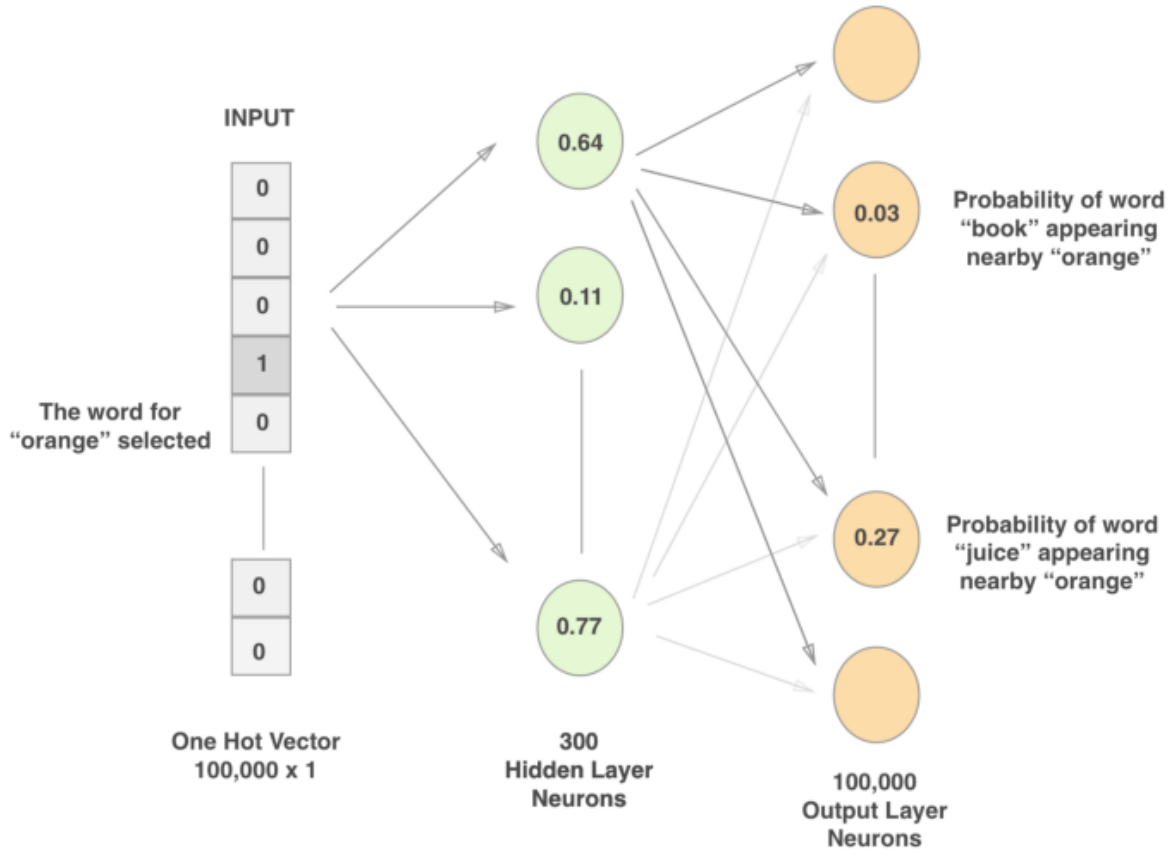
Το word2vec είναι μία μη-επιβλεπόμενη τεχνική, που λαμβάνει ως είσοδο μία συλλογή κειμένων και δίνει ως έξοδο word embeddings. Η εκπαίδευση του embedding κάθε λέξης γίνεται με βάση το συγκείμενο της πλαίσιο, στο corpus κειμένων που χρησιμοποιείται και η αναπαράσταση γίνεται σε έναν διανυσματικό χώρο N. (Κάθε λέξη αναπαριστάται ως ένα διάνυσμα διάστασης N σε αυτόν τον χώρο, όπου το N=300 συνήθως). Το συγκείμενο πλαίσιο ορίζεται ως το παράθυρο των γειτονικών λέξεων.

Το Word2Vec δεν είναι ένας αλγόριθμος αλλά έχουν προταθεί δύο διαφορετικά τεχνικές μάθησης: Το Combined Bag of Words (CBOW) και το Skip-gram μοντέλο. Και οι δύο μέθοδοι είναι νευρωνικά δίκτυα λίγων επιπέδων (shallow feed-forward NN with one hidden layer), που μαθαίνουν βάρη που χρησιμοποιούνται ως word embeddings. Το Skip-gram παίρνει ως είσοδο μία λέξη και προσπαθεί να προβλέψει το συγκείμενο πλαίσιο, ενώ το CBOW εκτελεί την αντίστροφη διαδικασία, λαμβάνει ως είσοδο ένα συγκείμενο πλαίσιο και προσπαθεί να προβλέψει την πιθανότητα μιας λέξης.



Σχήμα 2. Οι δύο αρχιτεκτονικές μοντέλων του Word2Vec. Το CBOW προβλέπει μία λέξη με βάση το συγκείμενο πλαίσιο, και το Skip-gram προβλέπει τις γύρω λέξεις μιας λέξης. ©arXiv:1301.371

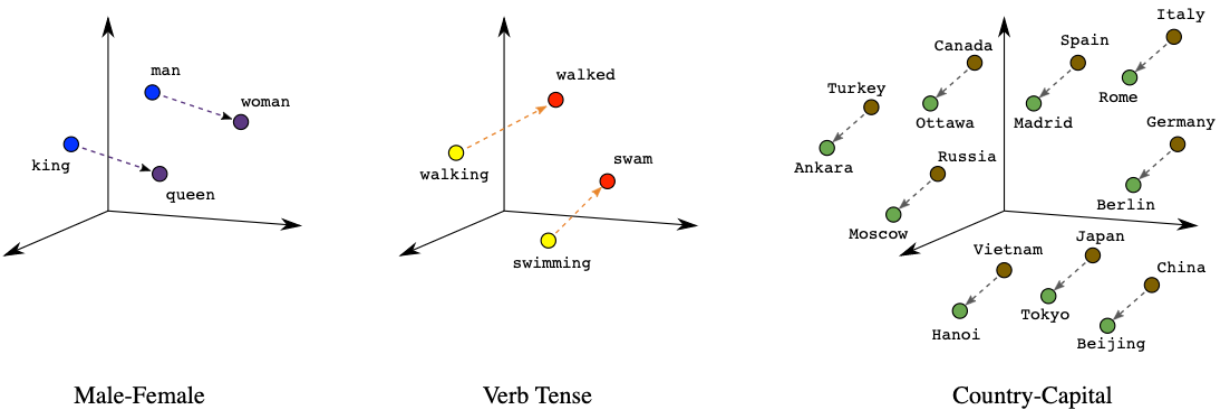
Παρακάτω δίνεται η οπτικοποίηση της λειτουργίας του μοντέλου Skip-gramγ μέσω ενός απλού παραδείγματος. Έστω ότι στο παράδειγμα, έχουμε ένα λεξιλόγιο με 100.000 μοναδικές λέξεις. Μία λέξη εισόδου, όπως π.χ. η λέξη 'orange' θα αναπαρίσταται με 1 στην θέση που αντιστοιχεί στο 'orange' και 0 σε όλες τις υπόλοιπες θέσεις. Η έξοδος του μοντέλου θα είναι ένα διάνυσμα διάστασης 100.000, όπου κάθε θέση θα περιέχει την πιθανότητα η αντίστοιχη λέξη να είναι ανήκει στο συγκείμενο πλαίσιο της λέξης 'orange.'



Σχήμα 3. Οπτικοποίηση του νευρωνικού δικτύου του μοντέλου Skip-gram για ένα τυχαίο παράδειγμα. ©towardsdatascience.com

Το πιο διάσημο αποτέλεσμα του word2vec ήταν η παρατήρηση ότι : ο τρόπος που οι λέξεις προβάλλονται στον διανυσματικό χώρο αποκαλύπτει σημασιολογικές σχέσεις διάφορων τύπων, όπως αρσενικό-θηλυκό, χρόνους ρημάτων, χώρα-πρωτεύουσα και κατ επέκταση, είναι δυνατά συμπεράσματα της μορφής :

model(king) - model(man) + model(woman) ≈ model(queen)



Σχήμα 4. Χρήσιμες αναλογίες των embeddings ©developers.google.com

2.2.2 GloVe

Το μοντέλο Glove (Global Vectors for Word Representation), προτάθηκε το 2014 από το πανεπιστήμιο Stanford και αποτελεί επέκταση του Word2Vev. Η βασική ιδέα είναι, ότι αντί να χρησιμοποιείται ένα 'παράθυρο' που ορίζει το τοπικό context, το GloVe βρίσκει τις σχέσεις μεταξύ των λέξεων από μία μήτρα συν-εμφάνισης λέξεων.

Η μήτρα συν-εμφάνισης λέξεων δίνει πληροφορίες για την συχνότητα με την οποία δύο λέξεις εμφανίζονται μαζί σε ένα μεγάλο corpus. Για την δημιουργία της, αρχικά ορίζεται ένα μέγεθος παραθύρου N (συνήθως 2 -10), και μετά με ένα πέρασμα σε όλο το κείμενο μετρώνται πόσες φορές κάθε ζεύγος λέξεων εμφανίζεται μαζί, δηλαδή χωρίζεται από μέχρι και N λέξεις.

$$X = \begin{array}{c} \\ I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} \begin{array}{c} \begin{array}{cccccccc} I & like & enjoy & deep & learning & NLP & flying & . \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Σχήμα 5. Παράδειγμα μήτρας συν-εμφάνισης λέξεων για N=2

Η μήτρα αυτή αποτελεί τον στόχο μάθησης του νευρωνικό δικτύου του GloVe. Πιο συγκεκριμένα, ο στόχος μάθησης του GloVe είναι να μάθει διανύσματα λέξεων τέτοια ώστε το εσωτερικό γινόμενο τους να ισούται με τον λογάριθμο της πιθανότητας της συν-εμφάνισης των λέξεων.

Η πιθανότητα της συν-εμφάνισης μιας λέξης j δοθέντος μιας λέξης i , είναι ο λόγος των φορών που η λέξη j εμφανίζεται στο συγκείμενο πλαίσιο της i, προς τον αριθμό των φορών που οποιαδήποτε άλλη λέξη εμφανίζεται στο context της i. Δηλαδή έχουμε :

$$P_{ij} = P(j|i) = \frac{X_{ij}}{\sum_{k \in context} X_{ik}}$$

, όπου X η μήτρα συν-εμφάνισης λέξεων και X_{kj} ο αριθμός φορών που η λέξη k εμφανίζεται στο συγκείμενο πλαίσιο της λέξης i.

Οπότε, ο λογάριθμος της πιθανότητας συνεμφάνισης των λέξεων θα είναι ίσος με το λογάριθμο της παραπάνω αναλογίας. Είναι γνωστό πως ο λογάριθμος μιας αναλογίας ισούται με την διαφορά των λογαρίθμων. Κατά αυτόν τον τρόπο, αυτός ο στόχος μάθησης συσχετίζει (τον λογάριθμο των) αναλογιών των πιθανοτήτων συνεμφάνισης με διανυσματικές διαφορες στο διανυσματικό χώρο. Καθώς οι αναλογίες αυτές μπορούν να κωδικοποιούν κάποια μορφή νοήματος, αυτή η πληροφορία κωδικοποιείται και ως διανυσματική διαφορά επίσης. Αυτός είναι και ο λόγος, που τα προκύπτοντα word embeddings αποδίδουν πολύ καλά σε tasks αναλογίας λέξεων.

Συμπερασματικά, οι δημιουργοί του GloVe έδειξαν ότι ο λόγος της πιθανότητας συν-εμφάνισης δύο λέξεων (παρά οι πιθανότητες συν-εμφάνισης από μόνες τους) είναι αυτές που περιέχουν πληροφορία και αυτή την πληροφορία προσπάθησαν να την κωδικοποιήσουν σε διανυσματικές διαφορές.

2.2.3 fastText

Το fastText που προτάθηκε το 2015 από την Facebook, αποτελεί επίσης, μία προέκταση του Word2Vec. Η βασική αρχή πίσω από το FastText, είναι ότι η μορφολογική δομή μιας λέξης φέρει σημαντική πληροφορία για το νόημα μιας λέξης. Στην πράξη, αυτό αποδίδεται με την συμπερίληψη n-gram χαρακτήρων στην εκπαίδευση.

Ως n-gram ορίζεται μία συνεχόμενη ακολουθία n αντικειμένων από ένα δεδομένο δείγμα κειμένου ή ομιλίας. Τα αντικείμενα μπορούν να είναι φωνήματα, συλλαβές, γράμματα, λέξεις ή ζεύγη βάσεων σύμφωνα με την εφαρμογή.

Κάθε λέξη λοιπόν, στο fastText αναπαρίσταται ως μία συλλογή n-gram χαρακτήρων, επιπρόσθετα από την ίδια την λέξη. Για παράδειγμα, για την λέξη ‘matter’ με n=3, οι n-grams χαρακτήρες όπως αναπαριστώνται από το FastText είναι : <ma, mat, att, tte, ter, er>. Τα < και > , προστίθενται ως συνοριακά σύμβολα, που ξεχωρίζουν το n-gram από την ίδια την λέξη.

