Τμήμα Πληροφορικής
Μεταπτυχιακό Δίπλωμα Ειδίκευσης
(Master of Science)

Διπλωματική Εργασία
(Master Thesis)

“Διανυσματικές Παραστάσεις Λέξεων που λαμβάνουν υπ' όψιν τη μορφολογία”

“Morphology Aware Word Embeddings”

Αναστάσιος Μαλινάκης

Φεβρουάριος 2016

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους γονείς μου οι οποίοι με στήριξαν στη διάρκεια του Μεταπτυχιακού Προγράμματος. Επίσης ευχαριστώ τον αναπληρωτή καθηγητή κ. Ίωνα Ανδρουτσόπουλο, ο οποίος με εισήγαγε στον πολύ ενδιαφέροντα κλάδο της Επεξεργασίας Φυσικής Γλώσσας. Εντέλει ευχαριστώ όλους τους καθηγητές και συμφοιτητές μου για τις εμπειρίες και τις γνώσεις που αποκόμισα τον προηγούμενο ενάμιση χρόνο. Θα ήθελα να ευχαριστήσω ιδιαίτερα το συμφοιτητή μου Θωμά Ασίκη, ο οποίος με βοήθησε σε μία μέθοδο αξιολόγησης στη διάρκεια της δικής του Διπλωματικής Εργασίας. Εντέλει, θα ήθελα να ευχαριστήσω τη φίλη μου, Χρύσα Μαυράκη, για την υποστήριξη και τη βοήθεια που μου πρόσφερε κατά την προετοιμασία της τελικής παρουσίασης της Διπλωματικής Εργασίας.

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1. Εισαγωγή

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Ένα βασικό πρόβλημα κατά την Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ) είναι η παράσταση της σημασίας ή / και άλλων χαρακτηριστικών των λέξεων σε μορφή κατανοητή από τον Ηλεκτρονικό Υπολογιστή. Τα περισσότερα συστήματα ΕΦΓ (π.χ. Επισημείωσης Μερών του Λόγου, Ανάλυσης Συναισθήματος κ.ά.) χρειάζονται τέτοιες παραστάσεις λέξεων. Γίνεται επομένως ευνόητο ότι η χρήση καλών παραστάσεων λέξεων μπορεί να οδηγήσει σε βελτίωση άλλων αλγορίθμων που τις χρησιμοποιούν.

Ένας από τους τρόπους παράστασης είναι οι “Διανυσματικές Παραστάσεις Λέξεων” (ΔΠΛ - word embeddings). Στην περίπτωση αυτή, κάθε λέξη αντιπροσωπεύεται από ένα αντίστοιχο διάνυσμα σταθερού πλήθους διαστάσεων, το οποίο περιέχει πραγματικούς αριθμούς. Ορισμένοι από τους αλγορίθμους παραγωγής τέτοιων παραστάσεων περιγράφονται στο Κεφάλαιο 2.

Η παρούσα Διπλωματική Εργασία πραγματεύεται εναλλακτικούς αλγορίθμους παραγωγής ΔΠΛ, λαμβάνοντας υπόψιν τη μορφολογία της εκάστοτε λέξης. Σε αντίθεση με διαδεδομένες προσεγγίσεις, οι οποίες θεωρούν κάθε λέξη ως αδιάσπαστη μονάδα, η παρούσα προσέγγιση αναγνωρίζει ότι οι λέξεις είναι συναθροίσεις μορφημάτων και ότι η παράσταση κάθε μίας λέξης πρέπει να είναι συνάρτηση των μορφημάτων που την αποτελούν. Για παράδειγμα, μία λέξη όπως η “κατασκηνώσαμε” μπορεί να διασπαστεί στα μορφήματα “κατα”, “σκην” και “ώσαμε”, ώστε να συναθροιστεί η πληροφορία των μορφημάτων κατά το σχηματισμό της τελικής ΔΠΛ, σε αντίθεση με διαδεδομένες προσεγγίσεις ΔΠΛ οι οποίες αγνοούν τα μορφήματα. Μεταξύ άλλων πλεονεκτημάτων, η χρήση των μορφημάτων επιτρέπει να παραχθούν καλύτερες διανυσματικές παραστάσεις σπάνιων λέξεων, των οποίων τα μορφήματα παρουσιάζονται πολύ συχνότερα σε κείμενα ως συνιστώσες άλλων λέξεων. Τα πειράματα έχουν διεξαχθεί σε κείμενα της Αγγλικής και της Ελληνικής γλώσσας.

Οι παραγόμενες ΔΠΛ αξιολογούνται με εσωτερικά (intrinsic) και εξωτερικά (extrinsic) μέτρα. Εσωτερικά μέτρα είναι οι επιδόσεις που επιτυγχάνουν σε σύνολα δεδομένων ομοιότητας λέξεων (word similarity) και αναλογίας λέξεων (word analogy). Εξωτερικά μέτρα θεωρούνται τα αποτελέσματα αλγορίθμων ΕΦΓ, οι οποίοι χρησιμοποιούν τις ΔΠΛ. Η περιπλοκή (perplexity) ενός Γλωσσικού Μοντέλου (Language Model) που χρησιμοποιεί ΔΠΛ είναι ένα παράδειγμα εξωτερικού μέτρου, καθώς επίσης και η ορθότητα (accuracy) στην Επισημείωση Μερών του Λόγου (ΕΜΛ – PoS Tagging), όταν η ΕΜΛ βασίζεται σε ΔΠΛ.

1.2 Κίνητρο

Οι ΔΠΛ αποτελούν βασικό δομικό στοιχείο πολλών σύγχρονων αλγορίθμων ΕΦΓ. Σε μορφολογικά πλούσιες γλώσσες, όπως τα Ελληνικά, συχνά δεν υπάρχουν αρκετές εμφανίσεις όλων των τύπων

(μορφών) των λέξεων (π.χ. “κατασκήνωσαν”, “κατασκηνώσαμε”) ακόμα και σε μεγάλα σώματα κειμένων κι έτσι ενδεχομένως να μην παράγονται ορθές διανυσματικές παραστάσεις όλων των τύπων, όταν δε λαμβάνονται υπόψη τα μορφήματά τους. Αυτό έχει ως αντίκτυπο οι αλγόριθμοι που χρησιμοποιούν τις ΔΠΛ να μην πετυχαίνουν ενδεχομένως τα καλύτερα δυνατά αποτελέσματα (εξωτερική αξιολόγηση). Επίσης, σε εσωτερικές (intrinsic) αξιολογήσεις, όπως στην ομοιότητα και στην αναλογία λέξεων, οι διανυσματικές παραστάσεις που δε λαμβάνουν υπόψιν τη μορφολογία δεν πετυχαίνουν πάντα καλά αποτελέσματα για σπάνιες λέξεις, όπως δείχνουν πειραματικά αποτελέσματα σε σύνολα σπάνιων λέξεων [Luong κ.ά. 2013].

Όπως προαναφέρθηκε, το πρόβλημα γίνεται πιο εμφανές σε μορφολογικά πλούσιες γλώσσες, όπως τα Ελληνικά, όπου μία λέξη σχεδόν πάντα έχει διαφορετική μορφή ανάλογα με την πτώση, το χρόνο, τον αριθμό κ.ο.κ. Σε τέτοιες περιπτώσεις, η μορφολογική ανάλυση βοηθά στην εξαγωγή διανυσματικών παραστάσεων ικανών να απομονώσουν την κυρίως σημασία των λέξεων, η οποία συνήθως εκφράζεται με το θέμα (stem), από τις κλιτικές παραλλαγές, οι οποίες συνήθως εκφράζονται με καταλήξεις στα Ελληνικά.

1.3 Συνεισφορά

Όπως προαναφέρθηκε, η παραγωγή ποιοτικών ΔΠΛ είναι πολύ σημαντική για τη μετέπειτα χρήση τους από άλλους αλγορίθμους ΕΦΓ. Η παρούσα εργασία μελετάει την παραγωγή ΔΠΛ λαμβάνοντας υπόψιν τη μορφολογία, περιορίζοντας τα μορφήματα μόνο σε θέματα και καταλήξεις σε αντίθεση με τις περισσότερες ως τώρα προσεγγίσεις. Μία ακόμη διαφοροποίηση είναι η εστίαση σε συλλαβές, ώστε να εξακριβωθεί αν μπορούν να χρησιμοποιηθούν αντί για μορφήματα. Επίσης, μελετάται η επίδραση της επιλογής του πλήθους των διαστάσεων των ΔΠΛ των θεμάτων και των καταλήξεων κατά την εσωτερική (intrinsic) αξιολόγηση.

Ακόμη, αναπτύχθηκε ένα νέο Ανατροφοδοτούμενο Νευρωνικό Δίκτυο (Recurrent Neural Network) για Γλωσσική Μοντελοποίηση το οποίο δέχεται ως είσοδο μορφήματα αντί για ολόκληρες λέξεις, με σκοπό τη βελτίωση της ταχύτητας εκπαίδευσης και ενδεχομένως της περιπλοκής.

Αν και τα αποτελέσματα των πειραμάτων δεν ήταν τα καλύτερα δυνατά, υπάρχει πολύς χώρος για βελτίωση και περαιτέρω μελέτη σε μελλοντικές εργασίες.

1.4 Διάρθρωση

Η Διάρθρωση της Διπλωματικής Εργασίας έχει ως εξής:

- **Κεφάλαιο 1: Εισαγωγή**. Περιληπτική παρουσίαση της εργασίας, των στόχων και της συνεισφοράς αυτής.
- **Κεφάλαιο 2: Υπόβαθρο**. Σύντομη παρουσίαση των τεχνικών γνώσεων που απαιτούνται για την κατανόηση της εργασίας. Επεξήγηση της μορφολογίας, παρουσίαση των Τεχνητών Νευρωνικών Δικτύων, ορισμένων συναρτήσεων οι οποίες χρησιμοποιήθηκαν, αλγορίθμων

παραγωγής ΔΠΛ κι εντέλει σύντομη παρουσίαση σχετικής προηγούμενης έρευνας.

- **Κεφάλαιο 3: Νέες μέθοδοι παραγωγής ΔΠΛ λαμβάνοντας υπόψιν τη μορφολογία**. Παρουσίαση των νέων αλγορίθμων οι οποίοι αναπτύχθηκαν στα πλαίσια της εργασίας για την εξαγωγή ΔΠΛ λαμβάνοντας υπόψιν τη μορφολογία των λέξεων.
- **Κεφάλαιο 4: Τρόποι Αξιολόγησης**. Παρουσίαση των μεθόδων αξιολόγησης των παραγόμενων ΔΠΛ.
- **Κεφάλαιο 5: Πειράματα**. Παρουσίαση των πειραμάτων τα οποία διενεργήθηκαν στη διάρκεια της εργασίας. Πιο συγκεκριμένα, παρουσίαση των δεδομένων, της προεπεξεργασίας, της διαδικασίας εκπαίδευσης και της διαδικασίας αξιολόγησης κι εντέλει των αποτελεσμάτων που προέκυψαν.
- **Κεφάλαιο 6: Συμπεράσματα και μελλοντική έρευνα**. Σύνοψη των συμπερασμάτων και των αποτελεσμάτων της εργασίας και παρουσίαση πιθανών βελτιώσεων για μελλοντική έρευνα.

2. Υπόβαθρο

2.1 Μορφολογία

Μορφολογία είναι η ανάλυση λέξεων στα συστατικά – μορφήματά τους. Για παράδειγμα, η μορφολογική ανάλυση της λέξης “ξανακατασκηνώσαμε” είναι ξανα-κατα-σκην-ώσαμε.

Η μορφολογία διαχωρίζεται σε δύο υποκατηγορίες:

- Παραγωγική Μορφολογία (Derivational Morphology), η οποία επικεντρώνεται στην παραγωγή νέων λέξεων (π.χ. ρημάτων από ουσιαστικά ή αντίστροφα, προσθήκη πρόθεσης σε ρήμα). Για παράδειγμα, στη λέξη “ξανακατασκηνώσαμε”, το πρόθεμα “ξανα-” δηλώνει επανάληψη κι επομένως ο αναγνώστης αντιλαμβάνεται ότι η ενέργεια “κατασηκνώσαμε” είχε επαναληφθεί στο παρελθόν.
- Κλιτική Μορφολογία (Inflectional Morphology), η οποία επικεντρώνεται στις μεταβολές που επιφέρονται στις λέξεις κατά την κλίση τους. Για παράδειγμα, οι λέξεις “μαθητής” και “μαθήτρια” είναι τύποι του ίδιου ουσιαστικού, αλλά διαφέρουν στο γένος.

Η παρούσα εργασία επικεντρώνεται στην παραγωγή ΔΠΛ που λαμβάνουν υπόψιν την κλιτική μορφολογία των λέξεων των Ελληνικών και των Αγγλικών. Επομένως, τα μορφήματα των καταλήξεων έχουν ιδιαίτερη σημασία για τους σκοπούς της εργασίας.

Πρέπει να σημειωθεί ότι το ίδιο το πρόβλημα της διάσπασης σε μορφήματα είναι αρκετά δύσκολο. Σε αρκετές περιπτώσεις δεν είναι εμφανής ακόμη και στον άνθρωπο η σωστή διάσπαση. Για παράδειγμα, δεν είναι εμφανές αν η λέξη “κατασκηνώσαμε” διασπάται σε “κατα-σκην-ώσαμε” ή “κατα-σκηνώ-σαμε” ή “κατα-σκη-νώ-σαμε”.

Όπως θα αναλυθεί λεπτομερώς στο Κεφάλαιο 3, υλοποιήθηκε ένας απλοϊκός αλγόριθμος διάσπασης σε αποκλειστικά δύο ψευδο-μορφήματα, ένα ψευρδο-θέμα και μία ψευδο-κατάληξη.

