ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη ρομπότ συνομιλίας με τεχνικές μηχανικής μάθησης

Διπλωματική εργασία

Μυρτώ Τσοκαναρίδου

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

Αθήνα, Ιούλιος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη ρομπότ συνομιλίας με τεχνικές μηχανικής μάθησης

Διπλωματική εργασία

Μυρτώ Τσοκαναρίδου

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 10η Ιουλίου 2019.

(υπογραφή)	*(υπογραφή)*	*(υπογραφή)*
……..	……..	……..
Γιώργος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.	Ανδρέας – Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Νικόλαος Παπασπύρου Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

(υπογραφή)

……..

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

Περίληψη

Η δυνατότητα ενός υπολογιστή να διεξάγει διάλογο σαν ανθρώπινο πρόσωπο είναι μία από τις μεγαλύτερες προκλήσεις -αν όχι η μεγαλύτερη- που αντιμετωπίζει ο τομέας της Τεχνητής Νοημοσύνης. Στην παρούσα εργασία, έχοντας υπ' όψιν το πόσο πολύπλοκο και ευρύ είναι αυτό το ζήτημα, επιχειρήθηκε η δημιουργία ενός ρομπότ συνομιλίας (chatbot) η λειτουργία του οποίου βασίζεται αποκλειστικά σε τεχνικές μηχανικής μάθησης.

Η αρχική δομή νευρωνικών δικτύων που χρησιμοποιήθηκε για την παραγωγή του chatbot είναι αυτή που χρησιμοποιείται για την κατασκευή συστήματος νευρωνικής μηχανικής μετάφρασης. Στην περίπτωση της μετάφρασης, ως σύνολο δεδομένων χρησιμοποιούνται ζεύγη προτάσεων σε δύο διαφορετικές γλώσσες με την δεύτερη να αποτελεί μετάφραση της πρώτης, ούτως ώστε μετά την εκπαίδευση το σύστημα να είναι ικανό να παράγει αρκετά ικανοποιητικές μεταφράσεις. Εμείς, από την άλλη, τροφοδοτούμε το σύστημα με ζεύγη προτάσεων στην ίδια γλώσσα, με την δεύτερη να αποτελεί απάντηση στην πρώτη. Φυσικά, μια πρόταση έχει -αν όχι μία μετάφραση- πάντως περιορισμένο αριθμό μεταφράσεων, ενώ οι απαντήσεις που μπορούν να δοθούν σε μία πρόταση είναι άπειρες. Επομένως, αφού η διεξαγωγή διαλόγου δεν έχει την κανονικότητα της παραγωγής μετάφρασης, δεν έχουμε την προσδοκία το σύστημα που εκπαιδεύσαμε να ανταποκρίνεται εξίσου ικανοποιητικά.

Για να βελτιώσουμε, λοιπόν, την απόδοση του, και χάρις στην επάρκεια των δεδομένων (περίπου 9.000.000 ζεύγη σχολίων από το reddit) εργαστήκαμε για την βελτίωση της απόδοσης του αποσκοπώντας στο να μείνει "εντός θέματος" στις συζητήσεις του. Αυτό το επιτύχαμε αξιοποιώντας την πληροφορία για την θεματική υπο-ενότητα (subreddit) στην οποία ανήκει το κάθε ζεύγος σχολίων την οποία αποδώσαμε κατάλληλα στο σύστημα με μεθόδους machine learning. Συγκεκριμένα, πραγματοποιήσαμε δύο προσεγγίσεις , μια με Hierarchical Agglomerative Clustering (HAC) και μία με Latent Dirichlet Allocation (LDA), με αποτέλεσμα την ταχύτερη σύγκλιση και στις δύο περιπτώσεις.

Λέξεις - κλειδιά: ρομπότ συνομιλίας, νευρωνική μηχανική μετάφραση, θεματικά ενήμερο ρομπότ συνομιλίας, ιεραρχική συσταδοποίηση, Latent Dirichlet Allocation, reddit

Abstract

The ability of a computer to engage in dialogue as a human being is one of the greatest challenges - if not the greatest - faced by the field of Artificial Intelligence. In this work, having in mind the complexity and broadness of this issue, we attempted to create a chatbot whose function is based solely on machine learning techniques.

The original neural network structure used to generate the chatbot is the one used to construct a neural machine translation system. In the case of translation, as a set of data, pairs of sentences in two different languages are used, with the second being a translation of the first so that the trained the system is able to produce quite satisfactory translations. On the other hand, in this work, we provide the system with pairs of sentences in the same language, with the second being the answer to the first. Of course, a sentence has - but not just one translation - a limited number of translations, while the answers that can be given to a certain sentence are infinite. Therefore, since dialogue has not the regularity of translation, we have no expectation that the system we have been training will respond equally satisfactorily.

So to improve performance, and due to the sufficiency of the data (about 9,000,000 pairs of comments from reddit), we have worked to improve its performance by aiming to stay on - topic in its discussions. This was accomplished using the information about the subreddit that includes each pair of comments that we appropriately attributed to the system using machine learning techniques. In particular, we have made two approaches, one with Hierarchical Agglomerative Clustering (HAC) and one with Latent Dirichlet Allocation (LDA), resulting in faster convergence in both cases.

Keywords: chatbot, neural machine translation, topic-informed chatbot, hierarchical agglomerative clustering, Latent Dirichlet Allocation, reddit

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή Γιώργο Στάμου που μου έδωσε την δυνατότητα να είμαι μέλος του Εργαστηρίου Ευφυών Συστημάτων και να εργαστώ σε ένα τόσο ενδιαφέρον θέμα. Οφείλω ένα μεγάλο “ευχαριστώ” στον Αλέξη Μανδαλιό για την καθοδήγηση, την βοήθεια και την υπομονή του καθ’ όλη την εκπόνηση της παρούσας εργασίας. Τέλος, είμαι πολύ ευγνώμων απέναντι στην οικογένειά μου για την κατανόηση και την συμπαράστασή της στις σπουδές μου όπως και σε κάθε μου προσπάθεια.

Πίνακας περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων

Κεφάλαιο 1 : Εισαγωγή

Στο παρακάτω κεφάλαιο παρουσιάζονται κάποιες πληροφορίες σχετικά με την τεχνητή νοημοσύνη, το σκοπό της και τον τρόπο με τον οποίο αυτός σχετίζεται με την παρούσα εργασία. Ειδικότερα, γίνεται αναφορά στα ρομπότ συνομιλίας και περιγραφή των χαρακτηριστικών κατασκευής και λειτουργίας του δικού μας ρομπότ. Στη συνέχεια, παρατίθενται κάποιοι ορισμοί, μαθηματικοί ή περιγραφικοί, βασικών εννοιών για την εργασία. Ακολουθεί περιγραφή της διάρθρωσης του υπόλοιπου κειμένου.

1.1 Εισαγωγικά στοιχεία

Η ιστορία της Τεχνητής Νοημοσύνης ξεκίνησε στην αρχαιότητα, με μύθους, ιστορίες και φήμες για τεχνητά όντα που ήταν προικισμένα με νοημοσύνη ή συνείδηση από σπουδαίους τεχνίτες. (“History of artificial intelligence”). Γι' αυτό που σήμερα αποκαλούμε τεχνητή νοημοσύνη, οι κεντρικές ιδέες προτάθηκαν από κλασικούς φιλοσόφους , οι οποίοι προσπάθησαν να περιγράψουν την διαδικασία της ανθρώπινης σκέψης ως τη μηχανική διαχείριση συμβόλων. Στην πράξη, η σοβαρή συζήτηση για την οικοδόμηση ηλεκτρονικού εγκεφάλου πυροδοτήθηκε από την εφεύρεση του προγραμματιζόμενου ψηφιακού υπολογιστή, την δεκαετία του 1940.

Στόχος της τεχνητής νοημοσύνης είναι να προσδώσει σε υπολογιστικά συστήματα ιδιότητες που να προσομοιάζουν την ανθρώπινη αντίληψη και συμπεριφορά. Για τον σκοπό αυτό, επιστήμονες και μηχανικοί εργάζονται σε υποπεδία της τεχνητής νοημοσύνης, όπως η αναπαράσταση γνώσης, η όραση υπολογιστών, η ρομποτική, ο αυτόματος σχεδιασμός και προγραμματισμός, η μηχανική μάθηση καθώς και η επεξεργασία φυσικής γλώσσας. Με τους δύο τελευταίους τομείς σχετίζεται και η παρούσα εργασία, η κατασκευή δηλαδή ενός προγράμματος - ρομπότ συνομιλίας το οποίο έχει “μάθει” να διαλέγεται βασισμένο σε δεδομένα, χωρίς να έχει ρητά προγραμματιστεί.

Αναλυτικότερα, με τον όρο ρομπότ συνομιλίας (chatbot) αναφερόμαστε σε ένα πρόγραμμα ή μια τεχνητή νοημοσύνη που διεξάγει μία συνομιλία με μεθόδους ακουστικές ή γραπτού κειμένου (“Chatbot”). Τέτοιου είδους προγράμματα σχεδιάζονται συνήθως προκειμένου να προσομοιάσουν πειστικά, τον τρόπο που ένας άνθρωπος συμπεριφέρεται ως συνομιλιτής και έτσι να επιτυγχάνουν στο τεστ του Turing, το οποίο έχει ως εξής: Ένας άνθρωπος μιλάει με άλλους ανθρώπους και ένας από αυτούς αντικαθίσταται με μηχανή. Η μηχανή περνάει το τεστ, εφόσον ο άνθρωπος δεν αντιληφθεί ότι κάποιος από τους συμπαίκτες του έχει αντικατασταθεί. Λίγα προγράμματα περνούν αυτό το τεστ και τα περισσότερα τείνουν να ξεγελάσουν τους κριτές, παρά να χρησιμοποιήσουν την υπάρχουσα υπολογιστική δύναμη στο έπακρο.

Τα ρομπότ συνομιλίας διακρίνονται σε κατηγορίες με βάση ορισμένα κριτήρια, όπως ο σκοπός τους. Υπάρχουν ρομπότ συνομιλίας ειδικού σκοπού, όπως για παράδειγμα αυτά που προσφέρουν τεχνική υποστήριξη ή που προωθούν προϊόντα, αλλά και άλλα που δεν προσανατολίζουν τον διάλογο κάπου, όπως αυτά που χρησιμοποιούνται πίσω από χαρακτήρες βιντεοπαιχνιδιών, ή για την εκμάθηση γλωσσών. Επιπλέον, ο τρόπος που το ρομπότ απαντάει μπορεί να βασίζεται είτε σε κάποιους κανόνες με τους οποίους το έχουμε προγραμματίσει είτε σε σύνολο δεδομένων στα οποία το έχουμε εκπαιδεύσει. Το δικό μας ρομπότ συνομιλίας έχει εκπαιδευτεί σε δεδομένα και δεν προσανατολίζει τον διάλογο προς καμιά κατεύθυνση. Έχει, ωστόσο, τα δικά του

“ενδιαφέροντα” τα οποία έχουν καθοριστεί από τις κατηγορίες θεμάτων στις οποίες εκπαιδεύτηκε, με αποτέλεσμα να μπορεί να χειριστεί με μεγαλύτερη ευκολία σχετικές συζητήσεις.

1.2 Ορισμοί – Βασικές έννοιες

Τεχνητό νευρωνικό δίκτυο : ένας τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία:

- Το δίκτυο προσλαμβάνει τη γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης.
- Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συναπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτάται.

Στην παρούσα εργασία, όπου αναφέρουμε νευρωνικό δίκτυο πρόκειται για τεχνητό.

Μοντέλο τεχνητού νευρώνα : βασική δομική μονάδα των νευρωνικών δικτύων που αποτελείται από έναν γραμμικό συνδυαστή ακολουθούμενο από έναν απότομο περιοριστή, ο οποίος εκτελεί τη συνάρτηση προσήμου. Ο κόμβος άθροισης του νευρωνικού μοντέλου υπολογίζει ένα γραμμικό συνδυασμό των εισόδων που εφαρμόζονται στις συνάψεις του, και ενσωματώνει επίσης μια εξωτερική πόλωση. Στη συνέχεια εφαρμόζεται ο απότομος περιοριστής, ο οποίος δίνει δυαδική έξοδο. Παραλλαγή του μοντέλου αποτελεί η χρήση ομαλού περιοριστή με τη βοήθεια συναρτήσεων, όπως η υπερβολική εφαπτομένη ή η σιγμοειδής.

Βαθύ νευρωνικό δίκτυο: δομή αποτελούμενη από τεχνητούς νευρώνες, οι οποίοι εκτείνονται σε παραπάνω από ένα επίπεδα. Τα ενδιάμεσα επίπεδα νευρώνων που παρεμβάλλονται μεταξύ εισόδου και εξόδου καλούνται κρυφά και όσο περισσότερα είναι τόσο βαθύτερο είναι το νευρωνικό δίκτυο. Επιπλέον, τα βαθιά νευρωνικά δίκτυα επιτελούν συχνά πολύπλοκες μαθηματικές διεργασίες όπως η συνέλιξη ή η αναδρομή.

Εκπαίδευση (Training) νευρωνικού δικτύου : διαδικασία κατά την οποία τα βάρη των διάφορων νευρώνων ενός δικτύου μεταβάλλονται με βάση ένα σύνολο εκπαίδευσης (αποτελούμενο από ζεύγη εισόδων - εξόδων) και με στόχο να δίνουν για τα διάφορα διανύσματα εισόδου, τα αντίστοιχα διανύσματα εξόδου – στόχου. Κατά τη διάρκειά της, το δίκτυο τροφοδοτείται με το σύνολο δεδομένων εκπαίδευσης για όσες φορές έχουμε ορίσει (εποχές).

Έλεγχος (Testing) νευρωνικού δικτύου: διαδικασία κατά την οποία υπολογίζεται το ποσοστό σωστών απαντήσεων που δίνει το νευρωνικό δίκτυο για ένα σύνολο ελέγχου (αποτελούμενο από ζεύγη εισόδων – εξόδων) τα οποίο δεν έχει συμμετέχει στην εκπαίδευσή του.

Αλγόριθμος οπισθοδιάδοσης βαρών (Back Propagation algorithm): αλγόριθμος που συντελεί στην ενίσχυση ή εξασθένηση των κατάλληλων νευρώνων, για την εκπαίδευση ενός νευρωνικού δικτύου. Λειτουργεί με γνώμονα την ελαχιστοποίηση κάποιας συνάρτησης σφάλματος μεταξύ των

πραγματικών εξόδων και των εξόδων – στόχων με τη μέθοδο καθόδου παραγώγου. Περιλαμβάνει τις εξής φάσεις:

- Φάση διάδοσης:
    - Τροφοδότηση των δεδομένων εισόδου και παραγωγή των εξόδων νευρωνικού δικτύου.
    - Υπολογισμός σφάλματος μεταξύ εξόδων που έχουν παραχθεί και εξόδων – στόχων.
    - Διάδοση των ενεργοποιήσεων εξόδου προς τα πίσω στο δίκτυο, με τη χρήση του μοτίβου εκπαίδευσης, προκειμένου να παραχθούν τα "δέλτα" (η διαφορά μεταξύ των πραγματικών εξόδων και των εξόδων στόχων) για όλους τους νευρώνες των εξόδων και των κρυφών επιπέδων.
- Φάση ανανέωσης βαρών:
    - Το δέλτα εξόδου και η συνάρτηση εισόδου του κάθε βάρους πολλαπλασιάζονται ώστε να βρεθεί η παράγωγος του βάρους.
    - Κάθοδος παραγώγου: Ένα ποσοστό της παραγώγου του κάθε βάρους αφαιρείται από το βάρος αυτό, και προκύπτουν έτσι τα νέα βάρη του δικτύου. Ο αριθμός που καθορίζει το ποσοστό αυτό λέγεται ρυθμός μάθησης και μπορεί να είναι είτε σταθερός είτε να μεταβάλλεται ανάλογα το στάδιο της εκπαίδευσης.

Τα βάρη μπορούν να ανανεώνονται είτε σε κάθε διάνυσμα εισόδου (στοχαστικά / απευθείας) είτε αφού τροφοδοτήσουμε το δίκτυο με περισσότερα διανύσματα εισόδου και υπολογίσουμε το μέσο όρο των "δέλτα" (μαζικά).

Βελτιστοποιητής adam (ADAptive Moment estimation): επαναληπτικός αλγόριθμος που χρησιμοποιείται για την ανανέωση βαρών (αντί της καθόδου παραγώγου) και βασίζεται στη χρήση του μέσου όρου των προηγούμενων παραγώγων καθώς και του μέσου όρου των τετραγώνων τους.

Αλγόριθμος μεγιστοποίησης προσδοκίας (Expectation Maximization EM): επαναληπτική μέθοδος για την εύρεση μέγιστης πιθανότητας ή μέγιστων εκ των υστέρων εκτιμήσεων παραμέτρων σε στατιστικά μοντέλα, όπου το μοντέλο εξαρτάται από μη παρατηρούμενες λανθάνουσες μεταβλητές. Σε κάθε επανάληψη του αλγορίθμου, εναλλάσσονται οι εκτελέσεις ενός βήματος προσδοκίας (expectation : E), το οποίο δημιουργεί μία συνάρτηση για την αναμενόμενη πιθανότητα -που αξιολογείται χρησιμοποιώντας την τρέχουσα εκτίμηση για τις παραμέτρους- και ένα βήμα μεγιστοποίησης (maximization : M), το οποίο υπολογίζει τις παραμέτρους που μεγιστοποιούν την αναμενόμενη λογαριθμική πιθανότητα που βρέθηκε στο βήμα E. Αυτές οι εκτιμήσεις παραμέτρων στη συνέχεια χρησιμοποιούνται για τον προσδιορισμό της κατανομής των λανθανουσών μεταβλητών στο επόμενο στάδιο E.

Overfitting: Φαινόμενο στην εκπαίδευση νευρωνικού δικτύου κατά το οποίο το σφάλμα μεταξύ αναμενόμενων και πραγματικών αποτελεσμάτων μειώνεται στο σύνολο εκπαίδευσης αλλά αυξάνεται στο σύνολο ελέγχου. Σ' αυτή την περίπτωση το δίκτυο πρακτικά απομνημονεύει τα δεδομένα, αντί να μαθαίνει από αυτά, πράγμα που είναι ανεπιθύμητο.

Dropout: Μέθοδος αντιμετώπισης του overfitting, κατά την οποία σε κάθε ανανέωση βαρών ένα ποσοστό των νευρώνων διατηρεί αμετάβλητα τα βάρη του.

Term frequency – Inverse document frequency (tf-idf): στατιστική μετρική που δείχνει πόσο σημαντικός είναι ένας όρος για ένα έγγραφο ενός συνόλου κειμένων. Αποτελεί το γινόμενο της σχετικής συχνότητας όρου (term frequency: tf) με την αντίστροφη συχνότητα εγγράφου (idf). Ειδικότερα , $tfidf(t,d,n) = tf(t) \cdot idf(d,n) = tf(t) \cdot [log(\frac{n}{df(d,t)}) + 1]$ όπου n το συνολικό πλήθος των εγγράφων, tf(t) η σχετική συχνότητα του όρου t σε ένα έγγραφο και df(d,t) το πλήθος των εγγράφων d που περιέχουν τον όρο t.

Clustering (συσταδοποίηση) : μέθοδος μη επιβλεπόμενης μηχανικής μάθησης κατά την οποία τα στοιχεία ενός συνόλου δεδομένων ομαδοποιούνται με βάση τα χαρακτηριστικά τους. Οι ομάδες παράγονται κατά τέτοιο τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (cluster) να είναι περισσότερο όμοια μεταξύ τους σε σχέση με εκείνα των άλλων ομάδων.

Εκ των προτέρων κατανομή πιθανότητας : κατανομή πιθανότητας, στην στατιστική Bayes, που εκφράζει τις πεποιθήσεις κάποιου για μία ποσότητα προτού ληφθούν υπ' όψιν αποδείξεις σχετικά με αυτήν.

Συζυγής εκ των προτέρων κατανομή πιθανότητας : η εκ των προτέρων κατανομή πιθανότητας όταν ανήκει στην ίδια οικογένεια κατανομών πιθανότητας με την ύστερη κατανομή.

Κατανομή Dirichlet: πιθανοτική κατανομή, η οποία ορίζεται ως το σύνολο των υπερπαραμέτρων $\boldsymbol{\alpha} = (a_1,\ldots,a_k)$ έτσι ώστε $\boldsymbol{\theta} \sim Dirichlet(\alpha_1,\ldots,\alpha_K)$αν $P(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} {\theta_i}^{a_i - 1}$ όπου η συνάρτηση βήτα: $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{K} a_i)}$.

Polytope: η γενίκευση ενός γεωμετρικού πολυέδρου από τις τρεις διαστάσεις σε περισσότερες.

Simplex: τα σημεία του κυρτού χώρου του polytope. Στη θεωρία πιθανοτήτων, τα σημεία του πρότυπου n-simplex στο R_{n+1} είναι ο χώρος των πιθανών παραμέτρων (πιθανοτήτων) της κατηγορικής κατανομής σε n + 1 πιθανά αποτελέσματα.

Συνάρτηση πιθανοφάνειας: συνάρτηση των παραμέτρων ενός στατιστικού μοντέλου. Δεδομένης μιας παραμετρικής οικογένειας συναρτήσεων πυκνότητας πιθανότητας $x \rightarrow f(x|\theta)$, όπου η θ είναι μια παράμετρος, η συνάρτηση πιθανοφάνειας είναι $\theta \rightarrow f(\theta|x)$, όπου x το παρατηρούμενο αποτέλεσμα ενός πειράματος.

1.3 Διάρθρωση του κειμένου

Το Κεφάλαιο 2 περιλαμβάνει το θεωρητικό υπόβαθρο που χρειάζεται για την κατανόηση ενός συστήματος νευρωνικής μηχανικής μετάφρασης καθώς και του ζητήματος της συσταδοποίησης δεδομένων και της θεματικής μοντελοποίησης κειμένου. Στο Κεφάλαιο 3 περιγράφονται οι προηγούμενες εργασίες

ερευνητών οι οποίες φάνηκαν χρήσιμες για την παρούσα εργασία. Στο Κεφάλαιο 4 αναλύεται η εξαγωγή θεμάτων από τα δεδομένα και οι τροποποιήσεις που πραγματοποιήσαμε στο σύστημα, οι προγραμματιστικές λεπτομέρειες των οποίων καταγράφεται στο Κεφάλαιο 5. Το Κεφάλαιο 6 περιλαμβάνει τα πειραματικά αποτελέσματα και την σύγκριση των τροποποιημένων συστημάτων με το αρχικό. Τέλος, στο Κεφάλαιο 7 παρουσιάζονται τα συμπεράσματα που προέκυψαν από την εργασία καθώς και προτάσεις για μελλοντικές επεκτάσεις της.

Κεφάλαιο 2 : Θεωρητικό υπόβαθρο

Στο παρακάτω κεφάλαιο καλύπτονται οι βασικές γνώσεις που αφορούν στην νευρωνική μηχανική μετάφραση -στη λογική της οποίας βασίστηκε και η κατασκευή του ρομπότ συνομιλίας- καθώς και τις μεθόδους του Hierarchical Agglomerative Clustering και της Latent Dirichlet Allocation τις οποίες χρησιμοποιήσαμε για να προσδώσουμε θεματική πληροφόρηση στο ρομπότ μας. Συγκεκριμένα, ξεκινάμε από τη γλωσσική μοντελοποίηση με τη βοήθεια των πιθανοτήτων, την περιγραφή του στατιστικού γλωσσικού μοντέλου, την ανάγκη για βελτίωσή του και την εξέλιξή του σε νευρωνικό γλωσσικό μοντέλο. Στη συνέχεια, ασχολούμαστε με το επαναληπτικό νευρωνικό δίκτυο, το οποίο είναι ικανό να επεξεργάζεται δεδομένα χρονικών ακολουθιών -όπως οι προτάσεις- και αποτελεί το κύριο συστατικό στοιχείο του συστήματος νευρωνικής μηχανικής μετάφρασης. Περιγράφουμε την αρχιτεκτονική του, τη διαδικασία εκπαίδευσής του, αναφερόμαστε στα προβλήματα που προκύπτουν κατά την πορεία της τελευταίας και τους τρόπους αντιμετώπισής τους. Στους τρόπους αυτούς εντάσσεται και η προσθήκη μνήμης μακρού – βραχέος όρου, για την οποία κάνουμε εκτενή περιγραφή. Έχοντας, λοιπόν, καλύψει τις βασικές προαπαιτούμενες έννοιες για τη νευρωνική μηχανική μετάφραση, περνάμε στην περιγραφή του συστήματός της, τις κύριες αρχές της, τον μηχανισμό προσοχής και τα είδη των λεξιλογίων της. Ακολουθεί περιγραφή της υλοποίησής της καθώς και του αυτόματου τρόπου αξιολόγησής της. Αναλύεται, επίσης, η λογική της συσταδοποίησης και οι διάφορες τεχνικές της. Τέλος, σχετικά με την Latent Dirichlet Allocation, αναλύουμε αφ' ενός τις ιδέες που τη διέπουν και τις μαθηματικές σχέσεις που την μοντελοποιούν και αφ' ετέρου την τεκμηρίωση μεταβλητών Bayes για την εφαρμογή της.

2.1 Γλωσσικό Μοντέλο

Η γλωσσική μοντελοποίηση παίζει καθοριστικό ρόλο στην μηχανική παραγωγή λόγου. Συγκεκριμένα, αποδίδει μια κατανομή πιθανότητας σε ακολουθίες συμβόλων (συνήθως λέξεις) ώστε να μπορεί κανείς να αποφανθεί αν μια ακολουθία λέξεων είναι πιο πιθανή ή πιο εύγλωττη από μια άλλη. Για να αποδοθεί αυτή η πιθανότητα χρησιμοποιούνται είτε πίνακες που καταγράφουν την συχνότητα συνδυασμών λέξεων (στατιστικό μοντέλο) είτε νευρωνικά δίκτυα (νευρωνικό πιθανοτικό μοντέλο).

2.1.1 Στατιστικό γλωσσικό μοντέλο

Το στατιστικό γλωσσικό μοντέλο αξιολογεί τις προτάσεις ως πιθανές με βάση την πιθανότητα ύπαρξης της κάθε λέξης, δεδομένων των προηγούμενών της, κάτι που προκύπτει από πίνακες σχετικών συχνοτήτων των “επόμενων λέξεων” για τους διάφορους συνδυασμούς προηγούμενων λέξεων. Πιο αναλυτικά, η πιθανότητα ύπαρξης μιας πρότασης με Τ λέξεις είναι το γινόμενο των δεσμευμένων πιθανοτήτων: πιθανότητα της 1ης * πιθανότητα της 2ης με δεδομένη την πρώτη λέξη * πιθανότητα της 3ης με δεδομένες την 1η και την 2η κ.ο.κ. , δηλ:

$$P(w_1^T) = \prod_{t=1}^{T} P(w_t | w_1^{t-1})$$

Για να μειωθεί η πολυπλοκότητα και η δυσκολία της μοντελοποίησης, αντί να ληφθούν υπ'

όψιν όλες οι λέξεις από την αρχή της πρότασης, επιλέγονται μόνο κάποιες προηγούμενες και έτσι έχουμε τα μοντέλα n-γραμμάτων τα οποία λειτουργούν με πίνακες με δεσμευμένες πιθανότητες για την επόμενη λέξη, με πολλούς συνδυασμούν (n-1) προηγούμενων λέξεων.

$$P(w_t|w_1^{t-1}) \approx P(w_t|w_{t-n+1}^{t-1})$$

Μ' αυτόν τον τρόπο αξιοποιούμε τους συνδυασμούς διαδοχικών λέξεων που εμφανίζονται (συχνά ή όχι) στο σώμα εκπαίδευσης. Τα πιο διαδεδομένα n-γράμματα λόγω του ότι εξισορροπούν την απλότητα και την διατήρηση πληροφορίας είναι τα 3-γράμματα.

2.1.2 Ελλείψεις στατιστικού μοντέλου

Υπάρχει ωστόσο πολλή περισσότερη πληροφορία στην ακολουθία λέξεων πριν από μια λέξη, σε σχέση με το ποιες είναι οι δύο τελευταίες λέξεις πριν από αυτήν. Για την αξιοποίηση της πληροφορίας αυτής, άρχισε η χρήση νευρωνικών γλωσσικών μοντέλων με στόχο το να λαμβάνονται υπ'όψιν περισσότερες από τις δύο τελευταίες λέξεις καθώς και την αξιοποίηση της ομοιότητας μεταξύ των λέξεων, όσον αφορά τον γραμματικό και συντακτικό τους ρόλο.