Με αυτόν τον τρόπο, μπορούν να υπολογίζονται και διανυσματικές αναπαραστάσεις λέξεων που δεν εμφανίζονται στο αρχικό corpus (out-of-vocabulary (OOV) λέξεις), αθροίζοντας τα διανύσματα των συστατικών τους n-grams, δοθέντων ότι τουλάχιστον ένα από αυτά είναι βρίσκεται στο training corpus.

Τέλος, το fastText εκπαίδευται να βρίσκει την λέξη που λείπει δοθέντος όλων των υπόλοιπων λέξεων σε μία πρόταση.

2.3 Sentence Embeddings

Με τον όρο 'Sentence Embeddings' αναφερόμαστε σε μία παρόμοια έννοια με πριν, με την διαφορά ότι τώρα έχουμε αποτύπωση μιας πρότασης σε ένα διανυσματικό χώρο. Το embedding μιας πρότασης μπορεί να κωδικοποιεί έναν αριθμό από παράγοντες συμπεριλαμβανομένου του σημασιολογικού νοήματος , την συντακτική δομή και το θέμα. Η κωδικοποίηση γίνεται με μοντέλα νευρωνικών δικτύων. Η βασική ιδέα είναι ότι εκπαιδεύουμε ένα νευρωνικό δίκτυο για να μάθει πως να συνδυάσει καλύτερα τα επιμέρους word embeddings.

Αντιθέτως με την Μηχανική Όραση, όπου τα convolutional νευρωνικά δίκτυα κυριαρχούν, υπάρχουν πολλοί τρόποι για να κωδικοποιηθεί μία πρόταση με νευρωνικά δίκτυα. Πρόσφατες έρευνες έχουν εξετάσει μη-επιβλεπόμενες καθώς και επιβλεπόμενες τεχνικές μάθησης με διαφορετικούς στόχους εκπαίδευσης, προκειμένου να μάθουν γενικού σκοπού και σταθερού μήκους, αναπαραστάσεις προτάσεων.

Οι μη-επιβλεπόμενες τεχνικές, μαθαίνουν τα sentence embeddings ως ένα υποπροϊόν της μάθησης να προβλέπουν μία συνεκτική διαδοχή προτάσεων ή συνεκτική διαδοχή υπο-προτάσεων μέσα σε μία πρόταση. Αυτές οι προσεγγίσεις (στην θεωρία) μπορούν να χρησιμοποιούν οποιαδήποτε κειμενικό σετ δεδομένων (text dataset), αρκεί να περιλαμβάνει προτάσεις/υπο-προτάσεις που παρατίθενται με συνεκτικό τρόπο. Τα Skip-thoughts διανύσματα είναι το αρχετυπικό παράδειγμα της μη επιβλεπόμενης μάθησης sentence embeddings. Είναι το ισοδύναμο του skip-gram για τις προτάσεις : προσπαθεί να προβλέψει τις περιβάλλουσες προτάσεις δοθέντος μιας πρότασης.

Οι επιβλεπόμενες τεχνικές μάθησης από την άλλη πλευρά, απαιτούν ένα dataset σημειωμένο για κάποια εφαρμογή. Το ποιος στόχος εκπαίδευσης μπορεί να μάθει sentence embeddings που μπορούν χρησιμοποιηθούν σε transfer tasks είναι ένα από τα βασικά ερωτήματα του συγκεκριμένου πεδίου έρευνας. Ένας από τους πιο δημοφιλής στόχος εκπαίδευσης είναι η Εξαγωγή Συμπερασμάτων σε Φυσική Γλώσσα, η οποία θα παρουσιαστεί πιο αναλυτικά παρακάτω.

Στην παρούσα εργασία, μελετήθηκαν προ-εκπαιδευμένοι κωδικοποιητές προτάσεων (pre-trained sentence encoders) που είναι σχεδιασμένοι να παίξουν το ίδιο ρόλο που παίζουν τα Word2Vec και GloVe για τις λέξεις.
Συγκεκριμένα, ασχολούμαστε με τα :

InferSent (Facebook, 2017) , USE (Google, 2018) , SentenceBERT (UKP- TUDA, 2019)

Η αναλυτική παρουσίαση τους γίνεται στην επόμενη ενότητα, “Προτεινόμενες Μέθοδοι”.

2.4 Μέτρο Ομοιότητας

Τα παραχθέντα word και sentence embeddings, δημιουργούνται με τρόπο ώστε να συλλαμβάνουν σημασιολογική πληροφορία για λέξεις και προτάσεις αντίστοιχα. Προκειμένου να έχουμε κάποια γνώση για το πόσο δύο embeddings μοιάζουν μεταξύ τους, και άρα κατ’ επέκταση και για το πόσο μοιάζουν δύο λέξεις/προτάσεις, πρέπει να τα συγκρίνουμε.

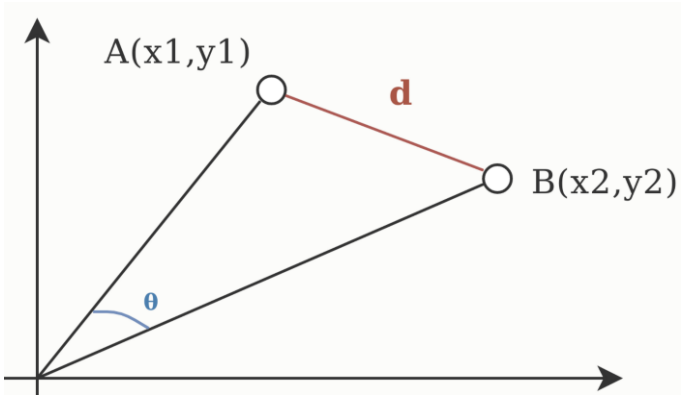
Γι’ αυτό χρησιμοποιούμε ένα μέτρο ομοιότητας, που λαμβάνει τα embeddings και επιστρέφει έναν αριθμό που να δείχνει την ομοιότητα τους. Ως μέτρο ομοιότητας, ορίζουμε μία πραγματική συνάρτηση που ποσοτικοποιεί την ομοιότητα ανάμεσα σε δύο αντικείμενα, ή διαφορετικά, μετράει πως/πόσο δύο αντικείμενα σχετίζονται ή μοιάζουν μεταξύ τους. Το μέτρο ομοιότητας συνήθως εκφράζεται ως ένας αριθμός ανάμεσα στο 0 και το 1, όπου μηδέν σημαίνει ελάχιστη ομοιότητα και ένα πολύ υψηλή. Όσο πιο κοντά στο 1, τόσο πιο πιθανό είναι ότι δύο συγκρινόμενα αντικείμενα να είναι πιο όμοια, και αντίστροφα.

Για τον υπολογισμό της ομοιότητας ανάμεσα σε δύο διανύσματα A = [x_1, x_2 , … , x_n] και B = [y_1, y_2 , … , y_n], δύο κοινά μέτρα ομοιότητας που μπορούμε να χρησιμοποιήσουμε είναι τα ακόλουθα :

* Ευκλείδεια Απόσταση

Η Ευκλείδεια απόσταση, ορίζεται ως η απόσταση ανάμεσα στα τέλη των διανυσμάτων, η διαφορετικά ως το μήκος του μονοπατιού που τα συνδέει.

Η απόσταση δίνεται από τον τύπο: $d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.



Σχήμα 6. Αναπαρασταση διανυσμάτων A και B, για n=2.

∗ Ομοιότητα Συνημιτόνου

Η ομοιότητα συνημιτόνου ορίζεται ως το συνημίτονο της γωνίας θ ανάμεσα στα διανύσματα. Ή εναλλακτικά, το συνημίτονο δύο διανυσμάτων ορίζεται ως το εσωτερικό τους γινόμενο, διαιρούμενο από το γινόμενο των μέτρων τους :

$$sim\,(A,\ B)\ =\ cos(\theta)\ =\ \frac{\mathrm{A}\cdot\mathrm{B}}{||\mathrm{A}||\ ||\mathrm{B}||}$$

Αν δύο διανύσματα είναι παράλληλα, δηλαδή σχηματίζουν γωνία 0 μοίρες, έχουν cos(0)=1, και άρα θεωρούνται όμοια. Διαφορετικά, ένα είναι κάθετα και είναι 90 μοίρες μακριά, cos(90)=0 και άρα η ομοιότητα είναι 0.

Η ομοιότητα συνημιτόνου υπολογίζει την κατεύθυνση δύο σημείων στο χώρο παρά το πόσο απέχουν μεταξύ τους. Αυτό σημαίνει ότι επηρεάζεται λίγο από το μέγεθος ή το πόσο μεγάλοι είναι οι αριθμοί.

ΚΕΦΑΛΑΙΟ 3

Προτεινόμενες Μέθοδοι Ομοιότητας Κειμένων

Στο κεφάλαιο αυτό αναλύονται οι μέθοδοι σημασιολογικής ομοιότητας κειμένων που μελετήθηκαν. Στην ενότητα 3.1 περιγράφονται οι μέθοδοι που βασίζονται στην χρήση word embeddings και στην ενότητα 3.2 οι μέθοδοι που βασίζονται στην χρήση sentence embeddings.

3.1. Μέθοδοι με χρήση Word Embeddings

3.1.1 Tf-Idf Διανύσματα

Η πρώτη προτεινόμενη μέθοδος και μία baseline προσέγγιση για την αριθμητική αναπαράσταση ενός κειμένου είναι τα Tf-Idf διανύσματα. Αν και δεν πρόκειται για καθαρά embeddings, αναπαριστούν διανυσματικά ένα κείμενο μέσω των Tf-Idf βαρών.

Η διαδικασία εισήχθη το 1972 σε μία εργασία του Karen Spärck Jones.

Με το Tf-Idf (Term Frequency - Inverse Document Frequency) μπορούμε να υπολογίσουμε ένα βάρος για κάθε λέξη, που δηλώνει την σημαντικότητα της λέξης στο κείμενο του corpus στο οποίο ανήκει. Η σημαντικότητα αυξάνεται αναλογικά με τον αριθμό των φορών που μία λέξη εμφανίζεται σε ένα κείμενο (term frequency = συχνότητα όρου) αλλά αντισταθμίζεται από την συχνότητα του όρου στο corpus (inverse document frequency = αντίστροφη συχνότητα κειμένου).

Η μαθηματική αναπαράσταση του βάρους ενός όρου t σε ένα έγγραφο d από το Tf-Ifd δίνεται από τον τύπο :

$$W(d,t) = TF(d,t) * log(\frac{N}{df(t)})$$

Όπου N ο αριθμός των εγγράφων στο corpus και df(t) είναι ο αριθμός των εγγράφων που περιέχουν τον όρο t στο corpus.

Κατ' αυτό τον τρόπο, μπορούμε να αναπαραστήσουμε ένα κείμενο ως διάνυσμα με διάσταση ίση με τον αριθμό των μοναδικών λέξεων στο corpus και τιμές τα tf-idf βάρη των λέξεων που ανήκουν στο κείμενο.

Οπότε, σε αυτήν την μέθοδο, το αποτέλεσμα της ομοιότητας θα βασίζεται στην σημαντικότητα των λέξεων που τα έγγραφα μοιράζονται.

Στην παρούσα εργασία, χρησιμοποιήθηκε η υλοποίηση του Tf-Idf της βιβλιοθήκης Gensim της Python.

```
from gensim.corpora import Dictionary
from gensim.models import TfidfModel
from gensim import similarities

dct = Dictionary(data_stem)
corpus = [dct.doc2bow(line) for line in data_stem]
model = TfidfModel(corpus)
vectors = model[corpus]
scores = similarities.MatrixSimilarity(vectors,num_features=len(dct)
```

Κώδικας 1. Υλοποίηση Tf-Idf με την βιβλιοθήκη gensim

3.1.2 Μέσος Όρος Word Embeddings

Μία άλλη απλή προσέγγιση για την αναπαράσταση ενός κειμένου διανυσματικά, είναι ο υπολογισμός του μέσου όρου των word embeddings όλων των λέξεων του.

Για την εύρεση των word embeddings, χρησιμοποιούμε προ-εκπαιδευμενα μοντέλα, που περιέχουν έτοιμες αναπαραστάσεις για ένα πολύ μεγάλο αριθμό λέξεων. Για λέξεις που δεν βρίσκονται στο λεξιλόγιο των μοντέλων, το διάνυσμα τίθεται ίσο με μηδέν.

Όπως αναφέρθηκε και παραπάνω, τα word embedding μοντέλα που χρησιμοποιούμε είναι τα Word2vec, GloVe, FastText, και τα προεκπαιδεύμενα μοντέλα τους αντίστοιχα είναι :

※'GoogleNews-vectors-negative300', έχει εκπαιδευτεί στο Google News Dataset (περίπου 100m λέξεις) και περιέχει περίπου 3 εκατομμύρια διανυσματικές αναπαραστάσεις λέξεων και φράσεων.

※' glove.42B.300d', έχει εκπαιδευτεί στο Common Crawl[1] και περιέχει διανυσματικές αναπαραστάσεις για 1.9 εκατομμύρια tokens.

※'wiki-news-300d-1M.vec', εκπαίδευτηκε πάνω στην WIkipedia και περιέχει 1 εκατομμύριο λέξεις.

Για την φόρτωση των μοντέλων χρησιμοποιήθηκε η βιβλιοθήκη gensim της Python.

```
from gensim.models import Word2Vec , KeyedVectors
model = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin',
binary=True)
```

Κώδικας 2. Φόρτωση προ-εκπαιδευμένων word embeddings με την βιβλιοθήκη gensim

[1] Common Crawl = "The Common Crawl corpus contains petabytes of data collected over 12 years of web crawling. The corpus contains raw web page data, metadata extracts and text extracts.", from the official web page.

