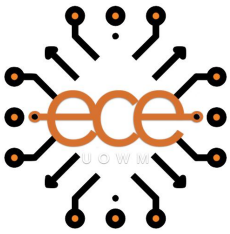
Πανεπιστήμιο Δυτικής Μακεδονίας
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών



Νικόλαος Λεύκος

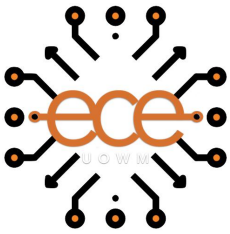
ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πλαίσιο σημασιολογικών επικοινωνιών με μοντέλα βασισμένα σε Transformer

Επιβλέπων: Δρ. Αλέξανδρος-Απόστολος Α. Μπουλογεώργος

Κοζάνη, Φεβρουάριος 2024

Πανεπιστήμιο Δυτικής Μακεδονίας
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών



Νικόλαος Λεύκος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πλαίσιο σημασιολογικών επικοινωνιών με μοντέλα βασισμένα σε Transformer

Επιβλέπων: Δρ. Αλέξανδρος-Απόστολος Α. Μπουλογεώργος

Κοζάνη, Φεβρουάριος 2024

Περίληψη

Με την εξέλιξη από την πρώτη γενιά (1G) στην πέμπτη γενιά (5G), ο ρυθμός μετάδοσης που επιτεύχθηκε βελτιώθηκε δεκάδες χιλιάδες φορές και η χωρητικότητα του συστήματος πλησιάζει σταδιακά το όριο Shannon. Επιπλέον, η εμφάνιση νέων ασύρματων εφαρμογών, παράγει πρωτοφανείς ποσότητες δεδομένων που θα επιβαρύνουν τη χωρητικότητα των σύγχρονων ασύρματων δικτύων. Καθώς τα δίκτυα έκτης γενιάς (6G) εισέρχονται στην εποχή των ευφυών επικοινωνιών, δίνεται έδαφος στις σημασιολογικές επικοινωνίες, ένα νέο παράδειγμα επικοινωνίας που επικεντρώνεται στην εξαγωγή και την εκμετάλλευση του νοήματος των δεδομένων κατά τη διάρκεια της επικοινωνίας. Σε αντίθεση με τις παραδοσιακές ασύρματες επικοινωνίες, οι σημασιολογικές επικοινωνίες μεταδίδουν μόνο τις απαραίτητες πληροφορίες, που σχετίζονται με τη συγκεκριμένη εργασία, στο δέκτη, γεγονός που οδηγεί σε σημαντική μείωση της κίνησης δεδομένων και βελτίωση της συνολικής απόδοσης του συστήματος. Επίσης, η επεξεργασία φυσικής γλώσσας, με την υποστήριξη της βαθιάς μάθησης, έχει σημειώσει μεγάλη επιτυχία στην ανάλυση και την κατανόηση μεγάλου αριθμού γλωσσικών κειμένων.

Στην παρούσα διπλωματική εργασία παρουσιάζεται η δημιουργία ενός διατερματικού πλαισίου σημασιολογικών επικοινωνιών για την αποστολή και λήψη δεδομένων κειμένου, με σκοπό την ανάδειξη των οφελών των σημασιολογικών επικοινωνιών στις ασύρματες επικοινωνίες. Η δημιουργία του εν λόγω πλαισίου περιλαμβάνει την υλοποίηση μοντέλων που αναλαμβάνουν την εξαγωγή των σημασιολογικών πληροφοριών ενός κειμένου και τη χρήση ενός μοντέλου για την ανακατασκευή των σημασιολογικών πληροφοριών. Η υλοποίηση των μοντέλων εξαγωγής σημασιολογικών πληροφοριών πραγματοποιήθηκε μέσω εκπαίδευσης και fine tuning προεκπαιδευμένων μοντέλων που βασίζονται σε transformer, χρησιμοποιώντας συγκεκριμένα σύνολα δεδομένων. Το μοντέλο που χρησιμοποιείται για την ανακατασκευή των σημασιολογικών πληροφοριών πρόκειται για ένα μεγάλο γλωσσικό μοντέλο βασισμένο σε transformer, το οποίο μέσω της εύρεσης και χρήσης κατάλληλης προτροπής, καθοδηγείται στη σωστή ανακατασκευή των σημασιολογικών πληροφοριών.

Αξιολογώντας το πλαίσιο σε μετρικές σημασιολογικών επικοινωνιών και εφαρμόζοντας το σε δεδομένα κειμένου, συμπεραίνουμε ότι το υλοποιημένο πλαίσιο μπορεί να χρησιμοποιηθεί για την αποστολή και λήψη δεδομένων κειμένου, ελαχιστοποιώντας τη μεταφορά περιττών πληροφοριών, μειώνοντας έτσι την κίνηση δεδομένων και βελτιστοποιώντας τη συνολική απόδοση του συστήματος.

Λέξεις κλειδιά: Σημασιολογικές επικοινωνίες, Επεξεργασία φυσικής γλώσσας, Εξαγωγή πληροφοριών, Βαθιά μάθηση, Transformers

Abstract

With the evolution from the first generation (1G) to the fifth generation (5G), the achieved transmission rate has improved tens of thousands of times and the system capacity is gradually approaching the Shannon limit. In addition, the emergence of new wireless applications produces unprecedented amounts of data that will strain the capacity of modern wireless networks. As sixth-generation (6G) networks enter the era of intelligent communications, they pave the way for semantic communications, a novel paradigm of communication, focused on extracting and exploiting the meaning of data during communication. Unlike traditional wireless communications, semantic communications transmit only the necessary information, related to the specific task, to the receiver, which leads to a significant reduction in data traffic and improvement of overall system performance. Also, natural language processing, with the support of deep learning, has been very successful in analyzing and understanding a large number of linguistic texts.

This diploma thesis presents the creation of an end-to-end semantic communications framework for sending and receiving text data, in order to highlight the benefits of semantic communications in wireless communications. The creation of this framework involves the implementation of models that undertake the extraction of semantic information of a text and the usage of a model to reconstruct semantic information. The implementation of semantic information extraction models was carried out through training and fine tuning pre-trained transformer-based models, in specific datasets. The model used to reconstruct semantic information is a large transformer-based language model, which through prompt engineering, is guided to the correct reconstruction of semantic information.

By evaluating the framework in semantic communications metrics and applying it to text data, we conclude that the implemented framework can be used to send and receive text data, while minimizing the transfer of unnecessary information, thereby reducing data traffic and optimizing overall system performance.

Keywords: Semantic communications, Natural language processing, Information extraction, Deep learning, Transformers

Δήλωση Πνευματικών Δικαιωμάτων

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο

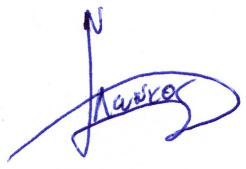
"Πλαίσιο σημασιολογικών επικοινωνιών με μοντέλα βασισμένα σε Transformer"

καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Αλέξανδρο-Απόστολο Α. Μπουλογεώργο

αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Υπογραφή Φοιτητή:



Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου προς τον κ. Αλέξανδρο Μπουλογεώργο για την πολύτιμη υποστήριξη που μου παρείχε, καθώς και για την εξαιρετική συνεργασία που αναπτύξαμε κατά την εκτέλεση της διπλωματικής εργασίας.

Θέλω επίσης να ευχαριστήσω την οικογένειά μου για τη στήριξη που μου προσέφερε καθ' όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Συντομογραφίες

BER	Bit Error Rate
BERT	Bidirectional Encoder Representation from Transformers
BLEU	Bilingual Evaluation Understudy
CR	Coreference Resolution
D2T	Data to Text
LLM	Large Language Model
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MLM	Masked Language Modelling
MSS	Mean Semantic Similarity
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
RC	Relation Classification
RDF	Resource Description Framework
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SemCom	Semantic Communications
SER	Symbol Error Rate
Seq2seq	Sequence to sequence
T5	Text-to-Text Transfer Transformer

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρα και στόχοι της διπλωματικής εργασίας

Τα κίνητρα πίσω από την υλοποίηση αυτής της διπλωματικής εργασίας είναι αρκετά. Πρώτον, η επιτακτική ανάγκη να αντιμετωπιστούν οι κλιμακούμενες απαιτήσεις για χωρητικότητα δεδομένων καθώς τα ασύρματα δίκτυα προχωρούν στην έκτη γενιά. Επιπλέον, ένα κίνητρο αποτελεί η αυξανόμενη χρήση συστημάτων σημασιολογικών επικοινωνιών σε ασύρματα δίκτυα. Προσφέροντας ένα νέο πλαίσιο σημασιολογικών επικοινωνιών, η διατριβή στοχεύει να προσθέσει στο υπάρχον τοπίο και να διερευνήσει περαιτέρω δυνατότητες σε αυτόν τον τομέα. Επιπλέον, ένα κίνητρο αποτελεί η πρόσφατη εμφάνιση νέων αρχιτεκτονικών βαθιάς μάθησης, όπως οι transformers, ικανών να επιτύχουν εξαιρετικά αποτελέσματα στην κατανόηση δεδομένων κειμένου. Η διατριβή στοχεύει στη διερεύνηση των πιθανών οφελών από την εφαρμογή ενός πλαισίου βασισμένου σε αυτές τις προηγμένες αρχιτεκτονικές στο πλαίσιο των σημασιολογικών επικοινωνιών.

Κεντρικός στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη και αξιολόγηση ενός πλαισίου σημασιολογικών επικοινωνιών. Το πλαίσιο αυτό στοχεύει στη βελτιστοποίηση της μεταφοράς δεδομένων εστιάζοντας στη μετάδοση μόνο βασικών πληροφοριών που σχετίζονται με συγκεκριμένες εργασίες, ενισχύοντας έτσι τη συνολική απόδοση του συστήματος. Πέρα από τη θεωρητική διερεύνηση, ο στόχος είναι να καταδειχθεί η πρακτική εφαρμογή των σημασιολογικών επικοινωνιών σε πραγματικά σενάρια. Επίσης, η διπλωματική εργασία στοχεύει στο να διερευνήσει και να αναλύσει τα πλεονεκτήματα και τις δυνατότητες ενσωμάτωσης της αρχιτεκτονικής βαθιάς μάθησης transformer στο πλαίσιο των σημασιολογικών επικοινωνιών.

1.2 Συνεισφορά της διπλωματικής εργασίας

Η συνεισφορά της παρούσας διπλωματικής εργασίας είναι πολύπλευρη. Πρώτον, προσπαθεί να συμβάλει στην εξέλιξη των δικτύων νέας γενιάς εισάγοντας ένα πλαίσιο σημασιολογικών επικοινωνιών. Η έμφαση του πλαισίου στη μείωση της κίνησης δεδομένων και τη βελτιστοποίηση της αποτελεσματικότητας της επικοινωνίας προσδοκεί να προσφέρει λύσει στις τρέχουσες και μελλοντικές προκλήσεις των ασύρματων δικτύων.

Επιπλέον, η εργασία εισάγει ελαφριά μοντέλα για την εξαγωγή σημασιολογικών πληροφοριών, η οποία επιτυγχάνεται με το fine-tuning του DistilBERT, μια μικρότερη και πιο αποδοτική έκδοση του BERT, σε σύγκριση με τα μεγαλύτερα μοντέλα transformer που χρησιμοποιούνται συνήθως σε πλαίσια σημασιολογικών

επικοινωνιών. Με αυτό τον τρόπο αναδεικνύεται πως τα μοντέλα αυτά διατηρούν υψηλή απόδοση ενώ είναι αποδοτικά ως προς τους πόρους.

Επιπλέον, τα μοντέλο που αναπτύχθηκε για την εξαγωγή σχέσεων, εκπαιδεύεται στο σύνολο δεδομένων T-REx. Αυτό το μεγάλο σύνολο δεδομένων περιλαμβάνει ένα ευρύ φάσμα διαφορετικών σχέσεων και έτσι το συνολικό πλαίσιο διαφοροποιείται από τα υπάρχοντα πλαίσια που συχνά εκτελούν εξαγωγή σχέσεων σε μικρότερα σύνολα δεδομένων που χρησιμοποιούνται για benchmarks ή σύνολα δεδομένων προσαρμοσμένα σε συγκεκριμένα θέματα, όπως επιστημονικές δημοσιεύσεις. Αυτή η επιλογή σχεδιασμού επιτρέπει στο μοντέλο εξαγωγής σχέσεων της διπλωματικής εργασίας να είναι πιο ευέλικτο και εφαρμόσιμο σε πιο γενικά καινούρια δεδομένα.

Επί προσθέτως, τα μοντέλα που αναπτύχθηκαν για την εξαγωγή σημασιολογικών πληροφοριών μπορούν να εξυπηρετήσουν το σκοπό της συμπλήρωσης και της δημιουργίας βάσεων γνώσης. Οι βάσεις γνώσεων είναι ζωτικής σημασίας για διάφορες εφαρμογές, συμπεριλαμβανομένης της ανάκτησης πληροφοριών, των συστημάτων απάντησης ερωτήσεων και των σημασιολογικών μηχανών αναζήτησης, ενισχύοντας την προσβασιμότητα και τη χρήση δομημένων πληροφοριών.

Τέλος, αυτή η διπλωματική εργασία αποτελεί μια πρωτοποριακή προσπάθεια για την αξιοποίηση του μεγάλου γλωσσικού μοντέλου Flan-T5 για την ανακατασκευή σημασιολογικών πληροφοριών σε κείμενο. Σε αντίθεση με τα υπάρχοντα μοντέλα που χρησιμοποιούν μοντέλα προτροπής όπως το GPT-3.5 για τη δημιουργία κειμένου, τα οποία συχνά έρχονται με περιορισμούς χρήσης, το Flan-T5 προσφέρει μια νέα και ελεύθερα προσβάσιμη προσέγγιση, επεκτείνοντας τη χρήση μεγάλων γλωσσικών μοντέλων σε αυτόν τον τομέα.

1.3 Οργάνωση της διπλωματικής εργασίας

Η δομή των επερχόμενων κεφαλαίων της διπλωματικής διαμορφώνεται ως εξής:

Το Κεφάλαιο 2 περιλαμβάνει την ανάλυση του θεωρητικού υπόβαθρου που είναι απαραίτητο για την κατανόηση των εννοιών, των μεθόδων και των διαδικασιών που ακολουθήθηκαν για την υλοποίηση των μοντέλων των κεφαλαίων 4 και 5 καθώς και για την αξιολόγηση αυτών των μοντέλων για την εξαγωγή συμπερασμάτων.

Στο Κεφάλαιο 3 αναφέρονται τα εργαλεία που ήταν απαραίτητα και χρησιμοποιήθηκαν για την υλοποίηση και την αξιολόγηση των μοντέλων των κεφαλαίων 4 και 5.

Στο Κεφαλαίο 4 αναλύονται εκτενώς οι τεχνικές και οι διαδικασίες που εφαρμόστηκαν για τη δημιουργία και τη χρήση των μοντέλων που ευθύνονται για την εξαγωγή σημασιολογικών πληροφοριών από δεδομένα κειμένου.

Στο Κεφάλαιο 5 εξηγούνται λεπτομερώς οι τεχνικές και οι διαδικασίες που εφαρμόστηκαν για την ανακατασκευή των σημασιολογικών πληροφοριών και την παραγωγή τους σε κείμενο. Επιπλέον, η ενότητα 5.2 περιλαμβάνει την αξιολόγηση του συνολικού πλαισίου σημασιολογικών επικοινωνιών που υλοποιήθηκε, σε ένα σύνολο μετρικών σημασιολογικών επικοινωνιών.

Στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα που προέκυψαν μετά την αξιολόγηση και τη χρήση του δημιουργημένου πλαισίου σε ένα μεγάλο πλήθος κειμένων. Τέλος, προτείνονται διάφορες πιθανές μελλοντικές επεκτάσεις που μπορεί να συμβάλουν στη βελτίωση του δημιουργημένου πλαισίου.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

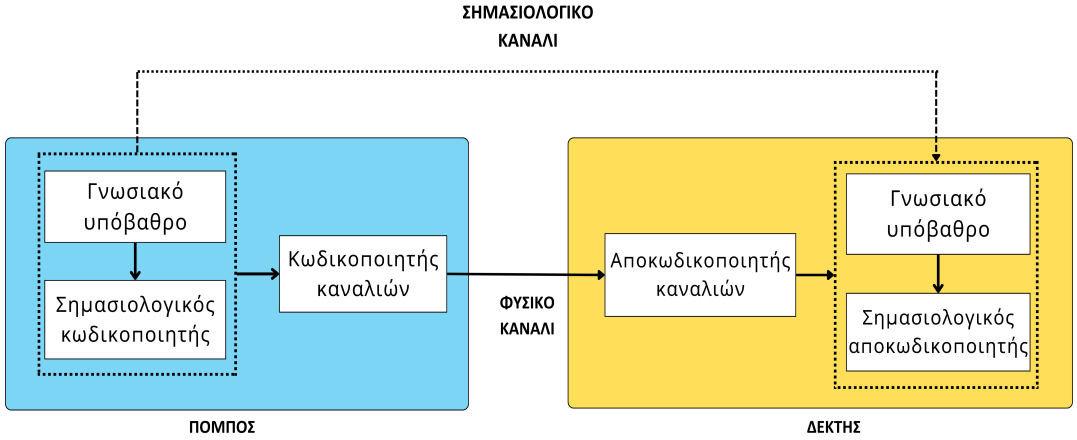
2.1 Σημασιολογικές Επικοινωνίες

Οι σημασιολογικές επικοινωνίες (semantic communications) αποτελούν ένα νέο παράδειγμα επικοινωνίας, που ως στόχο έχει να μεταφέρει το νόημα ή τη σημασιολογία ενός μηνύματος στους προοριζόμενους χρήστες, αντί να επικεντρώνεται αποκλειστικά στην ακριβή λήψη μεμονωμένων συμβόλων ή bit, όπως στις παραδοσιακές ασύρματες επικοινωνίες [1]. Σε ένα συμβατικό σύστημα επικοινωνίας, η πληροφορία που στέλνει ο δέκτης μετατρέπεται σε ακολουθία bit για επεξεργασία. Στον δέκτη, η ακολουθία bit που αντιπροσωπεύει την πληροφορία ανακτάται με ακρίβεια. Σε αυτό το σύστημα ο ρυθμός μετάδοσης bit/συμβόλου οριοθετείται από τη χωρητικότητα Shannon. Οι σημασιολογικές επικοινωνίες αντιθέτως μεταδίδουν μόνο το νόημα της πληροφορίας.

Μία από τις κρίσιμες διαφορές είναι η εισαγωγή της σημασιολογικής κωδικοποίησης, η οποία συλλαμβάνει τα σημασιολογικά χαρακτηριστικά, ανάλογα με τις εργασίες ή τις ενέργειες που πρέπει να εκτελεστούν στον δέκτη. Έτσι, θα μεταδοθούν μόνο αυτά τα σημασιολογικά χαρακτηριστικά, γεγονός που μειώνει σημαντικά τους απαιτούμενους πόρους επικοινωνίας. Οι εργασίες στον δέκτη θα μπορούσαν να είναι η ανακατασκευή δεδομένων ή κάποιες πιο έξυπνες εργασίες, όπως η ταξινόμηση εικόνων και η μετάφραση γλώσσας [2]. Σε αντίθεση με τις παραδοσιακές ασύρματες επικοινωνίες, στις σημασιολογικές επικοινωνίες δεν απαιτείται η ακολουθία αποκωδικοποίησης στην πλευρά του δέκτη να είναι αυστηρά συνεπής με την ακολουθία κωδικοποίησης στην πλευρά του αποστολέα. Αυτό που απαιτείται είναι οι σημασιολογικές πληροφορίες που ανακτώνται στην πλευρά του δέκτη να ταιριάζουν με τις σημασιολογικές πληροφορίες που μεταδίδονται στην πλευρά του αποστολέα [3].

Στο Σχήμα 2.1 απεικονίζονται τα κύρια μέρη ενός συστήματος σημασιολογικών επικοινωνιών. Στο σύστημα αυτό, περιλαμβάνονται το σημασιολογικό επίπεδο και το επίπεδο της μετάδοσης. Το σημασιολογικό επίπεδο αφορά την επεξεργασία σημασιολογικών πληροφοριών για την απόκτηση σημασιολογικής αναπαράστασης. Για παράδειγμα, σε ένα σύστημα σημασιολογικών επικοινωνιών για την αποστολή και λήψη δεδομένων κειμένου, ο σημασιολογικός κωδικοποιητής είναι υπεύθυνος για την εξαγωγή των σημασιολογικών πληροφοριών των προτάσεων και την κωδικοποίηση τους σε μία πιο αποδοτική μορφή, ενώ ο σημασιολογικός αποκωδικοποιητής είναι υπεύθυνος για την ερμηνεία και την εξαγωγή των ληφθέντων σημασιολογικών πληροφοριών. Το επίπεδο μετάδοσης εγγυάται την επιτυχή λήψη συμβόλων στον δέκτη μετά τη διέλευση από το μέσο μετάδοσης. Αυτό πραγματοποιείται από τον κωδικοποιητή και τον αποκωδικοποιητή καναλιών. Ο σημασιολογικός πομπός και ο σημασιολογικός δέκτης είναι εξοπλισμένοι με ορισμένες βασικές γνώσεις για τη

διευκόλυνση της εξαγωγής σημασιολογικών χαρακτηριστικών, όπου οι βασικές γνώσεις θα μπορούσαν να είναι διαφορετικές για την εκάστοτε εφαρμογή.



Σχήμα 2.1: Τα κύρια μέρη ενός συστήματος σημασιολογικών επικοινωνιών

Όταν το κανάλι επικοινωνίας έχει θόρυβο, τότε το μήνυμα που λαμβάνει ο δέκτης συνήθως είναι παραμορφωμένο. Σύμφωνα με τη θεωρία του Shannon, τα σφάλματα που προκαλούνται λόγω του θορύβου, μπορούν να μετρηθούν από τον λόγο δυφιακών σφαλμάτων (bit error rate) ή από τον λόγο συμβολικών σφαλμάτων (symbol error rate). Στο σημασιολογικό επίπεδο, τα σφάλματα που προκαλούνται λόγω σημασιολογικού θορύβου μπορούν να μετρηθούν μέσω της σημασιολογικής αναντιστοιχίας. Ο σημασιολογικός θόρυβος αναφέρεται στη διαταραχή που επηρεάζει την ερμηνεία του μηνύματος, η οποία θα μπορούσε να αντιμετωπιστεί ως αναντιστοιχία σημασιολογικών πληροφοριών μεταξύ του πομπού και του δέκτη [2]. Η εξαγωγή των σημασιολογικών πληροφοριών, η σημασιολογική κωδικοποίηση τους και η ερμηνεία τους επιτυγχάνεται με τη χρήση τεχνικών που υπάγονται στους κλάδους της επεξεργασίας φυσικής γλώσσας (Natural Language Processing) και της βαθιάς μάθησης (deep learning).

2.1.1 Επεξεργασία Φυσικής Γλώσσας (natural language processing)

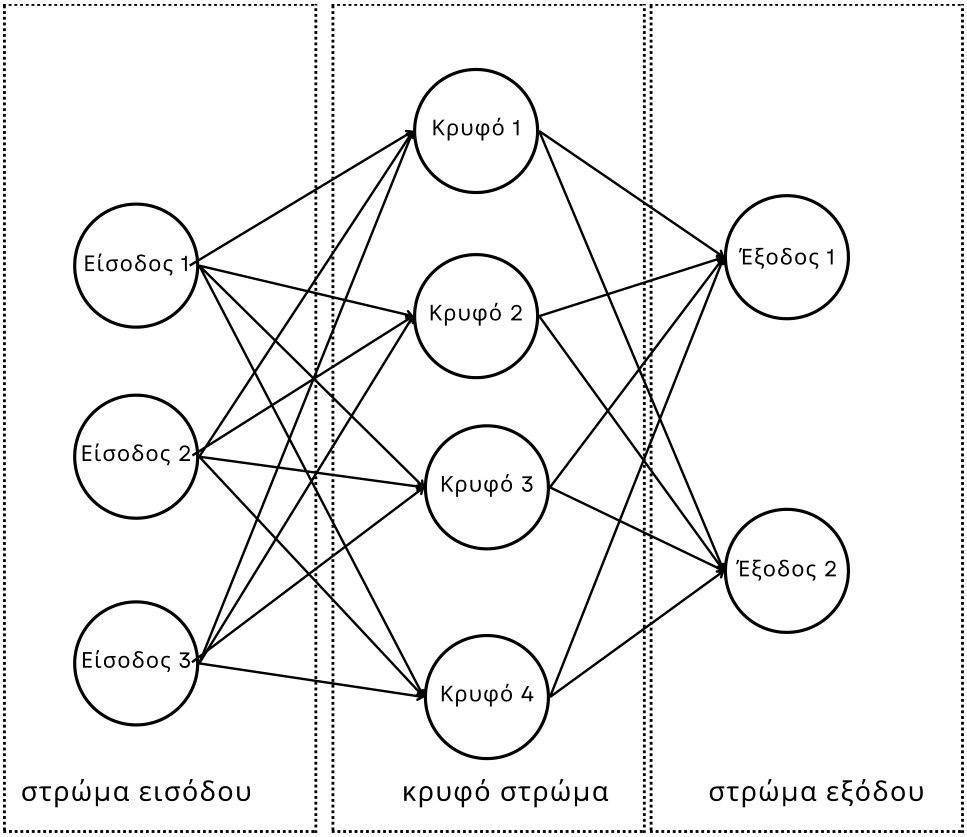
Η επεξεργασία φυσικής γλώσσας (NLP) είναι ένα διεπιστημονικό υποπεδίο της επιστήμης των υπολογιστών και της γλωσσολογίας. Ασχολείται πρωτίστως με την παροχή της δυνατότητας στους υπολογιστές να υποστηρίζουν και να χειρίζονται την ανθρώπινη γλώσσα. Περιλαμβάνει την επεξεργασία συνόλων δεδομένων φυσικής γλώσσας όπως σώματα κειμένων, χρησιμοποιώντας προσεγγίσεις μηχανικής μάθησης είτε βασισμένες σε κανόνες είτε πιθανοτικές (στατιστικές και πιο πρόσφατα, βασισμένες σε νευρωνικά δίκτυα) [4].

2.1.2 Βαθιά Μάθηση

Η βαθιά μάθηση είναι ένας κλάδος της μηχανικής μάθησης που περιλαμβάνει τη χρήση τεχνητών νευρωνικών δικτύων με πολλαπλά επίπεδα για την εκμάθηση και κατανόηση σύνθετων αναπαραστάσεων δεδομένων. Το επίθετο "βαθιά" αναφέρεται στη χρήση πολλαπλών επιπέδων στο δίκτυο. Οι μέθοδοι που χρησιμοποιούνται μπορούν να είναι είτε επιβλεπόμενοι (supervised), είτε ημι-επιβλεπόμενοι (semi-supervised) είτε μη επιβλεπόμενοι (unsupervised). Αρχιτεκτονικές βαθιάς μάθησης όπως βαθιά νευρωνικά δίκτυα (deep neural networks), αναδρομικά νευρωνικά δίκτυα (recurrent neural networks), συνελικτικά νευρωνικά δίκτυα (convonutional neural networks) και transformers έχουν εφαρμοστεί σε τομείς όπως η όραση υπολογιστών, η αναγνώριση ομιλίας, η επεξεργασία φυσικής γλώσσας, η μηχανική μετάφραση, η βιοπληροφορική κ.α. και παράγουν αποτελέσματα συγκρίσιμα και σε ορισμένες περιπτώσεις καλύτερα των ανθρώπινων επιδόσεων [4].

2.1.3 Νευρωνικά δίκτυα

Νευρωνικό δίκτυο είναι το εμπνευσμένο από τον ανθρώπινο εγκέφαλο υπολογιστικό μοντέλο, σχεδιασμένο να αναγνωρίζει μοτίβα και να λαμβάνει αποφάσεις με βάση δεδομένα. Αποτελείται από διασυνδεδεμένους κόμβους (νευρώνες) οργανωμένους σε στρώματα: στρώμα εισόδου, κρυφά στρώματα και στρώμα εξόδου [5].



Σχήμα 2.2: Αρχιτεκτονική νευρωνικού δικτύου

Κάθε νευρώνας υπολογίζει ένα σταθμισμένο άθροισμα (weighted sum) των εισόδων του, προσθέτει έναν όρο μεροληψίας (bias) και εφαρμόζει μια συνάρτηση ενεργοποίησης (activation function) για να παράγει την έξοδο του.