Το ζήτημα που προκύπτει με τη χρήση n-γραμμάτων με μεγάλο n είναι το λεγόμενο πρόβλημα,ή κατά τους (Bengio κ.α., 2003) "κατάρα της διαστατικότητας": Η από κοινού κατανομή πιθανότητας μεταξύ πολλών τυχαίων διακριτών μεταβλητών (λέξεων, από κοινού στην πρόταση) απαιτεί τον προσδιορισμό πολλών παραμέτρων, το πλήθος των οποίων αυξάνει εκθετικά με το πλήθος των μεταβλητών. Για παράδειγμα, αν θέλουμε να προβλέψουμε την εμφάνιση μιας πρότασης αποτελούμενης από 10 λέξεις όταν το λεξιλόγιό μας έχει 100.000 λέξεις, τότε οι ελεύθερες παράμετροι που πρέπει να προσδιορίσουμε είναι 10^{50}-1. Προκειμένου να αντιμετωπιστεί το ζήτημα με την διαστατικότητα, εισήχθη η ιδεά της εκμάθησης κατανεμημένων αναπαραστάσεων για λέξεις, η οποία συνδυάζεται με την ταυτόχρονη έκφραση της συνάρτησης πιθανότητας με χρήση πολυστρωματικού νευρωνικού δικτύου, όπως παρουσιάζεται στο νευρωνικό πιθανοτικό μοντέλο των (Bengio κ.α., 2003).

2.1.3 Νευρωνικό πιθανοτικό γλωσσικό μοντέλο

Κύριο χαρακτηριστικό των νευρωνικών γλωσσικών μοντέλων είναι η αναπαράσταση των λέξεών τους ως "λεκτικά διανύσματα" υψηλών διαστάσεων πραγματικών τιμών και η έκφραση συνάρτησης πιθανότητας για τις ακολουθίες λέξεων αναφορικά με τις αναπαραστάσεις αυτές. Το μοντέλο μαθαίνει ταυτόχρονα τόσο τις αναπαραστάσεις των λέξεων όσο και τη συνάρτηση πιθανότητας των αποτελούμενων από αυτές ακολουθιών.

Το λεκτικό διάνυσμα αναπαριστά διαφορετικές ιδιότητες της λέξης: κάθε λέξη αντιστοιχίζεται σε ένα σημείο ενός διανυσματικού χώρου, με το πλήθος των συντεταγμένων του να είναι πολύ μικρότερο από το μέγεθος του λεξιλογίου (100 με 1000 φορές). Θα μπορούσε είτε να αρχικοποιηθεί με βάση την εκ των προτέρων γνώση σημασιολογικών χαρακτηριστικών, είτε όχι. Η συνάρτηση πιθανότητας εκφράζεται ως γινόμενο των δεσμευμένων πιθανοτήτων της επόμενης λέξης, με δεδομένες (κάποιες) προηγούμενες. Οι παράμετροί της ενημερώνονται επαναληπτικά προκειμένου να μεγιστοποιήσουν την λογαριθμική πιθανότητα των δεδομένων εκπαίδευσης ή με βάση κάποιο κριτήριο κανονικοποίησης (πχ. ποινή εξασθένισης βαρών). Η γενίκευση σ' αυτό το μοντέλο προκύπτει αβίαστα διότι "όμοιες" λέξεις αναμένεται να έχουν όμοια λεκτικά διανύσματα και επίσης, η συνάρτηση πιθανότητας είναι μία ομαλή (smooth) συνάρτηση αυτών των παραμέτρων, οπότε μια μικρή αλλαγή στα χαρακτηριστικά θα δώσει μια μικρή αλλαγή στην πιθανότητα.

Το σύνολο δεδομένων είναι μια ακολουθία $w_1 ... w_T$ λέξεων με w_t να ανήκουν σε σύνολο V που αποτελεί το λεξιλόγιο : ένα μεγάλο αλλά πεπερασμένο σύνολο λέξεων. Το αντικείμενο του νευρωνικού μοντέλου είναι η εκμάθηση ενός καλού μοντέλου $f(w_t, \dots, w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1})$ με

την έννοια ότι δίνει καλή εκτίμηση πιθανότητας για προτάσεις εκτός του συνόλου εκπαίδευσης. Η αξιολόγηση των διαφόρων κατανομών γίνεται με το μέγεθος της περιπλοκής, δηλαδή τον γεωμετρικό μέσο των αντίστροφων κατανομών πιθανότητας $\frac{1}{\hat{P}(w_t|w_1^{t-1})}$ ο οποίος ισούται επίσης με την εκθετική συνάρτηση της μέσης αρνητικής λογαριθμικής συνάρτησης πιθανοφάνειας. Ο μόνος περιορισμός στο μοντέλο είναι ότι για κάθε επιλογή ακολουθίας, οι προβλεπόμενες πιθανότητες για όλες τις λέξεις του λεξιλογίου οφείλουν να έχουν άθροισμα ίσο με 1, με τις πιθανότητες θετικές. Από το γινόμενο αυτών των υπό συνθήκη πιθανοτήτων παίρνουμε ένα μοντέλο για την από κοινού πιθανότητα των ακολουθιών των λέξεων.

Η συνάρτηση $f(w_t, \ldots, w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1})$ μπορεί να διασπαστεί σε δύο τμήματα. Το πρώτο είναι μια απεικόνιση C κάθε στοιχείου i του V σε ένα πραγματικό διάνυσμα $C(i) \in R^m$και αναπαριστά τα κατανεμημένα διανύσματα χαρακτηριστικών που σχετίζονται με κάθε λέξη του λεξιλογίου και καλούνται διανύσματα ενσωμάτωσης (embedding vectors). Στην πράξη ο C αποτελεί έναν πίνακα $|V| \times m$. Το δεύτερο τμήμα είναι το πώς κατανέμεται η πυκνότητα πιθανότητας στις λέξεις οι οποίες εκφράζονται με την C: μια συνάρτηση g αντιστοιχίζει μια ακολουθία εισόδου αποτελούμενη από διανύσματα ενσωμάτωσης για τις n τελευταίες λέξεις σε μια υπό συνθήκη κατανομή πιθανότητας στις λέξεις του V για την επόμενη λέξη w_t. Η έξοδος της g είναι ένα διάνυσμα του οποίου το i-οστό στοιχείο εκτιμά την πιθανότητα $\hat{P}(w_t = i|w_1^{t-1})$.

Με τα δύο αυτά τμήματα σχετίζονται ορισμένες παράμετροι. Οι παράμετροι της απεικόνισης C είναι απλά τα ίδια τα διανύσματα ενσωμάτωσης ενώ η συνάρτηση g μπορεί να υλοποιηθεί με ένα νευρωνικό δίκτυο είτε πρόσθιας τροφοδότησης, είτε επαναληπτικό ή με μια άλλη παραμετροποιημένη συνάρτηση έχοντας παραμέτρους $\boldsymbol{\omega}$. Έτσι το σύνολο παραμέτρων είναι το $\boldsymbol{\theta} = (\boldsymbol{C}, \boldsymbol{\omega})$ και η εκπαίδευση επιτυγχάνεται με την εύρεση $\boldsymbol{\theta}$ το οποίο μεγιστοποιεί την κανονικοποιημένη λογαριθμική πιθανότητα $L = \frac{1}{T}\sum_t logf(w_t, w_{t-1}, \ldots, w_{t-n+1}; \boldsymbol{\theta}) + R(\boldsymbol{\theta})$ όπου η R($\boldsymbol{\theta}$) είναι ένας όρος ποινής (κανονικοποίησης). Οι (Bengio κ.α., 2003) χρησιμοποίησαν για την g δίκτυο πρόσθιας τροφοδότησης και η συνάρτηση ποινής τους ήταν μια συνάρτηση εξασθένισης βαρών που εφαρμόστηκε στα βάρη του νευρωνικού και στον πίνακα C.

Η έλλειψη της προσέγγισης αυτής ήταν ότι ένα δίκτυο πρόσθιας τροφοδότησης πρέπει να χρησιμοποιεί διάνυσμα "συμφραζομένων" (πρακτικά προηγούμενων λέξεων) καθορισμένου μεγέθους το οποίο οφείλει να προσδιορίζεται επί τούτω. Αυτό, συνήθως, σημαίνει ότι το νευρωνικό δίκτυο βλέπει πέντε με δέκα προηγούμενες λέξεις για να προβλέψει την επόμενη. Οι άνθρωποι, ωστόσο, μπορούμε να εκμεταλλευόμαστε επιτυχώς περισσότερο γλωσσικό περιεχόμενο, και να συγκρατούμε διαφορετικό μέγεθος πληροφορίας κατά περίπτωση. Η ικανότητα αυτή επιχειρήθηκε να δοθεί στο νευρωνικό γλωσσικό μοντέλο με τη χρήση επαναληπτικού νευρωνικού δικτύου, από τους (Mikolov κ.α. 2010).

2.2 Επαναληπτικό νευρωνικό δίκτυο

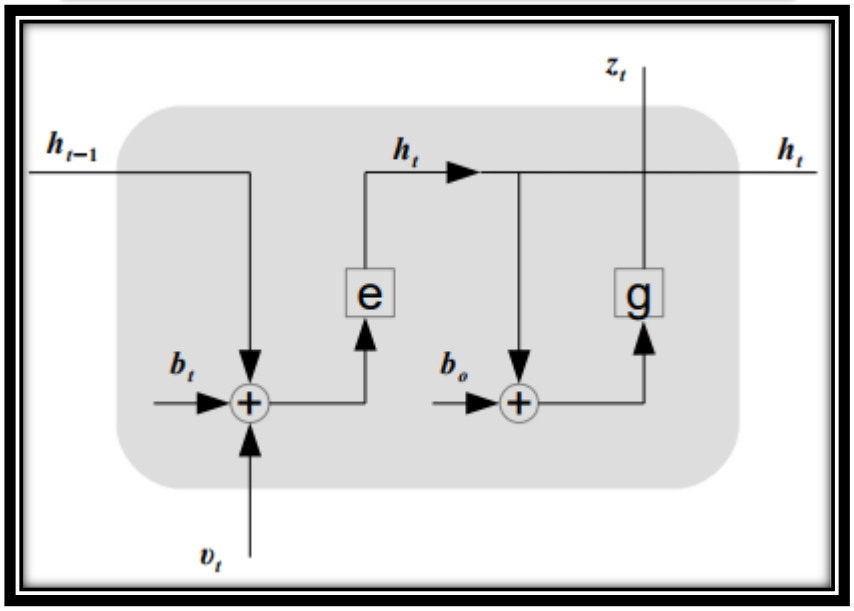
Το επαναληπτικό νευρωνικό δίκτυο (Recurrent Neural Network : RNN) αποτελεί μία αρχιτεκτονική μεγάλων δυνατοτήτων που μπορεί να χειριστεί δεδομένα ακολουθιών και έχει εφαρμοστεί σε ζητήματα γλωσσικής μοντελοποίησης με μεγάλη επιτυχία. Η διεπαφή του επαναληπτικού δικτύου είναι απλή και συνήθης στα νευρωνικά, δέχεται ένα διάνυσμα σταθερού μεγέθους ως είσοδο και παράγει διάνυσμα σταθερού μεγέθους στην έξοδο (για παράδειγμα, κατανομή πιθανότητας). Το ξεχωριστό με τα επαναληπτικά δίκτυα είναι η ικανότητα χειρισμού χρονικών ακολουθιών τέτοιων διανυσμάτων, όπως παρουσιάζεται αναλυτικά στη συνέχεια.

2.2.1 Αρχιτεκτονική και εκπαίδευση του επαναληπτικό νευρωνικού δικτύου

Το απλό επαναληπτικό νευρωνικό δίκτυο (Elman, 1990) αποτελεί ένα μη γραμμικό δυναμικό σύστημα το οποίο παραμετροποιείται από τρεις πίνακες βαρών και τρία σταθερά διανύσματα

$[W_{hv}, W_{hh}, W_{oh}, b_h, b_o, h_0]$ των οποίων η σύμπτυξη **θ** περιγράφει πλήρως το επαναληπτικό νευρωνικό δίκτυο.

Εικόνα 1: Μονάδα επαναληπτικού νευρωνικού δικτύου



Δοσμένης της ακολουθίας εισόδου $(v_1, \ldots, v_T)$την οποία συμβολίζουμε ως v_1^T, το επαναληπτικό δίκτυο υπολογίζει μια ακολουθία κρυφών καταστάσεων h_1^T και μια ακολουθία εξόδων z_1^T , με τον εξής αλγόριθμο:

Αλγόριθμος 1: Εμπρόσθια τροφοδότηση επαναληπτικού δικτύου

1: ***for*** t ***from*** 1 ***to*** T***do***
2: $v_t \leftarrow W_{hv} v_t + W_{hh} h_{t-1} + b_h$
3: $h_t \leftarrow e(v_t)$
4: $o_t \leftarrow W_{oh} h_t + b_o$
5: $z_t \leftarrow g(o_t)$
6: ***end for***

Όπου οι e(.) και οι g(.) είναι οι μη γραμμικότητες κρυφής κατάστασης και εξόδου του επαναληπτικού νευρωνικού δικτύου και το h_0 είναι ένα διάνυσμα παραμέτρων που αποθηκεύουν την αρχική κρυφή κατάσταση. Η συνάρτηση κόστους του επαναληπτικού νευρωνικού δικτύου είναι συνήθως το άθροισμα των συναρτήσεων κόστους ανά χρονική στιγμή:

$$L(z, y) = \sum_{t=1}^{T} L(z_t; y_t)$$

όπου y η επιθυμητή έξοδος. Οι παράγωγοι των επαναληπτικών νευρωνικών δικτύων υπολογίζονται εύκολα μέσω του αλγορίθμου οπισθοδιάδοσης στο χρόνο:

Αλγόριθμος 2: Οπισθοδιάδοση στο χρόνο βαρών επαναληπτικού δικτύου

1: $\boldsymbol{for}\ t\ \boldsymbol{from}\ T\ \boldsymbol{downto}\ 1\ \boldsymbol{do}$
2: $do_t \leftarrow g'(o_t).dL\dfrac{(z_t; y_t)}{dz_t}$
3: $db_o \leftarrow db_o + do_t$
4: $dW_{oh} \leftarrow dW_{oh} + do_t h_t^T$
5: $dh_t \leftarrow dh_t + W_{oh}^T do_t$
6: $dz_t \leftarrow e'(z_t).dh_t$
7: $dW_{hv} \leftarrow dW_{hv} + dz_t v_t^T$
8: $db_h \leftarrow db_h + dz_t$
9: $dW_{hh} \leftarrow dW_{hh} + dz_t h_{t-1}^T$
10: $dh_{t-1} \leftarrow W_{hh}^T dz_t$
11: $\boldsymbol{end\ for}$
12: $\boldsymbol{Return}\ d\theta = [dW_{hv}, dW_{hh}, dW_{oh}, db_h, db_o, dh_0]$

2.2.2 Δυσκολίες κατά την εκπαίδευση του επαναληπτικού νευρωνικού δικτύου

Η δυσκολία στην εκπαίδευση του επαναληπτικού νευρωνικού δικτύου έγκειται στην επαναληπτική μη γραμμική φύση του. Τα δύο συχνότερα προβλήματα που προκύπτουν όταν πραγματευόμαστε πολύ μεγάλες ακολουθίες είναι αυτά της διόγκωσης και της συρρίκνωσης των παραγώγων όπως περιγράφονται από τους (Bengio κ.α., 1994). Εν συντομία, η διόγκωση των παραγώγων αναφέρεται στο φαινόμενο όταν οι παράγωγοι αυξάνονται εκθετικά καθώς η οπισθοδιάδοση προχωρά στο χρόνο, μετατρέποντάς την σε "κολοβή" οπισθοδιάδοση στο χρόνο, η οποία είναι ανίκανη να συγκρατήσει εξαρτήσεις μεγάλης έκτασης στις ακολουθίες.

Η βασική αιτία των προβλημάτων αυτών έγκειται στην γραμμή 10 του αλγόριθμου οπισθοδιάδοσης βαρών στο χρόνο (Αλγόριθμος 2). Αν αγνοήσουμε τις ενδιάμεσες απώλειες, ένα σήμα που έχει οπισθοδιαδοθεί από την τρέχουσα κατάσταση, για K βήματα, θα γίνει $dh_{t-k} = \prod_{i=1}^{K}(W_{hh}^T . dz_t)$. Βέβαια, η z_t είναι μια συνάρτηση (πχ. tanh ή σιγμοειδής) η οποία είναι ομαλή, άρα η συμπεριφορά του δικτύου καθορίζεται από τα χαρακτηριστικά του επαναληπτικού πίνακα W_{hh} και οι περισσότερες αναλύσεις τον εξετάζουν αναφορικά με την μεγαλύτερη ιδιοτιμή του. Αν αυτή είναι πολύ μεγάλη ή υπερβολικά μικρή, υπάρχει ο κίνδυνος διόγκωσης ή συρρίκνωσης παραγώγων αντίστοιχα. (Bengio κ.α. 1994, Hochreiter and Schmidhuber, 1997, Martens and Sutskever, 2011, Pascanu κ.α. 2013).

Διόγκωση παραγώγων

Γενικά, είναι αρκετά εύκολο να χειριστούμε το πρόβλημα διόγκωσης παραγώγων με την εφαρμογή διάφορων μορφών ψαλιδίσματος παραγώγων. Η πρώτη προσέγγιση προτάθηκε από τον (Mikolov 2012) με τη μορφή του "ανά στοιχείο" ψαλιδίσματος: κάθε χρονική στιγμή, κατά τη διάρκεια της οπισθοδιάδοσης, κάθε συντεταγμένη του dh που είναι μεγαλύτερη από ένα κατώφλι τ, όπου τ > 0, ή μικρότερη από -τ, τίθεται ίση με τ ή -τ αντίστοιχα. Μια άλλη προσέγγιση είναι το ψαλίδισμα νόρμας παραγώγου που προτάθηκε από τους (Pascanu κ.α. 2013). Σ' αυτή την περίπτωση, η νόρμα του διανύσματος της παραγώγου g που έχει υπολογιστεί "ανά υποσύνολο των δεδομένων" συγκρίνεται με ένα θετικό κατώφλι τ. Αν το ξεπερνά, τότε το διάνυσμα της παραγώγου αντικαθίσταται από την ποσότητα $\tau\frac{g}{||g||}$. Σε πολλές περιπτώσεις οι δύο προσεγγίσεις συνδυάζονται.

Συρρίκνωση παραγώγων

Για το ζήτημα της συρρίκνωσης παραγώγων έχουν, επίσης, προταθεί αρκετοί τρόποι αντιμετώπισης, οι περισσότεροι από τους οποίους προσπαθούν να μετριάσουν τη μη γραμμικότητά των δικτύων καθιστώντας ωστόσο την εκπαίδευση λιγότερο βαθιά. Η καλύτερη προσέγγιση φαίνεται να είναι η προσθήκη μνήμης μακρού και βραχέος όρου (Long Short – Term Memory LSTM) που

εφευρέθηκε από τους (Hochreiter και Schmidhuber, 1997) και βελτιώθηκε από τους (Gers κ.α., 2000), με την οποία θα ασχοληθούμε αναλυτικά στην ακόλουθη υποενότητα.

2.2.3 Δίκτυο Long - Short Term Memory (LSTM)

Το δίκτυο Long - Short Term Memory (LSTM) είναι μια αρχιτεκτονική επαναληπτικού νευρωνικού δικτύου, η οποία χειρίζεται κομψά τις συρρικνούμενες παραγώγους χρησιμοποιώντας μονάδες μνήμης. Αυτές οι γραμμικές μονάδες έχουν μια μονής κατεύθυνσης σύνδεση και ένα ζεύγος βοηθητικών "μονάδων – πυλών" που ελέγχουν τη ροή της πληροφορίας από και προς τη μονάδα μνήμης. Όταν οι μονάδες – πύλες είναι κλειστές, οι παράγωγοι μπορούν να ρέουν δια μέσω της μονάδας μνήμης χωρίς μεταβολή για αόριστο χρονικό διάστημα, ξεπερνώντας, έτσι, το πρόβλημα συρρίκνωσης παραγώγων. Αν και οι πύλες ποτέ δεν απομονώνουν τη μονάδα μνήμης, στην πράξη, αυτή η εξήγηση δείχνει ότι το LSTM χειρίζεται το πρόβλημα συρρίκνωσης παραγώγων σε μερικές τουλάχιστον καταστάσεις, και πράγματι εύκολα λύνει ένα πλήθος σύνθετων προβλημάτων με παθολογικές χρονικές εξαρτήσεις μεγάλης έκτασης, οι οποίες στο παρελθόν θεωρούνταν άλυτες από τα απλά επαναληπτικά νευρωνικά δίκτυα.

Στον παρακάτω πίνακα φαίνονται τα διανύσματα που συμμετέχουν στη συγκρότηση του LSTM, το οποίο έστω ότι αποτελείται από N μονάδες μνήμης:

Πίνακας 1: Μεταβλητές που συμμετέχουν στη μνήμη μακρού - βραχέος όρου

Όνομα μεταβλητής	Περιγραφή
i_t^g	Διάνυσμα πυλών εισόδου. Παίρνει τιμές στο $[0,1]^N$
i_t	Διάνυσμα εισόδων στις μονάδες μνήμης. Παίρνει τιμές στο $[-1,1]^N$
o_t	Διάνυσμα πυλών εξόδου. Παίρνει τιμές στο $[0,1]^N$
f_t	Διάνυσμα πυλών λήθης. Παίρνει τιμές στο $[0,1]^N$
v_t	Διάνυσμα εισόδου. Παίρνει τιμές στο R^v
h_t	Κλασικό διάνυσμα κρυφής κατάστασης. Παίρνει τιμές στο $[-1,1]^h$
m_t	Διάνυσμα κατάστασης μονάδας μνήμης. Παίρνει τιμές στο R^N
$\tilde{m}_t$	Διάνυσμα κατάστασης μονάδας μνήμης διαθέσιμο στην υπόλοιπη μνήμη μακρού – βραχέος όρου. Παίρνει τιμές στο R^N
z_t	Διάνυσμα εξόδου

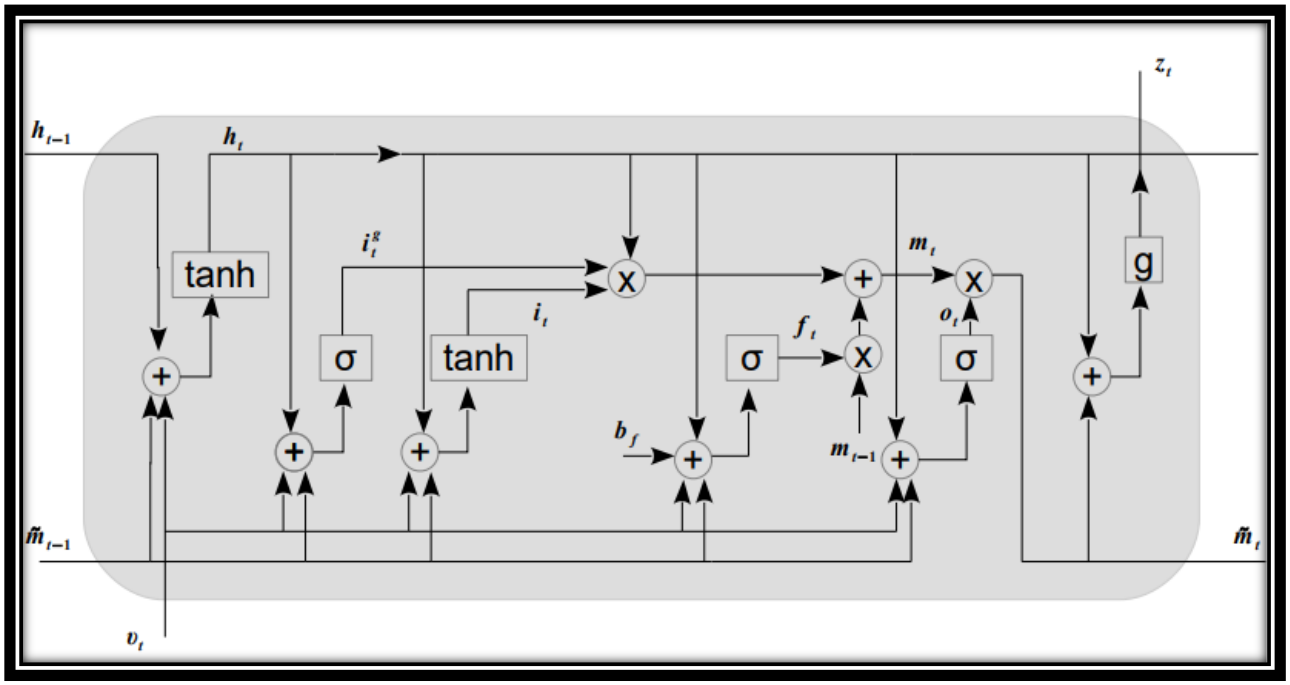
Η εξέλιξη των διανυσμάτων αυτών ορίζεται από τις παρακάτω εξισώσεις:

$$h_t = tanh(W_{hh}h_{t-1} + W_{hv}v_t + W_{hm}\tilde{m}_{t-1})$$
$$i_t^g = sigmoid(W_{igh}h_t + W_{igv}v_t + W_{igm} + \tilde{m}_{t-1})$$

$i_t = tanh(W_{ih}h_t + W_{iv}v_t + W_{im}\tilde{m}_{t-1})$
$o_t = sigmoid(W_{oh}h_t + W_{ov}v_t + W_{om}\tilde{m}_{t-1})$
$f_t = sigmoid(b_f + W_{fh}h_t + W_{f_v}v_t + W_{fm}\tilde{m}_{t-1})$
$m_t = m_{t-1} \circ f_t + i_t \circ i_t^g$ η πύλη εισόδου επιτρέπει στη μονάδα μνήμης να ενημερωθεί
$\tilde{m}_t = m_t \circ o_t$ η πύλη εξόδου καθορίζει αν η πληροφορία μπορεί να βγει από τη μονάδα
$z_t = g(W_{yh}h_t + W_{ym}\tilde{m}_t)$

Οι τελευταίες τρεις εξισώσεις είναι που ορίζουν τις μονάδες μνήμης και περιγράφουν πώς οι πύλες εισόδου και εξόδου φυλάσσουν τα περιεχόμενα της μονάδας μνήμης. Περιγράφουν, επίσης, πώς η πύλη της λήθης μπορεί να κάνει τη μονάδα μνήμης να ξεχάσει τα περιεχόμενά της. Η σταθερά b_f έχει εισαχθεί για τις πύλες λήθης προκειμένου εκείνες να έχουν τιμή περίπου 1 στα αρχικά στάδια της εκπαίδευσης κάτι που πραγματοποιείται αρχικοποιώντας την b_f σε μια μεγάλη τιμή (όπως 5). Αν δεν γίνει αυτό, θα είναι δυσκολότερη η εκπαίδευση μεγάλης έκτασης εξαρτήσεων επειδή οι μικρότερες τιμές των πυλών λήθης θα δημιουργήσουν πρόβλημα συρρίκνωσης παραγώγων.

Εικόνα 2: Μονάδα μνήμης μακρού – βραχέος όρου



2.3 Μηχανική μετάφραση

Η νευρωνική μηχανική μετάφραση αποτελεί πρακτικά ένα πιθανοτικό γλωσσικό μοντέλο, με την ιδιαιτερότητα ότι η κατανομή πιθανότητας που αυτή μας προσφέρει για μία μετάφραση (στη γλώσσα - στόχο) είναι δεσμευμένη και καθορίζεται από την αμετάφραστη πρόταση (στη γλώσσα – πηγή) , με την οποία έχουμε προηγουμένως τροφοδοτήσει το σύστημά μας.

2.3.1 Βασικές αρχές της μηχανικής μετάφρασης

Η μηχανική μετάφραση βασίζεται στην ύπαρξη ενός συστήματος κωδικοποιητή –

αποκωδικοποιητή επαναληπτικών νευρωνικών δικτύων (Cho κ.α. 2014), (Sutskever κ.α. 2014). Το σύστημα αυτό, πρακτικά, αποκωδικοποιεί πιθανοτικά μια πρόταση – στόχο, δεδομένης της κωδικοποιημένης πρότασης εισόδου, με τις προτάσεις αυτές να μπορούν να είναι και διαφορετικού μεγέθους. Δεδομένου ενός ζευγαριού προτάσεων (S,T), S είναι η πρόταση εισόδου της ξένης γλώσσας και T η μετάφραση – στόχος, που θα θέλαμε για έξοδο όπου S = (s_1, s_2,… ,s_{m-1}, s_m) και T=(t_1 , t_2 , … , t_{n-1} , t_n) είναι οι λέξεις στο ζευγάρι των προτάσεων. Η αρχιτεκτονική κωδικοποιητή – αποκωδικοποιητή χρειάζεται να βρει τις πιθανότητες μετάφρασης για κάθε λέξη στην T, όπως φαίνεται στην παρακάτω εξίσωση:

$$p(T|S) = \prod_{j=1}^{n} p(t_j|t_{1:j-1}, S)$$

ενώ η δεσμευμένη πιθανότητα δίνεται από τον αποκωδικοποιητή ο οποίος χρησιμοποιεί την συνάρτηση softmax προκειμένου να παράγει ως έξοδο την κατανομή πιθανότητας σε όλες τις λέξεις του στόχου:

$$p(t_j = t|t_{1:j-1}, S) = softmax\left(f(h_j)\right)$$

όπου η f είναι μια συνάρτηση, η οποία μπορεί να μετατρέψει το περιεχόμενο της μετάφρασης- στόχου h_j σε ένα διάνυσμα μεγέθους ίδιου με το λεξιλόγιο. Το περιεχόμενο h_j ορίζεται ως εξής:

$$h_j = g(t_{j-1}, h_{j-1}, c) \quad (2.3.2.1)$$

με το c να είναι το διάνυσμα περιεχομένου της πρότασης εισόδου το οποίο έχει παραχθεί από τον κωδικοποιητή. Με t_{j-1} συμβολίζουμε την διανυσματική λεκτική αναπαράσταση της (j-1)-οστής λέξης και με h_{j-1} την κρυφή κατάσταση για την στιγμή j-1. Τέλος, η συνάρτηση g είναι μια μη γραμμική συνάρτηση ενεργοποίησης. Έτσι, χρησιμοποιούμε την πρόταση εισόδου και τις προηγούμενες μεταφρασμένες λέξεις για να προβλέψει την επόμενη λέξη.