3.1. 3 Smooth Inverse Frequency

Αν και ο μέσος όρος των word embeddings μία πρότασης θεωρείται καλό baseline, τείνει να δίνει πολύ βάρος σε λέξεις που είναι μη-σημαντικές σημασιολογικά (π.χ. 'και', 'μπορεί' , 'θα', 'τότε', κτλ.) . Το 2016, ο Arora κ.α., πρότειναν τα SIF embeddings, ως λύση σε αυτό το πρόβλημα. Η μέθοδος SIF υπολογίζει τα embeddings των προτάσεων ως έναν σταθμισμένο μ.ο. των embeddings των λέξεων της, συνδυασμένο με ένα ήπιο 'denoising'. Τα δύο βασικά μέρη της μεθόδου είναι :

1. Weighting - Στάθμιση
2. Common component removal - Αφαίρεση Κοινών Συνιστωσών

Algorithm 1 Sentence Embedding

Input: Word embeddings $\{v_w : w \in \mathcal{V}\}$, a set of sentences $\mathcal{S}$, parameter a and estimated probabilities $\{p(w) : w \in \mathcal{V}\}$ of the words.
Output: Sentence embeddings $\{v_s : s \in \mathcal{S}\}$
1: **for all** sentence s in $\mathcal{S}$ **do**
2: $v_s \leftarrow \frac{1}{|s|} \sum_{w \in s} \frac{a}{a+p(w)} v_w$
3: **end for**
4: Form a matrix X whose columns are $\{v_s : s \in \mathcal{S}\}$, and let u be its first singular vector
5: **for all** sentence s in $\mathcal{S}$ **do**
6: $v_s \leftarrow v_s - uu^\top v_s$
7: **end for**

Κώδικας 3. Ψευδοκώδικας Υπολογισμού SIF Embeddings ©'A Simple But Tough-To-Beat Baseline For Sentence Embeddings''

Μέθοδος Για τον Υπολογισμό των SIF Embeddings:

I. Κάθε word embedding σταθμίζεται κατά a/(a + p(w)), όπου α είναι μία υπερ-παράμετρος, ορισμένη από τους συγγραφείς στο 0.001 , και p(w) η συχνότητα της λέξης w στο corpus. Η υπερ-παράμετρος a προσαρμόζει ποιες λέξεις είναι ποσοτικά "συνηθισμένες " και "ασυνήθιστες".

II. Στην συνέχεια, για κάθε πρόταση υπολογίζεται ο μ.ο. αυτών των σταθμισμένων word embeddings.

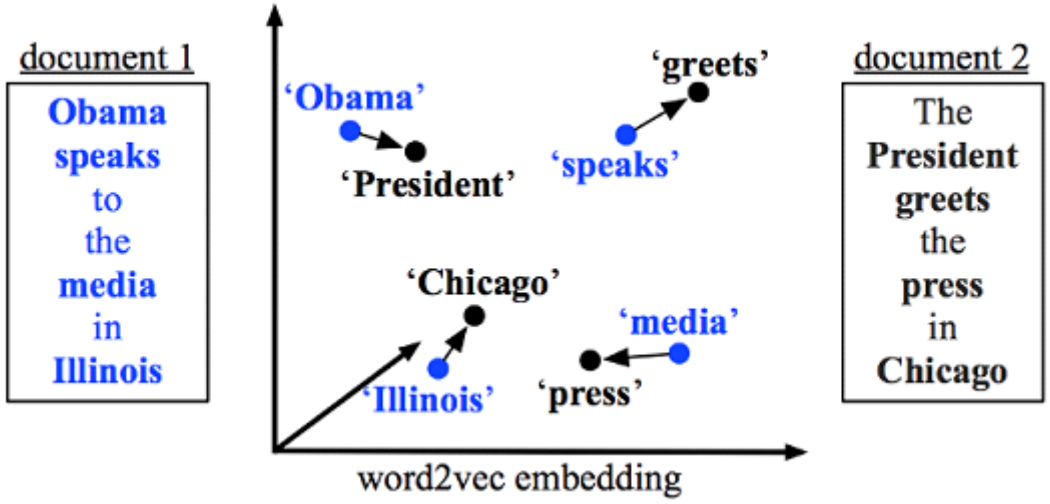
III. Τέλος, υπολογίζονται οι κύριες συνιστώσες των προκύπτοντων sentence embeddings. και αφαιρείται από αυτά τα sentence embeddings η πρώτη συνιστώσα τους. Ως αποτέλεσμα, το SIF υποβαθμίζει ασήμαντες λέξεις όπως 'but', 'just', κτλ και κρατάει πληροφορία που συνεισφέρει περισσότερο στην σημασιολογία της πρότασης.

3.1.4 Word Mover's Distance

Ένας διαφορετικός τρόπος αξιοποίησης των word embeddings για την εύρεση της ομοιότητας, είναι μέσω του ορισμού μιας συνάρτησης απόστασης.

Στην παρούσα εργασία χρησιμοποιούμε το Word Mover's Distance (WMD), μία συνάρτηση απόστασης μεταξύ κειμένων, που παρουσιάστηκε το 2015, στην εργασία "From Word Embeddings To Document Distances".

Το WMD χρησιμοποιεί τα word embeddings των λέξεων δύο κειμένων για να μετρήσει την ελάχιστη απόσταση που οι λέξεις στο πρώτο κείμενο χρειάζονται να "ταξιδέψουν" στο σημασιολογικό χώρο για να "φτάσουν" τις λέξεις στο δεύτερο κείμενο.



Σχήμα 7. Μία απεικόνιση του WMD. © 'From Word Embeddings To Document Distances'

Παραπάνω, απεικονίζεται το παράδειγμα που παρουσιάστηκε στην πρωτότυπη εργασία. Για τις προτάσεις "Obama speaks to the media in Illinois" και "The President greets the press in Chicago", παρατηρούμε γρήγορα πως, αν και έχουν μηδενική λεκτική ομοιότητα, οι λέξεις τους είναι όμοιες σημασιολογικά, και άρα βρίσκονται κοντά στον διανυσματικό χώρο. Το WMD βασιζόμενο σε αυτή την απλή παρατήρηση, υπολογίζει τις αντίστοιχες αποστάσεις των λέξεων τους στον διανυσματικό χώρο και εάν είναι μικρές 'βρίσκει' πως τα κείμενα είναι όμοια.

Το WMD ωστόσο, είναι μία πολύ αργή μέθοδος. Συγκεκριμένα, είναι O(n*m) , όπου n = το μήκος της πρότασης 1, m = το μήκος της πρότασης 2.

```
model.wmdistance( ['Obama', 'speaks', 'media', 'Illinois' ] , ['President', 'greets', 'press', 'Chicago'] )
```

Κώδικας 4. Υπολογισμός παραπάνω παραδείγματος με το μοντέλο που ήδη έχουμε φορτώσει από τον Κώδικα 2.

3.2 Μέθοδοι βασιζόμενοι σε Sentence Embeddings

3.2.1 InferSent

Το 2017, η ομάδα έρευνας του Facebook κυκλοφόρησε το InferSent, το πρώτο μοντέλο για την μάθηση καθολικών (universal) αναπαραστάσεων προτάσεων με επιβλεπόμενη μάθηση. Συγκεκριμένα, το InferSent είναι ένας προ-εκπαιδευμένος κωδικοποιητής προτάσεων που μεταφέρεται καλά σε διαφορές εφαρμογές της ΕΦΓ.

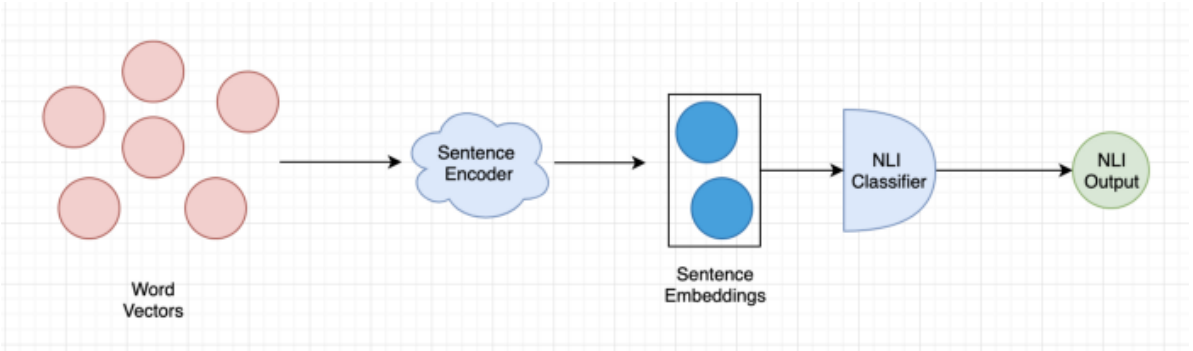
Η βασική και καινοτόμα ιδέα των συγγραφέων προκειμένου να λάβουν καθολικές αναπαραστάσεις, ήταν να εκπαιδεύσουν το μοντέλο τους, σε ένα task εξαγωγής συμπερασμάτων. Η Εξαγωγή Συμπερασμάτων στην Φυσική Γλώσσα (Natural Language Inference) είναι the task της ταξινόμησης ζευγαριών προτάσεων ως συνεπαγωγή /entailment , αντίφαση / contradiction , ή ουδέτερο / neutral (καμία από τις δύο προηγούμενες).

Τα Πειράματα που δημοσιεύτηκαν μαζί με την εργασία, έδειξαν πως τα sentence embeddings που μαθαίνονται με αυτό τον τρόπο, παρουσιάζουν τα καλύτερα αποτελέσματα μεταφερσιμότητας που είχαν μετρηθεί μέχρι τότε.

3.2.1.1 Λειτουργία Μοντέλου

Αρχικά, για ένα ζεύγος προτάσεων, γίνεται ο χωρισμός των επιμέρους λέξεων, και βρίσκονται τα αντίστοιχα Glove Embeddings. Στην συνέχεια, η αρχιτεκτονική του νευρωνικού δικτύου του InferSent αποτελείται από δύο μέρη :

1. Το πρώτο είναι ο κωδικοποιητής προτάσεων, που δεχόμενος ως είσοδο τις διανυσματικές αναπαραστάσεις των λέξεων κάθε πρότασης, κωδικοποιεί τις προτάσεις σε διανύσματα.
2. Το δεύτερο μέρος, είναι ένας NLI ταξινομητής, που παίρνει ως είσοδο της διανυσματικές αναπαραστάσεις των προτάσεων που δημιουργήθηκαν προηγούμενος, και δίνει ως έξοδο μία κλάση μεταξύ των : entailment, contradiction και neutral.

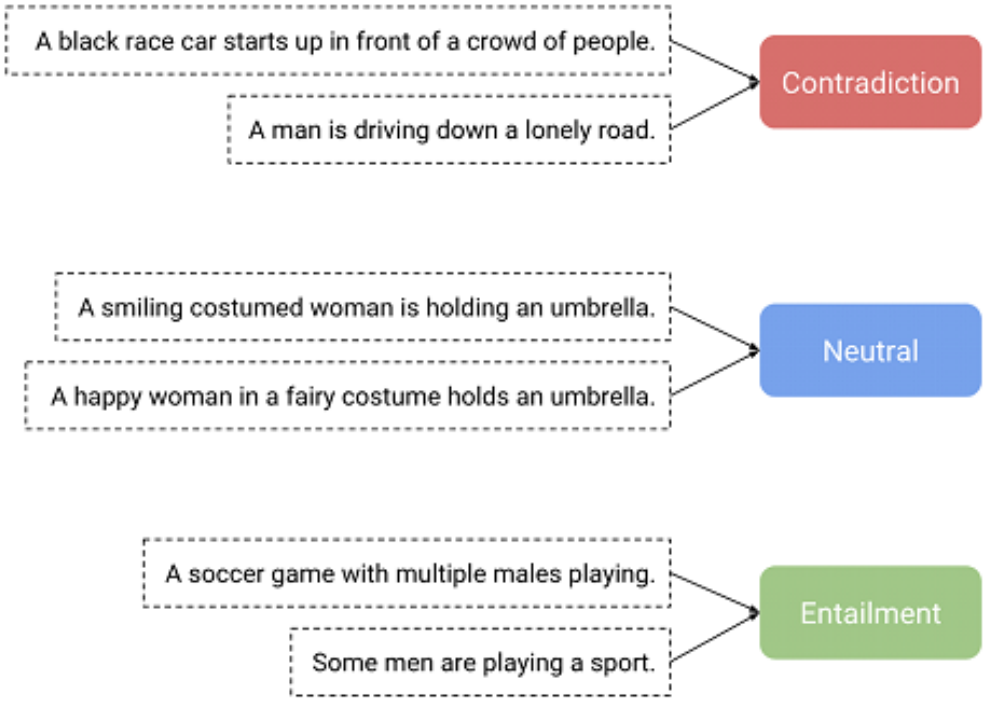


Σχήμα 8. Γενική Ροή Του InferSent.

3.2.1.2 Εκπαίδευση Μοντέλου

Η εκπαίδευση του μοντέλου έγινε για τον classifier, και συγκεκριμένα πως δοθέντων δύο προτάσεων θα κάνει την σωστή πρόβλεψη για το είδος της σχέσης τους. Το dataset που χρησιμοποιήθηκε για το σκοπό αυτό είναι το Stanford Natural Language Inference (SNLI).

