

**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**



- Πτυχιακή Εργασία -

**Εφαρμογή αρχιτεκτονικών βαθιάς μάθησης στην αξιολόγηση μηχανικής
μετάφρασης**

Ζαφείριος Καραμποϊκης

ΑΜ : Π2013084

Επιβλέπουσα : Κάτια-Λύδα Κερμανίδου

Κέρκυρα 2023

Επιβλέπουσα

Κάτια-Λύδα Κερμανίδου, *Αναπληρώτρια Καθηγήτρια*
Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Θεμιστοκλής Έξαρχος, *Αναπληρωτής Καθηγητής*
Ιόνιο Πανεπιστήμιο
Άννα Σωτηροπούλου, *ΕΔΙΠ*,
Ιόνιο Πανεπιστήμιο

Περίληψη

Στόχος της συγκεκριμένης εργασίας είναι η διερεύνηση της εφαρμογή αρχιτεκτονικών βαθιάς μάθησης σε συνδυασμό με την χρήση της μορφής Bag Of Words ως εργαλείο αξιολόγησης μηχανικής μετάφρασης. Αρχικά έγινε μια σύντομη ιστορική αναδρομή στον τομέα της μετάφρασης και στη συνέχεια μελετήθηκαν τεχνικές μηχανικής μετάφρασης καθώς και υφιστάμενες τεχνικές αξιολόγησης αυτών, αυτόματες αλλά και ανθρώπινες. Στα πλαίσια διεξαγωγής πειραμάτων, μερικές από τις υφιστάμενες τεχνικές εφαρμόστηκαν στο σετ δεδομένων που δόθηκε ώστε να συγκριθούν τα αποτελέσματα τους με αυτά της εφαρμογής βαθιών αρχιτεκτονικών με δεδομένα στην μορφή Bag Of Words, μέθοδος για την οποία εκτελέστηκαν στο σύνολο 1440 πειράματα. Παρόλο που η ανάλυση των τελικών αποτελεσμάτων υποδεικνύει ότι η συγκεκριμένη προσέγγιση δεν έχει ιδιαίτερο ερευνητικό ενδιαφέρον, μπορούν από αυτά να εξαχθούν πορίσματα και προτάσεις που ίσως βοηθήσουν μελλοντικές προσεγγίσεις.

Περιεχόμενα

A Εισαγωγή.....	5
B Μηχανική Μετάφραση.....	6
B.1 Μηχανική Μετάφραση Βασισμένη σε Κανόνες.....	6
B.2 Στατιστική Μηχανική Μετάφραση.....	7
Γ Νευρωνική Μηχανική Μετάφραση.....	8
Γ.1 RNN Encoder-Decoder Framework.....	8
Γ.2 LSTM Encoder-Decoder Framework.....	8
Γ.3 Μοντέλα Προσοχής.....	8
Γ.3.1 Transformer.....	9
Γ.3.2 GPT-4, PaLM 2 και άλλα LLMs.....	9
Δ Αξιολόγηση Μηχανικής Μετάφρασης.....	10
Δ.1 Τι είναι η αξιολόγηση Μηχανικής Μετάφρασης?.....	10
Δ.2 Ανθρώπινη Αξιολόγηση.....	10
Δ.2.1 Adequacy and Fluency.....	10
Δ.2.2 Απευθείας Αξιολόγηση.....	10
Δ.2.3 Multidimensional Quality Metrics.....	10
Δ.3 Αυτόματη Αξιολόγηση.....	11
Δ.3.1 WER (Word Error Rate).....	11
Δ.3.2 BLEU.....	11
Δ.3.3 METEOR (Metric for Evaluation of Translation with Explicit Ordering).....	12
Δ.3.4 Translation Error Rate.....	12
Δ.3.5 ROUGE (Recall-Oriented Understudy for Gisting Evaluation).....	13
Δ.3.6 CHRF (Character n-gram F-score).....	13
Δ.3.7 BERTScore.....	13
Δ.3.8 COMET.....	13
Δ.3.9 Σχήματα μάθησης.....	13
Ε Έρευνα.....	14
E.1 Εργαλεία που χρησιμοποιήθηκαν.....	14
E.2 Για το αρχικό Dataset.....	14
E.3 Υλοποίηση.....	15
E.3.1 Επεξεργασία και Προεπεξεργασία Δεδομένων.....	15
E.3.2 Επισημείωση μετρήσεων και αξιολόγησης αυτών.....	16
E.3.2.1 Μέτρηση Bleu.....	16
E.3.2.2 Μέτρηση ROUGE.....	16
E.3.2.3 FastText.....	16
E.3.3 Bag of Words σε Νευρωνικό Δίκτυο.....	17
ΣΤ Πειράματα και αποτελέσματα.....	21
ΣΤ.1 Πειράματα και αποτελέσματα μετρήσεων BLEU.....	21
ΣΤ.2 Πειράματα και αποτελέσματα μετρήσεων ROUGE.....	21
ΣΤ.3 Πειράματα και αποτελέσματα με την χρήση της βιβλιοθήκης fastText.....	22
ΣΤ.4 Πειράματα και αποτελέσματα με το συνδυασμό Bag Of Words και TNΔ.....	22
Ζ Συμπεράσματα – Προτάσεις για το μέλλον.....	23
Ζ.1 Σύγκριση με αποτελέσματα προηγούμενων ερευνών.....	23
Ζ.2 Εκπαίδευση μοντέλων με μεγαλύτερο πλήθος νευρώνων στα κρυφά στρώματα.....	23

Z.3 Συνδυασμός με περισσότερα δεδομένα.....	23
Z.4 Διαχωρισμός σετ εκπαίδευσης και σετ αξιολόγησης με αλγόριθμο τυχαιότητας.....	23
H Πίνακες – Σχήματα – Αναφορές.....	24
H.1 NLTK Stopwords.....	40
H.2 Additional Stopwords.....	40
Θ Συντμήσεις.....	40
I Γλωσσάρι Ξενικών Όρων.....	42
ΙΑ Βιβλιογραφία.....	44

A Εισαγωγή

Η αξιολόγηση μηχανικής μετάφρασης είναι ζωτικής σημασίας για την ανάπτυξη και παραμετροποίηση εφαρμογών μηχανικής μετάφρασης. Καθώς τα συστήματα μηχανικής μετάφρασης γίνονται όλο και πιο περίπλοκα και αποτελεσματικά, η αξιολόγηση τους εξασφαλίζει ότι τα αποτελέσματα τους πληρούν τα πρότυπα ποιότητας και ευφράδειας. Η αξιολόγηση μηχανικής μετάφρασης, παίζει σημαντικό ρόλο σε όλα τα στάδια της ανάπτυξης εφαρμογών μηχανικής μετάφρασης, από την εύρεση αδυναμιών των συστημάτων αυτών μέχρι την σύγκριση και δημοσίευση πειραμάτων, ακόμη και στο ποια μοντέλα θα φτάσουν τελικά την παραγωγή και τον τελικό χρήστη[1]. Κατά την εκπόνηση αυτής της πτυχιακής εργασίας, δοκιμάστηκε για την αξιολόγηση μηχανικής μετάφρασης η μετατροπή του dataset σε μορφή Bag Of Words σε συνδυασμό με την χρήση νευρωνικών δικτύων. Στη συνέχεια έγινε σύγκριση των αποτελεσμάτων με αυτά προϋπαρχόντων τεχνικών όπως η BLEU και η ROUGE, αλλά και η εξαγωγή embeddings χρησιμοποιώντας προεκπαιδευμένα μοντέλα, και σύγκριση αυτών. Με βάση τα πειράματα που έγιναν δεν φαίνεται η συγκεκριμένη τεχνική να έχει κάποιο ερευνητικό ενδιαφέρον.

B Μηχανική Μετάφραση

Η μηχανική μετάφραση αποτελεί την αυτοματοποιημένη διαδικασία μετάφρασης κειμένου ή προφορικού λόγου από μία γλώσσα σε μία άλλη. Σε αντίθεση με την ανθρώπινη μετάφραση, όπου ένας γλωσσολόγος διερμηνεύει το νόημα και τη σημασία για να μεταφράσει το περιεχόμενο, η μηχανική μετάφραση βασίζεται σε αλγορίθμους και υπολογιστικά μοντέλα. Ο τομέας της μηχανικής μετάφρασης περιλαμβάνει πολλαπλές τεχνικές, από τεχνικές βασισμένες σε κανόνες και στατιστικές προσεγγίσεις, έως και πιο ανεπτυγμένες με τη χρήση βαθιών νευρωνικών δικτύων. Σε μια παγκόσμια κοινωνία που γίνεται ολοένα και πιο διασυνδεδεμένη, η ανάγκη να ξεπεράσουμε τα γλωσσικά εμπόδια εμφανίζεται όλο και συχνότερα. Η Μηχανική Μετάφραση παίζει σημαντικό ρόλο στην διευκόλυνση διεθνών επιχειρήσεων, πολιτισμικών ανταλλαγών αλλά και την πρόσβαση σε πληροφορίες που υπάρχουν σε άλλες γλώσσες. Από την εκπαίδευση και την υγεία μέχρι την τεχνολογία και τη διασκέδαση, οι εφαρμογές της μηχανικής μετάφρασης είναι πολλές και αυξανόμενες.

Στα μέσα της δεκαετίας του 1930, οι Georges Astrouni και Petr Troyanskii κάνουν τις πρώτες αιτήσεις κατοχύρωσης πατέντας για μηχανές μετάφρασης, οι ιδέες τους όμως δεν έγιναν γνωστές μέχρι τα τέλη της δεκαετίας του 1950. Η Ιστορία της Μηχανικής Μετάφρασης ξεκινάει με τα πρώτα χρόνια χρήσεων υπολογιστικών μηχανών. Με την εμφάνιση των πρώτων ηλεκτρονικών αριθμομηχανών, ξεκίνησε και η έρευνα για τη χρήση τους ως βοηθούς στην μετάφραση φυσικής γλώσσας. Ως αρχή μπορεί να χαρακτηριστεί ένα γράμμα από το ίδρυμα Rockefeller στον Norbert Wiener το 1947. Το 1954 δόθηκε η πρώτη δημόσια επίδειξη για την δυνατότητα της επίτευξης Μηχανικής Μετάφρασης, σε συνεργασία του πανεπιστημίου της Georgetown με την IBM, και παρά το πολύ περιορισμένο λεξιλόγιο και την επίσης περιορισμένη γραμματική, ήταν αρκετό για να αποσπάσει μεγάλη χρηματοδότηση αλλά και να εμπνεύσει την εκκίνηση αντίστοιχων έργων Μηχανικής Μετάφρασης σε όλο τον κόσμο[2]. Η αρχική της φάση χαρακτηριζόταν από μεθόδους βασισμένες σε κανόνες, όπου οι μεταφραστές προγραμματίζαν γλωσσολογικούς κανόνες μέσα στο σύστημα. Το τέλος αυτής της εποχής, ακολούθησε η περίοδος της Στατιστικής Μηχανικής Μετάφρασης στα τέλη των 80s και 90s, όπου στατιστικά μοντέλα και μεγάλα δίγλωσσα corpora χρησιμοποιήθηκαν για την πραγματοποίηση αυτοματοποιημένης μετάφρασης. Ο ερχομός του 21ου αιώνα έφερε την μετάβαση στην Νευρωνική Μηχανική Μάθηση με τη χρήση βαθιών νευρωνικών δικτύων για την παράδοση αποτελεσμάτων με μεγαλύτερη ευφράδεια και ακρίβεια. Κάθε εποχή προσέφερε στον τομέα της μηχανικής μετάφρασης και οι τωρινές τεχνολογίες αποτελούν απόδειξη της συνεχούς προόδου και καινοτομίας σε αυτόν.

B.1 Μηχανική Μετάφραση Βασισμένη σε Κανόνες

Η βασισμένη σε κανόνες μηχανική μετάφραση(RBMT) είναι μια τεχνική που καταπιάνεται από τις βάσεις γλωσσολογικών κανόνων και λεξιλογίων που έχουν φτιαχτεί και για την γλώσσα της πηγής αλλά και για τη γλώσσα της τελικής μετάφρασης. Η RBMT λειτουργεί με βάση ένα ορισμένο σύνολο γραμματικών, συντακτικών και λεκτικών κανόνων. Η διαδικασία της μετάφρασης περιλαμβάνει την απορρόφηση του κειμένου της πηγής έτσι ώστε να διαμορφωθεί μια ενδιάμεση αναπαράσταση του. Αυτή η αναπαράσταση τότε χρησιμοποιείται για την δημιουργία της τελικής μετάφρασης [3]. Ένα από τα πλεονεκτήματα είναι ότι όταν οι κανόνες είναι πολύ προσεγμένοι, μπορεί να δημιουργεί συνεχώς ορθά δομημένες μεταφράσεις. Βέβαια η δημιουργία τέτοιων κανόνων απαιτεί βαθιές γνώσεις των δύο γλωσσών αλλά και πολύωρη εργασία. Αυτή η μέθοδος κυριάρχησε στα αρχικά βήματα της ανάπτυξης της μηχανικής μετάφρασης, πριν την ανάδειξη μεθόδων που βασίζονταν παραπάνω σε δεδομένα, όπως η στατιστική μηχανική μετάφραση.

B.2 Στατιστική Μηχανική Μετάφραση

Η Στατιστική Μηχανική Μετάφραση αφορά τεχνικές που βασίζονται πολύ περισσότερο σε δεδομένα, αντί να βασίζονται σε πολύ προσεκτικά δομημένους κανόνες, τα Στατιστικά συστήματα λειτουργούν αναλύοντας μεγάλα σύνολα δίγλωσσων κειμένων όπως παράλληλα corpora, όπου οι προτάσεις του ενός κειμένου αποτελούν την μετάφραση της αντίστοιχης πρότασης από το άλλο. Μέσω αυτής της ανάλυσης, τα στατιστικά μοντέλα δημιουργούν στατιστικά μοτίβα και πιθανότητες για την αντιστοίχιση λέξεων και φράσεων. Ένα σημαντικό πλεονέκτημα της ΣΜΜ είναι η προσαρμοστικότητα της. Αν δοθούν αρκετά δεδομένα, μπορεί να προσαρμοστεί σε διαφορετικές λεκτικές συνήθειες αλλά και τομείς[4].

Γ Νευρωνική Μηχανική Μετάφραση

Η Νευρωνική Μηχανική Μετάφραση αποτελεί τομέα της μηχανικής μετάφρασης στον οποίο γίνεται χρήση Τεχνητών Νευρωνικών Δικτύων. Για την επίτευξη της μετάφρασης, τα τεχνητά νευρωνικά δίκτυα πρέπει να εκπαιδευτούν πάνω σε μεγάλα δίσλωσσα σετ δεδομένων, δημιουργώντας έτσι μια εσωτερική αναπαράσταση της γλώσσας. Τα αρχικά μοντέλα Νευρωνικής Μηχανικής Μετάφρασης βασίζονταν σε RNNs ή LSTMs[5]. Στη συνέχεια, έγιναν δημοφιλέστερα τα μοντέλα που χρησιμοποιούν μηχανισμούς προσοχής, χάρις τους οποίους το μοντέλο μπορεί να “στρέψει την προσοχή του” προς διαφορετικά τμήματα των δεδομένων εισόδου ώστε να παράξει την κάθε λέξη του τελικού αποτελέσματος. Ένα από τα καθοριστικά μοντέλα που χρησιμοποιούν μια τέτοια μέθοδο είναι η αρχιτεκτονική Transformer[6]. Σημαντικά παραδείγματα της εφαρμογής του μοντέλου του Transformer αποτελεί το BERT[7] και οι διάφορες παραλλαγές του. Αυτά τα μοντέλα έχουν παρουσιάσει αξιοσημείωτη ευφράδεια και κατανόηση στις μεταφράσεις τους.

Γ.1 RNN Encoder-Decoder Framework

Η αρχιτεκτονική encoder-decoder εφαρμόζει μία διαδικασία αποτελούμενη από δύο βήματα, σχεδιασμένη έτσι ώστε να διαχειρίζεται εισόδους διαφόρων μεγεθών και να δημιουργεί ακολουθίες. Στο πρώτο βήμα, ο encoder επεξεργάζεται τα δεδομένα εισόδου λέξη προς λέξη, και ανανεώνει την εσωτερική του κατάσταση σε κάθε βήμα. Αυτή η διαδικασία ολοκληρώνεται με τον encoder να δημιουργεί ένα συμπαγές context vector, που έχει ως στόχο να συμπεριλάβει ολόκληρη την σημασιολογική ουσία της πηγαίας πρότασης. Αυτό το διάνυσμα χρησιμοποιείτε για την γεφύρωση με το επόμενο βήμα, κατά το οποίο ο decoder ξεκινά με βάση αυτό να δημιουργεί την τελική μεταφρασμένη πρόταση. Σε κάθε βήμα της αποκωδικοποίησης ο decoder προβλέπει το επόμενο token(είτε αυτό αποτελεί ολόκληρη λέξη είτε μέρος αυτής), με βάση την κατάσταση του την δεδομένη στιγμή. Το κάθε token επηρεάζει και τις επόμενες προβλέψεις. Στην ουσία ο encoder δημιουργεί μία συμπαγής αναπαράσταση της πηγής την οποία ο decoder αποκωδικοποιεί για να καταλήξει στην τελική μετάφραση[8]. Φυσικά το πρόβλημα των vanishing και exploding gradients[9], που αντιμετωπίζουν τα RNNs υφίσταται και στο Encoder-Decoder framework και ενώ η βασική αρχιτεκτονική που περιγράφεται παραπάνω αποτέλεσε θεμέλιο για τα μελλοντικά Seq2Seq μοντέλα, τα προβλήματα τα οποία αντιμετωπίζει με μεγαλύτερες προτάσεις λόγω των vanishing/exploding gradients ή με μικρές διαφοροποιήσεις στο περιεχόμενο, περιορίζουν τις δυνατότητες του. Εξαιτίας αυτών των περιορισμών ο τομέας οδηγήθηκε στην εξερεύνηση πιο προχωρημένων λύσεων.

Γ.2 LSTM Encoder-Decoder Framework

Τα Δίκτυα Μακράς Βραχύχρονης Μνήμης(LSTM) έπαιξαν καθοριστικό ρόλο στην ανάπτυξη του τομέα της Μηχανικής Μετάφρασης χάρις την δυνατότητα τους να επεξεργαστούν αποτελεσματικότερα μακρόχρονες εξαρτήσεις σε ακολουθίες, σε σχέση με τα RNNs. Όσων αφορά την νευρωνική Μηχανική Μετάφραση, τα LSTMs τυπικά υλοποιούνται σε μια δομή τύπου Encoder-Decoder, όπως και τα απλούστερα RNNs. Ο ξεχωριστός μηχανισμός θηρών των LSTMs, τους επιτρέπει να διαχειρίζονται καλύτερα την ροή της πληροφορίας, καθιστώντας τα έτσι πολύ καταλληλότερα για τη διαχείριση μακρύτερων ακολουθιών, καθώς καταφέρνουν να ξεπεράσουν προβλήματα όπως το αυτό της εξαφάνισης/εκρηκτικής κλίσης, το οποίο αποτελεί μεγάλο πρόβλημα για την αρχιτεκτονική των τυπικών RNNs[10]. Αυτή η δυνατότητα είναι άκρως σημαντική στον τομέα της μηχανικής μετάφρασης, στον οποίο η κατανόηση του περιεχομένου μεγαλύτερων προτάσεων είναι απαραίτητη για την σύνθεση σωστών μεταφράσεων[11].