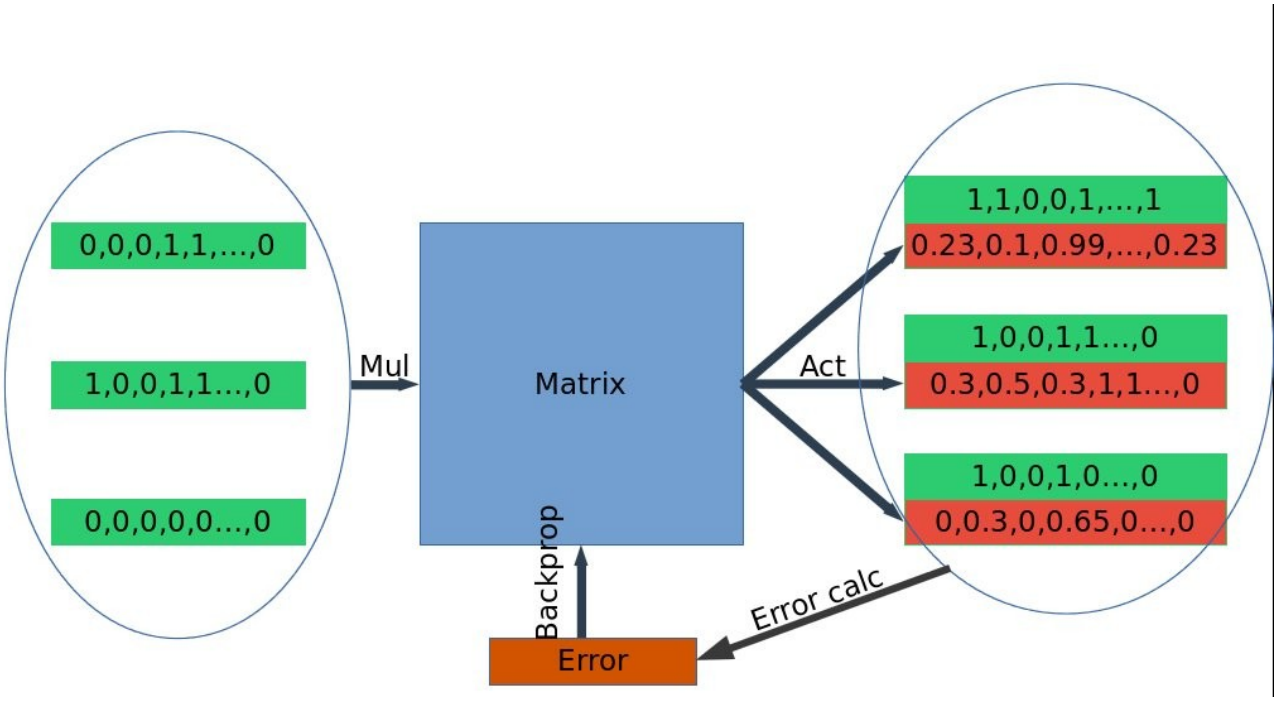
2.2 Νευρωνικά Δίκτυα

Αν και προτάθηκαν δεκαετίες νωρίτερα, τα (Τεχνητά) Νευρωνικά Δίκτυα έγιναν πρόσφατα πάλι ιδιαίτερα δημοφιλή στην ΕΦΓ και τη Μηχανικής Μάθηση. Καταλυτικό ρόλο σε αυτήν την εξέλιξη είχε η δραματική βελτίωση του υλισμικού (hardware) των υπολογιστών που είχε ως αποτέλεσμα τη μείωση του παλαιότερα απαγορευτικού χρόνου εκπαίδευσης, η ύπαρξη μεγάλων συνόλων δεδομένων, καθώς και βελτιώσεις στους αλγορίθμους εκπαίδευσης των Νευρωνικών Δικτύων. Έκτοτε, μέθοδοι που χρησιμοποιούν Νευρωνικά Δίκτυα έχουν αποδείξει να παράγουν συχνά τα καλύτερα (state of the art) αποτελέσματα σε πολλά προβλήματα της ΕΦΓ.

Κατά την εκπαίδευση, συνήθως το Νευρωνικό Δίκτυο “μαθαίνει” να μετασχηματίζει τα δεδομένα

εισόδου (inputs) στα επιθυμητά δεδομένα εξόδου (outputs / targets). Ο μετασχηματισμός γίνεται μέσω πολλαπλασιασμών πινάκων που βρίσκονται στα στρώματα (layers) του Νευρωνικού Δικτύου και της εφαρμογής συνήθως μη γραμμικών συναρτήσεων ενεργοποίησης (activation functions) στις εξόδους των στρωμάτων. Ο σκοπός της εκπαίδευσης είναι να βρεθούν τα σωστά βάρη στους εν λόγω πίνακες ώστε τα δεδομένα εισόδου να μετασχηματίζονται όσο το δυνατόν πιο κοντά στα δεδομένα εξόδου.

Η έξοδος από το τελευταίο στρώμα του Νευρωνικού Δικτύου (output layer) αποτελεί την πρόβλεψη του Δικτύου για τα τρέχοντα δεδομένα εισόδου. Σε αυτό το σημείο, υπολογίζεται το σφάλμα από τα πραγματικά δεδομένα εξόδου. Το σφάλμα μπορεί να υπολογιστεί με διάφορες συναρτήσεις, όπως Διαφορά Τετραγώνων (mean square error), hinge loss, διασταυρωμένη εντροπία (cross entropy) κ.ά., ανάλογα με τη φύση του προβλήματος. Έπειτα, το σφάλμα διαχέεται πίσω στα βάρη του δικτύου ώστε να ενημερωθούν με τη χρήση παραγωγίσεων, όπως αυτές προκύπτουν από το γνωστό κανόνα της αλυσίδας (chain rule). Η διαδικασία αυτή ονομάζεται error back-propagation [LeCun κ.ά. 1998].



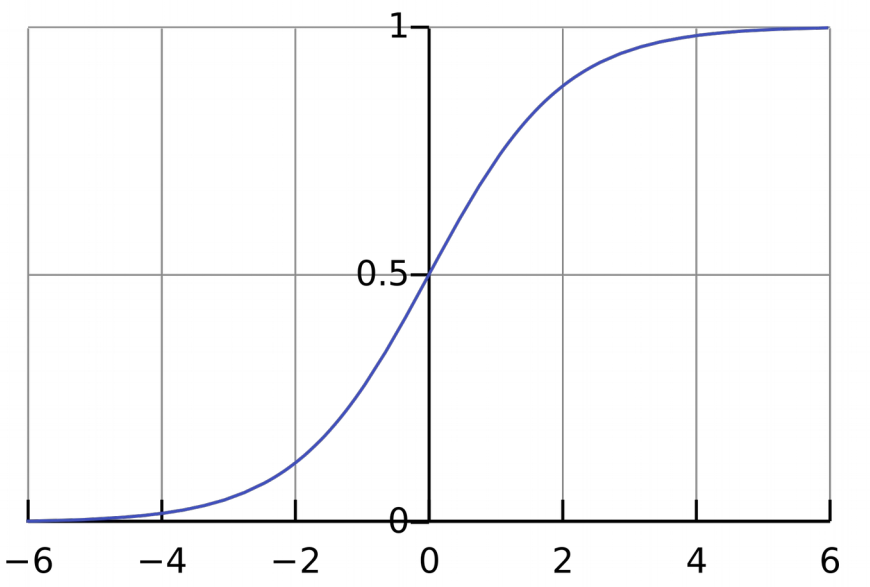
Εικόνα 1: Εκπαίδευση Τεχνητού Νευρωνικού Δικτύου

Παρακάτω, θα περιγραφούν περιληπτικά διάφορα στοιχεία των Νευρωνικών Δικτύων τα οποία χρησιμοποιήθηκαν στην εργασία.

2.2.1 Σιγμοειδής Συνάρτηση

Η σιγμοειδής (ή λογιστική) συνάρτηση (Logistic function) είναι από τις πιο συχνά χρησιμοποιούμενες συναρτήσεις ενεργοποίησης στα Νευρωνικά Δίκτυα. Είναι μη γραμμική συνάρτηση με πεδίο ορισμού το σύνολο των φυσικών αριθμών και πεδίο τιμών το (0, 1).

Χρησιμοποιείται κατά κόρον στο στρώμα εξόδου για τη μετατροπή των τιμών εξόδου σε πιθανότητες, όταν υπάρχουν μόνο δύο κατηγορίες (κλάσεις). Η συνάρτηση ορίζεται ως εξής: $S(t)=\frac{1}{1+e^{-t}}$ και η γραφική παράσταση είναι η παρακάτω:



Εικόνα 2: Σιγμοειδής Συνάρτηση

2.2.2 Συνάρτηση Softmax

Η συνάρτηση Softmax αποτελεί μία γενίκευση της σιγμοειδούς συνάρτησης για μετασχηματισμό τιμών σε πιθανότητες, όταν υπάρχουν παραπάνω από δύο κατηγορίες (κλάσεις). Εφαρμόζεται στο δίανυσμα που προκύπτει από το στρώμα εξόδου και μετασχηματίζει τις τιμές σε πιθανότητες οι οποίες αθροίζουν στη μονάδα. Αν το διάνυσμα εξόδου είναι μήκους 2, η softmax ταυτίζεται με τη σιγμοειδή συνάρτηση. Η συνάρτηση softmax ορίζεται ως εξής:

$$\sigma(z)_j=\frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{, για j = 1,…,K.}$$

Η softmax χρησιμοποιείται ευρέως σε προβλήματα κατηγοριοποίησης, όπου μόνο μία κατηγορία είναι η σωστή. Για παράδειγμα, αν κάθε λέξη του λεξιλογίου θεωρηθεί ως κατηγορία, κατά τη Γλωσσική Μοντελοποίηση, ο στόχος είναι να μεγιστοποιηθεί η πιθανότητα που προκύπτει από τη softmax για την επόμενη λέξη / κατηγορία, δεδομένων όλων των προηγούμενων λέξεων.

Ένα μείζον πρόβλημα κατά την εκπαίδευση μοντέλων τα οποία χρησιμοποιούν τη softmax είναι ότι είναι αρκετά χρονοβόρα, λόγω της κανονικοποίησης, ιδιαίτερα όταν εμπλέκονται εκατομμύρια ή δισεκατομμύρια κατηγορίες στα διανύσματα εξόδου. Λόγω αυτού, νέες προσεγγίσεις τόσο στην παραγωγή ΔΠΛ όσο και κατά τη Γλωσσική Μοντελοποίηση αποφεύγουν την εφαρμογή της

softmax σε όλο το μήκος του λεξιλογίου ([Ji κ.ά. 2015]).

2.2.3 Συναρτήσεις Σφάλματος

Οι συναρτήσεις σφάλματος εφαρμόζονται στο διάνυσμα εξόδου από το τελικό στρώμα του δικτύου, αφού αυτό έχει υποστεί τυχόν μετασχηματισμό από συνάρτηση ενεργοποίησης. Οι συναρτήσεις σφάλματος, δεδομένων των ορθών κατηγοριών και των αποτελεσμάτων του δικτύου που περιέχονται στο διάνυσμα εξόδου, παράγουν ένα νέο διάνυσμα, το διάνυσμα σφάλματος, το οποίο μετά την εκπαίδευση διαχέεται πίσω στα διανύσματα / βάρη του δικτύου ώστε αυτά να ενημερωθούν.

Έχουν προταθεί πολλές συναρτήσεις σφάλματος και η επιλογή μίας εξ αυτών εξαρτάται από το εκάστοτε πρόβλημα. Στην παρούσα εργασία, χρησιμοποιήθηκε κυρίως η (κατηγορική) διασταυρωμένη εντροπία (categorical cross entropy) ως συνάρτηση σφάλματος. Η συνάρτηση αυτή είναι κατάλληλη για προβλήματα κατηγοριοποίησης όταν παράγεται μία κατανομή πιθανοτήτων για τις κατηγορίες στο στρώμα εξόδου.

2.2.4 Ανατροφοδοτούμενα Νευρωνικά Δίκτυα

Τα Ανατροφοδοτούμενα Τεχνητά Νευρωνικά Δίκτυα είναι μία κατηγορία Νευρωνικών Δικτύων, στα οποία η έξοδος δεν εξαρτάται μόνο από την είσοδο αλλά και από την προηγούμενη έξοδο (ή) και εσωτερική κατάσταση του δικτύου. Με τον τρόπο αυτό σχηματίζεται ένας κατευθυνόμενος κύκλος μεταξύ των εξόδων του δικτύου και των εισόδων του. Σε πιο περίπλοκες δομές Ανατροφοδοτούμενων Νευρωνικών Δικτύων, όπως στα Long Short-Term Memory (LSTMs), ενδέχεται να υπάρχουν περισσότερες από μία εσωτερικές κρυφές καταστάσεις του δικτύου [Hochreiter κ.ά. 1997].

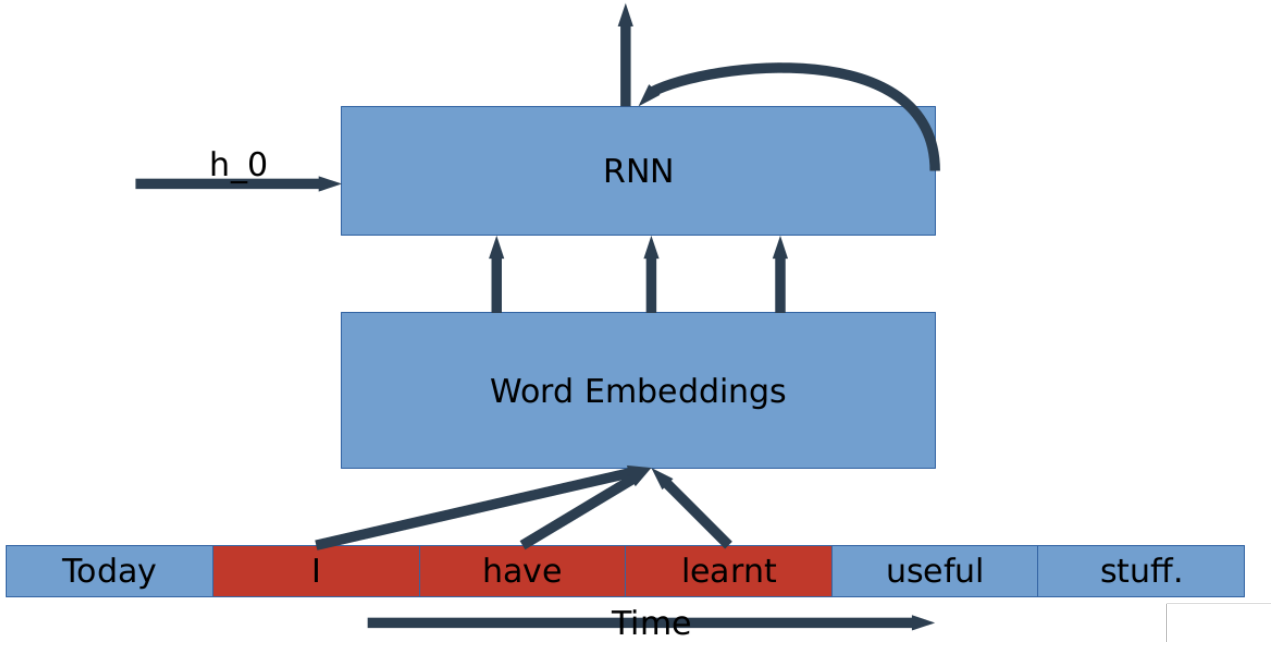
Τα Ανατροφοδοτούμενα Νευρωνικά Δίκτυα επομένως, είναι δίκτυα με “μνήμη”, αφού οι είσοδοι του παρελθόντος επηρεάζουν τις προβλέψεις του μέλλοντος. Αυτό τα καθιστά ιδανικά για προβλήματα όπως η Γλωσσική Μοντελοποίηση, όπου ο στόχος μπορεί να θεωρηθεί πως είναι η πρόβλεψη της επόμενης λέξης δεδομένων όλων των προηγούμενων.