Το SNLI είναι μία συλλογή από 570 χιλιάδες ζευγάρια προτάσεων στα Αγγλικα, γραμμένα από ανθρώπους, που χειρωνακτικά ταξινομήθηκαν με έναν εκ των χαρακτηρισμών : entailment, contradiction, and neutral.



Σχήμα 9. Παράδειγμα SNLI dataset.

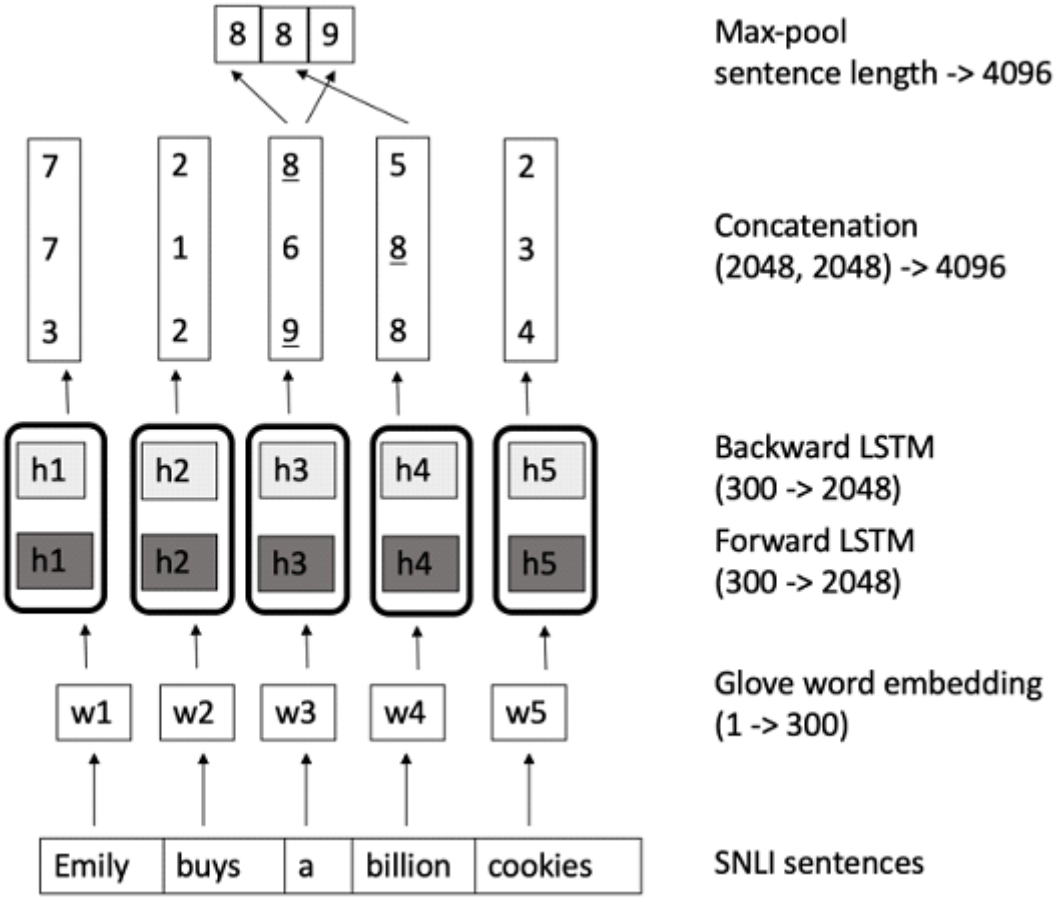
Η βασική ιδέα των συγγραφέων είναι ότι η σημασιολογική φύση αυτού του task, μπορεί να το καταστήσει έναν καλό υποψήφιο για την μάθηση αναπαραστάσεων προτάσεων, που συλλαμβάνουν καθολικά χρήσιμα χαρακτηριστικά. Καθώς το μοντέλο εκπαιδεύεται για να μάθει την σημασιολογική ομοιότητα για τα ζεύγη των προτάσεων, τα sentence embeddings μπορούν να χρησιμοποιηθούν για τον υπολογισμό της σημασιολογικής ομοιότητας ανάμεσα σε προτάσεις.

3.2.1.3 Κωδικοποιητής Μοντέλου

Εξερευνώντας διάφορες αρχιτεκτονικές, οι συγγραφείς κατέληξαν πως η αρχιτεκτονική με την καλύτερη επίδοση για τον κωδικοποιητή είναι ένα αμφίδρομο LSTM (bi-directional Long Short-Term Memory) δίκτυο με max-pooling.

Ένα αμφίδρομο LSTM δίκτυο επεξεργάζεται μία ακολουθία λέξεων και σε κανονική (forward LSTM) και αντίστροφη σειρά (backward LSTM). Έτσι, για μία ακολουθία n λέξεων, ένα bi-LSTM υπολογίζει ένα σύνολο n διανύσματα και κάθε διάνυσμα είναι μία συνένωση ενός forward LSTM και ενός backward LSTM. Κατόπιν, ένα επίπεδο σμίκρυνσης εφαρμόζεται σε κάθε συνενωμένο διάνυσμα, επιλέγοντας την μέγιστη τιμή από κάθε διάσταση των κρυφών επιπέδων (max pooling), για τον σχηματισμό του σταθερού μήκους τελικό διάνυσμα.

Η χρησιμοποίηση ενός αμφίδρομου LSTM έναντι ενός απλού LSTM επιτρέπει περισσότερη εκφραστικότητα με την χρήση περισσότερου συγκείμενου. Τα αμφίδρομα LSTM δίκτυα είναι ένας τρόπος κρίσης της "χρησιμότητας" μιας λέξης χρησιμοποιώντας και τις προηγούμενες και τις επόμενες λέξεις σε μία πρόταση.



Σχήμα 10. Κωδικοποιητής InferSent : Bi-LSTM + Max-Pooling.

3.2.1.4 Κώδικας

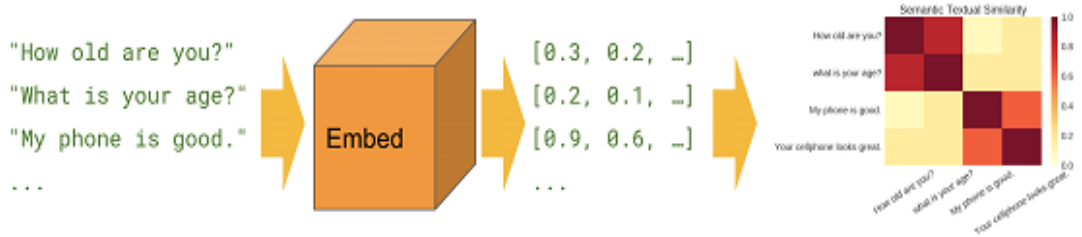
```
import torch
from models import InferSent

model_version = 1
MODEL_PATH = "infersent%s.pkl" % model_version
params_model = {'bsize': 64, 'word_emb_dim': 300, 'enc_lstm_dim': 2048,
                'pool_type': 'max', 'dpout_model': 0.0, 'version': model_version}

model = InferSent(params_model)

model.load_state_dict(torch.load(MODEL_PATH))

W2V_PATH = 'glove.840B.300d.txt'
model.set_w2v_path(W2V_PATH)

model.build_vocab(sentences, tokenize=True)

embeddings = model.encode(sentences , bsize=128)
```

Κώδικας 5. Υλοποίηση InferSent σε Python.

3.2.2 USE

Ο Universal Sentence Encoder, USE , κυκλοφόρησε ένα χρόνο αργότερα, το 2018, από την Google. Πρόκειται για μία συλλογή μοντέλων για την κωδικοποίηση κειμένου, με μήκος μεγαλύτερο από αυτό μιας λέξης (όπως προτάσεις, φράσεις, ή μικρές παράγραφοι) σε embedding διανύσματα.Κύριος στόχος των παραγόμενων sentence embeddings είναι η μεταφερσιμότητα τους σε διάφορες εργασίες της ΕΦΓ, όπως είναι ταξινόμηση κειμένου, σημασιολογική ομοιότητα, και ομαδοποίηση.

Μία συλλογή από προ-εκπαιδευμένα μοντέλα του USE είναι διαθέσιμα στο Tensorflow-hub.



Σχήμα 11. Βασική απεικόνιση λειτουργίας του USE για την εύρεση της σημασιολογικής ομοιότητας μεταξύ προτάσεων. ©tensorflow.com

Στην παρούσα εργασία χρησιμοποιήθηκαν δύο εξ' αυτών, τα ίδια με αυτά που παρουσιάστηκαν και στην αρχική εργασία : Το πρώτο μοντέλο εκπαιδεύεται με έναν κωδικοποιητή Transformer και το δεύτερο με ένα Deep Averaging Network (DAN). Τα δύο μοντέλα έχουν ένα αντιστάθμισμα ακρίβειας και απαιτήσεων σε υπολογιστικούς πόρους.

Και τα δύο μοντέλα, έχουν εκπαιδευτεί σε πολλές και διαφορετικές πηγών δεδομένων (supervised και unsupervised) και σε διαφορετικά tasks, με σκοπό να συλλέξουν όσο περισσότερη γενική σημασιολογική πληροφορία είναι δυνατό. Ενδεικτικά, οι unsupervised πηγές είναι : η Wikipedia, ειδήσεις από το διαδίκτυο, σελίδες ερωτήσεων-απαντήσεων στο διαδίκτυο, και φόρουμ συζητήσεων, ενώ ως supervised corpus χρησιμοποιήθηκε το SNLI.

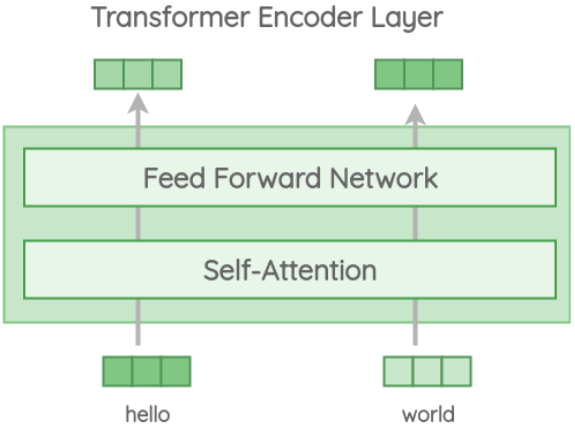
Παρακάτω, παρουσιάζονται εν συντομία και τα δύο μοντέλα.

3.2.2.1 Κωδικοποιητής Transformer

Το πρώτο μοντέλο που θα μελετήσουμε χρησιμοποιεί τον κωδικοποιητή Transformer. Το μοντέλο Transformer εισήχθη από την Google στην εργασία “Attention is All You Need” το 2018. Είναι μία καινοτόμα αρχιτεκτονική που στοχεύει στο να λύσει sequence-to-sequence εργασίες, ενώ χειρίζεται με ευκολία τις εξαρτήσεις μεγάλης εμβέλειας . Τα μοντέλα sequence-to-sequence (Seq2Seq) αφορούν την μετατροπή ακολουθιών (sequences) από έναν τομέα (π.χ. προτάσεις στα Αγγλικά) σε ακολουθίες σε έναν άλλο τομέα (π.χ. οι ίδιες προτάσεις μεταφρασμένες στα Ελληνικά).

Ο Transfomer περιλαμβάνει δύο διαφορετικούς μηχανισμούς : έναν κωδικοποιητή που “διαβάζει” το κείμενο εισόδου και έναν αποκωδικοποιητή που παράγει μία πρόβλεψη για το task. Το μοντέλο Trasnformer, βασίζεται στην ιδέα του self-attention, δηλαδή δεν επεξεργάζεται μία ακολουθία εισόδου λέξη-λέξη, αλλά λαμβάνει ως είσοδο ολόκληρη την ακολουθία, και έτσι επεξεργάζεται τις λέξεις σε σχέση με όλες τις υπόλοιπες λέξεις σε μία πρόταση.

Το μοντέλο USE που βασίζεται στο μοντέλο Transformer, κάνει χρήση του τμήματος του κωδικοποιητή της αρχικής αρχιτεκτονικής του. Πιο συγκεκριμένα, αποτελείται από μία στοίβα N=6 πανομοιότυπων επιπέδων. Κάθε επίπεδο έχει ένα self-attention τμήμα ακολουθούμενο από ένα feed-forward δίκτυο. Τα word-embeddings που δίνονται ως έξοδο, προστίθενται στοιχείο-στοιχείο και διαιρούνται από την τετραγωνική ρίζα του μήκους της πρότασης, ώστε να ληφθεί υπόψη η διαφορά στο μήκος των προτάσεων. Ως έξοδο λαμβάνουμε ένα sentence embedding διάστασης 512.