Γ.3 Μοντέλα Προσοχής

Οι μηχανισμοί προσοχής, οδήγησαν σε σημαντικές βελτιώσεις στον τομέα της νευρωνικής μηχανικής μετάφρασης. Αυτά τα μοντέλα βασίζονται στον στάνταρ Encoder-Decoder μηχανισμό αλλά ξεπερνούν το

πρόβλημα που αντιμετωπίζεται σε αυτόν, όπου η πηγαία πρόταση συμπιέζεται σε ένα συγκεκριμένου μεγέθους διάνυσμα περιεχομένου, πολλές φορές οδηγώντας έτσι σε χάσιμο πληροφορίας για μακρύτερες προτάσεις. Ο μηχανισμός προσοχής επιτρέπει στον decoder να "συγκεντρωθεί" σε διαφορετικά τμήματα της πηγαίας πρότασης σε κάθε βήμα αποκωδικοποίησης, ορίζοντας δυναμικά το βάρος της συσχέτισης του κάθε token. Καθώς ο decoder δημιουργεί κάθε τελικό token, ο μηχανισμός προσοχής υπολογίζει ένα σταθμισμένο άθροισμα των κρυφών σταδίων του encoder με βάση την σχετικότητα τους, με αποτέλεσμα ένα διάνυσμα περιεχομένου ειδικά σχεδιασμένο για το συγκεκριμένο βήμα αποκωδικοποίησης. Αυτό το προσαρμόσιμο context εξασφαλίζει ότι ο αποκωδικοποιητής έχει πρόσβαση σε ολόκληρη την πηγαία πρόταση και μπορεί να προσαρμοστεί στις σχετικές λεπτομέρειες, βελτιώνοντας έτσι δραματικά την ποιότητα της μετάφρασης, ειδικά για μεγαλύτερες ακολουθίες. Πρωτοπόρο μοντέλο προσοχής για τη μηχανική μετάφραση θεωρείται αυτό των Bahdanau et al.'s, το οποίο απέδειξε την αποτελεσματικότητα του στην απόδοση σύμφωνης πληροφορίας μεταξύ πηγής και τελικής μετάφρασης[5].

Γ.3.1 Transformer

Ενώ το μοντέλο των Bahdanau et al. έκανε την αρχή με την εισαγωγή του μηχανισμού προσοχής ως μια συμπληρωματική βελτίωση στην τότε standard μορφή του Encoder-Decoder μοντέλου, το 2017 προτάθηκε από τους Vaswani et al η αρχιτεκτονική Transformer. Η κεντρική διαφορά της αρχιτεκτονικής του Transformer με αυτή του standard Encoder-Decoder, είναι ότι στους Transformers καταργούνται εντελώς οι μηχανισμοί αναδρομής, αφήνοντας μόνο το μηχανισμό προσοχής[6]. Ενώ ο αρχικός σχεδιασμός των transformers είχε σαν στόχο την χρήση τους στον τομέα της μηχανικής μετάφρασης, η δομή αυτή έχει κυριαρχήσει τα τελευταία χρόνια γενικότερα στον τομέα της τεχνητής νοημοσύνης καθώς χρησιμοποιήθηκε στην ανάπτυξη μοντέλων όπως το GPT-4[12] και το DALL-E[13], τα οποία έχουν επηρεάσει δραστικά τον τελευταίο χρόνο την κοινή γνώμη σχετικά με την πορεία της τεχνητής νοημοσύνης και το ρόλο αυτής στην καθημερινότητα μας και ειδικότερα στον εργασιακό τομέα.

Γ.3.2 GPT-4, PaLM 2 και άλλα LLMs

Τα LLMs που κυριαρχούν αυτή τη στιγμή, έχουν παρουσιάσει εκπληκτικά αποτελέσματα όσον αφορά την δημιουργία φυσικής γλώσσας, όμως οι δυνατότητες τους όσον αφορά την μετάφραση, τον τομέα για τον οποίο αναπτύχθηκε αρχικά η αρχιτεκτονική Transformer, δεν έχουν εξεταστεί εκτενώς. Παρόλο που η αρχιτεκτονική τους δεν ακολουθεί το encoder – decoder framework, όπως τα περισσότερα NMT συστήματα, έχουν δείξει σημάδια ικανότητας μετάφρασης. Οι έρευνες των Hendy et al δείχνουν ότι με την απλή χρήση των LLMs μπορεί να επιτευχθούν αρκετά ικανοποιητικές μεταφράσεις, ενώ χρησιμοποιώντας τα σε μια υβριδική προσέγγιση σε συνδυασμό με κλασσικά NMT συστήματα μπορεί να επιτευχθεί κορυφαία ποιότητα αποτελεσμάτων.[14]

Δ Αξιολόγηση Μηχανικής Μετάφρασης

Δ.1 Τι είναι η αξιολόγηση Μηχανικής Μετάφρασης?

Η αξιολόγηση Μηχανικής Μετάφρασης (είτε γίνεται αυτόματα είτε από ανθρώπους) έχει ως σκοπό να εκτιμήσει την ποιότητα της μετάφρασης που πραγματοποιήθηκε αυτόματα. Για την αυτόματη αξιολόγηση πρέπει να χρησιμοποιηθούν μετρήσεις που δείχνουν την ποιότητα της μετάφρασης. Βέβαια, η ποιότητα της μετάφρασης είναι καθαρά υποκειμενική. Καθώς δεν υπάρχει αντικειμενικό και μετρήσιμο “καλό”, κάθε μέτρηση πρέπει να παραθέτει βαθμολογίες της ποιότητας που συμβαδίζουν με την ανθρώπινη κρίση αυτής της ποιότητας. Δηλαδή μια μέτρηση της ποιότητας θα πρέπει να είναι υψηλή όταν και οι άνθρωποι θεωρούν υψηλή την ποιότητα της μετάφρασης, και χαμηλή όταν οι άνθρωποι θεωρούν χαμηλή την ποιότητα της μετάφρασης. Η ανθρώπινη κρίση αποτελεί το σημείο αναφοράς για την επιλογή αυτόματων μετρήσεων, καθώς οι άνθρωποι αποτελούν τους τελικούς αποδέκτες της μετάφρασης.

Δ.2 Ανθρώπινη Αξιολόγηση

Η διαδικασία της μετάφρασης(είτε είναι αυτόματη, είτε ανθρώπινη), αποτελεί διαδικασία η οποία απαιτεί ιδιαίτερη και εκτενής προσοχή, έτσι ώστε να επιτευχθεί η σωστή αναμετάδοση μηνυμάτων μεταξύ των διαφορετικών γλωσσών. Παρά την εμφάνιση αρκετών νέων εργαλείων και μεθόδων για την αξιολόγηση της ποιότητας της μετάφρασης, η ανθρώπινη αξιολόγηση εξακολουθεί να έχει τα καλύτερα αποτελέσματα. Ενώ η μηχανική μετάφραση έχει δει μεγάλες βελτιώσεις με εργαλεία όπως το Google Translate, το DeepL και τα μοντέλα βασισμένα στο BERT[7], η πρόοδος στον τομέα της αξιολόγησης της μηχανικής μετάφρασης δεν έχει κάνει τα ίδια άλματα. Εδώ έρχεται πρώτη η ανθρώπινη κρίση. Οι αλγόριθμοι μπορεί να αστοχήσουν στην κατανόηση ιδιωμάτων και συναισθημάτων τα οποία οι άνθρωποι μπορούν να κατανοήσουν ενστικτωδώς. Στη συνέχεια ακολουθούν μερικές μέθοδοι ανθρώπινης αξιολόγησης.

Δ.2.1 *Adequacy and Fluency*

Σύμφωνα με τους D. Jurafsky, J.H. Martin , οι μεταφράσεις αξιολογούνται σε δύο διαστάσεις. Την πιστότητα(δηλαδή το πόσο καλά συλλαμβάνουν το περιεχόμενο της πρότασης) και την ευφράδεια(δηλαδή το πόσο ευφραδής είναι η τελική μετάφραση στην γλώσσα την οποία στοχεύουμε)[15].

Δ.2.2 *Απευθείας Αξιολόγηση*

Με την απευθείας αξιολόγηση(Direct Assessment), οι αξιολογητές βαθμολογούν την ποιότητα της μετάφρασης με βάση μια προκαθορισμένη κλίμακα. Μπορεί να προσφέρει γρηγορότερες αξιολογήσεις για την παραμετροποίηση ενός μοντέλου. Η απευθείας αξιολόγηση μπορεί να κατηγοριοποιηθεί σε monolingual(οι άνθρωποι-αξιολογητές βλέπουν μόνο την τελική μετάφραση/candidate), bilingual(οι άνθρωποι-αξιολογητές βλέπουν την τελική μετάφραση/candidate αλλά και την πηγαία πρόταση) και reference-based(οι άνθρωποι-αξιολογητές βλέπουν την μετάφραση/candidate αλλά και την “σωστή” μετάφραση αναφοράς) [16].

Δ.2.3 *Multidimensional Quality Metrics*

Το MQM αποτελεί δομή ενός ευέλικτου συστήματος πάνω στο οποίο μπορούν να δηλωθούν μετρικές αξιολόγησης ποιότητας μετάφρασης. Ο αρχικός σκοπός είναι η αξιολόγηση της μετάφρασης, αλλά μπορεί να

αξιολογηθεί και η ποιότητα της πηγής. Το MQM έχει σχεδιαστεί για να μπορεί να ταιριάζει με κάθε είδος μεταφρασμένου κειμένου, δεν μπορεί να συμβάλει όμως στην δημιουργία ενός μοντέλου αξιολόγησης το οποίο θα είναι ιδανικό για κάθε την αξιολόγηση κάθε μετάφρασης [17]. Η βασική ιδέα του MQM προϋποθέτει την ύπαρξη μιας ιεραρχικής λίστας τύπων σφαλμάτων(Error Types), η οποία προέκυψε μετά από εκτενή έρευνα προϋπαρχόντων μετρικών αξιολόγησης. Στο αρχικό επίπεδο της λίστας, οι τύποι σφαλμάτων είναι οι Ορολογία, Ακρίβεια, Γλωσσολογικές συμβάσεις(προηγούμενως Ευφράδεια), Στυλ, Τοπικές συμβάσεις, Καταλληλότητα κοινού(προηγούμενως Πιστότητα), Σχεδιασμός και Σήμανση. Κάτω όμως από αυτές τις αρχικές κατηγορίες, υπάρχει ένα δεύτερο επίπεδο υποκατηγοριών. Για παράδειγμα υποκατηγορίες του τύπου Ορολογία αποτελούν οι Λάθος ορολογία, η Ασυνεπής χρήση ορολογίας αλλά και η Ασυνέπεια ορολογίας σε σχέση με την πηγή της. Σε κάθε σφάλμα αποδίδεται ένας βαθμός σοβαρότητας(π.χ. “Minor”, “Major”, “Critical”). Ανάλογα με τον κάθε τομέα του αντικειμένου του κειμένου, οι αξιολογητές μπορεί να επιλέξουν διαφορετικό υποσύνολο τύπου σφαλμάτων για την αξιολόγηση της μετάφρασης, κάτι που καθιστά το MQM προσαρμόσιμο σε διαφορετικές ανάγκες. Στη συνέχεια, οι αξιολογητές μπορούν να βαθμολογήσουν τα μεταφρασμένα κείμενα με βάση τους επιλεγμένους τύπους σφαλμάτων και την σοβαρότητα των σφαλμάτων, ώστε να καταλήξουν στην τελική αξιολόγηση τους.

Ενώ η ανθρώπινη αξιολόγηση εξακολουθεί ακόμη και στις μέρες μας να αποτελεί την καλύτερη λύση, ενδέχεται να παρουσιάζει ορισμένες δυσκολίες. Πρώτη είναι η αντικειμενικότητα, καθώς δύο διαφορετικοί αξιολογητές μπορεί να βαθμολογήσουν την ίδια μετάφραση διαφορετικά λόγω διαφορετικού γλωσσολογικού υπόβαθρου και εμπειριών. Φυσικά, η ανθρώπινη αξιολόγηση απαιτεί την ανάγκη απασχόλησης ανθρώπων αξιολογητών που μπορεί να είναι ακριβή και ιδιαίτερα δύσκολη, ειδικά για μεγαλύτερα έργα.

Δ.3 Αυτόματη Αξιολόγηση

Η δυνατότητα αυτοματοποιημένης αξιολόγησης μηχανικής μετάφρασης αποτελεί πλέον απαραίτητο κριτήριο για την ανάπτυξη αλγορίθμων μηχανικής μετάφρασης, καθώς μπορεί να προσφέρει γρηγορότερα έως και άμεσα αποτελέσματα με κόστος πολύ χαμηλότερο από αυτό της ανθρώπινης αξιολόγησης. Βέβαια καμία αυτόματη μετρική αξιολόγησης δεν μπορεί να αντικαταστήσει πλήρως την ανθρώπινη κρίση, η οποία αυτή τη στιγμή αποτελεί την χρυσή αναφορά.

Δ.3.1 WER (Word Error Rate)

Η WER αποτελεί μία από τις πρώτες αυτοματοποιημένες μετρήσεις που χρησιμοποιήθηκαν στην αξιολόγηση μηχανικής μετάφρασης και αποτελεί την standard μετρική αξιολόγησης στον τομέα της αυτόματης αναγνώρισης προφορικού λόγου. Βέβαια, καθώς στην μετάφραση μπορεί να είναι σωστά παραπάνω από ένα αποτελέσματα, σε αντίθεση με την αναγνώριση προφορικού λόγου όπου η σωστή απάντηση είναι πάντα μόνο μία, είναι λογικό οι ίδιες μετρικές να μην είναι κατάλληλες και για τους δύο τομείς. Η WER δεν θεωρείται πλέον αρκετά αξιόπιστη μέθοδος για την αξιολόγηση μηχανικής μετάφρασης καθώς δεν μπορεί να λειτουργήσει με περισσότερες από μία προτάσεις ως references αλλά και επειδή αδυνατεί να αναγνωρίσει την αλλαγή της σειράς λέξεων και φράσεων στη μετάφραση[18].

Δ.3.2 BLEU

Η BLEU είναι μια μέθοδος αξιολόγησης μηχανικής μετάφρασης που δεν εξαρτάται από τη γλώσσα της μετάφρασης. Επίσης είναι γρήγορη και αποτελεσματική με χαμηλό κόστος[19]. Είκοσι χρόνια μετά την δημοσίευση του αρχικού άρθρου η μέτρηση BLEU χρησιμοποιείται ακόμη για την αξιολόγηση μεταφράσεων. Μάλιστα η αρχική δημοσίευση έχει βραβευτεί με το “NAACL Test of Time award”[20]. Η μέθοδος αυτή

δέχεται τις πιθανές μεταφράσεις(Candidates) και τις αξιολογεί συγκρίνοντας τες με ανθρώπινες μεταφράσεις(References). Για τον υπολογισμό της BLEU, αρχικά επιλέγετε ο αριθμός λέξεων σε σειρά, με βάση τον οποίο θα γίνει η αξιολόγηση (ή οι αριθμοί λέξεων σε σειρά). Π.χ. για δύο λέξεις σε σειρά μετράμε τα 2-gramms ή αλλιώς bigramms έτσι ώστε να μετρήσουμε κατά πόσο πλησιάζει η μηχανική μετάφραση την ανθρώπινη.

Παράδειγμα :

Candidates

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

References

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Για την Candidate 1, η μέθοδος θα εντοπίσει όλα τα 2-gramms της πρότασης.

It is, is a, a guide, guide to, to action, action which, which ensures, ensures that, that the, the military, military always, always obeys, obeys the, the commands, commands of, of the , the party

Επομένως έχουμε ένα σύνολο από δεκαεπτά διαφορετικά bigramms. Από αυτά, εάν αναζητήσουμε και στα τρία references, μπορούμε να βρούμε τα δέκα, επομένως έχουμε bigramm precision = 10/17

Ομοίως, για Candidate 2, έχουμε δεκατρία 2-gramms, από τα οποία μπορούμε να βρούμε μόνο ένα bigramm(it is) στα references. Οπότε η πρόταση Candidate 2 έχει BLUE Score = 1/13

4.3.3 METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Ενώ αναπτύχθηκε ως μία εναλλακτική λύση για την BLEU, η METEOR λαμβάνει υπόψιν τη στελέχωση(stemming) αλλά και την σειρά των λέξεων. Επίσης επιχειρεί να αντιμετωπίσει κάποιες από τις αδυναμίες της BLEU όπως την αδυναμία αντιστοίχισης συνώνυμων λέξεων μεταξύ reference και μεταφρασμένων προτάσεων[21].

4.3.4 Translation Error Rate

Η TER μετράει τον αριθμό των διορθώσεων που θα απαιτούνταν να πραγματοποιηθούν από έναν άνθρωπο ώστε μια μεταφρασμένη πρόταση να μετατραπεί σε μία πρόταση από το σετ των references. Λαμβάνει υπόψιν τις εισαγωγές, διαγραφές, αντικαταστάσεις και εναλλαγές[22]. Η μέτρηση TER προκύπτει από την παρακάτω εξίσωση :

$$TER = \frac{\text{Αριθμός Αλλαγών}}{\text{Μέσος Αριθμός Λέξεων ανά πρόταση των References}}$$

Κάθε αλλαγή υπολογίζεται σαν μία, ανεξάρτητα από το αν αποτελεί εισαγωγή, διαγραφή, αντικατάσταση ή εναλλαγή. Η TER μπορεί να υπολογιστεί είτε όταν υπάρχει μόνο μία πρόταση ως reference, είτε πολλαπλές. Στην περίπτωση των πολλαπλών reference προτάσεων, η TER υπολογίζεται για κάθε μία από αυτές, και τελικά

αυτή που λαμβάνεται υπόψιν είναι αυτή με τη χαμηλότερη TER μέτρηση. Εναλλακτικά η TER μπορεί να συνδυαστεί με ταίριασμα συνωνύμων και stemming, βελτιώνοντας έτσι ακόμη περισσότερο τα αποτελέσματα της. Αυτή η παραλλαγή της ονομάζεται TERp[18].

4.3.5 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Ενώ αρχικά αναπτύχθηκε για την αξιολόγηση περιλήψεων, η ROUGE χρησιμοποιείται και ως μέθοδος αξιολόγησης μηχανικής μετάφρασης, ειδικά σε περιπτώσεις που το πηγαίο κείμενο μπορεί να καταλήξει σε πολλές διαφορετικές σωστές μεταφράσεις. Όπως και η BLEU βασίζεται στον αριθμό των κοινών n-grams αλλά εστιάζει στην ανάκληση [23].

4.3.6 CHRF (Character n-gram F-score)

Η CHRF βασίζεται στον αριθμό κοινών N-grams μεταξύ του αποτελέσματος της μηχανικής μετάφρασης και των reference προτάσεων, και υπολογίζει τα F-Scores (harmonic mean of precision and recall) σε επίπεδο χαρακτήρων, που μπορεί να φανεί πολύ χρήσιμο στην περίπτωση μορφολογικά πλούσιων γλωσσών[24].

4.3.7 BERTScore

Η BERTScore αποτελεί μια πιο πρόσφατη μετρική αξιολόγησης, που χρησιμοποιεί σημασιολογικά embeddings από το μοντέλο BERT για να αξιολογήσει τις ομοιότητες μεταξύ μηχανικά μεταφρασμένων προτάσεων και ανθρώπινα μεταφρασμένων προτάσεων. Σε κάποια σενάρια δείχνει να συμβαδίζει περισσότερο με την ανθρώπινη αξιολόγηση σε σχέση με παλαιότερες μετρικές όπως η BLEU[25].

