ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΣΧΕΔΙΑΣΜΟΥ & ΑΝΑΠΤΥΞΗΣ ΣΥΣΤΗΜΑΤΩΝ
ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ

Γνωσιακές Μηχανές με Χρήση Μεθόδων Μηχανικής Μάθησης

(Cognitive Engines Using Machine Learning Methods)

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Δημήτριος Α. Παπαδόπουλος

Χανιά, 2022

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΣΧΕΔΙΑΣΜΟΥ & ΑΝΑΠΤΥΞΗΣ ΣΥΣΤΗΜΑΤΩΝ
ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ

Γνωσιακές Μηχανές με Χρήση Μεθόδων Μηχανικής Μάθησης

(Cognitive Engines Using Machine Learning Methods)

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

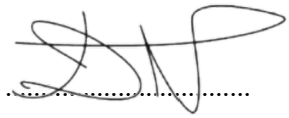
Δημήτριος Α. Παπαδόπουλος

Τριμελής Συμβουλευτική Επιτροπή : Νικόλαος Ματσατσίνης (Επιβλέπων)
Νικόλαος Παπαδάκης
Νικόλαος Δάρας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 22α Ιουλίου 2022.

Ματσατσίνης Νικόλαος Καθηγητής, Σχολή Μηχανικών Παραγωγής & Διοίκησης, Πολυτεχνείο Κρήτης	Παπαδάκης Νικόλαος Αναπληρωτής Καθηγητής, Στρατιωτική Σχολή Ευελπίδων	Δάρας Νικόλαος Καθηγητής, Στρατιωτική Σχολή Ευελπίδων
Βαρβαρίγου Θεοδώρα Καθηγήτρια, Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο	Τσαφαράκης Στέλιος Αναπληρωτής Καθηγητής, Σχολή Μηχανικών Παραγωγής & Διοίκησης, Πολυτεχνείο Κρήτης	Δουλάμης Αναστάσιος Αναπληρωτής Καθηγητής, Σχολή Αγρονόμων και Τοπογράφων Μηχανικών, Εθνικό Μετσόβιο Πολυτεχνείο
Κοσμόπουλος Δημήτριος Αναπληρωτής Καθηγητής, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Πατρών		

Χανιά, 2022



Δημήτριος Α. Παπαδόπουλος
Διπλωματούχος Μηχανολόγος Μηχανικός Ε.Μ.Π
Διδάκτωρ Πολυτεχνείου Κρήτης

Η παρούσα Διδακτορική Διατριβή υπεβλήθη για τη μερική εκπλήρωση των υποχρεώσεων απόκτησης του Διδακτορικού Διπλώματος του Πολυτεχνείου Κρήτης.

Η έγκριση της Διδακτορικής Διατριβής από τη Σχολή Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

Η ερευνητική εργασία υποστηρίχτηκε από το Ελληνικό Ίδρυμα Έρευνας και Καινοτομίας (ΕΛ.ΙΔ.Ε.Κ.) στο πλαίσιο της Δράσης «Υποτροφίες ΕΛ.ΙΔ.Ε.Κ. Υποψηφίων Διδακτόρων» (*2η Προκήρυξη Υποτροφιών ΕΛ.ΙΔ.Ε.Κ. για Υποψήφιους Διδάκτορες, Αριθμός Υποτροφίας: 50, 82208/40701*).

Στο **#039**.