2.3.2 Μηχανική μετάφραση με μηχανισμό προσοχής

Οι (Bahdanau κ.α. 2015) πρότειναν μια καινούρια αρχιτεκτονική για την μηχανική μετάφραση, η οποία περιείχε επαναληπτικό νευρωνικό δίκτυο διπλής κατεύθυνσης ως κωδικοποιητή, καθώς και έναν αποκωδικοποιητή, ο οποίος πραγματοποιεί αναζήτηση σε όλη την έκταση της πρότασης εισόδου κατά τη μετάφραση.

Η παραλλαγή στη σχέση που μας δίνει ο αποκωδικοποιητής έχει να κάνει με το διάνυσμα περιεχομένου της πρότασης εισόδου c. Πλέον αυτό το διάνυσμα δεν είναι ίδιο για όλες τις λέξεις της μετάφρασης - στόχου αλλά εξειδικεύεται ώστε στην j-οστή λέξη της μεταφρασμένης πρότασης αντιστοιχεί ένα c_j. δηλ. για την j-οστή λέξη του αποκωδικοποιητή, με την παραλλαγή αυτή, έχουμε:

$$h_j = g(t_{j-1}, h_{j-1}, c_j)$$

Το διάνυσμα περιεχομένου c_j εξαρτάται από τις κρυφές καταστάσεις που προκύπτουν από τις λέξεις της πρότασης εισόδου, και συγκεκριμένα αποτελούν έναν γραμμικό συνδυασμό τους. Το ιδιαίτερο με τον μηχανισμό προσοχής είναι ότι οι συντελεστές αυτού του γραμμικού συνδυασμού μαθαίνονται με ήπια ανάθεση από το δίκτυο κατά την διάρκεια της εκπαίδευσης, ως εξής:

$$e_{ij} = \alpha(h_{j-1}, h_i)$$

όπου το e_{ij} είναι το μοντέλο ανάθεσης για να αξιολογήσουμε την συσχέτιση μεταξύ της j-οστής λέξης της μετάφρασης και της i-οστής λέξης της πρότασης εισόδου. Όσο μεγαλύτερο βαθμό συσχέτισης αποδώσουμε, τόσο περισσότερο το σύστημά μας θα εστιάσει την προσοχή του στην i-οστή λέξη της εισόδου για να παράγει την j-οστή λέξη της μετάφρασης-στόχου.

Το διάνυσμα περιεχομένου της j-οστής λέξης, δεδομένων των συσχετίσεών της με όλες τις λέξεις της εισόδου, επομένως, είναι:

$$c_j = \sum_{i=1}^{m} \alpha_{ij} h_i$$

όπου α_{ij} τα κανονικοποιημένα βάρη για τις λέξεις της εισόδου, που προκύπτουν ως εξής:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{i=1}^{m} exp(e_{ij})}$$

Αξίζει να αναφερθεί ότι για το μηχανισμό προσοχής, η σημασία του κωδικοποιητή διπλής κατεύθυνσης είναι πολύ μεγάλη. Ο συνήθης κωδικοποιητής (μονής κατεύθυνσης) διαβάζει την ακολουθία εισόδου ξεκινώντας από το αρχικό σύμβολο και ολοκληρώνοντας στο τελευταίο. Ωστόσο, στην παρούσα περίπτωση χρειαζόμαστε το σύστημά μας να λαμβάνει υπ 'όψιν του, όχι μόνο τις προηγούμενες λέξεις αλλά κι εκείνες που ακολουθούν.

Το επαναληπτικό νευρωνικό δίκτυο διπλής κατεύθυνσης αποτελείται από ένα δίκτυο "προς τα εμπρός" και ένα "προς τα πίσω". Το "προς τα εμπρός" διαβάζει την πρόταση εισόδου μήκους T λέξεων με την κανονική της σειρά (από την πρώτη λέξη προς την τελευταία) και υπολογίζει μια ακολουθία "προς τα εμπρός" κρυφών καταστάσεων $h_i = (h_1, \ldots, h_T)$. Το "προς τα πίσω" δίκτυο διαβάζει την πρόταση με αντίστροφη σειρά (από το τέλος προς την αρχή) και έχει ως αποτέλεσμα μια ακολουθία από "προς τα πίσω" κρυφές καταστάσεις $h^b_i = (h^b_1, \ldots, h^b_T)$. Το κρυφό επίπεδο της i-οστής λέξης εισόδου προκύπτει με συνδυασμό των δύο ειδών κρυφών καταστάσεων (για παράδειγμα $[(h_i)^T; (h^b_i)^T]^T$). Λόγω της τάσης του επαναληπτικού νευρωνικού δικτύου να αναπαριστά καλύτερα τις πρόσφατες εισόδους, το συνδυασμένο κρυφό επίπεδο τείνει να εστιάζεται στις κοντινές λέξεις της i-οστής.

Μια παραλλαγή του μηχανισμού προσοχής από τον (Luong, 2016) αποτελεί η χρήση απλά των υψηλότερων επιπέδων του κωδικοποιητή και αποκωδικοποιητή (και όχι η παράθεση των επιπέδων από το κάθε κατεύθυνσης δίκτυο) και από την αρχική προσέγγιση του κρυφού επιπέδου της λέξης – στόχου, γίνεται ο υπολογισμός του e με πολλαπλασιαστική μορφή (αντί για παράθεση των κρυφών καταστάσεων πηγής και στόχου), δηλ.$e_{ij} = h_j W_a h_i$ ώστε στη συνέχεια να βρεθεί το τελικό κρυφό επίπεδο με μηχανισμό προσοχής.

2.3.3 Λεξιλόγιο νευρωνικής μηχανικής μετάφρασης και παραλλαγή του

Η νευρωνική μηχανική μετάφραση απαιτεί για τις δύο γλώσσες (πηγής και στόχου) από ένα λεξιλόγιο. Το λεξιλόγιο αυτό έχει μέγεθος V και περιέχει τις V συχνότερες λέξεις της κάθε γλώσσας, οι οποίες αντιμετωπίζονται ως μοναδικές. Όλες τις υπόλοιπες, το σύστημα τις βλέπει σαν άγνωστες, τις αποδίδει σε ένα κοινό σύμβολο στο λεξιλόγιο και στη συνέχεια τις τροφοδοτεί με το ίδιο διάνυσμα στον κωδικοποιητή/αποκωδικοποιητή. Το γεγονός αυτό καθιστά το σύστημα ελλιπές στον χειρισμό λέξεων που του είναι άγνωστες ή σπανίζουν.

Για την αντιμετώπιση του προβλήματος αυτού προτάθηκε από τους (Sennrich κ.α. 2016) η κωδικοποίηση με ακολουθίες τμημάτων λέξεων, εμπνευσμένη από την τεχνική κωδικοποίησης ζεύγους byte (Byte Pair Encoding, BPE). Η κωδικοποίηση ζεύγους byte είναι μια απλή τεχνική συμπίεσης δεδομένων που επαναληπτικά αντικαθιστά τα πιο συχνά ζεύγη bytes σε μια ακολουθία με

ένα μοναδικό, αχρησιμοποίητο byte. Αυτό υιοθετήθηκε για την κατάτμηση λέξεων όπου αντί να ενοποιούνται ζεύγη bytes, εννοποιούνται χαρακτήρες ή ακολουθίες χαρακτήρων.

Η διαδικασία παραγωγής του λεξιλογίου έχει ως εξής: Πρώτα, αρχικοποιείται το λεξιλόγιο συμβόλων με χαρακτήρες και κάθε λέξη αναπαρίσταται ως ακολουθία αποτελούμενη από χαρακτήρες και από έναν χαρακτήρα λήξης συμβόλου (πχ. "_"), ο οποίος επιτρέπει την επαναφορά της αρχικής αναπαράστασης μετά τη μετάφαση. Τα ζεύγη συμβόλων μετρώνται κατ' επανάληψιν και κάθε εμφάνιση του πιο συχνού ζεύγους (πχ. 'A','B') αντικαθίσταται με ένα νέο σύμβολο (που αντιστοιχεί στο 'AB'). Κάθε ενοποίηση παράγει ένα νέο σύμβολο το οποίο αναπαριστά ένα n-γραμμα χαρακτήρων. Τα συχνά n-γράμματα (δηλ. οι συχνότερες λέξεις και τα συχνότερα τμήματα λέξεων) καταλήγουν ενοποιημένες σε ένα σύμβολο και αποτελούν -μαζί με τους αρχικούς χαρακτήρες- το λεξιλόγιο της μηχανικής μετάφρασης.

Το σύστημα έχει λοιπόν την ικανότητα γενίκευσης και άρα παραγωγής νέων σύνθετων λέξεων, οι οποίες έχουν ως συνθετικά τμήματα λέξεων (προθέσεις, καταλήξεις, ρίζες λέξεων). Επιπλέον, του δίνεται η δυνατότητα απόδοσης λέξεων η οποίες απαιτούν απλά αντιστοιχίες μεταξύ των χαρακτήρων του αλφαβήτου της κάθε γλώσσας (όπως ονόματα) καθώς και φωνολογικές αντιστοιχίες (όπως τοπωνύμια ή διεθνείς επιστημονικούς όρους).

2.3.4 Συνολική περιγραφή της υλοποίησης της μηχανικής μετάφρασης βήμα προς βήμα

Εκπαίδευση

Ενσωμάτωση

Δεδομένης της κατηγορηματικής φύσης των λέξεων, το μοντέλο πρέπει πρώτα να ψάξει στα διανύσματα ενσωμάτωσης (γλώσσας πηγής και στόχου) για να ανακτήσει τις αντίστοιχες λεκτικές αναπαραστάσεις. Τα βάρη του πίνακα ενσωμάτωσης – ένα σύνολο βαρών ανά γλώσσα – μαθαίνονται από το σύστημα κατά τη διάρκεια της εκπαίδευσης εφόσον το σύνολο δεδομένων είναι αρκετά μεγάλο (αν και υπάρχουν και σύνολα έτοιμων προεκπαιδευμένων διανυσμάτων). Στα πλαίσια της ενσωμάτωσης, λοιπόν, έχουμε έναν πίνακα βαρών ενσωμάτωσης κωδικοποιητή και μια λειτουργία αντιστοίχισης λέξης λεξιλογίου σε διάνυσμα αναπαράστασης, καθώς και έναν πίνακα ενσωμάτωσης αποκωδικοποιητή και μια λειτουργία αντιστοίχισης από το διάνυσμα αναπαράστασης στη λέξη λεξιλογίου.

Κωδικοποίηση

Έχοντας, λοιπόν, τις λεκτικές αναπαραστάσεις τροφοδοτούμε με αυτές το κύριο δίκτυο το οποίο αποτελείται από δύο πολυστρωματικά επαναληπτικά νευρωνικά δίκτυα, έναν κωδικοποιητή για τη γλώσσα – πηγή και έναν αποκωδικοποιητή για τη γλώσσα – στόχο. Διαθέτουμε, λοιπόν, ένα κύτταρο LSTM με τόσες μονάδες όσες και οι συντεταγμένες των διανυσμάτων αναπαράστασης. Με τη χρήση αυτού του κυττάρου δημιουργούμε ένα δυναμικό δίκτυο κωδικοποίησης που αφού δεχτεί όλες τις λέξεις της ακολουθίας εισόδου υπολογίζει τις εξόδους του κωδικοποιητή καθώς και την κρυφή του κατάσταση. Σημαντικό είναι να αναφερθεί ότι το μήκος της ακολουθίας εισάγεται σαν είσοδος στο δυναμικό δίκτυο προκειμένου να αποφύγουμε τους περιττούς υπολογισμούς.

Αποκωδικοποίηση

Ο αποκωδικοποιητής χρειάζεται κι αυτός να έχει πρόσβαση στην πληροφορία πηγής και ένας απλός τρόπος για να επιτευχθεί αυτό είναι να τον αρχικοποιήσουμε με την τελευταία κρυφή κατάσταση του κωδικοποιητή. Η αποκωδικοποίηση γίνεται σε δύο στάδια, το βοηθητικό στάδιο και το στάδιο αποκωδικοποίησης. Στο πρώτο στάδιο, η βοηθητική συνάρτηση εκπαίδευσης παίρνει σαν είσοδο τον πίνακα ενσωμάτωσης αποκωδικοποίησης και δειγματοληπτεί την πρόταση στόχο, δίνοντάς την δείγμα – δείγμα στον βασικό αποκωδικοποιητή. Στο δεύτερο στάδιο, λειτουργεί ο βασικός αποκωδικοποιητής ο οποίος είναι ένα δίκτυο αποτελούμενο από κύτταρα μνήμης μακρού – βραχέος όρου με πλήθος μονάδων ίσο με τις παραμέτρους του διανύσματος ενσωμάτωσης. Αυτός, λοιπόν, παίρνει σαν είσοδο την κατάσταση του κωδικοποιητή καθώς και το περιεχόμενο της μνήμης. Το μέγιστο μήκος της πρότασης αποκωδικοποίησης είναι ανάλογο του μήκους της πρότασης

κωδικοποίησης με σταθερά αναλογίας α (συνήθως α = 1,5). Σε κάθε χρονικό βήμα στην πλευρά του αποκωδικοποιητή η πληροφορία περνάει από τα στρώματα του δικτύου και δίνει μια κρυφή κατάσταση η οποία χρησιμοποείται για να υπολογιστεί η κατανομή πρόβλεψης για τη μεταφρασμένη λέξη. Η κατανομή αυτή στη συνέχεια μέσω ενός επιπέδου προβολής καθορίζει ποια από τις λέξεις του λεξιλογίου στόχου είναι η μεταφρασμένη λέξη του βήματος αυτού.

Κόστος

Έχοντας, λοιπόν, τόσο τη μετάφραση του αποκωδικοποιητή όσο και τη μετάφραση στόχο μπορούμε να υπολογίσουμε μια συνάρτηση σφάλματος που μας δείχνει πόσο καλά λειτουργεί το σύστημά μας, η οποία είναι η διασταυρούμενη εντροπία. Έτσι το κόστος κατά την εκπαίδευση είναι το άθροισμα των διασταυρούμενων εντροπιών μεταφρασμένων λέξεων και λέξεων στόχου, για όλες τις λέξεις της πρότασης.

Υπολογισμός παραγώγου και βελτιστοποίηση

Στη συνέχεια παραγωγίζουμε τη συνάρτηση κόστους και ψαλιδίζουμε την παράγωγο ώστε η νόρμα της να μην υπερβαίνει την τιμή 5. Εφαρμόζουμε αλγόριθμο καθόδου παραγώγου και ανανεώνουμε τους πίνακες βαρών του συστήματος με οπισθοδιάδοση.

Αξιολόγηση

Μεταξύ των επαναλήψεων της εκπαίδευσης διεξάγεται μια διαδικασία, κατά την οποία το σύστημα αξιολογείται ως προς την ποιότητα της μετάφρασης. Ξεχωρίζονται λοιπόν δύο ενδεικτικά τμήματα του συνόλου δεδομένων, τα λεγόμενα σύνολα ανάπτυξης και ελέγχου. Μέσω του συνόλου ανάπτυξης βαθμολογείται η παραγωγή μετάφρασης, η οποία έχει πρόσβαση στην πληροφορία της μετάφρασης στόχου, ενώ με το σύνολο ελέγχου αξιολογείται η παραγωγή μετάφρασης χωρίς πρόσβαση στη μετάφραση – στόχο. Οι διάφορες τιμές των μετρικών αξιολόγησης, αντιστοιχίζονται στα σημεία ελέγχου, πράγμα που μας δίνει τη δυνατότητα να παρακολουθούμε την πορεία της εκπαίδευσης στο χρόνο, και τη σύγκλισή της. Επιπλέον, χάρις σε αυτά μπορούμε να επιλέγουμε και τις παραμέτρους του συστήματος που μεγιστοποιούν τις μετρικές αυτές.

Τεκμηρίωση

Αφού, λοιπόν, πραγματοποιηθεί η εκπαίδευση μπορούμε να δώσουμε στο σύστημα άγνωστες προτάσεις, τις οποίες δεν έχει ξαναδεί, στη γλώσσα-πηγή και να μας τις μεταφράσει στη γλώσσα στόχο. Η διαδικασία αυτή ονομάζεται τεκμηρίωση. Η διαφορά μεταξύ εκπαίδευσης και τεκμηρίωσης έγκειται στο στάδιο της αποκωδικοποίησης. Στην τεκμηρίωση έχουμε προτάσεις γλώσσας πηγής αλλά δεν έχουμε τις μεταφράσεις – στόχους και έτσι ο αποκωδικοποιητής δεν τροφοδοτείται από τις λέξεις της πρότασης στόχου. Αντ' αυτού, για να προβλέψει την επόμενη μεταφρασμένη λέξη χρησιμοποιεί την προηγούμενη που ο ίδιος έχει παράγει. Έχει, ωστόσο, αποδειχθεί ότι αν χρησιμοποιήσει την πληροφορία από παραπάνω από μία καλύτερες μεταφράσεις της προηγούμενης λέξης, η απόδοση του συστήματος βελτιώνεται αισθητά. Η αποκωδικοποίηση, λοιπόν, κατά την οποία λαμβάνουμε υπ' όψιν μόνο το μέγιστο του διανύσματος κατανομής και λαμβάνουμε υπ' όψιν την μία και μοναδική μετάφραση της λέξης στη συνέχεια καλείται "άπληστη αποκωδικοποίηση". Καλύτερα αποτελέσματα εμφανίζει η δεματική αποκωδικοποίηση, η οποία λαμβάνει υπ' όψιν της τις Β μέγιστες τιμές του διανύσματος κατανομής, και άρα τα Β κρυφά επίπεδα και τις Β μεταφράσεις της προηγούμενης λέξης για να παράγει την επόμενη, (όπου Β το μέγεθος του "δέματος", δηλαδή το πλήθος των μεταφράσεων που λαμβάνουμε υπ' όψιν) .

2.3.5 Αυτόματη αξιολόγηση μηχανικής μετάφρασης

Έχοντας, λοιπόν, μιλήσει για την αξιολόγηση του γλωσσικού νευρωνικού μοντέλου με το μέγεθος της περιπλοκής, σκόπιμη είναι και η περιγραφή της αξιολόγησης του συστήματος

νευρωνικής μηχανικής μετάφρασης με υποκαταστάτη δίγλωσσης αξιολόγησης (Papineni κ.α. 2002). Η έννοια "υποκατάστατης" οφείλεται στο γεγονός ότι το σύστημα αξιολόγησης πρακτικά υποκαθιστά την αξιολόγηση από άνθρωπο και βασίζεται στην ιδέα ότι όσο εγγύτερα είναι μια μετάφραση σε μια επαγγελματική ανθρώπινη μετάφραση, τόσο καλύτερη είναι, με κριτήρια την επάρκεια, την ευφράδεια, και την πιστότητα της.

Για την εφαρμογή της αξιολόγησης, το σύστημα απαιτεί ένα σώμα ανθρώπινων ποιοτικών μεταφράσεων που περιέχει μία μετάφραση, ή και παραπάνω, για κάθε πρόταση και βρίσκει αντιστοιχίες λέξεων και φράσεων μεταξύ μεταφράσεων αναφοράς και υποψήφιας μετάφρασης . Το σύστημα κατασκευάζει n-γράμματα λέξεων (συνήθως με n=1,...,4) για όλες τις προτάσεις (υποψήφια και μεταφράσεις αναφοράς) και υπολογίζει τις ελάχιστες φορές τις οποίες το n-γραμμα εμφανίστηκε και στα δύο είδη προτάσεων. Το γεγονός ότι υπολογίζονται οι ελάχιστες φορές εμφάνισης του n-γράμματος, αποσκοπεί στο να θεωρηθεί θετικό από το σύστημα η ύπαρξη των σωστών φράσεων τόσες φορές όσες και στη μετάφραση και να αποφευχθούν καλές βαθμολογίες σε μεταφράσεις που επαναλαμβάνουν λέξεις ή φράσεις (ακόμα κι αν αυτές ανήκουν στις μεταφράσεις αναφοράς). Η συνάρτηση, λοιπόν, η οποία υπολογίζει την εγγύτητα της μηχανικής με την ιδανική μετάφραση (ή τις ιδανικές μεταφράσεις) είναι η τροποποιημένη ακρίβεια n-γράμματος:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}\,(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')}$$

Η ανάγκη για ύπαρξη n-γραμμάτων με n>1 σχετίζεται με το ότι ίδια 1-γράμματα εξασφαλίζουν επάρκεια, ενώ ίδια n-γράμματα με n>1 εξασφαλίζουν ευφράδεια του κειμένου. Για το τελικό αποτέλεσμα σε κείμενο παίρνουμε τον σταθμισμένο μέσο λογάριθμο των τροποποιημένων ακριβειών με ομοιόμορφα βάρη (ο οποίος μας δίνει όμοια αποτελέσματα με τον γεωμετρικό μέσο) και τον πολλαπλασιάζουμε με την ποινή συντομίας, η οποία αν η υποψήφια μετάφραση c είναι μικρότερη από την μικρότερη από τις αναφοράς r είναι ίση με $BP = e^{1-\frac{r}{c}}$.

Άρα, εν τέλει, η μετρική αξιολόγησης μηχανικής μετάφρασης είναι:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

ή λογαριθμικά:

$$log(BLEU) = min\left(1 - \frac{r}{c}\right) + \sum_{n=1}^{N} w_n log p_n$$

όπου N=4 και w_n=1/N.

2.4 Συσταδοποίηση

Η συσταδοποίηση (clustering) είναι μια συνήθης τεχνική για την ανάλυση στατιστικών δεδομένων, η οποία χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένων της μηχανικής μάθησης, και της εξόρυξης δεδομένων. Αποτελεί διαδικασία ομαδοποίησης παρόμοιων αντικειμένων σε διαφορετικές ομάδες, ή πιο συγκεκριμένα, το διαχωρισμό ενός συνόλου δεδομένων σε υποσύνολα, με βάση κάποιο ορισμένο κριτήριο ανομοιότητας (Madhulatha 2012).

Η συσταδοποίηση είναι η χαρακτηριστικότερη τεχνική μη επιβλεπόμενης μάθησης και όπως κάθε άλλη τεχνική αυτού του είδους, ασχολείται με την εύρεση μιας δομής σε μια συλλογή μη ετικετοποιημένων (unlabeled) δεδομένων. Συνεπώς, μία συστάδα (cluster) είναι μια συλλογή από

αντικείμενα που είναι "παρόμοια" μεταξύ τους και είναι "ανόμοια" με τα αντικείμενα που ανήκουν σε άλλες συστάδες.

Καθοριστικό ρόλο στην δημιουργία των clusters διαδραματίζει ο βαθμός που τα στοιχεία του συνόλου δεδομένων (τα οποία αναπαρίστανται ως διανύσματα) απέχουν μεταξύ τους. Ως μετρική απόστασης λοιπόν,μεταξύ δύο δεδομένων u,v, μια συνήθης επιλογή είναι η ευκλείδεια απόσταση $\left|\left|u-v\right|\right|_2$, ενώ στην ειδική περίπτωση που το σύνολο δεδομένων αποτελείται από έγγραφα, επιλέγεται κατ' εξοχήν η συνημιτονοειδής απόσταση $1-\frac{uv}{\left|\left|u\right|\right|_2\left|\left|v\right|\right|_2}$.

Οι αλγόριθμοι συσταδοποίησης μπορούν να είναι είτε ιεραρχικοί είτε διαχωριστικοί (partitional). Οι πρώτοι, σε κάθε τους στάδιο, δημιουργούν συστάδες χρησιμοποιώντας τις συστάδες που έχουν ήδη καθοριστεί στο προηγούμενο στάδιο και δεν επιτρέπουν μεταπηδήσεις στοιχείων που έχουν συσταδοποιηθεί σε άλλες ομάδες. Οι δεύτεροι, αντίθετα, επιτρέπουν μεταπηδήσεις στοιχείων σε άλλες ομάδες καθ' όλη τη διαδικασία εφαρμογής τους, εφόσον αυτά πληρούν ορισμένα απαιτούμενα κριτήρια.

2.4.1 Διαχωριστική συσταδοποίηση

Συσταδοποίηση κέντρων (k-means)

Στην k-means συσταδοποίηση, οι συστάδες αντιπροσωπεύονται από ένα κεντρικό διάνυσμα, το οποίο μπορεί να μην είναι αναγκαστικά μέλος του συνόλου δεδομένων. Για τα k clusters, όπου k προκαθορισμένος αριθμός, το ζητούμενο του αλγορίθμου είναι η εύρεση των κέντρων των k ομάδων και η ανάθεση των αντικείμενων στο πλησιέστερο κέντρο, έτσι ώστε να ελαχιστοποιηθούν οι τετραγωνικές αποστάσεις από αυτό. Οι παραλλαγές των k-μέσων περιλαμβάνουν συχνά τέτοιες βελτιστοποιήσεις όπως την επιλογή των καλύτερων από πολλαπλές διαδρομές, αλλά και τον περιορισμό των κεντρικών διανυσμάτων σε μέλη του συνόλου δεδομένων (k-medoids), την επιλογή των αρχικών κέντρων λιγότερο τυχαία ή επιτρέποντας την ανάθεση ασαφών ομάδων.

Συσταδοποίηση κατανομών

Στην συσταδοποίηση που βασίζεται σε κατανομές, η κάθε ομάδα ορίζεται ως σύνολο αντικειμένων που ανήκουν με μεγάλη πιθανότητα στην ίδια κατανομή. Μια εξέχουσα μέθοδος είναι γνωστή ως Gaussian Mixture Models (που χρησιμοποιούν τον αλγόριθμο Expectation - Maximization). Εδώ, το σύνολο δεδομένων συνήθως διαμορφώνεται με ένα σταθερό αριθμό Gaussian κατανομών (για αποφυγή του overfitting) που αρχικοποιούνται τυχαία και των οποίων οι παράμετροι βελτιστοποιούνται επαναληπτικά ώστε να ταιριάζουν καλύτερα το σύνολο δεδομένων.

Συσταδοποίηση πυκνοτήτων

Στην συσταδοποίηση πυκνοτήτων, στόχος είναι να εντοπιστούν συστάδες αυθαίρετου σχήματος με κριτήριο την πυκνότητα των περιοχών - clusters. Η βασική ιδέα είναι ότι, για κάθε σημείο ενός συμπλέγματος, η γειτονιά μιας δεδομένης ακτίνας πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό σημείων, δηλαδή, η πυκνότητα στη γειτονιά πρέπει να υπερβαίνει κάποιο προκαθορισμένο όριο.

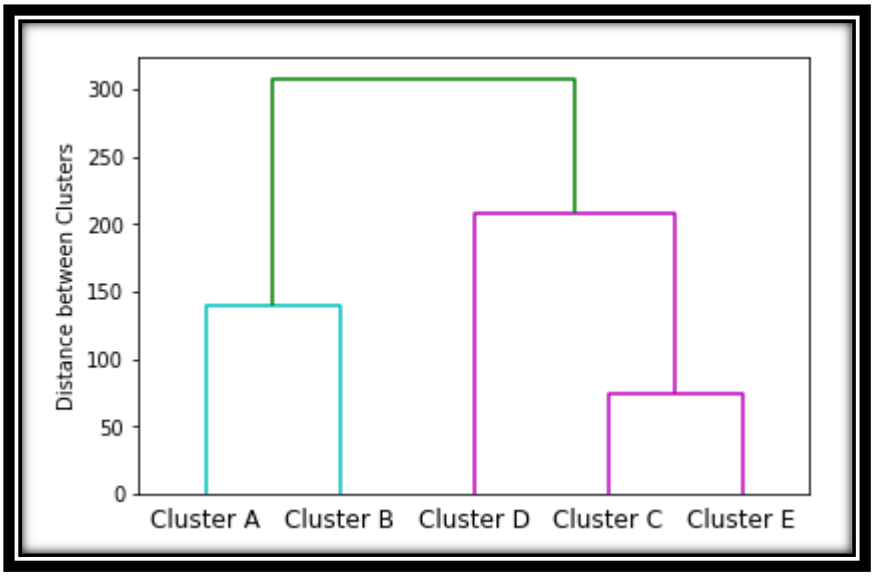
2.4.2 Ιεραρχική συσταδοποίηση

Οι ιεραρχικοί αλγόριθμοι μπορούν να είναι είτε συσσωρευτικοί (bottom - up) ή διαιρετικοί (top-down). Οι συσσωρευτικοί (agglomerative) αλγόριθμοι αρχίζουν θεωρώντας κάθε στοιχείο ως ξεχωριστό cluster, τα οποία στη συνέχεια συγχωνεύονται διαδοχικά σε μεγαλύτερες ομάδες. Οι διαιρετικοί αλγόριθμοι αρχίζουν με ολόκληρο το σύνολο ως ένα cluster και το χωρίζουν διαδοχικά σε μικρότερα σύνολα.