4.3.8 COMET

Η COMET αποτελεί ένα framework νευρωνικής αξιολόγησης μηχανικής μετάφρασης βασισμένο στο PyTorch το οποίο μπορούν άλλοι ερευνητές να παραμετροποιήσουν περαιτέρω ώστε να καταλήξουν σε δικές τους μετρικές νευρωνικής αξιολόγησης. Προκύπτει από την εκπαίδευση σε μεγάλα σετ δεδομένων τα οποία περιλαμβάνουν πηγαίες και μεταφρασμένες προτάσεις, candidate προτάσεις αλλά και scores που έχουν προκύψει από ανθρώπινες αξιολογήσεις[26]. Αρχικά, χρησιμοποιώντας ένα προεκπαιδευμένο μοντέλο(π.χ. το XLM-RoBERTa), εξάγονται τα embeddings για κάθε τμήμα κειμένου από τις πηγαίες προτάσεις, τις μεταφράσεις που προέκυψαν από τη μηχανική μετάφραση(Candidates) αλλά και τα αποτελέσματα της ανθρώπινης μετάφρασης(References). Στη συνέχεια αυτά τα embeddings, επεξεργάζονται από ένα pooling layer, για να δημιουργήσουμε νέα embeddings για κάθε κομμάτι κειμένου, τα οποία τελικά χρησιμοποιούνται για την τελική πρόβλεψη του score συσχέτισης με την ανθρώπινη αξιολόγηση. Έτσι τα μοντέλα που έχουν βασιστεί στο COMET framework μαθαίνουν να προβλέπουν την ανθρώπινη αξιολόγηση μιας μηχανικής μετάφρασης. Γνωστές παραλλαγές του COMET αποτελούν η COMMET-22[27] και η COMETkiwi[28]. Η διαφορά των δύο τελευταίων είναι ότι η COMETkiwi κάνει χρήση μόνο πηγαίου κειμένου, ενώ η COMET-22 χρησιμοποιεί και μεταφρασμένο κείμενο[29].

4.3.9 Σχήματα μάθησης

Άλλη μία μέθοδο αυτόματης αξιολόγησης μηχανικής μετάφρασης αποτελεί αυτή με τη χρήση σχημάτων μάθησης, όπως το NoDeeLe, το οποίο συγκρίνει τα αποτελέσματα δύο συστημάτων νευρωνικής μηχανικής μετάφρασης συνδυάζοντας τα με το πηγαίο κείμενο, τις ανθρώπινες μεταφράσεις(Reference Translations) και με προϋπάρχουσες αυτόματες μετρικές αξιολόγησης. Το NoDeeLe, αρχικά μετατρέπει τα μεταφρασμένα

κείμενα(είτε candidates είτε references) σε αριθμητικά διανύσματα τα οποία στη συνέχεια εισάγει, σε συνδυασμό με τις αυτόματες μετρικές αξιολόγησης (BLEU, METEOR, WER), καθώς και ένα σενάριο από features σε ένα βαθύ τεχνητό νευρωνικό δίκτυο με σκοπό να το εκπαιδεύσει και τελικά να πετύχει αποτελέσματα που συμβαδίζουν με την ανθρώπινη αξιολόγηση ως και κατά 85% [30].

E Έρευνα

E.1 Εργαλεία που χρησιμοποιήθηκαν

Το μεγαλύτερο μέρος του κώδικα υλοποιήθηκε στη γλώσσα Python3, και με τη χρήση διαφόρων βιβλιοθηκών συμπεριλαμβανομένων των: Tensorflow[31](συμπεριλαμβάνει και τη βιβλιοθήκη Keras[32]), nltk[33], sklearn[34], fastText[35], numpy[36], pandas[37]. Επίσης χρησιμοποιήθηκε η git για version control και κάποιες bash εντολές για την εύκολη αφαίρεση κάποιων ειδικών χαρακτήρων. Φυσικά, πολλά ερωτήματα τεχνικής φύσεως που γεννήθηκαν κατά την υλοποίηση, βρήκαν απάντηση σε διάφορα online forums, ή ιστοσελίδες όπως το StackOverflow[38]. Επίσης για το σχεδιασμό διαφορών γραφημάτων χρησιμοποιήθηκε η offline έκδοση του εργαλείου draw.io[39]. Για τη συγγραφή της εργασίας χρησιμοποιήθηκαν οι εφαρμογές του LibreOffice (Writer και Calc) σε συνδυασμό με το πρόσθετο Zotero για τη διαχείριση της βιβλιογραφίας.

E.2 Για το αρχικό Dataset

Το dataset που χρησιμοποιήθηκε αποτελείται από τρία παράλληλα κείμενα στην Ελληνική γλώσσα που προέκυψαν από την αντίστοιχη μετάφραση στην Αγγλική γλώσσα. Το καθένα περιέχει ελληνικές μεταφράσεις 2686 αγγλικών προτάσεων.

- 1.data_gr_for_keras_S1.csv
- 2.data_gr_for_keras_S2.csv
- 3.data_gr_for_keras_R.csv

Τα τρία αυτά CSV αποτελούν το Dataset, περιέχοντας στο πρώτο κελί κάθε γραμμής ολόκληρη την πρόταση. Το S1 περιέχει τα αποτελέσματα που προέκυψαν από τη μετάφραση του στατιστικού μοντέλου, το S2 περιέχει τα αποτελέσματα που προέκυψαν από τη μετάφραση του νευρωνικού μοντέλου και το R περιέχει τα αποτελέσματα που προέκυψαν από ανθρώπινη μετάφραση(References). Φυσικά κάποιες λέξεις δεν έχουν πάρει κάποια μετάφραση στην Ελληνική αλλά έχουν παραμείνει στο σετ δεδομένων με την αρχική τους μορφή στην Αγγλική γλώσσα(π.χ. REST, αναφέρεται στο REST API). Συνολικά το dataset, πριν την μετατροπή των κεφαλαίων γραμμμάτων σε πεζά και την αφαίρεση ειδικών χαρακτήρων, αποτελείται από 13654 διαφορετικές λέξεις. Στο παράρτημα Συχνότερα Εμφανιζόμενες Λέξεις πριν την Προεπεξεργασία φαίνονται οι είκοσι συχνότερα εμφανιζόμενες λέξεις με την λέξη ‘να’ να έχει 4008 εμφανίσεις. Μετά την μετατροπή όλων των κεφαλαίων γραμμμάτων σε πεζά και την αφαίρεση ειδικών χαρακτήρων, το dataset αποτελείται πλέον από 12.589 διαφορετικές λέξεις, από τις οποίες πολλές επαναλαμβάνονται. Οι περισσότερες λέξεις εμφανίζονται από μία φορά, ενώ συχνότερα απ’ όλες εμφανίζεται η λέξη ‘να’(4031 φορές), που αρχικά δεν συμπεριλαμβανόταν στα stopwords του πακέτου NLTK. Οι περισσότερες λέξεις εμφανίστηκαν από μία φορά(5.237 λέξεις), από δύο φορές εμφανίστηκαν 2.202 λέξεις, από τρεις φορές εμφανίστηκαν 1.420 λέξεις και από τέσσερις φορές εμφανίστηκαν 672 λέξεις. Αριθμούς εμφανίσεων, από 253 έως και 4.031 φορές έκαναν 53 λέξεις. Στο σχήμα Κατανομή Αριθμού Εμφανίσεων Λέξεων (χωρίς τις ακραίες περιπτώσεις) φαίνεται η κατανομή των εμφανίσεων των λέξεων, χωρίς τις ακραίες περιπτώσεις που αναφέρονται παραπάνω, ώστε το ραβδόγραμμα να είναι πιο ευανάγνωστο από το Κατανομή Αριθμού Εμφανίσεων των Λέξεων (με τις ακραίες περιπτώσεις). Βλέποντας τα παραπάνω διαγράμματα σε συνδυασμό με το Κατανομή Εμφανίσεων κάθε Λέξης (όλες οι λέξεις) παρατηρείται η ανισότητα των εμφανίσεων των λέξεων μεταξύ τους. Υπάρχει επίσης ένα τέταρτο αρχείο, το data_gr_for_keras_ANNOT1.csv το οποίο περιλαμβάνει την εκτίμηση ειδικών γλωσσολόγων για την κοντινότερη στην ανθρώπινη μετάφραση σε κάθε περίπτωση. Στις περισσότερες περιπτώσεις θεωρήθηκε ότι η μετάφραση του νευρωνικού μοντέλου ήταν κοντινότερη. Αυτό σημαίνει ότι για τη σύγκριση των αποτελεσμάτων με αυτά του αρχείου, αρκεί να έχει προτιμηθεί σε όλες τις περιπτώσεις η μετάφραση του νευρωνικού μοντέλου ως “κοντινότερη” στην ανθρώπινη(ZeroR) ώστε να επιτευχθεί συμφωνία με την

ανθρώπινη αξιολόγηση σε ποσοστό 62.59%. Βέβαια, καθώς για την περίπτωση των νευρωνικών δικτύων απαιτείται ο διαχωρισμός του σετ δεδομένων σε σετ αξιολόγησης και σετ εκπαίδευσης, επιβάλλεται να διερευνηθούν τα αποτελέσματα του ZeroR και στα υποσύνολα του σετ δεδομένων που προκύπτουν. Για το δεύτερο μισό του σετ δεδομένων, με τον ZeroR ξανά με την επιλογή των μεταφράσεων του νευρωνικού μοντέλου ως κοντινότερες στην ανθρώπινη, επιτυγχάνεται συμφωνία με την ανθρώπινη αξιολόγηση στο 60.24%. Αντίστοιχα για το τελευταίο 10% των δεδομένων, με τον ZeroR επιτυγχάνεται συμφωνία στο 53.7%, αυτή τη φορά όμως με την επιλογή των μεταφράσεων του στατιστικού μοντέλου ως κοντινότερες στην ανθρώπινη. Στον πίνακα Αποτελέσματα ZeroR φαίνονται όλα τα αποτελέσματα του ZeroR για το σύνολο και τα υποσύνολα του σετ δεδομένων και εκεί μπορεί να παρατηρηθεί ότι και στα δύο υποσύνολα του αρχικού σετ, ο ZeroR επιτυγχάνει χαμηλότερη ορθότητα. Επίσης κάτι που πρέπει να ληφθεί υπόψιν είναι το γεγονός ότι στο τελευταίο 10% του σετ δεδομένων ο ZeroR επιτυγχάνει την υψηλότερη βαθμολογία με διαφορετική επιλογή από αυτήν που επιτυγχάνει την υψηλότερη βαθμολογία στο 100% και στο δεύτερο 50% του συνόλου. Ταυτόχρονα, για την υλοποίηση της προσέγγισης με την εκπαίδευση του TND, δεν χρησιμοποιήθηκε κάποιος αλγόριθμος τυχαίας επιλογής για το διαχωρισμό του σετ εκπαίδευσης και του σετ αξιολόγησης. Από το συνδυασμό των παραπάνω γεγονότων προκύπτει το συμπέρασμα ότι στην περίπτωση του διαχωρισμού του σετ σε ποσοστό 9 προς 1, το κάθε μοντέλο θα έχει εκπαιδευτεί σε ένα σετ εκπαίδευσης που είναι ως ένα βαθμό προκατειλημμένο προς τις μεταφράσεις του νευρωνικού δικτύου ενώ το σετ αξιολόγησης θα είναι προκατειλημμένο προς τις μεταφράσεις του στατιστικού μοντέλου.

E.3 Υλοποίηση

Η αρχική προσέγγιση γύρω από την υλοποίηση αποτελούσε τον χωρισμό των εργασιών σε δύο κύριες διαδικασίες. Πρώτα αυτή της προεπεξεργασίας/φιλτραρίσματος του αρχικού σετ δεδομένων, και στην συνέχεια αυτή της δημιουργίας μετρήσεων για κάθε πρόταση και τελικά την αξιολόγηση. Καθώς τα δύο μέρη αυτά μπορούν κάποιες φορές να λειτουργήσουν ανεξάρτητα, έγιναν προσπάθειες ώστε να παραχθούν διάφοροι συνδυασμοί αυτών, παρατηρώντας έτσι διαφορετικά αποτελέσματα. Στο σχήμα Flow όλων των αρχείων και scripts που χρησιμοποιήθηκαν φαίνεται ένα υπόδειγμα της ροής και των διαφορετικών scripts που χρησιμοποιήθηκαν μέχρι να επιτευχθεί το κάθε τελικό αποτέλεσμα.

E.3.1 Επεξεργασία και Προεπεξεργασία Δεδομένων

Για να ξεπεραστούν προβλήματα με το encoding, το περιεχόμενο των τριών αρχικών αρχείων μεταφέρθηκε από μορφή csv σε text. Από αυτό το σημείο και μετά, όλη η υλοποίηση βασίστηκε σε αυτά τα αρχεία. Καθώς οι προτάσεις των γραμμών 1613 και 1619 δημιουργούσαν πρόβλημα στην επεξεργασία των δεδομένων, αφαιρέθηκαν τελείως από το set δεδομένων.

Διάφορες τεχνικές που χρησιμοποιήθηκαν κατά την επεξεργασία των δεδομένων αποτελούσαν οι :

- Μετατροπή όλων των κεφαλαίων γραμμάτων σε πεζά. (Συμβολίζεται με I(πεζό αγγλικό L) στην στήλη “Σετ Δεδομένων” στο παράρτημα BOW-TND - 20 Υψηλότερες Βαθμολογίες)
- Αφαίρεση λέξεων που εμφανίζονται σπάνια στο σύνολο του dataset. (όπου $A > x$ στην στήλη “Σετ Δεδομένων” στο παράρτημα BOW-TND - 20 Υψηλότερες Βαθμολογίες)
- Αφαίρεση λέξεων που εμφανίζονται πολύ συχνά στο σύνολο του dataset.(όπου $A < x$ υποδεικνύει ότι όσες λέξεις έκαναν x ή περισσότερες από x εμφανίσεις, αφαιρέθηκαν από το dataset)
- Αφαίρεση των ελληνικών stopwords της NLTK βιβλιοθήκης από το σετ δεδομένων. Μπορεί να παρατηρηθεί ότι πολλές από τις λέξεις στη λίστα των stopwords δεν τονίζονται. Καθώς στο set δεδομένων που παρέχεται, οι αντίστοιχες λέξεις είναι τονισμένες, οι παραπάνω δεν αναμένεται από μόνες τους να δημιουργήσουν σημαντική διαφορά.(Συμβολίζεται με No StopWords, στη στήλη “Σετ Δεδομένων” στο παράρτημα BOW-TND - 20 Υψηλότερες Βαθμολογίες)

- Προσθήκη νέων stopwords. Επειδή πολλά από τα αρχικά stopwords δεν εμφανίζονται στο dataset ακριβώς με την μορφή που δίνεται από το πακέτο NLTK, έγινε μια προσπάθεια να συμπεριληφθούν νέα stopwords με κάποιες διορθώσεις στον τονισμό τους. (Συμβολίζεται με No StopWords+, στη στήλη “Σετ Δεδομένων” στο παράρτημα BOW-TNΔ - 20 Υψηλότερες Βαθμολογίες)

E.3.2 Επισημείωση μετρήσεων και αξιολόγησης αυτών

Για τη σύγκριση των διάφορων μετρήσεων με την ανθρώπινη αξιολόγηση, η κάθε τεχνική που χρησιμοποιήθηκε, υπολόγιζε ποια από τις δύο μεταφράσεις έμοιαζε περισσότερο με την ανθρώπινη. Στη συνέχεια γίνεται σύγκριση των αποτελεσμάτων της μεθόδου, με το αρχείο `data_gr_for_keras_ANNOT1`, έτσι ώστε να υπολογιστεί η ορθότητα (accuracy) των αποτελεσμάτων.

E.3.2.1 Μέτρηση Bleu

Μία από τις μετρήσεις που χρησιμοποιήθηκαν για την αξιολόγηση της μηχανικής μετάφρασης, και στη συνέχεια συγκρίθηκαν με την ανθρώπινη αξιολόγηση ήταν και η BLEU. Το αρχείο `bleu.py` δέχεται σαν είσοδο τα παράλληλα κείμενα σε μορφή CSV και υπολογίζει το bleu score για την κάθε μετάφραση του στατιστικού μοντέλου και την κάθε μετάφραση του νευρωνικού μοντέλου σε σύγκριση με την ανθρώπινη μετάφραση ως reference, χρησιμοποιώντας την μέθοδο `sentence_bleu` που έρχεται μαζί με τη βιβλιοθήκη `nlk`. Αναλόγως με το ποια μετάφραση είχε υψηλότερο score, γίνεται και η ανάλογη καταγραφή σε ένα νέο αρχείο, ώστε αυτό να συγκριθεί στη συνέχεια με το αρχείο `data_gr_for_keras_ANNOT1` που προέκυψε από την ανθρώπινη αξιολόγηση ως προς την ακρίβεια. Σε πολλές περιπτώσεις, οι δύο μεταφράσεις μπορεί να είχαν ισόβαθμο bleu score. Όταν για αυτές τις περιπτώσεις, επιλέχθηκε ως κοντινότερη η μετάφραση του στατιστικού μοντέλου (1) τότε επιτεύχθηκε και η υψηλότερη ακρίβεια (accuracy). Το ποσοστό αυτό ήταν 65.68% σε συνδυασμό με την αφαίρεση των αρχικών stopwords και τη χρήση μόνο πεζών γραμμάτων, υπολογίζοντας το bleu score για όλα τα 1-gramms και 2-gramms των προτάσεων (όλες οι δοκιμές με την BLEU φαίνονται στον πίνακα Αποτελέσματα BLEU). Φυσικά η BLEU με bigramms αξιολογήθηκε και πάνω στα δύο υποσύνολα του σετ δεδομένων. Στο σχήμα Τεχνική Ροή Δημιουργίας και Αξιολόγησης Μέτρησης BLEU μπορούμε να δούμε την διαδικασία που ακολουθήθηκε για τη δημιουργία των τελικών αποτελεσμάτων.

E.3.2.2 Μέτρηση ROUGE

Για τον υπολογισμό της μέτρησης ROUGE δημιουργήθηκε το script `rouge_implement.py` στο οποίο δηλώνεται η `calculate_rouge` function που υπολογίζει το score μεταξύ δύο προτάσεων που δέχεται σαν είσοδο (hypothesis, reference) χρησιμοποιώντας την `get_scores` μέθοδο. Στη συνέχεια, η `main` function δέχεται τα τρία σύνολα μεταφράσεων, και εκμεταλλεύομενη την `calculate_rouge` function που δηλώθηκε νωρίτερα, υπολογίζει τα Rouge Scores (Rouge-1, Rouge-2, Rouge-L) για τις στατιστικές και νευρωνικές μεταφράσεις, δημιουργώντας έτσι ένα τελικό αρχείο το οποίο θα συγκριθεί με την ανθρώπινη αξιολόγηση. Η υψηλότερη ακρίβεια (accuracy) επιτεύχθηκε χρησιμοποιώντας τη ρύθμιση `["rouge-1"]` `["f"]` (l για longest common subsequence και f για f1 score) η οποία ήταν 66.24%. Τα αποτελέσματα φαίνονται στον πίνακα Αποτελέσματα Μετρήσεων ROUGE. Φυσικά η ρύθμιση που πέτυχε την υψηλότερη ορθότητα αξιολογήθηκε και πάνω στα δύο υποσύνολα του σετ δεδομένων, ώστε να συγκριθεί αργότερα με τα αποτελέσματα του TNΔ. Στο διάγραμμα ροής Τεχνική Ροή Δημιουργίας και Αξιολόγησης Μέτρησης ROUGE φαίνεται η διαδικασία με βάση την οποία προέκυψαν τα τελικά αποτελέσματα της συγκεκριμένης μεθοδολογίας.