Ένα μεγάλο πρόβλημα κατά την εκπαίδευση των Ανατροφοδοτούμενων Νευρωνικών Δικτύων είναι ότι λόγω της επαναληπτικής τους φύσης, το διάνυσμα σφάλματος γίνεται συχνά (λόγω διαδοχικών παραγωγίσεων) σχεδόν μηδενικό ή αποκτά άπειρο μέτρο [Bengio κ.ά. 1994]. Σε αυτές τις περιπτώσεις, ο υπολογιστής αδυνατεί να παραστήσει με ακρίβεια τις τιμές με αποτέλεσμα το διάνυσμα ενημέρωσης κατά την οπισθοδρόμηση σφάλματος να είναι μηδενικό, ή “άπειρο”. Και στις δύο περιπτώσεις, το δίκτυο δεν εκπαιδεύεται σωστά.

Για την αντιμετώπιση του παραπάνω προβλήματος, τα κρυφά διανύσματα των Ανατροφοδοτούμενων Νευρωνικών Δικτύων στην παρούσα εργασία αρχικοποιούνται με τη μέθοδο Glorot Uniform, όπως περιγράφεται στο [Glorot κ.ά. 2010]. Επίσης, έχει βρεθεί [Pascanu κ.ά. 2013] ότι θέτοντας ένα μέγιστο όριο στην τιμή των παραγώγων οι οποίες ενημερώνουν τα βάρη του

δικτύου, εξαλείφεται ο κίνδυνος εμφάνισης άπειρου μέτρου του διανύσματος σφάλματος και αυτό λειτουργεί καλά στην πράξη.

Στην παρακάτω εικόνα φαίνεται μία απεικόνιση ενός Ανατροφοδοτούμενου Νευρωνικού Δικτύου. Η αρχική κρυφή κατάσταση παριστάνεται με ένα διάνυσμα h_0. Το δίκτυο δέχεται ως είσοδο διαδοχικά παράθυρα μήκους τριών λέξεων (κόκκινο χρώμα). Οι λέξεις μετατρέπονται στις αντίστοιχες ΔΠΛ τους κι έπειτα μετασχηματίζονται από το RNN. Η έξοδος της πρώτης λέξης (του παραθύρου της) αποτελεί μαζί με το παράθυρο της επόμενης λέξης την είσοδο για τον υπολογισμό της εξόδου της επόμενης κ.ο.κ. Με αυτόν τον τρόπο όλες οι προηγούμενες λέξεις επηρεάζουν την έξοδο των επόμενων.

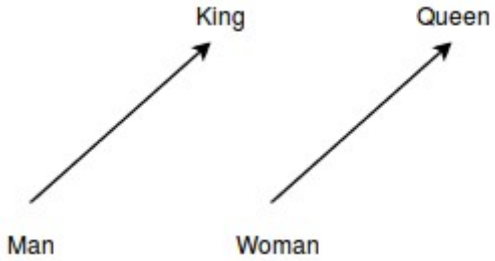


Εικόνα 3: Απεικόνιση Ανατροφοδοτούμενου Νευρωνικού Δικτύου

2.3 Διανυσματικές Παραστάσεις Λέξεων

Όπως δηλώνει η ονομασία, οι λέξεις παριστάνονται ως διανύσματα. Τα εν λόγω διανύσματα είναι σταθερού μήκους πλήθους διαστάσεων και περιέχουν πραγματικούς αριθμούς. Συνηθισμένα μήκη διανυσμάτων είναι 50, 100, 150, 200, 300. Δεν είναι πάντα απαραίτητο ότι μεγαλύτερου μήκους διανύσματα θα δώσουν καλύτερα αποτελέσματα στα προβλήματα που τα χρησιμοποιούν.

Ένα από τα χαρακτηριστικά γνωρίσματα των ΔΠΛ όταν εκπαιδευτούν σωστά είναι η διατήρηση γραμμικών συσχετισμών ανάμεσα στις λέξεις. Έτσι, για παράδειγμα, το διάνυσμα που προκύπτει από την πράξη “king – man + woman” είναι πολύ κοντά στο διάνυσμα της λέξης “queen” [Mikolov κ.ά. 2013α].



Εικόνα 4: Παραλληλισμός ΔΠΛ

2.3.1 Παραγωγή Διανυσματικών Παραστάσεων Λέξεων με τη χρήση Επαναληπτικού Νευρωνικού Δικτύου

Μία από τις πρώτες ευρέως γνωστές εξελίξεις στην παραγωγή διανυσματικών παραστάσεων λέξεων έγινε από τον Bengio και τους συνεργάτες του [Bengio κ.ά. 2003]. Οι διανυσματικές παραστάσεις τους παράγονται κατά την εκπαίδευση ενός Ανατροφοδοτούμενου Νευρωνικού Δικτύου, το οποίο εκπαιδεύεται για Γλωσσική Μοντελοποίηση. Επομένως, είναι δευτερεύον προϊόν του Νευρωνικού Δικτύου, αφού δεν ήταν αυτός ο πρωταρχικός στόχος εκπαίδευσής του.

Η εκπαίδευση μεγάλων σε μέγεθος λεξιλογίου Ανατροφοδοτούμενων Νευρωνικών Δικτύων είναι γενικά αρκετά χρονοβόρα διαδικασία. Όπως θα αναλυθεί παρακάτω, στο πρόσφατο παρελθόν προτάθηκαν εναλλακτικά μοντέλα παραγωγής διανυσματικών παραστάσεων λέξεων, τα οποία επικεντρώνονται στο συγκεκριμένο έργο, αποφεύγοντας βαθιά (με παραπάνω από ένα κρυφό στρώμα), κι επομένως χρονοβόρα, μοντέλα όπως τα Ανατροφοδοτούμενα Νευρωνικά Δίκτυα.

2.3.2 Παραγωγή Διανυσματικών Παραστάσεων Λέξεων με τη χρήση “ρηχών” Νευρωνικών Δικτύων

Όπως προαναφέρθηκε, η παραγωγή ΔΠΛ μέσω βαθέων Νευρωνικών Δικτύων είναι αρκετά χρονοβόρα διαδικασία. Για την αντιμετώπιση αυτού του φαινομένου, οι [Mikolov κ.ά. 2013γ] εισήγαγαν δύο νέους αλγορίθμους για τον υπολογισμό ΔΠΛ, τον “skip-grams” και τον “continuous bag of words” (CBoW). Αμφότεροι αυτοί οι αλγόριθμοι λειτουργούν σαρώνοντας ένα μεγάλο σώμα κειμένων εκπαίδευσης (corpus) με κυλιόμενο παράθυρο. Το μέγεθος του παραθύρου καθορίζεται ως υπερ-παράμετρος. Το μοντέλο Skip-Grams επιχειρεί να μάθει ΔΠΛ που επιτρέπουν την πρόβλεψη των υπολοίπων λέξεων (συμφραζομένων) κάθε παραθύρου, δεδομένης της κεντρικής λέξης του παραθύρου. Η πρόβλεψη προκύπτει (σε γενικές γραμμές) υπολογίζοντας το εσωτερικό γινόμενο του διανύσματος της κεντρικής λέξης και κάθε λέξης του λεξιλογίου και εφαρμόζοντας softmax στα αποτελέσματα των εσωτερικών γινομένων. Εξάγεται έτσι μία κατανομή πιθανότητας για όλο το λεξιλόγιο, για κάθε μη κεντρική θέση του παραθύρου. Αντίθετα, στόχος του CboW είναι να αποδώσει μεγάλη πιθανότητα στην κεντρική λέξη δεδομένων των συμφραζομένων.

?	?	υπερψήφισε	?	?

Εικόνα 5: Skip-Grams

Η	Βουλή	?	το	νόμο

Εικόνα 6: CboW

Για την επιτάχυνση των υπολογισμών και τη βελτίωση των αποτελεσμάτων, μπορούν να χρησιμοποιηθούν δύο τεχνικές σε οποιονδήποτε από τους δύο παραπάνω αλγορίθμους. Οι τεχνικές αυτές ονομάζονται ιεραρχικό softmax και αρνητική δειγματοληψία (negative sampling) ([Mikolov κ.ά. 2013β]).

Η αρνητική δειγματοληψία, η οποία χρησιμοποιήθηκε και στην παρούσα εργασία για την εκπαίδευση των ΔΠΛ, αποφεύγει την εφαρμογή της συνάρτησης softmax σε όλο το στρώμα εξόδου κι επομένως σε ολόκληρο το λεξιλόγιο. Αντί αυτού, επιλέγεται ένα καθορισμένο πλήθος (συνήθως 3-10) τυχαίων λέξεων από το λεξιλόγιο, οι οποίες δεν εμφανίζονται στο τρέχον παράθυρο εκπαίδευσης. Η δειγματοληψία ευνοεί τις σπανιότερες λέξεις του λεξιλογίου, υποδειγματοληπτώντας τις πιο συχνές, με το σκεπτικό ότι οι σπανιότερες λέξεις έχουν περισσότερη “ανάγκη” παραδειγμάτων εκπαίδευσης. Στη συνέχεια, ο στόχος είναι να μεγιστοποιηθεί η πιθανότητας πρόβλεψης της σωστής κεντρικής λέξης και να ελαχιστοποιηθεί η πιθανότητα πρόβλεψης των τυχαίων λέξεων ως κεντρικές λέξεις του παραθύρου για τον αλγόριθμο CboW, ή αντίστοιχα, να μεγιστοποιηθούν οι πιθανότητες πρόβλεψης των σωστών συμφραζομένων και να ελαχιστοποιηθούν εκείνες των τυχαίων λέξεων ως συμφραζόμενα για τον αλγόριθμο Skip-Grams.

2.3.3 Παραγωγή Διανυσματικών Παραστάσεων Λέξεων λαμβάνοντας υπόψιν τη μορφολογία

Οι μέθοδοι παραγωγής ΔΠΛ που παρουσιάστηκαν δε λαμβάνουν υπόψιν τη μορφολογία των λέξεων. Αυτό πρακτικά σημαίνει ότι θεωρούν όλες τις λέξεις ως ανεξάρτητες και ασύνδετες μεταξύ τους, ακόμα και αν η μορφολογική αλλαγή είναι πολύ μικρή, όπως για παράδειγμα η αλλαγή πτώσης (“καλή” -> “καλής”). Σε πολύ πλούσιες μορφολογικά γλώσσες όπως η ελληνική, ενέχεται ο κίνδυνος να μην παρασταθούν ορθά ορισμένες μορφολογικές παραλλαγές μιας λέξεις λόγω της σπανιότητάς τους στο σώμα εκπαίδευσης. Πολύ συχνά, οι λέξεις αυτές είναι παραλλαγές άλλων πολύ συχνότερων λέξεων, αλλά οι προηγούμενες μέθοδοι ΔΠΛ αδυνατούν να εκμεταλλευτούν αυτήν την πληροφορία για να τις παραστήσουν καλύτερα.

Τα μοντέλα τα οποία έχουν προταθεί και λαμβάνουν υπόψιν τη μορφολογία διαφέρουν αρκετά ως προς τη δομή και την οπτική γωνία τους.

- Το MorphemeCboW των Qiu και των συνεργατών του [Qiu κ.ά. 2014] αποτελεί μία παραλλαγή του CboW που παρουσιάστηκε προηγουμένως, στο οποίο δημιουργούνται διανυσματικές παραστάσεις και για τα μορφήματα των λέξεων. Τα μορφήματα προκύπτουν

από εξωτερικά εργαλεία, όπως το Morfessor (http://morfessor.readthedocs.org/en/latest/). Εντέλει, οι ΔΠΛ των λέξεων προκύπτουν προσθέτοντας (με βάρη) τα διανύσματα των μορφημάτων των λέξεων και τις ΔΠΛ που προκύπτουν με την αρχική μορφή του CboW.

- Το csmRNN [Luong κ.ά. 2013] χρησιμοποιεί Αναδρομικά Νευρωνικά Δίκτυα (Recursive Neural Networks) για τη δημιουργία διανυσματικών παραστάσεων αγνώστων (ή σπανίων) λέξεων από διανυσματικές παραστάσεις μορφημάτων (τις οποίες επίσης μαθαίνει). Τα μορφήματα παράγονται πάλι από το Morfessor. Στη συνέχεια, οι ΔΠΛ δίνονται σε ένα Ανατροφοδοτούμενο Νευρωνικό Δίκτυο, το οποίο χρησιμοποιείται για Γλωσσική Μοντελοποίηση.
- Το C2W [Ling κ.ά. 2015] χρησιμοποιεί αμφίδρομα Επαναληπτικά Νευρωνικά Δίκτυα (πιο συγκεκριμένα δύο αμφίδρομα LSTM) για να δημιουργήσεις ΔΠΛ από διανυσματικές παραστάσεις χαρακτήρων (πρώτο LSTM) και για να εκτελέσει κατόπιν εργασίες όπως Γλωσσική Μοντελοποίηση ή Επισημείωση Μερών του Λόγου (δεύτερο LSTM). Έτσι, καμία λέξη δεν είναι λέξη εκτός λεξιλογίου και η κατανάλωση μνήμης περιορίζεται θεαματικά. Παρόλ' αυτά, ο χρόνος εκπαίδευσης είναι αρκετά υψηλός. Έχει δειχθεί ότι εξάγει πάρα πολύ καλά αποτελέσματα στην εκπαίδευση ΔΠΛ, στη Γλωσσική Μοντελοποίηση, αλλά και στην Επισημείωση Μερών του Λόγου.

3. Νέες μέθοδοι παραγωγής Διανυσματικών Παραστάσεων Λέξεων λαμβάνοντας υπόψιν τη μορφολογία

3.1 Επισκόπηση

Στη διάρκεια της παρούσας Διπλωματικής Εργασίας αναπτύχθηκαν δύο συστήματα παραγωγής ΔΠΛ. Αρχικά, ακολουθείται και στα δύο συστήματα η ίδια διαδικασία για την εξαγωγή ψευδομορφημάτων, διαχωρίζοντας σε θέματα και καταλήξεις, με τρόπο ο οποίος θα αναλυθεί λεπτομερώς σε επόμενη υποενότητα.

Το πρώτο σύστημα, το οποίο ονομάζεται “stem-suffix algorithm”, εκτελεί το εξωτερικό πρόγραμμα word2vec δύο φορές, μία φορά για τα θέματα και μία για τις καταλήξεις. Το τελικό διάνυσμα κάθε λέξης προκύπτει από συνένωση του διανύσματος του θέματος και του διανύσματος της κατάληξης, τα οποία παράγονται από τις δύο εκτελέσεις του word2vec.