Οι συστάδες οι οποίες παράγονται από την ιεραρχική συσταδοποίηση συχνά απεικονίζονται

με δενδρογράμματα (Εικόνα 3). Τα δενδρογράμματα αποτελούν γραφικές αναπαραστάσεις οι οποίες απεικονίζουν τον τρόπο με τον οποίο συντίθεται κάθε σύμπλεγμα, σχεδιάζοντας μια σύνδεση σχήματος U για κάθε δύο συστάδες που ενώνονται σε μία στους συσσωρευτικούς αλγορίθμους (ή που προέκυψαν από μία, στους διαιρετικούς). Το ύψος της κορυφής του συνδέσμου U είναι η απόσταση μεταξύ των clusters που αποτελούν διαμμέρισή του.

Εικόνα 3: Δενδρόγραμα



Ιεραρχική συσσωρευτική συσταδοποίηση

Κατά την ιεραρχική συσταδοποίηση, αρχικά όλα τα στοιχεία του συνόλου δεδομένων αποτελούν χωριστά clusters, και τα στοιχεία με τη μικρότερη απόσταση ενώνονται σε μία συστάδα. Για την επιλογή των συστάδων που θα ενωθούν σε κάθε επανάληψη χρησιμοποιούνται διάφορες μέθοδοι συνένωσης, με κριτήριο είτε το πόσο απέχουν τα πλησιέστερα σημεία των δύο συστάδων (single method) ή τα πιο απόμακρα (complete method) ή τον μέσο όρο των αποστάσεων των στοιχείων τους (average method). Έτσι, σε κάθε επανάληψη ο αλγόριθμος έχει ως αποτέλεσμα το πλήθος των clusters μειώνεται κατά ένα. Η παραγωγή συστάδων που επιλέγεται βασίζεται σε κάποιο κριτήριο όπως για παράδειγμα η απόσταση μεταξύ των συστάδων που έχουν απομείνει να μην υπερβαίνει έναν προκαθορισμένο αριθμό.

Ιεραρχική διαιρετική συσταδοποίηση

Αυτή η παραλλαγή της ιεραρχικής συσταδοποίησης ονομάζεται συσσώρευση από πάνω προς τα κάτω (top down) ή διαχωριστική συσταδοποίηση (divisive). Αρχίζουμε στην κορυφή με όλα τα έγγραφα σε ένα cluster. Το cluster χωρίζεται χρησιμοποιώντας έναν αλγόριθμο επίπεδης συσταδοποίησης. Αυτή η διαδικασία εφαρμόζεται αναδρομικά μέχρις ότου κάθε έγγραφο βρίσκεται στη δική του ομάδα. Η διαιρετική συσταδοποίηση είναι εννοιολογικά πιο πολύπλοκη από τη συσσώρευση από τη βάση προς την κορυφή, δεδομένου ότι πρέπει να χρησιμοποιήσουμε έναν δεύτερο, επίπεδο αλγόριθμο συσταδοποίησης ως “υπορουτίνα” . Έχει το πλεονέκτημα ότι είναι πιο αποτελεσματική αν δεν δημιουργήσουμε μια πλήρη ιεραρχία αλλά έχουμε προκαθορισμένο αριθμό clusters.

2.5 Latent Dirichlet Allocation

Η Latent Dirichlet Allocation (LDA) αποτελεί γενετικό πιθανοτικό μοντέλο σε ένα σώμα κειμένων (Blei κ.α. 2003) και βασίζεται στην παραδοχή ότι όλα τα έγγραφα που ανήκουν σε ένα σώμα κειμένου χαρακτηρίζονται από κατανομές σε λανθάνοντα θέματα, ενώ τα θέματα χαρακτηρίζονται από πιθανοτικές κατανομές σε λέξεις.

2.5.1 Μαθηματική μοντελοποίηση της LDA

Το μοντέλο της LDA είναι ένα ιεραρχικό μοντέλο Bayes τριών επιπέδων το οποίο κατά την μοντελοποίηση σωμάτων κειμένου, θεωρεί την παρακάτω διαδικασία παραγωγής για ένα σώμα με D έγγραφα και K θέματα:

1. Για κάθε θέμα k, σχεδιάζει την κατανομή του θέματος σε λέξεις β_k , k = 1 … K
2. Για κάθε έγγραφο d, σχεδιάζει την κατανομή του εγγράφου σε θέματα , θ_d ~ Dirichlet(α) , d = 1 … D
3. Για την κάθε λέξη i στο έγγραφο d :
   (α) Σχεδιάζει έναν θεματικό δείκτη z_{di} ~ multinomial(θ)
   β) Επιλέγει τις λέξεις w_{ij} ~ multinomial($\beta_{z_{di}}$)

Αρκετές παραδοχές γίνονται χάριν απλοποίησης σ' αυτό το βασικό μοντέλο. Πρώτον, η διαστατικότητα k της κατανομής Dirichlet (και άρα η διαστατικότητα της θεματικής μεταβλητής z) θεωρείται γνωστή και συγκεκριμένη. Δεύτερον, οι λεκτικές πιθανότητες παραμετροποιούνται από έναν πίνακα β, διαστάσεων $k \times V$, όπου $\beta_{ij} = p(w^j = 1|z^i = 1)$, τον οποίο προς το παρόν χρησιμοποιούμε σαν μια συγκεκριμένη ποσότητα που πρέπει να εκτιμηθεί. Τέλος, η εκτίμηση Poisson δεν είναι καθοριστική για τίποτα που ακολουθεί ενώ μπορούν να χρησιμοποιηθούν πιο ρεαλιστικές κατανομές στην έκταση του εγγράφου. Επιπλέον, αξίζει να σημειωθεί ότι το πλήθος λέξεων N είναι ανεξάρτητο από όλες τις άλλες μεταβλητές παραγωγής δεδομένων (θ και z) . Είναι, λοιπόν, μια βοηθητική μεταβλητή και γενικά θα αγνοήσουμε την τυχαιότητά της στην επακόλουθη ανάπτυξη.

Μια k-διάστατη τυχαία μεταβλητή Dirichlet θ μπορεί να πάρει τιμές στο (k-1)-simplex και έχει την παρακάτω πυκνότητα πιθανότητας σε αυτό το simplex:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

όπου η παράμετρος α είναι ένα διάνυσμα k διαστάσεων με στοιχεία $\alpha_i > 0$, όπου η Γ(x) είναι η συνάρτηση Γάμμα. Η Dirichlet είναι μια κατάλληλη κατανομή στο simplex – ανήκει στην οικογένεια εκθετικών , έχει πεπερασμένων διαστάσεων στατιστικά επαρκείς παραμέτρους και είναι συζυγής της πολυωνυμικής κατανομής. Δεδομένων των παραμέτρων α και β, η από κοινού κατανομή ενός μίγματος θεμάτων θ, ενός συνόλου N θεμάτων z και ενός συνόλου N λέξεων w δίνεται από την σχέση:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta),$$

όπου η $p(z_n|\theta)$ είναι απλά η θ_i για το μοναδικό i για το οποίο ισχύει $z_n^i = 1$. Ολοκληρώνοντας ως προς θ και αθροίζοντας πάνω στο z, παίρνουμε την οριακή κατανομή ενός εγγράφου:

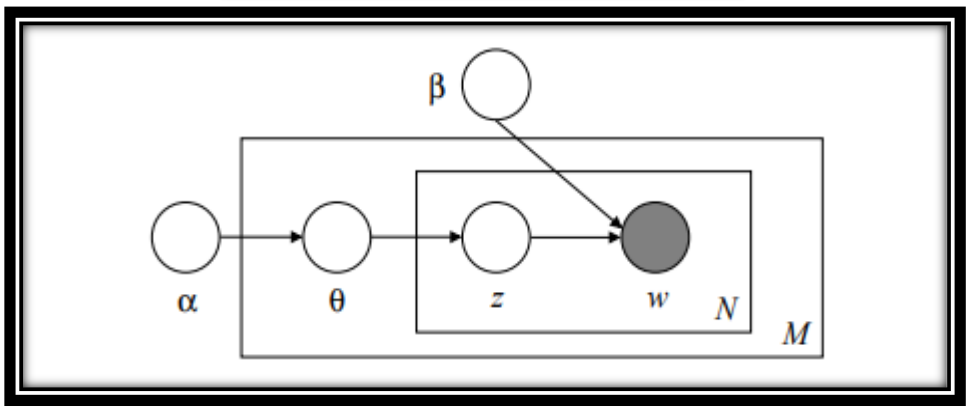
$$p(\boldsymbol{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

Τελικά, παίρνοντας το γινόμενο των οριακών πιθανοτήτων ξεχωριστών εγγράφων, λαμβάνουμε την πιθανότητα του σώματος:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

Το μοντέλο λανθάνουσας αντιστοιχίας Dirichlet αναπαρίσταται ως ένα πιθανοτικό γραφικό μοντέλο στην Εικόνα 4. Όπως φαίνεται ξεκάθαρα στην εικόνα, υπάρχουν τρία επίπεδα στην LDA αναπαράσταση. Οι παράμετροι α και β είναι παράμετροι επιπέδου σώματος κειμένου, που εκτιμώνται να έχουν δειγματοληπτηθεί μια φορά στη διαδικασία της παραγωγής ενός σώματος. Οι μεταβλητές θ_d είναι μεταβλητές επιπέδου εγγράφου, που έχουν δειγματοληπτηθεί μια φορά ανά έγγραφο. Τέλος, οι μεταβλητές z_{dn} και w_{dn} είναι μεταβλητές επιπέδου λέξεων και έχουν δειγματοληπτηθεί μια φορά για κάθε λέξη κάθε εγγράφου.

Εικόνα 4: Πλάκα αναπαράστασης LDA



Είναι σημαντικό να διακρίνουμε την LDA από ένα απλό πολυωνυμικό μοντέλο ομαδοποίησης Dirichlet. Ένα κλασικό μοντέλο ομαδοποίησης θα περιλάμβανε ένα μοντέλο δύο επιπέδων στο οποίο μια Dirichlet δειγματοληπτείται μια φορά για ένα σώμα κειμένου, μια πολυωνυμική μεταβλητή ομαδοποίησης επιλέγεται μια φορά για κάθε έγγραφο στο σώμα κειμένου και ένα σύνολο λέξεων επιλέγεται για το έγγραφο με συνθήκη τη μεταβλητή ομαδοποίησης. Όπως με πολλά μοντέλα ομαδοποίησης, ένα τέτοιου είδους μοντέλο περιορίζει ένα έγγραφο ώστε να σχετίζεται μόνο με ένα θέμα. Η λανθάνουσα αντιστοιχία Dirichlet , από την άλλη πλευρά , περιλαμβάνει τρία επίπεδα και ειδικά ο θεματικός κόμβος δειγματοληπτείται κατ' επανάληψιν μέσα στο έγγραφο. Σ' αυτό, λοιπόν, το μοντέλο τα έγγραφα μπορούν να συσχετιστούν με πολλαπλά θέματα το καθένα.

Το βασικό ζήτημα υλοποίησης που προκύπτει με την LDA είναι ο υπολογισμός της ύστερης κατανομής των κρυφών μεταβλητών δεδομένου ενός εγγράφου:

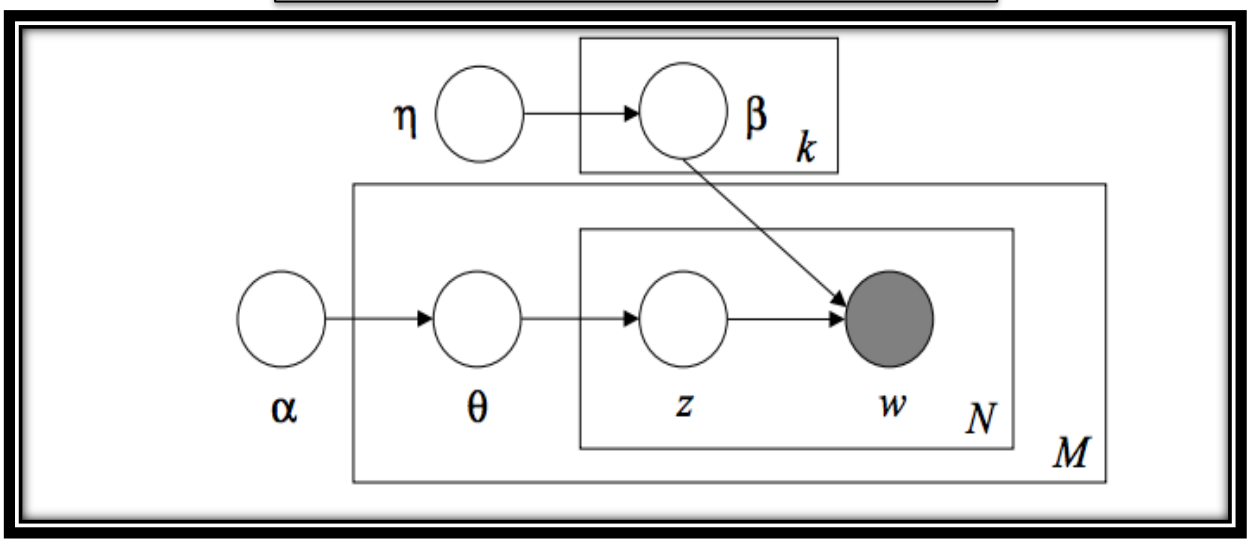
$$p(\theta,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta) = \frac{p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)}{p(\boldsymbol{w}|\alpha,\beta)}$$

Δυστυχώς, αυτή η κατανομή είναι γενικά δύσχρηστη στον υπολογισμό της. Έτσι, για να κανονικοποιήσουμε την κατανομή αθροίζουμε πάνω στις κρυφές μεταβλητές και γράφουμε με όρους των παραμέτρων του μοντέλου:

$$p(\boldsymbol{w}|\alpha,\beta) = \frac{\Gamma(\sum_i a_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

μια συνάρτηση που είναι δύσχρηστη εξ' αιτίας της εξάρτησης μεταξύ θ και β στην άθροιση πάνω σε λανθάνοντα θέματα. Παρά την δυσκολία στην ακριβή υλοποίηση, λοπόν, υπάρχουν αρκετοί προσεγγιστικοί αλγόριθμοι για υλοποίση της LDA. Για την αρχικοποίηση της κατανομής του εγγράφου σε θέματα εισάγεται μία ακόμα παράμετρος η, από την οποία προκύπτει η θεματική κατανομή εγγράφου ως $\beta_k \sim$ Dirichlet (η) (Εικόνα 5).

Εικόνα 5: Πλάκα αναπαράστασης εξομαλυμένης LDA



Για την εκτίμηση των παραμέτρων η μεταγενέστερη κατανομή είναι:

$$p(\boldsymbol{z},\theta,\beta|w,\alpha,\eta) = \frac{p(\boldsymbol{z},\theta,\beta|\alpha,\eta)}{p(\boldsymbol{w}|\alpha,\eta)} \quad (2.5.1.1)$$

Μπορούμε να αναλύσουμε ένα σώμα εγγράφων με LDA, εξετάζοντας την μεταγενέστερη κατανομή των θεμάτων β, των θεματικών ποσοστών θ, και των θεματικών αναθέσεων z δεσμευμένων στα έγγραφα. Αυτό αποκαλύπτει λανθάνουσα δομή στη συλλογή η οποία μπορεί να χρησιμοποιηθεί για πρόβλεψη ή για διερεύνηση δεδομένων. Αυτή η μεταγενέστερη πληροφορία δεν μπορεί να υπολογιστεί άμεσα και ένας συνήθης τρόπος προσέγγισής της είναι η τεκμηρίωση μεταβλητών, είτε διατρέχοντας όλα τα δεδομένα είτε με απ'ευθείας ενημέρωση (Hoffman κ.α. 2010) .

2.5.2 Τεκμηρίωση μεταβλητών Bayes για την LDA

Η τεκμηρίωση μεταβλητών Bayes χρησιμοποιεί την μεταβλητή φ για να παραμετροποιήσει τις ύστερες θεματκές αναθέσεις z ανά λέξη $q(z_{di} = k) = \varphi_{dw_{di}k}$, τη μεταβλητή γ για την παραμετροποίηση των ύστερων θεματικών βαρών θ ανά έγγραφο $q(\theta_d) = Dirichlet(\theta_d, \gamma_d)$ και την μεταβλητή λ για την παραμετροποίηση των ύστερων θεμάτων β $q(\beta_k) = Dirichlet(\beta_k; \lambda_k)$. Έτσι, η

κατανομή (2.4.1.1) προσεγγίζεται με μια απλούστερη κατανομή $q(z,\theta,\beta|\lambda,\varphi,\gamma)$ και οι παράμετροι λ,φ,γ βελτιστοποιούνται ούτως ώστε να μεγιστοποιούν το κατώτατο όριο απόδειξης (Evidence Lower Bound ELBO) :

$$logp(w|\alpha,\eta) \geqslant L(w,\varphi,\gamma,\lambda) \equiv E_q[logp(w,z,\theta,\beta|\alpha,\eta)] - E_q[logq(z,\theta,\beta)]$$

η οποία παραγοντοποιείται ως εξής:

$$L(w,\varphi,\gamma,\lambda) = \sum_d E_q[logp(w_d|\theta_d,z_d,\beta)] + E_q[logp(z_d|\theta_d)] - E_q[logq(z_d)] + E_q[logp(\theta_d|\alpha)] - E_q[logq(\theta_d)] + \frac{\left(E_q[logp(\beta|\eta)] - E_q[logq(\beta)]\right)}{D}$$

Αναλύουμε τις παραπάνω αναμενόμενες τιμές ως συναρτήσεις μεταβλητών παραμέτρων. Αυτό αποκαλύπτει ότι από τις λέξεις μας ενδιαφέρει μόνο η μεταβλητή n_{dw} , δηλαδή το πλήθος των φορών που η λέξη w εμφανίζεται στο έγγραφο d. Χρησιμοποιώντας την τεκμηρίωση μεταβλητών Bayes τα έγγραφα μπορούν να συνοψιστούν από το πλήθος των διάφορων λέξεών τους:

$$L = \sum_d \sum_w n_{dw} \sum_k \varphi_{dwk}\left(E_q[log\theta_{dk}] + E_q[log\beta_{kw}] - log\varphi_{dwk}\right) - logГ\left(\sum_k \gamma_{dk}\right) + \sum_k (\alpha - \gamma_{dk})E_q[log\theta_{dk}] + logГ(\gamma_{dk}) + \frac{\left(\sum_k - \log Г(\sum_w \lambda_{kw}) + \sum_w (\eta - \lambda_{kw})E_q[log\beta_{kw}] + logГ(\lambda_{kw})\right)}{D} + logГ(K\alpha) - KlogГ(\alpha) + \frac{\left(logГ(W\eta) - WlogГ(\eta)\right)}{D} \triangleq \sum_d l(n_d,\varphi_d,\gamma_d,\boldsymbol{\lambda})$$

(2.5.2.1)

όπου W το μέγεθος του λεξιλογίου, D το πλήθος των εγγράφων και $l(n_d, \varphi_d, \gamma_d, \lambda)$ η συμβολή του εγγράφου d στο όριο ELBO. Η L μπορεί να βελτιστοποιηθεί με αύξηση συντεταγμένων στις μεταβλητές παραμέτρους φ, γ, λ :

$$\varphi_{dwk} \propto expE_q[log\theta_{dk}] + E_q[log\beta_{kw}]\,,\ \gamma_{dk} = \alpha + \sum_w n_{dw}\varphi_{dwk}\ ,\ \lambda_{kw} = \eta + \sum_d n_{dw}\varphi_{dwk}$$

(2.5.2.2)

όπου η Ψ δηλώνει τη συνάρτηση δίγαμμα (την πρώτη παράγωγο του λογάριθμου της συνάρτησης γάμμα).

Οι επαναλήψεις στις εξισώσεις (2.5.2.2) εγγυώνται σύγκλιση σε ένα σταθερό σημείο του ορίου ELBO. Σε αναλογία με τον αλγόριθμο Expectation – Maimization (EM) μπορούμε να χωρίσουμε αυτές τις επαναλήψεις σε βήμα "Ε" όπου κατ' επανάληψιν ενημερώνονται τα γ και φ ως την σύγκλιση, διατηρώντας το λ σταθερό – και σε βήμα "Μ" όπου αναβαθμίζεται το λ δεδομένου του φ. Πρακτικά, ο αλγόριθμος αυτός συγκλίνει σε μια καλύτερη λύση αν επαναρχικοποιούμε το γ και το φ πριν από κάθε βήμα ΄'Ε'.

Αλγόριθμος 3: Δεματική τεκμηρίωση μεταβλητών Bayes για τη Λανθάνουσα Αντιστοιχία Dirichlet

1: $Initialize\ \boldsymbol{\lambda}\ randomly$
2: $\boldsymbol{while}\ relative\ improvement\ in\ L(\boldsymbol{w},\boldsymbol{\varphi},\boldsymbol{\gamma},\boldsymbol{\lambda}) > 0.00001\ \boldsymbol{do}$

3: $E\ step$:
4: $\boldsymbol{for}\ d\ =\ 1\ to\ D\ \boldsymbol{do}$
5: $Initialize\ \gamma_{dk} = 1$
6: $\boldsymbol{repeat}$
7: $Set\ \varphi_{dwk} \propto [expE_q(log\theta_{dk}) + E_q(log\beta_{kw})]$
8: $Set\ \gamma_{dk} = a + \sum_w \varphi_{dwk} n_{dw}$
9: $\boldsymbol{until}\ \frac{1}{K}\sum_k |changein\gamma_{dk}| < 0.00001$
10: $\boldsymbol{end\ for}$
11: $M\ step$:
12: $Set\ \lambda_{kw} =\ \eta + \sum_d n_{dw}\varphi_{dwk}$
13: $\boldsymbol{end\ while}$

Ο παραπάνω αλγόριθμος δουλεύει με "δέματα" (batches), δηλαδή διατρέχει πρώτα ολόκληρο το σύνολο δεδομένων για να κάνει τους διάφορους υπολογισμούς και μετά ενημερώνει τις παραμέτρους. Υπάρχει, ωστόσο, μια παραλλαγή του, η οποία πραγματοποιεί απ' ευθείας ενημέρωση της παραμέτρου λ σε κάθε στοιχείο (ή υποσύνολο : mini-batch) του συνόλου δεδομένων και για μεγάλα σύνολα δεδομένων συγκλίνει γρηγορότερα.

Μια καλή συνθήκη για τα θέματα λ είναι το L να είναι όσο το δυνατόν υψηλότερο μετά την προσαρμογή των μεταβλητών παραμέτρων γ και φ στο E-βήμα του αλγορίθμου. Ο στόχος είναι να θέσουμε $\boldsymbol{\lambda}$ που να μεγιστοποιεί το $L(\boldsymbol{n}, \boldsymbol{\lambda}) \triangleq \sum_d l(n_d, g(n_d, \boldsymbol{\lambda}), \varphi(n_d, \boldsymbol{\lambda}), \boldsymbol{\lambda})$ όπου η $l(n_d, \gamma_d, \varphi_d, \boldsymbol{\lambda})$ είναι η συμμετοχή του εγγράφου d στο όριο ELBO της εξίσωσης (2.5.2.1).

Η "απ' ευθείας LDA" περιγράφεται στον αλγόριθμο 4. Καθώς το τ-οστό διάνυσμα του πλήθους των λέξεων n_t παρατηρείται, πραγματοποιείται ένα E-βήμα για να βρεθούν οι τοπικά βέλτιστες τιμές των γ_t και φ_t , διατηρώντας το λ σταθερό. Στη συνέχεια υπολογίζουμε το $\tilde{\lambda}$, την τιμή δηλαδή του λ που θα ήταν βέλτιστη (δεδομένου του φ_t) εάν ολόκληρο το σώμα κειμένου μας αποτελούνταν από το έγγραφο n_t επαναλαμβανόμενο D φορές (όπου D το μέγεθος του σώματος κειμένων). Στη συνέχεια, ενημερώνουμε το λ χρησιμοποιώντας σταθμισμένο μέσο της προηγούμενης τιμής του και του $\tilde{\lambda}$. Το βάρος που αποδίδεται στο $\tilde{\lambda}$ δίνεται από τη σχέση $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$, όπου το $\kappa \in (0.5,1]$ελέγχει το ρυθμό με τον οποίο οι παλιές τιμές του $\tilde{\lambda}$ ξεχνιούνται και το $\tau_0 \geq 0$επιβραδύνει τις αρχικές επαναλήψεις του αλγορίθμου. Η συνθήκη $\kappa \in (0.5,1]$ εγγυάται την σύγκλιση. Όταν η ενημέρωση του $\boldsymbol{\lambda}$ δεν γίνεται στοιχείο προς στοιχείο αλλά σε μικρά υποσύνολα του συνόλου δεδομένων αποτελούμενα από S παρατηρήσεις, το $\tilde{\lambda}_{\kappa w}$ υπολογίζεται ως εξής: $\tilde{\lambda}_{\kappa w} = \eta + \frac{D}{S}\sum_s n_{tsk}\varphi_{tskw}$, όπου n_{ts} είναι το s-οστό έγγραφο στο υποσύνολο t.

Αλγόριθμος 4: Απευθείας τεκμηρίωση μεταβλητών Bayes για τη Λανθάνουσα Αντιστοιχία Dirichlet

1: $Define\ \rho_t \triangleq (\tau_0 + t)^{-\kappa}$
2: $Initialize\ \boldsymbol{\lambda}\ randomly$
3: $\boldsymbol{for}\ t = 0\ \boldsymbol{to}\ \infty\ \boldsymbol{do}$
4: $E\ step$:
5: $Initialize\ \gamma_{tk} = 1$
6: $\boldsymbol{repeat}$
7: $Set\ \varphi_{dwk} \propto [expE_q(log\theta_{dk}) + E_q(log\beta_{kw})]$
8: $Set\ \gamma_{dk} = a + \sum_w \varphi_{dwk} n_{dw}$

9: $\boldsymbol{until}\ \frac{1}{K}\sum_k |change\ in\ \gamma_{dk}| < 0.00001$
10: $M\ step$:
11: $Compute\ \tilde{\lambda}_{kw} = \eta + \mathrm{D}n_{dw}\varphi_{dwk}$
12: $Set\ \lambda = (1-\rho_t)\lambda + \rho_t\tilde{\lambda}$
13: $\boldsymbol{end\ for}$

Στην δεματική LDA, οι εκτιμήσεις των υπερπαραμέτρων α και η προσεγγίζονται, δοσμένων των γ και λ, χρησιμποιώντας γραμμικού χρόνου μέθοδο Newton – Raphson. Παρόμοια, μπορούμε να ενσωματώσουμε τις ενημερώσεις για τα α και η στην απ' ευθείας LDA:

$$a \leftarrow a - \rho_t \tilde{a}(\gamma_t), \qquad \eta \leftarrow \eta - \rho_t \tilde{\eta}(\lambda)$$

όπου η $\tilde{a}(\gamma_t)$είναι η αντίστροφη Hessian μήτρα επί την κλίση $\nabla_a l(n_t, \gamma_t, \varphi_t, \lambda)$και η $\tilde{\eta}(\lambda)$είναι η αντίστροφη Hessian μήτρα επί την κλίση $\nabla_a L$.