Η συνάρτηση ενεργοποίησης είναι μια μαθηματική συνάρτηση που εφαρμόζεται στην έξοδο ενός νευρώνα σε ένα νευρωνικό δίκτυο, εισάγοντας μη γραμμικότητα και επιτρέποντας στο δίκτυο να μάθει πολύπλοκα μοτίβα.

Το δίκτυο μαθαίνει προσαρμόζοντας τα βάρη και τις μεροληψίες του μέσω μιας διαδικασίας που περιλαμβάνει τη διάδοση προς τα εμπρός (forward propagation), τον υπολογισμό απωλειών (loss computation), την οπισθοδιάδοση (backpropagation)και την ενημέρωση των παραμέτρων χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης.

1. Διάδοση προς τα εμπρός (forward propagation): Τα δεδομένα εισόδου περνούν μέσα από ένα νευρωνικό δίκτυο για τον υπολογισμό των προβλέψεων εξόδου. Κατά τη διάρκεια αυτής της διαδικασίας, κάθε νευρώνας υπολογίζει ένα σταθμισμένο άθροισμα των εισόδων του, προσθέτει έναν όρο μεροληψίας και χρησιμοποιεί τη συνάρτηση ενεργοποίησης για να παράγει μια έξοδο. Οι έξοδοι από όλους τους κόμβους μέσα σε ένα επίπεδο χρησιμοποιούνται στη συνέχεια ως είσοδοι για το επόμενο στρώμα, συνεχίζοντας έως ότου το τελικό στρώμα εξόδου δημιουργήσει τις προβλέψεις.

2. Υπολογισμός απωλειών: Η διαφορά μεταξύ των προβλέψεων ενός νευρωνικού δικτύου και των πραγματικών τιμών-στόχων υπολογίζεται χρησιμοποιώντας μια συνάρτηση απωλειών, η οποία είναι μια μαθηματική έκφραση που μετρά αυτή τη διαφορά. Ο στόχος κατά τη διάρκεια της εκπαίδευσης είναι να ελαχιστοποιηθεί αυτή η διαφορά, βελτιώνοντας έτσι την απόδοση του μοντέλου.

3. Οπισθοδιάδοση (backpropagation): Η οπισθοδιάδοση είναι μία τεχνική που χρησιμοποιείται για τον προσδιορισμό του τρόπου προσαρμογής των βαρών και των μεροληψιών του μοντέλου προκειμένου να ελαχιστοποιηθούν οι απώλειες. Περιλαμβάνει τον υπολογισμό των κλίσεων της συνάρτησης απωλειών σε σχέση με κάθε βάρος και μεροληψία εφαρμόζοντας τον κανόνα αλυσίδας του διαφορικού λογισμού. Μια κλίση μετρά το μεγαλύτερο ρυθμό μεταβολής της συνάρτησης απωλειών σε σχέση με τα βάρη και τις μεροληψίες του μοντέλου.

4. Ενημέρωση παραμέτρων: Αυτή η διαδικασία προσαρμόζει τις παραμέτρους προς την αντίθετη κατεύθυνση από τις κλίσεις που υπολογίζονται κατά τη διάρκεια της οπισθοδιάδοσης για να μειώσει το σφάλμα μεταξύ των προβλέψεων του δικτύου και των πραγματικών τιμών-στόχων.

Μέσω της επανάληψης, διαπερνώντας ολόκληρο το σύνολο δεδομένων πολλές φορές (εποχές), το νευρωνικό δίκτυο βελτιώνει τα βάρη και τις μεροληψίες του κατά τη διάρκεια κάθε περάσματος. Αυτή η συνεχής προσαρμογή συμβάλλει στην ελαχιστοποίηση των απωλειών και ενισχύει την ικανότητα του δικτύου να κάνει ακριβείς προβλέψεις με βάση τα δεδομένα εισόδου.

2.1.4 Βαθιά Μάθηση και Σημασιολογικές Επικοινωνίες

Η βαθιά μάθηση χρησιμοποιείται στις σημασιολογικές επικοινωνίες για να καλύψει τις θεμελιώδεις ανάγκες των συστημάτων τους, όπως η εξαγωγή σημασιολογικών πληροφοριών και η ανακατασκευή των δεδομένων, αλλά και για να βελτιστοποιήσει τις λειτουργίες των συστημάτων σημασιολογικών επικοινωνιών που έχουν δημιουργηθεί.

Για παράδειγμα, οι Xie et al. [6] προτείνουν ένα σύστημα σημασιολογικών επικοινωνιών βασισμένο σε βαθιά μάθηση που ονομάζεται DeepSC, το οποίο στοχεύει στη μεγιστοποίηση της χωρητικότητας του συστήματος και στην ελαχιστοποίηση των σημασιολογικών σφαλμάτων στη μετάδοση κειμένου. Στο συγκεκριμένο σύστημα οι σημασιολογικές πληροφορίες εξάγονται αποτελεσματικά με την υλοποίηση ενός πλαισίου που βασίζεται στην αρχιτεκτονική των transformers.

Οι Wang et al. [7] προτείνουν ένα πλαίσιο σημασιολογικών επικοινωνιών για ασύρματα δίκτυα, όπου οι σταθμοί βάσης εξάγουν σημασιολογικές πληροφορίες από δεδομένα κειμένου και τις μεταδίδουν στους χρήστες χρησιμοποιώντας γράφους γνώσης (knowledge graphs). Στη συνέχεια, οι χρήστες ανακτούν το αρχικό κείμενο χρησιμοποιώντας ένα μοντέλο παραγωγής κειμένου από γράφο. Η απόδοση του

συστήματος υπολογίζεται χρησιμοποιώντας μια μετρική σημασιολογικής ομοιότητας. Το πλαίσιο βελτιστοποιείται με τη διατύπωση του προβλήματος ως εργασία βελτιστοποίησης για τη μεγιστοποίηση της συνολικής μέσης σημασιολογικής ομοιότητας (mean semantic similarity-MSS) όλων των κειμένων που ανακτώνται από τους χρήστες, ικανοποιώντας παράλληλα τις απαιτήσεις της μετάδοσης. Αυτό το πρόβλημα βελτιστοποίησης περιλαμβάνει τη βελτιστοποίηση της εκχώρησης resource blocks (RB) και τον προσδιορισμό των μερικών σημασιολογικών πληροφοριών που μεταδίδονται σε κάθε χρήστη. Για την επίλυση αυτού του προβλήματος βελτιστοποίησης, το έγγραφο προτείνει έναν αλγόριθμο attention policy gradient (APG). Αυτός ο αλγόριθμος αξιολογεί τη σημασία κάθε σημασιολογικής τριπλέτας στη σημασιολογική πληροφορία και αναλύει τη σχέση μεταξύ της κατανομής σημαντικότητας των τριπλετών και του συνολικού MSS.

Οι Zhang et al. [8] αναπτύσσουν ένα ενοποιημένο σύστημα σημασιολογικών επικοινωνιών (U-DeepSC) που μπορεί να χειριστεί πολλαπλές εργασίες με διαφορετικούς τύπους δεδομένων. Οι συγγραφείς προτείνουν ένα δυναμικό σχήμα για την προσαρμογή του αριθμού των μεταδιδόμενων συμβόλων και χαρακτηριστικών με βάση τις συνθήκες της εργασίας και του καναλιού. Ο κύριος σκοπός του συστήματος U-DeepSC είναι να χειρίζεται ταυτόχρονα εργασίες που αφορούν διαφορετικούς τύπους δεδομένων (εικόνα, κείμενο, ομιλία). Το U-DeepSC στοχεύει στην εξαγωγή πληροφοριών που σχετίζονται με τη συγκεκριμένη εργασία μεταξύ όλων των εργασιών. Για αυτό το σκοπό, εισάγονται ενσωματώσεις διανυσμάτων εργασιών (task embedding vectors) και πίνακες ερωτημάτων εργασιών (task query matrices). Επιπλέον, το σύστημα έχει σχεδιαστεί για να προσαρμόζει δυναμικά το επίβαρο μετάδοσης με βάση τις συνθήκες του καναλιού και τη συγκεκριμένη εργασία. Αυτή η προσαρμοστικότητα επιτρέπει την αποτελεσματικότερη χρήση των πόρων και τη βελτίωση της απόδοσης σε διάφορες εργασίες.

2.1.5 Μετρικές αξιολόγησης συστημάτων σημασιολογικών επικοινωνιών

Στα συμβατικά συστήματα επικοινωνίας, οι μετρικές αξιολόγησης είναι το BER και το SER. Αυτές οι μετρικές δεν είναι κατάλληλες για την αξιολόγηση των συστημάτων SemCom, καθώς το επίκεντρο των επικοινωνιών μετατοπίζεται από την ακριβή μετάδοση συμβόλων στην αποτελεσματική ανταλλαγή σημασιολογικών πληροφοριών. Παρακάτω αναλύονται οι μετρικές αξιολόγησης που χρησιμοποιούνται στα συστήματα σημασιολογικών επικοινωνιών για δεδομένα κειμένου.

BLEU SCORE: Η βαθμολογία BLEU (bilingual evaluation understudy) είναι μια ευρέως χρησιμοποιούμενη μετρική η οποία μετράει την ποιότητα του κειμένου μετά από μηχανική μετάφραση. Η βαθμολογία BLEU έχει αξιοποιηθεί για τη μέτρηση συστημάτων σημασιολογικών επικοινωνιών για μετάδοση κειμένου. Η βαθμολογία BLEU μεταξύ της μεταδιδόμενης πρότασης s και της ληφθείσας πρότασης $\hat{s}$ υπολογίζεται ως:

$$\log BLEU = min\left(1 - \frac{l\hat{s}}{ls}, 0\right) + \sum_{n=1}^{N} u_n \log P_n$$

όπου ls και $l\hat{s}$ είναι το μήκος λέξεων των προτάσεων s και $\hat{s}$ αντίστοιχα, ενώ το u_n ορίζει τα βάρη των n-grams και το P_n είναι η βαθμολογία n-grams που ορίζεται ως:

$$P_n = \frac{\sum_k min\left(C_k(\hat{s}), C_k(s)\right)}{\sum_k min\left(C_k(\hat{s})\right)}$$

όπου C_k είναι η συνάρτηση μέτρησης συχνότητας για το k-οστό στοιχείο στο n-οστό gram. Η βαθμολογία BLEU μετρά τη διαφορά των n-grams μεταξύ δύο προτάσεων, όπου τα n-grams αναφέρονται στον αριθμό των λέξεων σε μια ομάδα λέξεων για σύγκριση. Για παράδειγμα, για την πρόταση "This is a dog", οι ομάδες λέξεων είναι "this", "is", "a" και "dog" για 1-gram. Για 2-grams, οι ομάδες λέξεων περιλαμβάνουν "this is", "is a" και "a dog". Ο ίδιος κανόνας ισχύει και για τα υπόλοιπα n-grams.

Το εύρος της βαθμολογίας BLEU κυμαίνεται μεταξύ 0 και 1. Όσο υψηλότερη είναι η βαθμολογία, τόσο μεγαλύτερη είναι η ομοιότητα μεταξύ των δύο προτάσεων. Ωστόσο, λίγες ανθρώπινες μεταφράσεις θα επιτύχουν τη βαθμολογία του 1, καθώς προτάσεις με διαφορετικές εκφράσεις ή λέξεις μπορεί να αναφέρονται στο ίδιο νόημα. Για παράδειγμα, οι προτάσεις "my bicycle was stolen" και "my bike was stolen" έχουν το ίδιο νόημα, αλλά η βαθμολογία BLEU δεν είναι 1, καθώς διαφέρουν όταν συγκρίνονται λέξη προς λέξη. Για να αποδοθεί ένα τέτοιο χαρακτηριστικό, η ομοιότητα προτάσεων (sentence similarity) προτάθηκε ως νέα μέτρηση για τη μέτρηση της σημασιολογικής ομοιότητας δύο προτάσεων [2].

Sentence similarity: Μια λέξη μπορεί να πάρει διαφορετικές σημασίες σε διαφορετικά πλαίσια. Για παράδειγμα, η λέξη ποντίκι έχει διαφορετική σημασία στη βιολογία και διαφορετική στους υπολογιστές. Η ομοιότητα μεταξύ της αρχικής πρότασης s και της ανακτηθείσας πρότασης $\hat{s}$ μετριέται ως εξής:

$$match(\hat{s}, s) = \frac{B_\Phi(s) \cdot B_\Phi(\hat{s})^T}{\left\|B_\Phi(s)\right\| \left\|B_\Phi(\hat{s})\right\|}$$

όπου το B_Φ αντιπροσωπεύει το BERT, ένα τεράστιο προ-εκπαιδευμένο μοντέλο, που περιλαμβάνει εκατομμύρια παραμέτρους, που χρησιμοποιούνται για την εξαγωγή της σημασιολογικής πληροφορίας. Η ομοιότητα της πρότασης που ορίζεται στον παραπάνω τύπο είναι ένας αριθμός μεταξύ 0 και 1, ο οποίος δείχνει πόσο παρόμοια είναι η αποκωδικοποιημένη πρόταση με τη μεταδιδόμενη πρόταση, με το 1 να αντιπροσωπεύει την υψηλότερη ομοιότητα και το 0 να αντιπροσωπεύει καμία ομοιότητα μεταξύ s και $\hat{s}$. Σε σύγκριση με τη βαθμολογία BLEU, το BERT έχει τροφοδοτηθεί από εκατομμύρια προτάσεις. Ως εκ τούτου, έχει ήδη μάθει τις σημασιολογικές πληροφορίες από αυτές τις προτάσεις και μπορεί να δημιουργήσει διαφορετικά σημασιολογικά διανύσματα για διαφορετικά πλαίσια αποτελεσματικά [6].

METEOR: Το METEOR (Metric for Evaluation of Translation with Explicit ORdering) είναι μια αυτόματη μετρική που χρησιμοποιείται συνήθως για την αξιολόγηση της ποιότητας των μεταφράσεων που δημιουργούνται από μηχανές. Έχει σχεδιαστεί για να αντιμετωπίζει ορισμένους περιορισμούς άλλων μετρικών όπως το BLEU. Η μετρική METEOR ενσωματώνει πληροφορίες των precision, recall και alignment με βάση την αντιστοίχιση unigram (1-gram) μεταξύ της μετάφρασης που δημιουργείται από τη μηχανή και ενός συνόλου μεταφράσεων αναφοράς που παράγονται από ανθρώπους. Η μέτρηση λαμβάνει επίσης υπόψιν το stemming και τη συνωνυμία χρησιμοποιώντας το WordNet.

Ο τύπος υπολογισμού του METEOR ορίζεται ως:

$$METEOR = (1 - Pen) \cdot F$$

Όπου Pen είναι ο συντελεστής ποινής και F είναι ο αρμονικός μέσος όρος που συνδυάζει precision (Pm) και recall (Rm) όπως δίνεται από τον τύπο:

$$F = \frac{(Pm \cdot Rm)}{a \cdot Pm + (1 - a) Rm}$$

Όπου α είναι μια υπερπαράμετρος σύμφωνα με το WordNet.

Οι τιμές του precision (Pm) και του recall (Rm) υπολογίζονται με βάση την αντιστοίχιση των unigrams μεταξύ της μηχανικής μετάφρασης και των μεταφράσεων αναφοράς. Ο όρος Pen αντιπροσωπεύει τις διαφορές στη σειρά των λέξεων [9].

ROUGE: Το ROUGE (Recall-Oriented Understudy for Gisting Evaluation) είναι ένα σύνολο μετρικών για την αξιολόγηση μοντέλων παραγωγής κειμένου (σύνοψη ή μηχανική μετάφραση). Το ROUGE βασίζεται στη μέτρηση της επικάλυψης μεταξύ της πρόβλεψης του μοντέλου και της αναφοράς που παράγεται από τον άνθρωπο. Το ROUGE έχει διάφορες παραλλαγές, όπου κάθε παραλλαγή αντιστοιχεί σε συγκεκριμένα n-grams που αλληλεπικαλύπτονται (ROUGE-1, ROUGE-2 κ.ο.κ.). Επίσης, μια τρίτη πιο χρησιμοποιούμενη παραλλαγή ROUGE, η ROUGE-L, χρησιμοποιεί τη μεγαλύτερη κοινή υποακολουθία (longest common subsequence-lcs) κοινών, όχι απαραίτητα διαδοχικών, ταξινομημένων λέξεων μεταξύ δύο προτάσεων. Το ROUGE-N περιλαμβάνει τον υπολογισμό των precision, recall και F-score με βάση τα n-grams που μοιράζονται μεταξύ της προβλεπόμενης περίληψης και της περίληψης αναφοράς. Το ROUGE-L υπολογίζει το F-score με βάση το lcs μεταξύ της προβλεπόμενης περίληψης και της περίληψης αναφοράς. Εξετάζει το precision και το recall του lcs για να αξιολογήσει την ποιότητα της παραγόμενης περίληψης σε σύγκριση με την περίληψη αναφοράς. Στις περισσότερες περιπτώσεις χρησιμοποιούνται τα ROUGE-1, ROUGE-2 και ROUGE-L. Ο υπολογισμός των ROUGE-N και ROUGE-L γίνεται ως εξής :

$$ROUGE - N\ Precision = \frac{Number\ of\ overlapping\ n-grams\ in\ the\ generated\ text}{Total\ number\ of\ n-grams\ in\ the\ generated\ text}$$

$$ROUGE - N\ Recall = \frac{Number\ of\ overlapping\ n-grams\ in\ the\ generated\ text}{Total\ number\ of\ n-grams\ in\ the\ reference\ text}$$

$$ROUGE-N\,F1=\frac{2\cdot ROUGE-N\,Precision\cdot ROUGE-N\,Recall}{ROUGE-N\,Precision+ROUGE-N\,Recall}$$

$$ROUGE-L=\frac{\left(1+\beta^2\right)\cdot Cl\cdot Pl}{Cl+\beta^2\cdot Pl}$$

Όπου Cl είναι το recall βάσει lcs, Plείναι το precision βάσει lcs και β είναι μια παράμετρος που ελέγχει τη σημασία του precision συγκριτικά με το recall [10].

BERTScore: Το BERTScore [46] είναι μια μετρική που χρησιμοποιείται για την αξιολόγηση της ποιότητας του κειμένου που δημιουργείται από μηχανή, συγκρίνοντάς το με το κείμενο αναφοράς . Έχει σχεδιαστεί ειδικά για την αξιολόγηση της ομοιότητας μεταξύ του παραγόμενου κειμένου και του κειμένου αναφοράς. Το όνομα "BERTScore" προέρχεται από το γεγονός ότι χρησιμοποιεί ενσωματώσεις με βάση το πλαίσιο, από το BERT. Το BERTScore λαμβάνει υπόψη όχι μόνο την επικάλυψη μεμονωμένων λέξεων, αλλά και το πλαίσιο στο οποίο εμφανίζονται οι λέξεις αυτές. Αυτό το καθιστά πιο ισχυρό στην αποτύπωση της σημασιολογικής ομοιότητας μεταξύ του παραγόμενου κειμένου και του κειμένου αναφοράς.

Ο τύπος που χρησιμοποιείται από το BERTSCORE για τον υπολογισμό της ομοιότητας δύο προτάσεων, βασίζεται στο άθροισμα των ομοιοτήτων συνημίτονου μεταξύ των ενσωματώσεων των token τους. Το BERTSCORE υπολογίζει τις βαθμολογίες precision και recall με βάση τις ομοιότητες συνημίτονου μεταξύ των ενσωματώσεων token των υποψήφιων προτάσεων και των προτάσεων αναφοράς. Οι βαθμολογίες precision και recall συνδυάζονται στη συνέχεια για τον υπολογισμό του F1 score, το οποίο είναι ένας αρμονικός μέσος όρος precision και recall. Οι τύποι για τον υπολογισμό των precision, recall και F1 score στο πλαίσιο της μετρικής BERTScore έχουν ως εξής:

$$RBERT=\frac{1}{|x|}\sum_{x_i\in x'}\max_{x_j'\in x'} x_i^T x_j'$$

$$PBERT=\frac{1}{|x'|}\sum_{x_i'\in x'}\max_{x_j\in x} x_i'^T x_j$$

$$FBERT=2\cdot\frac{PBERT\cdot RBERT}{PBERT+RBERT}$$

Σε αυτούς τους τύπους, το x αντιπροσωπεύει την πρόταση αναφοράς, το x' αντιπροσωπεύει την υποψήφια πρόταση, το $|x|$ είναι το μήκος της πρότασης αναφοράς και το $|x'|$ είναι το μήκος της υποψήφιας πρότασης. Η ομοιότητα συνημίτονου μεταξύ των διακριτικών συμβολίζεται με : $x_i^T x_j'$.

2.2 Εξαγωγή πληροφοριών από δεδομένα κειμένου

Σε ένα σύστημα σημασιολογικών επικοινωνιών ο στόχος είναι να αποστέλλεται το νόημα ενός μηνύματος αντί ολόκληρου του μηνύματος. Αυτό επιτυγχάνεται από το σημασιολογικό επίπεδο με τη διαδικασία της εξαγωγής σημασιολογικών πληροφοριών. Απλούστερα, στον κλάδο του NLP, αυτή η διαδικασία ονομάζεται ως εξαγωγή πληροφοριών (information extraction). Πιο συγκεκριμένα, ως προς τα δεδομένα κειμένου, αυτή η διαδικασία περιλαμβάνει την εξαγωγή σημασιολογικών σχέσεων (relation extraction) μεταξύ ονοματικών οντοτήτων (named entities). Πιο αναλυτικά, δεδομένου ενός ακατέργαστου κειμένου x, στόχος είναι η μετατροπή του κειμένου σε ένα σύνολο από σημασιολογικές τριπλέτες της μορφής <οντότητα 1 , σχέση r , οντότητα 2>. Οι οντότητες 1 και 2 μπορεί να είναι λέξεις, φράσεις ή άλλες συντακτικές μονάδες στο κείμενο, ενώ η σχέση r είναι ένας προκαθορισμένος τύπος $r \in R$ που περιγράφει τη σχέση μεταξύ των δύο οντοτήτων [11]. Συνήθως η οντότητα 1 ονομάζεται και ως οντότητα κεφαλή (head entity) και συνήθως είναι το υποκείμενο μίας πρότασης, ενώ η οντότητα 2 ονομάζεται και ως οντότητα ουρά (tail entity) και συνήθως είναι το αντικείμενο μίας πρότασης.

Τα τελευταία χρόνια, η μηχανική μάθηση και τα βαθιά νευρωνικά δίκτυα, καθώς και τα προ-εκπαιδευμένα γλωσσικά μοντέλα παρέχουν υπερσύγχρονες λύσεις στον κλάδο της εξαγωγής πληροφοριών. Από τη σκοπιά της δημιουργίας συστημάτων εξαγωγής πληροφοριών αυτή συμβαίνει ακολουθώντας κυρίως τις παρακάτω τρεις προσεγγίσεις: επιβλεπόμενη (supervised), ημι-επιβλεπόμενη (semi-supervised) και απομακρυσμένα επιβλεπόμενη (distant supervised).

Η εξαγωγή πληροφοριών ακολουθώντας την επιβλεπόμενη προσέγγιση επιτυγχάνεται θεωρώντας ένα σταθερό σύνολο σημασιολογικών σχέσεων. Αυτή η προσέγγιση απαιτεί την παράλληλη εκπαίδευση μεγάλων σωμάτων κειμένου με τις αντιστοιχισμένες σημασιολογικές τριπλέτες τους. Η εκπαίδευση των μοντέλων γίνεται με βάση τα διαθέσιμα επισημασμένα σύνολα δεδομένων (annotated datasets) που είναι διαθέσιμα δημοσίως. Η δημιουργία αυτών των συνόλων δεδομένων είναι μία αρκετά χρονοβόρα και δύσκολη διαδικασία, για αυτό και συνήθως τα σύνολα δεδομένων που είναι διαθέσιμα είναι μικρού μεγέθους.

Η εξαγωγή πληροφοριών ακολουθώντας την ημι-επιβλεπόμενη προσέγγιση είναι ένας συνδυασμός επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Σε αυτήν την προσέγγιση, μια μικρή ποσότητα επισημασμένων δεδομένων χρησιμοποιείται για την εκπαίδευση ενός μοντέλου και στη συνέχεια, το μοντέλο χρησιμοποιείται για την επισήμανση μεγαλύτερης ποσότητας μη επισημασμένων δεδομένων έτσι ώστε να δημιουργηθεί ένα μεγαλύτερο σύνολο δεδομένων. Τα επισημασμένα και μη δεδομένα συνδυάζονται στη συνέχεια για να εκπαιδεύσουν ένα νέο μοντέλο, το οποίο χρησιμοποιείται για την εξαγωγή πληροφοριών από νέο κείμενο.

Η εξαγωγή πληροφοριών ακολουθώντας την απομακρυσμένα επιβλεπόμενη προσέγγιση είναι μία μέθοδος που στόχο έχει να αποκτηθούν οι αντιστοιχίσεις

κειμένου-τριπλετών χωρίς την ανθρώπινη παρέμβαση και αντιστοίχιση. Στην απομακρυσμένα επιβλεπόμενη προσέγγιση, οι τριπλέτες από μια υπάρχουσα γνωσιακή βάση (knowledge base) αντιστοιχίζονται σε ένα σώμα ελεύθερου κειμένου, όπως άρθρα της Wikipedia ή άρθρα ειδήσεων.

Κάθε προσέγγιση έχει τα δικά της πλεονεκτήματα και μειονεκτήματα και η επιλογή της προσέγγισης εξαρτάται από την εκάστοτε εργασία και τους διαθέσιμους πόρους. Η επιβλεπόμενη μάθηση απαιτεί μεγάλο όγκο επισημασμένων δεδομένων, ενώ η προσέγγιση της απομακρυσμένα επιβλεπόμενης μάθησης μπορεί να παρέχει αυτόματα μεγάλο όγκο δεδομένων εκπαίδευσης, αλλά μπορεί να περιέχει θορυβώδη δείγματα. Η ημι-επιβλεπόμενη μάθηση μπορεί να είναι ένας καλός συμβιβασμός μεταξύ των δύο, αλλά απαιτεί προσεκτική επιλογή των επισημασμένων δεδομένων και της μη επιβλεπόμενης μεθόδου μάθησης που χρησιμοποιείται [12].

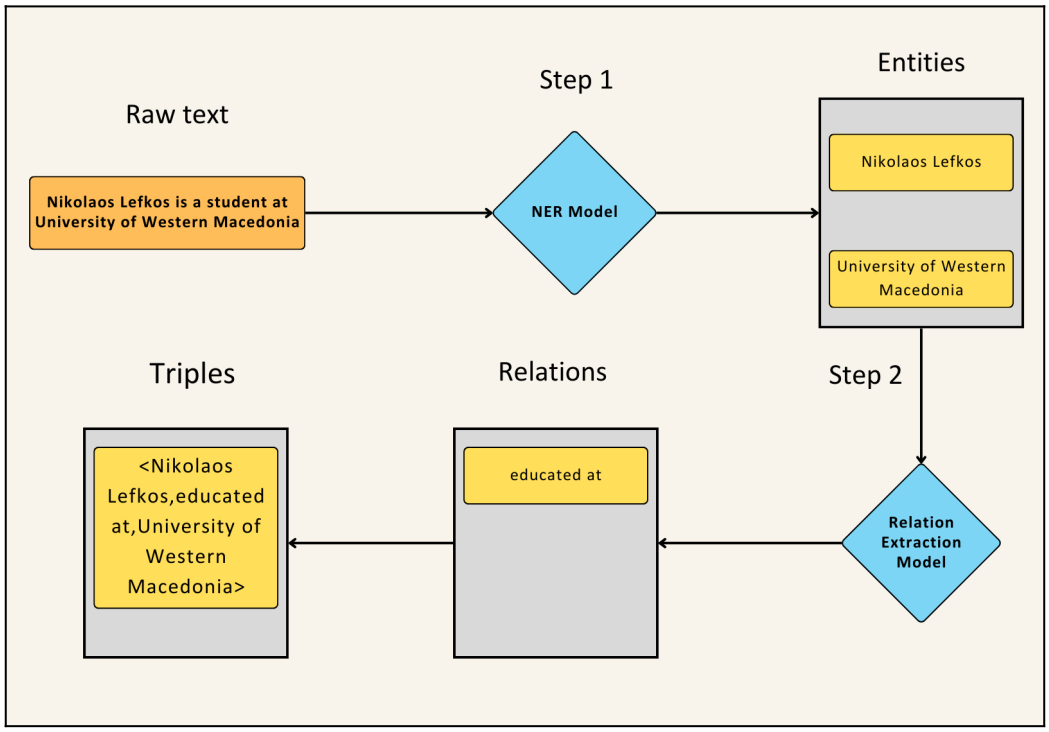
Για να επιτευχθεί η εξαγωγή πληροφοριών ακολουθούνται προσεγγίσεις που μπορούν να χωριστούν σε δύο κατηγορίες:

1) Σωληνωτή προσέγγιση (pipeline-based approach)
2) Συνδυαστική προσέγγιση (joint approach)

Στην παρούσα διπλωματική εργασία ακολουθείται η σωληνωτή προσέγγιση, οι μέθοδοι της οποίας αποτυπώνονται λεπτομερώς παρακάτω. Οι τρόποι με τους οποίους μπορεί

να επιτευχθεί εξαγωγή πληροφοριών ακολουθώντας τη συνδυαστική προσέγγιση θα αναφερθούν συνοπτικά στη συνέχεια.