Ευχαριστίες

Η εκπόνηση μιας διδακτορικής διατριβής συνήθως αποδεικνύεται μια εξόχως απαιτητική δοκιμασία, καθώς από τη φύση της αποτελεί διαδρομή μοναχική, με πολλές στιγμές αυτό-αμφισβήτησης, προβληματισμού και απογοήτευσης. Εξαίρεση αποτελούν κατά κύριο λόγο δύο περίοδοι: αυτή της έναρξης, που χαρακτηρίζεται από την "παιδική αθωότητα", και αυτή της λήξης (κατά την οποία γράφονται αυτές οι γραμμές), όπου το αίσθημα της ανακούφισης έρχεται δεύτερο μόνο από αυτό της ικανοποίησης. Σε όλα τα ενδιάμεσα στάδια, ο ερευνητής χρειάζεται τόσο την πρακτική όσο και την ψυχική συνδρομή και υποστήριξη μιας μεγάλης ομάδας ανθρώπων, προκειμένου να παραμείνει προσηλωμένος στο στόχο του, διατηρώντας το σθένος και την πίστη στον εαυτό του. Η ενότητα αυτή αποτελεί μια ένδειξη ευγνωμοσύνης σε όλους τους συνοδοιπόρους μου σε αυτό το ταξίδι.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα της διατριβής μου, κ. Νικόλαο Ματσατσίνη, Καθηγητή της Σχολής Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, ο οποίος μου παρείχε ανελλιπώς ουσιαστική και πολύτιμη βοήθεια καθ' όλη τη διάρκεια της έρευνάς μου. Οι υψηλές απαιτήσεις και οι καίριες παρατηρήσεις του αποτέλεσαν σημαντικό κίνητρο για τη βελτίωση των αποτελεσμάτων και του τρόπου συγγραφής της παρούσας διατριβής.

Οι ευχαριστίες επεκτείνονται στον συνεπιβλέποντα κ. Νικόλαο Παπαδάκη, Αναπληρωτή Καθηγητή της Στρατιωτικής Σχολής Ευελπίδων, ο οποίος αποτέλεσε πηγή έμπνευσης στη διαμόρφωση της ερευνητικής μου μεθοδολογίας και μου προσέφερε ανεκτίμητη επιστημονική και ηθική καθοδήγηση.

Επίσης, εκφράζω την ευγνωμοσύνη μου στο έτερο μέλος της τριμελούς μου επιτροπής Καθηγητή κ. Νικόλαο Δάρα, για τις επιστημονικές και πρακτικές του συμβουλές και υποστήριξη, καθώς και στα υπόλοιπα μέλη της επταμελούς εξεταστικής μου επιτροπής για την πρόθυμη συμμετοχή τους στην κρίση της διδακτορικής μου διατριβής.

Το Ελληνικό Ίδρυμα Έρευνας και Καινοτομίας (ΕΛΙΔΕΚ) υπήρξε ουσιαστικός αρωγός στην εκπλήρωση των ονείρων μου, και επομένως δε θα μπορούσα να παραλείψω να το ευχαριστήσω για την επιλογή χρηματοδότησης της διδακτορικής μου διατριβής.

Ένα τεράστιο μερίδιο ευγνωμοσύνης ανήκει δικαιωματικά στην οικογένειά μου, η οποία δε σταμάτησε ποτέ να με στηρίζει, τόσο υλικά όσο και πνευματικά από τη στιγμή της γέννησής μου. Η απλόχερη αγάπη τους και η συνεχής τους εμψύχωση για τη συνέχιση των σπουδών

μου τους καθιστούν πρότυπα για τη ζωή μου. Η πορεία μου μέχρι εδώ δε θα ήταν ίδια χωρίς την πειθαρχία που μου εμφύσησαν και τις ατέλειωτες ώρες βοήθειας που μου προσέφεραν. Ελπίζω κάποια στιγμή να ανταποδώσω τις θυσίες τους.

Τέλος, ακρογωνιαίος λίθος σε αυτή μου τη προσπάθεια υπήρξε η ακούραστη και υπομονετική μου σύντροφος Κατερίνα, η οποία τα τελευταία δώδεκα χρόνια αποτελεί αναπόσπαστο κομμάτι της ζωής μου. Από τα φοιτητικά μας χρόνια και τις ανέμελες συζητήσεις μας για το μέλλον, μέχρι τα πρώτα μας επαγγελματικά βήματα και την κοινή μας πια ερευνητική πορεία, αποτελεί το έτερον μου ήμισυ με κάθε έννοια, καθώς και το πιο ασφαλές λιμάνι σε κάθε τρικυμία. Οι φορές που μου προσέφερε την απέραντη αγάπη της, τις ειλικρινείς συμβουλές και το χρόνο της, υπέμεινε τα νεύρα και τα ξεσπάσματά μου και συνέβαλε ουσιαστικά στην ολοκλήρωση των ερευνητικών μου στόχων είναι πάρα πολλές για να χωρέσουν σε μια εισαγωγική ενότητα. Η παρούσα διατριβή δε θα μπορούσε να ολοκληρωθεί χωρίς τη βοήθειά της και είναι αφιερωμένη σε αυτήν. Σ' ευχαριστώ.