E.3.2.3 FastText

Για την υλοποίηση της αξιολόγησης με τη χρήση του FastText, χρησιμοποιήθηκαν τα αρχικά αρχεία σε μορφή κειμένου σαν input στο ανάλογο script(`fastText_valuation.py`), στο οποίο εφαρμόζεται η χρήση του προεκπαιδευμένου μοντέλου που προσφέρεται από το FastText για την ελληνική γλώσσα(`cc.el.300.bin`). Με βάση αυτό το pretrained μοντέλο και τη μέθοδο `get_sentence_vector` γίνεται εξαγωγή ενός διανύσματος για κάθε πρόταση του dataset. Η μέθοδος `get_sentence_vector` επιστρέφει το διάνυσμα της πρότασης που δέχεται σαν παράμετρο, βρίσκοντας το μέσο όρο των διανυσμάτων των λέξεων της πρότασης αυτής. Στη συνέχεια, το κάθε διάνυσμα που προέκυψε από τις προτάσεις της στατιστικής και της νευρωνικής μετάφρασης, συγκρίνεται με αυτό που προέκυψε από την αντίστοιχη πρόταση της ανθρώπινης μετάφρασης. Για να πραγματοποιηθεί αυτή η σύγκριση, χρησιμοποιήθηκε η μέθοδος `get_cosine_similarity` της βιβλιοθήκης `sklearn`. Αυτή η μέθοδος επιστρέφει την ομοιότητα συνημίτονου μεταξύ δύο διανυσμάτων. Έτσι δημιουργείται ένα τελικό αρχείο το οποίο συμπληρώνεται με 0 και 1, ανάλογα με το ποιο διάνυσμα είχε υψηλότερη ομοιότητα συνημίτονου με το διάνυσμα που προκύπτει από την ανθρώπινη μετάφραση. Τέλος αυτό το αρχείο συγκρίνεται με το αρχείο που έχει προκύψει από την ανθρώπινη αξιολόγηση (`data_gr_for_keras_ANNOT1`). Τα αποτελέσματα συμφωνούν σε μεγαλύτερο βαθμό με τα ανθρώπινα(65.13%), αφού έχει γίνει μετατροπή όλων των γραμμών σε πεζά, και έχει επιλεγεί ως “σωστή” η μετάφραση του νευρωνικού μοντέλου στις περιπτώσεις όπου τα δύο διανύσματα είχαν την ίδια ομοιότητα συνημίτονου με το διάνυσμα της ανθρώπινης μετάφρασης. Φυσικά έχει γίνει η αντίστοιχη αξιολόγηση και στα υποσύνολα του σετ δεδομένων. Όλα τα αποτελέσματα που προέκυψαν από τις μετρήσεις με τη ROUGE φαίνονται στον πίνακα Αποτελέσματα FastText και Ομοιότητας Συνημίτονου. Στο σχήμα Τεχνική Ροή Δημιουργίας και Αξιολόγησης Μετρήσεων από το fastText φαίνεται η ροή που οδήγησε στα τελικά αποτελέσματα.

E.3.3 Bag of Words σε Νευρωνικό Δίκτυο

E.3.3.1 Δημιουργία σετ δεδομένων σε μορφή Bag of Words

Αυτή η υλοποίηση αποτελούταν από δύο κύρια μέρη τα οποία δεν συνδυάστηκαν με τα προαναφερόμενα αλλά μόνο μεταξύ τους. Το πρώτο αποτελεί την δημιουργία της Bag of Words με βάση το αρχικό σετ δεδομένων. Το δεύτερο περιλαμβάνει την χρήση ενός νευρωνικού δικτύου που δέχεται το σετ δεδομένων σε μορφή B.O.W. σαν είσοδο, και τη χρησιμοποιεί για να προβλέψει εάν η στατιστική ή η νευρωνική μετάφραση είναι κοντινότερη στην ανθρώπινη. Για να επιτευχθεί η τελική μορφοποίηση του σετ δεδομένων σε μορφή bag of words, τα αρχικά αρχεία κειμένου μετατράπηκαν σε μορφή CSV όπου κάθε κελί περιείχε μία λέξη της πρότασης. Στη συνέχεια, έπρεπε να συνδυαστούν όλα μαζί σε ένα pandas Dataframe από το οποίο έγινε εξαγωγή των διαφορετικών λέξεων σε μία λίστα που ονομάστηκε Dictionary. Στη συνέχεια δημιουργήθηκε ένας numpy πίνακας διαστάσεων $X \times Y$, όπου X το πλήθος των διαφορετικών λέξεων του Dataset και Y το πλήθος των προτάσεων συνολικά. Αυτό το κομμάτι αρχικά είχε υλοποιηθεί επίσης με pandas Dataframe, αλλά με την αλλαγή σε numpy array ξεπεράστηκαν προβλήματα που εμφανίστηκαν λόγω έλλειψης μνήμης. Εδώ η κάθε κολώνα συμβολίζει μία λέξη του dataset, και η τιμή στο κελί το πόσες φορές εμφανίζεται αυτή η λέξη στην αντίστοιχη πρόταση. Αυτό επιτεύχθηκε ελέγχοντας μία μία τις τιμές του συνδυαστικού dataframe που δημιουργήθηκε νωρίτερα, και αυξάνοντας κάθε φορά κατά μία μονάδα, την τιμή του αντίστοιχου κελιού στο numpy array για την συγκεκριμένη πρόταση και λέξη. Ταυτόχρονα μπορούν να οριστούν κατώφλια απαιτούμενων εμφανίσεων μιας λέξης έτσι ώστε να αποφευχθεί η εκπαίδευση χρησιμοποιώντας λέξεις που εμφανίζονται πολύ συχνά ή πολύ σπάνια στο σετ δεδομένων. Έτσι δημιουργείται ένα DataFrame που περιέχει ακέραιους σε κάθε κελί ξεκινώντας από το 0. Στη συνέχεια ξαναχωρίζεται σε τρία μέρη (ένα για κάθε μετάφραση), έτσι ώστε να εισαχθεί σαν είσοδος στο νευρωνικό δίκτυο που θα εκπαιδευτεί ώστε να αναγνωρίζει ποια από τις δύο μεθόδους μετάφρασης

είναι κοντινότερη στην ανθρώπινη. Αυτό γίνεται στο επόμενο script, όπου κάθε τριάδα μεταφράσεων σε μορφή Bag Of Words εισάγεται ταυτόχρονα σαν ένα record στο νευρωνικό δίκτυο με το απαραίτητο label από το data_gr_for_keras_ANNOT1 αρχείο.

Για παράδειγμα, ένα dataset που αποτελείται μόνο από τις τρεις παρακάτω παράλληλες προτάσεις:

S1 : Με την έκφραση, σειρά μαθημάτων

S2 : Στην έκφραση “Κριτική Σκέψη”

R : Στην έκφραση “Κριτική Σκέψη”

Το λεξικό που προκύπτει από το παραπάνω σετ δεδομένων, φαίνεται στον παρακάτω πίνακα. Ο αριθμός που συνοδεύει την κάθε λέξη δείχνει τον αριθμό της στήλης που θα αντιστοιχεί στην συγκεκριμένη λέξη στην τελική μορφή Bag Of Words του σετ δεδομένων.

στην	έκφραση	κριτική	σκέψη	με	την	σειρά	μαθημάτων
0	1	2	3	4	5	6	7

Με βάση τα παραπάνω, η μορφή BOW για την κάθε πρόταση που προκύπτει φαίνεται παρακάτω
S1 : 0,0,0,0,1,1,1,1 S2 : 1,1,1,1,0,0,0,0 R:1,1,1,1,0,0,0,0

Έτσι, το input που εισάγεται στο νευρωνικό δίκτυο, αποτελείται από την ένωση των τριών παραπάνω προτάσεων $[[0,0,0,0,1,1,1,1], [1,1,1,1,0,0,0,0], [1,1,1,1,0,0,0,0]]$ συνοδευόμενη από το label 0 που υποδεικνύει ότι η πρόταση του S2 είναι κοντινότερη στην ανθρώπινη μετάφραση από την πρόταση του S1. Καθώς το λεξικό αποτελείται από συνολικά 12.349 διαφορετικές λέξεις, το μήκος της κάθε ξεχωριστής πρότασης θα συμβολίζεται από τουλάχιστον τόσα ψηφία(λιγότερα όταν χρησιμοποιούνται φίλτρα για την μείωση του όγκου των δεδομένων, όπως η αφαίρεση σπάνια εμφανιζόμενων λέξεων). Φυσικά, η κάθε γραμμή της BOW δομής, αποτελείται από τρεις προτάσεις, οπότε το κάθε record θα έχει $3 \times 12.349 = 37.049$ μεταβλητές. Επομένως, μόνο το αρχικό layer θα έχει 37.049 νευρώνες. Με βάση τα παραπάνω, το τελικό σχήμα της BOW θα είναι ένας δισδιάστατος πίνακας διαστάσεων 37.049×2.686 , όπου 2.686 οι συνολικές τριάδες προτάσεων.

E.3.3.2 Υλοποίηση και Εκπαίδευση Τεχνητού Νευρωνικού Δικτύου

Καθώς η βιβλιοθήκη tensorflow προσφέρει πολλαπλές επιλογές για τον τρόπο με τον οποίο στήνεται και εκπαιδεύεται ένα νευρωνικό δίκτυο, ο κώδικας που γράφτηκε υλοποιεί τις παραπάνω διαδικασίες μέσω ενός βρόχου αλλάζοντας κάθε φορά τις παραμέτρους που σχετίζονται με την συνάρτηση ενεργοποίησης για κάθε κρυφό στρώμα, τον αριθμό των νευρώνων σε κάθε κρυφό στρώμα, τον αριθμό των εποχών για τις οποίες θα εκπαιδευτεί το ΤΝΔ, αλλά και το μέγεθος του σετ εκπαίδευσης. Για την συνάρτηση ενεργοποίησης των κρυφών στρωμάτων έγιναν δοκιμές με τρεις διαφορετικές επιλογές, την ReLU, LeakyReLU και tanh. Όσον αφορά το διαχωρισμό του corpus σε σετ εκπαίδευσης και τεστ σετ, δοκιμάστηκαν και ο διαχωρισμός κατά το ήμισυ, αλλά και σε ποσοστό εννιά προς ένα. Η εκπαίδευση κάθε φορά γινόταν μεταξύ τεσσάρων και επτά εποχών. Τέλος τα κρυφά στρώματα μπορεί να αποτελούνταν από 128, 256, 1024, 2048 ή 3840 νευρώνες.

```
#####Parameterization#####
runNumber=0
activ_Functions=['tanh','relu','LeakyReLU']
lengthsoftrainSet=[1343,2416]#2416=(90/10) #1343=(50/50)
```

```
epochChoices=[6,4,5,7]
deepLayerNeurons=[128,256,1024,2048,3840]
#####
```

Υπόδειγμα Κώδικα Α: Παράμετροι δημιουργίας και εκπαίδευσης ΤΝΔ

Πολλαπλασιάζοντας τις παραπάνω διαφορετικές μεταβλητές προκύπτει το συμπέρασμα ότι κάθε τρέξιμο του script(nn.py) θα έχει ως αποτέλεσμα την εκπαίδευση και αξιολόγηση 120 διαφορετικών ΤΝΔ. Συνολικά το τρέξιμο του script έγινε πολλαπλές φορές για διαφορετικές μορφές στη δομή του ΤΝΔ αλλά και διαφορετικά σετ δεδομένων. Ποιο συγκεκριμένα, δοκιμάστηκαν δομές δύο και τεσσάρων κρυφών στρωμάτων, με ή χωρίς τον συνδυασμό με dropout layers. Επίσης έγιναν δοκιμές με τρία διαφορετικά dataset, από τα οποία και τα τρία είχαν επεξεργαστεί ώστε να μην περιέχουν κεφαλαία γράμματα και τα stopwords που προτείνονται από τη βιβλιοθήκη NLTK. Σε ένα από αυτά είχαν επίσης αφαιρεθεί τα νέα stopwords που δημιουργήθηκαν κατά την εκπόνηση αυτής της εργασίας και φαίνονται στο Η.2, καθώς και όσες λέξεις εμφανίστηκαν συνολικά στα τρία παράλληλα κείμενα λιγότερες από τρεις ή περισσότερες από διακόσιες φορές. Στα άλλα δύο έχουν αφαιρεθεί όσες λέξεις εμφανίστηκαν λιγότερες από πέντε φορές, και στο ένα από αυτά τα δύο, όσες λέξεις εμφανίστηκαν περισσότερες από χίλιες φορές. Όλες οι διαφορετικές ρυθμίσεις με τους μέσους όρους που προέκυψαν για την καθεμία φαίνονται στον πίνακα Μέσος όρος ορθότητας ΤΝΔ κάθε διαθέσιμης ρύθμισης. Στο Παράδειγμα δήλωσης δομής ΤΝΔ χωρίς Dropout Layers φαίνεται ένα παράδειγμα κώδικα που χρησιμοποιήθηκε για τη δημιουργία ενός νευρωνικού δικτύου με τη χρήση του πακέτου keras, έτσι ώστε να περιέχει 9.243 νευρώνες στο αρχικό στρώμα. Σε κάθε περίπτωση, οι νευρώνες στο αρχικό στρώμα εξαρτώνται από τις διαφορετικές λέξεις στο τελικό σετ δεδομένων. Έτσι προκύπτει ότι για τις διαφορετικές δοκιμές, το μέγεθος του αρχικού στρώματος μπορεί να είναι κάθε φορά διαφορετικό. Στα κρυφά στρώματα, στα οποία ο αριθμός των νευρώνων του καθενός ορίζεται από τη μεταβλητή midNeurons, δηλώνεται επίσης μία από τις τρεις διαφορετικές συναρτήσεις ενεργοποίησης, καθώς και η λέξη κλειδί Dense, που δηλώνει ότι κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες του επόμενου στρώματος. Στο τελευταίο στρώμα, το οποίο αποτελείται από μόλις ένα νευρώνα καθώς το ΤΝΔ καλείται να προβλέψει μεταξύ δύο διαφορετικών επιλογών, χρησιμοποιείται η σιγμοειδής σαν Activation Function.

```
model = keras.Sequential([
    keras.layers.Flatten(input_shape=(3,3081)),
    keras.layers.Dense(midNeurons,activation=activation_function),
    keras.layers.Dense(midNeurons, activation=activation_function),
    keras.layers.Dense(1, activation="sigmoid") ])
```

Υπόδειγμα Κώδικα Β: Παράδειγμα δήλωσης δομής ΤΝΔ χωρίς Dropout Layers

Με την ίδια λογική, έγιναν δοκιμές και με την παρακάτω δομή (Παράδειγμα δήλωσης δομής ΤΝΔ με Dropout Layers), στην οποία χρησιμοποιείται ένα dropout layer μετά από κάθε στρώμα του νευρωνικού δικτύου.

```

model = keras.Sequential([
    keras.layers.Flatten(input_shape=(3,3081)),
        keras.layers.Dense(midNeurons,activation=activation_function),
        keras.layers.Dropout(0.5),
        keras.layers.Dense(midNeurons, activation=activation_function),
        keras.layers.Dropout(0.5),
        keras.layers.Dense(midNeurons, activation=activation_function),
        keras.layers.Dropout(0.5),
        keras.layers.Dense(midNeurons, activation=activation_function),
        keras.layers.Dropout(0.5),
        keras.layers.Dense(1, activation="sigmoid") ])

```

ΥπόδειγμαΚώδικα Γ: Παράδειγμα δήλωσης δομής ΤΝΔ με Dropout Layers

Από όλες αυτές τις διαφορετικές ρυθμίσεις, προέκυψαν συνολικά 1.440 πειράματα των οποίων η μέση ορθότητα ήταν 54.47%. Τελικά το μοντέλο εκπαιδεύεται και μετά ελέγχεται ως προς την ορθότητα του, πάνω στο σετ εκπαίδευσης που έχουμε ορίσει νωρίτερα, όπως φαίνεται στο υπόδειγμα Δήλωση δημιουργίας, εκπαίδευσης και αξιολόγησης ΤΝΔ. Στην δήλωση της μεθόδου compile χρησιμοποιείται η λέξη κλειδί “adam” για την επιλογή του optimizer και η λέξη κλειδί binary_crossentropy για την επιλογή της loss function που θα χρησιμοποιηθεί κατά την εκπαίδευση και αξιολόγηση του μοντέλου. Στην παράμετρο metrics δηλώνονται οι μετρήσεις με βάση τις οποίες θα αξιολογηθεί το μοντέλο. Για τη συγκεκριμένη υλοποίηση χρησιμοποιήθηκε αρχικά η ορθότητα(accuracy), αλλά και η ακρίβεια(Precision) και η ανάκληση(Recall) κυρίως για εκπαιδευτικούς σκοπούς. Για την ακρίβεια και την ανάκληση, χρησιμοποιήθηκε ο ορισμός που δίνεται από την βιβλιοθήκη Keras.

```

model.compile(optimizer="adam",
    loss="binary_crossentropy",
    metrics=["accuracy", Precision(), Recall()])
model.fit(train_data, train_labels, epochs)
test_loss, test_acc = model.evaluate(test_data, test_labels)

```

ΥπόδειγμαΚώδικα Δ: Δήλωση δημιουργίας, εκπαίδευσης και αξιολόγησης ΤΝΔ

Τα είκοσι υψηλότερα αποτελέσματα καθώς και οι ρυθμίσεις που χρησιμοποιήθηκαν για την επίτευξη του καθενός φαίνονται στον πίνακα BOW-TNΔ - 20 Υψηλότερες Βαθμολογίες. Τα υψηλότερα αποτελέσματα, σε συμφωνία στο 63.49% με την ανθρώπινη αξιολόγηση, προέκυψαν όταν το ΤΝΔ υλοποιήθηκε με τέσσερα κρυφά στρώματα με 256 νευρώνες το καθένα, συνοδευόμενα από dropout layers και εκπαιδεύτηκε σε έξι εποχές. Επίσης στο σετ δεδομένων χρησιμοποιήθηκαν μόνο πεζά γράμματα και αφαιρέθηκαν οι λέξεις που εμφανίστηκαν λιγότερες από πέντε φορές. Καθώς σε προγενέστερες δοκιμαστικές προσπάθειες, η εκπαίδευση του ΤΝΔ μέσω του συγκεκριμένου βρόγχου οδηγούσε σε μη αναμενόμενα αλλά και μη αναπαράξιμα αποτελέσματα, έγινε χρήση της εντολής “del model” πριν το τέλος του βρόγχου αλλά και μιας συνάρτησης στην αρχή του βρόγχου για την μετατροπή των τυχαίων αριθμών που μπορεί να χρησιμοποιούν οι διάφορες βιβλιοθήκες για τυχαιότητα, σε σταθερές. Με αυτό τον τρόπο, τα αποτελέσματα κάθε πειράματος έγιναν πλήρως αναπαράξιμα.