Σχήμα 12.Μοντέλο USE με κωδικοποιητή Tranformer ©amitness.com

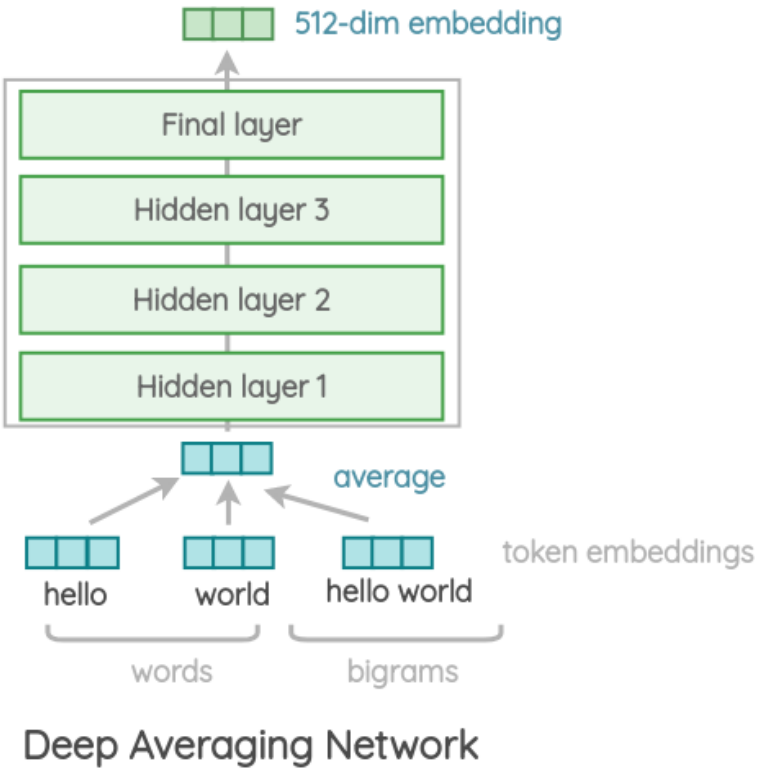
Αυτός ο κωδικοποιητής έχει καλύτερη ακρίβεια αλλά υψηλότερες απαιτήσεις σε μνήμη και υπολογιστικούς πόρους λόγω της πολύπλοκης αρχιτεκτονικής του. Επίσης, ο χρόνος υπολογισμού κλιμακώνεται δραματικά με την αύξηση του μήκους της πρότασης εισόδου, καθώς το self-attention έχει πολυπλοκότητα $O(n^2)$, όπου n το μήκος της πρότασης.

```
import tensorflow_hub as hub
model = hub.load('https://tfhub.dev/google/universal-sentence-encoder-large/5')
embeddings = model(texts)
```

Κώδικας 6. Φόρτωση USE μοντέλου με κωδικοποιητή Transformer

3.2.2.2 Κωδικοποιητής DAN

Με το μοντέλο DAN, πρώτα υπολογίζεται ο μέσος όρος των embeddings των λέξεων και των bi-grams μαζί. Κατόπιν, περνάνε από ένα 4-επιπέδων feed-forward νευρωνικό δίκτυο για την δημιουργία των sentence embeddings.



Σχήμα 13.Μοντέλο USE με κωδικοποιητή DAN ©amitness.com

Το κύριο πλεονέκτημα του κωδικοποιητή DAN είναι ότι ο χρόνος υπολογισμού είναι γραμμικός στο μήκος της πρότασης εισόδου.

3.2.3 SentenceBERT

Το μοντέλο SentenceBERT, προτάθηκε το 2019 σαν τροποποίηση του γνωστού προ-εκπαιδευμένου μοντέλου BERT, με τρόπο ώστε να χρησιμοποιείται για την εξαγωγή σημασιολογικών embeddings.

3.2.3.1 Το μοντέλο BERT

Το μοντέλο BERT, “Bidirectional Encoder Representations from Transformers”, δημιουργήθηκε και δημοσιεύτηκε το 2018 από την Google, και αποτέλεσε αμέσως την state-of-the-art τεχνική στην ΕΦΓ για την δημιουργία embeddings.

Ένα εκπαιδευμένο BERT μοντέλο παίρνει ως είσοδο μία πρόταση και δίνει ως έξοδο διανύσματα για κάθε λέξη της πρότασης, που περιλαμβάνουν πληροφορία για την θέση τους και τις περιβάλλουσες λέξεις. Το μοντέλο BERT κάνει χρήση του μοντέλου Transformer για την δημιουργία embeddings λέξεων που λαμβάνουν υπόψη ολόκληρο το συγκείμενο πλαίσιο μιας λέξης.

Το BERT φτιάχτηκε για να καταλαβαίνει την πρόθεση πίσω από τα αιτήματα αναζήτησης.

Αυτό που κάνει το BERT καινοτόμο, είναι ότι πρόκειται για την πρώτη βαθιά αμφίδρομη, μη επιβλεπόμενη γλωσσική αναπαράσταση, προ-εκπαιδευμένη χρησιμοποιώντας μόνο απλό κείμενο (συγκεκριμένα εκπαιδεύτηκε στην Wikipedia).

Μπορούμε να χρησιμοποιήσουμε το μοντέλο BERT για να εξάγουμε υψηλής ποιότητας γλωσσικά χαρακτηριστικά από τα δεδομένα μας, ή διαφορετικά μπορούμε να το προσαρμόσουμε με ακρίβεια σε συγκεκριμένα tasks με δικά μας δεδομένα.

Ωστόσο, για sentence-pair regression tasks πρόβλεψης συνεχούς τιμής, απαιτεί ως είσοδο και τις δύο προτάσεις μαζί, χωρίζομενες από ένα ειδικό token [SEP]. Αυτό προκαλεί μία τεράστια υπολογιστική επιβάρυνση από την μια πλευρά. Και από την άλλη, με αυτό τον τρόπο, δεν παράγονται ανεξάρτητα embeddings προτάσεων. Επίσης, το μοντέλο BERT δεν εκπαιδεύεται για την σημασιολογική ομοιότητα προτάσεων κατευθείαν όπως τα μοντέλα USE και InferSent.

Έτσι, έχουν αναπτυχθεί μοντέλα και υλοποιήσεις που χρησιμοποιούν το μοντέλο BERT, για την εξαγωγή sentence embeddings με σημασιολογική πληροφορία. Στην παρούσα εργασία, θα χρησιμοποιήσουμε το SentenceBERT που προτάθηκε από το Open Source by UKILab το 2019.

3.2.3.2 Αρχιτεκτονική Μοντέλου SentenceBERT

Το SentenceBERT χρησιμοποιεί Siamese BERT-Networks για την εξαγωγή σημασιολογικών embeddings προτάσεων, που μπορούν να συγκριθούν χρησιμοποιώντας απλά συνημίτονο.

Στην αρχιτεκτονική του SentenceBERT, είναι ενσωματωμένες 4 βασικές έννοιες :

1. **Attention** - επιτρέπει στον αλγόριθμο να δημιουργήσει τα embeddings εστιάζοντας μόνο στα πιο σημαντικά μέρη της εισόδου

2. **Transformers** - an attention based model with positional encodings

3. **BERT** - αποτελείται από 24 layers από Transformer blocks.

4. **Siamese Network** - μία κλάση νευρωνικών δικτύων που περιλαμβάνει δύο ή περισσότερα πανομοιότυπα υποδίκτυα. Πανομοιότυπα, διότι έχουν την ίδια ρύθμιση παραμέτρων, with the same parameters and weights parameter updating is mirrored across both sub-networks. Εκπαιδεύεται για να μαθαίνει μία συνάρτηση ομοιότητας, και έτσι χρησιμοποιείται για να συγκρινει την ομοιότητα μεταξύ δύο εισόδων.

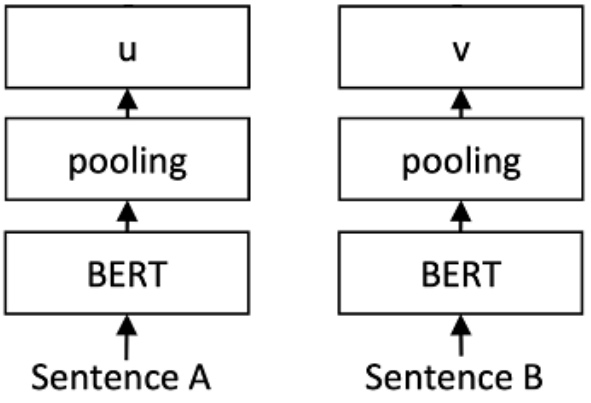
3.2.3.3 Λειτουργία Μοντέλου

Το Sentence-BERT χρησιμοποιεί ως αρχιτεκτονική ένα Siamese δίκτυο για να παρέχει δύο προτάσεις ως εισόδους. Αυτό του επιτρέπει να επεξεργάζεται δύο προτάσεις με τον ίδιο τρόπο ταυτόχρονα.

Οι προτάσεις στην συνέχεια, περνάνε σε BERT μοντέλα και ένα επίπεδο σμίκρυνσης (pooling layer) για την δημιουργία των embeddings τους.

Το επίπεδο σμίκρυνσης μας επιτρέπει να δημιουργούμε μία fixed-size αναπαράσταση για τις προτάσεις εισόδου που ποικοίλουν το μήκος.

Τέλος, τα embeddings των προτάσεων χρησιμοποιούνται για να βρεθεί η ομοιοτητα τους μέσω της συνάρτησης ομοιότητας συνημιτόνου.



Σχήμα 14 .Απεικόνιση αρχιτεκτονικής δικτύου SentenceBERT.
©https://www.sbert.net/

3.2.3.3 Προ-εκπαιδευμένα μοντέλα

Στην παρούσα εργασία, χρησιμοποιήθηκαν 3 διαφορετικά προ-εκπαιδευμένα SentenceBERT μοντέλα, ειδικά βελτιστοποιημένα στο task της σημασιολογικής ομοιότητας.
Οι συγγραφείς του SentenceBERT χρησιμοποίησαν το SNLI dataset και το Multi-Genre NLI (MG-NLI) dataset για την δημιουργία μίας συλλογής από 1,000,000 ζευγάρια προτάσεων για την εκπαίδευση των μοντέλων.

Τα μοντέλα που χρησιμοποιούμε είναι τα ακόλουθα :

※ *"stsb-bert-base"*

※ *"stsb-roberta-base"*

※ *"stsb-distilbert-base"*

```
from sentence_transformers import SentenceTransformer

sbert_model = SentenceTransformer('stsb-bert-base') #load the model
sentence_embeddings = sbert_model.encode(texts)
```

Κώδικας 7. Φόρτωση 'stsb-bert-base' και δημιουργία embeddings.

ΚΕΦΑΛΑΙΟ 4

Πειραματική Αξιολόγηση

Σε αυτό το κεφάλαιο αξιολογούνται πειραματικά τα μοντέλα υπολογισμού της σημασιολογικής ομοιότητα κειμένων που παρουσιάστηκαν παραπάνω. Στην ενότητα 4.1 περιγράφεται το dataset μικρών κειμένων που χρησιμοποιήθηκε, δίνονται παραδείγματα ζευγαριών των κειμένων και περιγράφεται το dataset από το οποίο λάβαμε το ground truth . Στην ενότητα 4.2 περιγράφεται η διαδικασία προεπεξεργασίας κειμένων που χρειάστηκε για την χρήση των μεθόδων που βασίζονται στα word embeddings. Στην ενότητα 4.3 και 4.4, γίνεται αναφορά στην επιλογή του μέτρου ομοιότητας και του μέτρου απόδοσης αντίστοιχα. Τέλος, στην ενότητα 4.5 παρατίθενται τα πειραματικά αποτελέσματα και γίνεται ο σχολιασμός του.

4.1 Dataset και Ground Truth

Προκειμένου να αξιολογήσουμε τις προτεινόμενες μεθόδους, δηλαδή την ακρίβεια τους, ένας απλός τρόπος είναι να συγκρίνουμε την απόδοση τους σε σχέση με την ανθρώπινη κρίση. Όσο πιο πολύ μία μέθοδος πλησιάζει την ανθρώπινη κρίση, τόσο πιο πολύ θεωρείται πιο ακριβής.

Το βασικό dataset μας, είναι ένα μικρό corpus από 50 ειδήσεις επιλεγμένες από πρωτοσέλιδες ειδήσεις του ηλεκτρονικού ταχυδρομείου της Αυστραλιανής Ραδιοτηλεόρασης. Το dataset συνοδεύεται επίσης από μία συλλογή από ανθρώπινες αξιολογήσεις για την ομοιότητα κάθε ζεύγους ειδήσεων. Τόσο το corpus όσο και οι ανθρώπινες αξιολογήσεις, συλλέχθηκαν από τον Lee και τους συνεργάτες του, για την εργασία τους : “An empirical evaluation of models of text document similarity” το 2005.

Τα 50 ειδησεογραφικά κείμενα κυμαίνονται σε μήκος από 51-126 λέξεις, και καλύπτουν ένα ευρύ φάσμα από θέματα. Σύμφωνα με παρατηρήσεις του πρωτότυπου paper, το corpus αυτό είναι ‘within the normal range of English text for word frequency spectrum and vocabulary growth’ και οπότε μπορεί να θεωρηθεί “αντιπροσωπευτικό ενός μέσου αγγλικού κειμένου”.

> “China said Sunday it issued new regulations controlling the export of missile technology, taking steps to ease U.S. concerns about transferring sensitive equipment to Middle East countries, particularly Iran. However, the new rules apparently do not ban outright the transfer of specific items - something Washington long has urged Beijing to do. (54 words)”

Παράδειγμα κειμένου του dataset.

Οι ανθρώπινες μετρήσεις αποτελούνται από 10 αξιολογήσεις για κάθε ζεύγος κειμένων.