2.2.1 Σωληνωτή προσέγγιση



Σχήμα 2.3: Παράδειγμα εξαγωγής πληροφοριών με σωληνωτή προσέγγιση

Στη σωληνωτή προσέγγιση οι διαδικασίες της εξαγωγής οντοτήτων ή αλλιώς αναγνώριση ονοματικών οντοτήτων (named entity recognition) και η εξαγωγή σχέσεων γίνονται ξεχωριστά και διαδοχικά. Όπως φαίνεται και στο Σχήμα 2.3, στο πρώτο στάδιο γίνεται η αναγνώριση των οντοτήτων, μέσω ενός μοντέλου εκπαιδευμένο για αυτή τη διαδικασία. Στη συνέχεια, ένα εκπαιδευμένο μοντέλο ταξινόμησης χρησιμοποιείται για τον προσδιορισμό της σχέσης μεταξύ κάθε πιθανού ζεύγους αναγνωρισμένων οντοτήτων. Τέλος, η πληροφορία αποθηκεύεται στη μορφή τριπλέτας. Η αλυσιδωτή προσέγγιση βασίζεται στην υπόθεση ότι οι οντότητες έχουν ήδη προσδιοριστεί και το μοντέλο ταξινόμησης στοχεύει στον προσδιορισμό της σχέσης μεταξύ ζευγών οντοτήτων.

2.2.2 Αναγνώριση Ονοματικών Οντοτήτων (Named Entity Recognition-NER)

Η αναγνώριση ονοματικών οντοτήτων είναι μια διαδικασία που λειτουργεί ως ακρογωνιαίος λίθος για πολλές δραστηριότητες που σχετίζονται με την εξαγωγή πληροφοριών. Μια ονοματική οντότητα είναι μια λέξη ή μια φράση που προσδιορίζει με σαφήνεια ένα στοιχείο από ένα σύνολο άλλων στοιχείων που έχουν παρόμοια χαρακτηριστικά. Παραδείγματα ονοματικών οντοτήτων είναι ονόματα οργανισμών,

προσώπων, τοποθεσιών, ημερομηνιών κ.λπ. Ως αναγνώριση ονοματικών οντοτήτων, ορίζουμε τη διαδικασία εντοπισμού και ταξινόμησης ονοματικών οντοτήτων από ένα κείμενο σε μία κατηγορία από ένα σύνολο προκαθορισμένων κατηγοριών οντοτήτων [13].

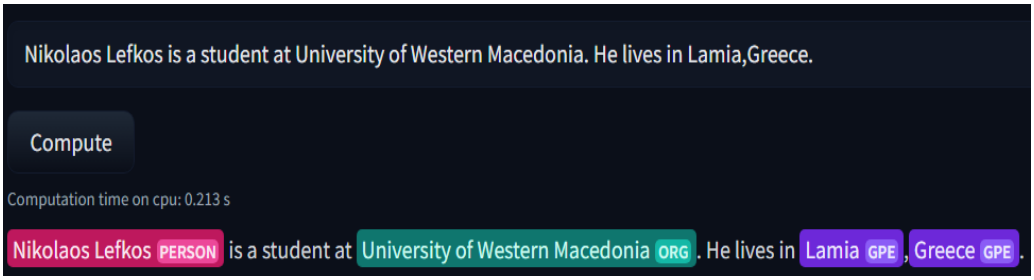
Η αναγνώριση ονοματικών οντοτήτων πρόκειται για μία εργασία ταξινόμησης token (token classification). Η ταξινόμηση token είναι μια εργασία του NLP που περιλαμβάνει την εκχώρηση μιας προκαθορισμένης ετικέτας ή κατηγορίας σε κάθε token σε μια ακολουθία κειμένου. Στο πλαίσιο του NLP, ένα token αναφέρεται σε μια μονάδα κειμένου, η οποία μπορεί να είναι τόσο σύντομη όσο ένας χαρακτήρας ή τόσο μεγάλη όσο μία λέξη. Το αν ένα token θα είναι χαρακτήρας, λέξη ή πρόταση εξαρτάται από την προσέγγιση που ακολουθείται στη διαδικασία του tokenization.

Tokenization ονομάζεται η διαδικασία διάσπασης ενός κειμένου σε μικρότερες μονάδες που ονομάζονται tokens. Η πιο κοινή προσέγγιση tokenization είναι το tokenization σε επίπεδο λέξης (word tokenization). Ο στόχος είναι να χωριστεί μια πρόταση ή ένα κομμάτι κειμένου στις λέξεις που την αποτελούν, διευκολύνοντας την ανάλυση ή την επεξεργασία της γλώσσας. Για παράδειγμα, το παρακάτω κείμενο θα μετατραπεί κατά αυτόν τον τρόπο έπειτα από το word tokenization.

Παράδειγμα 1
Παράδειγμα κειμένου: "Η εκπαίδευση είναι η δύναμη που μπορεί να αλλάξει τον κόσμο." Κείμενο μετά το tokenization: ["Η", "εκπαίδευση", "είναι", "η", "δύναμη", "που", "μπορεί", "να", "αλλάξει", "τον", "κόσμο", "."]

Σε αυτό το παράδειγμα, το κείμενο έχει διασπαστεί σε μεμονωμένες λέξεις. Κάθε λέξη έχει αποτυπωθεί σε μια λίστα, και το σημείο στίξης (στο συγκεκριμένο παράδειγμα το "." στο τέλος) αντιμετωπίζεται ως ξεχωριστό τμήμα. Αυτή η διαδικασία δημιουργεί μία δομή που επιτρέπει την επεξεργασία του κειμένου σε επίπεδο λέξεων.

Στην αναγνώριση ονοματικών οντοτήτων, ο στόχος είναι να εντοπιστούν και να ταξινομηθούν συγκεκριμένα tokens (λέξεις ή φράσεις) από ένα κείμενο, σε προκαθορισμένες κατηγορίες όπως Πρόσωπο, Τοποθεσία, Οργανισμός, Ημερομηνία κ.λπ. Κάθε token του κειμένου ταξινομείται είτε ως ονοματική οντότητα είτε ως μη οντότητα. Εάν πρόκειται για ονοματική οντότητα, ταξινομείται περαιτέρω σε μια συγκεκριμένη κατηγορία.



Σχήμα 2.4: Παράδειγμα αναγνώρισης ονοματικών οντοτήτων

2.2.2.1 Δημιουργία συστημάτων NER

Η υλοποίηση των NER συστημάτων μπορεί να πραγματοποιηθεί ακολουθώντας διάφορες μεθοδολογίες [14]. Γενικά, οι κατευθύνσεις με τις οποίες μπορεί να υλοποιηθεί ένα σύστημα NER είναι οι παρακάτω:

NER βάσει κανόνων (rule-based approach): Αυτή η προσέγγιση περιλαμβάνει τη δημιουργία ενός συνόλου κανόνων με βάση τη γλώσσα, τη σύνταξη και τις γνώσεις γύρω από ένα συγκεκριμένο τομέα για να επιτευχθεί η αναγνώριση ονοματικών οντοτήτων. Ενώ αυτή η προσέγγιση μπορεί να είναι εξαιρετικά ακριβής, περιορίζεται σε έναν συγκεκριμένο τομέα και απαιτεί ανθρώπινη παρέμβαση για το σχεδιασμό των κανόνων.

NER μέσω επιβλεπόμενης μάθησης (supervised learning approach): Αυτή η προσέγγιση περιλαμβάνει την εκπαίδευση ενός μοντέλου μηχανικής μάθησης σε επισημασμένα δεδομένα για την αναγνώριση ονοματικών οντοτήτων. Το μοντέλο μαθαίνει να αναγνωρίζει μοτίβα στα δεδομένα και μπορεί να γενικεύσει σε νέα δεδομένα. Αυτή η προσέγγιση μπορεί να επιτύχει υψηλή ακρίβεια, αλλά απαιτεί μεγάλο όγκο επισημασμένων δεδομένων για την εκπαίδευση των μοντέλων.

NER μέσω μη επιβλεπόμενης μάθησης (unsupervised learning approach): Αυτή η προσέγγιση περιλαμβάνει τεχνικές ομαδοποίησης και συσχέτισης για την αναγνώριση ονοματικών οντοτήτων σε μη επισημασμένα δεδομένα. Αυτή η προσέγγιση είναι χρήσιμη όταν τα επισημασμένα δεδομένα είναι σπάνια ή μη διαθέσιμα, αλλά ενδέχεται να μην επιτυγχάνουν το ίδιο επίπεδο ακρίβειας με την επιβλεπόμενη μάθηση.

NER μέσω βαθιάς μάθησης: Αυτή η προσέγγιση περιλαμβάνει τη χρήση νευρωνικών δικτύων, όπως τα συνελικτικά νευρωνικά δίκτυα (CNNs), τα αναδρομικά νευρωνικά δίκτυα (RNNs) ή οι transformers για την αναγνώριση ονοματικών οντοτήτων. Αυτές οι υλοποιήσεις μπορούν να επιτύχουν υψηλή ακρίβεια και να μάθουν πολύπλοκα μοτίβα στα δεδομένα, αλλά απαιτούν μεγάλο όγκο επισημασμένων δεδομένων και υπολογιστικών πόρων.

Στην παρούσα διπλωματική εργασία δημιουργείται ένα μοντέλο NER με τη μέθοδο της επιβλεπόμενης μάθησης βασισμένη στη βαθιά μάθηση. Τα βήματα που συνήθως ακολουθούνται για τη δημιουργία ενός τέτοιου μοντέλου είναι τα παρακάτω:

1. Συλλογή και προεπεξεργασία δεδομένων: Πρώτα επιλέγεται ένα επισημασμένο σύνολο δεδομένων. Το σύνολο δεδομένων περιέχει σώματα κειμένων όπου κάθε λέξη αντιστοιχίζεται με μία ετικέτα που αντιστοιχεί σε μία κατηγορία ονοματικής οντότητας. Έπειτα γίνεται η προεπεξεργασία των δεδομένων κειμένου. Αυτή περιλαμβάνει το tokenization, το lowercasing και το χειρισμό τυχόν ειδικών χαρακτήρων ή θορύβου.

2. Διαχωρισμός δεδομένων: Το σύνολο δεδομένων διαχωρίζεται σε σύνολα εκπαίδευσης (train data), επικύρωσης (validation data) και δοκιμών (test data). Αυτό βοηθά στην εκπαίδευση του μοντέλου, στη ρύθμιση των υπερπαραμέτρων και στην αξιολόγηση της απόδοσής του σε καινούρια δεδομένα.

3. Tokenization και ενσωμάτωση: Τα κείμενα του συνόλου δεδομένων περνούν από τη διαδικασία του tokenization σε επίπεδο λέξης ή σε επίπεδο μονάδων υπολέξεων (subword units). Τα tokens μετατρέπονται σε αριθμητικά διανύσματα χρησιμοποιώντας ενσωματώσεις λέξεων (word embedding, π.χ. Word2Vec, GloVe) ή ενσωματώσεις ως προς το πλαίσιο (context embedding, π.χ. BERT, GPT).

4. Επιλογή αρχιτεκτονικής μοντέλου: Επιλέγεται η κατάλληλη αρχιτεκτονική για το μοντέλο βαθιάς μάθησης. Οι κοινές επιλογές περιλαμβάνουν αναδρομικά νευρωνικά δίκτυα (RNN), δίκτυα μακροπρόθεσμης βραχυπρόθεσμης μνήμης (LSTM) ή μοντέλα που βασίζονται σε transformers όπως το BERT. Το μοντέλο σχεδιάζεται για το χειρισμό του μεταβλητού μήκους των ακολουθιών εισόδου.

5. Εξαγωγή χαρακτηριστικών (feature extraction): Γίνεται εκμετάλλευση των ενσωματώσεων για την εξαγωγή χαρακτηριστικών που θα περιέχουν πληροφορία με νόημα.

6. Προσθήκη επιπέδου για την αναγνώριση ονοματικών οντοτήτων: Προστίθεται ένα επιπλέον επίπεδο στο μοντέλο που αναγνωρίζει ονοματικές οντότητες για κάθε token στην ακολουθία. Συνήθως, αυτό το επίπεδο υλοποιείται ως επίπεδο softmax με τον αριθμό των κόμβων να αντιστοιχεί στον αριθμό των κατηγοριών οντοτήτων.

7. Επιλογή συνάρτησης απωλειών (loss function) και ρύθμιση υπερπαραμέτρων (hyperparameter tuning): Επιλέγεται η κατάλληλη συνάρτηση απωλειών για την εκπαίδευση του μοντέλου. Συνήθως για τους σκοπούς του NER επιλέγεται η categorical cross-entropy loss function. Πραγματοποιείται ρύθμιση των παραμέτρων του μοντέλου όπως το learning rate και το batch size.

8. Εκπαίδευση μοντέλου: Γίνεται εκπαίδευση του μοντέλου στα δεδομένα του συνόλου εκπαίδευσης χρησιμοποιώντας αλγόριθμους backpropagation και βελτιστοποίησης όπως ο Adam ή ο SGD.

9. Αξιολόγηση μοντέλου: Πραγματοποιείται η αξιολόγηση του μοντέλου στα δεδομένα δοκιμής για να υπολογιστεί η απόδοση του μοντέλου σε καινούρια δεδομένα. Για την αξιολόγηση του NER χρησιμοποιούνται μετρικές όπως η ακρίβεια (precision), η ανάκληση (recall) και η βαθμολογία F1.

10. Χρήση μοντέλου: Το μοντέλο μπορεί να χρησιμοποιηθεί σε καινούρια δεδομένα για την αναγνώριση ονοματικών οντοτήτων.

2.2.2.2 Μετρικές αξιολόγησης συστημάτων NER

Τα συστήματα NER συνήθως αξιολογούνται συγκρίνοντας τα αποτελέσματά τους με τις ανθρώπινες επισημάνσεις. Το NER περιλαμβάνει τον προσδιορισμό τόσο των ορίων των οντοτήτων όσο και των τύπων των οντοτήτων. Με την "αξιολόγηση ακριβούς αντιστοίχισης" (exact-match evaluation), μια οντότητα θεωρείται σωστά αναγνωρισμένη μόνο εάν και τα όρια και ο τύπος ταιριάζουν με την εδαφική αλήθεια (ground truth). Αυτό περιλαμβάνει τον υπολογισμό των precision, recall και F-score ή F1 score τα οποία υπολογίζονται με βάση τον αριθμό των True Positive (TP), False Positive (FP) και False Negative (FN):

- True Positive: οντότητες που αναγνωρίζονται από το NER μοντέλο και ταιριάζουν με την εδαφική αλήθεια.
- False Positive: οντότητες που αναγνωρίζονται από το NER μοντέλο αλλά δεν ταιριάζουν με την εδαφική αλήθεια.
- False Negative: οντότητες επισημασμένες στην εδαφική αλήθεια που δεν αναγνωρίζονται από το NER μοντέλο.

Το precision μετρά την ικανότητα ενός συστήματος NER να αναγνωρίζει μόνο σωστές οντότητες.

$$precision = \frac{TP}{TP + FP}$$

Το recall μετρά την ικανότητα ενός συστήματος NER να αναγνωρίζει όλες τις οντότητες σε ένα σώμα κειμένου.

$$recall = \frac{TP}{TP + FN}$$

Το F-score είναι ο αρμονικός μέσος όρος μεταξύ precision και recall και είναι η μετρική που χρησιμοποιείται συχνότερα.

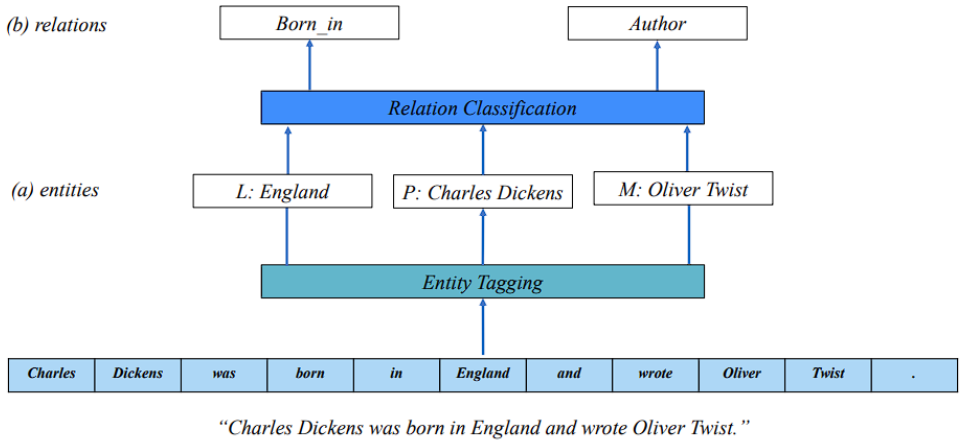
$$F - score = \frac{precision \cdot recall}{precision + recall}$$

2.2.3 Εξαγωγή Σχέσεων

Στην περίπτωση της σωληνωτής προσέγγισης η διαδικασία της εξαγωγής σημασιολογικών σχέσεων προσεγγίζεται ως μία εργασία ταξινόμησης σχέσεων (relation classification-RC) μεταξύ οντοτήτων. Η RC είναι μια εργασία NLP που

περιλαμβάνει τον προσδιορισμό σημασιολογικής σχέσης μεταξύ δύο δεδομένων οντοτήτων, εάν υπάρχει, και την ταξινόμηση της σε μία εκ των προκαθορισμένων σχέσεων [15].

Στο πλαίσιο της εξαγωγής πληροφοριών από ακατέργαστο κείμενο, η ταξινόμηση σχέσεων διαδραματίζει κρίσιμο ρόλο στον εντοπισμό και την κατηγοριοποίηση των σχέσεων μεταξύ οντοτήτων. Αυτή η εργασία είναι απαραίτητη για διάφορες εφαρμογές NLP, όπως η ανάκτηση πληροφοριών (information retrieval), κατασκευή γραφημάτων γνώσης (knowledge graph construction), απάντηση ερωτήσεων (question answering) κ.α.



Σχήμα 2.5: Παράδειγμα εξαγωγής σχέσεων μεταξύ οντοτήτων [15].

Στο Σχήμα 2.5 απεικονίζεται η εξαγωγή σχέσεων μεταξύ επισημασμένων οντοτήτων για μία πρόταση κειμένου. Στο συγκεκριμένο παράδειγμα πρώτα πραγματοποιείται η επισήμανση των οντοτήτων της πρότασης (entity tagging) και έπειτα εκτελείται η ταξινόμηση σχέσεων μεταξύ των επισημασμένων οντοτήτων. Σε αυτό το παράδειγμα, η σχέση «Born_in» συνδέει το άτομο «Charles Dickens» με την τοποθεσία «England» και η σχέση «Author» συνδέει το ίδιο πρόσωπο με το έργο «Oliver Twist».

2.2.3.1 Δημιουργία συστημάτων RC

Η δημιουργία των RC συστημάτων μπορεί να συμβεί με αρκετές διαφορετικές μεθόδους, αλλά αυτές που έχουν επικρατήσει και είναι οι πιο αξιόπιστες είναι οι μέθοδοι της επιβλεπόμενης και απομακρυσμένα επιβλεπόμενης μάθησης που βασίζονται σε βαθιά νευρωνικά δίκτυα. Σε αντίθεση με τις παραδοσιακές μεθόδους, οι μέθοδοι επιβλεπόμενης και απομακρυσμένα επιβλεπόμενης μάθησης χρησιμοποιούν αλγόριθμους μηχανικής μάθησης για την αυτόματη εκμάθηση χαρακτηριστικών από μεγάλα επισημασμένα σύνολα δεδομένων ή αυτόματα επισημασμένα σύνολα

δεδομένων. Αυτό τους επιτρέπει να συλλάβουν πιο σύνθετα μοτίβα και σχέσεις μεταξύ οντοτήτων και να επιτύχουν μεγαλύτερη ακρίβεια στην εξαγωγή σχέσεων [16].

Στην παρούσα διπλωματική εργασία δημιουργείται ένα μοντέλο RC με τη μέθοδο της επιβλεπόμενης μάθησης βασισμένη σε βαθιά νευρωνικά δίκτυα. Αυτή η προσέγγιση περιλαμβάνει την εκπαίδευση μοντέλων σε επισημασμένα σύνολα δεδομένων, όπου τα δεδομένα εισόδου αποτελούνται από προτάσεις ή αποσπάσματα κειμένου που περιέχουν ζεύγη οντοτήτων και η έξοδος είναι η ταξινόμηση της σχέσης μεταξύ των οντοτήτων. Πιο αναλυτικά τα βήματα που συνήθως ακολουθούνται για τη δημιουργία ενός τέτοιου μοντέλου είναι τα παρακάτω [15]:

1. Επιλογή επισημασμένων δεδομένων εκπαίδευσης: Κάθε παράδειγμα εκπαίδευσης περιέχει τις οντότητες ενδιαφέροντος και την αντίστοιχη επισημασμένη σχέση τους. Αυτή η σχέση ανήκει σε ένα σύνολο προκαθορισμένων ετικετών σχέσεων που το μοντέλο στοχεύει να ταξινομήσει.

2. Εξαγωγή χαρακτηριστικών: Χρησιμοποιούνται τεχνικές εξαγωγής χαρακτηριστικών για την αναπαράσταση του κειμένου εισόδου. Αυτές οι τεχνικές μπορεί να περιλαμβάνουν ενσωματώσεις λέξεων, ετικέτες μέρους του λόγου (Part Of Speech Tags), συντακτικά χαρακτηριστικά και άλλες γλωσσικές αναπαραστάσεις που καταγράφουν τις σχετικές πληροφορίες για την ταξινόμηση σχέσεων.

3. Εκπαίδευση ταξινομητή: Μόλις εξαχθούν τα χαρακτηριστικά, ένας ταξινομητής, όπως μια μηχανή διανυσμάτων υποστήριξης (SVM), η λογιστική παλινδρόμηση ή ένα νευρωνικό δίκτυο, εκπαιδεύεται στα επισημασμένα δεδομένα για να μάθει τη χαρτογράφηση μεταξύ των χαρακτηριστικών εισόδου και των ετικετών σχέσης. Το μοντέλο στοχεύει στη γενίκευση από τα δεδομένα εκπαίδευσης για την ακριβή ταξινόμηση των σχέσεων σε καινούριες περιπτώσεις.

4. Αξιολόγηση και δοκιμή: Μετά την εκπαίδευση, το μοντέλο αξιολογείται σε ξεχωριστό σύνολο δοκιμών για να εκτιμηθεί η απόδοσή του στην ταξινόμηση σχέσεων. Οι συνήθεις μετρήσεις αξιολόγησης περιλαμβάνουν το precision, την ανάκληση (recall), τη βαθμολογία F1 και την ακρίβεια (accuracy), οι οποίες παρέχουν πληροφορίες σχετικά με την ικανότητα του μοντέλου να προσδιορίζει σωστά τους τύπους σχέσεων.

2.2.3.2 Μετρικές αξιολόγησης συστημάτων RC

Οι μετρικές αξιολόγησης που χρησιμοποιούνται για την αξιολόγηση μοντέλων εξαγωγής σχέσεων όταν αυτή αντιμετωπίζεται ως εργασία ταξινόμησης σχέσεων περιλαμβάνει τη μέτρηση των precision, recall και F-score.

2.2.4 Συνδυαστική προσέγγιση

Στη συνδυαστική προσέγγιση οι διεργασίες της αναγνώρισης ονοματικών οντοτήτων και η εξαγωγή σχέσεων πραγματοποιούνται ταυτόχρονα. Σε αυτή την προσέγγιση, το μοντέλο στοχεύει να βρει τόσο οντότητες όσο και σχέσεις σε μια πρόταση εξάγοντας έγκυρες σημασιολογικές τριπλέτες. Αυτή η προσέγγιση βασίζεται στο γεγονός ότι οι οντότητες και οι σχέσεις συνδέονται στενά σε πραγματικές εφαρμογές και στοχεύει στην καταγραφή των εξαρτήσεων μεταξύ οντοτήτων και σχέσεων. Η δημιουργία αυτών των μοντέλων πραγματοποιείται ακολουθώντας διαφορετικές προσεγγίσεις, όπως προσεγγίσεις βάσει εύρους (span-based), προσεγγίσεις ακολουθίας προς ακολουθία (Seq2Seq), προσεγγίσεις question answering (QA) και προσεγγίσεις ταξινόμησης ακολουθίας.

2.2.5 Πλεονεκτήματα και μειονεκτήματα της κάθε προσέγγισης

Πλεονεκτήματα σωληνωτής προσέγγισης:

Αρθρωτότητα(modularity): Η σωληνωτή προσέγγιση επιτρέπει την ευελιξία στο σχεδιασμό και την ενημέρωση των δύο ξεχωριστών διεργασιών. Κάθε βήμα, όπως η αναγνώριση οντοτήτων και η ταξινόμηση σχέσεων, μπορεί να βελτιστοποιηθεί ανεξάρτητα.

Ερμηνευσιμότητα: Τα αποτελέσματα κάθε διεργασίας είναι ερμηνεύσιμα και μπορούν να αναλυθούν ξεχωριστά. Αυτό μπορεί να είναι χρήσιμο για τον εντοπισμό σφαλμάτων και την κατανόηση της συμπεριφοράς κάθε διεργασίας.

Μειονεκτήματα σωληνωτής προσέγγισης:

Διάδοση σφαλμάτων: Τα σφάλματα από ένα αρχικό στάδιο μπορούν να μεταδοθούν σε επόμενα στάδια, οδηγώντας σε αρνητικό αντίκτυπο στη συνολική απόδοση. Για παράδειγμα, εάν οι οντότητες που έχουν αναγνωριστεί στο πρώτο στάδιο είναι λάθος, αυτό θα επηρεάσει την ακρίβεια της εξαγωγής σχέσεων.

Έλλειψη γενικής εικόνας: Οι εγγενείς σχέσεις μεταξύ της αναγνώρισης οντοτήτων και της εξαγωγής σχέσεων δεν μπορούν να αποτυπωθούν ικανοποιητικά επειδή κάθε βήμα λειτουργεί ανεξάρτητα.

Πλεονεκτήματα συνδυαστικής προσέγγισης:

Απόκτηση γενικής εικόνας: Η αναγνώριση οντοτήτων και η ταξινόμηση σχέσεων εκτελούνται ταυτόχρονα και με αυτό τον τρόπο συλλαμβάνονται ικανοποιητικά οι εξαρτήσεις μεταξύ οντοτήτων και σχέσεων.

Χειρισμός επικαλυπτόμενων οντοτήτων: Σύνθετες προτάσεις που μπορεί να περιέχουν επικαλυπτόμενες οντότητες χειρίζονται αποτελεσματικά.

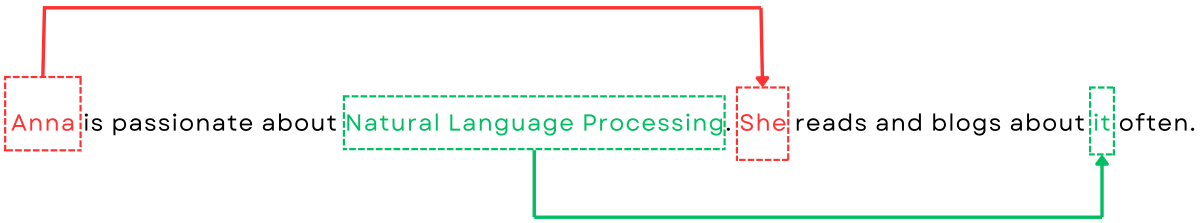
Μειονεκτήματα συνδυαστικής προσέγγισης:

Πολυπλοκότητα: Τα μοντέλα συνδυαστικής προσέγγισης είναι πιο δύσκολο να υλοποιηθούν και απαιτούν προσεκτικό σχεδιασμό και συντονισμό. Επίσης, απαιτούνται περισσότερα δεδομένα εκπαίδευσης και υπολογιστικοί πόροι.