ΣΤ Πειράματα και αποτελέσματα

Η σύγκριση όλων των αποτελεσμάτων μετρήθηκε ως προς την ορθότητα(accuracy). Τελικά, απ' όλα τα πειράματα που πραγματοποιήθηκαν κατά την εκπόνηση της συγκεκριμένης εργασίας, η μεγαλύτερη συμφωνία με την ανθρώπινη αξιολόγηση στο σύνολο του σετ δεδομένων επιτεύχθηκε με τη ROUGE(66.24%). Το νούμερο αυτό είναι μόλις τέσσερις μονάδες πάνω από το 62.59% που θα μπορούσαμε να επιτύχουμε με την χρήση του ZeroR. Με την εισαγωγή του κειμένου μορφής Bag of Words σε νευρωνικό δίκτυο επιτεύχθηκε στο μέγιστο ακρίβεια 63.49%. Στο αντίστοιχο υποσύνολο του σετ δεδομένων, με τη ROUGE επιτεύχθηκε ορθότητα σε ποσοστό 64.26% που αποτελεί την υψηλότερη που καταγράφηκε στο συγκεκριμένο υποσύνολο. Με την εξαγωγή διανυσμάτων από το ελληνικό μοντέλο του fastText η μέγιστη ορθότητα που επιτεύχθηκε στο ίδιο υποσύνολο του σετ ήταν 62.84%, ενώ στο συνολικό σετ δεδομένων ήταν 65.13%, χωρίς βέβαια να γίνει χρήση κάποιας τεχνικής κλαδέματος πέρα από την απλή αφαίρεση ειδικών χαρακτήρων.

ΣΤ.1 Πειράματα και αποτελέσματα μετρήσεων BLEU

Σε όλα τα πειράματα που έγιναν με τη μέτρηση BLEU, τα κεφαλαία γράμματα μετατράπηκαν σε πεζά. Επίσης, καθώς σε πολλαπλά παραδείγματα φάνηκαν να μην υπάρχουν συνδυασμοί από 2-gramms, 3-gramms ή 4-gramms, σε κάθε περίπτωση τα πειράματα έγιναν και με τους μικρότερους δυνατούς συνδυασμούς. Αυτό έγινε καθώς αν μια υποψήφια πρόταση candidate1 έχει κάποιο κοινό 1-gram με την πρόταση reference αλλά όχι κάποιο κοινό bi-gramm, αν υπολογίσουμε το BLEU score με βάση μόνο τα bi-gramms τότε η candidate1 θα βαθμολογηθεί με bleu_score=0. Όσο ο αριθμός απαιτούμενων λέξεων για κάθε n-gramm αυξάνεται, η εύρεση κοινών n-gramms γίνεται όλο και δυσκολότερη, με αποτέλεσμα όλο και περισσότερες προτάσεις να βαθμολογούνται με 0. Συμπεριλαμβάνοντας πάντα και τους μικρότερους συνδυασμούς n-gramms για τον υπολογισμό του BLEU score, μπορεί να αποφευχθεί αυτό το φαινόμενο. Τα αποτελέσματα όλων των πειραμάτων που έγιναν φαίνονται στον πίνακα Αποτελέσματα BLEU. Στην στήλη N-Gramms βλέπουμε τους συνδυασμούς από n-gramms που χρησιμοποιήθηκαν για τον υπολογισμό του BLEU score της κάθε υποψήφιας πρότασης. Για παράδειγμα, στα πειράματα στα οποία για τον υπολογισμό των BLEU scores χρησιμοποιήθηκαν μόνο τα unigramms, η στήλη N-Gramms είναι συμπληρωμένη με το λεκτικό "1", στα παραδείγματα που συνδυάστηκαν unigramms και bi-gramms είναι συμπληρωμένη με το λεκτικό "1,2" κ.ο.κ.. Ανεξάρτητα από τους διάφορους συνδυασμούς που εξετάστηκαν, τα αποτελέσματα της μέτρησης BLEU συμβάδιζαν με την ανθρώπινη αξιολόγηση κατά μέσο όρο κατά 63.4% στο συνολικό σετ δεδομένων. Στην καλύτερη περίπτωση, συμβάδιζαν με την ανθρώπινη αξιολόγηση κατά 65.68%. Η λίστα με τις λέξεις για εξαίρεση του πακέτου nltk δεν χρησιμοποιήθηκε σε μόλις ένα πείραμα, και καθώς στο αντίστοιχο πείραμα που χρησιμοποιήθηκε η λίστα υπήρχε κατά τρεις μονάδες αυξημένη συμφωνία με την ανθρώπινη μετάφραση, η λίστα χρησιμοποιήθηκε για όλα τα υπόλοιπα πειράματα.

ΣΤ.2 Πειράματα και αποτελέσματα μετρήσεων ROUGE

Για τα πειράματα με την μέτρηση ROUGE εξερευνήθηκαν διάφορες παραλλαγές της. Καθώς η βιβλιοθήκη rouge για την γλώσσα python προσφέρει εύκολα την εναλλαγή μεταξύ τριών επιλογών για τον αριθμό λέξεων που θα αποτελούν ένα n-gramm, αλλά και τον τρόπο βαθμολόγησης του τελικού αποτελέσματος(μεταξύ ανάκλησης "r", ακρίβειας(precision) "p", και f-1 score "F") έγιναν πειράματα και με τους εννιά δυνατούς διαφορετικούς συνδυασμούς. Παρατηρώντας τα Αποτελέσματα Μετρήσεων ROUGE μπορούν να προκύψουν τα παρακάτω συμπεράσματα. Με την ρύθμιση "rouge-l" επιτεύχθηκε ο υψηλότερος μέσος όρος (~0.23% υψηλότερα από ότι με τη χρήση unigramms και σχεδόν 2% υψηλότερα απ' ότι με τη χρήση bi-gramms). Ταυτόχρονα από τις τρεις επιλογές του scoring μηχανισμού, η f1 μέτρηση είχε κατά μέσο όρο την μεγαλύτερη

συμφωνία με την ανθρώπινη αξιολόγηση. Από τα παραπάνω προκύπτει το συμπέρασμα ότι το υψηλότερο αποτέλεσμα που προέκυψε με τη ρύθμιση ["rouge-l"]["f"] (66.24% συμφωνία με την ανθρώπινη αξιολόγηση) δεν αποτελεί προϊόν τυχαιότητας αλλά επιβεβαίωση των παραπάνω στοιχείων.

ΣΤ.3 Πειράματα και αποτελέσματα με την χρήση της βιβλιοθήκης fastText

Τα αποτελέσματα των πειραμάτων με τη συγκεκριμένη μεθοδολογία φαίνονται στον πίνακα Αποτελέσματα FastText και Ομοιότητας Συνημίτονου. Συνολικά έγιναν τέσσερα πειράματα, με ή χωρίς μετατροπή των κεφαλαίων γραμμάτων σε πεζά και ορίζοντας ως επιλογή είτε την μετάφραση του στατιστικού μοντέλου είτε του νευρωνικού μοντέλου, στις περιπτώσεις όπου οι δύο προτάσεις είχαν ισόβαθμη ομοιότητα συνημίτονου με την πρόταση την ανθρώπινης μετάφρασης. Ακόμη και στην καλύτερη περίπτωση, η συμφωνία της συγκεκριμένης μεθοδολογίας ήταν 65.13% και μάλιστα με ένα μεγάλο ποσοστό ισοβαθμίας μεταξύ των διανυσμάτων των δύο μεταφράσεων. Για την ακρίβεια, η συγκεκριμένη μέθοδος κατέληξε σε ισοβαθμία σε 187 περιπτώσεις, δηλαδή σχεδόν στο 7% των δεδομένων. Αυτό φαίνεται και από την μεγάλη διαφορά που προκύπτει αν όταν στην περίπτωση ισοβαθμίας επιλεχθούν οι μεταφράσεις του στατιστικού μοντέλου ως κοντινότερες στις ανθρώπινες, αντί για τις μεταφράσεις του νευρωνικού μοντέλου, όπου η συμφωνία με τις ανθρώπινες αξιολογήσεις είναι μόλις στο 59.95%. Συγκρίνοντας την συγκεκριμένη μέθοδο με τις μετρικές αξιολόγησης ROUGE και BLEU, μπορεί να παρατηρηθεί ότι τα αποτελέσματα της είναι χαμηλότερα παρά την πολυπλοκότερη διαδικασία που απαιτείται για την δημιουργία του προεκπαιδευμένου μοντέλου με βάση το οποίο έγινε η δημιουργία των διανυσμάτων.

ΣΤ.4 Πειράματα και αποτελέσματα με το συνδυασμό Bag Of Words και TNA

Με τη συγκεκριμένη μέθοδο έγιναν και τα περισσότερα πειράματα καθώς οι παράμετροι του TNA επέτρεπαν τις περισσότερες αλλαγές και προσαρμογές. Τα κορυφαία αποτελέσματα των διάφορων πειραμάτων φαίνονται στο BOW-TNA - 20 Υψηλότερες Βαθμολογίες. Στην καλύτερη περίπτωση, η συμφωνία με την ανθρώπινη αξιολόγηση έφτασε το 63.48%, όταν στο ίδιο σετ δεδομένων με τον ZeroR μπορούμε να πετύχουμε ορθότητα σε ποσοστό 57.97%. Δεδομένου της πολυπλοκότητας της υλοποίησης αλλά και της απαιτούμενης υπολογιστικής δύναμης, η συγκεκριμένη μέθοδος δεν έχει παράξει αρκετά καλά αποτελέσματα για να αξιοποιηθεί ως εργαλείο σε πραγματικές συνθήκες, ειδικά όταν μια ιδιαίτερα γρηγορότερη και οικονομικότερη μετρική όπως η ROUGE βρίσκεται κοντινότερα στην ανθρώπινη αξιολόγηση. Παρόλα αυτά μπορούμε να κάνουμε κάποιες παρατηρήσεις πάνω στα αποτελέσματα που υπάρχουν, οι οποίες ίσως προσφέρουν κάποια ένδειξη προς κάποια μελλοντική κατεύθυνση για την αναζήτηση βελτιώσεων της συγκεκριμένης μεθοδολογίας ή κάποιας άλλης που αποσκοπεί στην επίλυση κάποιου παρόμοιου προβλήματος. Όσον αφορά τη δομή του νευρωνικού δικτύου, έγιναν δοκιμές με δύο ή τέσσερα κρυφά στρώματα. Το διάγραμμα διασποράς αυτών των δοκιμών με τα αποτελέσματα τους φαίνεται στο Σκορ ανά αριθμό κρυφών στρωμάτων και φαίνεται ότι η χρήση τεσσάρων κρυφών στρωμάτων δεν βοήθησε στην βελτίωση πρόβλεψης της ανθρώπινης αξιολόγησης σε σχέση με την χρήση δύο κρυφών στρωμάτων παρά την μεγαλύτερη απαίτηση πόρων που απαιτούν τα μοντέλα με περισσότερα στρώματα. Άλλη μία παράμετρος γύρω από την οποία έγιναν διάφορες δοκιμές αποτελούσε ο αριθμός των εποχών κατά τις οποίες θα εκπαιδευτεί το TNA. Από το διάγραμμα διασποράς Σκορ ανά Εποχή (BOW - TNA) και τον πίνακα Μέσος όρος ορθότητας TNA κάθε διαθέσιμης ρύθμισης φαίνεται ότι εκτός από την ρύθμιση με τέσσερις εποχές που φαίνεται να είχε ένα μικρό ποσοστό υψηλότερης ορθότητας(0.2%) σε σχέση με τις άλλες δοκιμές, ο αριθμός των εποχών δεν έπαιξε ιδιαίτερο ρόλο για το μέσο όρο ορθότητας των TNA, εφόσον ήταν από 5 ως 7 εποχές. Όσον αφορά τον αριθμό νευρώνων ανά στρώμα, στο διάγραμμα διασποράς Σκορ ανά αριθμό νευρώνων σε κάθε κρυφό στρώμα σε συνδυασμό με τους μέσους όρους που έχουν προκύψει για την ορθότητα, φαίνεται ότι οι μέσοι όροι των επιλογών που δοκιμάστηκαν δεν είχαν μεγάλες διαφορές μεταξύ τους. Ωστόσο το γεγονός ότι οι

δύο υψηλότεροι μέσοι όροι γύρω από αυτή την παράμετρο, επιτεύχθηκαν με 256 και 1024 νευρώνες στα κρυφά στρώματα, αλλά και το ότι με αυτό των αριθμών νευρώνων επιτεύχθηκαν τα 11 από τα 12 υψηλότερα score των πειραμάτων, υποδεικνύει ότι η χρυσή τομή όσον αφορά την συγκεκριμένη παράμετρο ίσως βρίσκεται κάπου μεταξύ αυτών των δύο αριθμών. Όσον αφορά την χρήση DropOut στρωμάτων επίσης φαίνεται να μην έπαιξε ιδιαίτερο ρόλο για το μέσο όρο ορθότητας, ωστόσο τα πέντε κορυφαία αποτελέσματα προέκυψαν με τη χρήση αυτών. Σχετικά με την συνάρτηση ενεργοποίησης, οι δοκιμές που έγιναν με τις tanh, relu και LeakyRelu δεν είχαν αξιοσημείωτες διαφορές μεταξύ τους, με την relu να επιτυγχάνονται ελαφρώς υψηλότερα αποτελέσματα. Τελικά, οι διαφορές μεταξύ των πειραμάτων που φαίνεται να επηρέασαν σε μεγαλύτερο βαθμό τα αποτελέσματα ήταν το φιλτράρισμα στο αρχικό σετ δεδομένων και ο τρόπος με τον οποίο διαχωρίστηκε το σετ εκπαίδευσης από το σετ δεδομένων. Από τους τρεις διαφορετικούς τρόπους μορφοποίησης του τελικού σετ δεδομένων, ο πρώτος φαίνεται να είχε τον υψηλότερο μέσο όρο ορθότητας, καθώς και τα πέντε κορυφαία αποτελέσματα. Για να αποκτήσει το συγκεκριμένο σετ δεδομένων την τελική του μορφή, δεν αφαιρέθηκαν καθόλου λέξεις που εμφανίζονταν συχνά, ούτε οι ανανεωμένες stopwords του H.2, αφαιρέθηκαν όμως οι προτεινόμενες λέξεις της βιβλιοθήκης NLTK για την Ελληνική γλώσσα καθώς και όσες λέξεις εμφανίστηκαν από πέντε ή λιγότερες φορές συνολικά στα παράλληλα κείμενα. Με αυτόν τον τρόπο σημειώθηκε μέσος όρος 0.64% υψηλότερος από το δεύτερο στη βαθμολογία σετ δεδομένων. Από τους τρόπους διαχωρισμού του σετ δεδομένων, ο διαχωρισμός 1προς1, δηλαδή αυτός που καταλήγει σε σετ αξιολόγησης ίδιου μεγέθους με το σετ εκπαίδευσης, πέτυχε μακράν υψηλότερα αποτελέσματα ως προς την συμφωνία με την ανθρώπινη αξιολόγηση, καθώς η διαφορά των μέσων όρων αυτής για τα δύο σετ δεδομένων ήταν στο 10%.

Z Συμπεράσματα – Προτάσεις για το μέλλον

Παρόλο που τα αποτελέσματα δεν είναι ενθαρρυντικά για την τεχνική με τη δομή Bag of Words που υλοποιήθηκε στα πλαίσια εκπόνησης της συγκεκριμένης εργασίας, μπορούν να προκύψουν σημειώσεις και συμπεράσματα τα οποία ίσως φανούν χρήσιμα σε μελλοντικά πειράματα. Για παράδειγμα, στις περισσότερες περιπτώσεις, όταν όλοι οι χαρακτήρες μετατράπηκαν σε μικρά γράμματα, παρατηρήσαμε καλύτερες επιδόσεις σε σχέση με τη διατήρηση των κεφαλαίων και πεζών γραμμάτων. Επίσης, για την εκπαίδευση του νευρωνικού δικτύου, όλες οι μεταβλητές γύρω από τη δομή του ΤΝΔ δεν έπαιξαν σημαντικό ρόλο, με τους μέσους όρους ορθότητας για το σύνολο των πειραμάτων για κάθε μία από αυτές τις τιμές να κυμαίνονται μεταξύ 53.73% και 55.17%. Αντιθέτως, το σετ δεδομένων που χρησιμοποιήθηκε για τα πειράματα και ο τρόπος με τον οποίο χωρίστηκε σε σετ εκπαίδευσης και σετ αξιολόγησης, φαίνεται ότι επηρέασαν τα αποτελέσματα των πειραμάτων σε πολύ μεγαλύτερο βαθμό.

Z.1 Σύγκριση με αποτελέσματα προηγούμενων ερευνών

Καθώς στη βιβλιογραφία υπάρχουν πολλοί διαφορετικοί τρόποι αξιολόγησης των διαφόρων μετρικών και σε πολλά διαφορετικά σετ δεδομένων, η άμεση και εύκολη σύγκριση των διαφορετικών αλγορίθμων αξιολόγησης μηχανικής μετάφρασης αποτελεί αρκετά πολύπλοκη διαδικασία. Για παράδειγμα, σε κάποια άρθρα οι συγγραφείς μετρούν την ακρίβεια των αποτελεσμάτων των αυτόματων τεχνικών σε σχέση με αυτά της ανθρώπινης αξιολόγησης ως προς την ακρίβεια[40], αλλά οι τεχνικές που χρησιμοποιούνται δεν συμπίπτουν με αυτές που χρησιμοποιήθηκαν κατά την εκπόνηση αυτής της εργασίας. Σε άλλες δουλειές μπορεί να χρησιμοποιούνται μετρήσεις διαφορετικές της Ορθότητας(accuracy) όπως ο Pearson συντελεστής συνδιακύμανσης[23]. Όλες αυτές οι διαφορετικές παράμετροι κάνουν την εξαγωγή συμπερασμάτων από διαφορετικές δουλειές αρκετά περίπλοκη και επιρρεπής στο σφάλμα, δίνουν όμως μια εικόνα των πόσων διαφορετικών παραμέτρων πρέπει να λάβει υπ όψιν ένας ερευνητής αλλά και του όγκου της δουλειάς που έχει γίνει ήδη στον τομέα. Τυπικά, η τάση που παρατηρείται στην βιβλιογραφία, τοποθετεί την BLEU χαμηλότερα από την ROUGE, κάτι που επιβεβαιώνεται και στα πειράματα που εκπονήθηκαν για την παρούσα εργασία, και τις τεχνικές που χρησιμοποιούν προεκπαιδευμένα γλωσσικά μοντέλα ψηλότερα. Κατά την εκπόνηση αυτής της εργασίας, αυτό δεν επιβεβαιώθηκε καθώς η μετρική ROUGE είχε μεγαλύτερη συμφωνία με την ανθρώπινη αξιολόγηση από της μετρήσης που προέκυψαν με την βοήθεια της βιβλιοθήκης fastText και του μοντέλου που διατίθεται με αυτή.

Z.2 Προτάσεις για το μέλλον και τη βελτίωση των αποτελεσμάτων

Z.2.1 Πειραματισμός και με διαφορετικούς αριθμούς νευρώνων νευρώνων στα κρυφά στρώματα

Στον πίνακα Μέσος όρος ορθότητας ΤΝΔ κάθε διαθέσιμης ρύθμισης παρατηρείται ότι οι δύο καλύτερες επιλογές για τον αριθμό νευρώνων για κάθε κρυφό στρώμα ήταν οι 256 ή 1024. Καθώς σε κανένα από τα πειράματα δεν χρησιμοποιήθηκαν στρώματα με πλήθος νευρώνων μεταξύ αυτών των δύο αριθμών, περαιτέρω πειραματισμός γύρω από την συγκεκριμένη μεταβλητή, π.χ. με 512 νευρώνες, ίσως αποτελεί προσέγγιση που μπορεί να οδηγήσει σε αποτελέσματα κοντινότερα στα ανθρώπινα.