Κεφάλαιο 3 : Προηγούμενες εργασίες

Το παρακάτω κεφάλαιο πραγματεύεται εργασίες ερευνητών ή ερευνητικών ομάδων, οι οποίες μας χρησίμευσαν για την παρούσα διπλωματική. Ειδικότερα, αρχικά αναφερόμαστε στον βασικό κώδικα για ρομπότ συνομιλίας που εκπαιδεύεται όμοια με το σύστημα μηχανικής μετάφρασης. Στη συνέχεια, περιγράφουμε τη διαδικασία προσθήκης θεματικής ενημέρωσης στο σύστημα μετάφρασης, από την οποία εμπνευστήκαμε την προσθήκη πληροφόρησης στο ρομπότ συνομιλίας. Τέλος, δεδομένου ότι οι διάφορες παράμετροι της LDA προσφέρουν αρκετή ελευθερία κατά την εφαρμογή της στο σύνολο κειμένων, αναφερόμαστε στα πορίσματα της θεωρητικής και πειραματικής μελέτης των παραγόντων αυτών. Τα τελευταία έθεσαν κάποιους περιορισμούς στην πλεγματική αναζήτηση που πραγματοποιήσαμε προκειμένου να βρούμε τις κατάλληλες παραμέτρους για την θεματική μοντελοποίηση των σχολίων μας, ενώ καθόρισαν και τον τρόπο χειρισμού του συνόλου δεδομένων των σχολίων ώστε το μοντέλο να επιτύχει.

3.1 Ανάπτυξη ρομπότ συνομιλίας με τη λογική της μηχανικής μετάφρασης

Η εργασία, από άποψη κώδικα, στηρίχθηκε στον κώδικα κατασκευής συστήματος μηχανικής μετάφρασης, στον οποίο είχαν ήδη γίνει κάποιες προσθήκες από ερευνητές προκειμένου να μπορεί να λειτουργήσει για δημιουργία ρομπότ συνομιλίας. Ο κώδικας του συστήματος νευρωνικής μηχανικής μετάφρασης που χρησιμοποιήσαμε είναι των Thang Luong, Eugene Brevdo, Rui Zhao (*https://github.com/tensorflow/nmt*). Πάνω σε αυτόν έχουν βασιστεί και οι Daniel Kukiela και Harrison Kinsley οι οποίοι έχουν προσθέσει στο σύστημα την δυνατότητα παραγωγής λεξιλογίου με τμήματα λέξεων και έχουν προσαρμόσει τον κώδικα ώστε να χρησιμοποιείται για δημιουργία ρομπότ συνομιλίας (*https://github.com/daniel-kukiela/nmt-chatbot*) .

3.2 Θεματικά ενήμερη μηχανική μετάφραση

Η ιδέα να δώσουμε "πληροφορία" στο ρομπότ συνομιλίας σχετικά με την θεματολογία της πρότασης του συνομιλητή του βασίστηκε στην εργασία των (Zhang κ.α. 2016) . Εκείνοι, κάνοντας το σύστημα μηχανικής μετάφρασης θεματικά ενήμερο, διαπίστωσαν ότι παρέχοντας την θεματική πληροφορία της πρότασης εισόδου, και των ήδη μεταφρασμένων λέξεων στον αποκωδικοποιητή, μπορεί να διατηρηθεί η ίδια θεματολογία στη διάρκεια της αποκωδικοποίησης και κατά συνέπεια να παραχθούν καλύτερες μεταφράσεις. Συγκεκριμένα, σαν παράδειγμα μπορούμε να δούμε την πολυσημία της λέξης "ουρά", η οποία μπορεί να αναφέρεται σε ένα ζώο, σε άτομα μπροστά σε ένα ταμείο, ή και σε δομή δεδομένων. Όταν, λοιπόν, μία λέξη εισόδου, έχει παραπάνω από μία έννοιες, η γενικότερη θεματολογία της πρότασης είναι που θα καθορίσει την μετάφρασή της.

Στα θεματικά μοντέλα, οι κατανομές των λέξεων ανά θεματολογία μπορούν να θεωρηθούν σαν διάνυσμα. Στην νευρωνική μηχανική μετάφραση, επίσης, οι λέξεις αντιστοιχίζονται σε αναπαράστασεις του διανυσματικού χώρου. Προκύπτει, λοιπόν, με τρόπο φυσικό η από κοινού χρήση τους. Οι λεκτικές εκφράσεις στο νευρωνικό μοντέλο παράγονται από τις όμοιες στο περιεχόμενο λέξεις του, ενώ τα θεματικά διανύσματα αναπαριστούν την θεματική πληροφορία σε επίπεδο εγγράφου. Για τον λόγο αυτό, βλέπουμε τις δύο αναπαραστάσεις ως συμπληρωματικές μεταξύ τους. Η σκέψη, λοιπόν, είναι ότι η ενσωμάτωση της θεματικής πληροφορίας θα ωφελήσει την ποιότητα της μετάφρασης γι' αυτό και την επιχειρούν είτε στην πρόταση εισόδου, είτε στην μετάφραση στόχο, είτε και στις δύο.

Ο κωδικοποιητής στην συνήθη μηχανική μετάφραση χρησιμοποιεί μόνο τα διανύσματα λεκτικής ενσωμάτωσης για να υπολογίσει το διάνυσμα περιεχομένου της πρότασης εισόδου. Με τη χρήση θεματικής πληροφόρησης στην πλευρά της εισόδου, ο αποκωδικοποιητής μπορεί να έχει μια σφαιρική εικόνα των θεμάτων εισόδου κατά την αποκωδικοποίηση. Επιπλέον, το μοντέλο προσοχής μπορεί αθόρυβα να δώσει προσοχή στην θεματική κατανομή καθεμιάς από τις λέξεις εισόδου. Έτσι, αρχικά υπολογίζονται οι θεματικές κατανομές (κατανομές των λέξεων στα διάφορα θέματα) για κάθε λέξη της πρότασης εισόδου. Στη συνέχεια, οι θεματικές αυτές κατανομές συνδυάζονται με την κρυφή κατάσταση καθεμιάς από τις λέξεις εισόδου. Τέλος, από τον συνδυασμό αυτόν προκύπτει το θεματικά ενήμερο διάνυσμα περιεχομένου της πρότασης εισόδου, ως εξής:

$$topic\_c_j = \sum_{i=1}^{m} \alpha_{ij} [h_i, \beta_i^S]$$

όπου β^S_i είναι οι κατανομή στα διάφορα θέματα της i-οστής λέξης της πρότασης εισόδου. Έτσι, η κρυφή κατάσταση για την j-οστή λέξη της μετάφρασης, από την (2.3.2.1), γίνεται:

$$h_j = g(t_{j-1}, h_{j-1}, topic\_c_j)$$

προκειμένου να έχουμε θεματικά ενήμερο μοντέλο σχετικά με την πρόταση εισόδου.

Ενώ, λοιπόν, η θεματική πληροφορία της πρότασης – πηγής ενισχύει την ακρίβεια των μεταφράσεων, η προσθήκη της θεματικής πληροφορίας των ήδη μεταφρασμένων λέξεων συμβάλλει στη συνοχή μεταξύ των λέξεων της πρότασης εξόδου. Μια φυσική επιλογή για να εξασφαλίσουμε τη θεματική αυτή συνοχή κατά την αποκωδικοποίηση είναι η χρήση ίδιας αρχιτεκτονικής με την ήδη υπάρχουσα αρχιτεκτονική του αποκωδικοποιητή για να λάβουμε το θεματικό κρυφό επίπεδο για την αντίστοιχη λέξη της μεταφρασμένης πρότασης. Έτσι στην προηγούμενη σχέση θα προσθέσουμε ως είσοδο και το θεματικό κρυφό επίπεδο h^{β}_{j-1} που έχει η μεταφρασμένη πρόταση ως και την λέξη της στη θέση (j-1):

$$h_j = g\left(t_{j-1}, h_{j-1}, c, h_{j-1}^{\beta^T}\right)$$

Συνεπώς, ο αποκωδικοποιητής μπορεί να χρησιμοποιήσει την θεματική γνώση που του παρέχεται από τις προηγούμενες μεταφρασμένες λέξεις προκειμένου να αυξήσει την πιθανότητα να επιλέξει λέξεις ίδιου θέματος.

Φυσικά, ο θεματικά πληροφορημένος κωδικοποιητής μπορεί να συνδυαστεί με τον θεματικά πληροφορημένο αποκωδικοποιητή και έτσι να έχουμε την εξής σχέση για το συνολικά θεματικά ενήμερο μεταφραστικό σύστημα:

$$h_j = g\left(t_{j-1}, h_{j-1}, topic\_c_j, h_{j-1}^{\beta^T}\right)$$

Σχετικά με τον τρόπο που μοντελοποιούνται τα θέματα στην μηχανική μετάφραση, οι Zhang κ.α. επέλεξαν να χρησιμοποιήσουν λανθάνουσα αντιστοιχία Dirichlet για την οποία καθόρισαν εξ' αρχής τον αριθμό των θεμάτων προς εκπαίδευση, βασισμένοι στο σύνολο εκπαίδευσης - εισόδου ή μεταφράσεων.

3.3 Περιοριστικοί παράγοντες θεματικής μοντελοποίησης

Μία ακόμα χρήσιμη για εμάς εργασία είναι εκείνη των (Tang κ.α. 2014), οι οποίοι βασισμένοι

σε ένα γεωμετρικό μοντέλο προσομοίωσης της λανθάνουσας αντιστοιχίας Dirichlet που οι ίδιοι πρότειναν, μελέτησαν τη σχέση των διαφόρων παραμέτρων με την επιτυχία της μοντελοποίησης. Τα συμπεράσματα της γεωμετρικής τους προσέγγισης επιβεβαιώθηκαν από την εφαρμογή σε έγγραφα προερχόμενα από Wikipedia, New York Times και Twitter. Ειδικότερα, οι παράγοντες οι οποίοι μελετήθηκαν είναι: το πλήθος των εγγράφων, το μήκος του καθενός από αυτά, το πλήθος των θεμάτων και οι υπερπαράμετροι Dirichlet α και β. Αξίζει να επισημάνουμε ότι οι παράγοντες αυτοί οφείλουν όλοι να πληρούν τις απαραίτητες προϋποθέσεις, καθώς η απόδοση της μοντελοποίησης περιορίζεται από την λιγότερο κατάλληλη παράμετρο κατ' αναλογία με το νόμο ελαχίστου Liebeg, όπου η κοντύτερη σανίδα του βαρελιού καθορίζει και το ύψος της στάθμης του.

Το πλήθος των εγγράφων βρέθηκε να παίζει μάλλον τον σημαντικότερο ρόλο, καθώς είναι θεωρητικά αδύνατο να εγγυηθούμε ταύτιση θεμάτων από έναν μικρό αριθμό εγγράφων, όση έκταση κι αν έχουν. Εφόσον τα έγγραφα αρκούν, η περαιτέρω αύξηση του αριθμού τους μπορεί να μην προκαλέσει αισθητή βελτίωση στην απόδοση, εκτός εάν κατ' αντιστοιχία αυξηθεί και το μέγεθος των εγγράφων. Στην πράξη, η λανθάνουσα αντιστοιχία Dirichlet επιτυγχάνει συγκρίσιμα αποτελέσματα μεταξύ μιας μεγάλης συλλογής και ενός δείγματος χιλιάδων κειμένων από αυτήν.

Το μήκος των εγγράφων παίζει, επίσης, καθοριστικό ρόλο: όταν τα έγγραφα είναι μικρά, ακόμα κι αν είναι πολλά, η λανθάνουσα αντιστοιχία Dirichlet δεν αναμένεται να επιτύχει. Εφόσον τα έγγραφα έχουν επαρκές μέγεθος, περεταίρω αύξησή του, δεν συντελεί σε αισθητή βελτίωση και έτσι πολύ μεγάλα έγγραφα έχουν συγκρίσιμα αποτελέσματα με επαρκή σε μέγεθος τμήματά τους.

Το πλήθος των θεμάτων που θεωρούμε σε σχέση με το πραγματικό πλήθος των θεμάτων του σώματος κειμένου μας, αποτελεί, επίσης, ένα σπουδαίο ζήτημα. Γενικά, το μοντέλο λειτουργεί ικανοποιητικά για μεγαλύτερο αριθμό θεμάτων από τον πραγματικό . Αυτός ο αριθμος, ωστόσο, οφείλει να μην είναι πολύ μεγαλύτερος από τον πραγματικό διότι τότε το μοντέλο παύει να είναι αποτελεσματικό.

Αναφορικά με τις υπερπαραμέτρους της λανθάνουσας αντιστοιχίας Dirichlet, εάν θεωρείται ότι κάθε έγγραφο άπτεται λίγων θεμάτων, η παράμετρος α των θεματικών κατανομών ανά έγγραφο πρέπει να τεθεί σε μικρή τιμή (πχ. 0.1). Αν είναι γνωστό ότι τα θέματα είναι λεκτικά αραιά, δηλ. δεν χρησιμοποιούν σε μεγάλο βαθμό κοινές λέξεις, η παράμετρος κατανομής των λέξεων ανά θέμα τίθεται σε μικρή τιμή (πχ. 0.01) και στην περίπτωση αυτή η εκπαίδευση είναι αποτελεσματική. Μεγάλα β σημαίνει ότι οι λέξεις του λεξιλογίου διαχέονται σε πολλά θέματα με τα τελευταία να είναι όμοια, κάτι που καθιστά την εκπαίδευση δύσκολη.

Κεφάλαιο 4 : Θεματική ανάλυση δεδομένων και τροποποίηση συστήματος

Στο παρακάτω κεφάλαιο αναλύουμε τον τρόπο με τον οποίο επεξεργαστήκαμε τα δεδομένα από το reddit ώστε να μπορέσουμε στη συνέχεια να προχωρήσουμε σε εκπαίδευση του ρομπότ συνομιλίας καθώς και σε θεματική μοντελοποίηση με λανθάνουσα αντιστοιχία Dirichlet. Παραθέτουμε τη διαδικασία άντλησης της θεματικής πληροφορίας από τα δεδομένα μας με δύο τρόπους, τόσο με Hierarchical Agglomerative Clustering (HAC) όσο και με Latent Dirichlet Allocation (LDA) . Τέλος, εξηγούμε τον τρόπο με τον οποίο η θεματική πληροφορία που προκύπτει από την μοντελοποίηση προστέθηκε στο αρχικό σύστημα.

4.1 Προεπεξεργασία δεδομένων

Στη συνέχεια, περιγράφουμε τη διαδικασία που ακολουθήσαμε από την άντληση των σχολίων ως την κατασκευή των αρχείων για την εκπαίδευση, την ανάπτυξη και τον έλεγχο του ρομπότ συνομιλίας αλλά και εκείνων που χρησιμοποιήσαμε για την εφαρμογή της λανθάνουσας αντιστοιχίας Dirichlet.

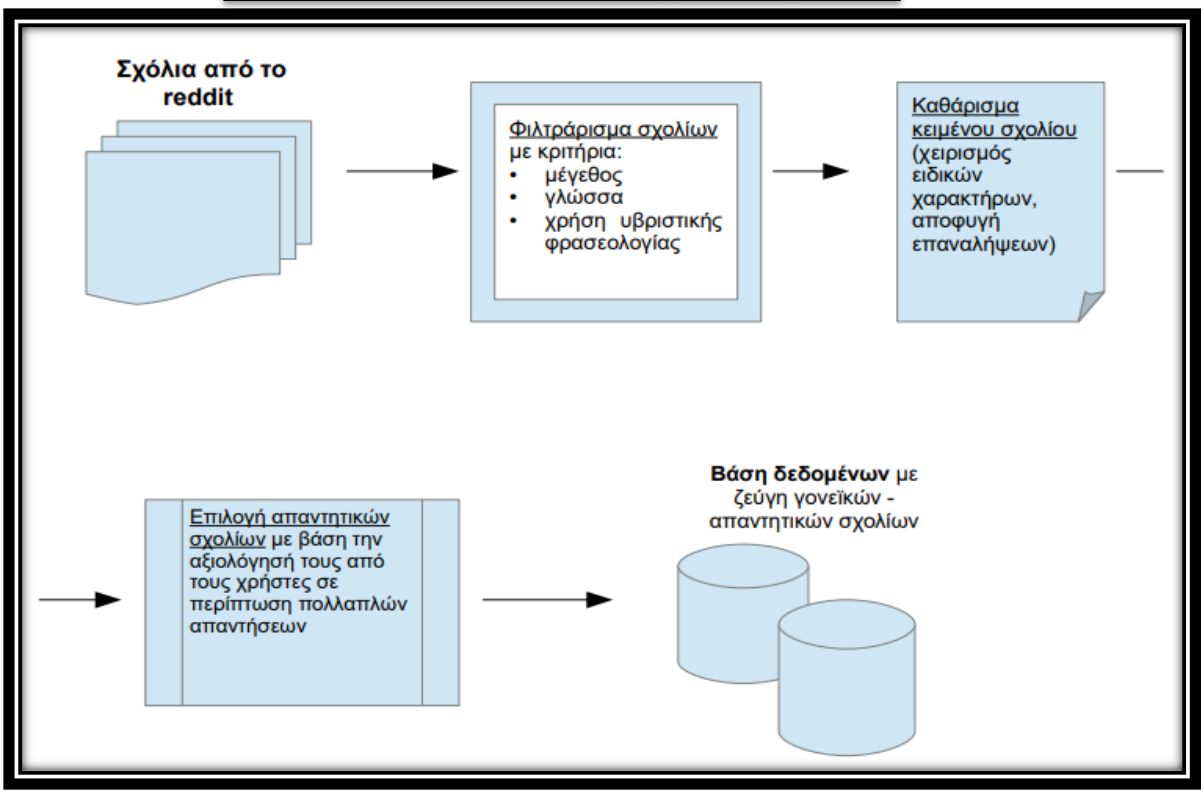
4.1.1 Άντληση δεδομένων, φιλτράρισμα και κατασκευή βάσης δεδομένων

Τα αρχεία των σχολίων τα οποία χρησιμοποιήσαμε ήταν διαθέσιμα στο https://files.pushshift.io/reddit/comments/ , σε μορφή JSON. Αποτελούν το σύνολο των σχολίων που έγιναν από χρήστες κατά τους μήνες του Σεπτεμβρίου και Οκτωβρίου του 2018. Το κάθε αντικείμενο JSON διαθέτει από 21 κλειδιά από τα οποία εμείς χρησιμοποιήσαμε το σώμα του κειμένου, τον κωδικό του, τη θεματική κατηγορία στην οποία ανήκει (subreddit) , την αξιολόγησή του από τους διάφορους χρήστες και τον κωδικό του σχολίου στο οποίο απαντά το σχόλιο.

Προκειμένου να είναι εύκολη η επεξεργασία των σχολίων, κατασκευάσαμε βάση δεδομένων που περιλαμβάνει τα γονεϊκά σχόλια και τις απαντήσεις τους. Κρατήσαμε μόνο τα γονεϊκά σχόλια που έχουν απάντηση και για αυτά που έχουν περισσότερες από μία απαντήσεις επιλέξαμε εκείνη με τη μεγαλύτερη βαθμολογία από τους αναγνώστες. Διατηρήσαμε μόνο τα σχόλια τα οποία δεν υπερβαίνουν τις 100 λέξεις και είναι στα αγγλικά. Ο περιορισμός στη γλώσσα είναι σημαντικός, αφ' ενός γιατί μια άγνωστη σε εμάς γλώσσα καθιστά τον έλεγχο του ρομπότ συνομιλίας δύσκολο, και αφ' ετέρου διότι περιορίζει τις διαθέσιμες θέσεις για όρους στο λεξιλόγιο. Επίσης, πέρα από τον έλεγχο που γίνεται στο reddit για ακατάλληλο περιεχόμενο, τοποθετήσαμε στη βάση δεδομένων μας σχόλια τα οποία δεν περιλαμβάνουν υβριστικό λεξιλόγιο. Η διαδικασία παραγωγής των ζευγών σχολίων απεικονίζεται σχηματικά στην Εικόνα 6.

Η βάση δεδομένων διαθέτει έναν πίνακα με 9.191.779 ζεύγη σχολίων, τα οποία κατανέμονται σε 28.363 θεματικές κατηγορίες. Ενδεικτικό στιγμιότυπο της παρουσιάζεται στην Εικόνα 7.

Εικόνα 6: Διαδικασία παραγωγής ζευγών σχολίων



Εικόνα 7: Ενδεικτικό τμήμα της βάσης δεδομένων

Table: comment_pairs

	id	parent_body	reply_body	subreddit	parent_length	reply_length
	Filter	Filter	Filter	Filter	Filter	Filter
1	1003093	I knew he's moved around a lo...	At one point I believed he was ...	nfl	44	16
2	4061797	This man must taste delicious.	Eddy Bull he's, edible. newlinec...	BlackPeopleTwitter	5	43
3	3733355	As someone who grew up with...	Whoa really nice answer :)	relationship_advice	84	5
4	4162680	Wait, its possible to fail? I thou...	Yes. The barrier has health for...	MonsterHunter	15	8
5	2796808	You haven't answered my ques...	Eh, he's right though. Flight is ...	EliteDangerous	22	94
6	3230486	Right? I was comparing gamer...	But PlayStation and Sony how	DestinyTheGame	27	5
7	557429	Trump's white... Meaning he is...	He’s orange, remember?	CringeAnarchy	8	4
8	4042400	Creep up right next to them an...	Don't do the creep up. Then yo...	IdiotsInCars	24	26
9	4153335	This warms the cockles of my ...	The reactions were great, his ...	youseeingthisshit	7	7
10	4446766	I also just think (taking LBGT o...	I support. Moblit is best husba...	ShingekiNoKyojin	55	30
11	4323473	Should have kept your fingers ...	Hes doomed, no medical profe...	WTF	11	8
12	3762485	I loved how unexpected it was....	Lol I saw some others post the...	skyrim	9	15
13	4314156	This is so common in Portland,...	My ex-MIL loves in Oregon City...	pics	48	89
14	88586	This but the cat titan.	Fox are like the cats of dogs.	titanfolk	6	7
15	2175824	In the new Jack Ryan series th...	They switched it from the book...	badeconomics	22	47
16	3630181	Wade Miley for wild card. Actu...	Miley has collapse written all o...	baseball	23	82
17	998407	seems like that would be the b...	At least.	The_Donald	12	2

4.1.2 Άντληση θεματικής πληροφορίας από τα δεδομένα μας

Για την προσθήκη θεματικής πληροφορίας στο chatbot θα χρησιμοποιήσουμε το subreddit στο οποίο ανήκει το κάθε σχόλιο. Είναι λογικό να σκεφτούμε, πως για να βγάλουμε κάποιο συμπέρασμα για τη θεματολογία του subreddit πρέπει να έχουμε αρκετά σχόλια που να ανήκουν σε αυτό. Παρατηρήσαμε, ωστόσο, ότι παραπάνω από τα μισά subreddits δεν έχουν ούτε 10 ζεύγη σχολίων, πράγμα που σημαίνει ότι μάλλον θα μας μπέρδευαν στην εξαγωγή συμπεράσματος για την θεματολογία παρά θα μας πρόσφεραν την επιθυμητή πληροφορία. Αυτό μας οδήγησε στο να βάλουμε ένα κατώφλι στα subreddits ώστε να απορρίπτουμε εκείνα με μικρότερο πλήθος σχολίων από αυτό. Ως κατώφλι πιλέξαμε τα 1.000 ζεύγη σχολίων. Τα ζεύγη σχολίων που έγιναν αποδεκτά είναι 7.619.664 κατανεμημένα σε 1.342 subreddits. Χαρακτηριστικά παραθέτουμε τα πιο δημοφιλή:

Πίνακας 2: Subbredits με τα περισσότερα σχόλια

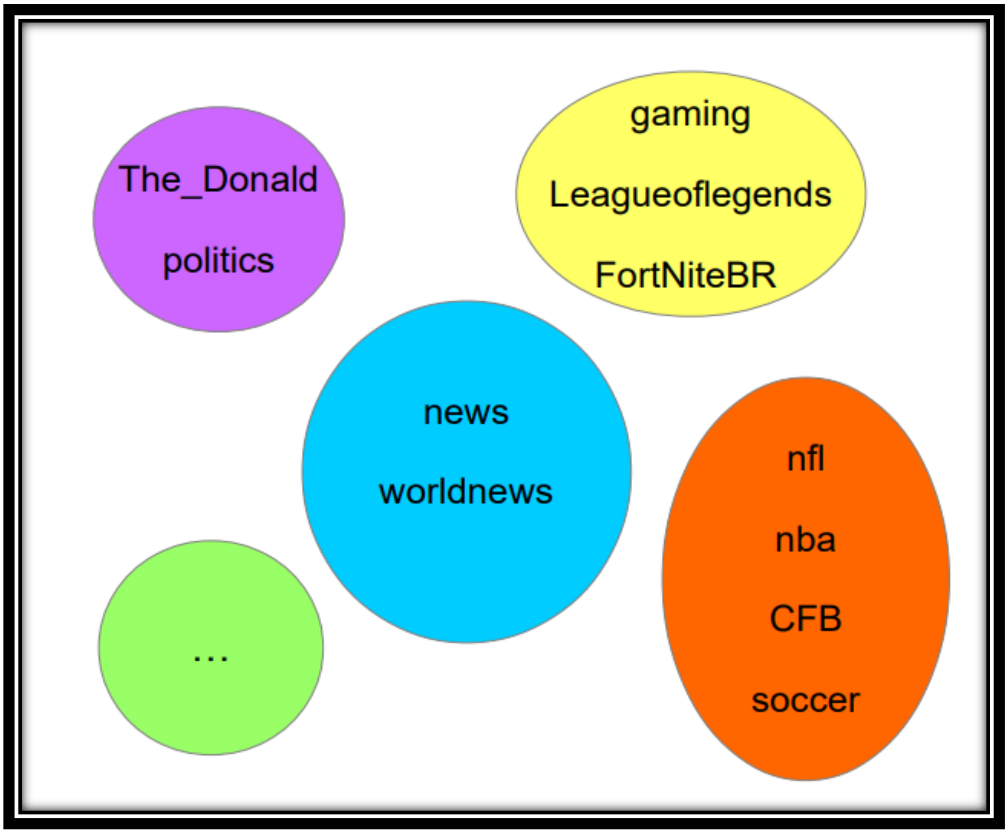
Subreddit	**Πλήθος ζευγών σχολίων**
AskReddit	362.951
politics	208.999
nfl	136.518
The_Donald	118.869
CFB	91.559
nba	88.889
funny	82.936
soccer	76.335
leagueoflegends	72.026
worldnews	70.917
news	66.335
todayilearned	62.490
wow	57.551
gaming	56.880
fantasyfootball	52.742

Δεδομένου ότι έχουμε την πληροφορία για το subreddit στο οποίο ανήκει το κάθε σχόλιό μας, σκοπεύουμε να την διαθέσουμε στο chatbot. Όπως έχει ήδη αναφερθεί, η κάθε λέξη του λεξιλογίου αναπαρίσταται με διάνυσμα ενσωμάτωσης. Στόχος μας είναι να παράξουμε ένα νέο διάνυσμα για κάθε λέξη, ένα "θεματικό διάνυσμα", το οποίο θα παρατεθεί με το διάνυσμα ενσωμάτωσης ώστε να αποτελέσουν την τελική αναπαράσταση της λέξης.

Μια προφανής σκέψη για εξαγωγή διανύσματος για μία λέξη θα ήταν ο υπολογισμός της tf-idf έχοντας ως όρους (terms) τις λέξεις του λεξιλογίου και ως έγγραφα (documents) όλα τα σχόλια που ανήκουν σε ένα subreddit. Έτσι θα προέκυπταν διανύσματα 1.342 θέσεων για κάθε μία από τις 15.000 λέξεις, γεγονός που καθιστά την επιλογή αυτή απαγορευτική λόγω σπατάλης πόρων. Άλλωστε, και μόνο από τα ονόματα των subreddits μπορεί κανείς εύλογα να υποθέσει ότι ορισμένα έχουν όμοια θεματολογία και μπορούν να ενοποιηθούν σε κοινή θεματική κατηγορία (Εικόνα 8). Το subreddit "politics" για παράδειγμα, πιθανότατα συγγενεύει θεματικά με το "The_Donald" ενώ μια

ενιαία ομάδα θα μπορούσαν να αποτελούν τα “soccer”, “CFB” και “nfl”.

Εικόνα 6: Subreddits που υποθέτουμε πως θα μπορούσαν να ενωθούν σε ενιαίες θεματικές ομάδες



Βασισμένοι, λοιπόν, στην υπόθεση ότι με ενοποίηση subreddits μπορούμε να παράγωγουμε διανύσματα μικρότερης διάστασης (κάποιων δεκάδων θέσεων) αποφασίσαμε να το επιχειρήσουμε με δύο διαφορετικές μεθόδους μη επιβλεπόμενης μάθησης: Hierarchical Agglomerative Clustering (HAC) και Latent Dirichlet Allocation (LDA).

4.1.3 Μορφοποίηση των σχολίων με tokenization

Το λεξιλόγιο που το σύστημά μας καλείται να κατασκευάσει από τα δεδομένα, έχει μέγεθος 15.000 όρων, οι οποίοι είναι οι συχνότερες λέξεις ή τμήματά τους (κωδικοποίηση BPE). Ο καθένας από αυτούς τους όρους αναπαρίσταται με ένα διάνυσμα 512 θέσεων, και παράγεται κατά την εκπαίδευση του συστήματος.