Δυσκολία στην ερμηνεία σφαλμάτων: Όταν εμφανίζονται σφάλματα είναι δύσκολο να προσδιοριστεί εάν προέρχονται από την αναγνώριση οντοτήτων, την ταξινόμηση σχέσεων ή και τα δύο. Αυτό μπορεί να κάνει την ανάλυση σφαλμάτων και τη βελτίωση του μοντέλου πιο περίπλοκη [11].

2.3 Επίλυση συναναφορών (Coreference Resolution-CR)

Ένα από τα ελαττώματα της σωληνωτής προσέγγισης είναι η διάδοση σφαλμάτων από την οποία μπορεί να επηρεαστεί το σύστημα ταξινόμησης σχέσεων. Μία τεχνική που μπορεί να βοηθήσει σε αυτό το πρόβλημα είναι η επίλυση συναναφορών. Η CR, είναι μια εργασία NLP που περιλαμβάνει τον προσδιορισμό όλων των αναφορών σε ένα κείμενο που αναφέρονται στην ίδια οντότητα. Αυτές οι αναφορές μπορεί να έχουν τη μορφή ονοματικών φράσεων, ονοματικών οντοτήτων ή αντωνυμιών. Ο στόχος της επίλυσης συναναφορών είναι να ομαδοποιήσει αυτές τις αναφορές μαζί και να τις αναθέσει σε μια ενιαία οντότητα, η οποία μπορεί να είναι ένα πρόσωπο, μέρος, πράγμα ή έννοια. Η CR είναι σημαντική για πολλές εφαρμογές NLP, επειδή βοηθά στη βελτίωση της ακρίβειας και της αποτελεσματικότητας αυτών των εφαρμογών. Η CR μπορεί να προστεθεί ως ένα επιπλέον στάδιο της σωληνωτής προσέγγισης και να βελτιώσει τη συνολική απόδοση συστημάτων όπως η μηχανική μετάφραση, η ανάλυση συναισθήματος και η εξαγωγή σχέσεων [17], [18].



Σχήμα 2.6: Παράδειγμα επίλυσης συναναφορών

Η CR μπορεί να βοηθήσει στην εξαγωγή σχέσεων προσδιορίζοντας τις οντότητες που εμπλέκονται σε μια σχέση και συνδέοντάς τις μεταξύ τους. Σε πολλές περιπτώσεις, οι σχέσεις μεταξύ οντοτήτων εκφράζονται χρησιμοποιώντας αντωνυμίες ή άλλες εκφράσεις αναφοράς, γεγονός που μπορεί να δυσκολέψει τον προσδιορισμό των ακριβών οντοτήτων που εμπλέκονται. Για παράδειγμα, εξετάστε την ακόλουθη πρόταση: “John met Mary at the park. He gave her a gift”. Σε αυτή την πρόταση, δεν είναι σαφές σε ποιον αναφέρεται η αντωνυμία "he" και η αντωνυμία "her". Ωστόσο,

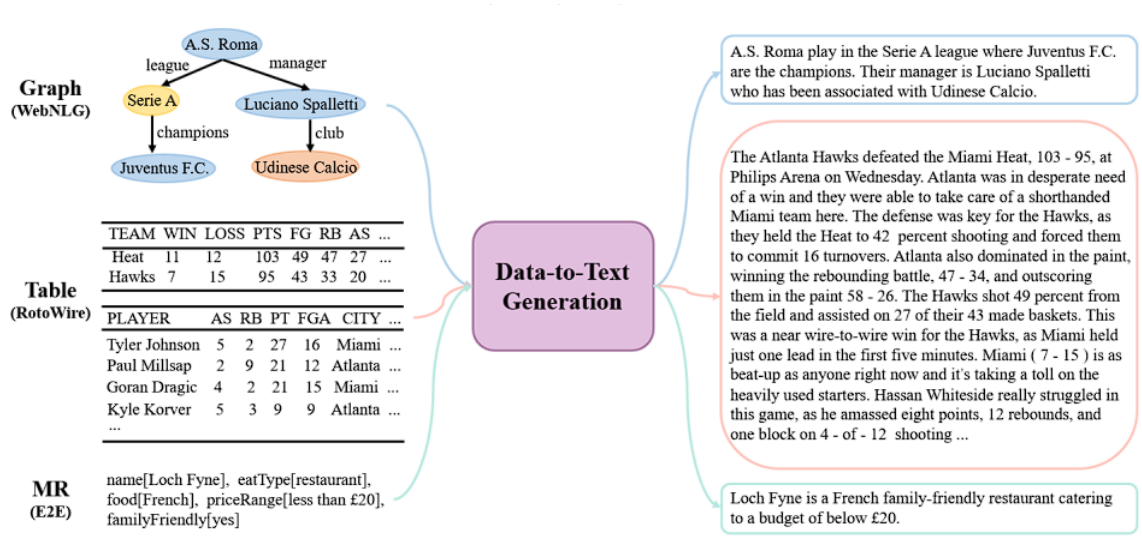
εκτελώντας CR μπορούμε να αναγνωρίσουμε ότι το «he» αναφέρεται στον John και το «her» αναφέρεται στη Mary. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για να εξαχθεί η σχέση που συνδέει τους John και Mary. Με τον ακριβή προσδιορισμό των οντοτήτων που εμπλέκονται σε μια σχέση, η επίλυση συναναφορών μπορεί να βοηθήσει στη βελτίωση της ακρίβειας της εξαγωγής σχέσεων. Αυτό είναι ιδιαίτερα σημαντικό σε εφαρμογές όπως η εξαγωγή πληροφοριών, όπου ο στόχος είναι η αυτόματη εξαγωγή δομημένης πληροφορίας από μη δομημένο κείμενο.

2.4 Παραγωγή κειμένου από δεδομένα

Μετά από την εξαγωγή των σημασιολογικών πληροφοριών από ένα κείμενο και τη κωδικοποίηση τους στη μορφή των σημασιολογικών τριπλετών μέσω του σημασιολογικού κωδικοποιητή, ο πομπός στέλνει αυτή την πληροφορία στο δέκτη μέσω ενός φυσικού καναλιού. Ο δέκτης χρησιμοποιώντας τον σημασιολογικό αποκωδικοποιητή είναι υπεύθυνος για την ανακατασκευή των σημασιολογικών τριπλετών σε μορφή κειμένου. Ο σημασιολογικός αποκωδικοποιητής πρέπει να αποδώσει ερμηνεία στις σημασιολογικές τριπλέτες που λαμβάνει και να δημιουργήσει ένα συνεκτικό κείμενο συναφές με το αρχικό κείμενο του πομπού. Αυτή η διαδικασία επιτυγχάνεται μέσω της παραγωγής κειμένου από δεδομένα (data-to-text generation) και ειδικά όταν τα δεδομένα είναι σε μορφή σημασιολογικών τριπλετών η διαδικασία μπορεί να οριστεί ως: παραγωγή κειμένου από τριπλέτες rdf (rdf-to-text generation).

Η παραγωγή κειμένου από δεδομένα (data-to-text generation-D2T) είναι η εργασία παραγωγής φυσικής γλώσσας (natural language generation-NLG) από δομημένα δεδομένα όπως γραφήματα, πίνακες και αναπαραστάσεις νοήματος (meaning representations-MR). Ο στόχος του D2T είναι να παράγει κείμενο που περιγράφει με ακρίβεια και ευχέρεια τα μη γλωσσικά δομημένα δεδομένα. Τα δομημένα δεδομένα αποθηκεύονται συνήθως σε διαφορετικές μορφές, συμπεριλαμβανομένων γραφημάτων, πινάκων και MR. Η εργασία του D2T περιλαμβάνει την ανάλυση και το φιλτράρισμα των δομημένων δεδομένων, την επιλογή όλων ή μέρους των δεδομένων για απόδοση νοήματος και την ακριβή και άπταιστη περιγραφή των επιλεγμένων δεδομένων μέσω φυσικής γλώσσας [19].

Η παραγωγή κειμένου από τριπλέτες RDF (RDF-to-text generation) είναι μία υποεργασία του D2T και πρόκειται για το έργο της δημιουργίας κειμένου φυσικής γλώσσας από μια συλλογή τριπλετών RDF (Resource Description Framework). Οι τριπλέτες RDF αποτελούνται από μια οντότητα υποκειμένου, μια σχέση και μια οντότητα αντικειμένου και μπορούν να αναπαρασταθούν ως κατευθυνόμενος γράφος. Ο στόχος του RDF-to-text generation είναι να δημιουργήσει κείμενο που μεταφέρει με ακρίβεια και πληρότητα τις πληροφορίες που περιέχονται στις τριπλέτες RDF [20].



Σχήμα 2.7: απεικόνιση του D2T από τρεις τύπους δομημένων δεδομένων. Επάνω: Γράφος. Μέση: Πίνακας. Κάτω: MR [19].

2.4.1 Δημιουργία συστημάτων D2T

Οι τεχνικές που χρησιμοποιούνται για την υλοποίηση μοντέλων D2T βασίζονται κυρίως στη μηχανική μάθηση. Μερικές από τις επικρατέστερες τεχνικές είναι οι παρακάτω [20]:

1. Μοντέλα ακολουθίας προς ακολουθία (Seq2seq): Αυτά τα μοντέλα, συχνά εξοπλισμένα με μηχανισμούς προσοχής και μηχανισμούς αντιγραφής, έχουν δείξει υποσχόμενες επιδόσεις σε εργασίες παραγωγής φυσικής γλώσσας, συμπεριλαμβανομένης της παραγωγής κειμένου από RDF triples. Λειτουργούν ισοπεδώνοντας (flattening) μια συλλογή τριπλετών RDF σε μια ακολουθία κειμένου, επιτρέποντας τη δημιουργία συνεκτικών και ενημερωτικών περιγραφών κειμένου.

2. Νευρωνικά Δίκτυα Γράφων (GNN): Τα GNNs έχουν προσελκύσει ενδιαφέρον στον τομέα της παραγωγή κειμένου από δεδομένα, ειδικά όταν τα δεδομένα εισόδου μπορούν να αναπαρασταθούν σε μορφή γράφου. Αυτά τα μοντέλα μπορούν να μάθουν ενσωματώσεις κόμβων και ενσωματώσεις γειτονικών κόμβων χρησιμοποιώντας δομή γράφου, καθιστώντας τα κατάλληλα για εργασίες όπου οι δομημένες πληροφορίες είναι σημαντικές, όπως η δημιουργία κειμένου από τριπλέτες RDF.

3. Προ-εκπαιδευμένα γλωσσικά μοντέλα (PLMs) μεγάλης κλίμακας: Οι πρόσφατες εξελίξεις στην επεξεργασία φυσικής γλώσσας έχουν φέρει στο προσκήνιο την εφαρμογή προ-εκπαιδευμένων γλωσσικών μοντέλων μεγάλης κλίμακας, όπως το T5, για εργασίες παραγωγής κειμένου από δεδομένα. Αυτά τα μοντέλα επιδεικνύουν βελτιωμένη απόδοση αξιοποιώντας εξωτερικές πηγές γνώσης, όπως η Wikipedia, για τη βελτίωση της ποιότητας παραγωγής κειμένου. Επίσης αυτά τα μοντέλα μπορούν να ρυθμιστούν με ακρίβεια (fine-tuning) σε συγκεκριμένο σύνολο δεδομένων.

4. Μεγάλα γλωσσικά μοντέλα (LLMs): Αυτά τα μοντέλα, όπως τα GPT-3, T5, έχουν αποδείξει την ικανότητα δημιουργίας συνεκτικού και σχετικού με τα συμφραζόμενα κειμένου φυσικής γλώσσας με βάση τις προτροπές εισόδου. Όταν πρόκειται για τη δημιουργία κειμένου από δεδομένα, η προτροπή ενός LLM περιλαμβάνει την παροχή δομημένων δεδομένων στο μοντέλο, όπως τριπλέτες RDF, σε συγκεκριμένη μορφή ή κωδικοποίηση και, στη συνέχεια, την προτροπή του μοντέλου να δημιουργήσει μια περιγραφή φυσικής γλώσσας με βάση αυτήν την είσοδο.

2.4.2 Μετρικές αξιολόγησης συστημάτων D2T

Για την αξιολόγηση της απόδοσης των μοντέλων παραγωγής κειμένου από δεδομένα χρησιμοποιείται μία πληθώρα μετρικών. Αυτές οι μετρήσεις κατηγοριοποιούνται σε μετρικές αυτόματης αξιολόγησης και μετρικές ανθρώπινης αξιολόγησης (human evaluation). Οι μετρικές αυτές ορίζονται παρακάτω:

Μετρικές αυτόματης αξιολόγησης: Χωρίζονται σε τρεις επιμέρους κατηγορίες:

1.Λεξικοκολογική ομοιότητα (Lexical Similarity): Μετρικές όπως οι BLEU, ROUGE, METEOR, NIST, CIDER CHRF, TTR, και Dist-n χρησιμοποιούνται για τη μέτρηση της λεξικολογικής ομοιότητας και ποικιλομορφίας στο παραγόμενο κείμενο.

2.Σημασιολογική ισοδυναμία (Semantic Equivalence): Οι πιο πρόσφατες μετρικές βασίζονται στην ομοιότητα των ενσωματώσεων προτάσεων και χρησιμοποιούνται για την αξιολόγηση της σημασιολογικής ισοδυναμίας μεταξύ παραγόμενου κειμένου και κειμένου αναφοράς. Αυτές είναι οι BERTScore, BLEURT, MoverScore, FrugalScore, FINE και ROUGH.

3.Πιστότητα (Faithfulness): Μετρήσεις όπως η επικύρωση γεγονότων(fact checking) και τα μοντέλα που βασίζονται σε παραγωγή συμπεράσματος φυσικής γλώσσας (Natural Language Inference-NLI) χρησιμοποιούνται για την αξιολόγηση της πιστότητας μεταξύ του παραγόμενου κειμένου και των δεδομένων εισόδου.

Μετρικές ανθρώπινης αξιολόγησης: Η ανθρώπινη αξιολόγηση εφαρμόζεται για την αξιολόγηση μετρήσεων όπως η ευφράδεια, η πιστότητα και η συνοχή στο παραγόμενο κείμενο. Αυτές οι μετρήσεις είναι ζωτικής σημασίας για την αξιολόγηση της ποιότητας των περιγραφών φυσικής γλώσσας που παράγονται από μοντέλα D2T.

2.5 Μοντέλα βασισμένα σε Transformer

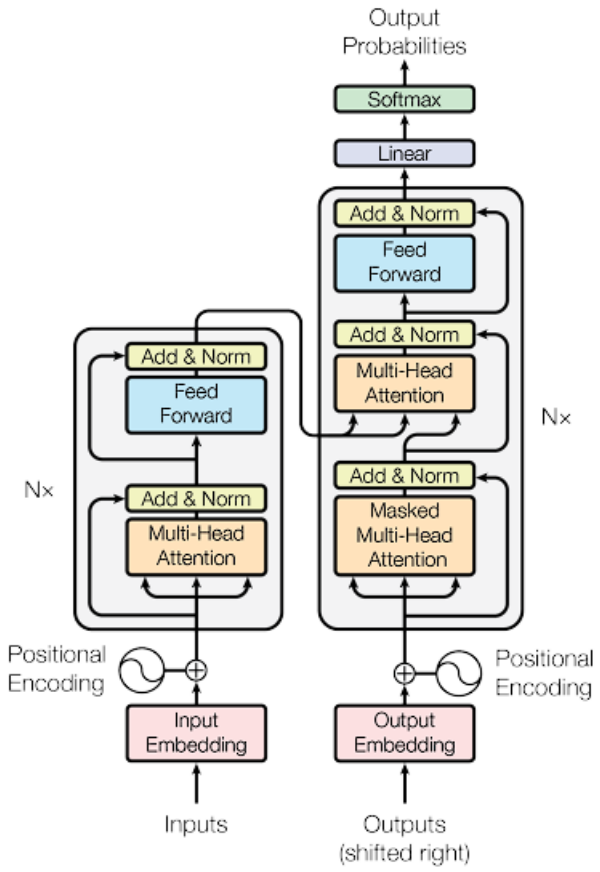
Οι εργασίες NLP που είναι απαραίτητες στις σημασιολογικές επικοινωνίες γίνονται κυρίως μέσω μοντέλων βαθιάς μάθησης. Τα τελευταία έτη, τα μοντέλα βαθιάς μάθησης που χρησιμοποιούνται ευρέως και είναι τα πιο αποτελεσματικά για εργασίες NLP είναι τα μοντέλα που βασίζονται στην αρχιτεκτονική Transformer.

2.5.1 Transformer

Ο transformer είναι μία αρχιτεκτονική νευρωνικού δικτύου σχεδιασμένη για το χειρισμό ακολουθιών δεδομένων, όπως προτάσεις, παράγραφοι ή χρονοσειρές. Παρουσιάστηκε στη δημοσίευση "Attention Is All You Need" από τους Vaswani et al. [21] για την αντιμετώπιση περιορισμών προηγούμενων μοντέλων ακολουθίας, όπως τα αναδρομικά νευρωνικά δίκτυα (RNNs) και τα συνελικτικά νευρωνικά δίκτυα (CNNs). Ο μηχανισμός αυτοπροσοχής (self-attention) του transformer του επιτρέπει να εξετάζει τις σχέσεις μεταξύ όλων των στοιχείων σε μια ακολουθία ταυτόχρονα, συλλαμβάνοντας πολύπλοκες εξαρτήσεις.

Το μοντέλο transformer σχεδιάστηκε κυρίως για εργασίες Seq2seq, όπως η μηχανική μετάφραση, όπου ο στόχος είναι να μετατραπεί μια ακολουθία λέξεων από μια γλώσσα σε μια άλλη γλώσσα. Ωστόσο, η αρχιτεκτονική του έχει αποδειχθεί ευέλικτη και αποτελεσματική για ένα ευρύ φάσμα άλλων εργασιών, συμπεριλαμβανομένης της δημιουργίας κειμένου, της σύνοψης κειμένου, της απάντησης ερωτήσεων, της ανάλυσης συναισθήματος και άλλων.

Το κυριότερο πλεονέκτημα του transformer έναντι των υπολοίπων αρχιτεκτονικών είναι η δυνατότητα του να μπορεί να συλλάβει εξαρτήσεις μεγάλης εμβέλειας σε ακολουθιακά δεδομένα πιο αποτελεσματικά. Αυτό επιτυγχάνεται μέσω του μηχανισμού self-attention ο οποίος επιτρέπει σε κάθε θέση στην ακολουθία εισόδου να παρακολουθεί όλες τις θέσεις, επιτρέποντας στο μοντέλο να αντιλαμβάνεται ευκολότερα το γενικό πλαίσιο. Αυτός είναι ο βασικός λόγος για τον οποίο οι transformers έχουν διακριθεί στην επεξεργασία φυσικής γλώσσας, όπου η κατανόηση της σημασίας μιας λέξης συχνά απαιτεί την εξέταση του πλαισίου ολόκληρης της πρότασης.



Σχήμα 2.8: Η αρχιτεκτονική transformer [21].

Ο τρόπος λειτουργίας του transformer χωρίζεται στα ακόλουθα βήματα:

1. Ενσωμάτωση εισόδου (input embedding): Η ακολουθία εισόδου (π.χ. μια πρόταση) μετατρέπεται σε διανύσματα χρησιμοποιώντας ενσωματώσεις. Αυτά τα διανύσματα αντιπροσωπεύουν τις λέξεις ή τα tokens με ουσιαστικό τρόπο.

2. Μηχανισμός αυτοπροσοχής (self-attention): Ο πυρήνας του transformer. Για κάθε λέξη στην είσοδο, το μοντέλο υπολογίζει τη σημασία της σε σχέση με όλες τις άλλες λέξεις. Αυτό γίνεται μέσω των εσωτερικών γινομένων των query, key και value τα οποία δημιουργούν βαθμολογίες προσοχής (attention scores). Οι βαθμολογίες προσοχής καθορίζουν πόση προσοχή πρέπει να δοθεί στις πληροφορίες κάθε λέξης κατά τη δημιουργία της αναπαράστασης.

3. Προσοχή πολλαπλών κεφαλών (multi-head attention): Αντί να βασίζεται σε έναν ενιαίο μηχανισμό αυτοπροσοχής, ο transformer έχει πολλαπλούς μηχανισμούς προσοχής (heads) οι οποίοι λειτουργούν παράλληλα. Κάθε κεφαλή επικεντρώνεται σε διαφορετικές πτυχές των δεδομένων, επιτρέποντας στο μοντέλο να συλλάβει διαφορετικούς τύπους σχέσεων.

4. Κωδικοποίηση θέσης (positional encoding): Δεδομένου ότι ο transformer δεν κατανοεί εγγενώς τη σειρά των λέξεων σε μια ακολουθία, προστίθενται

κωδικοποιήσεις θέσης στις ενσωματώσεις. Αυτό βοηθά το μοντέλο να διαφοροποιήσει τις θέσεις διαφορετικών λέξεων.

5. Κωδικοποιητής και αποκωδικοποιητής: Η αρχιτεκτονική του transformer αποτελείται από δύο μέρη: τον κωδικοποιητή και τον αποκωδικοποιητή. Ο κωδικοποιητής επεξεργάζεται την ακολουθία εισόδου και καταγράφει το νόημά της, ενώ ο αποκωδικοποιητής δημιουργεί την ακολουθία εξόδου με βάση τις πληροφορίες του κωδικοποιητή και τα tokens που δημιουργήθηκαν προηγουμένως.

6. Αυτοπροσοχή αποκωδικοποιητή (decoder self-attention) και προσοχή κωδικοποιητή-αποκωδικοποιητή (encoder-decoder attention): Στον αποκωδικοποιητή, η αυτο-προσοχή χρησιμοποιείται για να επικεντρωθεί στα δημιουργημένα tokens κατά τη δημιουργία της ακολουθίας εξόδου. Επιπλέον, η προσοχή κωδικοποιητή-αποκωδικοποιητή βοηθά τον αποκωδικοποιητή να ευθυγραμμιστεί με σχετικά μέρη της ακολουθίας εισόδου.

7. Νευρωνικό δίκτυο πρόσω τροφοδότησης με βάση τη θέση (Position-wise Feed forward Network): Μετά τον μηχανισμό προσοχής, κάθε επίπεδο περιλαμβάνει ένα position-wise feedforward νευρωνικό δίκτυο. Αυτό το δίκτυο εφαρμόζεται ανεξάρτητα σε κάθε θέση της ακολουθίας, προσθέτοντας έναν μη γραμμικό μετασχηματισμό στις αναπαραστάσεις.

8. Παραγωγή εξόδου: Ο αποκωδικοποιητής χρησιμοποιεί την έξοδο του τελικού στρώματος για να δημιουργήσει την ακολουθία εξόδου. Αυτή η έξοδος μετατρέπεται σε κατανομή πιθανότητας από το λεξιλόγιο χρησιμοποιώντας μια συνάρτηση softmax. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο προσπαθεί να ελαχιστοποιήσει τη διαφορά μεταξύ της προβλεπόμενης κατανομής και της πραγματικής κατανομής της ακολουθίας στόχου [22].

Τα μοντέλα όπως τα BERT, GPT, T5, BART, XLNet βασίζονται στον transformer και ανήκουν στην οικογένεια μοντέλων transformer. Τα μοντέλα αυτά αξιοποιούν είτε τον κωδικοποιητή είτε τον αποκωδικοποιητή είτε και τα δύο μέρη του transformer, αναλόγως την εργασία NLP που επιτελούν.

2.5.2 BERT

Το bidirectional encoder representations from transformers (BERT), αποτελεί ένα πλαίσιο μηχανικής μάθησης ανοιχτού κώδικα σχεδιασμένο για τον κλάδο της NLP, το οποίο δημιουργήθηκε το 2018 από τη Google [23]. Το BERT αξιοποιεί ένα νευρωνικό δίκτυο βασισμένο σε transformers για να κατανοήσει και να δημιουργήσει ανθρώπινη γλώσσα. Το BERT χρησιμοποιεί μόνο τον κωδικοποιητή του transformer διότι το μοντέλο εστιάζει στην κατανόηση των ακολουθιών εισόδου παρά στη δημιουργία ακολουθιών εξόδου. Τα παραδοσιακά γλωσσικά μοντέλα επεξεργάζονται το κείμενο διαδοχικά, είτε από αριστερά προς τα δεξιά είτε από δεξιά προς τα αριστερά. Αυτή η μέθοδος περιορίζει την επίγνωση του μοντέλου μόνο στο άμεσο πλαίσιο που

προηγείται της λέξης-στόχου. Το BERT χρησιμοποιεί μία αμφίδρομη προσέγγιση λαμβάνοντας υπόψη τόσο το αριστερό όσο και το δεξί πλαίσιο των λέξεων σε μια πρόταση. Έτσι, αντί να αναλύει το κείμενο διαδοχικά, το BERT εξετάζει όλες τις λέξεις σε μια πρόταση ταυτόχρονα.

Το BERT είναι ένα προεκπαιδευμένο γλωσσικό μοντέλο. Η προ-εκπαίδευση σημαίνει ότι το BERT εκπαιδεύεται αρχικά σε μεγάλο σώμα μη επιβλεπόμενων ή ημι-επιβλεπόμενων δεδομένων. Συγκεκριμένα, το BERT είναι εκπαιδευμένο σε δύο διαφορετικές, αλλά σχετικές, εργασίες NLP: το masked language modeling (MLM) και το next sentence prediction. Το next sentence prediction είναι η διαδικασία κατά την οποία το μοντέλο προσπαθεί να προβλέψει το αν δύο δεδομένες προτάσεις έχουν μια λογική, διαδοχική σύνδεση ή αν η σχέση τους είναι απλά τυχαία [24].

2.5.2.1 MLM

Το MLM είναι η διαδικασία κατά την οποία μία λέξη μίας πρότασης κρύβεται (καλύπτεται) και στη συνέχεια το μοντέλο προσπαθεί να προβλέψει ποια λέξη έχει κρυφτεί με βάση τα συμφραζόμενα της κρυφής λέξης. Αυτό επιτυγχάνεται ακολουθώντας τα παρακάτω βήματα:

1. Κάλυψη λέξεων (words masking): Πριν το BERT μάθει από τις προτάσεις, κρύβει μερικές λέξεις (περίπου το 15%) και τις αντικαθιστά με ένα ειδικό σύμβολο [MASK].

2. Πρόβλεψη κρυμμένων λέξεων: Το BERT προσθέτει ένα στρώμα ταξινόμησης πάνω από την έξοδο του κωδικοποιητή. Τα διανύσματα εξόδου από το επίπεδο ταξινόμησης πολλαπλασιάζονται με τον πίνακα ενσωμάτωσης, μετατρέποντάς τα στη διάσταση λεξιλογίου. Αυτό το βήμα βοηθά στην ευθυγράμμιση των προβλεπόμενων αναπαραστάσεων με τον χώρο λεξιλογίου. Η πιθανότητα κάθε λέξης στο λεξιλόγιο υπολογίζεται χρησιμοποιώντας τη συνάρτηση ενεργοποίησης softmax. Αυτό το βήμα δημιουργεί μια κατανομή πιθανότητας σε ολόκληρο το λεξιλόγιο για κάθε θέση μάσκας.

3. Εκμάθηση του BERT στο MLM: Η συνάρτηση απωλειών που χρησιμοποιείται κατά τη διάρκεια της εκπαίδευσης λαμβάνει υπόψη μόνο την πρόβλεψη των καλυμμένων λέξεων. Το μοντέλο τιμωρείται για την απόκλιση μεταξύ των προβλέψεών του και των πραγματικών τιμών των καλυμμένων λέξεων.

4. Δίνεται ιδιαίτερη προσοχή στις καλυμμένες λέξεις: Ο κύριος στόχος του BERT κατά τη διάρκεια της εκπαίδευσης είναι να προβλέψει αυτές τις καλυμμένες λέξεις σωστά. Αυτή η στρατηγική βοηθά το BERT να κατανοήσει το νόημα και το πλαίσιο των λέξεων.

2.5.2.2 Next Sentence Prediction

Το BERT προβλέπει εάν η δεύτερη πρόταση συνδέεται με την πρώτη. Αυτό γίνεται μετατρέποντας την έξοδο του [CLS] token σε ένα διάνυσμα σχήματος 2×1, χρησιμοποιώντας ένα επίπεδο ταξινόμησης και υπολογίζοντας την πιθανότητα εάν η δεύτερη πρόταση ακολουθεί την πρώτη χρησιμοποιώντας τη συνάρτηση softmax.