Z.2.2 Συνδυασμός με περισσότερα δεδομένα

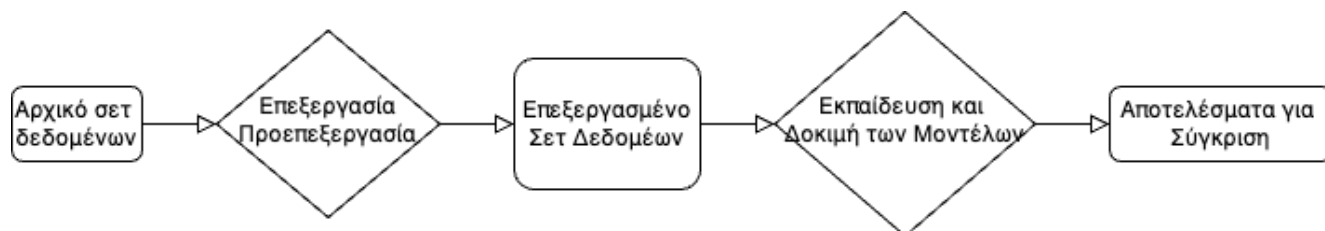
Θα μπορούσε στο μέλλον το συγκεκριμένο dataset να συνδυαστεί και με άλλα, έτσι ώστε να παρατηρηθεί εάν ένα μεγαλύτερο σετ δεδομένων θα οδηγούσε σε καλύτερα αποτελέσματα, ειδικά στην προσέγγιση που

χρησιμοποιήθηκε η δομή Bag Of Words σαν εισαγωγή σε τεχνητό νευρωνικό δίκτυο.

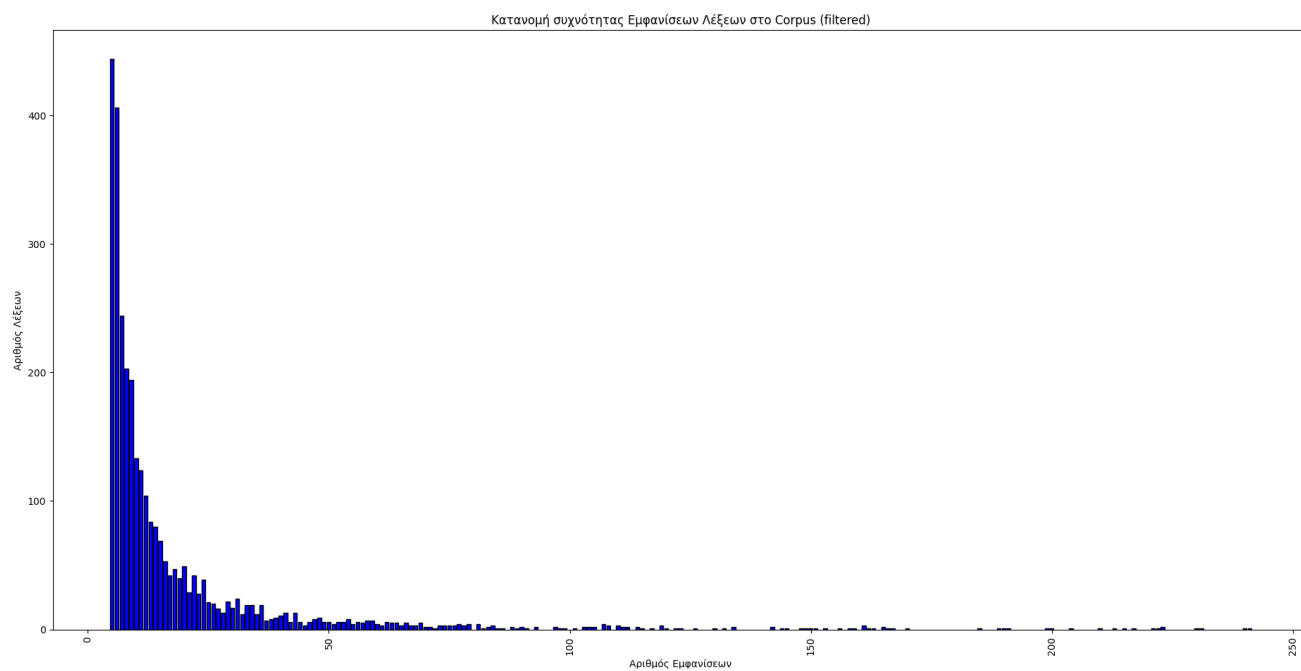
Z.2.3 Διαχωρισμός σετ εκπαίδευσης και σετ αξιολόγησης με αλγόριθμο τυχειότητας

Όπως μπορεί να παρατηρηθεί στα πειράματα του ΤΝΔ από τους μέσους όρους στον πίνακα Μέσος όρος ορθότητας ΤΝΔ κάθε διαθέσιμης ρύθμισης, τα μοντέλα που εκπαιδεύτηκαν στο 50% του σετ δεδομένων είχαν υψηλότερα αποτελέσματα από αυτά που εκπαιδεύτηκαν στο 90%. Θεωρητικά, τα περισσότερα δεδομένα του υποσυνόλου του 90% του σετ δεδομένων θα έπρεπε να είχαν συμβάλει σε καταλληλότερη εκπαίδευση του ΤΝΔ με υψηλότερα αποτελέσματα ορθότητας. Στην πράξη όμως η προκατάληψη των δύο υποσυνόλων σε διαφορετικές απαντήσεις ως ορθές είχε τα αντίστροφα αποτελέσματα. Μία προσέγγιση που ίσως βοηθούσε στο να ξεπεραστεί το παραπάνω εμπόδιο, και στη συνέχεια σε υψηλότερα αποτελέσματα είναι ο τυχαίος διαχωρισμός του σετ εκπαίδευσης και σετ δεδομένων, αντί αυτού του διαχωρισμού με βάση την υφιστάμενη θέση στο σετ, εφόσον έχει παρατηρηθεί αυτή η ανισότητα στην κατανομή των απαντήσεων.

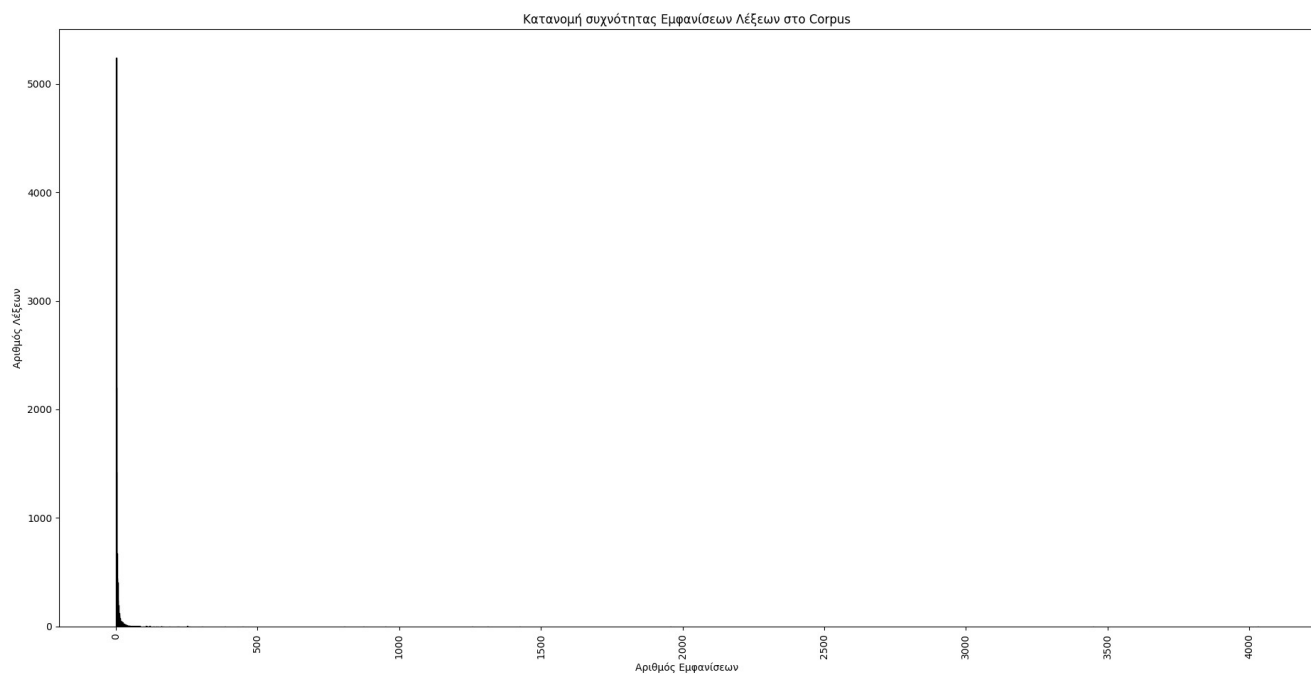
Η Πίνακες – Σχήματα – Αναφορές



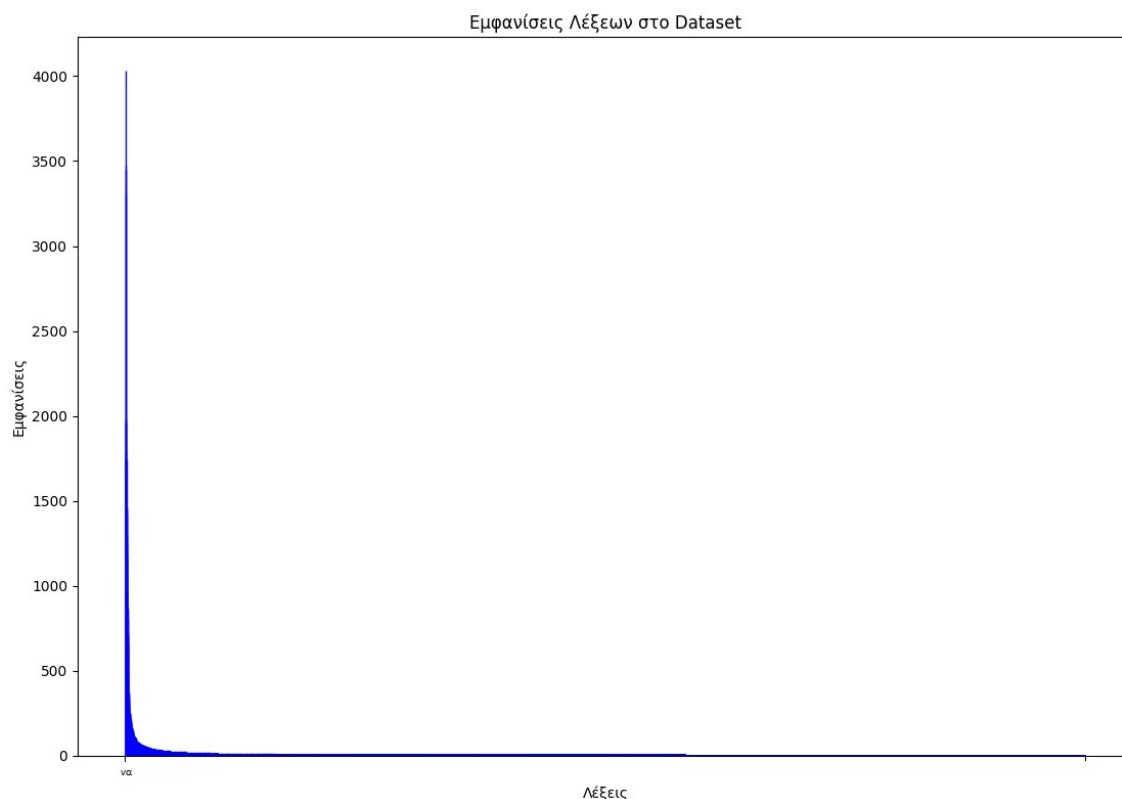
Σχήμα A: Flow Μεθοδολογίας Υλοποίησης



Σχήμα B: Κατανομή Αριθμού Εμφανίσεων Λέξεων (χωρίς τις ακραίες περιπτώσεις)



Σχήμα Γ: Κατανομή Αριθμού Εμφανίσεων των Λέξεων (με τις ακραίες περιπτώσεις)



Σχήμα Δ: Κατανομή Εμφανίσεων κάθε Λέξης (όλες οι λέξεις)

Πίνακας Α: Αποτελέσματα ZeroR

ZeroR	Υποσύνολο Σετ Αξιολόγησης	Ορθότητα	Ακρίβεια	Ανάκληση
Μόνο 0	Όχι	0.63	0	0
Μόνο 1	Όχι	0.37	0.37	1
Μόνο 0	50-50	0.6	0	0
Μόνο 1	50-50	0.4	0.4	1
Μόνο 0	90-10	0.46	0	0
Μόνο 1	90-10	0.54	0.54	1

Πίνακας Β: BOW-TNΔ - 20 Υψηλότερες Βαθμολογίες

Πείραμα #	Ορθότητα	Διαχωρισμός Σει	Συνάρτηση Ενεργοποίησης	Νευρώνες Κρυφών Στρώματων	Εποχές	Κρυφά Στρώματα	Σει Δεδομένων	DropOut Στρώματα
144	0.63	50-50	tanh	256	6	4	No StopWords, A>5 ,l	Y
187	0.63	50-50	relu	1024	7	4	No StopWords, A>5 ,l	Y
151	0.63	50-50	relu	256	4	4	No StopWords, A>5 ,l	Y
997	0.63	50-50	relu	256	5	2	No StopWords, A>5 ,l	Y
1003	0.63	50-50	relu	256	7	2	No StopWords, A>5 ,l	Y
1243	0.63	50-50	relu	256	7	4	No StopWords, 1000>A>5 appearance, l	N
985	0.63	50-50	relu	256	6	2	No StopWords, A>5 ,l	Y
134	0.63	50-50	LeakyReLU	128	5	4	No StopWords, A>5 ,l	Y
181	0.63	50-50	relu	1024	5	4	No StopWords, A>5 ,l	Y
301	0.63	50-50	relu	1024	5	4	No StopWords, 1000>A>5 appearance, l	Y
1105	0.63	50-50	relu	256	6	4	No StopWords, A>5 ,l	N
1123	0.63	50-50	relu	256	7	4	No StopWords, A>5 ,l	N
1213	0.63	50-50	relu	128	5	4	No StopWords, 1000>A>5 appearance, l	N

55	0.63	50-50	relu	1024	4	2	No StopWords, A>5 ,l	N
61	0.63	50-50	relu	1024	5	2	No StopWords, A>5 ,l	N
991	0.63	50-50	relu	256	4	2	No StopWords, A>5 ,l	Y
1237	0.63	50-50	relu	256	5	4	No StopWords, 1000>A>5 appearance, l	N
91	0.63	50-50	relu	2048	7	2	No StopWords, A>5 ,l	N
211	0.63	50-50	relu	2048	7	4	No StopWords, A>5 ,l	Y
962	0.63	50-50	LeakyReLU	128	6	2	No StopWords, A>5 ,l	Y

Πίνακας Γ: Αποτελέσματα BLEU

Μόνο Πεζά γράμματα	Αφαίρεση StopWords	Απάντηση σε περίπτωση ίδιου score	N-Gramms	Ορθότητα (%)	Διαχωρισμός Σετ
Y	Y	1	1	65.46	-
Y	N	1	1	62.56	-
Y	Y	1	1,2	65.68	-
Y	Y	0	1,2	62.07	-
Y	Y	1	1,2,3	65.09	-
Y	Y	0	1,2,3	61.36	-
Y	Y	1	1,2,3,4	64.38	-
Y	Y	0	1,2,3,4	60.54	-
Y	Y	1	2	60	9 προς 1
Y	Y	1	2	62.92	1 προς 1

Πίνακας Δ: Αποτελέσματα FastText και Ομοιότητας Συνημίτονου

Απάντηση σε περίπτωση ίδιου score	Μόνο Πεζά γράμματα	Ορθότητα (%)	Διαχωρισμός Σετ
0	N	59.69	-
1	N	64.83	-
1	Y	65.13	-
0	Y	59.95	-
1	Y	56.29	9 προς 1
1	Y	62.84	1 προς 1
0	Y	50	9 προς 1
0	Y	59.05	1 προς 1

Πίνακας Ε: Αποτελέσματα Μετρήσεων ROUGE

Μόνο πεζά γράμματα	Αφαίρεση StopWords	Απάντηση σε περίπτωση ίδιου score	Ορθότητα % (accuracy)	Διαχωρισμός Σετ	
Y	Y	1	66.24	-	["rouge-l"]["f"]
Y	Y	0	63.08	-	["rouge-l"]["f"]
Y	Y	0	62.6	-	["rouge-1"]["f"]
Y	Y	1	65.61	-	["rouge-1"]["f"]
Y	Y	0	60.99	-	["rouge-2"]["f"]
Y	Y	1	63.11	-	["rouge-2"]["f"]
Y	Y	1	62.07	-	["rouge-l"]["r"]
Y	Y	0	63.52	-	["rouge-l"]["r"]
Y	Y	0	62.97	-	["rouge-1"]["r"]
Y	Y	1	62.22	-	["rouge-1"]["r"]
Y	Y	1	60.91	-	["rouge-2"]["r"]
Y	Y	0	62.15	-	["rouge-2"]["r"]
Y	Y	1	65.69	-	["rouge-l"]["p"]
Y	Y	0	62.74	-	["rouge-l"]["p"]
Y	Y	0	62.56	-	["rouge-1"]["p"]
Y	Y	1	65.65	-	["rouge-1"]["p"]
Y	Y	0	61.21	-	["rouge-2"]["p"]

Y	Y	1	63.19	-	["rouge-2"]["p"]
Y	Y	1	60.37	9010	["rouge-l"]["f"]
Y	Y	0	53.7	9010	["rouge-l"]["f"]
Y	Y	1	64.26	5050	["rouge-l"]["f"]
Y	Y	0	61.43	5050	["rouge-l"]["f"]

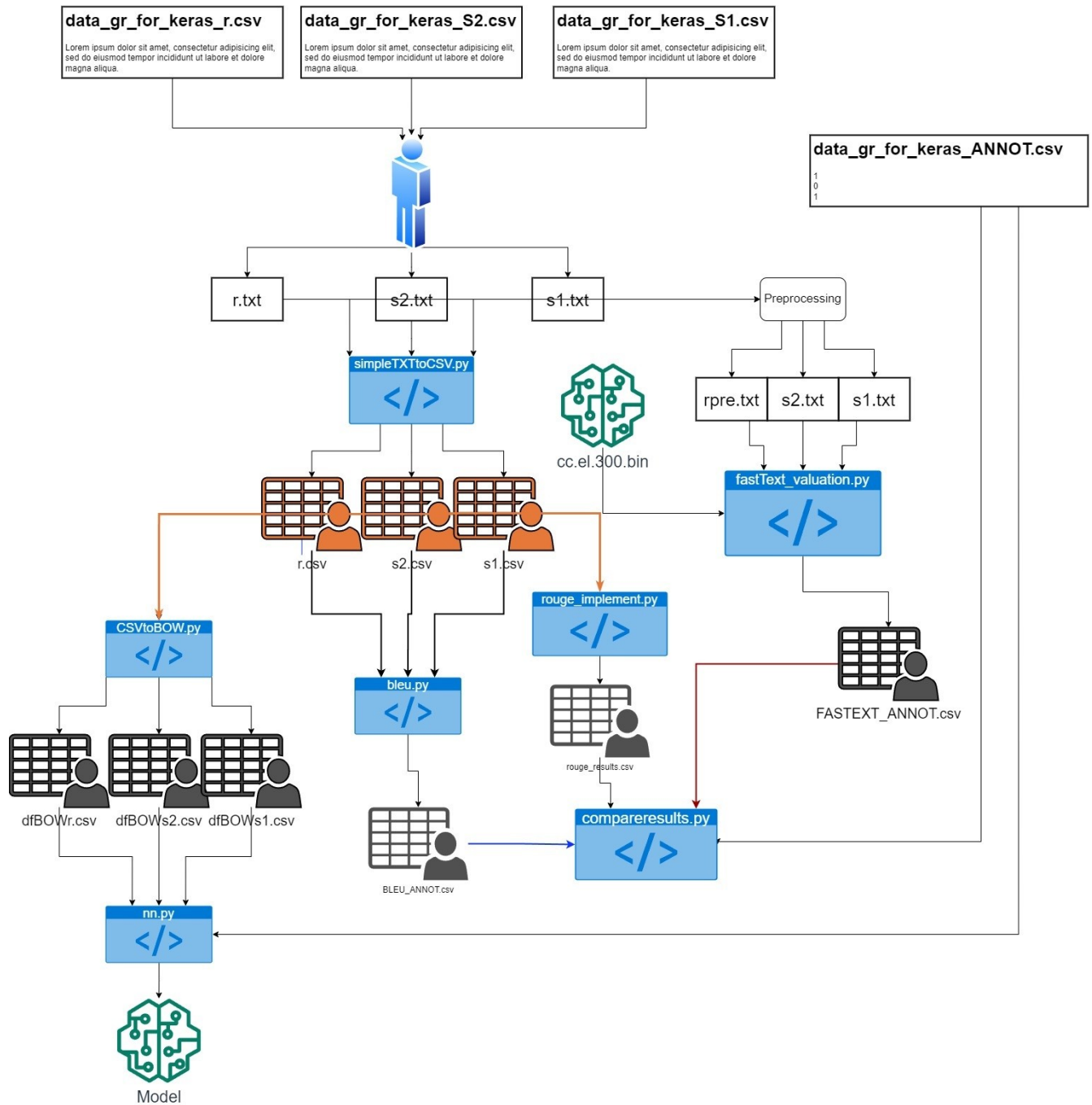
Πίνακας ΣΤ: Συχνότερα Εμφανιζόμενες Λέξεις πριν την Προεπεξεργασία