Αρχικά, εφόσον στόχος μας είναι να παράγουμε διανύσματα για κάθε token του λεξιλογίου οφείλαμε πριν ακόμα χρησιμοποιήσουμε τα δεδομένα για επεξεργασία να εφαρμόσουμε σε αυτά tokenization. Κάποια ενδεικτικά tokens φαίνονται παρακάτω.

Πίνακας 3: Ενδεικτικά tokens λεξιλογίου

_was _be _on	_I’ve ans _interesti	es se _based	_like ! _as	_Some _fl _differenc	_teams _takes _~	_wonder _doesn’t _god	_running _add per

_with	ng	_similar	-	e	_situation	_can’t	ap
_have	_become	_happens	_can	_His	_phone	_future	_early
_but	_non	_cut	_'	ee	_given	_main	_Most
_they	_dead	_totally	_my	_public	_die	nd	_thread
0	_God	_u	_so	ings	_sound	_control	_account
_not	_shot	_whatever	_at	ir	_history	_sorry	_sad
_are	_three	_St	_if	ized	_worst	_X	_news

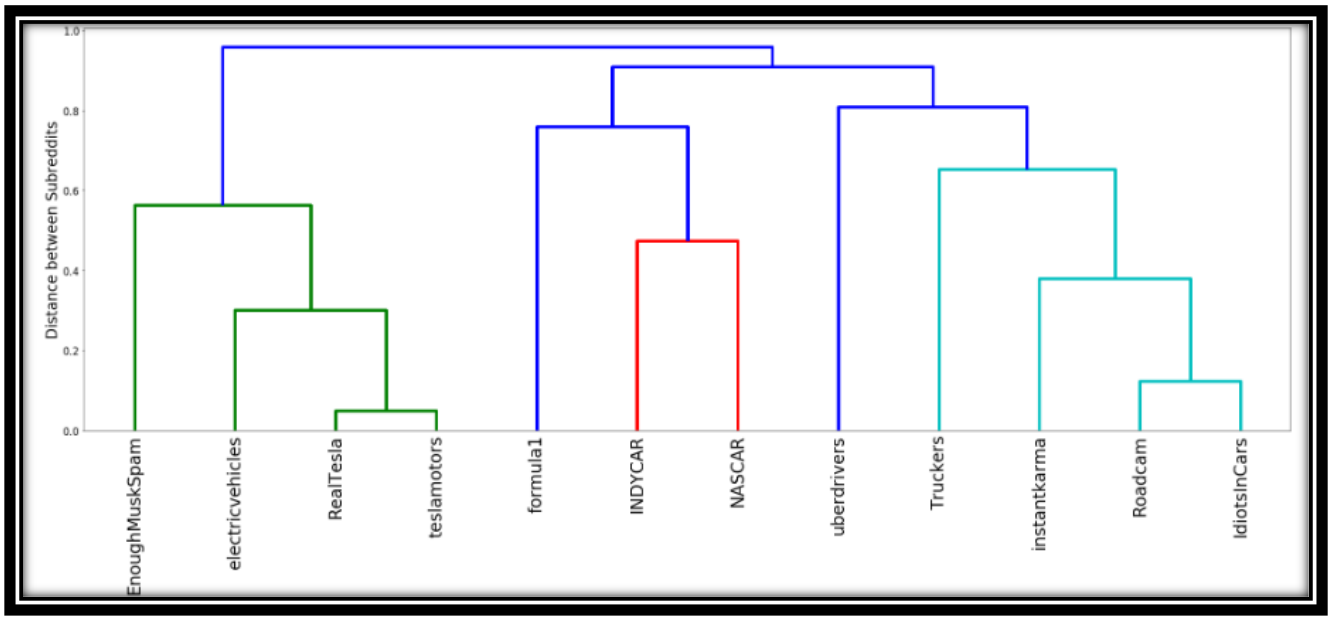
4.2 Κατασκευή θεματικών διανυσμάτων με Hierarchical Agglomerative Clustering (HAC)

Όπως αναφέρθηκε παραπάνω, αυτό που ήταν απαγορευτικό στην παραγωγή tfidf διανυσμάτων για κάθε token με βάση τα subreddits ήταν το πλήθος τους. Στοχεύουμε, λοιπόν, με τη βοήθεια του HAC να ενοποιήσουμε τα 1.342 subreddits σε 40 ομάδες, με το σύνολο των σχολίων της καθεμιάς να αποτελεί ένα νέο document. Με τη βοήθεια αυτών, θα υπολογίσουμε έπειτα τα νέα tfidf διανύσματα 40 θέσεων για κάθε token-term.

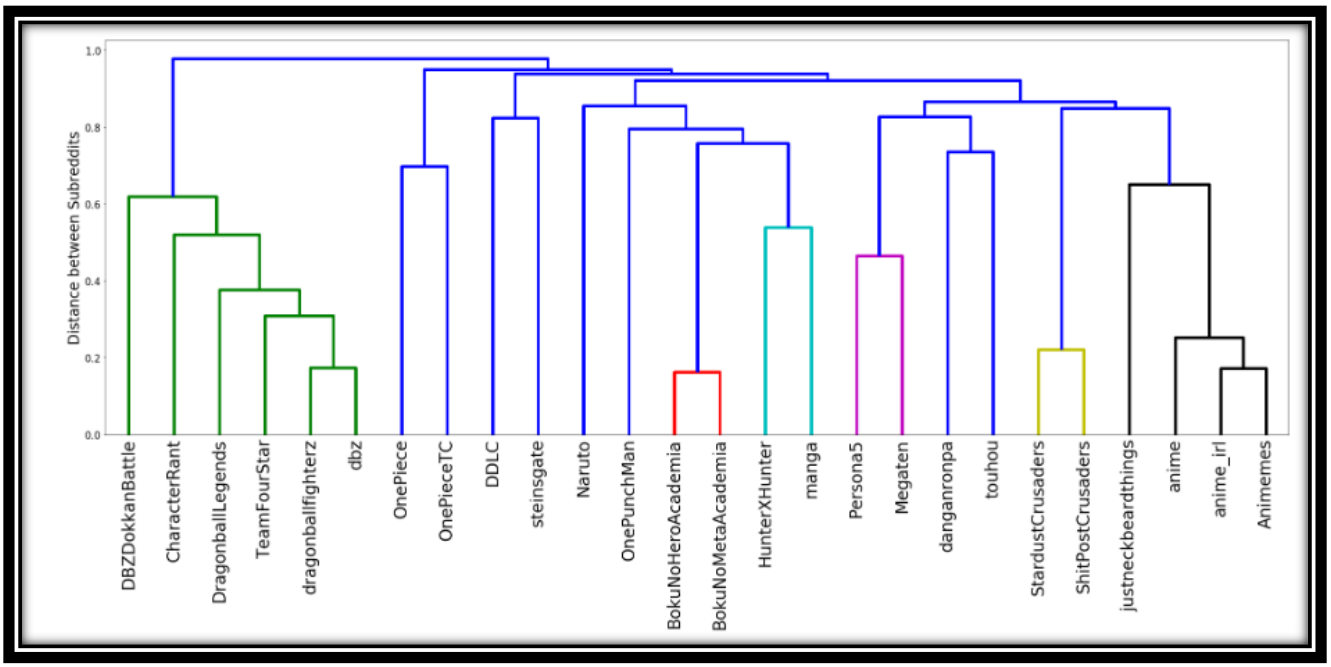
Για να γίνει η συσταδοποίηση των subreddits οφείλουμε να έχουμε ένα διάνυσμα για το κάθε subreddit. Δεδομένου ότι έχουμε τον πίνακα term-document (token-subreddit) διαπιστώνουμε ότι κάθε στήλη του αποτελεί ένα επιθυμητό διάνυσμα που αντιπροσωπεύει το αντίστοιχο subreddit το οποίο έχει 15.000 συντεταγμένες (features) που αντιστοιχούν στα 15.000 tokens του λεξιλογίου.

Η μέθοδος του HAC, λοιπόν, βλέπει αρχικά τα 1.342 subreddits ως 1.342 clusters με 15.000 χαρακτηριστικά (features) και αρχίζει και συνενώνει ζεύγη clusters. Υπολογίζει ανά δύο clusters την συνημιτονοειδή ομοιότητα των features τους, την οποία θεωρεί ως μέτρο ανομοιότητάς τους. Οι δύο clusters που έχουν την μικρότερη ανομοιότητα, συνενώνονται σε μία ομάδα. Έτσι, οι clusters είναι τώρα 1.341. Η διαδικασία επαναλαμβάνεται μέχρι οι clusters να μειωθούν στις 40. Αξίζει, επίσης, να σημειωθεί ότι ως ανομοιότητα μεταξύ δύο ομάδων subreddits παίρνουμε την μέγιστη ανομοιότητα μεταξύ των subreddits που τις αποτελούν. Στις Εικόνες 9, 10, 11 και 12 παρατίθενται χαρακτηριστικά τα δενδρογράμματα κάποιων clusters. Καταλήγουμε με 40 Clusters καθένα από τα οποία περιέχει τα 15.000 tokens ορισμένες φορές. Κατασκευάζουμε, λοιπόν, τα tfidf vectors για κάθε λέξη-token, τα οποία έχουν μήκος 40 συντεταγμένων (αντίστοιχων των 40 clusters – εγγράφων).Η διαδικασία παραγωγής θεματικών διανυσμάτων μέσω HAC απεικονίζεται στην Εικόνα 13.

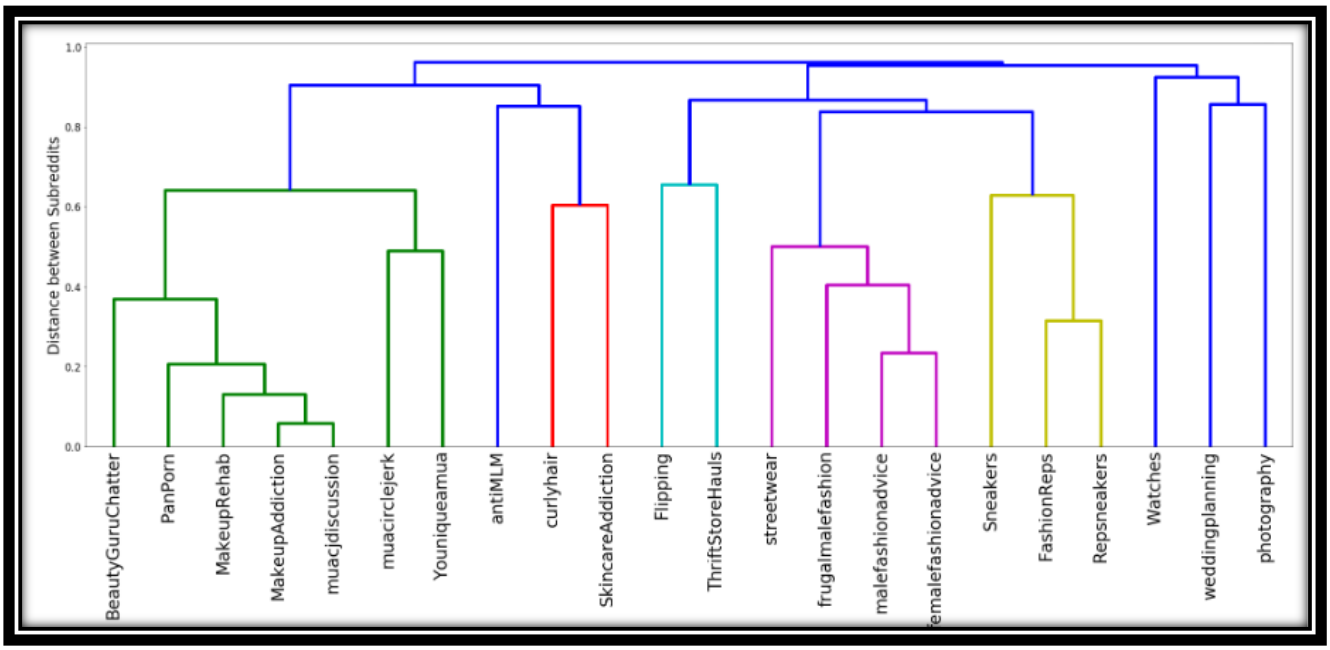
Εικόνα 7: Δενδρόγραμμα Cluster σχετικού με τα αυτοκίνητα



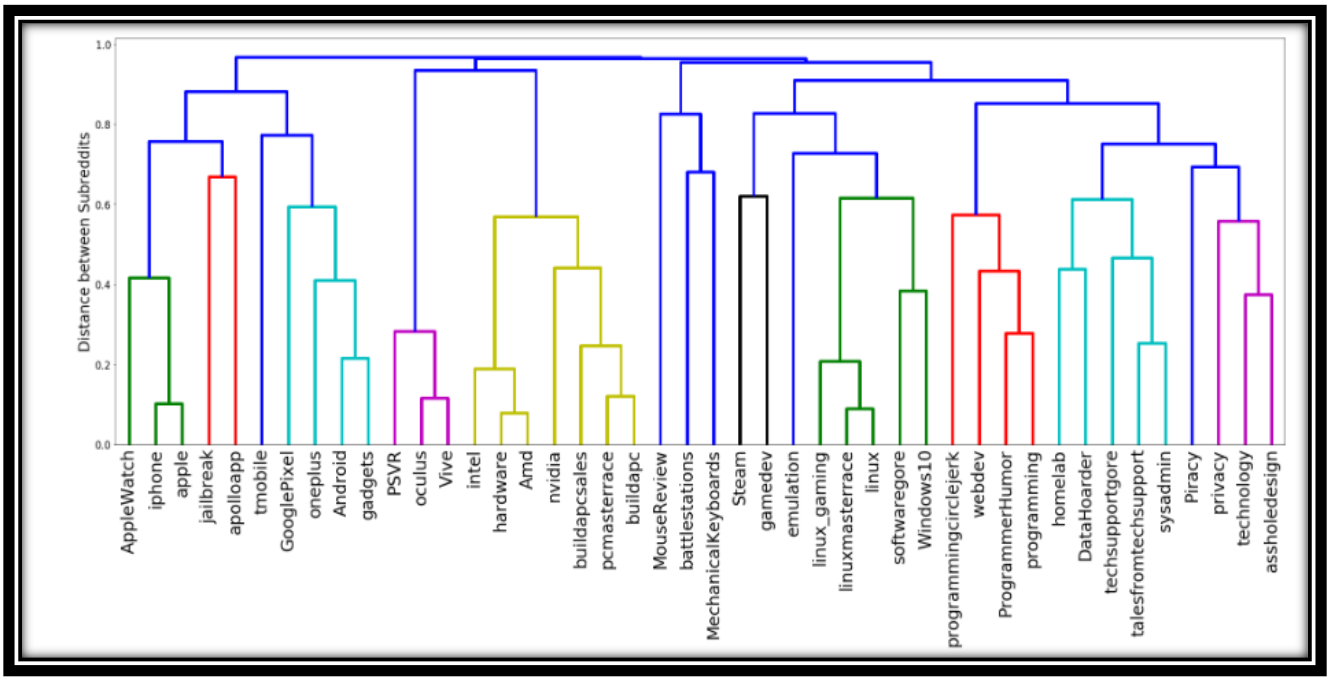
Εικόνα 8: Δενδρόγραμμα Cluster σχετικού με anime



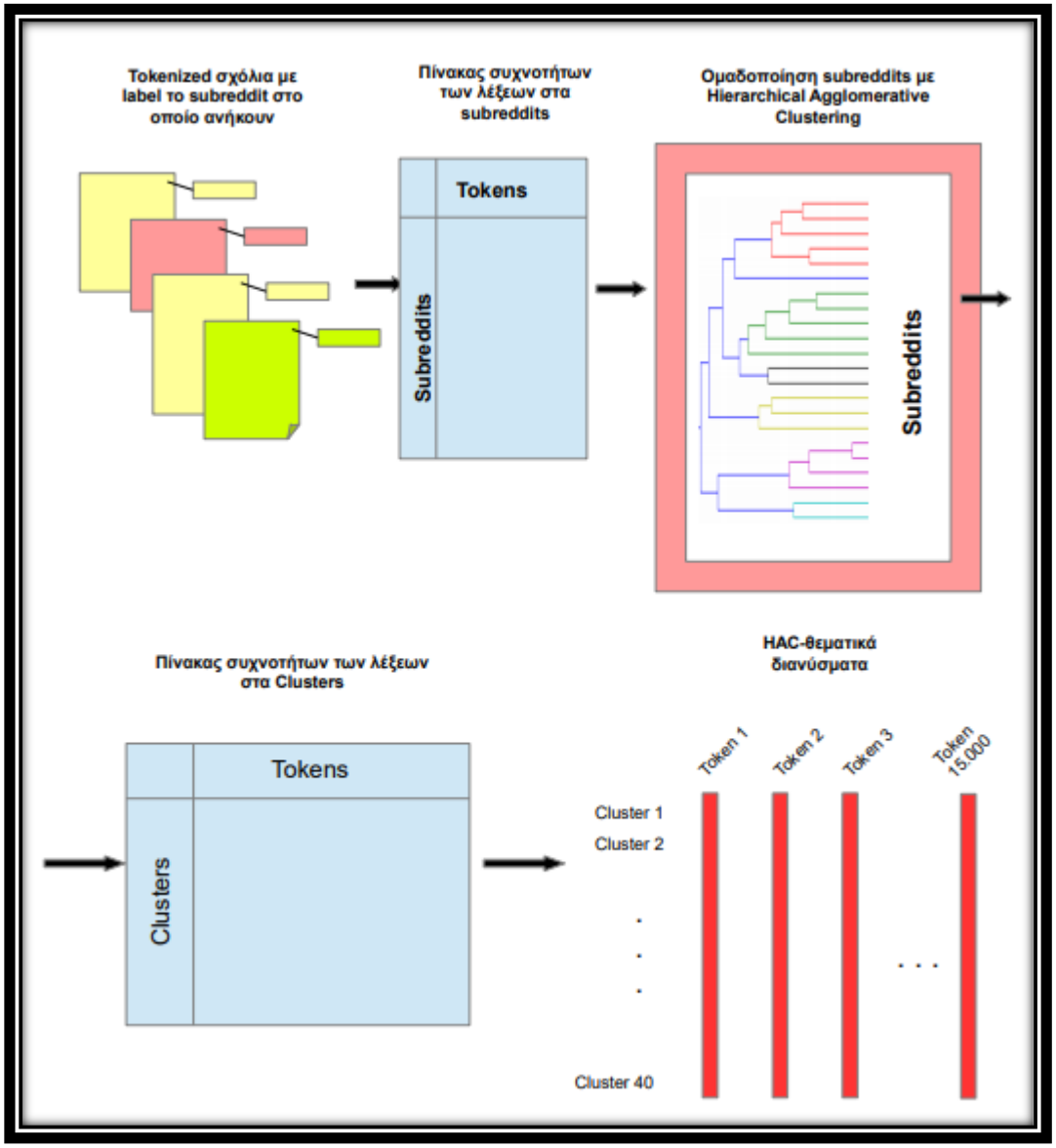
Εικόνα 9: Δενδρόγραμμα Cluster σχετικού με τη μόδα



Εικόνα 10: Δενδρόγραμμα Cluster σχετικού με υπολογιστές



Εικόνα 11: Παραγωγή θεματικών διανυσμάτων με HAC



4.3 Κατασκευή θεματικών διανυσμάτων με Latent Dirichlet Allocation (LDA)

Ο άλλος τρόπος με τον οποίο αποφασίσαμε να κατασκευάσουμε θεματικά διανύσματα ήταν με LDA. Το σύνολο των δεδομένων, όφειλε, επομένως, να έρθει σε κατάλληλη μορφή για την εφαρμογή και αυτής της τεχνικής.

Όπως έχει προαναφερθεί και στην ενότητα 3.3 το μέγεθος των εγγράφων αποτελεί καθοριστικό παράγοντα για την επιτυχία της μοντελοποίησης με LDA. Τα σχόλια σε ένα φόρουμ όπως το reddit υστερούν σ' αυτό το θέμα λόγω της βραχύτητάς τους. Το μέγεθός τους είναι μικρότερο των 100 λέξεων. Για το λόγο αυτό, αποφασίσαμε να κατασκευάσουμε ψευδοέγγραφα ενοποιώντας σχόλια με κριτήριο την κατηγορία στην οποία ανήκουν. Το κάθε έγγραφο αποτελείται από ενοποιημένα σχόλια, γονεϊκά και απαντήσεις, με το μέγεθός τους να κειμένεται μεταξύ 2.000 και 2.200 λέξεων. Κατασκευάσαμε συλλογή 10.000 ψευδοεγγράφων. Το πλήθος ψευδοεγγράφων ανά κατηγορία είναι ανάλογο του συνόλου των λέξεων των σχολίων που ανήκουν σε αυτήν.

Το γεγονός ότι το κάθε ψευδοέγγραφο έχει ληφθεί από ένα μόνο subreddit στηρίζει την εκτίμησή μας ότι άπτεται ενός ή λίγων θεμάτων, τα οποία δεν έχουν πολλές κοινές λέξεις (πέρα από τις συνήθεις ουδέτερες λέξεις οι οποίες παραβλέπονται κατά τη μοντελοποίηση). Πραγματοποιούμε, λοιπόν πλεγματική αναζήτηση, με βάση τα αποτελέσματα της οποίας, οι υπερπαράμετροι της λανθάνουσας αντιστοιχίας τίθενται ίσες με α = 0.7 και η = 0.01.

Κατασκευάσαμε tfidf vectors για το κάθε ψευδοέγγραφο και εφαρμόσαμε LDA για πλήθος θεμάτων ίσο με 40, αριθμός που αποτελεί έναν καλό συνδυασμό απόδοσης και οικονομίας πόρων οπότε και επιλέχθηκε.

Οι πέντε πιο αντιπροσωπευτικές λέξεις της κάθε θεματικής κατηγορίας φαίνονται παρακάτω:

Πίνακας 4: Αντιπροσωπευτικές λέξεις των θεμάτων που προέκυψαν από την LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
_game	_gun	_golf	_accounts	_fight	_$	_boss	_trump
_play	_shoot	_mark	_cases	ana	_pay	_attack	_race
_players	_guns	_beef	_calls	_Khabib	_city	_level	_votes
_map	_bugs	_Ferrari	_dismiss	_fighting	_money	_+	_flag
_PC	_Fallout	_Turn	_punishment	_fights	_car	_War	_Dems

Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
_Pokemon	_Halloween	_Toronto	Vy	_game	_WWE	SU	^
_her	_suit	q	Iny	_games	_match	_wrestling	_^
anda	_horror	_LeBron	_Eminem	]	ack	_neither	_u
elle	_origin	_ND	_tracks	_card	_Warriors	_Lewis	_upvote
_Thanos	_MCU	_finals	Ble	_cards	_Rock	_CF	_memes

Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24
_Star	_car	^	_rep	_wood	_movie	_her	_phone
lo	_cars	~	Zer	_Bojack	_show	_she	_Apple
_Wars	_water	_K	Ore	_Bama	_episode	_women	_X
_Paul	_ride	ak	Oud	_Lakers	_movies	_sex	_app
_star	_Conor	_anime	_Liverpool	_AC	_scene	_She	_tech

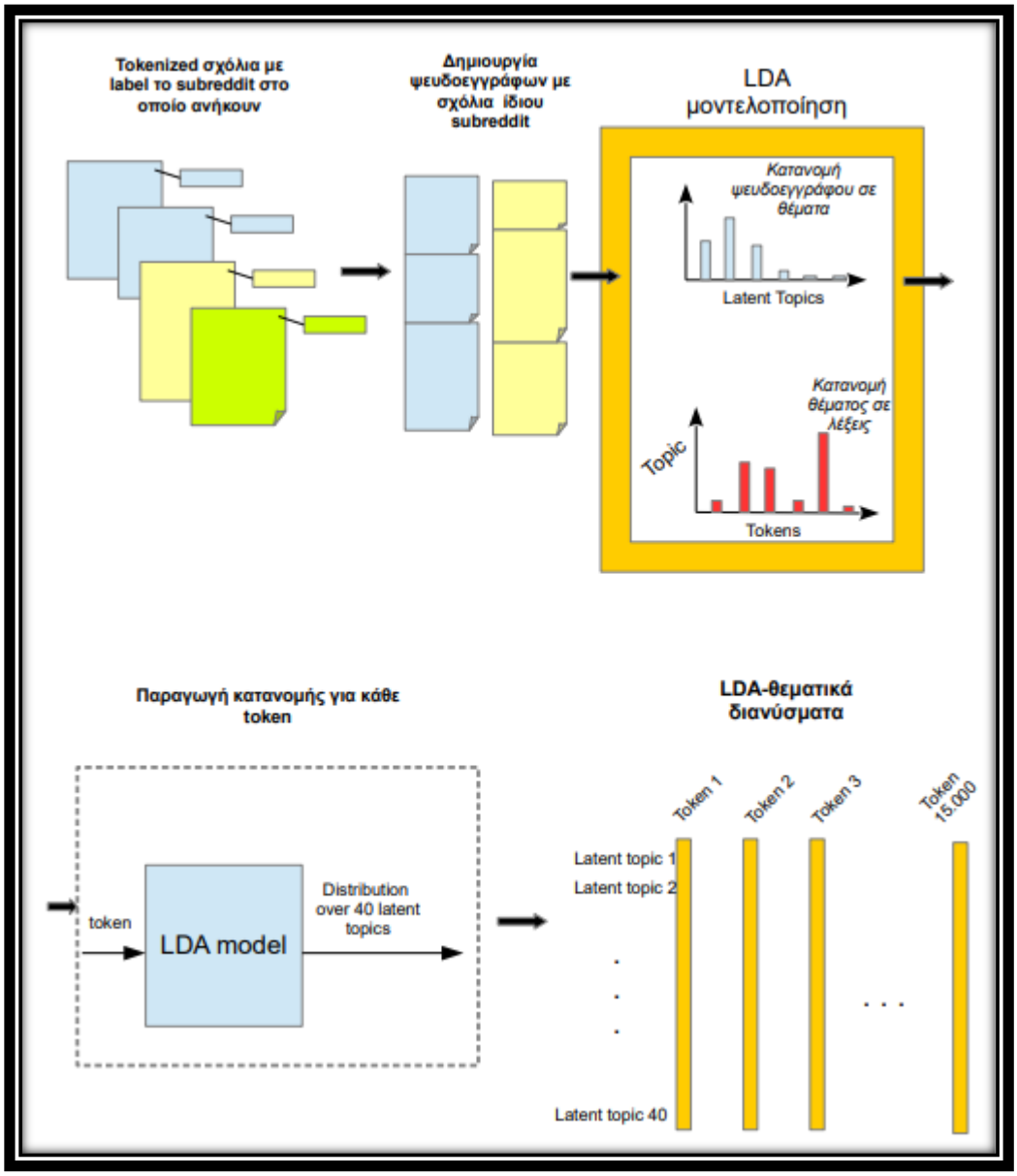
Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30	Topic 31	Topic 32
_THE	_dog	_bull	_ship	_student	_music	_her	_eat
ING	_cat	_puts	_ships	_students	_song	_asks	_food
_meme	_dogs	_negative	_rifle	_class	_album	_furry	_weight
^	_cats	_retirement	_myth	_employer	_she	_fur	_fat
_YOU	_pet	ipp	Iles	_exam	_her	_she	_eating

Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
_school	_Trump	_plane	X	_US	_rap	_team	_church
_OP	_vote	_flight	_mod	_country	_Ice	_season	_white
_drugs	_political	_lift	7	_countries	_KD	_game	_religion
_alcohol	_Kavanaugh	_air	_code	_UK	_Br	_teams	ism
_weed	_party	_lights	_buy	_government	_Kim	_fans	_God

Ως αποτέλεσμα έχουμε ένα μοντέλο κατανομής των ψευδοεγγράφων σε λανθάνοντα θέματα (συνολικά 40) και κατανομής θεμάτων σε λέξεις (πλήθους 15.000). Στο μοντέλο αυτό, λοιπόν, τοποθετούμε σαν εισόδους – έγγραφα τις λέξεις του λεξιλογίου και έχουμε ένα διάνυσμα (40 συντεταγμένων) κατανομής του κάθε token σε 40 λανθάνοντα θέματα.

Ολοκληρωμένα, η διαδικασία παραγωγής θεματικών διανυσμάτων με LDA φαίνεται παρακάτω:

Εικόνα 12: Παραγωγή θεματικών διανυσμάτων με LDA



4.4 Τροποποίηση συστήματος για την προσθήκη της θεματικής πληροφορίας

Έχοντας παράξει θεματικά διανύσματα για κάθε token του λεξιλογίου, αυτό που μένει είναι να τα διοχετεύσουμε στο σύστημα. Αυτό γίνεται πραγματοποιώντας concatenation (συνένωση) με τα embeddings των λέξεων. Δηλαδή, τα tokens πλέον αναπαρίστανται με διανύσματα μεγέθους 512 (αρχικό) + 40 (θεματικό διάνυσμα) = 552 θέσεων.