Το δεύτερο σύστημα, το οποίο ονομάζεται “morph-CboW”, αποτελεί παραλλαγή του συστήματος “morphemeCBoW” των Qiu και των συνεργατών του. Σε αντίθεση με το morphemeCBoW, δε χρησιμοποιήθηκε το Morfessor για την εξαγωγή μορφημάτων. Επίσης, το αναπτυχθέν σύστημα περιορίζεται αποκλειστικά σε δύο ψευδο-μορφήματα, ένα για θέμα και ένα για κατάληξη, ενώ το morphemeCBoW δεν έχει όριο στο πλήθος των δυνατών μορφημάτων που μπορούν να περιλαμβάνονται σε μία λέξη. Δεδομένου ότι η παρούσα Εργασία επικεντρώνεται στη μελέτη της κλιτικής μορφολογίας των λέξεων η οποία παρουσιάζεται κυρίως στις καταλήξεις, στα Ελληνικά και στα Αγγλικά, ο περιορισμός σε δύο μορφήματα φαινόταν καλή επιλογή.

Η αξιολόγηση των παραχθέντων ΔΠΛ γίνεται με τα εξής μέτρα αξιολόγησης:

- Ευστοχία (accuracy) στην αναλογία λέξεων.
- Spearman's correlation στην ομοιότητα λέξεων.
- Περιπλοκή (perplexity) κατά τη Γλωσσική Μοντελοποίηση.
- Ευστοχία κατά την ΕΜΛ.

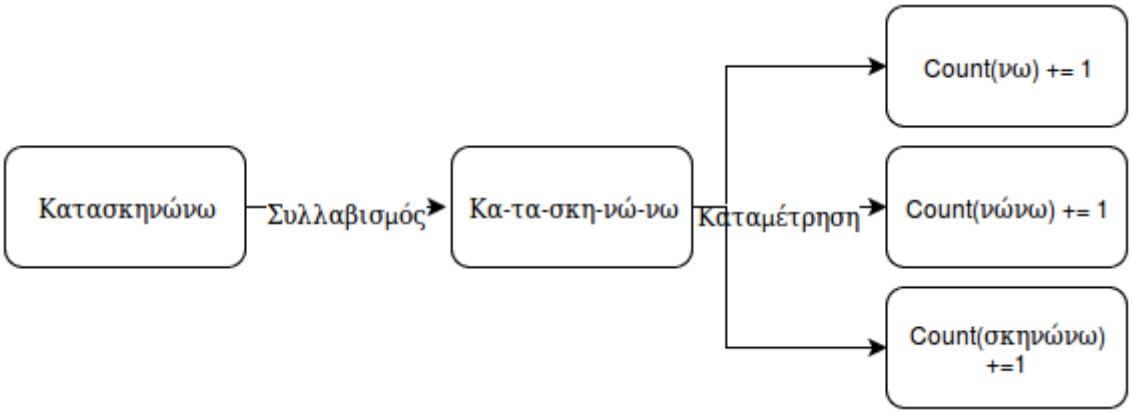
Είναι εφικτό να βελτιωθούν τα ΔΠΛ και κατά τη Γλωσσική Μοντελοποίηση, αλλά στην παρούσα εργασία η Γλωσσική Μοντελοποίηση χρησιμοποιείται μόνο για την αξιολόγηση ΔΠΛ που έχουν ήδη δημιουργηθεί και δεν τις τροποποιεί.

Όλος ο πηγαίος κώδικας της εργασίας είναι γραμμένος στη γλώσσα Python (3.4+). Τα Νευρωνικά Δίκτυα υλοποιήθηκαν με τις βιβλιοθήκες keras (http://keras.io/) και theano [Bastien κ.ά. 2012], οι οποίες είναι ανοιχτού κώδικα και ελεύθερα διαθέσιμες.

3.2 Διάσπαση λέξεων σε μορφήματα

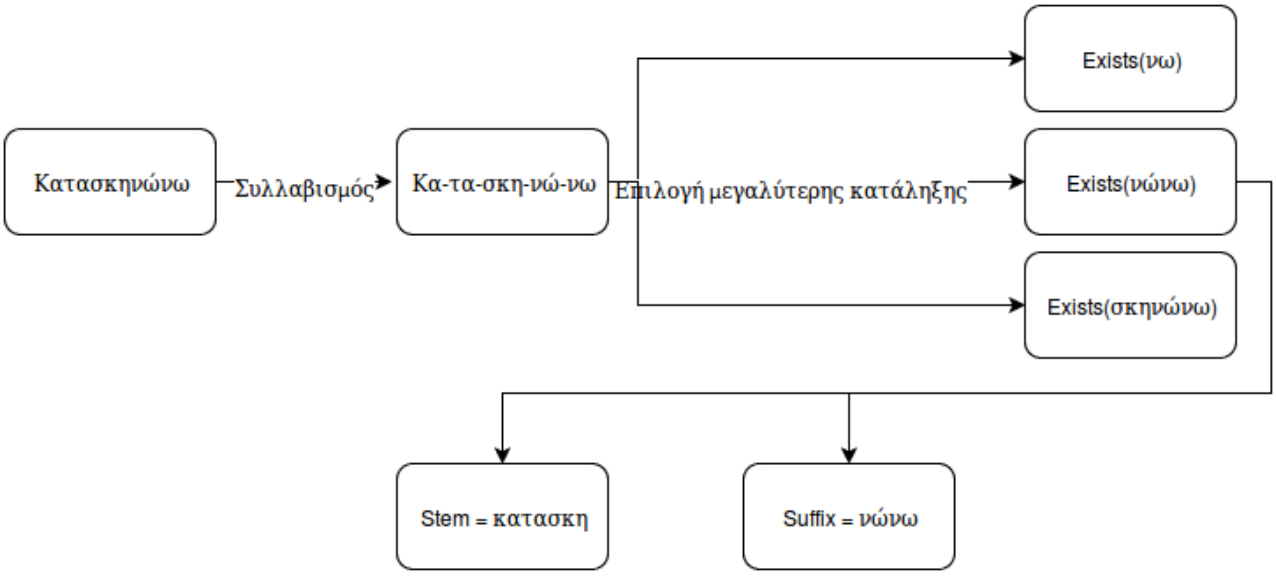
Όπως προαναφέρθηκε, δε χρησιμοποιήθηκε κάποιο εξωτερικό εργαλείο για την ανάλυση των λέξεων σε μορφήματα. Ο λόγος είναι ότι προπαρασκευαστικά πειράματα με το Morfessor στα Ελληνικά δεν διασπούσαν σωστά πολλές λέξεις, παρόλο που το σύνολο εκπαίδευσης ήταν επαρκές. Για τις ανάγκες της εργασίας, όλες οι λέξεις διασπώνται σε ψευδο-μορφήματα καταλήξεων και θεμάτων. Το πρώτο βήμα για τη διάσπαση είναι η διάσπαση σε συλλαβές. Έπειτα, δημιουργείται ένα σύνολο από υποψήφιες καταλήξεις, μετρώντας το πλήθος εμφανίσεων συνενώσεων έως των k τελευταίων συλλαβών (όπου k είναι ακέραιος αριθμός και υπερπαράμετρος εκπαίδευσης) για κάθε **μοναδική** λέξη στο σώμα εκπαίδευσης. Αφού ολοκληρωθεί η καταμέτρηση, από τις υποψήφιες καταλήξεις αφαιρούνται όσες δεν εμφανίστηκαν σε παραπάνω από t μοναδικές λέξεις (όπου t είναι ακέραιος αριθμός και υπερπαράμετρος εκπαίδευσής). Εντέλει, για τη διάσπαση μιας λέξης βρίσκεται πρώτα η μεγαλύτερη σε μήκος κατάληξη που ταιριάζει με την κατάληξη της λέξης, η οποία θεωρείται η κατάληξη της λέξης. Το υπόλοιπο κομμάτι της λέξης θεωρείται το θέμα. Σε περίπτωση που δε βρεθεί κατάληξη ή ολόκληρη η λέξη είναι κατάληξη, ως θέμα και κατάληξη θεωρείται ολόκληρη η λέξη.

Στην παρακάτω εικόνα φαίνεται για παράδειγμα η λέξη “κατασκηνώνω”. Από το συλλαβισμό προκύπτει η διάσπαση σε “κα”, “τα”, “σκη”, “νώ”, “νω”. Ο αλγόριθμος έχει δεχτεί ως παράμετρο να συνενώνει μέχρι τις τρεις τελευταίες συλλαβές για υποψήφιες καταλήξεις. Έτσι, αυξάνεται ο μετρητής της υποψήφιας κατάληξης “νω”, της “νώνω” (“νώ” + “νω”) και της “σκηνώνω” (“σκη” + “νώ” + “νω”).



Εικόνα 7: Καταμέτρηση πιθανών καταλήξεων

Συνεχίζοντας το παράδειγμα, αφού έχει ολοκληρωθεί η καταμέτρηση των υποψήφιων καταλήξεων διασχίζοντας μία πρώτη φορά όλο το κείμενο εκπαίδευσης, απορρίπτονται οι υποψήφιες καταλήξεις οι οποίες δεν ξεπέρασαν σε εμφανίσεις ένα όριο. Στην επόμενη εικόνα φαίνεται η τελική διάσπαση της λέξης “κατασκηνώνω” σε θέμα και κατάληξη. Αρχικά η λέξη διασπάται σε συλλαβές όμοια με προηγουμένως. Έπειτα, επιλέγεται η μεγαλύτερη κατάληξη. Πιο συγκεκριμένα, ο αλγόριθμος εξετάζει αν οι καταλήξεις “νω”, “νώνω” και “σκηνώνω” είχαν ξεπεράσει το όριο εμφανίσεων. Σε αυτήν την περίπτωση οι καταλήξεις “νω” και “νώνω” το είχαν ξεπεράσει ενώ η κατάληξη “σκηνώνω” όχι. Επομένως, ως κατάληξη θεωρείται το “νώνω” και ως θέμα η υπόλοιπη λέξη, δηλαδή το “κατασκη”.



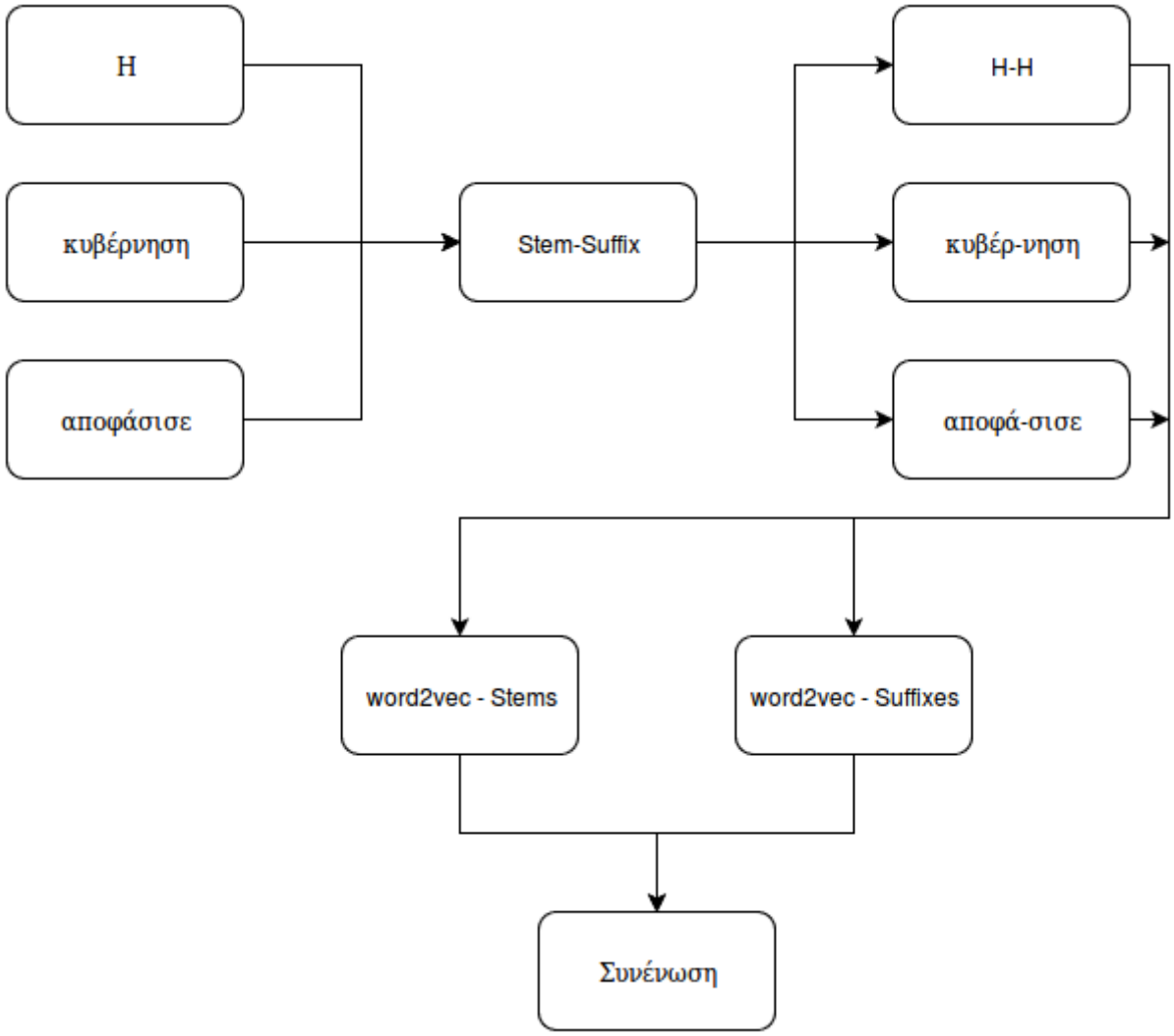
Εικόνα 8: Διάσπαση λέξης σε μορφήματα

3.3 Μοντέλο “Stem-Suffix Alg”

Το μοντέλο Stem-Suffix εκπαιδεύεται ως εξής: Αρχικά, όλες οι λέξεις του σώματος εκπαίδευσης διασπώνται σε θέματα και καταλήξεις με τον αλγόριθμο που περιγράφηκε παραπάνω. Έτσι, δημιουργούνται δύο νέες μορφές του σώματος εκπαίδευσης, οι οποίες περιέχουν μόνο τα θέματα και μόνο τις καταλήξεις των λέξεων αντίστοιχα. Έπειτα, αυτές οι δύο μορφές δίνονται ως είσοδος στο εξωτερικό πρόγραμμα word2vec, το οποίο παράγεις διανυσματικές παραστάσεις λέξεων και καταλήξεων. Οι τελικές ΔΠΛ για κάθε λέξη προκύπτουν από συνένωση της διανυσματικής παράστασης του θέματος και της κατάληξης.