Λέξη	Εμφανίσεις
να	4008
είναι	3156
το	3111
και	3060
που	1956
από	1768
για	1749
του	1745
την	1723
η	1577
με	1413
ένα	1357
σε	1320
τα	1196
της	1171
μια	1134
ότι	1063
θα	1015
αυτό	956
των	931

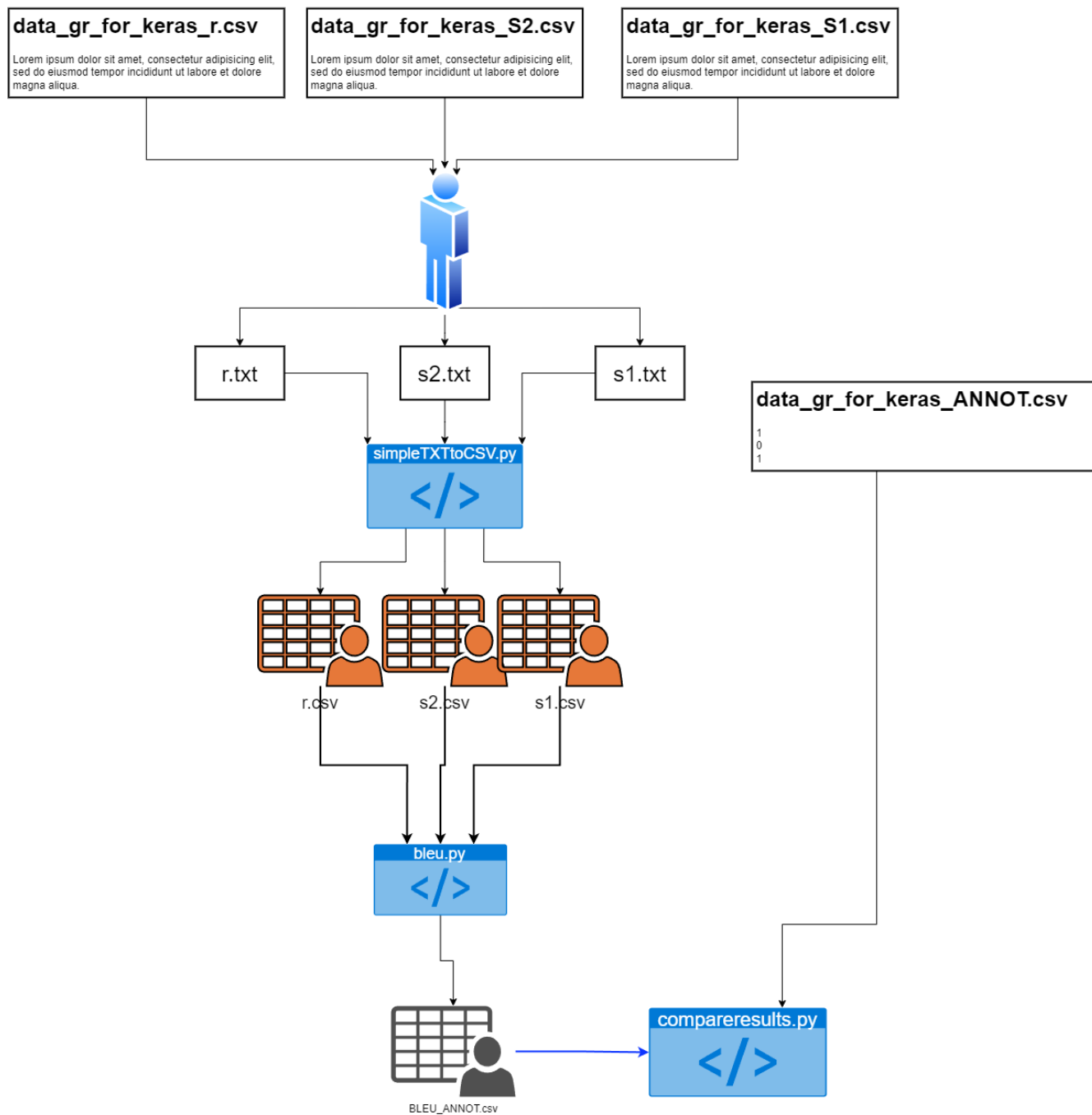
Πίνακας Ζ: Συχνότερα Εμφανιζόμενες Λέξεις μετά την Προεπεξεργασία

Λέξη	Εμφανίσεις
να	4031
και	3475
το	3449

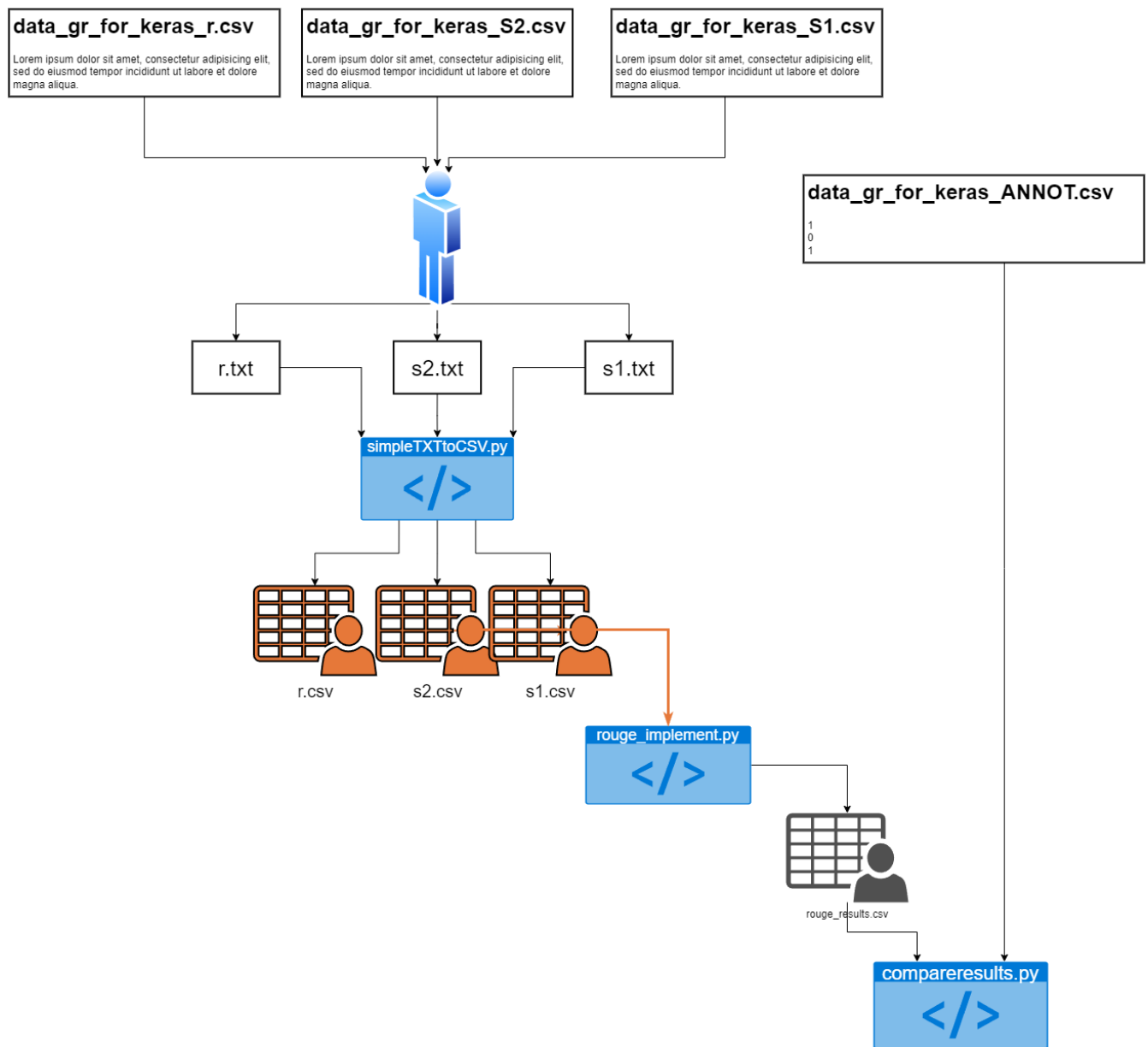
είναι	3297
η	1983
που	1961
για	1957
από	1801
του	1748
την	1740
με	1464
ένα	1452
σε	1426
τα	1313
μια	1258
αυτό	1214
της	1171
θα	1118
ότι	1071
οι	952



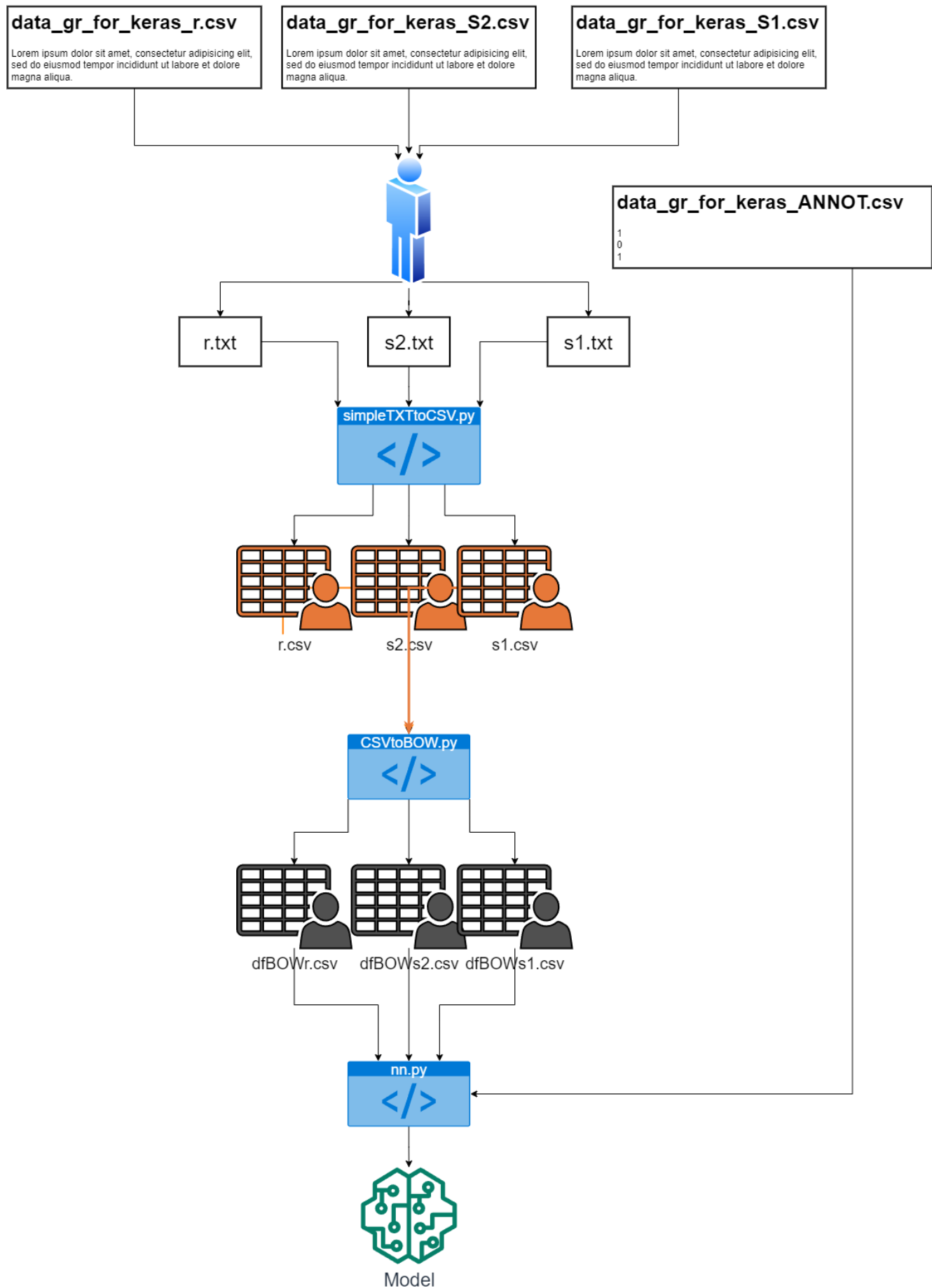
Σχήμα E.Flow όλων των αρχείων και scripts που χρησιμοποιήθηκαν



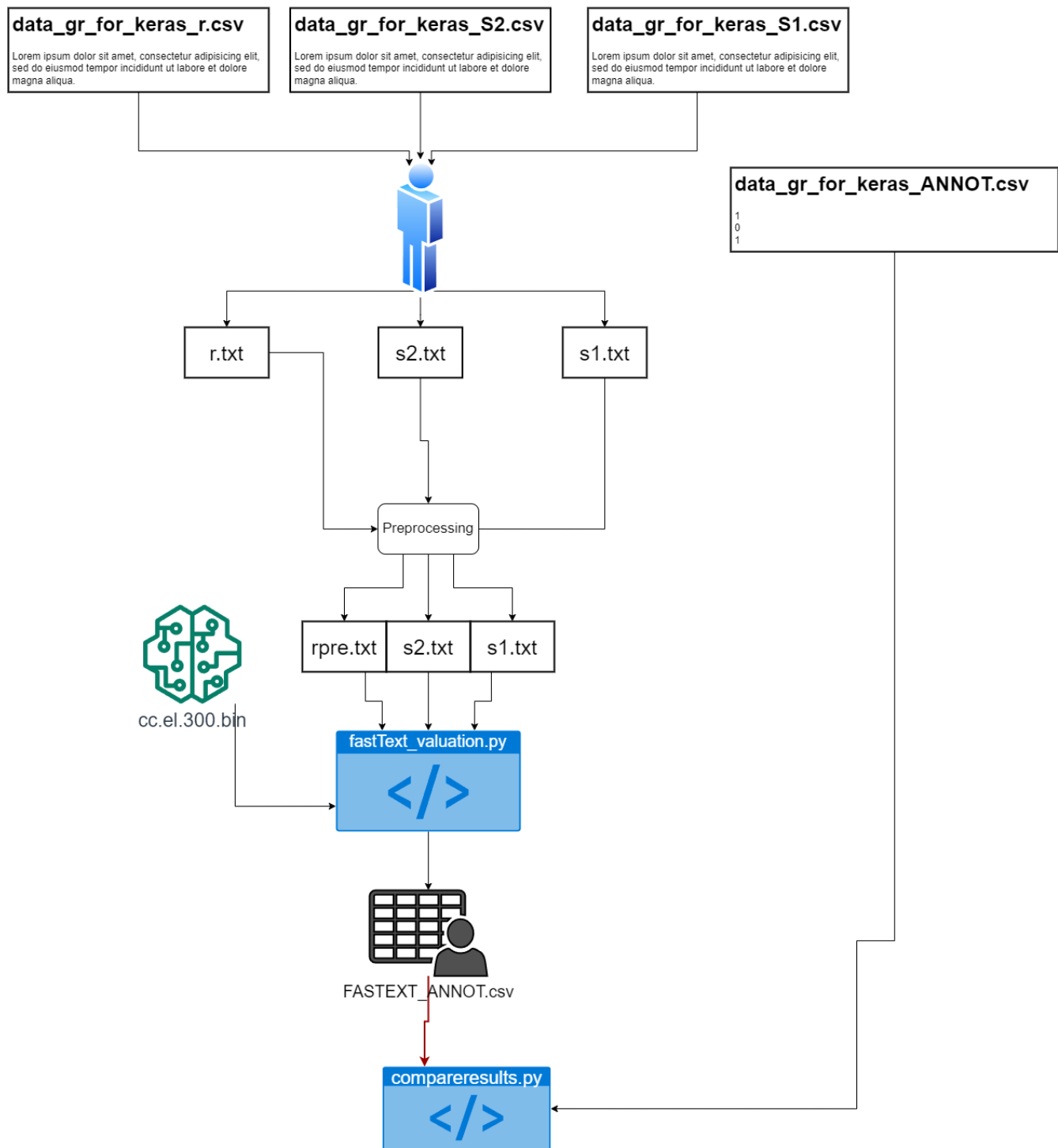
Σχήμα ΣΤ: Τεχνική Ροή Δημιουργίας και Αξιολόγησης Μέτρησης BLEU



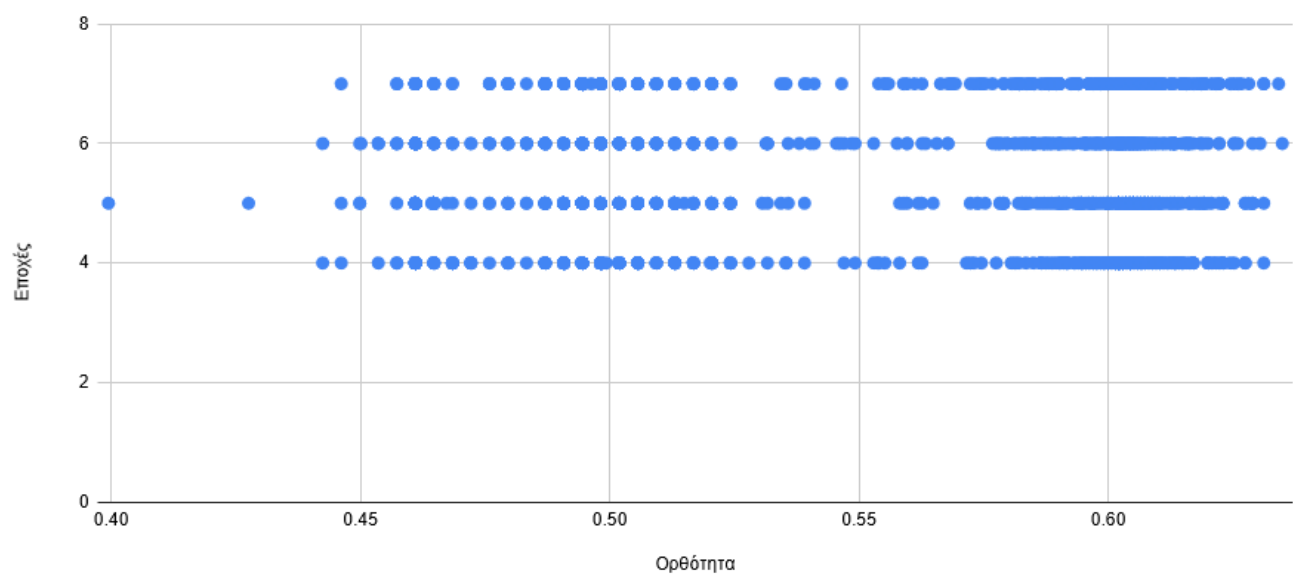
Σχήμα Z: Τεχνική Ροή Δημιουργίας και Αξιολόγησης Μέτρησης ROUGE