Οι αξιολογήσεις είναι σε κλίματα 5 σημείων (όπου 1 = “υψηλά ανόμοια” και 5 = “υψηλά όμοια”). Πάλι σύμφωνα με παρατηρήσεις του αρχικού paper, “οι αξιολογήσεις δεν διαφέρουν σημαντικά ανάμεσα στα υποκείμενα ή λόγω αριστερά-δεξιά θέσης”, οπότε για την εύρεση του ground truth, για κάθε ζεύγος κειμένων βρέθηκε ο μ.ο. των αντίστοιχων παρατηρήσεων, και κανονοποιήθηκε στην κλίμακα 0-1. Τέλος, οι συσχετίσεις μεταξύ των διαφορετικών ανθρώπινων βαθμολογιών είναι περίπου 0.6.

Παρακάτω παρουσιάζονται δύο ζευγάρια κειμένων. Το πρώτο είναι ένα παράδειγμα ομοιότητας με υψηλή συμφωνία μεταξύ ανάμεσα στους ανθρώπους, και το δεύτερο ένα παράδειγμα ανομοιότητας.

“The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist. (80 words)”
Queensland senator Andrew Bartlett has launched a last-minute bid to rescue the Australian Democrats from a split that threatens to destroy the party. With nominations for the party leadership to close on Wednesday night, Senator Bartlett met last night with deputy leader Aden Ridgeway to offer him a place on a unity ticket and set up a reform process to begin healing the party's wounds. Party sources said Senator Ridgeway, who turned against former leader Natasha Stott Despoja, is still expected to contest the leadership against one of her two supporters: Senator Bartlett or Brian Greig, installed as interim leader by the party's executive last Thursday. (105 words)”

Πίνακας 1. Παραδείγματα ομοιότητας κειμένων, με μ.ο. ανθρώπινων αξιολογήσεων 5.

“The European Parliament is spoiling for a fight with Israel. It has voted to review the EU's diplomatic links with the Jewish state, to impose an arms embargo and to threaten wider trade sanctions. Many MEPs want to go further and dispatch a European military force to the region in order to "protect the Palestinian people". (58 words)”
“Australia's Commonwealth Bank on Wednesday said it plans to cut about 1,000 jobs even as it reported its profit rose 11 percent last fiscal year. Workers reacted angrily to the planned cuts, which Australia's second largest bank said were designed to control costs. The cuts will take effect this financial year. The bank reported net profit of 2.66 billion Australian dollars ($1.4 billion) in the year to June 30, up from 2.4 billion Australian dollars in the previous year. (79 words)”

Πίνακας 2. Παραδείγματα ανομοιότητας κειμένων, με μ.ο. ανθρώπινων αξιολογήσεων 1.1.

4.2 Προεπεξεργασία Κειμένων

Για την χρήση των μεθόδων word embeddings, προηγήθηκαν βασικά βήματα προεπεξεργασίας του dataset, όπως λημματοποίηση, αφαίρεση stop-word, σημείων στίξης , αριθμών και κάποιων χαρακτήρων. Το WordNetLemmatizer χρησιμοποιήθηκε για την λημματοποίηση των λέξεων και για να φορτώσουμε τις stopwords της Αγγλικής Γλώσσας χρησιμοποιήσαμε την βιβλιοθήκη NLTK library.

Παρακάτω παρουσιάζονται τα βασικά στάδια επεξεργασίας ενός κειμένου που πραγματοποιήθηκαν. Στάδιο 1 : αρχικό κείμενο. Στάδιο 2 : μετατροπή του κειμένου σε πεζό, Στάδιο 3 : διαχωρισμός των λέξεων μεταξύ τους, και αφαίρεση περιττών χαρακτήρων. Στάδιο 3 : λημματοποίηση.

“Prince William has told friends his mother was right all along to suspect her former protection officer of spying on her and he doesn't want any detective intruding on his own privacy. William and Prince Harry are so devastated by the treachery of Ken Wharfe, whom they looked on as a surrogate father, they are now refusing to talk to their own detectives. (63 words)”
“prince william has told friends his mother was right all along to suspect her former protection officer of spying on her and he doesn't want any detective intruding on his own privacy. william and prince harry are so devastated by the treachery of ken wharfe, whom they looked on as a surrogate father, they are now refusing to talk to their own detectives.”
['prince', 'william', 'told', 'friends', 'mother', 'right', 'along', 'suspect', 'former', 'protection', 'officer', 'spying', 'want', 'detective', 'intruding', 'privacy', 'william', 'prince', 'harry', 'devastated', 'treachery', 'ken', 'wharfe', 'looked', 'surrogate', 'father', 'refusing', 'talk', 'detectives']
['prince', 'william', 'tell', 'friend', 'mother', 'right', 'along', 'suspect', 'former', 'protection', 'officer', 'spy', 'want', 'detective', 'intrude', 'privacy', 'william', 'prince', 'harry', 'devastate', 'treachery', 'ken', 'wharfe', 'look', 'surrogate', 'father', 'refuse', 'talk', 'detective']

Πίνακας 2. Αναπαράσταση σταδίων επεξεργασίας κειμένων που χρειάστηκαν για την μελέτη των μεθόδων που βασίζονται σε Word Embeddings.

4.3 Μέτρο Ομοιότητας

Πειραματικά χρησιμοποιήσαμε και τα δύο μέτρα ομοιότητας που παρουσιάστηκαν στην εργασία, και στις μετρήσεις μας δεν είχαμε κάποια αισθητή διαφορά. Τελικά, για τις μετρήσεις που έγιναν χρησιμοποιηθηκε το μέτρο ομοιότητας συνημιτόνου.

4.4 Μέτρο Απόδοσης

Μια κοινά αποδεκτή προσέγγιση, για την αποτίμηση της ακρίβειας των μεθόδων σημασιολογικής ομοιότητας, είναι αυτός του συσχετισμού των αποτελεσμάτων τους με τις αντίστοιχες ανθρώπινες μετρήσεις στο ίδιο task. Η συνάρτηση συσχέτισης ποσοτικοποιεί την ισχύ των γραμμικών μονοτικών σχέσεων μεταξύ δύο χαρακτηριστικών.

Στην παρούσα εργασία, υιοθετούμε τον συντελεστή συσχέτισης Pearson, ως μέτρο της δύναμης της σχέσης ανάμεσα στις ανθρώπινες αξιολογήσεις της ομοιότητας και τις υπολογιστικές τιμές. Ο συντελεστής συσχέτισης Pearson, ορίζεται ως :

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

,όπου X, Y τυχαίες μεταβλητές, με αναμενόμενες μεταβλητές μ_X, μ_Y και τυπικές αποκλίσεις σ_X , σ_Y.

Στην παρούσα εργασία, υπολογίζουμε την συσχέτιση Pearson ανάμεσα στις ανθρώπινες μετρήσεις και τις ομοιότητες συνημιτόνου των διανυσματικών αναπαραστάσεων, για κάθε έγγραφο και βρίσκουμε τον μ.ο. .

Ένα μέτρο ομοιότητας αναγνωρίζεται ότι έχει καλύτερη απόδοση εάν έχει υψηλότερη βαθμολογία συσχέτισης (όσο πιο κοντά στο 1.0 τόσο πιο καλά) με τις ανθρώπινες μετρήσεις, ενώ αναγνωρίζεται ότι δεν σχετίζεται με την ανθρώπινη εκτίμηση εάν η συσχέτιση είναι 0.

4.5 Πειραματικά Αποτελέσματα και Ανάλυση

Τα συγκεντρωτικά πειραματικά αποτελέσματα για κάθε μέθοδο που παρουσιάστηκε:

ΜΕΘΟΔΟΣ		**PEARSON C.C.**
tf-idf vectors		0.69860
avg. word embed	*word2vec*	**0.74974**
	glove	0.70909
	fasttext	0.70306
sif		0.70137
wmd		0.71030
InferSent		0.68134
USE	*Transformer model*	**0.82307**
	DAN model	0.78589
SentenceBERT	*'stsb-distilbert-base'*	0.68752
	''stsb-bert-base'	**0.70488**
	'stsb-roberta-base'	0.66598

Πίνακας 4. Συγκεντρωτικός Πίνακας Αποτελεσμάτων

Γενικά Σχόλια :

Το πρώτο μισό του πίνακα αφορά μεθόδους που βασίζονται στην χρήση word embeddings, και το δεύτερο μισό στην χρήση sentence embeddings.

Αρχικά, για τα word embeddings, παρατηρούμε ότι καμία από τις μεθόδους δεν έχει ιδιαίτερα κακή συσχέτιση και όλες βρίσκονται στην ίδια κλίμακα απόδοσης.

Την χαμηλότερη απόδοση έχουν τα tf-idf διανύσματα και την αισθητά καλύτερα ο μ.ο. των Word2Vec embeddings.

Περί των tf-idf διανυσμάτων, αν και πρόκειται για μία baseline προσέγγιση, φαίνεται να μπορούν να δώσουν χρήσιμα αποτελέσματα. Η καλή αυτή απόδοση, πιθανώς

οφείλεται στην φύση του ειδησεογραφικού λεξιλογίου του dataset μας, που είναι αρκετά αυστηρό, κυριολεκτικό, μικρό και περιεκτικό, και περιέχει αρκετά ονόματα.

Για την πολύ καλή απόδοση των Word2Vec embeddings, μπορούμε να εικάσουμε εύκολα, πως οφείλεται στο γεγονός πως το μοντέλο που χρησιμοποιούμε είναι εκπαιδευμένο σε ειδησεογραφικές ειδήσεις, και αρά τα embeddings περιέχουν πληροφορία πιο πλούσια σημασιολογικά για το θέμα μας .

Τέλος, οι προσεγγίσεις SIF και WMD, δίνουν μία καλή απόδοση επίσης, ξεπερνώντας ακόμη και τις μεθόδους για την δημιουργία sentence embeddings.

Οι sentence embeddings προσεγγίσεις με την σειρά τους, διαφέρουν αρκετά μεταξύ τους στην απόδοση. Το InferSent, έχει την χαμηλότερη απόδοση συνολικά, το SentenceBERT αν και υπόσχεται state-of-the-art αποτελέσματα, δίνει μία μέτρια απόδοση, ενώ το USE, και τα δύο μοντέλα του δίνουν απόδοση που ξεπερνά όλες τις άλλες μεθόδους, και μπορεί θεωρηθεί περισσότερο από ικανοποιητική από ένα προ-εκπαιδευμένο μοντέλο.

Η αισθητά καλύτερη απόδοση του USE πιθανώς να συμβαίνει γιατί έχει χρησιμοποιήσει πολύ μεγαλύτερη βάση από datasets, και αρκετά διαφορετικά μεταξύ τους, για να εκπαιδευτεί, τόσο supervised όσο και supervised, και συγχρόνως επωφελείται της τεχνολογίας Transformer.

Τέλος, καμία από τις μεθόδους δεν έχει λάβει υπόψη το corpus που χρησιμοποιείται. Το τελευταίο βήμα της μελέτης μας, είναι λοιπόν να πάρουμε ένα μοντέλο και να δούμε πως αποδίδει όταν εκπαιδεύεται πάνω στο corpus μας.

Διαλέξαμε το SentenceBERT να κάνουμε fine-tuning, ώστε να μάθει καλύτερα βάρη το μοντέλο. Η διαδικασία παρουσιάζεται παρακάτω.

4.6 Fine-Tuning SentenceBERT

i. Μέθοδος

Η εργασία του Lee συνοδεύεται και από ένα δεύτερο dataset με επιπλέον 314 ειδήσεις. Αυτό το αρχείο χρησιμοποιήθηκε για να κάνουμε fine-tuning το μοντέλο SentenceBERT.

Βήμα 1 - Training Data :

Αρχικά, πρέπει να εισάγουμε στο μοντέλο μας τα training data.

Για' αυτό χρησιμοποιούμε την κλάση `InputExample`, για να αποθηκεύσουμε τα παραδείγματα προς εκπαίδευση

Βήμα 2 - Loss function :

Για να βελτιστοποιήσουμε το δίκτυό μας, πρέπει κάπως να πούμε στο δίκτυό μας ποια ζεύγη προτάσεων είναι παρόμοια και πρέπει να είναι κοντά στο διανυσματικό χώρο και ποια ζεύγη είναι ανόμοια και πρέπει να βρίσκονται πολύ μακριά στο διανυσματικό χώρο.

Η ομοιότητα αυτών των embeddings υπολογίζεται χρησιμοποιώντας ομοιότητα συνημίτονο και το αποτέλεσμα συγκρίνεται με το gold similarity score. Αυτό επιτρέπει στο δίκτυό μας να τελειοποιηθεί και να αναγνωρίσει την ομοιότητα των προτάσεων.

Χρησιμοποιήσαμε την συνάρτηση `tf.keras.losses.CosineSimilarity`.

Βήμα 3 - Fit the model :

```
sbert_model.fit(train_objectives=[(train_dataloader, train_loss)], epochs=1,
warmup_steps=100)
```

Εκπαιδεύουμε το μοντέλο καλώντας `model.fit()`.

ii. Αποτελέσματα

Με το fine-tuning του SentenceBert, πετυχαίνουμε απόδοση 0.91098, που είναι και η υψηλότερη απόδοση που μετρήσαμε από όλα τα πειράματα που διεξήγαμε.

Μπορούμε να παρατηρήσουμε, πως αν και ορισμένα pre-trained μοντέλα δίνουν πολύ υψηλές αποδόσεις, δεν μπορούμε να δημιουργήσουμε μοντέλα που να αντιλαμβάνονται στο μέγιστο δυνατό την σημασιολογία ενός κειμένου, εάν δεν ληφθεί υπόψη το ίδιο το κείμενο από το μοντέλο.