Το μήκος (πλήθος διαστάσεων) των διανυσματικών παραστάσεων του θέματος και της κατάληξης ορίζεται ως υπερπαράμετρος στο μοντέλο και μπορεί να μην είναι η ίδια. Όπως θα αναλυθεί στο κεφάλαιο των πειραμάτων, αυτή η υπερπαράμετρος επηρεάζει αρκετά τα αποτελέσματα.

Στην παρακάτω εικόνα περιγράφεται ο τρόπος με τον οποίο εκπαιδεύονται τα μοντέλα για τη φράση “η κυβέρνηση αποφάσισε”. Αρχικά, κάθε λέξη διασπάται σε θέματα και καταλήξεις, δηλαδή στο συγκεκριμένο παράδειγμα η λέξη “η” έχει θέμα “η” και κατάληξη “η” (επειδή είναι μονοσύλλαβη), η λέξη “κυβέρνηση” έχει θέμα “κυβέρ” και κατάληξη “νηση” και η λέξη “αποφάσισε” έχει θέμα “αποφά” και κατάληξη “σισε”. Επομένως, η μορφή του σώματος εκπαίδευσης που περιέχει μόνο τα θέματα θα περιέχει τη φράση “η κυβέρ αποφά” και η μορφή του σώματος που περιέχει μόνο τις καταλήξεις θα περιέχει τη φράση “η νηση σισε”.



Εικόνα 9: Εκπαίδευση του Stem-Suffix Alg

3.4 Μοντέλο Morph-CboW

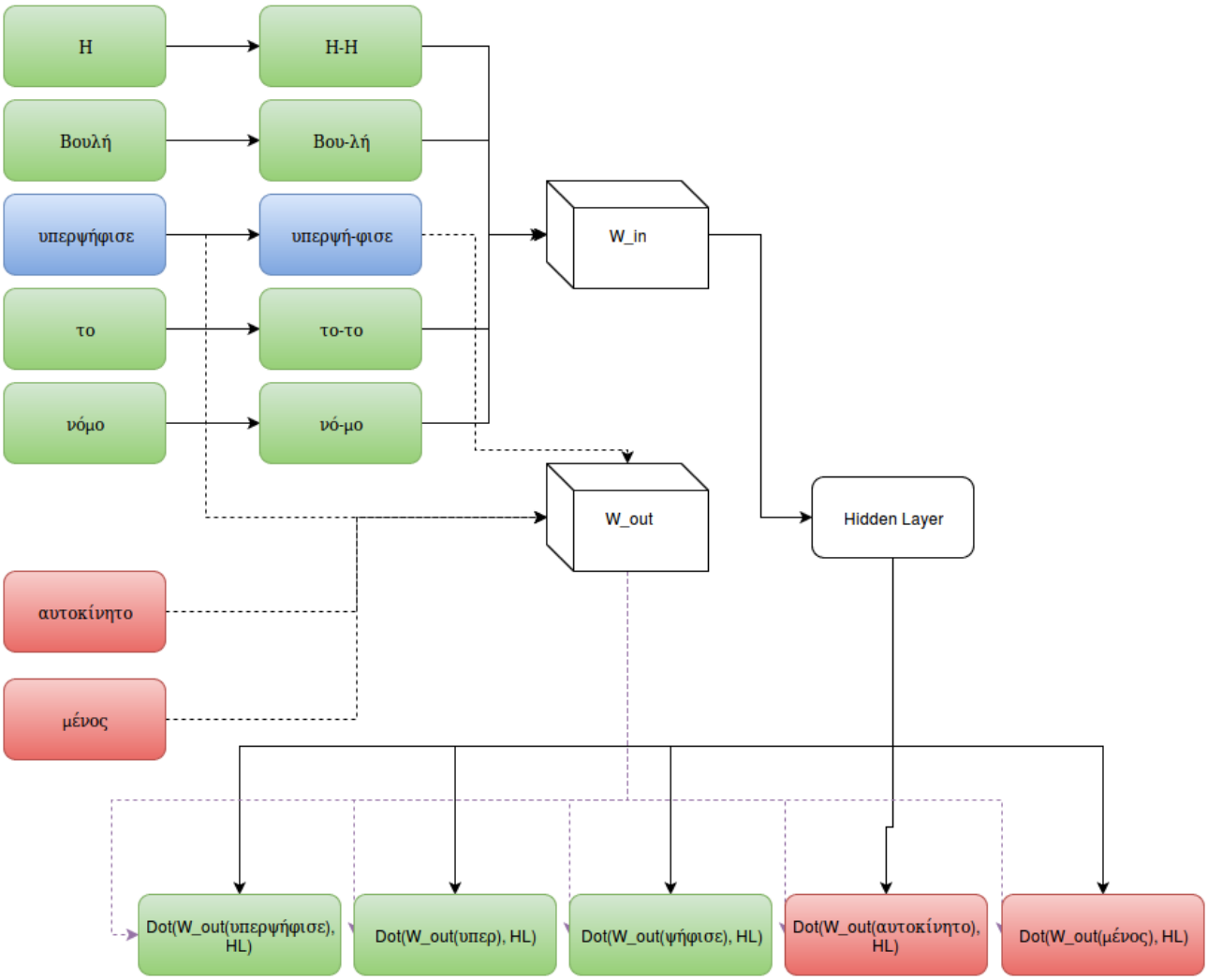
Το μοντέλο Morph-CboW αποτελεί παραλλαγή του μοντέλου “MorphemeCBoW” του μοντέλου που προτάθηκε από τον Qiu και τους συνεργάτες του. Στην παρούσα υλοποίηση ο αριθμός των μορφημάτων περιορίζεται σε δύο, δηλαδή στα μορφήματα που προκύπτουν από τη διάσπαση σε θέμα και κατάληξη με τον τρόπο που παρουσιάστηκε στην ενότητα 3.1. Κατά την εκπαίδευση, χρησιμοποιείται αρνητική δειγματοληψία, όπως και στο αρχικό μοντέλο. Επίσης, οι ΔΠΛ των λέξεων και των μορφημάτων των συμφραζομένων ζυγίζονται με διαφορετικά βάρη, τα οποία μεταβάλλονται κατά την εκπαίδευση, όπως και στο αρχικό μοντέλο.

Πιο αναλυτικά, η διαδικασία της εκπαίδευσης έχει ως εξής: Αρχικά δημιουργείται το σύνολο των πιθανών καταλήξεων όπως περιγράφεται στην ενότητα 3.1. Έπειτα, δημιουργούνται δύο πίνακες, ο W_in και ο W_out, οι οποίοι έχουν w + m γραμμές, όπου w το πλήθος των μοναδικών λέξεων του σώματος εκπαίδευσης και m το πλήθος όλων των μοναδικών μορφημάτων των λέξεων του σώματος εκπαίδευσης. Οι πίνακες έχουν d στήλες, όπου d είναι η διάσταση των ΔΠΛ. Στόχος της εκπαίδευσης είναι η κωδικοποίηση των ΔΠΛ στον πίνακα W_in.

Όπως και στο αρχικό CboW, υπάρχει ένα κυλιώμενο παράθυρο σταθερού μήκους στις λέξεις του σώματος εκπαίδευσης και σε κάθε βήμα το μοντέλο πρέπει να μεγιστοποιήσει την πιθανότητα εμφάνισης της κεντρικής λέξης δεδομένης της ύπαρξης των συμφραζομένων. Η διαφοροποίηση με την ύπαρξη των μορφημάτων είναι ότι στα συμφραζόμενα λαμβάνονται υπόψιν όχι μόνο ολόκληρες οι λέξεις, αλλά και όλα τα μορφήματα που τις αποτελούν. Έτσι, στο κρυφό στρώμα συνεισφέρουν τόσο οι λέξεις όσο και τα μορφήματα, αφού οι λέξεις βεβαρηθούν με ένα συντελεστή για αυτές και τα μορφήματα με έναν άλλο συντελεστή.

Επίσης, ο στόχος της πρόβλεψης σε αυτό το μοντέλο δεν είναι μόνο η κεντρική λέξη, αλλά και τα κεντρικά μορφήματα. Επιπλέον, δεδομένου ότι γίνεται χρήση αρνητικής δειγματοληψίας, επιλέγονται τυχαίες λέξεις ή μορφήματα που δεν υπάρχουν στο τρέχον παράθυρο και επιδιώκεται η ελαχιστοποίηση της πιθανότητας εμφάνισής τους ως κεντρική λέξη ή μόρφημα.

Στο παράδειγμα της παρακάτω εικόνας το παράθυρο είναι μήκους 4 (2 λέξεις πριν την κεντρική και 2 μετά). Ο αριθμός των αρνητικών λέξεων είναι 2. Η φράση είναι “Η Βουλή υπερψήφισε το νόμο” και οι αρνητικές λέξεις / μορφήματα είναι οι “αυτοκίνητο” και “μένος”.



Εικόνα 10: Παράδειγμα Morph-CBoW

3.5 Βελτίωση κατά τη Γλωσσική Μοντελοποίηση

Για τις ανάγκες της γλωσσικής μοντελοποίησης, η οποία χρησιμοποιείται ως τρόπος αξιολόγησης ΔΠΛ όπως θα αναλυθεί με μεγαλύτερη λεπτομέρεια στο κεφάλαιο 4, αναπτύχθηκε ένα Ανατροφοδοτούμενο Νευρωνικό Δίκτυο το οποίο αντί να επιδιώκει να μεγιστοποιήσει την πιθανότητα της επόμενης λέξης δεδομένων των προηγούμενων, όπως στην κλασική γλωσσική μοντελοποίηση, επιδιώκει να μεγιστοποιήσει την πιθανότητα του ζεύγους των μορφημάτων που απαρτίζουν την επόμενη λέξη δεδομένων των προηγούμενων. Για την αξιολόγηση, οι ΔΠΛ των λέξεων / μορφημάτων πρέπει να παραμείνουν “παγωμένες” ώστε να είναι αξιόπιστα τα αποτελέσματα. Δεδομένου όμως ότι πειράματα έχουν δείξει ότι μπορούν να παραχθούν ΔΠΛ ως υποπροϊόν της Γλωσσικής Μοντελοποίησης [Bengio κ.ά. 2003], είναι εφικτό να βελτιωθούν οι ΔΠΛ αν τους επιτραπεί να ενημερώνονται κατά την οπισθοδρόμηση σφάλματος. Παρόλ' αυτά, όπως έχει προαναφερθεί, η εκπαίδευση με Ανατροφοδοτούμενα Νευρωνικά Δίκτυα είναι εξαιρετικά χρονοβόρα διαδικασία και για να βελτιωθούν οι ΔΠΛ η εκπαίδευση πρέπει να γίνει σε σώμα κειμένου εφάμιλλου μεγέθους με εκείνο στο οποίο εκπαιδεύτηκαν αρχικά οι ΔΠΛ.

Περισσότερες λεπτομέρειες για την εκπαίδευση και τη δομή του Ανατροφοδοτούμενου Νευρωνικού Δικτύου το οποίο αναπτύχθηκε παραθέτονται στο κεφάλαιο 4.

4. Τρόποι Αξιολόγησης

4.1 Επισκόπηση

Οι ΔΠΛ αξιολογούνται με δύο τρόπους, εσωτερικά (intrinsic) και εξωτερικά (extrinsic).

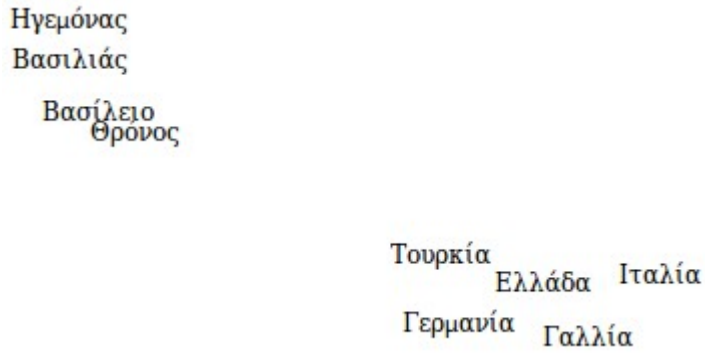
Οι εσωτερικές αξιολογήσεις εξετάζουν άμεσα τις ίδιες τις ΔΠΛ, χωρίς δηλαδή να εξετάζουν αν οι ΔΠΛ βελτιώνουν την επίδοση ενός μεγαλύτερου συστήματος. Εστιάζουν στην αξιολόγηση της σημασιολογίας που έχουν “μάθει” οι ΔΠΛ. Οι εσωτερικές αξιολογήσεις που χρησιμοποιήθηκαν είναι η ομοιότητα λέξεων (word similarity) και η αναλογία λέξεων (word analogy). Υπάρχουν επιπλέον μέθοδοι αξιολόγησης των ΔΠΛ κι έχει αποδειχτεί ότι δεν είναι αναγκαίο ένας αλγόριθμος παραγωγής ΔΠΛ που επιτυγχάνει καλύτερα αποτελέσματα αξιολόγησης με μία μέθοδο να τα πηγαίνει εξίσου καλά και με άλλες [Schnabel κ.ά. 2015].

Οι εξωτερικές αξιολογήσεις γίνονται με συστήματα τα οποία επικεντρώνονται στην επίλυση κάποιου προβλήματος της ΕΦΓ και χρησιμοποιούν τις παραχθείσες ΔΠΛ για την παράσταση των λέξεων. Επομένως, σε αυτήν την περίπτωση η αξιολόγηση είναι έμμεση, αλλά συνήθως πιο ουσιώδης μιας και γίνεται σε ένα πραγματικό πρόβλημα της ΕΦΓ.