Για να καταλάβει το μοντέλο τη σύνδεση μεταξύ των προτάσεων γίνονται τα παρακάτω:

- Εισάγεται ένα [CLS] token στην αρχή της πρώτης πρότασης και ένα [SEP] token στο τέλος κάθε πρότασης.
- Σε κάθε token προστίθεται μία ενσωμάτωση πρότασης που υποδεικνύει εάν το token ανήκει στην πρώτη ή στη δεύτερη πρόταση.
- Σε κάθε token της ακολουθίας εισάγεται μία ενσωμάτωση θέσης.

Το BERT εκπαιδεύεται ταυτόχρονα στο task του MLM και στο task του next sentence prediction. Το μοντέλο στοχεύει στην ελαχιστοποίηση της συνδυασμένης συνάρτησης απωλειών των δύο εργασιών, οδηγώντας σε ένα ισχυρό γλωσσικό μοντέλο με βελτιωμένες δυνατότητες κατανόησης του πλαισίου εντός των προτάσεων και των σχέσεων μεταξύ των προτάσεων. Το MLM βοηθά το BERT να κατανοήσει το πλαίσιο μέσα σε μια πρόταση και το next sentence prediction βοηθά το BERT να κατανοήσει τη σύνδεση ή τη σχέση μεταξύ ζευγών προτάσεων.

2.5.2.2 Fine-tuning BERT

Μετά τη φάση της προ-εκπαίδευσης, το μοντέλο BERT, οπλισμένο με τις ενσωματώσεις του ως προς το πλαίσιο (contextual embeddings), μπορεί στη συνέχεια να ρυθμιστεί μέσω του fine-tuning για συγκεκριμένες εργασίες NLP. Αυτό το βήμα προσαρμόζει το μοντέλο σε πιο στοχευμένες εφαρμογές, προσαρμόζοντας στη γενική κατανόηση της γλώσσας στη συγκεκριμένη εργασία.

Το fine-tuning είναι μια διαδικασία στη μηχανική μάθηση στην οποία ένα προ-εκπαιδευμένο μοντέλο, το οποίο έχει ήδη εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων για μια συγκεκριμένη εργασία, εκπαιδεύεται περαιτέρω σε ένα μικρότερο σύνολο δεδομένων για μια σχετική ή πιο συγκεκριμένη εργασία(π.χ. αναγνώριση ονοματικών οντοτήτων, ταξινόμηση κειμένου). Ο στόχος του fine-tuning είναι να προσαρμόσει τα χαρακτηριστικά και τις αναπαραστάσεις του προ-εκπαιδευμένου μοντέλου στις ανάγκες της εργασίας-στόχου, επιτρέποντας στο μοντέλο να αποδίδει καλά σε έναν συγκεκριμένο τομέα ή σε ένα σύνολο εργασιών.

2.5.2.3 Ταξινόμηση ακολουθίας μέσω BERT

Το BERT μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης ακολουθίας προσθέτοντας ένα επίπεδο ταξινόμησης στην κορυφή της εξόδου του transformer για το token [CLS]. Το [CLS] token αντιπροσωπεύει τις συγκεντρωτικές πληροφορίες από ολόκληρη την ακολουθία εισόδου. Αυτή η συγκεντρωτική αναπαράσταση μπορεί στη συνέχεια να χρησιμοποιηθεί ως είσοδος για το επίπεδο ταξινόμησης για να γίνει η πρόβλεψη για τη συγκεκριμένη εργασία.

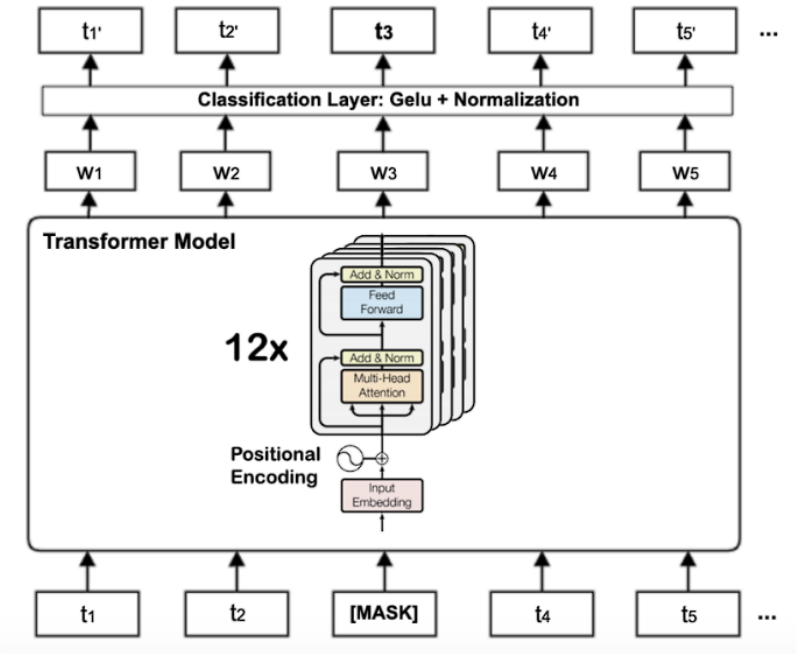
2.5.2.4 Αναγνώριση ονοματικών οντοτήτων μέσω BERT

Το BERT μπορεί να χρησιμοποιηθεί για την αναγνώριση ονοματικών οντοτήτων. Ένα μοντέλο NER που βασίζεται σε BERT εκπαιδεύεται λαμβάνοντας το διάνυσμα εξόδου κάθε token από τον transformer και τροφοδοτώντας το σε ένα επίπεδο ταξινόμησης. Το επίπεδο προβλέπει την ετικέτα ονοματικής οντότητας για κάθε token, υποδεικνύοντας τον τύπο της οντότητας που αντιπροσωπεύει.

2.5.2.5 Αρχιτεκτονική και μεγέθη μοντέλων BERT

Η αρχιτεκτονική του BERT αποτελείται από πολλαπλά στρώματα αμφίδρομων κωδικοποιητών transformer:

- Το BERTBASE έχει 12 στρώματα στη στοίβα κωδικοποιητών ενώ το BERTLARGE έχει 24 επίπεδα στη στοίβα κωδικοποιητών. Αυτά είναι περισσότερα από την αρχιτεκτονική transformer που περιγράφεται στην αρχική εργασία (6 επίπεδα κωδικοποιητή).
- Οι αρχιτεκτονικές BERT (BASE και LARGE) έχουν επίσης μεγαλύτερα δίκτυα πρόσω τροφοδότησης (768 και 1024 κρυφές μονάδες αντίστοιχα) και περισσότερες κεφαλές προσοχής (12 και 16 αντίστοιχα) από την αρχιτεκτονική transformer.
- Το BERTBASE περιέχει 110 εκατομμύρια παραμέτρους ενώ το BERTLARGE έχει 340 εκατομμύρια παραμέτρους.
- Το BERTBASE έχει εκπαιδευτεί σε 16GB δεδομένων τα οποία αποτελούνται από το σώμα κειμένου Toronto boks corpus και την αγγλική μορφή της Wikipedia.
- ΤοBERTBASE έχει εκπαιδευτεί σε χρόνο 12 GPU ημερών χρησιμοποιώντας 8 V100 κάρτες γραφικών.



Σχήμα 2.9: Η αρχιτεκτονική του BERTBASE [25].

Στο Σχήμα 2.9 απεικονίζεται η αρχιτεκτονική του μοντέλου BERT. Ο συμβολισμός "12x" υποδεικνύει ότι το BERT αποτελείται από δώδεκα πανομοιότυπα στρώματα κωδικοποιητή στοιβαγμένα το ένα πάνω στο άλλο. Κάθε ένα από αυτά τα επίπεδα περιλαμβάνει έναν μηχανισμό αυτο-προσοχής, ο οποίος βοηθά το μοντέλο να σταθμίσει τη σημασία διαφορετικών λέξεων σε μια πρόταση και ένα νευρωνικό δίκτυο πρόσω τροφοδότησης , το οποίο επεξεργάζεται τα δεδομένα εισόδου.

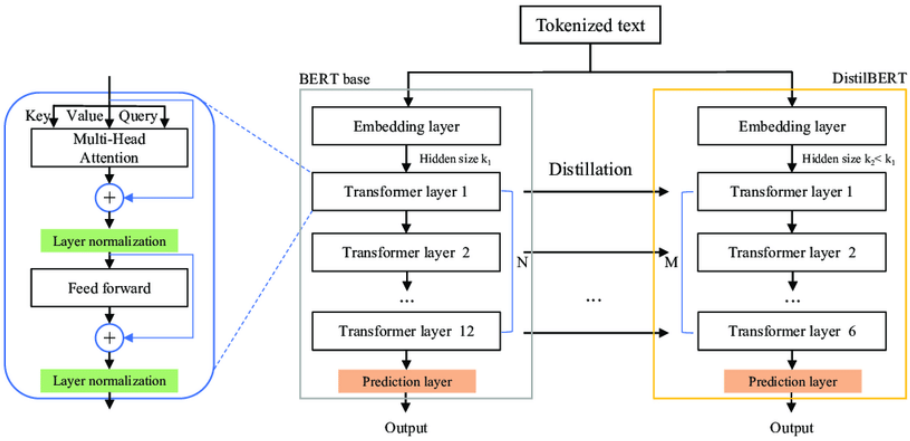
Στο κάτω μέρος της εικόνας, απεικονίζεται η κωδικοποίηση θέσης η οποία προστίθεται στα token της ακολουθίας εισόδου (t1 έως t5). Η κωδικοποίηση θέσης χρησιμοποιείται για να δώσει στο μοντέλο πληροφορίες σχετικά με τη σειρά των λέξεων, καθώς ο μηχανισμός αυτο-προσοχής από μόνος του δεν λαμβάνει υπόψη τη σειρά ακολουθίας. Οι έξοδοι του BERT στη συνέχεια περνούν μέσα από ένα επίπεδο ταξινόμησης το οποίο περιλαμβάνει τη συνάρτησης ενεργοποίησης Gelu και την κανονικοποίηση, για να παραχθούν οι τελικές προβλέψεις για κάθε διακριτικό (t1’ έως t5’).

2.5.2.6 Το μοντέλο DistilBERT

Το DistilBERT [26] είναι μια μικρότερη, ταχύτερη και ελαφρύτερη έκδοση του μοντέλου BERT, σχεδιασμένο για εργασίες NLP. Αναπτύσσεται μέσω μιας διαδικασίας που ονομάζεται απόσταξη γνώσης, η οποία περιλαμβάνει τη μεταφορά γνώσης από ένα μεγαλύτερο μοντέλο (στην περίπτωση αυτή το BERT) σε ένα μικρότερο μοντέλο (DistilBERT). Αυτή η διαδικασία επιτρέπει στο DistilBERT να επιτύχει μείωση του μεγέθους κατά 40% και αύξηση της ταχύτητας κατά 60%, διατηρώντας παράλληλα το 97% των δυνατοτήτων κατανόησης γλωσσών των BERT.

Οι βασικές διαφορές στην αρχιτεκτονική του DistilBERT σε σύγκριση με τα προηγούμενα μοντέλα BERT είναι:

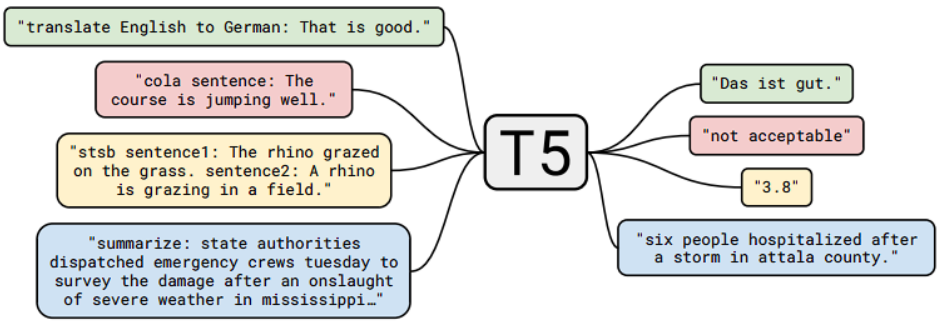
1. Λιγότερα στρώματα: Το DistilBERT έχει λιγότερα στρώματα από το BERT, γεγονός που συμβάλλει στο μικρότερο μέγεθος και την αυξημένη ταχύτητά του. Συγκεκριμένα αποτελείται από 6 στρώματα κωδικοποιητών σε αντίθεση με τα 12 του BERTBASE.
2. Μειωμένος αριθμός μηχανισμών προσοχής: Το DistilBERT έχει λιγότερους μηχανισμούς προσοχής σε σύγκριση με το BERT, γεγονός που συμβάλλει επίσης στο μικρότερο μέγεθος και την αυξημένη ταχύτητά του.
3. Το DistilBERT περιέχει 66 εκατομμύρια εκπαιδευόμενες παραμέτρους , αντί των 110 εκατομμυρίων του BERTBASE.
4. Η εκπαίδευση του DistilBERT έγινε σε χρόνο 3.5 GPU ημερών χρησιμοποιώντας 8 V100 κάρτες γραφικών.



Σχήμα 2.10: Η αρχιτεκτονική του DistilBERT [27].

2.5.3 T5

Το Text-To-Text Transfer Transformer (T5), είναι ένα μοντέλο NLP που προτάθηκε από ερευνητές της Google AI το 2019 [28]. Σε αντίθεση με τα παραδοσιακά μοντέλα NLP που έχουν σχεδιαστεί για συγκεκριμένες εργασίες όπως η μετάφραση, η σύνοψη ή η απάντηση ερωτήσεων, το T5 είναι ένα ευέλικτο και ενοποιημένο πλαίσιο που διατυπώνει όλες τις εργασίες NLP ως ένα πρόβλημα text-to-text. Η βασική ιδέα πίσω από το T5 είναι να μετατρέψει κάθε εργασία NLP σε εργασία δημιουργίας κειμένου. Αυτό σημαίνει ότι τόσο η είσοδος όσο και η έξοδος για οποιαδήποτε εργασία αντιπροσωπεύονται ως ακολουθίες κειμένου. Για παράδειγμα, για μια εργασία μετάφρασης, η είσοδος μπορεί να είναι: "Translate English to French: [sentence]" και η έξοδος θα είναι η μεταφρασμένη πρόταση στα Γαλλικά. Ομοίως, για μια εργασία σύνοψης, η είσοδος μπορεί να είναι: "Summarize the following text: [text]" και η έξοδος θα είναι η σύνοψη του κειμένου.



Σχήμα 2.11: Το πλαίσιο του μοντέλου T5 όπου κάθε εργασία NLP διατυπώνεται ως ένα πρόβλημα text-to-text [28].

Το μοντέλο T5 βασίζεται στην αρχιτεκτονική transformer. Αποτελείται από μια δομή κωδικοποιητή-αποκωδικοποιητή, όπως και άλλα μοντέλα Seq2seq. Χρησιμοποιεί μια στοίβα στρωμάτων transformer τόσο στον κωδικοποιητή όσο και στον αποκωδικοποιητή, επιτρέποντάς του να συλλάβει και να επεξεργαστεί τις ιεραρχικές αναπαραστάσεις των ακολουθιών εισόδου και εξόδου.

Ο κωδικοποιητής επεξεργάζεται την ακολουθία εισόδου, η οποία μπορεί να είναι ένας συνδυασμός αποτελούμενος από την περιγραφή της εργασίας και των δεδομένων εισόδου. Εφαρμόζει μηχανισμούς αυτο-προσοχής για να συλλάβει τις σχέσεις μεταξύ λέξεων και του πλαισίου των πληροφοριών. Ο μηχανισμός αυτο-προσοχής επιτρέπει στο μοντέλο να χειρίζεται διαφορετικά μέρη της ακολουθίας εισόδου, λαμβάνοντας υπόψη τις εξαρτήσεις και τους συσχετισμούς μεταξύ των λέξεων.

Ο αποκωδικοποιητής, από την άλλη πλευρά, δημιουργεί την ακολουθία εξόδου. Λαμβάνει την τελική κρυφή κατάσταση του κωδικοποιητή ως είσοδο και χειρίζεται τα σχετικά μέρη της ακολουθίας εισόδου κατά τη διάρκεια της διαδικασίας αποκωδικοποίησης. Ο αποκωδικοποιητής δημιουργεί την έξοδο βήμα προς βήμα, χρησιμοποιώντας τον μηχανισμό προσοχής για να εστιάσει σε διαφορετικά μέρη της εισόδου όσο προβλέπει το επόμενο token.

Τόσο τα επίπεδα κωδικοποιητή όσο και αποκωδικοποιητή στο T5 χρησιμοποιούν αυτο-προσοχή πολλαπλών κεφαλών, η οποία επιτρέπει στο μοντέλο να συλλάβει διαφορετικές εξαρτήσεις εντός των ακολουθιών εισόδου και εξόδου. Ο μηχανισμός προσοχής ενισχύει την ικανότητα του μοντέλου να χειρίζεται εξαρτήσεις μεγάλης εμβέλειας και να καταγράφει αποτελεσματικά το πλαίσιο της πληροφορίας.

Το T5 ενσωματώνει επίσης κωδικοποίηση θέσης για την κωδικοποίηση των πληροφοριών θέσης της ακολουθίας εισόδου. Αυτή η κωδικοποίηση θέσης βοηθά το μοντέλο να κατανοήσει τη σειρά και τη θέση των token στην ακολουθία, η οποία είναι ζωτικής σημασίας για την καταγραφή της διαδοχικής φύσης της γλώσσας [29].

2.5.3.1 Fine-tuning T5

Το fine-tuning του μοντέλου T5 περιλαμβάνει την προετοιμασία συνόλων δεδομένων, αποτελούμενα από ζεύγη εισόδου-εξόδου που σχετίζονται με την εργασία-στόχο. Αυτά τα ζεύγη ακολουθούν τη μορφή text-to-text, όπου το κείμενο εισόδου αντιπροσωπεύει την περιγραφή της εργασίας και τα δεδομένα εισόδου και το κείμενο εξόδου αντιπροσωπεύει την επιθυμητή έξοδο-στόχο. Εκπαιδεύοντας το T5 σε τέτοια δεδομένα, το μοντέλο μαθαίνει να αντιστοιχίζει το κείμενο εισόδου στο επιθυμητό κείμενο εξόδου και να προσαρμόζεται αποτελεσματικά στη συγκεκριμένη εργασία.

2.5.4 Prompt engineering

Το T5 δέχεται ως είσοδο μία προτροπή-οδηγία η οποία αποτελεί την περιγραφή της εργασίας που πρέπει το T5 να υλοποιήσει στα δεδομένα εισόδου. Η εύρεση της κατάλληλης προτροπής είναι μια σημαντική πτυχή του fine-tuning για το T5. Η διαδικασία της εύρεσης της κατάλληλης προτροπής προς ένα μοντέλο όπως το T5 ονομάζεται prompt engineering. Οι προτροπές είναι οδηγίες ή ενδείξεις για συγκεκριμένες εργασίες που παρέχονται για να καθοδηγήσουν τη συμπεριφορά του μοντέλου κατά τη διάρκεια της εξαγωγής συμπερασμάτων. Με τον προσεκτικό σχεδιασμό προτροπών, η έξοδος του μοντέλου μπορεί να επηρεαστεί και να κατευθυνθεί προς πιο ακριβείς και σωστές ως προς το πλαίσιο απαντήσεις. Το prompt engineering επιτρέπει τον καλύτερο έλεγχο της συμπεριφοράς του μοντέλου και διασφαλίζει ότι το μοντέλο ευθυγραμμίζεται με τις επιθυμητές απαιτήσεις της εργασίας.

2.5.5 Μεγάλα γλωσσικά μοντέλα(Large Language Models-LLM's)

Ένα μεγάλο γλωσσικό μοντέλο (LLM) είναι ένα γλωσσικό μοντέλο αξιοσημείωτο για την ικανότητά του να επιτυγχάνει παραγωγή γλώσσας γενικού σκοπού. Τα LLM αποκτούν αυτή την ικανότητα μαθαίνοντας στατιστικές σχέσεις από έγγραφα κειμένου κατά τη διάρκεια μιας υπολογιστικά δαπανηρής αυτο-επιβλεπόμενης και ημι-επιβλεπόμενης εκπαιδευτικής διαδικασίας. Τα LLM είναι τεχνητά νευρωνικά δίκτυα και η αρχιτεκτονική με την οποία κατασκευάζονται συνήθως είναι βασισμένη σε transformers.

Τα LLM μπορούν να χρησιμοποιηθούν για την παραγωγή κειμένου, μια μορφή γεννητικής τεχνητής νοημοσύνης (generative AI-GenAI), λαμβάνοντας ένα κείμενο εισαγωγής και προβλέποντας επανειλημμένα το επόμενο token ή λέξη. Μέχρι το 2020, το fine tuning ήταν ο μόνος τρόπος με τον οποίο ένα μοντέλο μπορούσε να προσαρμοστεί ώστε να είναι σε θέση να ολοκληρώσει συγκεκριμένες εργασίες. Μοντέλα μεγαλύτερου μεγέθους, όπως το GPT-3, ωστόσο, μπορούν άμεσα να επιτύχουν

παρόμοια αποτελέσματα μέσω του prompt enginnering. Τα LLM αποκτούν γνώσεις σχετικά με τη σύνταξη, τη σημασιολογία και την «οντολογία» που είναι έμφυτη στα σώματα κειμένων της ανθρώπινης γλώσσας, αλλά και σχετικά με τις ανακρίβειες και τις μεροληψίες που υπάρχουν στα σώματα κειμένων [4].

Κεφάλαιο 3

Εργαλεία υλοποίησης

1. Google Colab: Η συγγραφή και η εκτέλεση του κώδικα Python για την υλοποίηση, την εκπαίδευση και την αξιολόγηση των μοντέλων της διπλωματικής εργασίας έγινε στο περιβάλλον του Google Colab. Το Google Colab είναι μια cloud-based πλατφόρμα που παρέχεται από την Google και προσφέρει ένα δωρεάν περιβάλλον Jupyter Notebook. Επιτρέπει στους χρήστες να γράφουν και να εκτελούν κώδικα Python συνεργατικά μέσω του προγράμματος περιήγησης χωρίς να απαιτείται καμία ρύθμιση. Το Google Colab παρέχει δωρεάν πρόσβαση σε μια εικονική μηχανή με υποστήριξη CPU και GPU. Οι χρήστες μπορούν να εκτελέσουν κώδικα Python, να εκτελέσουν μοντέλα μηχανικής μάθησης και να εκτελέσουν ανάλυση δεδομένων χωρίς να χρειάζεται να εγκαταστήσουν λογισμικό τοπικά.

Συγκεκριμένα, για την εκπαίδευση του μοντέλου REBERT, η οποία ήταν υπολογιστικά απαιτητική, αποκτήθηκε η συνδρομητική έκδοση του Google Colab, Google Colab Pro. Το Google Colab Pro είναι μια premium έκδοση του Google Colab που προσφέρει πρόσθετες δυνατότητες και πόρους για χρήστες που απαιτούν περισσότερη υπολογιστική ισχύ και προηγμένες λειτουργίες. Το μοντέλο REBERT εκπαιδεύτηκε με τη χρήση της V100 GPU, ενώ το μοντέλο NERBERT εκπαιδεύτηκε με την A100 GPU του Google Colab Pro.

2. Tensorflow: Το TensorFlow χρησιμοποιήθηκε για διάφορες βασικές εργασίες, συμπεριλαμβανομένης της φόρτωσης μοντέλου, του tokenization, της προετοιμασίας συνόλου δεδομένων, της μεταγλώττισης και της εκπαίδευσης των μοντέλων. Το TensorFlow ενσωματώνει το Keras API για την κατασκευή και την εκπαίδευση μοντέλων. Το Keras παρέχει μια φιλική προς το χρήστη διεπαφή για την κατασκευή νευρωνικών δικτύων και χρησιμοποιείται ευρέως για την απλότητα και την ευελιξία του. Η συμβατότητα του TensorFlow με τη βιβλιοθήκη Hugging Face Transformers διευκολύνει την απρόσκοπτη ενσωμάτωση προ-εκπαιδευμένων μοντέλων και την ανάπτυξη προσαρμοσμένων μοντέλων για τις επιθυμητές εργασίες του κάθε μοντέλου. Το TensorFlow είναι μια βιβλιοθήκη μηχανικής εκμάθησης ανοιχτού κώδικα που αναπτύχθηκε από την Google. Παρέχει εργαλεία για την κατασκευή και την εκπαίδευση μοντέλων μηχανικής μάθησης, ιδιαίτερα νευρωνικών δικτύων.

3. Hugging Face Transformers: Το Hugging Face Transformers δημιουργήθηκε από την εταιρεία Hugging Face και είναι μια βιβλιοθήκη ανοιχτού κώδικα που παρέχει ένα ολοκληρωμένο σύνολο προ-εκπαιδευμένων μοντέλων NLP. Η βιβλιοθήκη είναι χτισμένη πάνω από το PyTorch και το TensorFlow, καθιστώντας την συμβατή και με τα δύο πλαίσια βαθιάς μάθησης. Η κύρια εστίασή του είναι σε μοντέλα που βασίζονται σε transformer. Το Hugging Face Transformers προσφέρει ένα ευρύ φάσμα προ-εκπαιδευμένων μοντέλων για εργασίες όπως ταξινόμηση κειμένου, αναγνώριση

ονοματικών οντοτήτων, απάντηση ερωτήσεων, μετάφραση γλώσσας και πολλά άλλα. Η βιβλιοθήκη περιλαμβάνει δημοφιλή μοντέλα όπως τα BERT, GPT, RoBERTa, DistilBERT και T5, μεταξύ άλλων. Επιπλέον παρέχει εύχρηστους tokenizers για διάφορα προ-εκπαιδευμένα μοντέλα. Αυτοί οι tokenizers χειρίζονται τη μετατροπή ακατέργαστου κειμένου σε ακολουθίες token συμβατές με το εκάστοτε μοντέλο. Το Hugging Face Transformers απλοποιεί τη χρήση προ-εκπαιδευμένων μοντέλων μέσω pipelines για συγκεκριμένες εργασίες. Αυτά τα pipelines καλύπτουν εργασίες όπως η δημιουργία κειμένου, η ταξινόμηση κειμένου, η αναγνώριση ονοματικών οντοτήτων, η μετάφραση και η σύνοψη, διευκολύνοντας την εφαρμογή μοντέλων σε πραγματικές περιπτώσεις χρήσης. Επίσης, στο Hugging Face Model Hub οι χρήστες μπορούν να αποθηκεύσουν, να μοιραστούν, να ανακαλύψουν και να διανείμουν μοντέλα εκπαιδευμένα με Hugging Face Transformers.

4. spaCy: Το spaCy αναπτύχθηκε από την Explosion AI και είναι μια βιβλιοθήκη NLP ανοιχτού κώδικα που έχει σχεδιαστεί για την αποτελεσματική επεξεργασία και ανάλυση δεδομένων ανθρώπινης γλώσσας. Το spaCy παρέχει μια ολοκληρωμένη σουίτα εργαλείων για εργασίες όπως tokenization, προσθήκη ετικετών μέρους ομιλίας, αναγνώριση ονοματικών οντοτήτων και συντακτική ανάλυση. Ένα από τα βασικά χαρακτηριστικά του spaCy είναι τα προ-εκπαιδευμένα στατιστικά μοντέλα του, τα οποία αξιοποιούν τεχνικές μηχανικής μάθησης για την εκτέλεση διαφόρων εργασιών NLP. Αυτά τα μοντέλα εκπαιδεύονται σε μεγάλα σώματα κειμένων και καλύπτουν πολλές γλώσσες, επιτρέποντας στους χρήστες να εφαρμόζουν προηγμένες τεχνικές NLP χωρίς την ανάγκη εκτεταμένων δεδομένων εκπαίδευσης. Το φιλικό προς το χρήστη API του spaCy επιτρέπει στους προγραμματιστές και τους ερευνητές να το ενσωματώσουν απρόσκοπτα στις εφαρμογές τους, καθιστώντας το ένα πολύτιμο εργαλείο για έργα που κυμαίνονται από την εξαγωγή πληροφοριών και την ανάλυση συναισθήματος έως την ανάπτυξη chatbot και την κατηγοριοποίηση εγγράφων. Επιπλέον, το spaCy υποστηρίζει μοντέλα βαθιάς μάθησης, παρέχοντας ευελιξία στους χρήστες που επιθυμούν να ενσωματώσουν τεχνικές αιχμής στις ροές εργασίας NLP.