Το σύστημα στην αρχική του έκδοση σχημάτιζε τις διανυσματικές αναπαραστάσεις των tokens κατά την εκπαίδευση άρα σε κάθε φάση ανανέωσης βαρών του δικτύου τα διανύσματα των λέξεων είχαν τροποποιηθεί. Πλέον, τα διανύσματα πέρα από το μεταβλητό τμήμα τους το οποίο αλλάζει κατά την

εκπαίδευση παράλληλα με τα βάρη του νευρωνικού συστήματος, έχουν και ένα σταθερό κομμάτι (το θεματικό) το οποίο φροντίσαμε να μην τροποποιείται κατά την εκπαίδευση.

Κεφάλαιο 5 : Ζητήματα υλοποίησης

Το παρακάτω κεφάλαιο αναφέρεται σε τεχνικά θέματα της εργασίας μας. Αρχικά, αναφερόμαστε στις επιλογές που πραγματοποιήσαμε στο προγραμματιστικό κομμάτι της επεξεργασίας των δεδομένων πριν την εκπαίδευση, της εξαγωγής θεμάτων και της τροποποίησης του συστήματος. Έπειτα, δίνονται στοιχεία που αφορούν την εκπαίδευση του συστήματος, τόσο στην απλή όσο και στις τροποποιημένες του εκδοχές.

5.1 Γλώσσα προγραμματισμού και βιβλιοθήκες

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε στην εργασία ήταν η Python. Αρχικά, ο κώδικας του συστήματος που είχαμε στην διάθεσή μας ήταν σε Python οπότε η τροποποίησή του με την προσθήκη των θεματικών διανυσμάτων έγινε στη γλώσσα αυτή. Αλλά και η κατασκευή των θεματικών διανυσμάτων έγινε στην ίδια γλώσσα, καθώς είναι εύχρηστη και διαθέτει ποικιλία βιβλιοθηκών για machine learning, επεξεργασία δεδομένων κειμένου και οπτικοποίησης αποτελεσμάτων.

Χαρακτηριστικά packages τα οποία χρησιμοποιήσαμε στα διάφορα στάδια της εργασίας είναι:

- Στο στάδιο της προεπεξεργασίας των δεδομένων χρησιμοποιήσαμε την sqlite3 προκειμένου να αποθηκεύσουμε, να οργανώσουμε τα δεδομένα μας και να αντλήσουμε ποσοτικές πληροφορίες όσον αφορά διάφορα χαρακτηριστικά τους. Επίσης στο στάδιο αυτό για το φιλτράρισμα των σχολίων χρησιμοποιήσαμε και τις langdetect, urlextract.
- Για την άλγεβρα που χρειάστηκε προκειμένου να κατασκευάσουμε τα tfidf διανύσματα των όρων (απαιτούμενο τόσο στην LDA όσο και στο Clustering), χρησιμοποιήθηκε η numpy.
- Η μέθοδος του Hierarhical Agglomerative Clustering υλοποιήθηκε με τη βοήθεια της scipy, ενώ για τη θεματική μοντελοποίηση μέσω LDA αξιοποιήθηκε η scikit-learn.
- Τέλος, οι απεικονίσεις που χρειάστηκαν σε διάφορα στάδια (δενδρογράμματα, αποτελέσματα εκπαίδευσης) έγιναν με τη βοήθεια της matplotlib.

5.2 Λεπτομέρειες εκπαίδευσης

Στη συνέχεια, αναφέρουμε τις λεπτομέρειες των εκπαιδεύσεων που πραγματοποιήσαμε. Για την υλοποίηση του κωδικοποιητή και αποκωδικοποιητή χρησιμοποιούνται από ένα δίκτυο μακρού – βραχέος όρου μνήμης, διπλής κατεύθυνσης, με δύο κρυφά στρώματα αποτελούμενα από 512 μονάδες το καθένα. Εφαρμόζεται ο μηχανισμός προσοχής (παραλλαγή Luong) με κανονικοποιημένη τη συνάρτηση που καθορίζει την συσχέτιση. Η αποκωδικοποίηση είναι δεματική με μέγεθος δέματος ίσο με 20.

Ο ρυθμός εκπαίδευσης είναι 0.001 και εκπαιδεύουμε το σύστημα για 3 εποχές. Τα βάρη του δικτύου ανανεώνονται μαζικά με την εφαρμογή του βελτιστοποιητή adam. Η ανανέωση των βαρών γίνεται μαζικά σε mini batches των 512 ζευγών σχολίων. Χρησιμοποιείται η τεχνική dropout με ποσοστό 0.2. Οι παράγωγοι κανονικοποιούνται ώστε να μην υπερβαίνουν το 5. Το λεξιλόγιο - μεγέθους 15.000 όρων- είναι κοινό για κωδικοποιητή και αποκωδικοποιητή, αφού οι προτάσεις

εισόδου και εξόδου είναι σε κοινή γλώσσα, τα αγγλικά. Το μέγεθος των word embeddings είναι 512 στην αρχική έκδοση και 552 στις τροποποιημένες. Τα σύνολα δεδομένων ανάπτυξης και ελέγχου αποτελούνται από 10.000 ζεύγη σχολίων το καθένα.

Οι εκπαιδεύσεις έγιναν στην GPU και είχαν διάρκεια περίπου 30 ώρες η κάθε μία.

Κεφάλαιο 6: Πειραματικά Αποτελέσματα

Στο παρόν κεφάλαιο δίνεται ποσοτική και ποιοτική ανάλυση των τριών συστημάτων που εκπαιδεύτηκαν, δηλαδή της αρχικής εκδοχής, της HAC εκδοχής καθώς και της LDA εκδοχής. Συγκεκριμένα, τα συστήματα που εκπαιδεύτηκαν αξιολογούνται ποσοτικά καθ' όλη τη διάρκεια της εκπαίδευσής τους, με βάση τις μετρικές της περιπλοκής και του BLEU που έχουν ήδη αναφερθεί. Σ' αυτή την αξιολόγηση, γίνεται εμφανής η ταχύτερη σύγκλιση των θεματικά πληροφορημένων συστημάτων. Στη συνέχεια, γίνεται ποιοτική ανάλυση με σύγκριση δειγμάτων απαντήσεων-εξόδων των τριών εκδοχών του chatbot.

6.1 Ποσοτικά αποτελέσματα

Οι μετρικές, μέσω των οποίων αξιολογούνται ποσοτικά οι απαντήσεις του chatbot είναι η περιπλοκή (perplexity) και το BLEU (bilingual evaluation understudy) , ο μαθηματικός ορισμός των οποίων περιγράφεται στις υποενότητες 2.1.3 και 2.3.5 αντίστοιχα. Οι μετρικές αυτές λαμβάνονται σε ορισμένα στάδια-βήματα της εκπαίδευσης, ώστε να βλέπουμε την απόδοση του συστήματος κατά τη διάρκειά της, ενώ υπολογίζονται τόσο στο σύνολο ανάπτυξης (development set) όσο και στο σύνολο ελέγχου (test set).

6.1.1 Περιπλοκή

Η περιπλοκή, σε κάθε εποχή μετράται ανά 1.000 training steps, καθώς και στο τέλος της εποχής. Στη συνέχεια, παρουσιάζονται τα αποτελέσματα των μετρήσεων τόσο για το σύνολο ανάπτυξης όσο και για το σύνολο ελέγχου καθώς και οι κοινές γραφικές παραστάσεις των τιμών της περιπλοκής της αρχικής έκδοσης και της κάθε τροποποίησης.

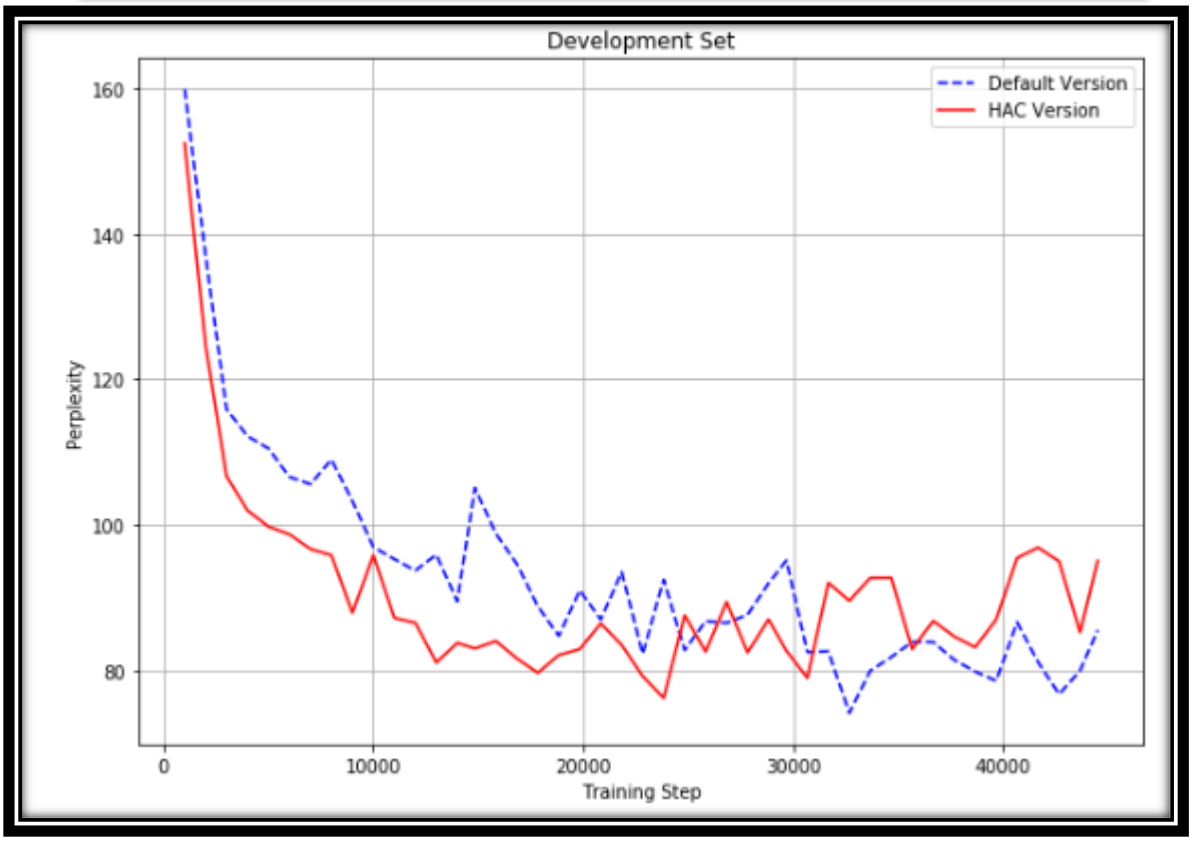
Πίνακας 5: Περιπλοκή συνόλου ανάπτυξης

training step	Perplexity of default chatbot	Perplexity of HAC chatbot	Perplexity of LDA chatbot
1000	159.81	152.38	146.46
2000	137.0	124.7	126.3
3000	115.9	106.7	111.6
4000	112.2	102.0	108.0
5000	110.5	99.78	95.88
6000	106.6	98.72	90.77
7000	105.6	96.71	85.21
8000	109.0	95.85	92.97
9000	103.3	87.96	87.64

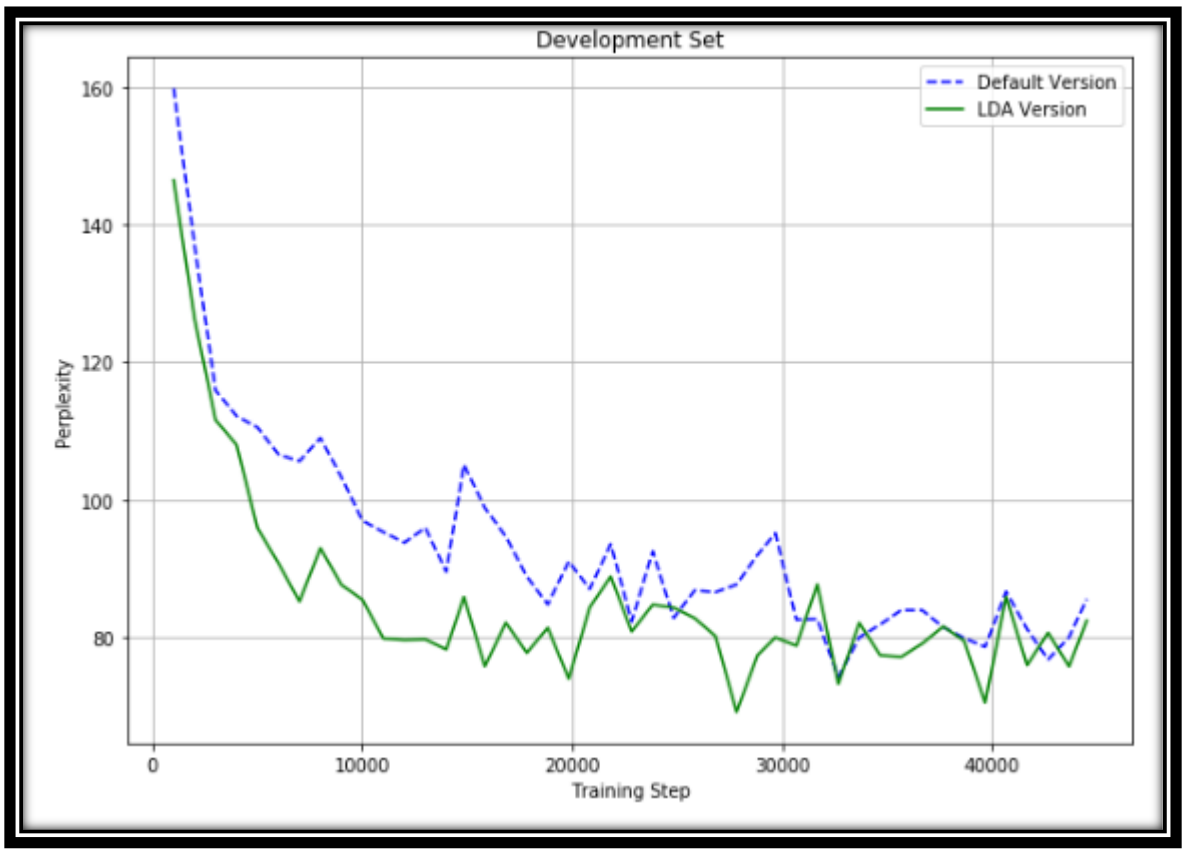
10000	96.97	95.86	85.48
11000	95.31	87.25	79.85
12000	93.76	86.58	79.63
13000	95.94	81.13	79.71
14000	89.53	83.82	78.23
14840	105.1	83.05	85.89
15840	98.83	84.08	75.79
16840	94.66	81.69	82.12
17840	88.84	79.72	77.76
18840	84.78	82.15	81.40
19840	91.05	82.97	73.97
20840	87.02	86.49	84.36
21840	93.56	83.54	88.84
22840	82.32	79.25	80.85
23840	92.50	76.26	84.74
24840	82.82	87.62	84.32
25840	86.82	82.62	82.80
26840	86.56	89.45	80.15
27840	87.69	82.51	69.16
28840	92.00	87.05	77.41
29690	95.13	82.75	79.99
30690	82.57	79.01	78.78
31690	82.67	92.08	87.70
32690	74.21	89.60	73.22
33690	80.00	92.75	82.11
34690	81.88	92.76	77.40
35690	83.95	82.98	77.17
36690	83.94	86.84	79.10
37690	81.49	84.71	81.58
38690	79.89	83.22	79.50
39690	78.61	87.03	70.51
40690	86.65	95.48	85.87
41690	81.18	96.92	75.95
42690	76.78	95.01	80.68

43690	80.00	85.27	75.76
44530	85.58	95.09	82.43

Εικόνα 13: Κοινή γραφική παράσταση της περιπλοκής της αρχικής εκδοχής και της HAC εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ανάπτυξης



Εικόνα 14: Κοινή γραφική παράσταση της περιπλοκής της αρχικής εκδοχής και της LDA εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ανάπτυξης



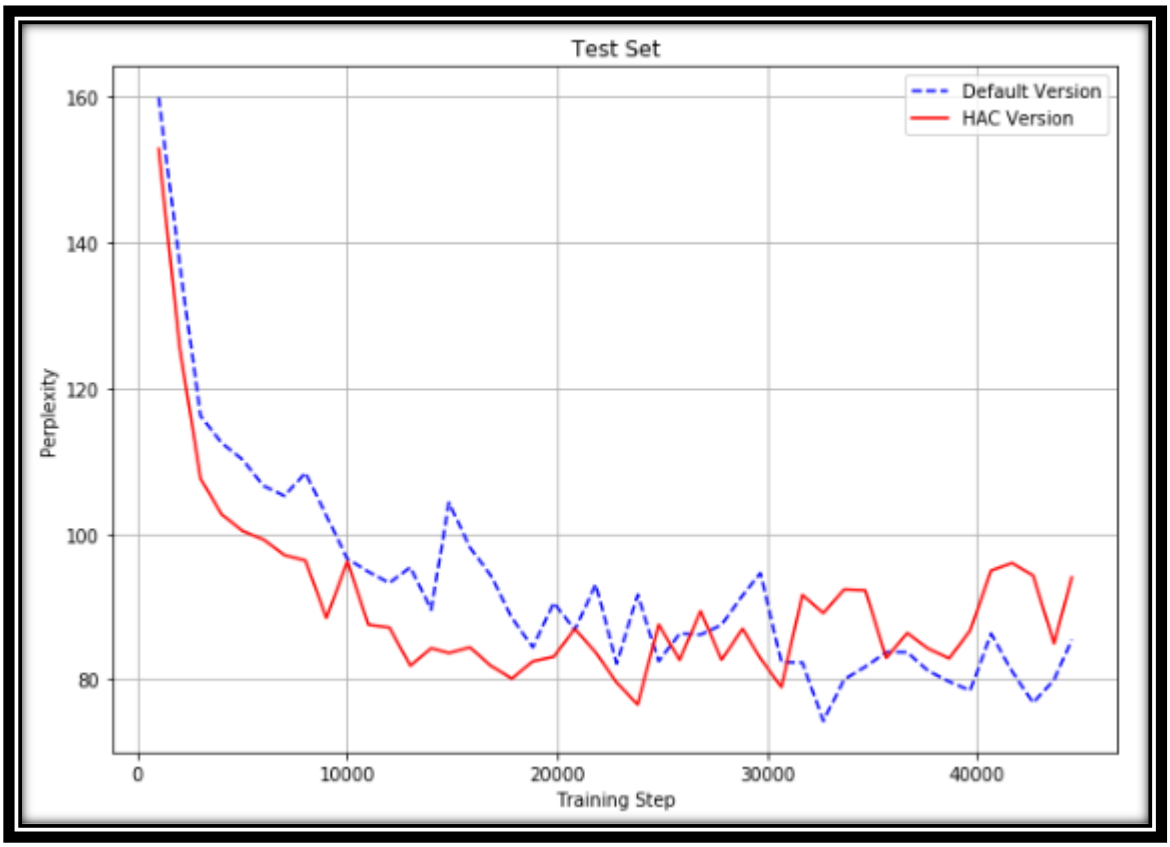
Πίνακας 6: Περιπλοκή συνόλου ελέγχου

training step	Perplexity of default chatbot	Perplexity of HAC chatbot	Perplexity of LDA chatbot
1000	159.98	152.9	147.07
2000	137.2	125.7	127.2
3000	116.21	107.6	112.8
4000	112.5	102.7	109.4
5000	110.2	100.4	97.21
6000	106.6	99.22	91.84
7000	105.2	97.08	86.34
8000	108.4	96.34	93.95
9000	102.5	88.43	89.01
10000	96.55	96.20	86.86

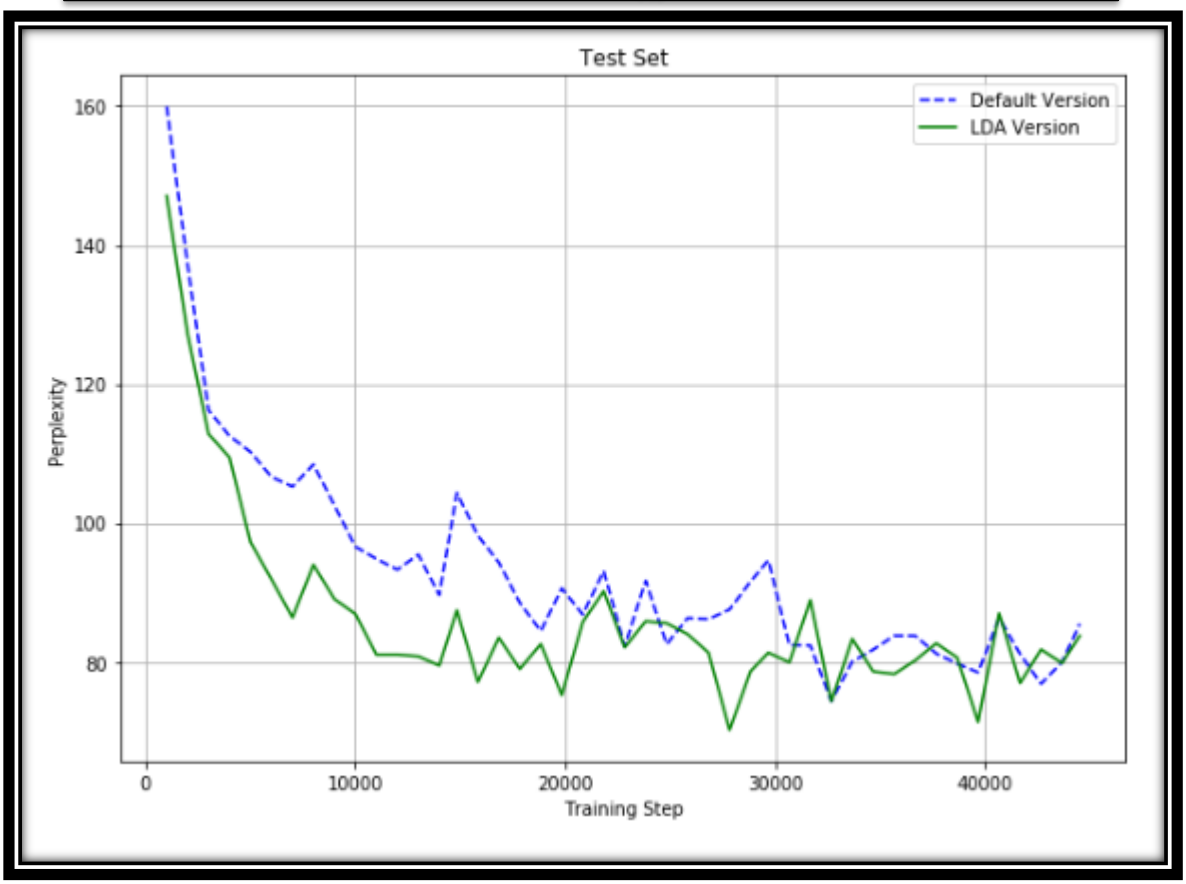
11000	94.80	87.53	81.01
12000	93.24	87.14	81.01
13000	95.42	81.88	80.76
14000	89.58	84.29	79.42
14840	104.3	83.61	87.40
15840	98.15	84.41	77.06
16840	94.28	81.87	83.44
17840	88.46	80.11	78.94
18840	84.40	82.49	82.53
19840	90.56	83.11	75.17
20840	86.77	86.95	85.70
21840	93.06	83.70	90.15
22840	82.19	79.56	82.05
23840	91.64	76.54	85.81
24840	82.49	87.55	85.52
25840	86.28	82.66	83.95
26840	86.13	89.39	81.27
27840	87.51	82.69	70.16
28840	91.53	86.95	78.61
29690	94.60	82.90	81.28
30690	82.38	78.94	79.89
31690	82.33	91.61	88.87
32690	74.29	89.11	74.27
33690	80.01	92.38	83.27
34690	81.73	92.25	78.58
35690	83.73	82.95	78.21
36690	83.72	86.36	80.17
37690	81.16	84.23	82.64
38690	79.71	82.87	80.58
39690	78.45	86.71	71.30
40690	86.27	94.97	86.95
41690	81.09	95.97	76.90
42690	76.79	94.29	81.75
43690	79.93	84.93	79.78

44530	85.43	94.01	83.72

Εικόνα 15: Κοινή γραφική παράσταση της περιπλοκής της αρχικής εκδοχής και της HAC εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ελέγχου



Εικόνα 16: Κοινή γραφική παράσταση της περιπλοκής της αρχικής εκδοχής και της LDA εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ελέγχου



Από τους παραπάνω πίνακες και τις γραφικές παραστάσεις παρατηρούμε ότι οι τροποποιημένες εκδοχές συγκλίνουν ταχύτερα. Για την ακρίβεια, για όλη την πρώτη εποχή (που ολοκληρώνεται στο βήμα 15.840) η περιπλοκή των τροποποιημένων εκδοχών είναι μικρότερη, κάτι που εξακολουθεί να συμβαίνει μέχρι τα μέσα της δεύτερης εποχής περίπου. Στη συνέχεια, και μέχρι το τέλος της εκπαίδευσης τα συστήματα φαίνονται να έχουν κοντινές τιμές περιπλοκής , με μικρές διακυμάνσεις γύρω από αυτές.

6.1.2 BLEU

Το BLEU (bilingual evaluation understudy) σε κάθε εποχή μετριέται κάθε 5.000 training steps καθώς και στο τέλος της εποχής. Παρακάτω φαίνονται τα αποτελέσματα των μετρήσεων τόσο για το σύνολο ανάπτυξης όσο και για το σύνολο ελέγχου καθώς και οι κοινές γραφικές παραστάσεις των τιμών του BLEU της αρχικής έκδοσης και της κάθε τροποποίησης.