4.2 Ομοιότητα και Αναλογία λέξεων

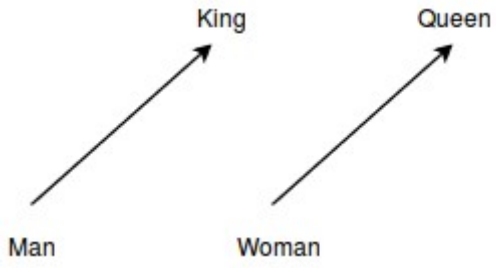
Η ομοιότητα και η αναλογία λέξεων είναι δύο εσωτερικοί τρόποι αξιολόγησης. Τόσο η αναλογία και (κυρίως) η ομοιότητα βασίζονται σε υποκειμενικά κριτήρια. Υπάρχουν πολλοί τρόποι για να υπολογιστεί η ομοιότητα μεταξύ μίας λέξης α με μία λέξη β. Στην παρούσα εργασία χρησιμοποιείται η ομοιότητα συνημιτόνου (cosine similarity) μεταξύ των ΔΠΛ των δύο λέξεων. Πριν υπολογιστεί η ομοιότητα συνημιτόνου, κάθε ΔΠΛ έχει κανονικοποιηθεί ώστε το άθροισμα των τετραγώνων των πραγματικών αριθμών που απαρτίζουν το εκάστοτε διάνυσμα να ισούται με τη μονάδα. Τα σύνολα αξιολόγησης συνήθως προκύπτουν από έρευνες στις οποίες συμμετέχουν πολλά άτομα παγκοσμίως, τα οποία καλούνται να απαντήσουν σε υποκειμενικές ερωτήσεις. Η τελική μετρική που χρησιμοποιείται για τα πειράματα είναι η συσχέτιση ρ του Spearman (Spearman's ρ correlation), ανάμεσα στην ομοιότητα συνημιτόνου που προέκυψε από το εκάστοτε μοντέλο και την ομοιότητα που έδωσαν οι ανθρώπινες κριτικές από τις έρευνες.

Για την εξαγωγή ενός συνόλου αξιολόγησης ομοιότητας, συνήθως οι ερωτήσεις είναι της δομής, “πόσο σχετική είναι η λέξη α με τη λέξη β σε κλίμακα από το x ως το y”. Οι ΔΠΛ μπορούν να απεικονιστούν στο δισδιάστατο χώρο όπως στην παρακάτω εικόνα (π.χ. χρησιμοποιώντας Ανάλυση Κυρίων Συνιστωσών - PCA). Σχετικές λέξεις πρέπει να βρίσκονται σε κοντινή απόσταση.



Εικόνα 11: Ομοιότητα Λέξεων

Αντίστοιχα, για την αναλογία λέξεων οι ερωτήσεις είναι της μορφής, “Η λέξη *x* είναι για τη λέξη *y* όπως ποια λέξη για τη λέξη *z*;”. Για παράδειγμα, “Ο βασιλιάς είναι για τον άντρα, όπως τι; για τη γυναίκα;”.



Εικόνα 12: Αναλογία Λέξεων

Απαιτείται αρκετός χρόνος και πόροι για τη δημιουργία συνόλων αξιολόγησης όπως τα παραπάνω και δυστυχώς δεν υπάρχουν ακόμη για την Ελληνική γλώσσα.

4.3 Γλωσσική Μοντελοποίηση

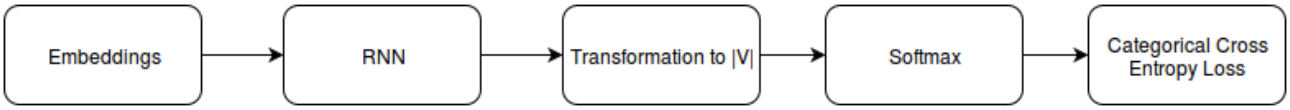
Η γλωσσική μοντελοποίηση είναι ένα από τα πιο συνηθισμένα και σημαντικά προβλήματα για την ΕΦΓ. Δοθέντος ενός μεγάλου σώματος εκπαίδευσης, εκπαιδεύεται ένα μοντέλο ώστε έπειτα να είναι σε θέση δοθείσης μίας ακολουθίας λέξεων να αποδώσει πιθανότητες σε όλες τις λέξεις του λεξιλογίου στις οποίες έχει εκπαιδευτεί. Η κάθε πιθανότητα από αυτές αντικατοπτρίζει τη βεβαιότητα του μοντέλου η αντίστοιχη λέξη να είναι η επόμενη λέξη της ακολουθίας που του δόθηκε. Ένα καλό γλωσσικό μοντέλο πρέπει να αποδίδει μεγάλη πιθανότητα στην πραγματικά επόμενη λέξη και μικρές πιθανότητες σε όλες τις άλλες. Τα πιο συνηθισμένα μέτρο αξιολόγησης ενός γλωσσικού μοντέλου είναι η διασταυρωμένη εντροπία ή αναλόγως η περιπλοκή (perplexity).

Παραδοσιακά, τα γλωσσικά μοντέλα υλοποιούνταν με ν-γράμματα και εξομάλυνση (για παράδειγμα Laplace, Kneser-Nay κ.ά.). Μάλιστα, τα γλωσσικά μοντέλα ν-γραμμάτων με εξομάλυνση Kneser-Nay αποδίδουν ακόμη και σήμερα πολύ καλά αποτελέσματα. Στην περίπτωση των ΔΠΛ, η γλωσσική μοντελοποίηση γίνεται με τη χρήση Νευρωνικών Δικτύων και πιο συνηθισμένα Ανατροφοδοτούμενων Νευρωνικών Δικτύων. Πρόσφατες έρευνες όμως έχουν αποδείξει ότι η χρήση (πολλαπλών) Ανατροφοδοτούμενων Νευρωνικών Δικτύων επιφέρουν

θεαματική μείωση στην περιπλοκή [Jozefowicz κ.ά. 2016].

Για τις ανάγκες της εργασίας δημιουργήθηκαν δύο μοντέλα αξιολόγησης ΔΠΛ. Το πρώτο μοντέλο είναι ένα από Ανατροφοδοτούμενο Νευρωνικό Δίκτυο το οποίο δέχεται ως είσοδο σειριακά παράθυρα λέξεων σταθερού μεγέθους κι εκπαιδεύεται ώστε να προβλέπει την επόμενη λέξη. Αρχικά οι λέξεις του παραθύρου παριστάνονται με τις τρέχουσες ΔΠΛ τους. Έπειτα, περνάνε ως είσοδος στο Ανατροφοδοτούμενο Νευρωνικό Δίκτυο το οποίο έχει ένα κρυφό διάνυσμα διάστασης |h|. Στη συνέχεια, πρέπει να γίνει μετατροπή πίσω στο μέγεθος ολόκληρου του λεξιλογίου. Αυτό επιτυγχάνεται με τη χρήση ενός πυκνού πίνακα W_out διαστάσεων |h| x |v|, όπου v το λεξιλόγιο. Εντέλει, οι προβλέψεις που προκύπτουν μετασχηματίζονται σε πιθανότητες με τη συνάρτηση softmax και το σφάλμα προκύπτει με τη χρήση κατηγορικής διασταυρωμένης εντροπίας.

Παρακάτω παρατίθεται μία απεικόνιση για τη δομή του Νευρωνικού Δικτύου.

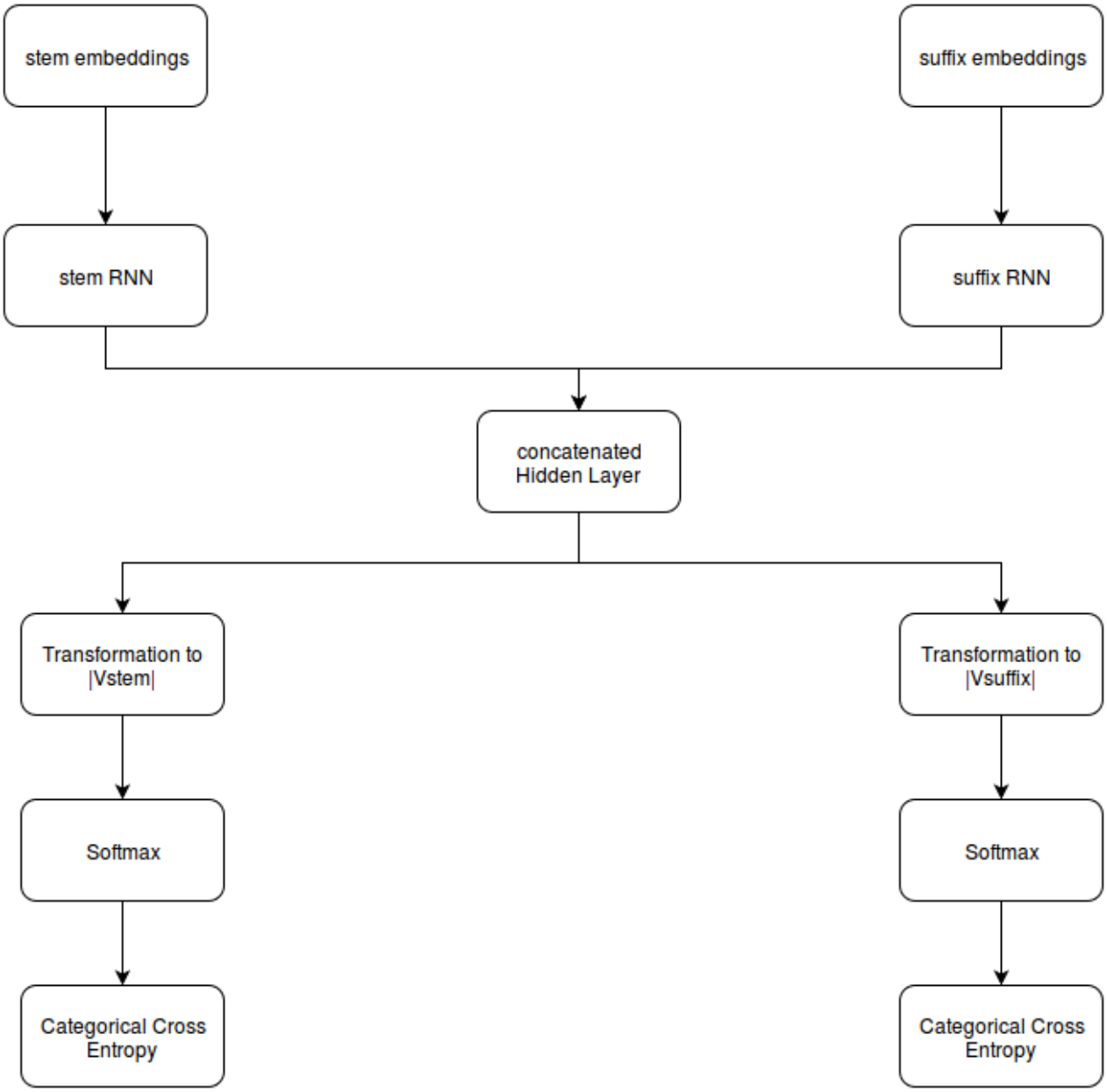


Εικόνα 13: Απλό Ανατροφοδοτούμενο Νευρωνικό Δίκτυο για Γλωσσική Μοντελοποίηση

Το πρόβλημα με αυτή την προσέγγιση είναι ο εξαιρετικά μεγάλος χρόνος εκπαίδευσης, ο οποίος οφείλεται στο πλήθος των πράξεων από τη συνάρτηση softmax, το οποίο με τη σειρά του αυξάνεται γραμμικά με το πλήθος του λεξιλογίου. Το πλήθος του λεξιλογίου σε μορφολογικά πλούσιες γλώσσες όπως στα Ελληνικά είναι πολύ υψηλό. Για το λόγο αυτό σχεδιάστηκε και αναπτύχθηκε ένα νέο Νευρωνικό Δίκτυο Γλωσσικής Μοντελοποίησης.

Η παραλλαγή αυτή, αντί για ολόκληρες λέξεις σε ένα παράθυρο δέχεται τα μορφήματα (θέμα και κατάληξη) τα οποία τις απαρτίζουν. Τα μορφήματα στη συνέχεια απεικονίζονται στις δοθείσες διανυσματικές παραστάσεις τους και περνάνε ως είσοδος σε δύο ξεχωριστά Ανατροφοδοτούμενα Νευρωνικά Δίκτυα, ένα για τα θέματα και ένα για τις καταλήξεις. Έπειτα, τα κρυφά επίπεδα των δύο Ανατροφοδοτούμενων Νευρωνικών Δικτύων συνενώνονται, ώστε να επιτευχθεί η επικοινωνία μεταξύ των διανυσματικών παραστάσεων των θεμάτων και των καταλήξεων. Ακολούθως, δημιουργούνται δύο πυκνοί πίνακες (W_out_stem και W_out_suffix), οι οποίοι μετατρέπουν το κοινό κρυφό επίπεδο πίσω στο μέγεθος του λεξιλογίου των θεμάτων και των καταλήξεων αντίστοιχα. Εντέλει, τα διανύσματα που προκύπτουν από αμφότερους τους πίνακες μετασχηματίζονται με τη συνάρτηση softmax και το σφάλμα προκύπτει με τη χρήση της κατηγορικής διασταυρωμένης εντροπίας, όπου παράγονται ξεχωριστές πιθανότητες για το θέμα και την κατάληξη της επόμενης λέξης. Η πιο πιθανή επόμενη λέξη είναι η συνένωση του πιο πιθανού επόμενου θέματος και της πιο πιθανής επόμενης κατάληξης.

Παρακάτω παρατίθεται μία απεικόνιση για τη δομή του Νευρωνικού Δικτύου.



Εικόνα 14: Ανατροφοδοτούμενο Νευρωνικό Δίκτυο μορφημάτων για Γλωσσική Μοντελοποίηση

4.5 Επισημείωση Μερών του Λόγου

Η Επισημείωση Μερών του Λόγου (ΕΜΛ) είναι ένα πρόβλημα της ΕΦΓ στο οποίο οι λέξεις μίας πρότασης πρέπει να καταταγούν στα μέρη του λόγου. Για παράδειγμα, στην πρόταση: “το αυτοκίνητο είναι κόκκινο”, η λέξη “είναι” πρέπει να καταταγεί ως “ρήμα”.

Στην παρούσα εργασία δεν υλοποιήθηκε σύστημα ΕΜΛ. Η αξιολόγηση έγινε στο σύστημα που αναπτύχθηκε στα πλαίσια της Διπλωματικής Εργασίας του Θωμά Ασίκη [Ασίκης 2016]. Διερευνήθηκε, δηλαδή, αν οι ΔΠΛ της παρούσας εργασίας βοηθούν στα συστήματα ΕΜΛ του Ασίκη.

5. Πειράματα

5.1 Δεδομένα

Η πηγή δεδομένων για τα πειράματα ήταν η βάση δεδομένων της Wikipedia για την Αγγλική και την Ελληνική γλώσσα. Τα δεδομένα υπέστησαν επεξεργασία ώστε να αφαιρεθούν οι ετικέτες XML και όλοι οι χαρακτήρες μετατράπηκαν σε πεζούς. Για αμφότερες τις γλώσσες, οι πρώτες 400.000 λέξεις χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης των Γλωσσικών Μοντέλων, οι επόμενες 20.000 λέξεις ως σύνολο επικύρωσης των Γλωσσικών Μοντέλων και οι επόμενες 20.000 λέξεις ως σύνολο αξιολόγησης των Γλωσσικών Μοντέλων. Όλες οι υπόλοιπες λέξεις χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης των ΔΠΛ. Παρακάτω παρατίθεται ένας συγκριτικός πίνακας ανάμεσα στις γλώσσες με στατιστικά στοιχεία από το σύνολο εκπαίδευσης των ΔΠΛ.