5. pandas: Η βιβλιοθήκη pandas είναι μια ισχυρή βιβλιοθήκη ανοιχτού κώδικα που χρησιμοποιείται για την ανάλυση και το χειρισμό δεδομένων. Παρέχει δομές δεδομένων για την αποτελεσματική αποθήκευση και χειρισμό μεγάλων, ετερογενών συνόλων δεδομένων και εργαλεία για την απρόσκοπτη εργασία με δομημένα δεδομένα. Οι κύριες δομές δεδομένων στην pandas είναι τα Series (μονοδιάστατοι επισημασμένοι πίνακες) και DataFrame (δισδιάστατες επισημασμένες δομές δεδομένων όπου οι στήλες μπορεί να είναι διαφορετικών τύπων). Το pandas χρησιμοποιείται ευρέως στην επιστήμη των δεδομένων, τη στατιστική και την οικονομική επιστήμη για εργασίες όπως ο καθαρισμός, η εξερεύνηση, ο μετασχηματισμός και η ανάλυση δεδομένων. Η βιβλιοθήκη απλοποιεί πολλές κοινές εργασίες χειρισμού δεδομένων και προσφέρει μια ευέλικτη και εύχρηστη διεπαφή για εργασία με δεδομένα σε μορφή πίνακα στην Python.

6. NumPy: Το NumPy είναι μια ισχυρή αριθμητική υπολογιστική βιβλιοθήκη στην Python που παρέχει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες μαζί με μια συλλογή μαθηματικών συναρτήσεων υψηλού επιπέδου για λειτουργία σε αυτούς τους

πίνακες. Είναι ένα θεμελιώδες πακέτο για επιστημονικούς υπολογισμούς στην Python και χρησιμεύει ως βάση για πολλές άλλες επιστημονικές βιβλιοθήκες.

7. Seqeval: Η Seqeval είναι μια βιβλιοθήκη Python σχεδιασμένη για την αξιολόγηση συστημάτων ταξινόμησης ακολουθίας. Είναι ιδιαίτερα χρήσιμη για την αξιολόγηση της απόδοσης μοντέλων που ασχολούνται με εργασίες όπως η αναγνώριση ονοματικών οντοτήτων, η προσθήκη ετικετών μέρους του λόγου και άλλα παρόμοια προβλήματα ταξινόμησης ακολουθίας. Η βιβλιοθήκη επικεντρώνεται στην αξιολόγηση των precision, recall και F1 score συγκρίνοντας τις προβλέψεις των συστημάτων επισήμανσης ακολουθίας με την εδαφική αλήθεια.

8. Scikit-learn: Το Scikit-learn, ή sklearn, είναι μια ευρέως χρησιμοποιούμενη βιβλιοθήκη μηχανικής μάθησης ανοιχτού κώδικα στην Python που παρέχει απλά και αποτελεσματικά εργαλεία για ανάλυση και μοντελοποίηση δεδομένων. Είναι χτισμένο πάνω από άλλες δημοφιλείς βιβλιοθήκες όπως το NumPy, το SciPy και το Matplotlib, καθιστώντας το ένα ισχυρό και ευέλικτο εργαλείο για διάφορες εργασίες μηχανικής μάθησης. Μια κρίσιμη πτυχή της μηχανικής μάθησης είναι η αξιολόγηση της απόδοσης των μοντέλων για να διασφαλιστεί η αποτελεσματικότητά τους και η γενίκευσή τους σε νέα, αόρατα δεδομένα. Το Scikit-learn προσφέρει μια ολοκληρωμένη σειρά λειτουργιών και κλάσεων για την αξιολόγηση μοντέλων, επιτρέποντας στους χρήστες να αξιολογήσουν την ποιότητα των μοντέλων τους μέσω διαφόρων μετρήσεων και τεχνικών.

Κεφάλαιο 4

Εξαγωγή πληροφοριών από κείμενο

4.1 Δημιουργία μοντέλου NER

Η δημιουργία του μοντέλου NER το οποίο ονομάσαμε ως NERBERT, πραγματοποιείται μέσω του fine-tuning του μοντέλου DistilBERT στο σύνολο δεδομένων OntoNotes. Αναλυτικά, τα βήματα και οι διαδικασίες που εφαρμόστηκαν για τη δημιουργία αυτού του μοντέλου είναι τα εξής:

4.1.1 Σύνολο δεδομένων

Το σύνολο δεδομένων που επιλέχτηκε για τη δημιουργία του NER μοντέλου είναι το OntoNotes [30] και συγκεκριμένα η ακόλουθη υλοποίηση [31] από τη βιβλιοθήκη συνόλων δεδομένων του Hugging Face: https://huggingface.co/datasets/SpeedOfMagic/ontonotes_english

Το OntoNotes είναι ένα μεγάλο πολύγλωσσο (Αγγλικά, Κινέζικα, Αραβικά) σώμα κειμένων που περιλαμβάνει διάφορα είδη κειμένου (ειδήσεις, τηλεφωνικές ομιλίες ,εκπομπές). Είναι επισημασμένο με ονοματικές οντότητες, συναναφορές, ετικέτες μέρους του λόγου και άλλες γλωσσικές πληροφορίες. Χρησιμοποιείται για την εκπαίδευση και την αξιολόγηση συστημάτων NLP. Η υλοποίηση που χρησιμοποιήθηκε περιέχει κείμενα με επισημάνσεις μόνο ως προς τις ονοματικές οντότητες.

Κάθε παράδειγμα του συγκεκριμένου συνόλου δεδομένων αποτελείται από 2 στήλες. Η πρώτη στήλη είναι η λίστα των tokens τα οποία αναπαριστούν τις λέξεις και τα σημεία στίξης της πρότασης κειμένου. Η δεύτερη στήλη περιέχει τα 'ner_tags'. Τα ner_tags είναι μια λίστα ετικετών κλάσης, όπου κάθε στοιχείο της λίστας αντιστοιχεί σε μία ετικέτα ονοματικής οντότητας για το αντίστοιχο token. Αυτή η λίστα είναι μία λίστα ακεραίων, όπου κάθε ακέραιος αποτελεί μία κωδικοποίηση της αντίστοιχης ετικέτας ονοματικής οντότητας.

Το πεδίο των ner_tags χρησιμοποιεί το σχήμα ετικετοποίησης BIO. Το ακρωνύμιο BIO αντιστοιχεί στα Beggining, Inside, Outside και υποδεικνύει τη θέση μίας λέξης μέσα σε μία ονοματική οντότητα.

Το πρόθεμα 'B-' χρησιμοποιείται για την προσθήκη ετικέτας στην πρώτη λέξη μιας ονοματικής οντότητας. Παράδειγμα: Το "B-PERSON" υποδεικνύει την αρχή του ονόματος ενός ατόμου.

Το πρόθεμα 'I-' χρησιμοποιείται για την προσθήκη ετικετών σε λέξεις μέσα σε μια ονοματική οντότητα, εξαιρουμένης της πρώτης λέξης. Παράδειγμα: Το "I-PERSON" υποδεικνύει μια λέξη μέσα στο όνομα ενός ατόμου.

Η ετικέτα 'O' χρησιμοποιείται για να υποδείξει ότι μια λέξη βρίσκεται εκτός οποιασδήποτε ονοματικής οντότητας ή δεν αποτελεί μέρος μιας ονοματικής οντότητας.

Παράδειγμα 2:
Πρόταση κειμένου: "John Smith works at ABC Corp in New York." Πρόταση μετά το BIO tagging: ” B-PERSON I-PERSON O O B-ORG I-ORG O B-LOC I-LOC"

Το σύνολο των ετικετών ορίζεται ως ένα αντικείμενο datasets.ClassLabel που αποτελείται από 37 κλάσεις. Οι 37 κλάσεις αναπαριστούν όλες τις διαφορετικές κατηγορίες των ονοματικών οντοτήτων οι οποίες είναι οι παρακάτω με την ακριβή αντιστοίχιση τους σε έναν ακέραιο αριθμό:

0. "O" (Outside)
1. "B-PERSON" (Beginning of a person)
2. "I-PERSON" (Inside a person)
3. "B-NORP" (Beginning of a nationalities or religious or political groups)
4. "I-NORP" (Inside a NORP)
5. "B-FAC" (Beginning of a facility)
6. "I-FAC" (Inside a facility)
7. "B-ORG" (Beginning of an organization)
8. "I-ORG" (Inside an organization)
9. "B-GPE" (Beginning of a geopolitical entity)
10. "I-GPE" (Inside a geopolitical entity)
11. "B-LOC" (Beginning of a location)
12. "I-LOC" (Inside a location)
13. "B-PRODUCT" (Beginning of a product)
14. "I-PRODUCT" (Inside a product)
15. "B-DATE" (Beginning of a date)
16. "I-DATE" (Inside a date)
17. "B-TIME" (Beginning of a time)
18. "I-TIME" (Inside a time)
19. "B-PERCENT" (Beginning of a percentage)
20. "I-PERCENT" (Inside a percentage)
21. "B-MONEY" (Beginning of a monetary value)
22. "I-MONEY" (Inside a monetary value)
23. "B-QUANTITY" (Beginning of a quantity)
24. "I-QUANTITY" (Inside a quantity)
25. "B-ORDINAL" (Beginning of an ordinal number)
26. "I-ORDINAL" (Inside an ordinal number)

27. "B-CARDINAL" (Beginning of a cardinal number)
28. "I-CARDINAL" (Inside a cardinal number)
29. "B-EVENT" (Beginning of an event)
30. "I-EVENT" (Inside an event)
31. "B-WORK_OF_ART" (Beginning of a work of art)
32. "I-WORK_OF_ART" (Inside a work of art)
33. "B-LAW" (Beginning of a law)
34. "I-LAW" (Inside a law)
35. "B-LANGUAGE" (Beginning of a language)
36. "I-LANGUAGE" (Inside a language)

Κάθε κλάση συσχετίζεται με έναν συγκεκριμένο τύπο ονοματικής οντότητας και το BIO tagging χρησιμοποιείται για την επισήμανση των οντοτήτων μέσα στο κείμενο.

Παράδειγμα του συνόλου δεδομένων OntoNotes:
{'tokens': ['Sunday', 'the', 'interview', 'with', 'Bob', 'Shapiro', '.'], 'ner_tags': [15, 0, 0, 0, 1, 2, 0]}

Το σύνολο δεδομένων είναι χωρισμένο σε τρία υποσύνολα: εκπαίδευσης, αξιολόγησης και δοκιμών τα οποία αποτελούνται από 59924, 13900 και 8262 παραδείγματα αντίστοιχα.

4.1.2 Προεπεξεργασία δεδομένων

Το πρώτο βήμα για την προεπεξεργασία των δεδομένων είναι το tokenization. Αυτό γίνεται με τη χρήση του tokenizer του μοντέλου DistilBERT. Ο tokenizer εκτελεί τις παρακάτω ενέργειες στα δεδομένα εισόδου:

1. WordPiece tokenization: Ο tokenizer του DistilBERT εκτελεί subword tokenization, δηλαδή σπάει κάθε λέξη σε μονάδες υπολέξεων, χρησιμοποιώντας το μοντέλο WordPiece.

2. Truncation: Με την παράμετρο truncation=True η ακολουθία εισόδου περικόπτεται εάν υπερβαίνει το μέγιστο όριο tokens που μπορεί να δεχθεί το DistilBERT.

3. Special tokens: Εισάγονται ειδικά tokens όπως: [CLS] (classification token), [SEP] (separator token) και [PAD] (padding token).

4. Token IDs: Κάθε token αντιστοιχίζεται στο αντίστοιχο ακέραιο ID του στο λεξιλόγιο.

5. Attention mask: Δημιουργείται μια μάσκα προσοχής για να υποδείξει ποια tokens είναι πραγματικές λέξεις και ποια είναι padding tokens.

Στο παραπάνω παράδειγμα φαίνεται ότι οι λέξεις είναι ήδη διαχωρισμένες σε token. Τα δεδομένα όμως πρέπει να περάσουν από τα παραπάνω βήματα που εκτελεί ο tokenizer για να είναι κατάλληλα για εισαγωγή στο DistilBERT. Ωστόσο, επειδή τα παραδείγματα είναι ήδη διαχωρισμένα σε token ορίζεται η παράμετρος is_split_into_words=True.

Η διαδικασία του tokenization προσθέτει ειδικά tokens ([CLS], [SEP]) στις ακολουθίες εισόδου. Επίσης, το subword tokenization δημιουργεί μια αναντιστοιχία μεταξύ της εισόδου και των ετικετών αφού μία λέξη που αντιστοιχεί σε μία μόνο ετικέτα μπορεί τώρα να έχει χωριστεί σε δύο υπολέξεις. Για αυτό το λόγο, τα tokens πρέπει να ευθυγραμμιστούν με τις ετικέτες. Αυτό γίνεται ως εξής:

1. Αντιστοίχιση όλων των token στην αντίστοιχη λέξη τους με τη μέθοδο word_ids: Για κάθε παράδειγμα, χρησιμοποιείται η μέθοδος word_ids για τη λήψη των δεικτών λέξεων που αντιστοιχούν σε κάθε token. Εάν ένα token είναι μέρος μιας λέξης, το word_id του δείχνει σε ποια λέξη της αρχικής πρότασης αναφέρεται.

2. Χειρισμός ειδικών token: Τα ειδικά token ([CLS], [SEP], [PAD]) που εισάγονται κατά τη διάρκεια του tokenization έχουν το None ως word_id τους. Αυτά τα token δεν αντιστοιχούν σε κάποια ετικέτα οπότε τους αποδίδεται η ετικέτα -100. Με αυτόν τον τρόπο αυτά τα token αγνοούνται από τη συνάρτηση απωλειών κατά την εκπαίδευση.

3. Μετατόπιση ετικέτας (λειτουργία shift_label): Για κάθε ετικέτα οντότητας στις αρχικές ετικέτες, η συνάρτηση ελέγχει εάν η ετικέτα είναι περιττός αριθμός (υποδεικνύοντας την αρχή μιας οντότητας). Εάν είναι περιττός αριθμός, η ετικέτα αυξάνεται κατά 1 για να γίνει ζυγός αριθμός. Ο σκοπός είναι να γίνει διάκριση μεταξύ των αρχικών και ενδιάμεσων token της ίδιας οντότητας.

4. Ευθυγράμμιση ετικετών (λειτουργία align_labels_with_tokens): Ορίζεται η μεταβλητή current_word σε None για να παρακολουθείται η τρέχουσα λέξη που υποβάλλεται σε επεξεργασία και η λίστα new_labels για να αποθηκευτούν οι καινούριες ευθυγραμμισμένες ετικέτες. Για κάθε λέξη (token) στην tokenized είσοδο: Εάν το word_id ισούται με None, υποδεικνύοντας ένα token που δεν αποτελεί μέρος των αρχικών λέξεων (π.χ. [CLS] token), η αντίστοιχη ετικέτα ορίζεται σε -100 για να αγνοηθεί κατά τη διάρκεια της εκπαίδευσης. Εάν το word_id είναι διαφορετικό από το current_word, ξεκινά μια νέα οντότητα και η ετικέτα για το τρέχον word_id εκχωρείται απευθείας στη λίστα new_labels. Εάν το word_id είναι ίδιο με την τρέχουσα λέξη, αυτό σημαίνει ότι η οντότητα συνεχίζει και η μετατοπισμένη ετικέτα (που λαμβάνεται με χρήση shift_label) εκχωρείται στη λίστα new_labels.

5. Data Collation: Έπειτα, χρησιμοποιείται η κλάση DataCollatorForTokenClassification από τη βιβλιοθήκη Transformers για το batching, padding και collating των tokenized εισόδων και ετικετών. Αυτό το βήμα είναι ζωτικής σημασίας για την προετοιμασία των δεδομένων σε μορφή κατάλληλη για εκπαίδευση μοντέλου. Με το padding διασφαλίζεται ότι οι ακολουθίες εισόδου έχουν ομοιόμορφα μήκη πριν δημιουργηθούν τα batches.

4.1.3 Διαμόρφωση μοντέλου

1. Αντιστοίχιση ετικετών ονοματικών οντοτήτων σε ακεραίους: Ορίζονται δύο λεξικά, id2label και label2id, τα οποία χρησιμοποιούνται για την αντιστοίχιση ακέραιων αριθμητικών δεικτών ετικετών στις αντίστοιχες ετικέτες τους και αντίστροφα. Ο σκοπός της δημιουργίας αυτών των λεξικών είναι να δημιουργηθεί μια βολική αντιστοίχιση μεταξύ των αριθμητικών δεικτών ετικετών και των αντίστοιχων ονομάτων ετικετών τους. Στις εργασίες ταξινόμησης token, τα μοντέλα συνήθως προβλέπουν ακέραιους δείκτες ως έξοδο και αυτά τα λεξικά διευκολύνουν τη μετατροπή μεταξύ αυτών των δεικτών και των αναγνώσιμων από τον άνθρωπο ονομάτων ετικετών.

2. Φόρτωση μοντέλου: Πραγματοποιείται η φόρτωση και η αρχικοποίηση ενός μοντέλου DistilBERT το οποίο είναι fine-tuned στην ταξινόμηση token. Αυτό συμβαίνει αρχικοποιώντας ένα αντικείμενο της κλάσης TFAutoModelForTokenClassification. Αυτή η κλάση είναι μία κλάση της βιβλιοθήκης Hugging Face Transformers σχεδιασμένη στο Tensorflow και ρυθμισμένη για εργασίες ταξινόμησης token. Κληρονομεί την αρχιτεκτονική του μοντέλου BERT που επιλέγεται και επιπλέον έχει μία κεφαλή ταξινόμησης token πάνω από τη βάση του DistilBERT, υπεύθυνη για την πρόβλεψη των ετικετών για κάθε token σε μια ακολουθία. Επίσης χρησιμοποιεί τη συνάρτηση ενεργοποίησης softmax για να μετατρέψει τα logits του μοντέλου σε βαθμολογίες πιθανότητας για κάθε κλάση.

Ως παραμέτρους δέχεται το όνομα ή το αναγνωριστικό του προ-εκπαιδευμένου μοντέλου BERT που θα φορτωθεί. Στην περίπτωση μας το μοντέλο είναι το distilbert-base-cased. Επίσης, δέχεται τον αριθμό των ετικετών που θα προβλέπει το μοντέλο (στην περίπτωση μας 37) και τα λεξικά τα οποία περιέχουν τις αντιστοιχίσεις μεταξύ δεικτών ετικετών και ονομάτων ετικετών.

3. Υπερπαράμετροι μοντέλου: Ορίζονται οι υπερπαράμετροι του μοντέλου, δηλαδή ο αριθμός του batch_size (32) ο αριθμός των εποχών (epoch) εκπαίδευσης (3), ο αριθμός των βημάτων εκπαίδευσης (len(tokenized_dataset["train"]) // batch_size) * num_train_epochs και το πρόγραμμα του ρυθμού εκπαίδευσης (learning rate). Στο πρόγραμμα του ρυθμού εκπαίδευσης ορίζεται ο ρυθμός εκπαίδευσης (3e-5) , ο αριθμός των βημάτων εκπαίδευσης και το weight_decay_rate (0.01) το οποίο αποτελεί έναν όρο κανονικοποίησης που τιμωρεί τα μεγάλα βάρη με σκοπό να αποφευχθεί το overfitting. Χρησιμοποιείται ο προεπιλεγμένος βελτιστοποιητής του TFAutoModelForTokenClassification ο οποίος είναι ο AdamW και η προεπιλεγμένη συνάρτηση απωλειών η οποία είναι η SparseCategoricalCrossentropy.

4.1.4 Εκπαίδευση μοντέλου

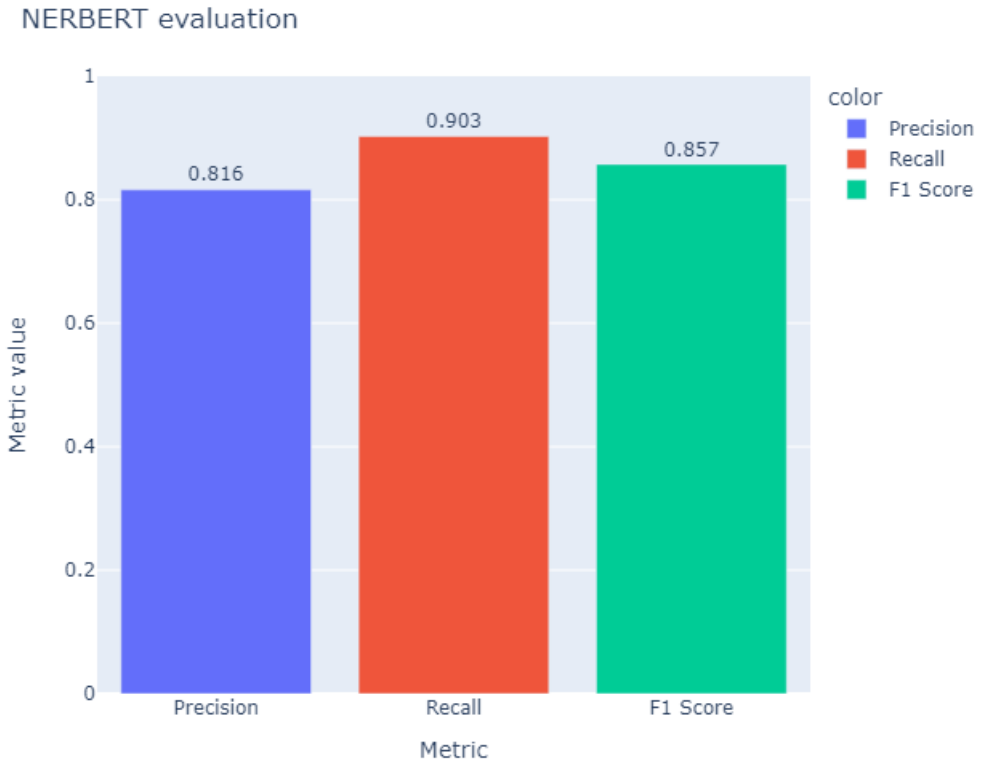
1. Προετοιμασία συνόλων δεδομένων εκπαίδευσης και επικύρωσης: Πραγματοποιείται προετοιμασία των συνόλων δεδομένων TensorFlow για εκπαίδευση (train_set) και επικύρωση (validation_set). Τα σύνολα αυτά περιέχουν τα tokenized δεδομένα, στα οποία εφαρμόζεται ανακάτεμα για το σύνολο εκπαίδευσης, ορίζεται το μέγεθος παρτίδας (batch_size) και χρησιμοποιείται ένα εργαλείο συλλογής δεδομένων (data_collator) για να δημιουργηθούν παρτίδες κατάλληλες για εκπαίδευση και επικύρωση.
2. Callbacks: Ορίζονται 2 callbacks τα οποία χρησιμοποιούνται κατά τη διάρκεια της εκπαίδευσης του μοντέλου. Το πρώτο είναι το PushToHubCallback το οποίο χρησιμοποιείται για την μεταφόρτωση του εκπαιδευμένου μοντέλου και των σχετικών αρχείων στο προσωπικό Hugging Face Model Hub έτσι ώστε να είναι ευκολότερη η μετέπειτα χρήση του μοντέλου για καινούρια παραδείγματα. Το δεύτερο είναι το KerasMetricCallback το οποίο χρησιμοποιείται για τον υπολογισμό και την καταγραφή μετρήσεων κατά τη διάρκεια της εκπαιδευτικής διαδικασίας. Η μέτρηση γίνεται μέσω της συνάρτησης compute_metrics η οποία χρησιμοποιεί το πλαίσιο seqeval για τον υπολογισμό των precision, recall και F1 score.
3. Μεταγλώττιση και εκπαίδευση μοντέλου: Αρχικά γίνεται μεταγλώττιση του μοντέλου, κατά την οποία ορίζονται ως παράμετροι οι υπερπαράμετροι του μοντέλου που ορίστηκαν νωρίτερα. Έπειτα, με τη μέθοδο .fit ξεκινάει η εκπαίδευση του μοντέλου στα σύνολα εκπαίδευσης και επικύρωσης και κατά την εκπαίδευση εφαρμόζονται τα callbacks που ορίστηκαν προηγουμένως.

4.1.5 Αξιολόγηση μοντέλου

Η αξιολόγηση έγινε για τις μετρικές των precision, recall και F1 score και τα αποτελέσματα είναι τα εξής:

Μετρική	**Τιμή**
Precision	0,8161
Recall	0,9027
F1 score	0,8572

Πίνακας 4.1: Μετρικές αξιολόγησης του μοντέλου NERBERT



Σχήμα 4.1: Μετρικές αξιολόγησης του μοντέλου NERBERT

4.2 Δημιουργία μοντέλου RC

Η δημιουργία του μοντέλου RC, το οποίο ονομάσαμε ως REBERT, επιτυγχάνεται μέσω του fine-tuning του μοντέλου DistilBERT στο σύνολο δεδομένων T-Rex. Μία τεχνική που αποδείχθηκε χρήσιμη για την βελτίωση της απόδοσης του μοντέλου, ως προς την ταξινόμηση σχέσεων μεταξύ οντοτήτων, είναι η επισήμανση της θέσης των οντοτήτων που εμπλέκονται σε μία σχέση μέσα στο κείμενο, χρησιμοποιώντας ειδικά markers [32], [33]. Αναλυτικά, τα βήματα που εφαρμόστηκαν για τη δημιουργία αυτού του μοντέλου είναι τα εξής:

4.2.1 Σύνολο δεδομένων

Το σύνολο δεδομένων που επιλέχτηκε για τη δημιουργία του RC μοντέλου είναι το T-Rex [34] και συγκεκριμένα η ακόλουθη υλοποίηση [35] από τη βιβλιοθήκη συνόλων δεδομένων του Hugging Face: https://huggingface.co/datasets/relbert/t_rex . Το T-Rex είναι ένα σύνολο δεδομένων μεγάλης κλίμακας το οποίο αποτελείται από αντιστοιχίσεις μεταξύ τριπλετών Γνωσιακής Βάσης (KB) και κειμένων φυσικής γλώσσας. Αποτελείται από 11 εκατομμύρια τριπλέτες ευθυγραμμισμένες με 3,09 εκατομμύρια περιλήψεις του Wikipedia, καλύπτοντας περισσότερα από 600 μοναδικά κατηγορήματα της γνωσιακής βάσης δεδομένων Wikidata. Το T-REx είναι δύο τάξεις μεγέθους μεγαλύτερο από το μεγαλύτερο διαθέσιμο σύνολο δεδομένων αντιστοιχίσεων και καλύπτει 2,5 φορές περισσότερα κατηγορήματα.

Η υλοποίηση που χρησιμοποιήθηκε αποτελεί μία φιλτραρισμένη έκδοση του T-Rex και περίεχει 1,5 εκατομμύρια περιλήψεις Wikipedia και 769 μοναδικά κατηγορήματα. Αυτό το σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης για την ταξινόμηση σχέσεων, επιτρέποντας στο μοντέλο να προβλέψει τις σχέσεις μεταξύ οντοτήτων σε νέα κείμενα.

Κάθε παράδειγμα του συγκεκριμένου συνόλου δεδομένων αποτελείται από 5 στήλες:

1. Relation: Αυτή η στήλη περιέχει την ετικέτα σχέσης μεταξύ δύο οντοτήτων. Η πλειοψηφία των σχέσεων στο σύνολο δεδομένων είναι στη μορφή PID. Αυτό το ID αντιστοιχεί σε μία μοναδική σχέση στη γνωσιακή βάση Wikidata (π.χ. P27 = country of citizenship). Σε μια εργασία ταξινόμησης σχέσεων, αυτές οι ετικέτες αντιπροσωπεύουν συνήθως διαφορετικούς τύπους σχέσεων που το μοντέλο στοχεύει να ταξινομήσει.
2. Head: Αυτή η στήλη περιέχει την πρώτη οντότητα που εμπλέκεται στη σχέση. Αυτή η οντότητα είναι συνήθως το υποκείμενο ή η κύρια οντότητα ενδιαφέροντος στη σχέση.
3. Tail: Αυτή η στήλη περιέχει τη δεύτερη οντότητα που εμπλέκεται στη σχέση. Η οντότητα ουρά είναι συνήθως το αντικείμενο ή η οντότητα που έχει συγκεκριμένη σχέση με την επικεφαλής οντότητα.
4. Title: Ο τίτλος του εγγράφου ή της πηγής από την οποία εξάγεται το παράδειγμα.
5. Text: Το πραγματικό περιεχόμενο του εγγράφου ή του αποσπάσματος κειμένου που περιέχει πληροφορίες σχετικά με τη σχέση μεταξύ των οντοτήτων κεφαλής και ουράς.