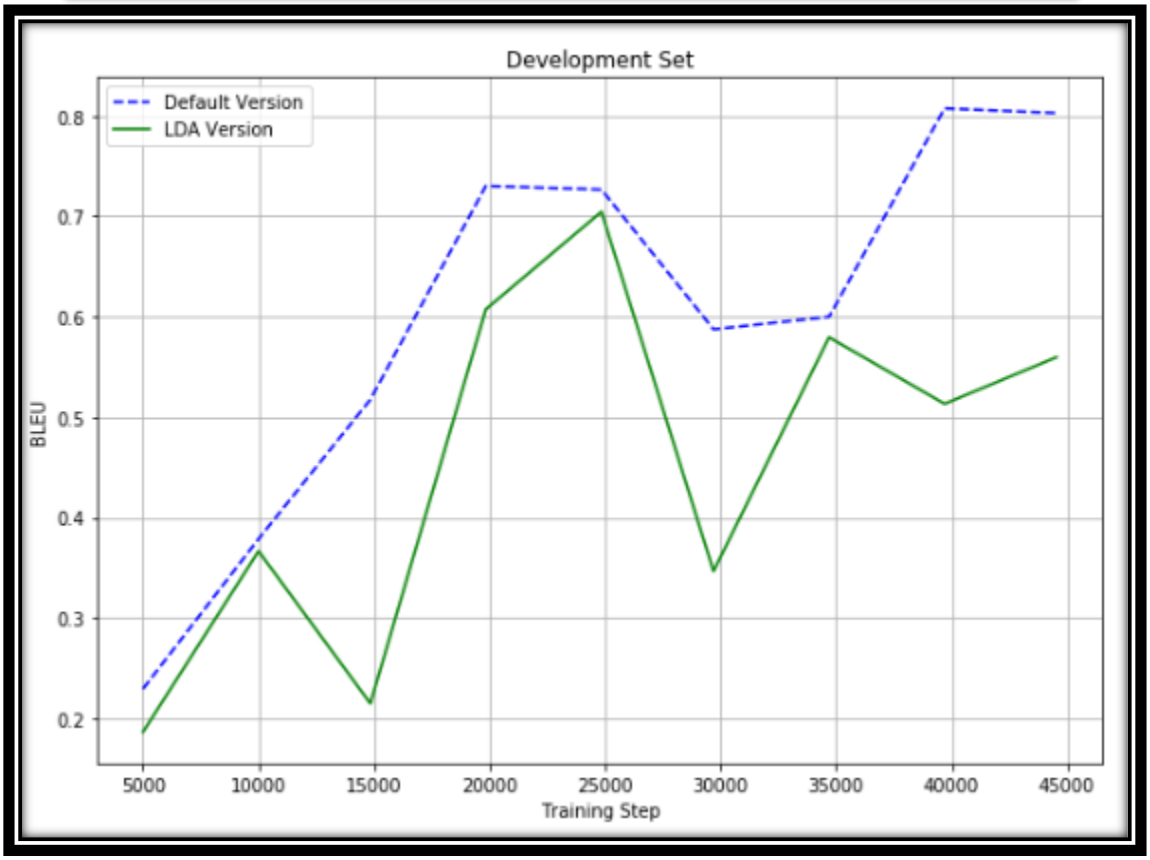
Πίνακας 7: BLEU συνόλου ανάπτυξης

Training Step	BLEU of default chatbot	BLEU of HAC chatbot	BLEU of LDA chatbot

5000	0.2294	0.3175	0.1865
10000	0.3786	0.6282	0.3665
14840	0.5174	0.5791	0.2148
19840	0.7301	0.8429	0.6073
24840	0.7264	0.9548	0.7043
29690	0.5873	0.8847	0.3465
34690	0.5997	0.8340	0.5793
39690	0.8072	0.7944	0.5131
44530	0.8026	0.7230	0.5595

Εικόνα 17: Κοινή γραφική παράσταση του BLEU της αρχικής εκδοχής και της HAC εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ανάπτυξης

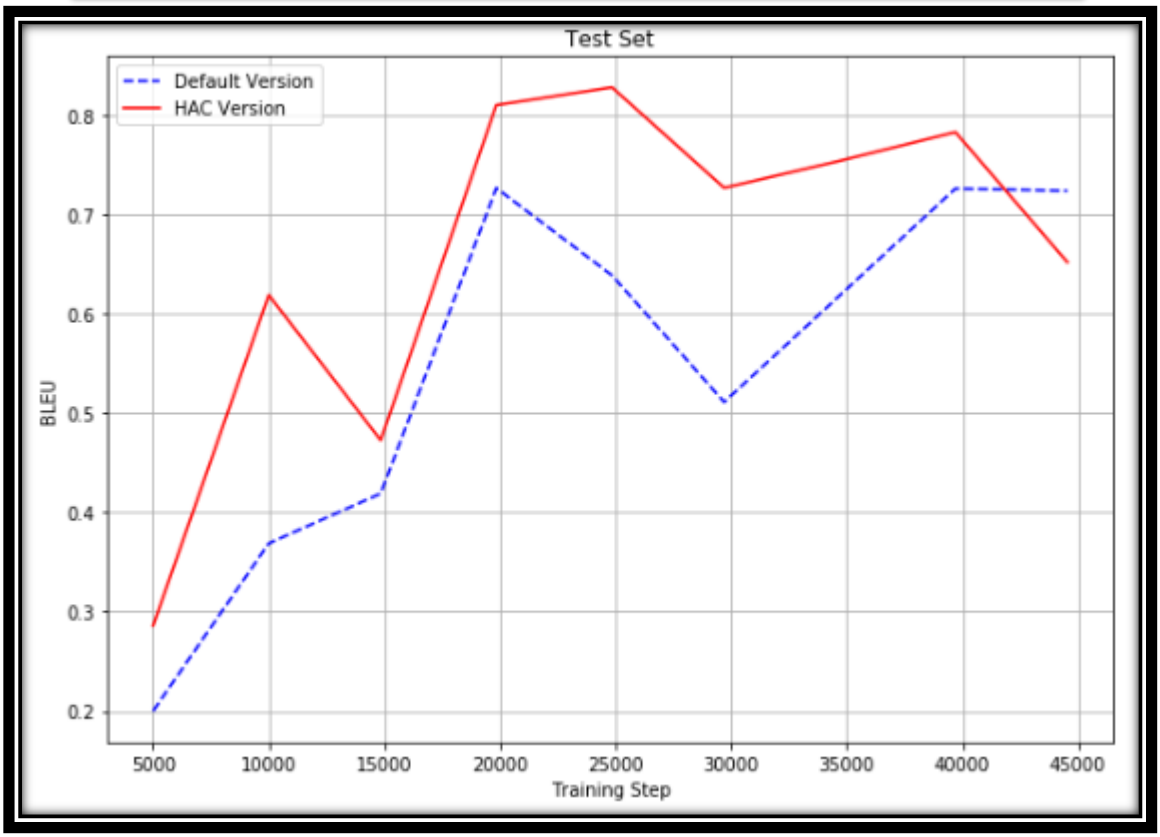
Εικόνα 18: Κοινή γραφική παράσταση του BLEU της αρχικής εκδοχής και της LDA εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ανάπτυξης



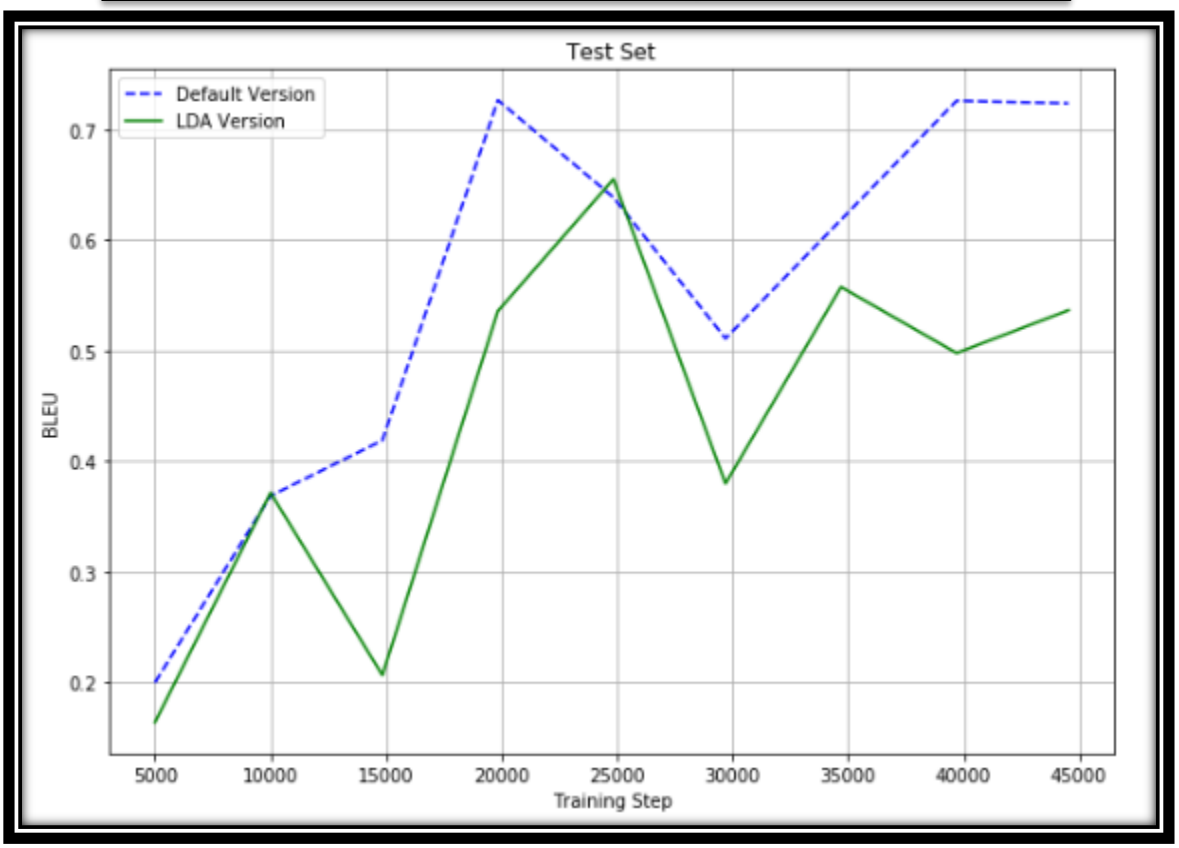
Πίνακας 8: BLEU συνόλου ελέγχου

Training Step	BLEU of default chatbot	BLEU of HAC chatbot	BLEU of LDA chatbot
5000	0.2000	0.2859	0.1641
10000	0.3687	0.6187	0.3715
14840	0.4189	0.4725	0.2066
19840	0.7264	0.8099	0.5358
24840	0.6385	0.8278	0.6550
29690	0.5110	0.7266	0.3798
34690	0.6183	0.7538	0.5575
39690	0.7259	0.7827	0.4978
44530	0.7236	0.6517	0.5365

Εικόνα 19: Κοινή γραφική παράσταση του BLEU της αρχικής εκδοχής και της HAC εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ελέγχου



Εικόνα 20: Κοινή γραφική παράσταση του BLEU της αρχικής εκδοχής και της LDA εκδοχής του chatbot συναρτήσει των βημάτων εκπαίδευσης, για το σύνολο ελέγχου



Παρατηρούμε ότι οι τιμές του BLEU όλων των εκδοχών του chatbot είναι πολύ μικρές αναφορικά με εκείνες που προκύπτουν σε ένα σύστημα μετάφρασης, γεγονός που σχετίζεται με το ότι οι σωστές μεταφράσεις μιας πρότασης έχουν κατά μεγάλη πιθανότητα κοινές λέξεις, κάτι που συνεπάγεται ότι μια καλή μετάφραση θα μοιάζει με τη μετάφραση αναφοράς άρα θα βαθμολογείται με μεγάλο BLEU. Αντίθετα, δύο αποδεκτές απαντήσεις σε μία πρόταση έχουν πολύ μικρότερη πιθανότητα να έχουν κοινές λέξεις, λόγω υποκειμενικότητας, το οποίο συνεπάγεται μια απάντηση του chatbot, αν και σωστή σημασιολογικά, θα αξιολογούνταν με χαμηλό BLEU score.

Με δεδομένο το παραπάνω, εξηγούνται και οι μεγάλες διακυμάνσεις του BLEU κατά την εκπαίδευση, ενώ ιδανικά θα περιμέναμε το BLEU να αυξάνεται κατά την εκπαίδευση. Όσον αφορά τις διάφορες εκδοχές του chatbot, η HAC εκδοχή εμφανίζει μεγαλύτερο BLEU από την αρχική ενώ η LDA μικρότερο, στο μεγαλύτερο διάστημα της εκπαίδευσης.

6.2 Ποιοτικά αποτελέσματα

Στη συνέχεια, παραθέτουμε απαντήσεις των τριών εκδοχών του chatbot σε ενδεικτικές προτάσεις. Οφείλουμε να σημειώσουμε πως οι παρακάτω προτάσεις δεν δίνουν πλήρη εικόνα της συγκριτικής απόδοσης των chatbots διότι σε άλλες προτάσεις θα είχαμε και διαφορετικές απαντήσεις.

Χρησιμοποιούμε τις προτάσεις των (Vinyals and Le 2015) για την ποιοτική αξιολόγηση των chatbots όσον αφορά την περίπτωση βασικού διαλόγου, απλών ερωτήσεων καθώς και ερωτήσεων γενικών γνώσεων. Στη συνέχεια, παραθέτουμε τρεις δικές μας επιπλέον κατηγορίες: η μία περιλαμβάνει ερωτήσεις θεματικών κατηγοριών, η άλλη έχει προτάσεις σχετικά με κάποιο ευχάριστο ή δυσάρεστο συμβάν και η τελευταία ορισμένες πολύ κοινότοπες εκφράσεις.

Βασικός διάλογος

Η κατηγορία αυτή περιλαμβάνει προτάσεις βασικού διαλόγου της καθημερινότητας.

Πίνακας 9: Απόκριση των εκδοχών του chatbot σε περιπτώσεις βασικού διαλόγου

Human's utterance	**Chatbot's response**		
	default	**HAC**	**LDA**
hello !	Username checks out.	Thanks!	Ha ha ha ha ha ha ha ha ha ha!
how are you ?	I don't know about you,	How am I?	Asking the real questions here.
what 's your name ?	I don't think so.	That's what I was thinking.	That's what I was thinking.
when were you born ?	I was born in 2008.	Asking the real questions here.	When was the last time you were born?
what year were you born ?	I was born in 2008.	I was born in 2016.	What year was this?
where are you from ?	Where I'm from.	I'm from South Carolina.	I'm from the Midwest
are you a man or a woman ?	I'm a woman.	I'm a woman.	No, I'm a woman.
why are we here ?	I don't think so.	I don't know.	I don't know.
okay, bye !	Hey, bye bye bye bye bye!	That's what I was thinking!	You're welcome!
see you later .	I second this.	That's what I was thinking.	Me too

Απλές ερωτήσεις

Οι παρακάτω ερωτήσεις είτε είναι πολλαπλής επιλογής είτε οι απαντήσεις της προκύπτουν από τα συμφραζόμενα της ερώτησης.

Πίνακας 10: Απόκριση των εκδοχών του chatbot σε απλές ερωτήσεις

Human's utterance	**Chatbot's response**		
	default	**HAC**	**LDA**
my name is david . what is my name ?	What is your name?	That's what I was thinking.	That's what I'm saying.
my name is john . what is my name ?	I don't think it's your name.	I don't know what you're talking about, but I've never heard of it.	That's what I was thinking!
are you a leader or a follower ?	I'm a follower.	I'm a follower.	No, I'm a leader.
are you a follower or a leader ?	I'm not a leader but I'm a follower.	I'm a follower and a follower.	I'm not a follower, but I'm a leader.

Ερωτήσεις γενικών γνώσεων

Παρακάτω φαίνονται κάποιες ερωτήσεις γενικών γνώσεων , οι οποίες έχουν τυχαία θεματολογία άσχετα με τα θέματα στα οποία βρήκαμε ότι κατανέμονται τα σχόλιά μας.

Πίνακας 11: Απόκριση των εκδοχών του chatbot σε ερωτήσεις γενικών γνώσεων

Human's utterance	**Chatbot's response**		
	default	**HAC**	**LDA**
who is skywalker ?	I don't know why you're getting down voted.	I'd like to know what you're talking about.	I don't know what you're talking about.
who is bill clinton ?	WHo i i i i i i iS	I'd like to know what you're talking about.	That's what I was thinking.
is sky blue or black ?	I'm pretty sure it's black.	I'm pretty sure you're right.	I can't tell if you're being sarcastic or not.
does a cat have a tail ?	My cat has a tail tail.	Asking the real questions here.	I don't think that's what it means.
does a cat have a wing ?	Cat has a wing.	Asking the real questions here.	You're right, it's a cat.

can a cat fly ?	I don't think so.	Can a dog fly?	Can confirm.
how many legs does a cat have ?	Asking the real questions here.	Asking the real questions.	Depends on how many legs you have.
how many legs does a spider have ?	Asking the real questions.	Asking the real questions.	Depends on how many legs you have.
how many legs does a centipede have ?	A centipede?	Asking the real questions here.	How many legs do you have?
what is the color of the sky ?	I think you're right.	That's what I was thinking.	I don't know what you're talking about.
what is the color of water ?	I don't think I've ever heard of water.	That's what I was thinking.	I don't know if you're joking or not, but you're correct.
what is the color of blood ?	I think you're right.	That's what I was thinking.	I don't know if you're joking or not.
what is the usual color of a leaf ?	I don't think it's a leaf.	That's what I was looking for.	I don't think it's a leaf.
what is the color of a yellow car ?	I don't think it's a yellow car.	It's a yellow car.	I don't think it's a yellow car.
how much is two plus two ?	I don't think I've ever seen one.	How much are you two?	Two plus two plus two plus two plus two plus two plus two.
how much is ten minus two ?	I'm pretty sure you're right.	How much are you counting?	100000000000000.

Ερωτήσεις θεματικών κατηγοριών

Οι παρακάτω ερωτήσεις είναι όμοιες με τις προηγούμενες, η επιλογή τους ωστόσο έχει βασιστεί στις θεματικές κατηγορίες που προέκυψαν από τις διαδικασίες της συσταδοποίησης και της LDA.

Πίνακας 12: Απόκριση των εκδοχών του chatbot στα θέματα στα οποία εκπαιδεύτηκε

Human's utterance	**Chatbot's response**		
	default	**HAC**	**LDA**
How to bake a chocolate cake?	How to bake cake?	How to bake a cake	Asking the real questions.
Should I use bitcoin?	That's what I was	That's what I was	That's what I'm

	thinking.	thinking too.	saying.
What is your favourite TV show?	I don't think I've ever seen a TV show.	I don't think I've ever heard of it.	It's my favorite show of all time.
Tell me about your favorite band!	My favorite band band of all time!	My favorite band of all time.	Thank you for your kind words.
What do you think about Donald Trump?	Donald Trump?	I don't think I've ever heard of it.	I don't think that's what he's saying.
Do you enjoy driving?	I don't think I enjoy driving. I	I don't like driving, but I can	I don't know if you're joking
Anime or manga?	Both.	Goblin slayer.	I don't think so.
What do you think about Messi?	I don't think it's a bad idea.	I don't think he's talking about Messi or Ronaldo.	I don't think that's what he's talking about.
How to get a haircut?	I don't know why you're getting downvotes.	How to get a haircut with a haircut?	How to get a haircut with a beard?
Are you a gamer?	I'm a gamer.	I'm a gamer.	No, but I'm not a gamer.
What do you think about Lakers?	I don't think he's a Lakers fan.	I don't think he's talking about the Lakers.	I don't think I've ever heard anyone say anything about the Lakers.
What is your favourite movie?	I don't think I've ever heard of that movie.	It's one of the best movies I've ever watched.	I'm in the same boat.

Προτάσεις προσωπικού περιεχομένου (θετικού/αρνητικού)

Οι παρακάτω προτάσεις αποτελούν ανακοινώσεις συμβάντων, είτε ευχάριστων είτε δυσάρεστων, ώστε να διαπιστώσουμε αν οι απαντήσεις του chatbot χαρακτηρίζονται από το αντίστοιχο - θετικό ή αρνητικό - πρόσημο.

Πίνακας 13: Απόκριση των εκδοχών του chatbot σε ευχάριστες/δυσάρεστες ανακοινώσεις

Human's utterance	**Chatbot's response**		
	default	**HAC**	**LDA**
I graduated!	Me too!	Me too!	Congratulations!

we have a baby!	I don't have a baby!	Yes!!!!!!!!!	I have a baby!
your grandma died.	She's dead.	That's what I was thinking!	Your grandma died of cancer.
I lost my job.	I lost my job because I lost my job.	I lost my job when I lost my job.	That's what I was going to say!

Κοινότοπες (ευγενικές) εκφράσεις

Τέλος έχουμε κοινότοπες εκφράσεις που χρησιμοποιούνται κατά κόρον στις συνομιλίες ώστε να διαπιστώσουμε αν το chatbot δίνει και τις αναμενόμενες κοινότοπες απαντήσεις.

Πίνακας 14: Απόκριση των εκδοχών του chatbot σε κοινότοπες εκφράσεις

Human's utterance	**Chatbot's response**		
	default	**HAC**	**LDA**
sorry.	No worries :)	No need to apolog	No worries.
I'm so sorry for your loss!	Thank you! :)	Thank you so much! I'm glad you're doing well!	Thank you for your kind words!
thanks for your advice.	You're welcome :)	You're welcome :)	Thank you for the advice.
thank you.	You're welcome :)	You're welcome :)	Thank you for your service.
congratulations!	Thank you so much!	Thank you :)	Thank you so much!
you look nice.	Thank you so much.	Thank you :)	Thank you for the compliment!
perfect job.	Thank you so much!	Thank you so much!	Username checks out.

Από τους παραπάνω πίνακες βλέπουμε ότι τα chatbot, στην πλειοψηφία των περιπτώσεων δίνουν απαντήσεις οι οποίες είναι ολοκληρωμένες και συντακτικά ορθές, όχι όμως και νοηματικά αποδεκτές. Επιπλέον, κάτι που παρατηρούμε έντονα είναι η εμφάνιση γενικών απαντήσεων που δηλώνουν συμφωνία, διαφωνία ή άγνοια όπως "That's what I was thinking", "I don't know what you're talking about". Συχνά επίσης, έχουμε επανάληψη τμήματος της ερώτησης (ή πρότασης) που του έχει τεθεί. Για παράδειγμα στις προτάσεις "We have a baby!" και "How to bake a chocolate cake?" η αρχική έκδοση απαντάει "I have a baby!" και "How to bake a cake?" αντίστοιχα.

Ειδικότερα, στον βασικό διάλογο τα chatbot δεν φαίνεται να αποδίδουν επαρκώς καλά αναφορικά με τη συχνότητα που αυτός προκύπτει στην καθημερινή ζωή. Αυτό δεν προκαλεί ιδιαίτερη εντύπωση δεδομένου ότι τα σχόλια στα οποία εκπαιδεύσαμε το σύστημα δεν αναμένουμε να περιλαμβάνουν βασικούς διαλόγους (όπως, π.χ. θα βρίσκαμε σε κάποιο chat).

Επιπλέον, παρατηρούμε καλύτερη απόκριση του chatbot όταν τίθενται ερωτήσεις από τις θεματικές κατηγορίες τις οποίες έχουμε εξάγει σε σχέση με τις τυχαίες απλές ερωτήσεις ή ερωτήσεις γνώσεων. Κάτι που επίσης εντοπίσαμε είναι ότι στην περίπτωση των τροποποιήσεων εμφανίζονται λέξεις συναφείς θεματικά με την ερώτηση, χωρίς αυτές να υπάρχουν στην ερώτηση. Για παράδειγμα, η HAC εκδοχή του chatbot στην ερώτηση “Anime or manga?” απαντά “Golbin slayer” και στην ερώτηση “What do you think about Messi?” λέει “I don’t think he’s talking about Messi or Ronaldo”. Επίσης, η LDA εκδοχή στην ερώτηση “How to get a haircut?” ρωτάει “How to get a haircut with a beard?”.

Κάτι που οφείλει επίσης να σημειωθεί είναι ότι το chatbot απαντά με επιτυχία στις κοινότοπες εκφράσεις (πχ. “thank you.” - “You’re welcome :)” ή “sorry.” - “No worries”) και αρκετά ικανοποιητικά στις περισσότερες διαζευκτικές ερωτήσεις (πχ. “Are you a leader or a follower?” - “I’m a follower.”)

Κεφάλαιο 7: Επίλογος

Το τελευταίο κεφάλαιο της εργασίας περιλαμβάνει τα συμπεράσματα στα οποία καταλήξαμε κατά την κατασκευή του ρομπότ συνομιλίας και των τροποποιήσεών του. Επίσης, αναφέρονται κάποια κύρια ζητήματα της ανάπτυξης συστήματος διαλόγων με μηχανική μάθηση στα οποία οφείλει να δοθεί έμφαση από τον ερευνητικό κλάδο. Τέλος, προτείνονται κάποιες δυνατές επεκτάσεις για καλύτερη απόδοση του συστήματος.

7.1 Συμπεράσματα

Αρχικά, συγκρίνοντας την αρχική εκδοχή του chatbot με τις τροποποιημένες, παρατηρήσαμε ότι οι τελευταίες εμφανίζουν ταχύτερη σύγκλιση, άρα πράγματι η προσθήκη θεματικών διανυσμάτων έχει νόημα με την έννοια ότι παρέχει πληροφορία στο σύστημα την οποία αυτό είναι ικανό να αξιοποιήσει. Παρ' όλα αυτά, καμία από τις εκδοχές δεν είναι σε αρκετά καλή μορφή ώστε η απόκρισή της να προσεγγίζει την ανθρώπινη.

Σε μια προσπάθεια παράλληλης εξέτασης του ζητήματος της μηχανικής μετάφρασης και της παραγωγής διαλόγου, διαπιστώνουμε ότι η δεύτερη χαρακτηρίζεται από πολύ μεγαλύτερο βαθμό υποκειμενικότητας. Τα δεδομένα στα οποία το ρομπότ συνομιλίας εκπαιδεύτηκε παρέχουν σε κάθε πρόταση (και όμοιες διατυπώσεις της) τόσες απαντήσεις όσες και οι διαφορετικές απόψεις των χρηστών, με αποτέλεσμα να μην υπάρχει η συνέπεια που υπάρχει στην παραγωγή μηχανικής μετάφρασης ή σε άλλα συστήματα νευρωνικών δικτύων, όπου όμοιες είσοδοι έχουν όμοιες εξόδους.

Αυτή η δυσκολία στο να προσδιοριστούν σωστές -ή έστω συχνές- απαντήσεις στις διάφορες προτάσεις εισόδου αντικατοπτρίζεται και στη δυσκολία εύρεσης κατάλληλων μετρικών για την μηχανική αξιολόγηση των συστημάτων παραγωγής διαλόγου. Έτσι, το BLEU που χρησιμοποιείται στη μηχανική μετάφραση και αποτελεί ένα μέτρο ομοιότητας της παραγόμενης πρότασης και των ανθρωπίνων προτάσεων, εδώ δεν μπορεί να θεωρηθεί αρκετά αντιπροσωπευτικό καθώς είναι πολύ πιθανό το ενδεχόμενο σωστές απαντήσεις σε μια πρόταση να μην περιλαμβάνουν κοινές λέξεις μεταξύ τους και άρα να αξιολογούνται ως λανθασμένες.

7.2 Δυνατές επεκτάσεις

Προκειμένου να είναι πληρέστερη η αξιολόγηση του ρομπότ συνομιλίας θα μπορούσε αυτή να γίνει απ' ευθείας από ανθρώπους σε ένα προκαθορισμένο σύνολο ερωτήσεων. Αν, ωστόσο, επιλεγεί η μηχανική αξιολόγηση, είναι σκόπιμο να χρησιμοποιηθούν περισσότερες της μίας απαντήσεις ώστε το BLEU να είναι πιο αντιπροσωπευτικό.

Επίσης, αξίζει να γίνει μια διεξοδική μελέτη αλληλεπίδρασης και προσδιορισμού των διαφόρων παραμέτρων του συστήματος. Μια πλεγματική αναζήτηση των κατάλληλων συνδυασμών παραμέτρων προϋποθέτει επίσης έναν τρόπο αξιολόγησης (αντίστοιχο του BLEU) τον οποίο εμπιστευόμαστε, ώστε με βάση αυτόν να επιλέξουμε το καλύτερο σύστημα.

Τέλος, παραθέτουμε μια σκέψη προκειμένου να αντιμετωπιστεί το ζήτημα της ποικιλίας απόψεων: Πολύ πιθανόν, αν εκπαιδεύαμε το σύστημα να μιμείται έναν συγκεκριμένο χρήστη του reddit, σε όλους τους μήνες που το reddit έχει διαθέσιμους, αυτό να ήταν πιο συνεπές. Φυσικά, εκεί

θα είχαμε περισσότερο περιορισμένα αποτελέσματα, όσον αφορά το πλήθος των σχολίων αλλά και τα subreddits στα οποία αυτό θα συμμετείχε.

Βιβλιογραφία

Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” *ICLR* 2015.

Yoshua Bengio, Réjean Ducharme , Pascal Vincent , Christian Jauvin. “A Neural Probabilistic Language Model.” *JMLR* 2003: 1137–1155.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. “Learning long-term dependencies with gradient descent is difficult.” *IEEE Transactions on Neural Networks* 1994: 157-166.

David M. Blei, Andrew Y. Ng, Michael I. Jordan. “Latent Dirichlet Allocation.” *JMLR* 2003: 993–1022.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau , Fethi Bougares , Holger Schwenk , Yoshua Bengio. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. .” *EMNLP*. 2014 .

Jeffrey L. Elman, “Finding structure in time.” *Cognitive Science*. 1990.

Felix A. Gers, Jürgen A. Schmidhuber, Fred A. Cummins. “Learning to forget: Continual prediction with LSTM.” *Neural Computation* 2000: 2451–2471.

Sepp Hochreiter, Jürgen Schmidhuber. “Long short-term memory.” *Neural Computation* 1997: 1735–1780.

Minh-Thang Luong. “Neural Machine Translation.” Ph.D. thesis. 2016.

T. Soni Madhulatha,. “ An Overview on Clustering Methods.” *IOSR Journal of Engineering* 2012.

James Martens, Ilya Sutskever. “Learning recurrent neural networks with Hessian free optimization.” *ICML*. 2011.

Tomáš Mikolov. *Statistical Language Models Based on Neural Networks*. Brno University of Technology, 2012. Ph.D. thesis.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan “Honza” Černocký, Sanjeev Khudanpur. “Recurrent neural network based language model.” *Interspeech*. 2010.

Razvan Pascanu, Tomáš Mikolov, and Yoshua Bengio. “ On the difficulty of training recurrent neural networks.” *ICML*. 2013.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: a Method for Automatic Evaluation of Machine Translation.” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2002. 311-318.

Rico Sennrich, Barry Haddow , Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016. 1715–1725.

Ilya Sutskever, Oriol Vinyals, Quoc V. Le. “Sequence to sequence learning with neural networks.” *NIPS*. 2014.

Jian Tang, Zhaoshi Meng, XuanLong Ngyuen, Qiaozhu Mei, Ming Zhang. “Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis.” *Proceedings of the 31 st International Conference on Machine Learning*. 2014.

Oriol Vinyals, Quoc V. Le. “A Neural Conversational Mode.” *ICML Deep Learning Workshop 2015*. 2015.

Wikipedia contributors. “History of artificial intelligence.” *Wikipedia, The Free Encyclopedia*, 21 May 2019. Web. 23 May 2019.

Wikipedia contributors. “Chatbot”. *Wikipedia, The Free Encyclopedia*, 22 May 2019. Web. 23 May 2019.

Jian Zhang, Liangyou Li, Andy Way, Qun Liu. “ Topic-Informed Neural Machine Translation.” *Proceedings of 26th International Conference on Computational Linguistics*. 2016.