	Ελληνικά	Αγγλικά
Συνολικό Πλήθος Λέξεων	38.429.249	83.130.306
Πλήθος Μοναδικών Λέξεων	212.491	165.069
Πλήθος Μοναδικών Θεμάτων	101.972	116.725
Πλήθος Μοναδικών Καταλήξεων	61.829	99.062

Πίνακας 1: Στατιστικά Στοιχεία Συνόλου Εκπαίδευσης ΔΠΛ

Είναι αξιοσημείωτο ότι παρόλο που τα δεδομένα για την Ελληνική γλώσσα είναι λιγότερο από τα μισά σε σχέση με την Αγγλική, οι μοναδικές λέξεις στα Ελληνικά υπερβαίνουν τις Αγγλικές κατά περίπου 50.000. Από αυτό, σε συνδυασμό με το γεγονός ότι το πλήθος μοναδικών θεμάτων και καταλήξεων στα Ελληνικά είναι μικρότερο από εκείνο των Αγγλικών, προκύπτει επίσης ότι τα θέματα και οι καταλήξεις στα Ελληνικά συνδυάζονται με μεγαλύτερη ποικιλία μεταξύ τους.

5.2 Εκπαίδευση

Η εκπαίδευση των αλγορίθμων που χρησιμοποιούν το word2vec εκτελείται σε συμβατική Κεντρική Μονάδα Επεξεργασίας (CPU). Η εκπαίδευση των αλγορίθμων οι οποίοι χρησιμοποιούν το keras και τη theano εκτελείται σε μεγάλο μέρος στην Κάρτα Γραφικών (GPU). Αν και οι χρόνοι εκπαίδευσης για τους αλγορίθμους των keras / theano μειώθηκαν αισθητά με τη χρήση της κάρτας γραφικών, παραμένουν κατά πολύ χειρότεροι σε σχέση με το word2vec. Αυτό οφείλεται αφενός στην επιλογή της γλώσσας προγραμματισμού (C για το word2vec έναντι python για keras / theano) κι αφετέρου στο γεγονός ότι το word2vec έχει βελτιστοποιηθεί για απόδοση. Παρ' όλ' αυτά, υπάρχει αρκετός χώρος για βελτίωση της απόδοσης στους αλγορίθμους για keras / theano.

Αλγόριθμος	Διάσταση	Χρόνος (σε λεπτά)
Mikolov's CboW	50	5
Mikolov's CboW	150	12
Mikolov's Cbow	300	30
Stem-Suffix	50	8
Stem-Suffix	150	16
Stem-Suffix	300	36
Morph-CboW	50	600 (10 ώρες)
Morph-CboW	150	~2000 (~32 ώρες)
Morph-CboW	300	~5500 (~90 ώρες)

Πίνακας 2: Μέτρηση χρόνων εκπαίδευσης για τα Ελληνικά

Δεν έγινε εκπαίδευση για τις διαστάσεις των 150 και 300 ΔΠΛ στο μοντέλο Morph-CboW, λόγω του μεγάλου χρόνου εκπαίδευσης. Ο χρόνος εκπαίδευσης αυξάνεται σχεδόν γραμμικά με το πλήθος των διάστασεων, το οποίο είναι αναμενόμενο. Δεν είναι όμως αναμενόμενο να αυξάνεται γραμμικά με το μέγεθος του λεξιλογίου. Αυτό το φαινόμενο είναι μία γνωστή ανεπάρκεια της βιβλιοθήκης theano.

5.3 Ομοιότητα Λέξεων

Για το πρόβλημα της ομοιότητας λέξεων χρησιμοποιήθηκαν πέντε σύνολα αξιολόγησης για την Αγγλική γλώσσα τα οποία έχουν χρησιμοποιηθεί ευρέως στην ΕΦΓ (EN-MC30, WordSim353, EN-RG65, SCWS, RareWords). Το τελευταίο σύνολο, δηλαδή το σύνολο των σπάνιων λέξεων (rare words), δημιουργήθηκε ειδικά για την επίδειξη των μειονεκτημάτων των καθιερωμένων ΔΠΛ όταν παριστάνονται (μορφολογικά) σπάνιες λέξεις, έναντι των ΔΠΛ που λαμβάνουν υπόψιν τη μορφολογία [Luong κ.ά. 2013].

Στον πίνακα που ακολουθεί φαίνονται τα αποτελέσματα των αλγορίθμων που χρησιμοποιήθηκαν στα παραπάνω σύνολα. Το μέτρο αξιολόγησης είναι η συσχέτιση ρ του Spearman πολλαπλασιασμένη με 100 με στρογγυλοποίηση δεύτερου δεκαδικού. Παρουσιάζονται δύο εκδοχές του μοντέλου Morph-CboW, μία χωρίς προεκπαίδευση (ΧΠ), όπου οι ΔΠΛ (των ολόκληρων λέξεων) αρχικοποιήθηκαν σε τυχαίους αριθμούς και μία με προεκπαίδευση (ΜΠ), όπου οι ΔΠΛ αρχικοποιήθηκαν με τις ΔΠΛ που προέκυψαν από το CboW. Οι διανυσματικές παραστάσεις των ψευδο-μορφημάτων (ψευδο-θέμα, ψευδο-κατάληξη) αρχικοποιούνται πάντα τυχαία.

	EN-MC30	WordSim353	EN-RG65	SCWS	Rare Words
CboW 50	49.40	43.26	36.53	52.42	14.34
CboW 150	**67.82**	**57.54**	**71.11**	63.40	21.83
CboW 200	65.11	56.00	71.00	**63.42**	**22.00**
Stem-Suffix 50	**67.82**	55.76	67.13	60.57	20.00
Stem-Suffix 150	62.37	49.48	50.43	54.45	17.68
Stem-Suffix 200	63.15	50.47	47.63	55.52	18.25
Morph-CboW 50 (ΧΠ)	10.66	5.80	-1.37	20.30	-3.77
Morph-Cbow 50 (ΜΡ)	66.98	56.08	66.14	60.80	19.37

Πίνακας 3: Αποτελέσματα συσχέτισης ρ του Spearman σε πειράματα ομοιότητας λέξεων

Γίνεται αμέσως εμφανές ότι μεγαλύτερο πλήθος διαστάσεων των ΔΠΛ δε συνεπάγεται απαραίτητα καλύτερα αποτελέσματα. Όπως φαίνεται, ο αλγόριθμος CboW επωφελείται περισσότερο από τις περισσότερες διαστάσεις, δεδομένου ότι για 50 διαστάσεις τα αποτελέσματα ήταν αρκετά χαμηλότερα σε σχέση με τις 150. Επίσης, για 50 διαστάσεις ο αλγόριθμος Morph-CboW αυτής της εργασίας με προεκπαίδευση (ΜΠ) παρήγαγε καλύτερα αποτελέσματα, συγκρίνοντας με τον CboW, κάτι που είναι ιδιαίτερα ενθαρρυντικό. Δυστυχώς δε στάθηκε δυνατό να συγκρίνουμε τον Morph-CboW με τον CboW για περισσότερες διαστάσεις, λόγω της χαμηλής ταχύτητας της υλοποίησης του Morph-CboW. Εντέλει, φαίνεται τεράστια διαφορά στον αλγόριθμο Morph-CboW όταν χρησιμοποιείται προεκπαίδευση (ΜΠ). Ενδεχομένως να χρειάζονται περισσότερες επαναλήψεις κατά την εκπαίδευση για να μπορέσει να παράγει καλύτερα αποτελέσματα χωρίς προεκπαίδευση (ΧΠ).

5.4 Αναλογία Λέξεων

Για το πρόβλημα της αναλογίας λέξεων χρησιμοποιήθηκε το σύνολο αξιολόγησης που εισήγαγαν ο Mikolov και οι συνεργάτες του [Mikolov κ.ά. 2013γ]. Είναι ένα σύνολο με τετράδες λέξεων για κάθε μία από τις οποίες το σύστημα δοθέντων των τριών πρώτων πρέπει να προβλέψει την τέταρτη. Το σύνολο χωρίζεται σε 14 κατηγορίες ως ακολούθως:

Όνομα Κατηγορίας	**Πλήθος**
Πόλη σε Πολιτεία (αμερικάνικη)	2467
Οικογένεια	506
Πρωτεύουσα σε χώρα (Παγκοσμίως)	4524
Πρωτεύουσα σε χώρα (Μεγάλες Χώρες)	506
Νόμισμα σε χώρα	866
Επίθετο σε επίρρημα	992
Αντίθετα	812
Συγκριτικός βαθμός	1332
Υπερθετικός βαθμός	1122
Γερούνδιο	1056
Εθνικότητα σε επίθετο	1599
Αόριστος χρόνος	1560
Πληθυντικός αριθμός (όχι σε ρήματα)	1332
Πληθυντικός αριθμός (σε ρήματα)	870

Πίνακας 4: Κατηγορίες και Πλήθος αναλογίας λέξεων

Στο πρόβλημα της αναλογίας λέξεων, οι ΔΠΛ που παρήχθησαν από τους αλγορίθμους αυτής της εργασίας δεν απέδωσαν καλά αποτελέσματα. Πιο συγκεκριμένα, το μοντέλο “stem-suffix” συνήθως ήταν χειρότερο από το CboW, ενώ το το μοντέλο “morph-cbow-alg” πολύ σπάνια επέστρεφε σωστή απάντηση. Στις κατηγορίες “Πόλη σε Πολιτεία (αμερικάνικη)”, “Πρωτεύουσα σε χώρα (παγκοσμίως)”, “Πρωτεύουσα σε χώρα (Μεγάλες Χώρες)” και “Εθνικότητα σε επίθετο” δεν επιστρέφονταν ποτέ σωστές απαντήσεις με οποιοδήποτε μοντέλο, συμπεριλαμβανομένου του κλασικού CboW. Ενδεχομένως ευθύνεται το σύνολο εκπαίδευσης ή οι επιλογές παραμέτρων του CboW.

Όσον αφορά το μοντέλο “stem-suffix”, μελετήθηκε η επίδραση του πλήθους των διαστάσεων των ΔΠΛ για το θέμα και την κατάληξη. Παρακάτω παρουσιάζονται τα αποτελέσματα πειραμάτων αναλογίας λέξεων για τρία διαφορετικά μοντέλα, όπου οι διαστάσεις των θεμάτων και καταλήξεων έχουν τεθεί αντίστοιχα σε (100-100), (150-50), (50-150). Επιπλέον, για λόγους σύγκρισης παρουσιάζονται τα αποτελέσματα του κλασικού CboW με 200 διαστάσεις, ώστε η σύγκριση να είναι δίκαιη.

Όπως φαίνεται, το κλασικό CboW πετυχαίνει καλύτερα αποτελέσματα σχεδόν σε όλες τις κατηγορίες και συνολικά προηγείται με αρκετή απόσταση από το Stem-Suffix (100-100). Επίσης, γίνεται εμφανές ότι οι διαστάσεις των θεμάτων είναι πολύ πιο σημαντικός παράγοντας για το πρόβλημα της αναλογίας λέξεων σε σχέση με τις διαστάσεις των καταλήξεων. Πράγματι, η διαφορά στο ποσοστό ευστοχίας μεταξύ του ισορροπημένου μοντέλου (100-100) κι εκείνου το οποίο ευνοεί τα θέματα (150-50) είναι αρκετά μικρότερη σε σύγκριση με εκείνο το οποίο ευνοεί τις καταλήξεις (50-150).

Κατηγορία	Mikolov's CboW	Stem-Suffix 100 - 100	Stem-Suffix 150 - 50	Stem-Suffix 50 - 150
Πόλη σε Πολιτεία (Αμερικάνικη)	0	0	0	0
Οικογένεια	**330**	201	197	191
Πρωτεύουσα σε χώρα (Παγκοσμίως)	0	0	0	0
Πρωτεύουσα σε χώρα (Μεγάλες Χώρες)	0	0	0	0
Νόμισμα σε χώρα	0	11	11	**13**
Επίθετο σε επίρρημα	206	281	**303**	199
Αντίθετα	191	**229**	216	204
Συγκριτικός βαθμός	**1030**	823	790	780
Υπερθετικός βαθμός	**475**	423	437	375
Γερούνδιο	**657**	507	499	433
Εθνικότητα σε επίθετο	0	0	0	0
Αόριστος χρόνος	**807**	316	340	236
Πληθυντικός αριθμός (όχι σε ρήματα)	**831**	470	433	460
Πληθυντικός αριθμός (σε ρήματα)	450	**452**	421	427
Ποσοστό Ευστοχίας	25.47%	19.00%	18.66%	16.98%
Ποσοστό Ευστοχίας εξαιρώντας τις μηδενικές κατηγορίες	47.64%	35.54%	34.91%	31.76%

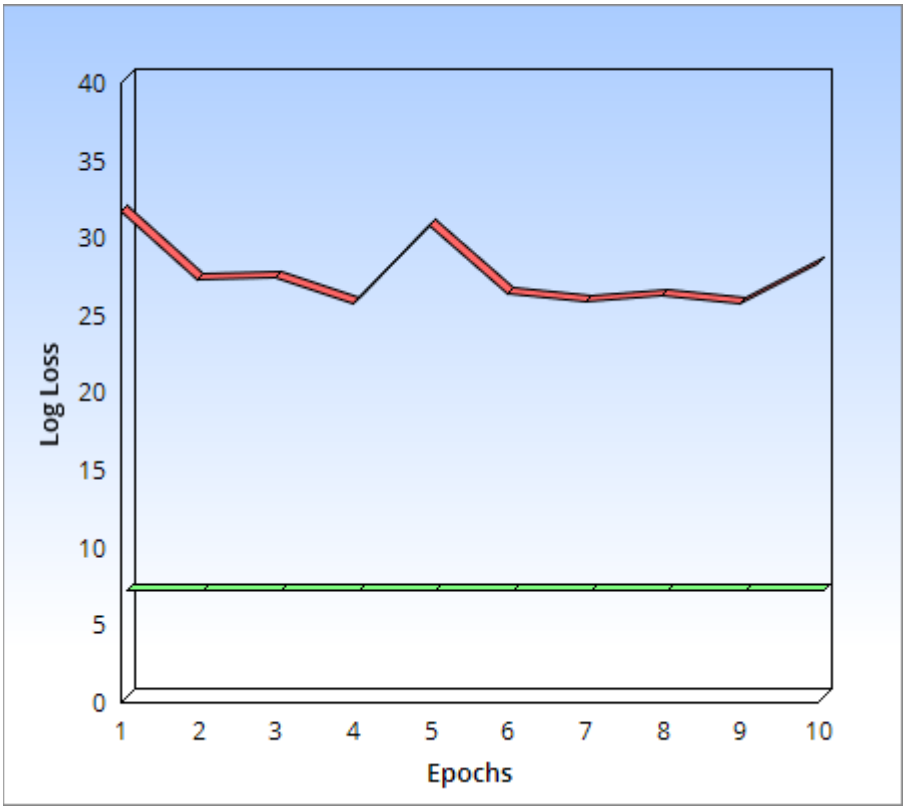
Πίνακας 5: Αποτελέσματα Αναλογίας Λέξεων

5.5 Γλωσσική Μοντελοποίηση

Όπως προαναφέρθηκε, η εκπαίδευση για τα μοντέλα της Γλωσσικής Μοντελοποίησης έγινε στις πρώτες 400.000 λέξεις των κειμένων για κάθε γλώσσα. Τα σύνολα επικύρωσης αποτελούνταν από τις επόμενες 20.000 λέξεις και τα σύνολα αξιολόγησης από τις επόμενες 20.000 λέξεις.