Παράδειγμα του συνόλου δεδομένων T-Rex:
{'relation': 'P54', 'head': 'Roger Nilsen', 'tail': 'Molde', 'title': 'Roger Nilsen', 'text': 'Roger Nilsen (born 8 August 1969 in Tromsø) is a Norwegian football coach and former defender. He played 32 matches and scored three goals for Norway. Nilsen played for the Norwegian clubs Tromsø, Viking, Molde and Bryne, and spent time abroad at 1. FC Köln, Sheffield United, Tottenham Hotspur and Grazer AK. He has later worked as assistant coach at Viking.'}

4.2.2 Προεπεξεργασία συνόλου δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε έχει το πλεονέκτημα ότι παρέχει μια μεγάλη πληθώρα παραδειγμάτων κειμένου και ένα μεγάλο αριθμό ξεχωριστών σχέσεων. Παρ'όλα αυτά, λόγω του μεγάλου αριθμού των μοναδικών σχέσεων (769), συναντήθηκε το πρόβλημα της ανισορροπίας κλάσεων (class imbalance). Η ανισορροπία κλάσεων, αναφέρεται σε μια κατάσταση στην οποία η κατανομή των κλάσεων στο σύνολο δεδομένων εκπαίδευσης δεν είναι ομοιόμορφη. Με άλλα λόγια,

ορισμένες κλάσεις έχουν σημαντικά περισσότερα ή λιγότερα παραδείγματα από άλλες. Αυτή η κατάσταση δημιουργεί αρκετά προβλήματα κατά την εκπαίδευση αλλά και κατά την μετέπειτα χρήση του μοντέλου. Επίσης, λόγω έλλειψης επαρκούς ποσότητας πόρων μνήμης, η εκπαίδευση ενός μοντέλου για όλα τα παραδείγματα του συνόλου δεδομένων και για όλες τις μοναδικές σχέσεις δεν ήταν εφικτή. Για αυτούς τους λόγους, έγιναν κάποια βήματα για την κατάλληλη προεπεξεργασία του συνόλου δεδομένων:

1. Φιλτράρισμα παραδειγμάτων με πάρα πολλά tokens: Με τη χρήση της μεθόδου filter, αφαιρούνται παραδείγματα όπου ο αριθμός των token στη στήλη "text", υπερβαίνει το καθορισμένο μέγιστο των 200 tokens. Σκοπός αυτού του βήματος, είναι να διασφαλιστεί ότι τα παραδείγματα στο σύνολο δεδομένων δεν έχουν υπερβολικό αριθμό token. Το DistilBERT, όπως και άλλα μοντέλα transformer, έχει ένα μέγιστο όριο token για ακολουθίες εισόδου. Εάν μια ακολουθία υπερβαίνει αυτό το όριο, πρέπει να περικοπεί ή να παραλειφθεί.
2. Φιλτράρισμα συνόλου δεδομένων με βάση τις 250 κορυφαίες σχέσεις: Πρώτα γίνεται καταμέτρηση σχετικά με το πόσα παραδείγματα αντιστοιχίζονται σε κάθε σχέση, μέσω της στήλης ‘relation’. Έπειτα, γίνεται μία κατανομή των σχέσεων με βάση την ποσότητα των παραδειγμάτων που αντιστοιχίζονται σε αυτή, σε φθίνουσα σειρά. Τέλος, γίνεται ένα φιλτράρισμα στο σύνολο δεδομένων όπου διατηρούνται μόνο τα παραδείγματα τα οποία αντιστοιχίζονται σε μία σχέση η οποία ανήκει στις 250 κορυφαίες σχέσεις. Επιλέγοντας τις κορυφαίες 250 σχέσεις με βάση τη συχνότητα, διασφαλίζεται ότι το μοντέλο εκτίθεται σε μια αντιπροσωπευτική κατανομή σχέσεων, βελτιώνοντας τη γενίκευση του μοντέλου. Το φιλτράρισμα του συνόλου δεδομένων με αυτόν τον τρόπο μπορεί να βοηθήσει στην αντιμετώπιση ζητημάτων που σχετίζονται με την ανισορροπία κλάσεων, ωθώντας το μοντέλο στο να επικεντρωθεί στις πιο σχετικές σχέσεις και να μειώσει τη συνολική πολυπλοκότητα του έργου αποκλείοντας λιγότερο συχνές σχέσεις.
3. Αφαίρεση παραδειγμάτων όπου και οι 2 οντότητες (head, tail) δεν περιέχονται στο κείμενο της στήλης text: Σκοπός αυτού του βήματος είναι να διασφαλιστεί ότι οι οντότητες head και tail που αναφέρονται σε κάθε παράδειγμα είναι πράγματι παρούσες στο κείμενο αυτού του παραδείγματος.
4. Διαχωρισμός σε σύνολα εκπαίδευσης και επικύρωσης με στρωματοποιημένη δειγματοληψία: Πραγματοποιείται διαχωρισμός του συνόλου δεδομένων σε σύνολα εκπαίδευσης και επικύρωσης με αναλογία 80-20. Ο διαχωρισμός γίνεται με στρωματοποιημένη δειγματοληψία με βάση τη στήλη ‘relation’. Αυτό σημαίνει ότι η κατανομή των παραδειγμάτων που αντιστοιχίζονται σε κάθε σχέση, θα είναι παρόμοια με το αρχικό σύνολο δεδομένων, τόσο στο σύνολο εκπαίδευσης, όσο και στο σύνολο επικύρωσης, συμβάλλοντας στη διατήρηση μιας ισορροπημένης αναπαράστασης των σχέσεων σε κάθε σύνολο.
5. Δημιουργία υποσυνόλων εκπαίδευσης και επικύρωσης με στρωματοποιημένη δειγματοληψία: Το σύνολο εκπαίδευσης μειώνεται στο μέγεθος των 300.000 παραδειγμάτων, ενώ το μέγεθος του συνόλου επικύρωσης μειώνεται στα 60.000 παραδείγματα. Η μείωση γίνεται με στρωματοποιημένη δειγματοληψία βασισμένη στη στήλη ‘relation’. Με αυτό τον τρόπο διατηρείται μια αναλογική αναπαράσταση των σχέσεων στα μικρότερο σύνολα δεδομένων εκπαίδευσης και επικύρωσης.

Σε αυτό το σημείο τα δεδομένα του συνόλου δεδομένων είναι έτοιμα για την απαραίτητη προεπεξεργασία και προετοιμασία προτού την εκπαίδευση του μοντέλου.

4.2.3 Προεπεξεργασία δεδομένων

1. Προσθήκη entity markers στη στήλη 'text': Η στήλη 'head' και η στήλη 'tail' κάθε παραδείγματος αντιπροσωπεύει τις οντότητες που περιέχονται στο κείμενο της στήλης 'text', ανάμεσα στις οποίες υπάρχει η σχέση της στήλης 'relation'. Αφού βρεθούν οι οντότητες head και tail μέσα στο text, γίνεται η αντικατάσταση τους με τις οντότητες head και tail επικαλυμμένες με ειδικά markers ([HEAD]'head'[/HEAD], [TAIL]'tail'[/TAIL]) που επιδεικνύουν τη θέση τους μέσα στο κείμενο.

Παράδειγμα πριν την προσθήκη entity markers:
{'relation': 'P19', 'head': 'Masato Yoshihara', 'tail': 'Fukuoka', 'title': 'Masato Yoshihara', 'text': 'Masato Yoshihara (吉原 正人 Yoshihara Masato, born October 27, 1991 in Nishi-ku, Fukuoka) is a Japanese football player who is currently playing for Cambodia Tiger in Cambodian League. A tall but technical striker, he has played for Japan at U-15 level before being called into the Avispa Fukuoka first team at age 18. Has experienced European football during a short study period at FC Girondins de Bordeaux, he is considered good with his feet for someone his size.'}

Παράδειγμα μετά την προσθήκη entity markers:
{'relation': 'P19', 'head': 'Masato Yoshihara', 'tail': 'Fukuoka', 'title': 'Masato Yoshihara', 'text': ' [HEAD]Masato Yoshihara[/HEAD] (吉原 正人 Yoshihara Masato, born October 27, 1991 in Nishi-ku, [TAIL]Fukuoka[/TAIL]) is a Japanese football player who is currently playing for Cambodia Tiger in Cambodian League. A tall but technical striker, he has played for Japan at U-15 level before being called into the Avispa Fukuoka first team at age 18. Has experienced European football during a short study period at FC Girondins de Bordeaux, he is considered good with his feet for someone his size.'}

2. Κωδικοποίηση ετικετών (label encoding): Με τη χρήση του LabelEncoder από τη βιβλιοθήκη scikit-learn γίνεται η μετατροπή όλων των ετικετών σχέσης της στήλης 'relation' σε αριθμητικές αναπαραστάσεις ούτως ώστε να μπορούν να χρησιμοποιηθούν ως μεταβλητές-στόχοι κατά τη διάρκεια της εκπαίδευσης του μοντέλου. Ο LabelEncoder εκχωρεί έναν μοναδικό ακέραιο αριθμό σε κάθε ξεχωριστή ετικέτα. Αυτό είναι απαραίτητο κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης, επειδή οι περισσότεροι αλγόριθμοι απαιτούν αριθμητικές εισόδους για μεταβλητές-στόχους.

3. Tokenization: Για το tokenization χρησιμοποιείται ο AutoTokenizer του DistilBERT. Για κάθε παράδειγμα των συνόλων εκπαίδευσης και επικύρωσης εξάγεται το κείμενο της στήλης 'text' και η ετικέτα της στήλης 'relation'. Ο tokenizer εκτελεί tokenization για το κείμενο της στήλης 'text' και χρησιμοποιείται truncation και padding με το μέγιστο μέγεθος token να είναι το 220. Δημιουργούνται τα input_ids και το attention_mask για κάθε παράδειγμα. Η ετικέτα του κάθε παραδείγματος κωδικοποιείται χρησιμοποιώντας τον LabelEncoder και η αριθμητική τιμή της εκχωρείται στο πεδίο 'labels' του κάθε παραδείγματος.

4.2.4 Διαμόρφωση μοντέλου

1. Φόρτωση μοντέλου: Πραγματοποιείται η φόρτωση και η αρχικοποίηση ενός μοντέλου DistilBERT το οποίο είναι fine-tuned στην ταξινόμηση ακολουθιών token. Αυτό συμβαίνει αρχικοποιώντας ένα αντικείμενο της κλάσης TFAutoModelForSequenceClassification. Αυτή η κλάση είναι μία κλάση της βιβλιοθήκης Hugging Face Transformers σχεδιασμένη στο Tensorflow και ρυθμισμένη για εργασίες ταξινόμησης ακολουθιών token. Κληρονομεί την αρχιτεκτονική του μοντέλου BERT που επιλέγεται και επιπλέον έχει ένα στρώμα το οποίο είναι μια κεφαλή ταξινόμησης που αντιστοιχίζει τις ενσωματώσεις εξόδου του transformer στον καθορισμένο αριθμό ετικετών εξόδου. Για την παραγωγή πιθανοτήτων για κάθε κλάση χρησιμοποιείται η συνάρτηση ενεργοποίησης softmax. Ως παραμέτρους δέχεται το όνομα ή το αναγνωριστικό του προ-εκπαιδευμένου μοντέλου BERT που θα φορτωθεί. Στην περίπτωση μας το μοντέλο είναι το distilbert-base-cased. Επίσης, δέχεται τον αριθμό των ετικετών που θα προβλέπει το μοντέλο (στην περίπτωση μας 250).
2. Προετοιμασία συνόλων δεδομένων εκπαίδευσης και επικύρωσης:
Πραγματοποιείται προετοιμασία των συνόλων δεδομένων TensorFlow για εκπαίδευση (train_set) και επικύρωση (validation_set). Τα σύνολα αυτά περιέχουν τα tokenized δεδομένα, στα οποία εφαρμόζεται ανακάτεμα για το σύνολο εκπαίδευσης, ορίζεται το μέγεθος παρτίδας (batch_size) και χρησιμοποιείται ένα εργαλείο συλλογής δεδομένων (data_collator) για να δημιουργηθούν παρτίδες κατάλληλες για εκπαίδευση και επικύρωση.

4.2.5 Μεταγλώττιση και εκπαίδευση μοντέλου

1. Callbacks: Δημιουργείται το PushToHubCallback το οποίο χρησιμοποιείται για την μεταφόρτωση του εκπαιδευμένου μοντέλου και των σχετικών αρχείων στο προσωπικό Hugging Face Model Hub έτσι ώστε να είναι ευκολότερη η μετέπειτα χρήση του μοντέλου για καινούρια παραδείγματα.
2. Μεταγλώττιση και εκπαίδευση: Κατά τη μεταγλώττιση ορίζεται ο Adam ως βελτιστοποιητής και η τιμή 3e-5 για το ρυθμό εκμάθησης. Στη συνέχεια ξεκινάει η εκπαίδευση του μοντέλου, αφού οριστούν τα σύνολα εκπαίδευσης και επικύρωσης, ο αριθμός των εποχών(5), το batch_size(128) και η συνάρτηση callback που δημιουργήθηκε προηγουμένως.

4.2.6 Αξιολόγηση μοντέλου

Η αξιολόγηση του μοντέλου REBERT γίνεται με τη χρήση της βιβλιοθήκης scikit-learn για τις μετρικές precision, recall και F1 score. Η αξιολόγηση γίνεται για weighted average και για macro average.
Η αξιολόγηση weighted average λαμβάνει υπόψη τη συμβολή κάθε κλάσης με βάση τη συχνότητά της στο σύνολο δεδομένων. Οι κλάσεις που έχουν μεγαλύτερη συχνότητα έχουν μεγαλύτερο αντίκτυπο στο μέσο όρο. Στο μοντέλο μας οι κλάσεις είναι οι σχέσεις της στήλης 'relation' .

Η αξιολόγηση macro average αντιμετωπίζει όλες τις κλάσεις ισότιμα, ανεξάρτητα από τη συχνότητά τους στο σύνολο δεδομένων. Η μετρική υπολογίζεται ανεξάρτητα για κάθε κλάση και στη συνέχεια, λαμβάνεται ο μέσος όρος. Η αξιολόγηση macro average παρέχει μια ισορροπημένη άποψη, καθώς κάθε τάξη συμβάλλει εξίσου στον μέσο όρο.
Οι τύποι για τον υπολογισμό των F1weighted και F1macro είναι:

$$F1\,weighted = \sum_{i=1}^{N} w_i \cdot F1_i$$

όπου N είναι ο αριθμός των κλάσεων, w_i είναι ο αριθμός των σωστών προβλέψεων για κάθε κλάση και $F1_i$ είναι το F1 score για κάθε ξεχωριστή κλάση.

$$F1\,macro = \frac{1}{N} \sum_{i=1}^{N} \cdot F1_i$$

όπου N είναι ο αριθμός των κλάσεων και $F1_i$ είναι το F1-score για κάθε ξεχωριστή κλάση.

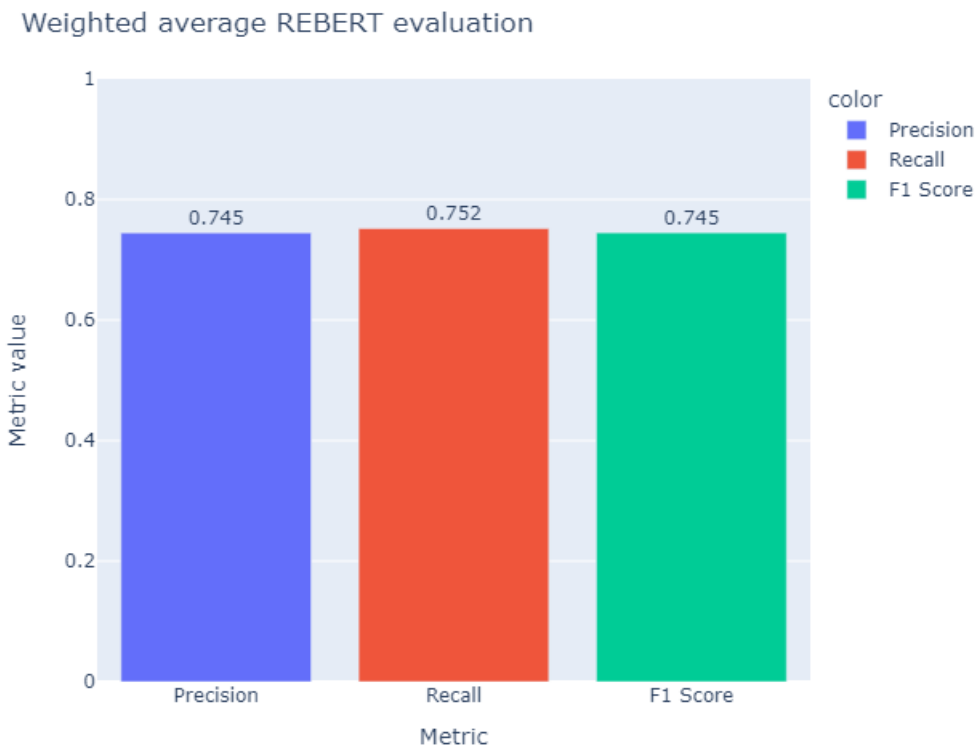
Τα αποτελέσματα που υπολογίστηκαν για την κάθε περίπτωση είναι τα εξής:

Μετρική	**Τιμή**
Precision	0,7445
Recall	0,7519
F1 score	0,7450

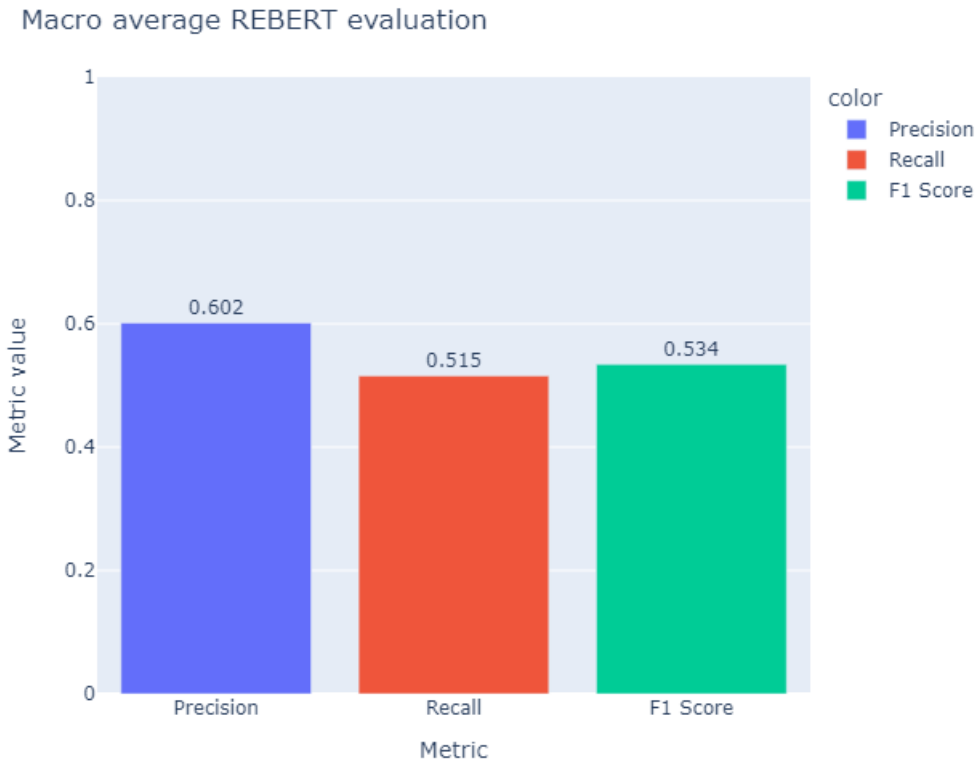
Πίνακας 4.2: Μετρικές αξιολόγησης του μοντέλου REBERT για weighted average

Μετρική	**Τιμή**
Precision	0,6016
Recall	0,5153
F1 score	0,5341

Πίνακας 4.3: Μετρικές αξιολόγησης του μοντέλου REBERT για macro average



Σχήμα 4.2: Μετρικές αξιολόγησης του μοντέλου REBERT για weighted average



Σχήμα 4.3: Μετρικές αξιολόγησης του μοντέλου REBERT για macro average

Η μετρική στην οποία πρέπει να δοθεί μεγαλύτερη έμφαση είναι το F1 score σε macro average λόγω της ανισορροπίας κλάσεων που παρατηρείται στο σύνολο δεδομένων.

4.3 Εξαγωγή πληροφοριών με σωληνωτή προσέγγιση

Τα μοντέλα που δημιουργήθηκαν, χρησιμοποιούνται διαδοχικά, ακολουθώντας την σωληνωτή προσέγγιση, για την επιτυχή εξαγωγή πληροφοριών από ένα κείμενο και την μετατροπή του κειμένου σε ένα σύνολο από σημασιολογικές τριπλέτες. Για τη βελτίωση της ακρίβειας και της αξιοπιστίας της εξαγωγής πληροφοριών, προστίθεται η επίλυση συναναφορών, ως ένα επιπλέον στάδιο του pipeline της σωληνωτής προσέγγισης. Η εξαγωγή πληροφοριών πραγματοποιείται σε επίπεδο πρότασης κειμένου, αφού μετά από εξαγωγή πληροφοριών σε επίπεδο πρότασης, παραγράφου και ολόκληρου κειμένου, παρατηρήθηκε μετά από ανθρώπινη αξιολόγηση, πως τα καλύτερα αποτελέσματα επιτυγχάνονται όταν η εξαγωγή πληροφοριών υλοποιείται σε επίπεδο πρότασης.

4.3.1 Επίλυση συναναφορών

Το κείμενο προς εξαγωγή πληροφοριών υποβάλλεται στη διαδικασία της επίλυσης συναναφορών. Αυτή επιτυγχάνεται χρησιμοποιώντας το API του πακέτου fastcoref σε συνδυασμό με τη βιβλιοθήκη spaCy. Το fastcoref [36] είναι ένα πακέτο python για γρήγορη, ακριβή και εύχρηστη επίλυση συναναφορών για αγγλικά κείμενα. Το πακέτο παρέχει ένα απλό API που καθιστά την πρόβλεψη οντοτήτων συναναφορών απλή και εύκολη στη χρήση. Το πακέτο υποστηρίζει οποιοδήποτε μήκος κειμένου ως είσοδο και εκτελεί αποτελεσματική ομαδοποίηση.Το πακέτο fastcoref παρέχεται ως στοιχείο της βιβλιοθήκης spaCy.

Για να εκτελεστεί η επίλυση συναναφορών αρχικά φορτώνεται ένα αγγλικό μοντέλο spaCy . Στη συνέχεια το στοιχείο fastcoref προστίθεται στο pipeline του μοντέλου spaCy. Το φορτωμένο μοντέλο spaCy εφαρμόζεται στο κείμενο εισαγωγής για τη δημιουργία ενός spaCy document (doc). Η παράμετρος component_cfg χρησιμοποιείται για τη διαμόρφωση του στοιχείου fastcoref. Σε αυτήν την περίπτωση, έχει οριστεί σε {'resolve_text': True}, πράγμα που σημαίνει ότι πρέπει να ληφθεί το επιλυμένο κείμενο. Με τον ορισμό της παραμέτρου {'resolve_text': True} επιτυγχάνεται η αποσαφήνιση και ενοποίηση των συναναφορών στο κείμενο αντικαθιστώντας τις αντωνυμίες και άλλες αναφορές με τις πραγματικές οντότητες στις οποίες αναφέρονται. Το κείμενο είναι πλέον επιλυμένο ως προς τις συναναφορές.

4.3.2 Εξαγωγή πληροφοριών σε επίπεδο πρότασης

1. Φόρτωση των μοντέλων NERBERT και REBERT: Φορτώνονται και αρχικοποιούνται τα εκπαιδευμένα μοντέλα NERBERT και REBERT, καθώς και ο tokenizer του REBERT, για να εκτελέσουν την αναγνώριση ονοματικών οντοτήτων και την ταξινόμηση των σχέσεων μεταξύ των οντοτήτων του κειμένου.

2. Διαχωρισμός κειμένου σε προτάσεις: Το επιλυμένο κείμενο διαχωρίζεται σε προτάσεις μέσω του spaCy document έτσι ώστε να γίνει η επεξεργασία της κάθε πρότασης του κειμένου.
3. Αναγνώριση ονοματικών οντοτήτων και δημιουργία ζευγών οντοτήτων για κάθε πρόταση: Για κάθε πρόταση εκτελείται η εύρεση και αναγνώριση των ονοματικών οντοτήτων μέσω του μοντέλου NERBERT. Για μεγαλύτερη αξιοπιστία διατηρούνται μόνο οι οντότητες που έχουν προβλεφθεί από το μοντέλο με confidence score >=0,8. Στη συνέχεια, μέσω ενός συνόλου γίνεται έλεγχος του κειμένου της κάθε οντότητας και αποθηκεύονται μόνο οι μοναδικές οντότητες κάθε πρότασης ούτως ώστε να αποφευχθούν οι διπλότυπες εμφανίσεις οντοτήτων. Έπειτα, δημιουργούνται όλα τα πιθανά ζεύγη οντοτήτων για κάθε πρόταση.
4. Εξαγωγή σχέσεων και δημιουργία τριπλετών: Για κάθε μοναδικό ζεύγος οντοτήτων, η πρόταση κειμένου τροποποιείται μέσω της προσθήκης entity markers για τις οντότητες του ζεύγους. Η πρώτη οντότητα του ζεύγους οντοτήτων επικαλύπτεται με head markers και η δεύτερη οντότητα του ζεύγους επικαλύπτεται με tail markers. Η τροποποιημένη πρόταση υποβάλλεται σε tokenization μέσω του tokenizer του REBERT. Στη συνέχεια το φορτωμένο μοντέλο REBERT χρησιμοποιείται για την πρόβλεψη πιθανής ετικέτας σχέσης για την τροποποιημένη πρόταση. Εάν το confidence score για την πρόβλεψη της ετικέτας σχέσης είναι μεγαλύτερο ή ίσο του 0,5, τότε η πρόβλεψη του μοντέλου θεωρείται αξιόπιστη και η ετικέτα σχέσης γίνεται δεκτή. Σε αυτή την περίπτωση, δημιουργείται η σημασιολογική τριπλέτα η οποία περιέχει τις οντότητες και τη μεταξύ τους σχέση και αποθηκεύεται σε μορφή tuple. Αυτή η τριπλέτα έχει τη μορφή: (head entity, relation, tail entity). Το tuple, αποτελούμενο από τα τρία στοιχεία, προστίθεται εντός μίας λίστας που περιέχει όλες τις τριπλέτες. Το τελικό βήμα είναι το φιλτράρισμα των πιθανών διπλότυπων εμφανίσεων τριπλετών εντός της λίστας των τριπλετών. Αφού ολοκληρωθεί και αυτό το βήμα, οι τριπλέτες αποθηκεύονται και αποστέλλονται στο δέκτη.

Κεφάλαιο 5

Παραγωγή κειμένου από σημασιολογικές τριπλέτες

Μετά την αποστολή των σημασιολογικών τριπλετών από τον πομπό στο δέκτη, σκοπός του δέκτη είναι η ανακατασκευή των σημασιολογικών τριπλετών σε μορφή κειμένου. Ο δέκτης πρέπει να αποδώσει ερμηνεία στις σημασιολογικές τριπλέτες που έχει λάβει και να δημιουργήσει ένα συνεκτικό κείμενο συναφές με το αρχικό κείμενο του πομπού.

Η δημιουργία ενός μοντέλου παραγωγής κειμένου από σημασιολογικές τριπλέτες μέσω της εκπαίδευσης του μοντέλου σε ένα επισημασμένο σύνολο δεδομένων δεν ήταν εφικτή, καθώς δε βρέθηκε κάποιο σύνολο δεδομένων που να μπορεί να καλύψει όλες τις πιθανές σχέσεις που μπορεί να προκύψουν από το μοντέλο ταξινόμησης σχέσεων. Επίσης, σε μερικά σύνολα δεδομένων που εξετάστηκαν και περιείχαν αρκετές από τις σχέσεις που εμφανίζονται και στο σύνολο δεδομένων της ταξινόμησης σχέσεων, παρατηρήθηκε ότι οι αντιστοιχίσεις μεταξύ τριπλετών και κειμένων δεν ήταν ακριβείς, υπό την έννοια ότι το αντιστοιχισμένο κείμενο δεν κάλυπτε με ικανοποιητικό τρόπο το περιεχόμενο των τριπλετών.