Ολοκληρώνοντας τον κύκλο των ευχαριστιών, συνειδητοποιώ αναδρομικά ότι το ταξίδι παρότι δύσκολο, παρά τα ξενύχτια και τις δύσκολες στιγμές του, υπήρξε πιο όμορφο απ' ότι φανταζόμουν, ίσως ακόμη και από τον τελικό προορισμό. Κι αυτό, χάρη στα παραπάνω πρόσωπα.

Όπως έγραψε και ο Κ. Καβάφης:

Ἡ Ἰθάκη σ' ἔδωσε τ' ὡραῖο ταξίδι.

Χωρὶς αὐτὴν δὲν θἄβγαινες στὸν δρόμο.

Ἄλλα δὲν ἔχει νὰ σὲ δώσει πιά.

Κι ἂν πτωχικὴ τὴν βρῆς, ἡ Ἰθάκη δὲν σὲ γέλασε.

Ἔτσι σοφὸς ποὺ ἔγινες, μὲ τόση πείρα,

ἤδη θὰ τὸ κατάλαβες ἡ Ἰθάκες τί σημαίνουν.

Και η έρευνα για κάτι που αγαπάς, δεν είναι παρά μια θάλασσα, γεμάτη ανεξερεύνητες Ιθάκες.

Με τιμή,

Δημήτρης Παπαδόπουλος

Περίληψη

Η σύγχρονη κοινωνία χαρακτηρίζεται από πρωτοφανή ανάπτυξη στο ρυθμό παραγωγής και διαμοιρασμού δεδομένων και πληροφοριών, ως απόρροια της ραγδαίας αύξησης της υπολογιστικής δύναμης, της διαθεσιμότητας και της δυνατότητας επεξεργασίας τεράστιου όγκου δεδομένων, προερχόμενων κυρίως από το Διαδίκτυο. Αυτός ο κατακλυσμός δεδομένων, ο οποίος συνήθως συναντάται με τη μορφή φυσικής γλώσσας, αναπόφευκτα μειώνει το συλλογικό εύρος προσοχής των παραληπτών, οδηγώντας περισσότερο στην αγχώδη και επιφανειακή κατανάλωσή τους, παρά στην ουσιαστική αφομοίωση και αξιολόγηση τους. Η διεθνής ερευνητική κοινότητα, μέσω εργαλείων και μεθοδολογιών επεξεργασίας φυσικής γλώσσας, προσπαθεί να απαντήσει στην ολοένα αυξανόμενη ζήτηση για αυτοματοποιημένη διαχείριση, αναπαράσταση και εξαγωγή πολύτιμης γνώσης από τις συνεχείς ροές δεδομένων που κατακλύζουν τον Παγκόσμιο Ιστό. Ωστόσο, το μεγαλύτερο μέρος της σημερινής έρευνας επικεντρώνεται σε μόλις 20 από τις περίπου 7000 γλώσσες του κόσμου, αφήνοντας τη συντριπτική πλειονότητα των γλωσσών υπό-μελετημένη. Οι γλώσσες αυτές χαρακτηρίζονται ως χαμηλών πόρων και συνήθως στερούνται αντίστοιχης προσοχής, ή/και δεδομένων για την ανάπτυξη αντίστοιχων μεθόδων. Μια από αυτές τις γλώσσες είναι και η ελληνική.