ΚΕΦΑΛΑΙΟ 5

Επίλογος

Στο κεφάλαιο αυτό συγκεντρώνονται τα συμπεράσματα τα οποία εξάγονται από τα κεφάλαια που έχουν προηγηθεί στην ενότητα 5.1. Έπειτα, στην ενότητα 7.2, προτείνονται δυνατές επεκτάσεις στο ερευνητικό κομμάτι που θα μπορούσαν να διεξαχθούν.

5.1 Συμπεράσματα

Στην παρούσα εργασία κάναμε μία σύγκριση ανάμεσα σε διαφορετικούς τρόπους προσεγγίσεις, για την εύρεση της σημασιολογικής ομοιότητας ανάμεσα σε δύο μικρά ειδησεογραφικά κείμενα. Χρησιμοποιήσαμε embeddings λέξεων και είδαμε πως μπορούμε να τα συνδυάσουμε για να πάρουμε το νόημα μιας ολόκληρης πρότασης καθώς και προ-εκπαιδευμένους κωδικοποιητές προτάσεων, για να δούμε πιο μοντέλο μπορεί να κωδικοποιησει καλύτερα το νόημα μιας πρότασης που δεν έχει ξαναδεί. Ο τρόπος αξιολόγησης έγινε μέσω του συντελεστή συσχέτισης Pearson. Όλες οι μέθοδοι, από τις πιο απλές, ως τις πιο σύγχρονες έδειξαν πολύ καλά αποτελέσματα.

5.2 Δυνατές Επεκτάσεις

Ενδιαφέρον παρουσιάζει η αξιολόγηση των παραπάνω μεθόδων για διαφορετικής φύσης dataset, πιο μεγάλα και πιο σύνθετα νοηματικά. Καθώς, το dataset που χρησιμοποιούμε, ήταν πολύ αυστηρό λεξιλογικά και περιεκτικό στην πληροφορία που μετέφερε, θα είχε ενδιαφέρον να εξετάσουμε πως οι αποδόσεις των παραπανω μοντέλων διαμορφώνονται για ένα λογοτεχνικό κείμενο, ή μια συλλογή από tweets.

Επιπλέον, στην παρούσα εργασία, είδαμε έναν τρόπο αξιοποίησης του μοντέλου BERT, μέσω του SentenceBERT, για την εκτίμηση της σημασιολογικής ομοιότητας. Ενδιαφέρον παρουσιάζει η εξερεύνηση και άλλων τρόπων αξιοποίησης του.

ΑΝΑΦΟΡΕΣ

Lee, M.D., Pincombe, B.M., & Welsh, M.B. (2005). A*n empirical evaluation of models of text document similarity*. In B.G. Bara, L.W. Barsalou & M. Bucciarelli, (Eds.), Proceedings of the 27th Annual Conference of the Cognitive Science Society, pp. 1254-1259. Mahwah, NJ: Erlbaum.URL : http://www.socsci.uci.edu/~mdlee/lee_pincombe_welsh_document.PDF

Pawar A. and Mago V. (2018). *Calculating the similarity between words and sentences using a lexical database and corpus statistics*. URL : https://arxiv.org/abs/1802.05667

Boom C.D. , Canneyt S.V., Bohez S., Demeester T. and Dhoedt B. (2015). *Learning Semantic Similarity for Very Short Texts*. URL : https://arxiv.org/abs/1512.00765

Conneau A. , Kiela D. , Schwenk H., Barrault L. and Bordes A. (2017). *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*. URL : https://arxiv.org/abs/1705.02364

Kushner M.J., Sun Y., Kolkin N.I. and Weinberger K.Q. (2015). *From Word Embeddings To Document Distances*. URL : http://proceedings.mlr.press/v37/kusnerb15.pdf

Arora S., Liang Y. and Ma T. (2017). *A Simple But Tough-To-Beat Baseline For Sentence Embeddings*. URL : https://openreview.net/pdf?id=SyK00v5xx

Cer D., Yang Y., Kong S., Hua N., Limtiaco N., St. John R, Constant N., Guajardo-Cespedes M., Yuan S., Tar S., Sung Y.H., Strope B. and Kurzweil R. (2018). *Universal Sentence Encoder*. URL : https://arxiv.org/abs/1803.11175

Reimers N. and Gurevych I. (2019). *Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks*. URL : https://arxiv.org/abs/1908.10084

Tversky A. (1977). *Features of Similarity*. Psychological Review, 84(4), 327–352. URL : https://psycnet.apa.org/record/1978-09287-001

Qua R., Fanga Y., Baib W. and Jiang J (2018). *Computing semantic similarity based on novel models of semantic representation using Wikipedia*.URL : https://www.sciencedirect.com/science/article/abs/pii/S0306457317309226?via%3Dihub

Pirró G. (2009). *A semantic similarity metric combining features and intrinsic information content*. URL : https://www.sciencedirect.com/journal/data-and-knowledge-engineering

Sánchez D., Batet M and Isern D. (2011). *Ontology-based information content computation*. URL : https://www.sciencedirect.com/journal/knowledge-based-systems

Guan N., Song D. amd Liao L. (2019). *Knowledge graph embedding with concepts*.

Harispe S. , Sánchez D., Ranwez S., Janaqi S. and Montmain J. (2013). *A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain*. URL : https://www.sciencedirect.com/journal/journal-of-biomedical-informatics

Zhu G. and Iglesias C.A. (2018). *Exploiting semantic similarity for named entity disambiguation in knowledge graphs*. URL : https://www.sciencedirect.com/science/article/pii/S0957417418300897?via%3Dihub

Zhu G. and Iglesias C.A. (2017). *Computing Semantic Similarity of Concepts in Knowledge Graphs*. URL : https://ieeexplore.ieee.org/document/7572993

Kiros R., Zhu Y., Salakhutdinov R., Zemel R.S. , Torralba A., Urtasun R. and Fidler S. (2015). *Skip-Thought Vectors*. URL : https://arxiv.org/abs/1506.06726

Pennington J., Socher R. and Manning C.D. (2014). *GloVe: Global Vectors forWord Representation*. URL : https://nlp.stanford.edu/pubs/glove.pdf

Mikolov T., Chen K., Corrado G. and Dean J. (2013). *Efficient Estimation of Word Representations in Vector Space*. URL : https://arxiv.org/abs/1301.3781

Bojanowski P., Grave E., Joulin A. and Mikolov T. (2016) Enriching Word Vectors with Subword Information. URL : https://arxiv.org/abs/1607.04606

Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, (2016). *Bag of Tricks for Efficient Text Classification*. URL : https://arxiv.org/abs/1607.01759

Collobert R. and Weston J. (2008). *A unified architecture for natural language processing: deep neural networks with multitask learning*. URL : https://dl.acm.org/doi/10.1145/1390156.1390177

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013). *Distributed Representations ofWords and Phrases and their Compositionality*. URL : https://arxiv.org/abs/1310.4546

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017). *Attention Is All You Need*. URL : https://arxiv.org/abs/1706.03762

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. URL : https://arxiv.org/abs/1810.04805

Emerson G. (2020). *What are the Goals of Distributional Semantics?* URL : https://arxiv.org/abs/2005.02982

Mikolov T., Yih W. and Zweig G. (2013). *Linguistic Regularities in Continuous Space Word Representations*. URL : https://www.aclweb.org/anthology/N13-1090/

Jonas Mueller and Aditya Thyagarajan (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. URL : https://dl.acm.org/doi/10.5555/3016100.3016291

Perone C.S. , Silveira R. and Paula T.S. (2018). *Evaluation of sentence embeddings in downstream and linguistic probing tasks.* URL : https://arxiv.org/abs/1806.06259

ΠΑΡΑΡΤΗΜΑ

Σε αυτό το παράρτημα παρατίθεται το dataset με τα 50 μικρά κείμενα που χρησιμοποιήθηκαν πειραματικά.

1. The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist. (80 words)

2. Cash-strapped financial services group AMP has shelved a $400 million plan to buy shares back from investors and will raise $750 million in fresh capital after profits crashed in the six months to June 30. Chief executive Paul Batchelor said the result was "solid" in what he described as the worst conditions for stock markets in 20 years. AMP's half-year profit sank 25 per cent to $303 million, or 27c a share, as Australia's largest investor and fund manager failed to hit projected 5 per cent earnings growth targets and was battered by falling returns on share markets. (98 words)

3. The United States government has said it wants to see President Robert Mugabe removed from power and that it is working with the Zimbabwean opposition to bring about a change of administration. As scores of white farmers went into hiding to escape a round-up by Zimbabwean police, a senior Bush administration official called Mr Mugabe's rule "illegitimate and irrational" and said that his re-election as president in March was won through fraud. Walter Kansteiner, the assistant secretary of state for African affairs, went on to blame Mr Mugabe's policies for contributing to the threat of famine in Zimbabwe. (98 words)

4. A radical armed Islamist group with ties to Tehran and Baghdad has helped al-Qaida establish an international terrorist training camp in northern Iraq, Kurdish officials say. Intelligence officers in the autonomous Kurdish region of Iraq told the Guardian that the Ansar al-Islam (supporters of Islam) group is harbouring up to 150 al-Qaida members in a string of villages it controls along the Iraq-Iran border. Most of them fled Afghanistan after the US-led offensive, but officials from the Patriotic Union of Kurdistan (PUK), which controls part of north-east Iraq, claim an "abnormal" number of recruits are making their way to the area from Jordan, Syria and Egypt. (106 words)

5. Washington has sharply rebuked Russia over bombings of Georgian villages, warning the raids violated Georgian sovereignty and could worsen tensions between Moscow and Tbilisi. "The United States regrets the loss of life and deplores the violation of Georgia's sovereignty," White House spokesman Ari Fleischer said. Mr Fleischer said US Secretary of State Colin Powell had delivered the same message to his Russian counterpart but that the stern language did not reflect a sign of souring relations between Moscow and Washington. (80 words)

6. A gay former student of a Melbourne Christian school is taking legal action under equal opportunity legislation, claiming the school discriminated against him because of his sexuality. Tim, 16, alleged a staff member at Hillcrest Christian College in Berwick told him he "had the devil in him", and constant bullying by students prompted the principal to tell him to hide his sexuality. He left the school several weeks ago and is continuing Year 10 by distance education after he said homophobic bullies threw rocks at his head, spat on him, called him names and slashed his belongings. (97 words)

7. Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden's terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa'ida, including bin Laden. The money enabled al-Qa'ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia. (97 words)

8. Palestinian hired gun Abu Nidal, whose violent death was reported last week from Baghdad, was murdered on the orders of Iraqi President Saddam Hussein after refusing to train al-Qa'ida fighters based in Iraq, reports said yesterday. Iraqi intelligence chief Taher Jalil Habbush said last Wednesday Abu Nidal had shot and killed himself after being discovered living illegally in Baghdad and facing interrogation for anti-Iraqi activities. But Western diplomats believe the radical militant was killed for refusing to reactivate his international terrorist network. (82 words)

9. Hunan province remained on high alert last night as thunderstorms threatened to exacerbate the flood crisis, now entering its fifth day and with 108 already dead and hundreds of thousands evacuated. On the flood frontline at Dongting Lake, the water level peaked at just under 35m on Saturday night, then eased about 3cm during the day under a hot sun, with temperatures reaching 35C. But with the lake still brimming at dangerously high levels, and spilling over the top of its banks in some places, locals were fearful that a thunderstorm and high winds forecast to hit the region last night would damage the dikes. About 1800km of dikes around the lake are all that stand between 10 million people in the surrounding farmland and disaster. (126 words)

10. A U.S.-British air raid in southern Iraq left eight civilians dead and nine wounded, the Iraqi military said Sunday. The military told the official Iraqi News Agency that the warplanes bombed areas in Basra province, 330 miles south of Baghdad. The U.S. Central Command in Florida said coalition aircraft used precision-guided weapons to strike two air defense radar systems near Basra "in response to recent Iraqi hostile acts against coalition aircraft monitoring the Southern No-Fly Zone." (76 words)

11. Iraq and Russia are close to signing a $40 billion economic cooperation plan, Iraq's ambassador said Saturday, a deal that could put Moscow at odds with the United States as it considers a military attack against Baghdad. The statement by Ambassador Abbas Khalaf came amid indications that Russia, despite its strong support for the

post-Sept. 11 antiterrorism coalition, is maintaining or improving ties with Iran and North Korea, which together with Iraq are the countries President Bush has labeled the "axis of evil." (83 words)

12. U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials. (67 words)

13. Drug squad detectives have asked the Police Ombudsman to investigate the taskforce that is examining allegations of widespread corruption within the squad. This coincides with the creation of a special unit within the taskforce to track the spending of at least 10 serving and former squad members. The corruption taskforce, codenamed Ceja, will check tax records and financial statements in a bid to establish if any of the suspects have accrued unexplained wealth over the past seven years. But drug squad detectives have countered with their own set of allegations, complaining to the ombudsman that the internal investigation is flawed, biased and over-zealous. (103 words)

14. Queensland senator Andrew Bartlett has launched a last-minutebid to rescue the Australian Democrats from a split that threatens to destroy the party. With nominations for the party leadership to close on Wednesday night, Senator Bartlett met last night with deputy leader Aden Ridgeway to offer him a place on a unity ticket and set up a reform process to begin healing the party's wounds. Party sources said Senator Ridgeway, who turned against former leader Natasha Stott Despoja, is still expected to contest the leadership against one of her two supporters: Senator Bartlett or Brian Greig, installed as interim leader by the party's executive last Thursday. (105 words)