Το τελευταίο διάστημα, αρκετές δημοσιεύσεις εξετάζουν τη δυνατότητα χρήσης μεγάλων γλωσσικών μοντέλων για την παραγωγή κειμένου από δεδομένα, μέσω του prompt enginneering αντί του fine-tuning σε προεκπαιδευμένα μοντέλα [37], [38], [39]. Τα ικανοποιητικά αποτελέσματα αυτών των δημοσιεύσεων μας οδήγησαν στη χρήση ενός μεγάλου γλωσσικού μοντέλου για την παραγωγή κειμένου από σημασιολογικές τριπλέτες. Συγκεκριμένα, το μοντέλο που χρησιμοποιείται είναι το flan-t5 [40] και μέσω της καθοδήγησης του χρησιμοποιώντας την κατάλληλη προτροπή, επιτυγχάνεται η εργασία της μετατροπής των σημασιολογικών τριπλετών σε κείμενο.

Το Flan-T5 είναι ένα εμπορικά διαθέσιμο LLM ανοιχτού κώδικα που δημιουργήθηκε από ερευνητές της Google και παρουσιάστηκε στη δημοσίευση ”Scaling Instruction-Finetuned Language Models” από τους Chung et al. [41]. Το flan-t5 έχει τελειοποιηθεί χρησιμοποιώντας τεχνικές fine-tuning που βασίζονται σε οδηγίες, για να βελτιώσει την απόδοσή του σε διάφορες εργασίες NLP. Το flan-t5 βασίζεται στην αρχιτεκτονική T5 και έχει σχεδιαστεί για να υπερέχει σε σενάρια zero-shot ή few-shot. Σε ένα σενάριο zero-shot, το μοντέλο καλείται να εκτελέσει μία διαδικασία χωρίς να του παρέχονται επισημασμένα παραδείγματα ή συγκεκριμένα δεδομένα εκπαίδευσης κατά την εκπαίδευση. Στο σενάριο few-shot, το μοντέλο παρέχεται μόνο με ένα μικρό αριθμό επισημασμένων παραδειγμάτων για μια συγκεκριμένη εργασία κατά τη διάρκεια της εκπαίδευσης. Το μοντέλο flan-t5 διατίθεται σε διαφορετικά μεγέθη, που κυμαίνονται από small έως xxl, το καθένα με διαφορετικό αριθμό παραμέτρων για να καλύψει διαφορετικές υπολογιστικές απαιτήσεις και ανάγκες απόδοσης.

5.1 Χρήση του μοντέλου flan-t5

Για να χρησιμοποιηθεί το flan-t5 για την εργασία της παραγωγής κειμένου από σημασιολογικές τριπλέτες, ακολουθούνται τα παρακάτω βήματα:

1. Φόρτωση μοντέλου: Πραγματοποιείται η φόρτωση και αρχικοποίηση του μοντέλου flan-t5-base και του tokenizer του μοντέλου. Αυτό συμβαίνει αρχικοποιώντας ένα αντικείμενο της κλάσης T5ForConditionalGeneration. Αυτή η κλάση είναι μία κλάση της βιβλιοθήκης Hugging Face Transformers, προσαρμοσμένη για μοντέλα T5. Η κλάση T5ForConditionalGeneration χρησιμοποιείται συνήθως με προ-εκπαιδευμένα βάρη, που ορίζονται μέσω του προ-εκπαιδευμένου μοντέλου T5 που θα φορτωθεί.
2. Ενημέρωση τριπλετών: Οι σημασιολογικές τριπλέτες βρίσκονται στη μορφή: (head, relation, tail), όπου στην πλειοψηφία των περιπτώσεων, η σχέση relation βρίσκεται στην κωδικοποιημένη μορφή pid.

Παράδειγμα τριπλέτας:
(Nikolaos Lefkos, P69, University of Western Macedonia)

Στο παραπάνω παράδειγμα το “P69” αντιστοιχεί στη σχέση “educated at”. Η ενημέρωση τριπλετών αφορά την αντικατάσταση των τιμών pid με τα αντίστοιχα ονόματά τους. Αυτό επιτυγχάνεται μέσω ενός json αρχείου (pid2name.json), το οποίο περιέχει τις αντιστοιχίσεις όλων των τιμών pid με τα αντίστοιχα ονόματα σχέσεων.
3. Prompt engineering: Η μεθοδολογία του prompt enginnering αναφέρεται στο στάδιο της δοκιμής διάφορων προτροπών ως είσοδο στο μοντέλο. Ανάμεσα σε πολλές προτροπές που ελέγχθηκαν, επιλέγεται αυτή που θεωρείται πιο κατάλληλη και παρουσιάζει τα καλύτερα αποτελέσματα στη διαδικασία παραγωγής κειμένου από σημασιολογικές τριπλέτες. Συγκεκριμένα, η προτροπή που χρησιμοποιείται είναι η εξής:

```
prompt = f"translate the following triples into text. Triples:{updated_triples}"
```

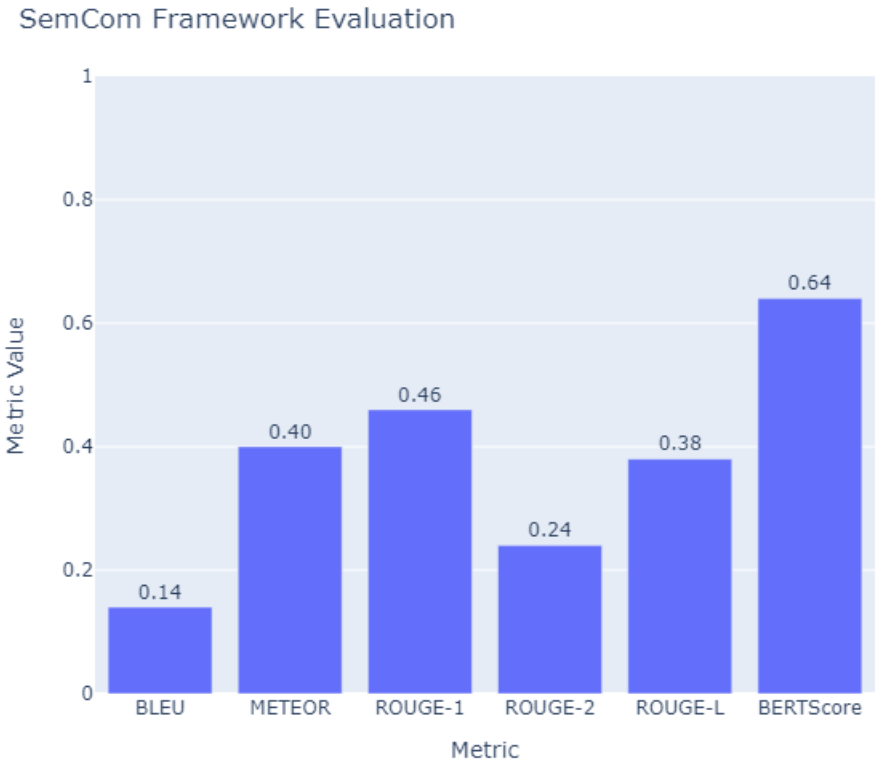
Με αυτήν την οδηγία, το μοντέλο ερμηνεύει το πρόβλημα ως μία εργασία μετάφρασης. Η μεταβλητή updated_triples περιλαμβάνει τη λίστα των ενημερωμένων σημασιολογικών τριπλετών που πρέπει να μεταφραστούν σε κείμενο.
4. Εφαρμογή του μοντέλου: Το τελικό βήμα είναι η χρήση του μοντέλου T5 για τη δημιουργία κειμένου με βάση την παρεχόμενη προτροπή. Αρχικά, το κείμενο της μεταβλητής prompt μαζί με τις ενημερωμένες τριπλέτες υποβάλλεται σε tokenization μέσω του T5Tokenizer και μετατρέπεται στη μορφή input_ids, η οποία είναι η μορφή που μπορεί να κατανοήσει το μοντέλο. Στη συνέχεια, το μοντέλο χρησιμοποιείται για τη δημιουργία κειμένου με βάση τα tokenized input_ids. Στο επόμενο βήμα, μέσω του T5Tokenizer, η παραγόμενη έξοδος του μοντέλου αποκωδικοποιείται σε ένα αναγνώσιμο κείμενο. Τέλος, το δημιουργημένο κείμενο εκτυπώνεται, δίνοντας τη δυνατότητα για ανάγνωση και αξιολόγηση.

5.2 Αξιολόγηση πλαισίου σημασιολογικών επικοινωνιών

Η αξιολόγηση του πλαισίου πραγματοποιήθηκε χρησιμοποιώντας τις μετρικές: BLEU, METEOR, ROUGE-1, ROUGE-2, ROUGE-L και BERTScore. Σε αυτή τη διαδικασία αξιολογείται η ποιότητα του κειμένου που έχει δημιουργήσει ο παραλήπτης με βάση τις σημασιολογικές τριπλέτες που έχει λάβει από τον αποστολέα. Η αξιολόγηση έγινε για 500 παραδείγματα κειμένου που αφορούν την πρώτη παράγραφο άρθρων της Wikipedia. Κατά την αξιολόγηση έγινε η σύγκριση του πρωταρχικού κειμένου, πριν ο αποστολέας το μετατρέψει σε σημασιολογικές τριπλέτες, με το τελικό κείμενο που δημιουργήθηκε από την ανακατασκευή των τριπλετών σε κείμενο. Συγκρίνοντας τα δύο κείμενα μεταξύ τους για το σύνολο των μετρικών, μπορούμε να εξάγουμε συμπεράσματα σχετικά με την ποιότητα του τελικού κειμένου και να συγκρίνουμε την ομοιότητά του με το αρχικό κείμενο. Τα αποτελέσματα της αξιολόγησης έπειτα από τον υπολογισμό του μέσου όρου για τα 500 παραδείγματα που εξετάστηκαν είναι τα εξής:

Μετρική	**Τιμή**
BLEU	0,14
METEOR	0,40
ROUGE-1	0,46
ROUGE-2	0,24
ROUGE-L	0,38
BERTScore	0,64

Πίνακας 5.1: Μετρικές αξιολόγησης πλαισίου σημασιολογικών επικοινωνιών



Σχήμα 5.1: Μετρικές αξιολόγησης πλαισίου σημασιολογικών επικοινωνιών

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία υλοποιήθηκε ένα διατερματικό πλαίσιο σημασιολογικών επικοινωνιών για δεδομένα κειμένου. Στο εν λόγω σύστημα, ο αποστολέας καταφέρνει να απομονώσει τη σημασιολογική πληροφορία ενός κειμένου και να μεταδώσει το νόημα του κειμένου αντί του ολόκληρου αρχικού μηνύματος. Η σημασιολογική πληροφορία του κειμένου κωδικοποιείται ως μία λίστα αποτελούμενη από σημασιολογικές τριπλέτες σε μορφή tuple. Ο παραλήπτης αναλαμβάνει την επαναφορά της σημασιολογικής πληροφορίας σε κανονικό κείμενο. Προκειμένου να επιτευχθεί η εξαγωγή της σημασιολογικής πληροφορίας, ο αποστολέας χρησιμοποιεί, ακολουθώντας τη σωληνωτή προσέγγιση, ένα μοντέλο αναγνώρισης ονοματικών οντοτήτων και στη συνέχεια, για τις αναγνωρισμένες οντότητες, χρησιμοποιεί ένα μοντέλο ταξινόμησης σχέσεων. Για το σκοπό αυτό, πραγματοποιήθηκε η δημιουργία αυτών των μοντέλων μέσω του fine-tuning προεκπαιδευμένων μοντέλων, βασισμένων στην αρχιτεκτονική transformer, σε επισημασμένα σύνολα δεδομένων.

Το μοντέλο αναγνώρισης ονοματικών οντοτήτων εκπαιδεύτηκε σε ένα επισημασμένο σύνολο δεδομένων που καλύπτει ένα ευρύ φάσμα κατηγοριών οντοτήτων, επιτρέποντας τη χρήση του για ποικίλα είδη κειμένου. Το μοντέλο ταξινόμησης σχέσεων εκπαιδεύτηκε σε ένα επισημασμένο σύνολο δεδομένων που καλύπτει σχέσεις μεταξύ οντοτήτων σε κείμενα, τα οποία είναι άρθρα της Wikipedia, επιτρέποντας τη χρήση του για κείμενα με αντίστοιχη δομή. Συνεπώς, μετά από αξιολόγηση του συνολικού πλαισίου, συμπεραίνεται ότι η εξαγωγή σημασιολογικών πληροφοριών είναι πιο αποτελεσματική όταν πραγματοποιείται σε κείμενα που αποτελούν άρθρα του Wikipedia ή έχουν παρόμοια δομή.

Όσον αφορά τον παραλήπτη, η μετατροπή της σημασιολογικής πληροφορίας σε κανονικό κείμενο γίνεται καθοδηγώντας το μεγάλο γλωσσικό μοντέλο Flan-T5. Αξιοποιώντας το γνωσιακό υπόβαθρο του μοντέλου σε εργασίες μετάφρασης, το μοντέλο παροτρύνεται να αντιμετωπίσει αυτήν την εργασία ως μια εργασία μετάφρασης της λίστας σημασιολογικών τριπλετών σε κείμενο. Μετά από αξιολόγηση και χρήση του πλαισίου, διαπιστώνεται ότι το πλαίσιο μπορεί να χρησιμοποιηθεί με αποτελεσματικότητα για την εξαγωγή της σημαντικής πληροφορίας ενός κειμένου, τη μεταφορά της από τον πομπό στο δέκτη σε συμπυκνωμένη μορφή και την ανάκτηση της σε μορφή κειμένου που διατηρεί παρόμοιο νόημα με το αρχικό κείμενο.

Η διαδικασία του fine-tuning σε προεκπαιδευμένα μοντέλα transformer επιτρέπει τον αποτελεσματικό συνδυασμό της γενικής γνώσης των μοντέλων αυτών με την προσαρμογή σε ειδικά πεδία, όπως στην περίπτωση των άρθρων του wikipedia. Επιπλέον η χρήση του μεγάλου γλωσσικού μοντέλου flan-t5 για τη δημιουργία

συνεκτικού κειμένου από σημασιολογικές τριπλέτες αξιοποιεί το γνωσιακό υπόβαθρο του συγκεκριμένου μοντέλου και την ικανότητα του να κατανοήσει τη σημασιολογία της πληροφορίας των σημασιολογικών τριπλετών, χωρίς την ανάγκη εκπαίδευσης ενός μοντέλου σε ένα ειδικό σύνολο δεδομένων.

6.2 Μελλοντικές επεκτάσεις

Λόγω του περιορισμένου αριθμού των token εισόδου που μπορεί να δεχθεί σαν προτροπή το flan-t5, η χρήση του πλαισίου είναι εφικτή μόνο για μικρά τμήματα κειμένου, π.χ. μία παράγραφος, αφού όσο αυξάνεται το μέγεθος ενός κειμένου αυξάνεται και ο αριθμός των σημασιολογικών τριπλετών που πρέπει να υποδειχθούν στο μοντέλο εντός της προτροπής. Οπότε, μία πιθανή επέκταση του πλαισίου, είναι η χρήση ενός διαφορετικού μεγάλου γλωσσικού μοντέλου από τον αποστολέα για την παραγωγή κειμένου από σημασιολογικές τριπλέτες, το οποίο θα μπορεί να δεχθεί μεγαλύτερο αριθμό tokens στην ακολουθία εισόδου. Αυτό θα επιτρέψει την αποστολή και λήψη μεγαλύτερων κειμένων.

Το τελευταίο διάστημα, αρκετές δημοσιεύσεις εξετάζουν το ενδεχόμενο της χρήσης μεγάλων γλωσσικών μοντέλων για το σκοπό της εξαγωγής πληροφοριών από δεδομένα κειμένου και το σχηματισμό σημασιολογικών τριπλετών, μέσω της χρήσης των κατάλληλων προτροπών-οδηγιών [42], [43], [44], [45]. Ως εκ τούτου, μία πιθανή επέκταση είναι η χρήση μεγάλων γλωσσικών μοντέλων για την εξαγωγή πληροφοριών από τον αποστολέα. Αυτή η προσέγγιση φαίνεται ελκυστική για χρήση σε κείμενα γενικού σκοπού, σε περιπτώσεις που ένα πλαίσιο σημασιολογικών επικοινωνιών θέλει να επιτύχει εξαγωγή πληροφοριών ενώ δεν υπάρχει η δυνατότητα εκπαίδευσης μοντέλων σε συγκεκριμένα σύνολα δεδομένων και χωρίς την απαίτηση υψηλής απόδοσης σε συγκεκριμένα είδη κειμένου (π.χ. επιστημονικά άρθρα).

Το προεκπαιδευμένο μοντέλο στο οποίο εφαρμόστηκε fine-tuning σε συγκεκριμένα σύνολα δεδομένων για τις εργασίες της αναγνώρισης ονοματικών οντοτήτων και της ταξινόμησης σχέσεων είναι το DistilBERT. Αυτό το μοντέλο αποτελεί την αποσταγμένη έκδοση του BERT. Μία επέκταση είναι η χρήση μεγαλύτερων προεκπαιδευμένων μοντέλων BERT όπως τα BERTBASE, BERTLARGE, RoBERTa για to fine-tuning και την εκπαίδευση τους στα σύνολα δεδομένων. Το μέγεθος των παραμέτρων αυτών των μοντέλων δύναται να οδηγήσει σε καλύτερα αποτελέσματα ως προς την απόδοση της εξαγωγής πληροφοριών.

Λόγω έλλειψης πόρων, η εκπαίδευση του μοντέλου ταξινόμησης σχέσεων περιορίστηκε σε ένα μικρότερο τμήμα του συνόλου δεδομένων, για μικρότερα επισημασμένα κείμενα και λιγότερες κατηγορίες σχέσεων. Μια πιθανή επέκταση θα ήταν η εκπαίδευση του μοντέλου σε ολόκληρο το σύνολο δεδομένων, πράγμα που θα οδηγούσε σε καλύτερη απόδοση όσον αφορά την εξαγωγή πληροφοριών, καθώς το μοντέλο θα αποκτούσε περισσότερες αναπαραστάσεις των δεδομένων κειμένου και των κατηγοριών σχέσεων. Επιπλέον, οι τεχνικές που χρησιμοποιήθηκαν για τη δημιουργία του μοντέλου ταξινόμησης σχέσεων μπορούν να εφαρμοστούν για τη

δημιουργία μοντέλων σε άλλα επισημασμένα σύνολα δεδομένων, προκειμένου να καλύψουν την εξαγωγή πληροφοριών για ποικίλα είδη κειμένου.

Στο πλαίσιο που δημιουργήθηκε, η εξαγωγή πληροφοριών πραγματοποιείται σε επίπεδο πρότασης για όλα τα ζεύγη οντοτήτων. Μία πιθανή επέκταση είναι η παράλληλη επεξεργασία και εξαγωγή πληροφοριών των κειμένων, χρησιμοποιώντας τεχνικές παράλληλου προγραμματισμού, κάτι που θα οδηγήσει σε ταχύτερη και πιο αποδοτική ανάλυση των δεδομένων. Ο παράλληλος προγραμματισμός επιτρέπει στο σύστημα να διαχειρίζεται παράλληλα διάφορα ζεύγη οντοτήτων. Αυτό οδηγεί σε μείωση του χρόνου επεξεργασίας και αυξημένη αποτελεσματικότητα στην εξαγωγή πληροφοριών.

Βιβλιογραφία

[1] E. Calvanese Strinati and S. Barbarossa, “6G networks: Beyond Shannon towards semantic and goal-oriented communications,” *Computer Networks*, vol. 190, p. 107930, May 2021, doi: 10.1016/J.COMNET.2021.107930.

[2] Z. Qin, S. Member, X. Tao, J. Lu, W. Tong, and G. Ye Li, “Semantic Communications: Principles and Challenges,” Dec. 2021, [Online]. Available: https://arxiv.org/abs/2201.01389v5

[3] Y. Liu, X. Wang, Z. Ning, M. Zhou, L. Guo, and B. Jedari, “A survey on semantic communications: technologies, solutions, applications and challenges,” *Digital Communications and Networks*, Jun. 2023, doi: 10.1016/J.DCAN.2023.05.010.

[4] “Wikipedia, the free encyclopedia.”

[5] “Neural Networks: Architecture, Components, and Learning Process | LinkedIn.” [Online]. Available: https://www.linkedin.com/pulse/neural-networks-architecture-components-learning-fricke-mcs-uxc/

[6] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, “Deep Learning Enabled Semantic Communication Systems,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, Jun. 2020, doi: 10.1109/tsp.2021.3071210.

[7] Y. Wang, M. Chen, W. Saad, T. Luo, S. Cui, and H. V. Poor, “Performance Optimization for Semantic Communications: An Attention-based Learning Approach,” *2021 IEEE Global Communications Conference, GLOBECOM 2021 - Proceedings*, 2021, doi: 10.1109/GLOBECOM46510.2021.9685056.

[8] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, “A Unified Multi-Task Semantic Communication System for Multimodal Data,” Sep. 2022, [Online]. Available: https://arxiv.org/abs/2209.07689v2

[9] T. M. Getu, G. Kaddoum, and M. Bennis, “Making Sense of Meaning: A Survey on Metrics for Semantic and Goal-Oriented Communication,” *IEEE Access*, vol. 11, pp. 45456–45492, 2023, doi: 10.1109/ACCESS.2023.3269848.

[10] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, “Semantic Similarity Metrics for Evaluating Source Code Summarization,” *IEEE International Conference on Program Comprehension*, vol. 2022-March, pp. 36–47, Apr. 2022, doi: 10.1145/nnnnnnn.nnnnnnn.

[11] X. Zhao *et al.*, “A Comprehensive Survey on Deep Learning for Relation Extraction: Recent Advances and New Frontiers,” *ACM Comput Surv*, vol. 1, Jun. 2023, [Online]. Available: https://arxiv.org/abs/2306.02051v2

[12] T. Nayak, N. Majumder, P. Goyal, and S. Poria, “Deep Neural Approaches to Relation Triplets Extraction: A Comprehensive Survey,” *Cognit Comput*, vol. 13, no. 5, pp. 1215–1232, Mar. 2021, doi: 10.1007/s12559-021-09917-7.

[13] J. Li, A. Sun, J. Han, and C. Li, “A Survey on Deep Learning for Named Entity Recognition,” *IEEE Trans Knowl Data Eng*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[14] B. Jehangir, S. Radhakrishnan, and R. Agarwal, “A survey on Named Entity Recognition — datasets, tools, and methodologies,” *Natural Language Processing Journal*, vol. 3, p. 100017, Jun. 2023, doi: 10.1016/J.NLP.2023.100017.

[15] S. Khalil and A. E. Bozorgi, “A Survey on Recent Named Entity Recognition and Relation Classification Methods with Focus on Few-Shot Learning Approaches,” Oct. 2023, [Online]. Available: https://arxiv.org/abs/2310.19055v1

[16] H. Wang, K. Qin, R. Y. Zakari, G. Lu, and J. Yin, “Deep Neural Network Based Relation Extraction: An Overview,” *Neural Comput Appl*, vol. 34, no. 6, pp. 4781–4801, Jan. 2021, doi: 10.1007/s00521-021-06667-3.

[17] K. Lata, P. Singh, and K. Dutta, “Mention detection in coreference resolution: survey,” *Applied Intelligence*, vol. 52, no. 9, pp. 9816–9860, Jul. 2022, doi: 10.1007/S10489-021-02878-2/TABLES/17.

[18] R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, “Anaphora and coreference resolution: A review,” *Information Fusion*, vol. 59, pp. 139–162, Jul. 2020, doi: 10.1016/J.INFFUS.2020.01.010.

[19] Y. Lin, T. Ruan, J. Liu, and H. Wang, “A Survey on Neural Data-to-Text Generation,” *IEEE Trans Knowl Data Eng*, 2023, doi: 10.1109/TKDE.2023.3304385.

[20] H. Gao *et al.*, “Triples-to-Text Generation with Reinforcement Learning Based Graph-augmented Neural Networks,” Nov. 2021, [Online]. Available: https://arxiv.org/abs/2111.10545v3

[21] A. Vaswani et al., “Attention Is All You Need,” Adv Neural Inf Process Syst, vol. 2017-December, pp. 5999–6009, Jun. 2017, [Online]. Available: https://arxiv.org/abs/1706.03762v7

[22] “Transformer Architecture Explained | Medium | GoPenAI.” [Online]. Available: https://blog.gopenai.com/transformer-architecture-explained-dde38acf1d1

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, [Online]. Available: https://arxiv.org/abs/1810.04805v2

[24] “Explanation of BERT Model - NLP - GeeksforGeeks.” [Online]. Available: https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/

[25] U. Khalid, M. O. Beg, and M. U. Arshad, “RUBERT: A Bilingual Roman Urdu BERT Using Cross Lingual Transfer Learning,” Feb. 2021, Accessed: Feb. 01, 2024. [Online]. Available: https://www.researchgate.net/publication/349546860_RUBERT_A_Bilingual_Roman_Urdu_BERT_Using_Cross_Lingual_Transfer_Learning

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Oct. 2019, [Online]. Available: https://arxiv.org/abs/1910.01108v4

[27] H. Adel *et al.*, “Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm,” *Mathematics*, vol. 10, no. 3, Feb. 2022, doi: 10.3390/math10030447.

[28] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, vol. 21, pp. 1–67, Oct. 2019, [Online]. Available: https://arxiv.org/abs/1910.10683v4

[29] “What is the T5-Model? | Data Basecamp.” [Online]. Available: https://databasecamp.de/en/ml-blog/t5-model

[30] “OntoNotes Release 5.0 - Linguistic Data Consortium.” [Online]. Available: https://catalog.ldc.upenn.edu/LDC2013T19

[31] “SpeedOfMagic/ontonotes_english · Datasets at Hugging Face.” [Online]. Available: https://huggingface.co/datasets/SpeedOfMagic/ontonotes_english

[32] W. Zhou and M. Chen, “An Improved Baseline for Sentence-level Relation Extraction,” Feb. 2021, [Online]. Available: https://arxiv.org/abs/2102.01373v4

[33] S. Wu and Y. He, “Enriching Pre-trained Language Model with Entity Information for Relation Classification,” *International Conference on Information and Knowledge Management, Proceedings*, pp. 2361–2364, May 2019, doi: 10.1145/3357384.3358119.

[34] “T-REx : A Large Scale Alignment of Natural Language with Knowledge Base Triples.” [Online]. Available: https://hadyelsahar.github.io/t-rex/

[35] “relbert/t_rex · Datasets at Hugging Face.” [Online]. Available: https://huggingface.co/datasets/relbert/t_rex

[36] S. Otmazgin, A. Cattan, and Y. Goldberg, “F-coref: Fast, Accurate and Easy to Use Coreference Resolution,” Sep. 2022, [Online]. Available: https://arxiv.org/abs/2209.04280v4

[37] S. Yuan and M. Färber, “Evaluating Generative Models for Graph-to-Text Generation,” *International Conference Recent Advances in Natural Language Processing, RANLP*, pp. 1256–1264, Jul. 2023, doi: 10.26615/978-954-452-092-2_133.

[38] M. Lorandi and A. Belz, “Data-to-text Generation for Severely Under-Resourced Languages with GPT-3.5: A Bit of Help Needed from Google Translate (WebNLG 2023).” pp. 80–86, 2023. [Online]. Available: https://aclanthology.org/2023.mmnlg-1.9

[39] A. Axelsson and G. Skantze, “Using Large Language Models for Zero-Shot Natural Language Generation from Knowledge Graphs,” Jul. 2023, [Online]. Available: https://arxiv.org/abs/2307.07312v2

[40] “google/flan-t5-base · Hugging Face.” [Online]. Available: https://huggingface.co/google/flan-t5-base

[41] H. W. Chung *et al.*, “Scaling Instruction-Finetuned Language Models,” Oct. 2022, [Online]. Available: https://arxiv.org/abs/2210.11416v5

[42] X. Xu, Y. Zhu, X. Wang, and N. Zhang, “How to Unleash the Power of Large Language Models for Few-shot Relation Extraction?,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 190–200, May 2023, doi: 10.18653/v1/2023.sustainlp-1.13.

[43] X. Li, F. Polat, and P. Groth, “Do Instruction-tuned Large Language Models Help with Relation Extraction?,” 2023, [Online]. Available: http://ceur-ws.org

[44] C. Peng *et al.*, “Model Tuning or Prompt Tuning? A Study of Large Language Models for Clinical Concept and Relation Extraction,” Oct. 2023, [Online]. Available: https://arxiv.org/abs/2310.06239v1

[45] S. Wadhwa, S. Amir, and B. C. Wallace, “Revisiting Relation Extraction in the era of Large Language Models,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 15566–15589, May 2023, doi: 10.18653/v1/2023.acl-long.868.

[46] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” 8th International Conference on Learning Representations, ICLR 2020, Apr. 2019, [Online]. Available: https://arxiv.org/abs/1904.09675v3