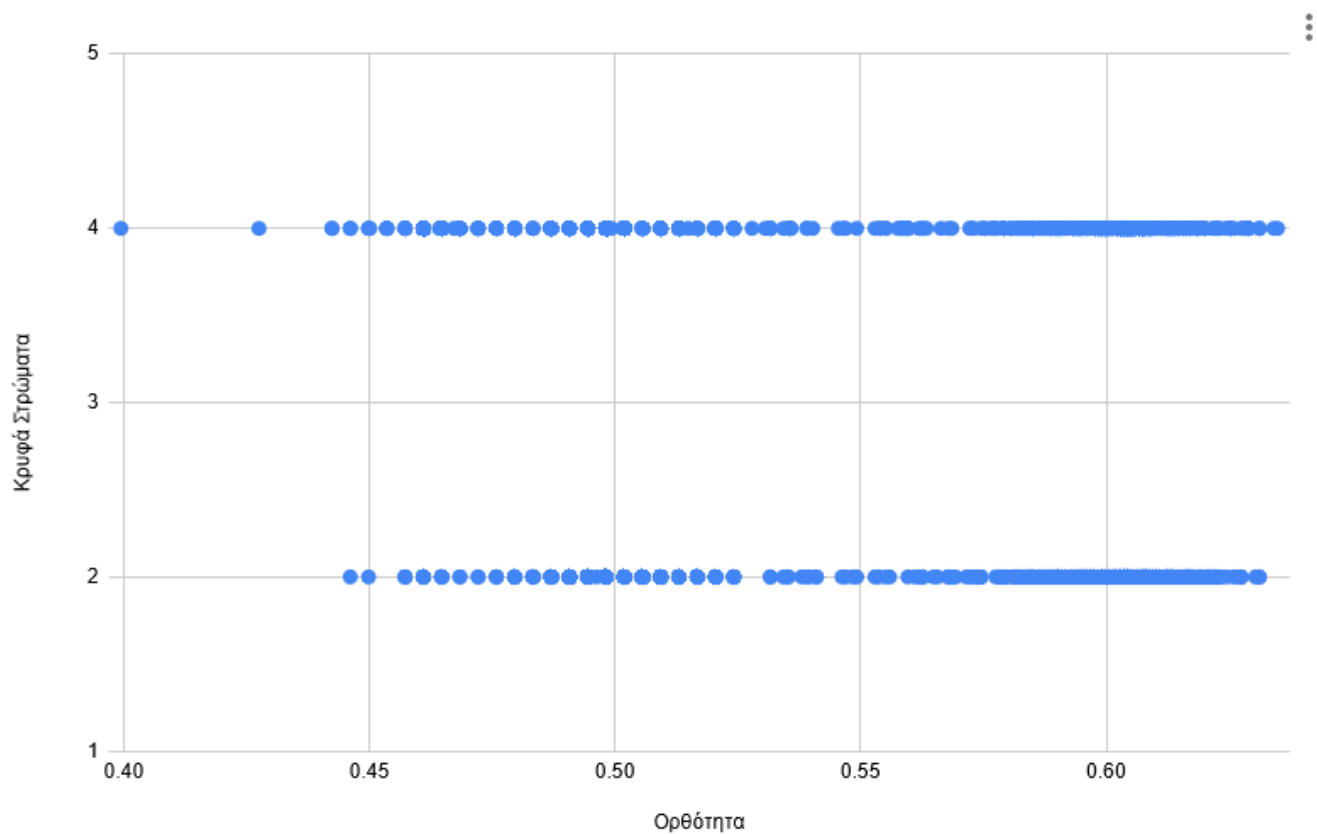
Σχήμα Η: Τεχνική Ροή Δημιουργίας και Αξιολόγησης Νευρωνικού Μοντέλου(Δεδομένα BOW)



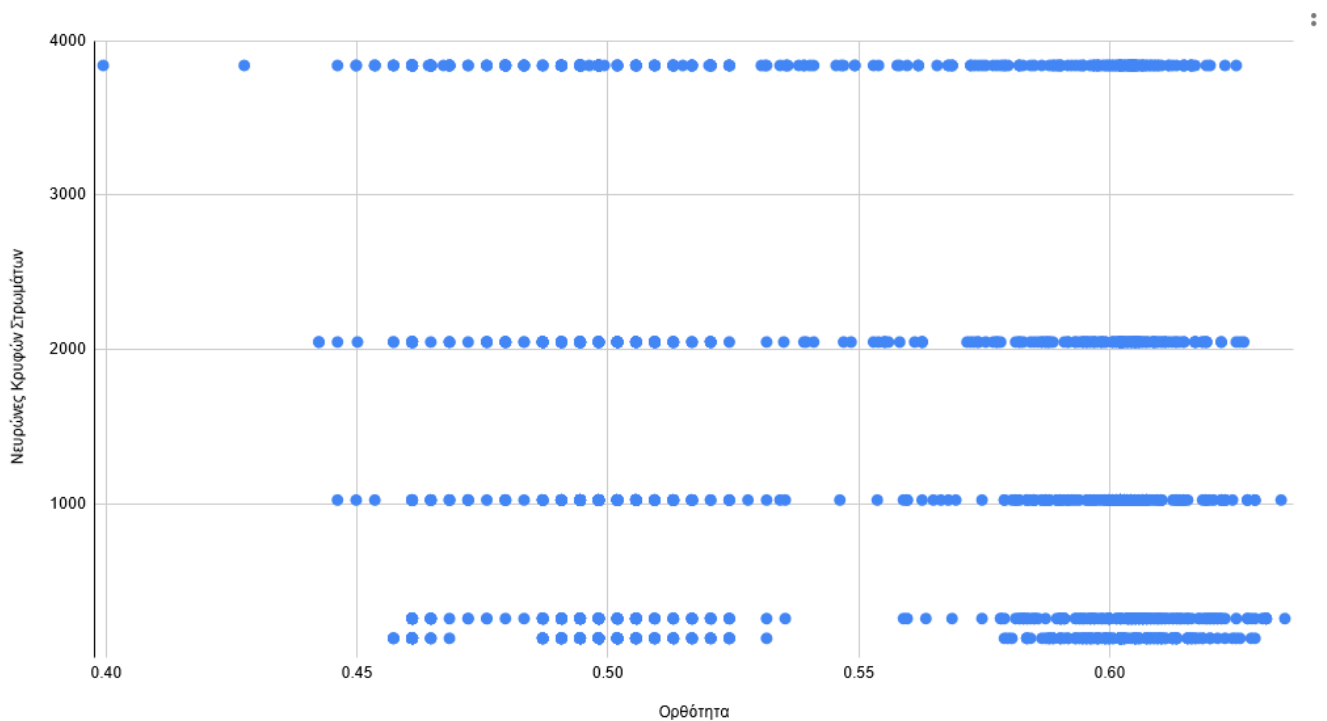
Σχήμα Θ: Τεχνική Ροή Δημιουργίας και Αξιολόγησης Μετρήσεων από το fastText



Σχήμα I: Σκορ ανά Εποχή (BOW - TNA)



Σχήμα ΙΑ: Σκορ ανά αριθμό κρυφών στρωμάτων



Σχήμα IB: Σκορ ανά αριθμό νευρώνων σε κάθε κρυφό στρώμα

Πίνακας H: Μέσος όρος ορθότητας TNA κάθε διαθέσιμης ρύθμισης

Ρύθμιση	Επιλογή Ρύθμισης	Μέση ορθότητα
Διαχωρισμός σετ	9 προς 1	49.23%
Διαχωρισμός σετ	1 προς 1	59.71%
Συνάρτηση ενεργοποίησης	tahn	54.18%
Συνάρτηση ενεργοποίησης	relu	54.9%
Συνάρτηση ενεργοποίησης	LeakyReLU	54.34%
Αριθμός Κρυφών Στρωμάτων	2	54.74%
Αριθμός Κρυφών Στρωμάτων	4	54.2%
Νευρώνες Κρυφών Στρωμάτων	128	54.69%
Νευρώνες Κρυφών Στρωμάτων	256	54.77%
Νευρώνες Κρυφών Στρωμάτων	1024	54.72%
Νευρώνες Κρυφών Στρωμάτων	2048	54.45%
Νευρώνες Κρυφών Στρωμάτων	3840	53.73%
Dropout Στρώματα	Με χρήση Dropout στρωμάτων	54.4%

LSTM – Long Short Term Memory

MQM – Multidimensional Quality Metrics

Seq2Seq – Sequence to Sequence

TER – Translation Error Rate

WER – Word Error Rate

MM – Μηχανική Μετάφραση

ΣΜΜ – Στατιστική Μηχανική Μετάφραση

TNΔ – Τεχνητά Νευρωνικά Δίκτυα

I Γλωσσάρι Ξενικών Όρων

Πίνακας Θ: Ξενικοί Όροι

Αγγλικός Όρος	Ελληνικός Όρος
Activation Function	Συνάρτηση Ενεργοποίησης
Array	Πίνακας
Artificial Neural Networks	Τεχνητά Νευρωνικά Δίκτυα
Bias	Παράγοντας προδιάθεσης, Πόλωση
Candidate	Υποψήφιος
Classification	Κατηγοριοποίηση
Context Vector	Διάνυσμα Περιεχομένου
Convolutional Layer	Συνελικτικό Στρώμα
Convolutional Neural Networks	Συνελικτικά Νευρωνικά Δίκτυα[41]
Corpora	Σώματα κειμένου
Cosine Similarity	Ομοιότητα Συνημίτονου
Dataset	Σετ δεδομένων
Decoder	Αποκωδικοποιητής
Dictionary	Λεξικό
Elementary Perceptron	Στοιχειώδες perceptron
Encoder	Κωδικοποιητής
Encoding	Κωδικοποίηση
Epoch	Εποχή
Error Type	Τύπος Σφάλματος
Exploding Gradient Problem	Πρόβλημα Εκρηκτικής Κλίσης
Feed forward	Πρόσθια Τροφοδότηση
Flattening	Στην προεπεξεργασία εικόνας είναι η διαδικασία κατά την οποία η πληροφορία για κάθε pixel μετατρέπεται από πολυδιάστατη σε μονοδιάστατη
Flow	Ροή
Framework	Πλαίσιο
Gradient	Κλίση
Hidden Layer	Κρυφό επίπεδο
Input Layer	Επίπεδο εισόδου
Layer	Επίπεδο, Στρώμα Τεχνητών Νευρώνων
Local Connectivity	Τοπική συνδεσιμότητα
Long Short Term Memory	Μακρά Βραχύχρονη Μνήμη

Machine Translation Evaluation	Αξιολόγηση Μηχανικής Μετάφρασης
Multidimensional Quality Metrics	Πολυδιάστατες Μετρικές Ποιότητας
Neuron	Νευρώνας
Node	Κόμβος
Output Layer	Επίπεδο εξόδου
Overfitting	Υπερπροσαρμογή / Υπερταίριασμα
Pixel	Εικονοκύτταρο
Pooling Layer	Συγκεντρωτικά Επίπεδα
Recall	Ανάκληση
Record	Εγγραφή/Καταχώρηση
Recurrent Neural Network	Αναδρομικά Νευρωνικά Δίκτυα
Reference	Αναφορά – Στην μηχανική μετάφραση με τον όρο Reference μπορεί να αναφερόμαστε σε μία σωστή πρόταση που έχει γίνει από άνθρωπο με σκοπό να την συγκρίνουμε με προτάσεις που έχουν προκύψει μετά από μηχανική μετάφραση
Sequence	Ακολουθία
Stemming	Στελέχωση[42]
Stopword	Λέξεις οι οποίες αφαιρέθηκαν από το σετ δεδομένων
Script	ένα μικρό κομμάτι κώδικα που εκτελεί μια επιμέρους εργασία[43]
Supervised Learning	Μάθηση με επίβλεψη
Transformer	Παραλλαγή του Encoder-Decoder Framework που χρησιμοποιεί κυρίως τον μηχανισμό προσοχής
Underfitting	Υποπροσαρμογή / Ατελής μάθηση
Unsupervised Learning	Μάθηση χωρίς επίβλεψη
Vanishing Gradient Problem	Πρόβλημα Εξαφάνισης Κλίσης
Variable	Παράμετρος
Weight	Βάρος

ΙΑ Βιβλιογραφία

- [1]R. Dale, H. Moisl, and H. Somers, Eds., *Handbook of Natural Language Processing*, 0 ed. CRC Press, 2000. doi: 10.1201/9780824746346.
- [2]J. Hutchins, “The history of machine translation in a nutshell,” 2006. [Online]. Available: <https://api.semanticscholar.org/CorpusID:231780839>
- [3]W. J. Hutchins and H. L. Somers, *An introduction to machine translation*, 2. printing. London: Academic Press, 1997.
- [4]. P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2009. [Online]. Available: <https://books.google.gr/books?id=kKYgAwAAQBAJ>
- [5]D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” 2014, doi: 10.48550/ARXIV.1409.0473.
- [6]A. Vaswani *et al.*, “Attention Is All You Need,” 2017, doi: 10.48550/ARXIV.1706.03762.
- [7]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018, doi: 10.48550/ARXIV.1810.04805.
- [8]K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.” arXiv, Sep. 02, 2014. Accessed: Aug. 19, 2023. [Online]. Available: <http://arxiv.org/abs/1406.1078>
- [9].R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks.” arXiv, Feb. 15, 2013. Accessed: Aug. 19, 2023. [Online]. Available: <http://arxiv.org/abs/1211.5063>
- [10]S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [11]I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks.” arXiv, Dec. 14, 2014. Accessed: Aug. 20, 2023. [Online]. Available: <http://arxiv.org/abs/1409.3215>
- [12]OpenAI, “GPT-4 Technical Report.” arXiv, Mar. 27, 2023. Accessed: Aug. 30, 2023. [Online]. Available: <http://arxiv.org/abs/2303.08774>
- [13]. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” 2022, doi: 10.48550/ARXIV.2204.06125.
- [14]A. Hendy *et al.*, “How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.” arXiv, Feb. 17, 2023. Accessed: Sep. 10, 2023. [Online]. Available: <http://arxiv.org/abs/2302.09210>
- [15]D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. in Prentice Hall series in artificial intelligence. Upper Saddle River, N.J: Prentice Hall, 2000.
- [16]“machinetranslate.org - Human Evaluation Metrics.” [Online]. Available: <https://machinetranslate.org/human-evaluation-metrics>
- [17]A. Lommel, H. Uszkoreit, and A. Burchardt, “Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics,” *Tradumàtica Tecnol. Trad.*, no. 12, pp. 455–463, Dec. 2014, doi: 10.5565/rev/tradumatica.77.
- [18]M. Snover, N. Madnani, B. Dorr, and R. Schwartz, “TERp system description,” in *MetricsMATR workshop at AMTA*, 2008, p. 108.
- [19]K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.
- [20]E. Reiter, “A Structured Review of the Validity of BLEU,” *Comput. Linguist.*, vol. 44, no. 3, pp. 393–401, Sep. 2018, doi: 10.1162/coli_a_00322.
- [21]S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [22] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A Study of Translation Edit Rate

- with Targeted Human Annotation,” in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, Aug. 2006, pp. 223–231. [Online]. Available: <https://aclanthology.org/2006.amta-papers.25>
- [23] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [24] M. Popović, “chrF: character n-gram F-score for automatic MT evaluation,” in *Proceedings of the tenth workshop on statistical machine translation*, 2015, pp. 392–395.
- [25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT.” arXiv, Feb. 24, 2020. Accessed: Sep. 10, 2023. [Online]. Available: <http://arxiv.org/abs/1904.09675>
- [26] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, “COMET: A Neural Framework for MT Evaluation.” arXiv, Oct. 19, 2020. Accessed: Sep. 20, 2023. [Online]. Available: <http://arxiv.org/abs/2009.09025>
- [27] R. Rei *et al.*, “COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task,” in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 578–585. [Online]. Available: <https://aclanthology.org/2022.wmt-1.52>
- [28] R. Rei *et al.*, “CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task.” arXiv, Sep. 13, 2022. Accessed: Sep. 21, 2023. [Online]. Available: <http://arxiv.org/abs/2209.06243>
- [29] X. Chen *et al.*, “Exploring Robustness of Machine Translation Metrics: A Study of Twenty-Two Automatic Metrics in the WMT22 Metric Task,” in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 530–540. [Online]. Available: <https://aclanthology.org/2022.wmt-1.46>
- [30] D. Mouratidis, M. Stasimioti, V. Sосoni, and K. L. Kermanidis, “NoDeeLe: A Novel Deep Learning Schema for Evaluating Neural Machine Translation Systems,” in *Proceedings of the Translation and Interpreting Technology Online Conference*, Held Online: INCOMA Ltd., Jul. 2021, pp. 37–47. [Online]. Available: <https://aclanthology.org/2021.triton-1.5>
- [31] K. Keeton, *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*. Berkeley, CA: USENIX Association, 2016.
- [32] F. Chollet and others, “Keras,” *Github*, [Online]. Available: <https://github.com/fchollet/keras>
- [33] S. Bird and others, “Natural Language Toolkit (NLTK).,” *Github*, [Online]. Available: <https://github.com/nltk/nltk>
- [34] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [35] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, “FastText.zip: Compressing text classification models,” *ArXiv Prepr. ArXiv161203651*, 2016.
- [36] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
- [37] T. pandas development team, “pandas-dev/pandas: Pandas.” Zenodo, Feb. 2020. doi: 10.5281/zenodo.3509134.
- [38] “StackOverflow,” 2023, [Online]. Available: <https://stackoverflow.com/>
- [39] “draw.io.” JGraph Ltd, 2023 2005. [Online]. Available: <https://www.drawio.com/>
- [40] ...D. Deutsch, J. Juraska, M. Finkelstein, and M. Freitag, “Training and Meta-Evaluating Machine Translation Evaluation Metrics at the Paragraph Level.” arXiv, Aug. 28, 2023. Accessed: Feb. 04, 2024. [Online]. Available: <http://arxiv.org/abs/2308.13506>
- [41] Θ. Πίσσας, “Αναγνώριση Ανθρώπινης Δράσης και Χειρονομιών χρησιμοποιώντας Συνελκτικά και Αναδρομικά Νευρωνικά Δίκτυα.” ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟΣΧΟΛΗ

ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ, Jul. 2017.

[42]Β. Ρουσκοπούλου, “Xml τεχνολογίες και εφαρμογή τους στην επεξεργασία της νεοελληνικής κοινής, σχεδιασμός κατασκευή και τεκμηρίωση ενός στελεχωτή stemmer για την ελληνική κοινή φυσική γλώσσα,” *Dr. Diss.*.

[43]I. Dicks, *Oxford Εικονογραφημένο Λεξικό Πληροφορικής Για όλους*. Athena: Aiora, 2008.