Γίνεται σύγκριση με ένα γλωσσικό μοντέλο τρι-γραμμάτων (3-grams) με τη χρήση εξωμάλυνσης Kneser-Nay. Το μοντέλο τρι-γραμμάτων εκπαιδεύτηκε με τη χρήση του λογισμικού SRILM. Το εναλλακτικό Ανατροφοδοτούμενο Νευρωνικό Δίκτυο που υλοποιήθηκε για το σκοπό αυτό προβλέπει μορφήματα κι επομένως ενδεχομένως να υπάρχουν περιπτώσεις όπου να προβλέπεται μόνο ένα σωστό μόρφημα της εκάστοτε λέξης. Για να είναι δίκαιη η σύγκριση με το μοντέλο τρι-γραμμάτων, απαιτείται να ανήκουν στο λεξιλόγιο του Ανατροφοδοτούμενου Νευρωνικού Δικτύου αμφότερα τα μορφήματα που απαρτίζουν την εκάστοτε λέξη. Σε αντίθετη περίπτωση θέτονται από κοινού σε “άγνωστη λέξη”. Επίσης, εκτός από το σύνολο εκπαίδευσης προορισμένο για τη Γλωσσική Μοντελοποίηση, η εκπαίδευση του μοντέλου τρι-γραμμάτων έγινε και στο σύνολο εκπαίδευσης των ΔΠΛ.

Η εκπαίδευση των μοντέλων με Νευρωνικά Δίκτυα έγινε σε δέκα εποχές. Για να αποφευχθεί η υπερπροσαρμογή των μοντέλων στα δεδομένα εκπαίδευσης, σε κάθε επανάληψη το παραγόμενο μοντέλο δοκιμαζόταν στο σύνολο επικύρωσης. Ως τελικό μοντέλο επιλεγόταν εκείνο το οποίο είχε τα καλύτερα αποτελέσματα στο σύνολο επικύρωσης. Λόγω έλλειψης χρόνου εκπαιδεύτηκαν γλωσσικά μοντέλα μόνο για ΔΠΛ με διάσταση 50. Παρακάτω παρουσιάζονται οι καμπύλες μάθησης κατά την εκπαίδευση του μοντέλου Stem-Suffix στα Αγγλικά. Η πράσινη (κάτω) καμπύλη παρουσιάζει το σφάλμα στο σύνολο εκπαίδευσης. Η κόκκινη (πάνω) καμπύλη παρουσιάζει το σφάλμα στο σύνολο επικύρωσης.



Εικόνα 15: Καμπύλες Μάθησης για τη Γλωσσική Μοντελοποίηση

Όπως φαίνεται, οι δύο καμπύλες απέχουν πολύ, το οποίο ενδεχομένως οφείλεται στο μικρό αριθμό εποχών, ίσως και στο μικρό πλήθος διαστάσεων. Αυτή η παρατήρηση επαληθεύεται από τα αποτελέσματα στο σύνολο αξιολόγησης. Σε μελλοντική εργασία θα μελετηθούν περαιτέρω αυτά τα προβλήματα.

Μοντέλο	Περιπλοκή
Αγγλικά Τριγράμματα	323.26
Ελληνικά Τριγράμματα	385.83
Αγγλικά – Stem-Suffix 50	2^{28}
Ελληνικά – Stem-Suffix 50	2^{30}
Αγγλικά – Morph-CboW 50	2^{27}
Ελληνικά – Morph-CboW 50	2^{28}

Πίνακας 6: Αποτελέσματα περιπλοκής κατά τη γλωσσική μοντελοποίηση

5.6 Επισημείωση Μερών του Λόγου

Για την αξιολόγηση στο πρόβλημα των ΕΜΛ, η οποία όπως προαναφέρθηκε έγινε στο σύστημα το οποίο αναπτύχθηκε στα πλαίσια της Εργασίας του Θωμά Ασίκη [Ασίκης 2016], χρησιμοποιήθηκαν οι ΔΠΛ οι οποίες προέκυψαν από την εκπαίδευση στην Ελληνική γλώσσα του πρώτου συστήματος (Stem-Suffix). Οι διαστάσεις των ΔΠΛ που δοκιμάστηκαν ήταν 50, 150 και 300.

Ο αλγόριθμος ταξινόμησης ο οποίος χρησιμοποιήθηκε είναι αυτός της Μέγιστης Εντροπίας. Τα αποτελέσματα ήταν ανταγωνιστικά με εκείνα των ΔΠΛ από το κλασικό word2vec. Το σύστημα δοκιμάστηκε σε δύο σύνολα ετικετών, ένα περιορισμένο και ένα εκτεταμένο. Στο εκτεταμένο σύνολο οι ετικέτες προσδιορίζουν περισσότερα γνωρίσματα των λέξεων, όπως γένος, αριθμός κ.ά. Προφανώς, η ανάθεση ετικετών στο περιορισμένο σύνολο θεωρείται πιο βατό πρόβλημα. Παρακάτω φαίνονται τα αποτελέσματα των πειραμάτων που εκτελέστηκαν για τα δύο σύνολα ετικετών.

Αλγόριθμος – Διάσταση ΔΠΛ	Ευστοχία – Περιορισμένο Σύνολο	Ευστοχία – Εκτεταμένο Σύνολο
Word2vec - 50	93.92%	79.11%
Word2vec - 150	94.52%	81.05%
Word2vec - 300	95.39%	83.37%
Stem-Suffix - 50	93.29%	71.32%
Stem-Suffix - 150	93.64%	73.35%
Stem-Suffix - 300	93.38%	74.67%
Stems - 25	90.31%	61.94%
Stems - 75	91.30%	62.52%
Stems - 150	91.17%	62.46%
Suffixes - 25	91.70%	72.05%
Suffixes - 75	93.16%	71.66%
Suffixes - 150	93.24%	72.81%

Πίνακας 7: Ακρίβεια κατά την Επισημείωση Μερών του Λόγου

Σε μία γλώσσα όπως στα Ελληνικά, οι καταλήξεις δίνουν πολλή πληροφορία για την ετικέτα μίας λέξης. Για παράδειγμα, μία λέξη η οποία έχει κατάληξη “-μένος”, είναι σχεδόν βέβαιο ότι είναι μετοχή, αρσενικού γένους και ενικού αριθμού. Αυτό επαληθεύεται από τα αποτελέσματα, όπου η ευστοχία για τις ΔΠΛ των καταλήξεων ήταν πάντα υψηλότερη από εκείνη για τις ΔΠΛ των θεμάτων.

6. Συμπεράσματα και μελλοντική έρευνα

6.1 Συμπεράσματα

Η παρούσα εργασία είχε ως στόχο τη βελτιωμένη παραγωγή ΔΠΛ για μορφολογικά πλούσιες γλώσσες, όπως τα Ελληνικά. Μελετήθηκε κατά πόσο μία απλοϊκή διάσπαση σε δύο μόνο μορφήματα, ένα θέμα και μία κατάληξη, για κάθε λέξη θα μπορούσε να επιτύχει καλά αποτελέσματα για αυτό το πρόβλημα. Αναπτύχθηκαν δύο νέοι αλγόριθμοι παραγωγής ΔΠΛ, οι οποίοι λαμβάνουν υπόψιν τη μορφολογία των λέξεων (stem-suffix και morph-CboW). Η αξιολόγηση έγινε στα προβλήματα της ομοιότητα λέξεων (Αγγλικά), της αναλογίας λέξεων (Αγγλικά), της Γλωσσικής Μοντελοποίησης (Ελληνικά και Αγγλικά) και της Επισημείωσης Μερών του Λόγου (Ελληνικά).

Όπως αποδείχθηκε, τα μοντέλα που αναπτύχθηκαν ήταν πολύ πιο χρονοβόρα από το αναμενόμενο. Επίσης, στα περισσότερα πειράματα δεν επετεύχθη βελτίωση των αποτελεσμάτων σε σχέση με διαδεδομένες προσεγγίσεις παραγωγής ΔΠΛ όπως το word2vec. Σε αυτό ενδεχομένως να συνέβαλε η επιλογή του αλγορίθμου διάσπασης, ο οποίος ενδεχομένως να ήταν υπερβολικά απλοϊκός. Επίσης, η επιλογή του περιορισμού αποκλειστικά σε δύο μορφήματα φαίνεται να μη λειτουργεί καλά στην πράξη.

Πάρα ταύτα, υπάρχουν πολλά περιθώρια για βελτίωση τόσο της απόδοσης όσο και των αποτελεσμάτων και δοκιμή επιπλέον αλγορίθμων διάσπασης. Επίσης, τα μοντέλα τα οποία αναπτύχθηκαν μπορούν με μικρές αλλαγές να χειριστούν παραπάνω από δύο μορφήματα. Επιπλέον, τα αποτελέσματα στην ομοιότητα λέξεων για το morph-CboW 50 (ΜΠ) ήταν αρκετά ενθαρρυντικά.

6.2 Μελλοντικές Προσεγγίσεις

Οι μεγάλες απαιτήσεις χρόνου εκπαίδευσης μπορούν να αντιμετωπιστούν εν μέρει με τη χρήση πολλαπλών νημάτων (threads). Δεδομένου ότι η υλοποίηση βασίζεται σε τρίτες βιβλιοθήκες, η απόδοση των μοντέλων εξαρτάται άμεσα από την απόδοση των βιβλιοθηκών αυτών.

Όσον αφορά τα ίδια τα μοντέλα, θα μπορούσαν να δοκιμαστούν εναλλακτικοί τρόποι εξαγωγής μορφημάτων. Επίσης, εφόσον το επιτρέψει ο χρόνος εκπαίδευσης, είναι ενδιαφέρον να μελετηθεί αν και κατά πόσο βελτιώνονται οι ΔΠΛ κατά τη Γλωσσική Μοντελοποίηση.

Ακόμη, θα μπορούσε να δοκιμαστεί η εισαγωγή ενός τρίτου βάρους κατά τον υπολογισμό του κρυφού στρώματος στον αλγόριθμο “morph-CboW”, ώστε αντί για την τρέχουσα υλοποίηση στην οποία υπάρχει ένα βάρος για ολόκληρες τις λέξεις και ένα για τα μορφήματα, να υπάρχει ένα αποκλειστικό βάρος για ολόκληρες τις λέξεις, τα θέματα και τις καταλήξεις.

Επίσης, θα είχε πολύ μεγάλο ενδιαφέρον να δοκιμαστούν οι αλγόριθμοι στην ομοιότητα και την αναλογία λέξεων για την Ελληνική γλώσσα, η οποία είναι μορφολογικά πλουσιότερη από την Αγγλική. Για να επιτευχθεί αυτό, ένα πρώτο βήμα θα μπορούσε να είναι η μετάφραση των Αγγλικών συνόλων εκπαίδευσης.

Εντέλει, θα μπορούσε στο εναλλακτικό μοντέλο Γλωσσικής Μοντελοποίησης να λαμβάνονται υπόψιν και οι ΔΠΛ ολόκληρων των λέξεων κατά την εκπαίδευση και την πρόβλεψη, ακόμη και αν στο τελικό στρώμα το μοντέλο συνεχίζει να προβλέπει μόνο ζεύγη θεμάτων και καταλήξεων ώστε να μην απαιτηθεί πολύ περισσότερος χρόνος.

Αναφορές

F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, Y. Bengio, 'Theano: new features and speed improvements', *Neural Information Processing Systems Workshop*, 2012

Y. Bengio, P. Simard, P. Frasconi, "Learning long-term dependencies with gradient descent is difficult", *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994

Y. Bengio, R. Ducharme, P. Vincent, "A neural probabilistic language model", *Journal of Machine Learning Research*, 3:1137-1155, 2003

X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", *International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, pp 249-256, 2010

S. Hochreiter, J. Schmidhuber, "Long short-term memory", *Neural computation*, 9(8):1735-1780, 1997

S. Ji, S. V. N. Vishwanathan, N. Satish, M. J. Anderson, P. Dubey, *BlackOut: Speeding up Recurrent Neural Network Language Models With Very Large Vocabularies*, http://arxiv.org/abs/1511.06909, 2015

R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, Y. Wu, *Exploring the Limits of Language Modeling*, http://arxiv.org/abs/1602.02410, 2016

Y. LeCun, L. Bottou, G. Orr, K. Muller, *Neural Networks: Tricks of the Trade*, Springer Berlin Heidelberg, 1998

W. Ling, T. Luis, L. Marujo, R. Astudillo, S. Amir, C. Dyer, A. Black, I. Trancoso, "Finding Function in Form: Compositional Character Models forOpen Vocabulary Word Representation", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp 1520-1530, 2015

M. Luong, R. Socher, D. Manning, "Better word representations with recursive neural networks for morphology", *Conference on Computational Natural Language Learning*, Sofia, Bulgaria, pp 104-113, 2013

T. Mikolov, W. Yih, G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, pp 746-751, 2013

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, Lake Tahoe, Nevada, USA, pp 3111-3119, 2013

T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient estimation of word representations in vector space*, http://arxiv.org/abs/1301.3781, 2013

R. Pascanu, T. Mikolov, Y. Bengio, "On the difficulty of training recurrent neural networks", *International Conference on Machine Learning*, Atlanta, Georgia, USA, pp 1310–1318, 2013

S. Qiu, Q. Cui, J. Bian, B. Gao, T. Liu, "Co-learning of Word Representations and Morpheme Representations", *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, pp 141–150, 2014

T. Schnabel, I. Labutov, D. Mimno, T. Joachims, "Evaluation methods for unsupervised word embeddings", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp 298-307, 2015

Ασίκης Θωμάς, *Επισημείωση Μερών του Λόγου με τη χρήση Μηχανικής Μάθησης*, Οικονομικό Πανεπιστήμιο Αθηνών, Αθήνα, 2016