15. Very few women have been appointed to head independent schools, thwarting efforts to show women as good leaders, according to the Victorian Independent Education Union. Although they make up two-thirds of teaching staff, women hold only one-third of principal positions, the union's general secretary, Tony Keenan, said. He believed some women were reluctant to become principals because of the long hours and the nature of the work. But in other cases they were shut out of the top position because of perceptions about their ability to lead and provide discipline. (90 words)

16. The Bush administration has drawn up plans to escalate the war of words against Iraq, with new campaigns to step up pressure on Baghdad and rally world opinion behind the US drive to oust President Saddam Hussein. This week, the State Department will begin mobilising Iraqis from across North America, Europe and the Arab world, training them to appear on talk shows, write opinion articles and give speeches on reasons to end President Saddam's rule. (75 words)

17. Beijing has abruptly withdrawn a new car registration system after drivers demonstrated "an unhealthy fixation" with symbols of Western military and industrial

strength - such as FBI and 007. Senior officials have been infuriated by a popular demonstration of interest in American institutions such as the FBI. Particularly galling was one man's choice of TMD, which stands for Theatre Missile Defence, a US-designed missile system that is regularly vilified by Chinese propaganda channels. (73 words)

18. The United Nations World Food Program estimates that up to 14 million people in seven countries - Malawi, Mozambique, Zambia, Angola, Swaziland, Lesotho and Zimbabwe - face death by starvation unless there is a massive international response. In Malawi, as many as 10,000 people may have already died. The signs of malnutrition - swollen stomachs, stick-thin arms, light-coloured hair - are everywhere. (62 words)

19. In Malawi, as in other countries in the region, AIDS is making the effects of the famine much worse. The overall HIV infection rate in Malawi is 19 per cent, but in some areas up to 35 percent of people are infected. A significant proportion of the young adult population is too sick to do any productive work. Malnutrition causes people to succumb to the disease much more quickly than they do in the West, and hunger forces women into prostitution in order to feed their families, making them more vulnerable to contracting the disease. Life expectancy has been reduced to 40 years. (103 words)

20. The United Nations was determined that its showpiece environment summit - the biggest conference the world has ever witnessed - should be staged in Africa. The venue, however, could not be further removed from the grim realities of life in the rest of Africa. Johannesburg's exclusive and formerly whites-only suburb of Sandton is the wealthiest neighbourhood in the continent. Just a few kilometres from Sandton begins the sprawling Alexandra township, where nearly a million people live in squalor. Organisers of the conference, which begins today, seem determined that the two worlds should be kept as far apart as possible. Tight security surrounds Sandton's convention centre and five-star hotels, where world leaders will debate poverty, the environment and sustainable development while enjoying lavish hospitality. (122 words)

21. The Iraqi capital is agog after the violent death of one of the world's most notorious terrorists, but the least of the Palestinian diplomat's worries was the disposal of Abu Nidal's body, which lay on a slab in an undisclosed Baghdad morgue. Abu Nidal's Fatah Revolutionary Council is held responsible for the death or injury of almost 1000 people in 20 countries across Europe and the Middle East in the three decades since he fell out with Yasser Arafat over what Abu Nidal saw as Arafat's willingness to accommodate Israel in the Palestinian struggle.

22. The Federal Government says changes announced today to the work for the dole scheme will benefit participants and taxpayers. Federal Employment Services Minister Mal Brough says that from July 1 those taking part in work for the dole will be able to perform extra hours to complete their mutual obligation more quickly to access training credits. (61 words)

23. The biowarfare expert under scrutiny in the anthrax attacks declared, "I am not the anthrax killer," and lashed out today against Attorney General John Ashcroft for calling him a "person of interest" in the investigation. For the second time in two weeks,

the scientist went before a throng of reporters outside his lawyer's office to profess his innocence and decry the attention from law enforcers that he contends has destroyed his life. (72 words)

24. China said Sunday it issued new regulations controlling the export of missile technology, taking steps to ease U.S. concerns about transferring sensitive equipment to Middle East countries, particularly Iran. However, the new rules apparently do not ban outright the transfer of specific items - something Washington long has urged Beijing to do. (54 words)

25. Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry. (57 words)

26. An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing. (61 words)

27. How did 2,300 allegedly unregistered missile warheads come to be stored on a Canadian businessman's anti-terrorism training facility in New Mexico? U.S. and Canadian officials are still trying to figure that out, but one security expert says the mystery is a "chilling" one. David Hudak, 41, was arrested in the United States more than a week ago when, according to court documents, agents searching his property found the warheads stored in crates that were marked "Charge Demolition." (77 words)

28. The Saudi Interior Ministry on Sunday confirmed it is holding a 21-year-old Saudi man the FBI is seeking for alleged links to the Sept. 11 hijackers. Authorities are interrogating Saud Abdulaziz Saud al-Rasheed "and if it is proven that he was connected to terrorism, he will be referred to the sharia (Islamic) court," the official Saudi Press Agency quoted an unidentified ministry official as saying. (65 words)

29. Sri Lanka's government will lift a four-year ban on Tamil Tiger rebels on Sept. 6, paving the way for peace talks with the insurgents scheduled for later that month in Thailand, a government minister said Saturday. "We will lift the ban as promised," Minister for Rehabilitation Jayalath Jayawardena told The Associated Press. The lifting of the ban is one of the key rebel conditions for resuming peace negotiations with the government after a hiatus of more than seven years. (79 words)

30. A man accused of making hidden-camera footage up the skirts of women also made child pornography of the worst kind, featuring the rape of children as young as 6, police said Friday. The latest allegations suggest there's nothing humorous about voyeurs who some may perceive to be making secret videos as a joke, Staff-Insp. Gary Ellis said. "Approximately 20 per cent of voyeurs have also committed sexual assault or rape," Ellis said, reading from a recently released federal government report on criminal

voyeurism. (83 words)

31. Police are combing through videotapes trying to spot the gunman dressed in black who shot a 30-year-old man to death at a downtown massage parlour. The victim was hit in the stomach and upper body and died about 3 1/2 hours later in hospital. The woman was not hurt. Police urged business owners to turn over any security-camera videotapes they might have that recorded people on the street at the time. Several such videos are now being reviewed. (78 words)

32. The Federal Government did not regret its actions 12 months on from the Tampa asylum seeker crisis, Small Business Minister Joe Hockey said today. Mr Hockey said the Government was not embarrassed by the Tampa issue, which began on August 27 of last year when the captain of the Norwegian cargo ship rescued more than 400 asylum seekers from an Indonesian ferry north of Christmas Island. (66 words)

33. At least three Democrats are considering splitting from the party while no-one has yet nominated to contest the leadership. Three of the "gang of four" senators who ousted Natasha Stott Despoja from the leadership are considering forming a new "progressive centre" party in the fallout from last week's turmoil. This would leave the Democrats with a rump of three or four members. West Australian Senator Andrew Murray said yesterday unless the Democrats left wing gave ground the party would split. (80 words)

34. A young humpback whale remained tangled in a shark net off the Gold Coast yesterday, despite valiant efforts by marine rescuers. With its head snared by the net and an anchor rope wrapped around its tail, the stricken whale was still swimming but hopes for its survival were fading. A second rescue attempt was planned for dawn today after rescuers braved heavy seas, strong wind and driving rain to try to free the whale. (74 words)

35. Prince William has told friends his mother was right all along to suspect her former protection officer of spying on her and he doesn't want any detective intruding on his own privacy. William and Prince Harry are so devastated by the treachery of Ken Wharfe, whom they looked on as a surrogate father, they are now refusing to talk to their own detectives. (63 words)

36. The spectre of Osama bin Laden rose again today, urging Afghans to launch a new Jihad, or holy war, and predicting the fall of the United States, in a hand-written "letter" posted on an Islamic website. There was no hard proof that the scruffy missive was genuine, but IslamOnline.net said it had been received by their correspondent in Jalalabad, eastern Afghanistan, from an Afghan source who asked to remain anonymous. The source claimed it was the "most recent letter" from the world's most wanted man. (85 words)

37. The Johannesburg Earth Summit is set to get under way with the promise that leaders will take action on the environment, debt and poverty. South African President Thabo Mbeki, speaking at the opening ceremony, said: "Out of Johannesburg and out of Africa must emerge something that takes the world forward." But the absence of US

President George W Bush was threatening to overshadow the summit. (65 words)

38. Robert Mugabe strengthened his hold on the Zimbabwean government yesterday by retaining the most combative hardliner ministers in a cabinet shuffle which offered little hope of a moderation of the land seizures and other policies that have kept Zimbabwe in crisis and brought international condemnation. (51 words)

39. They dress in black and disguise their identities with bandannas and sunglasses. Their logo is an image of the Southern Cross constellation, superimposed with a pair of crossed boomerangs, which resembles a swastika. The Blackshirts are former husbands aggrieved by their treatment at the hands of their ex-wives and the courts, who regard themselves as the vanguard of a "men's rights" movement in Australia and say that their actions will be remembered as marking a turning-point in history. (78 words)

40. The real level of world inequality and environmental degradation may be far worse than official estimates, according to a leaked document prepared for the world's richest countries and seen by the Guardian. It includes new estimates that the world lost almost 10% of its forests in the past 10 years; that carbon dioxide emissions leading to global warming are expected to rise by 33% in rich countries and 100% in the rest of the world in the next 18 years; and that more than 30% more fresh water will be needed by 2020. (93 words)

41. Researchers conducting the most elaborate wild goose chase in history are digesting the news that a bird they have tracked for over 4,500 miles is about to be cooked. Kerry, an Irish light-bellied Brent goose, was one of six birds tagged in Northern Ireland in May by researchers monitoring the species' remarkable migration. Last week, however, he was found dead in an Inuit hunter's freezer in Canada, still wearing his £3,000 satelite tracking device. Kerry was discovered by researchers on the remote Cornwallis Island. They picked up the signal and decided to try to find him. (96 words)

42. Russia defended itself against U.S. criticism of its economic ties with countries like Iraq, saying attempts to mix business and ideology were misguided. "Mixing ideology with economic ties, which was characteristic of the Cold War that Russia and the United States worked to end, is a thing of the past," Russian Foreign Ministry spokesman Boris Malakhov said Saturday, reacting to U.S. Defense Secretary Donald Rumsfeld's statement that Moscow's economic relationships with such countries sends a negative signal. (77 words)

43. Pope John Paul II urged delegates at a major U.N. summit on sustainable growth on Sunday to pursue development that protects the environment and social justice. In comments to tourists and the faithful at his summer residence southeast of Rome, the pope said God had put humans on Earth to be his administrators of the land, "to cultivate it and take care of it." "In a world ever more interdependent, peace, justice and the safekeeping of creation cannot but be the fruit of a joint commitment of all in pursuing the common good," John Paul said. (96 words)

44. The Russian defense minister said residents shouldn't feel threatened by the

growing number of Chinese workers seeking employment in the country's sparsely populated Far Eastern and Siberian regions. There are no exact figures for the number of Chinese working in Russia, but estimates range from 200,000 to as many as 5 million. Most are in the Russian Far East, where they arrive with legitimate work visas to do seasonal work on Russia's low-tech, labor-intensive farms. (75 words)

45. Australian spies listened to conversations between Norway's ambassador and its foreign office during the Tampa crisis, a soon to be published book will reveal. Phone calls were tapped by the Defence Signals Directorate when Norwegian ambassador Ove Thorsheim visited the freighter during the stand-off. A book, Tampa, to be published in Norway in October, recounts the events which triggered Australia's Pacific Solution and transformed Tampa Captain Arne Rinnan into a homeland hero. (72 words)

46. Batasuna, a political party that campaigns for an independent Basque state, faces a double blow today: the Spanish parliament is expected to vote overwhelmingly in favour of banning the radical group, while a senior investigative judge is poised to suspend Batasuna's activities on the grounds that they benefit Eta, the outlawed Basque separatist group. (56 words)

47. The river Elbe surged to an all-time record high Friday, flooding more districts of the historic city of Dresden as authorities scrambled to evacuate tens of thousands of residents in the worst flooding to hit central Europe in memory. In the Czech Republic, authorities were counting the cost of the massive flooding as people returned to the homes and the Vlava river receded, revealing the full extent of the damage to lives and landmarks. (74 words)

48. The European Parliament is spoiling for a fight with Israel. It has voted to review the EU's diplomatic links with the Jewish state, to impose an arms embargo and to threaten wider trade sanctions. Many MEPs want to go further and dispatch a European military force to the region in order to "protect the Palestinian people". (58 words)

49. Australia's Commonwealth Bank on Wednesday said it plans to cut about 1,000 jobs even as it reported its profit rose 11 percent last fiscal year. Workers reacted angrily to the planned cuts, which Australia's second largest bank said were designed to control costs. The cuts will take effect this financial year. The bank reported net profit of 2.66 billion Australian dollars ($1.4 billion) in the year to June 30, up from 2.4 billion Australian dollars in the previous year. (79 words)

50. Labor needed to distinguish itself from the Government on the issue of asylum seekers, Greens leader Bob Brown has said. His Senate colleague Kerry Nettle intends to move a motion today - on the first anniversary of the Tampa crisis - condemning the Government over its refugee policy and calling for an end to mandatory detention. "We Greens want to bring the Government to book over its serial breach of international obligations as far as asylum seekers in this country are concerned," Senator Brown said today. (86 words)