Είναι πρόδηλη η ανάγκη ανάπτυξης μέσων για την ελληνική γλώσσα τα οποία θα εστιάζουν στη διύλιση δεδομένων που προκύπτουν από τη διάχυση της πληροφορίας στο ευρύ κοινό μέσω του Διαδικτύου. Η παρούσα διδακτορική εργασία αποτελεί προσπάθεια κάλυψης της παραπάνω ανάγκης, με το σχεδιασμό μιας σύγχρονης γνωσιακής μηχανής εξαγωγής πληροφοριών από ελεύθερο κείμενο, ανίχνευσης λανθανουσών συσχετίσεων και προτύπων, που θα αξιοποιεί τον πληροφοριακό πλούτο ελληνικών διαδικτυακών πηγών ώστε να αναγνωρίζει, να ακολουθεί και να συνδυάζει την αλληλουχία εμφάνισης προγενέστερα ασυσχέτιστων δεδομένων (γεγονότων, ειδήσεων, απόψεων κτλ.), επιτρέποντας αφενός την αποτύπωση της πληροφορίας σε δομημένη μορφή και αφετέρου την αξιοποίησή της για τον έλεγχο των ισχυρισμών ενός χρήστη.

Συγκεκριμένα, η εργασία αξιοποιεί μηχανισμούς αυτοματοποιημένης άντλησης και προεπεξεργασίας δεδομένων από πηγές του Ιστού, μέσω κινητών πρακτόρων, με σκοπό την εξαγωγή πληροφοριών σε δομημένη μορφή και την εκμετάλλευσή τους για εργασίες διερευνητικής ανάλυσης και διαμόρφωσης αρχικών υποθέσεων. Ακόμη, μελετώνται και αναπτύσσονται εξελιγμένες γνωσιακές τεχνικές για την εξαγωγή σημασιολογικών

συμπερασμάτων μέσω του εντοπισμού και συσχέτισης εννοιολογικών οντοτήτων, με απώτερο στόχο την ανακάλυψη συσχετίσεων μεταξύ φαινομενικά ασύνδετων γεγονότων, προσώπων και πράξεων. Το τελικό προϊόν της εργασίας περιλαμβάνει το σχεδιασμό και υλοποίηση μεθοδολογιών εξαγωγής πληροφορίας από αδόμητο κείμενο καθώς και δυναμικού ελέγχου των ισχυρισμών ενός χρήστη (σε ελεύθερο κείμενο) βάσει της συγκεντρωθείσας πληροφορίας. Τα παραπάνω συνοδεύονται από την ανάπτυξη αντίστοιχων μοντέλων μηχανικής μάθησης που υποστηρίζουν τις παραπάνω εργασίες για την ελληνική γλώσσα.

Οι μηχανισμοί που προκύπτουν από την ανάπτυξη των προαναφερθεισών μεθοδολογιών επιτρέπουν την αποτύπωση κειμένου σε δομημένη μορφή (σχεσιακών ν-πλειάδων), για την καλύτερη διαχείριση της εξαχθείσας πληροφορίας μέσω βάσεων δεδομένων καθώς και για τον εμπλουτισμό της μέσω συσχετίσεων με εξωτερικές γνωσιακές βάσεις. Επιπλέον, καθίσταται δυνατή η δυνατότητα επικύρωσης ή απόρριψης ενός οποιουδήποτε ισχυρισμού, μέσω του συνδυασμού ετερογενών πληροφοριών από πολλαπλές πηγές σε πραγματικό χρόνο, αξιοποιώντας την προτεινόμενη μεθοδολογία κατασκευής σχετικών τεκμηρίων.

Στις ενότητες που ακολουθούν, περιγράφονται οι κύριοι στόχοι και τα ερευνητικά ερωτήματα της διατριβής (Κεφάλαιο 1), παρατίθεται το απαραίτητο θεωρητικό υπόβαθρο και σχετικό ερευνητικό έργο που αφορά την ανάπτυξη των επιμέρους συνιστωσών της προτεινόμενης μεθοδολογίας (Κεφάλαιο 2). Ακολουθεί η λεπτομερής περιγραφή της μεθοδολογίας για κάθε ερευνητικό άξονα (Κεφάλαιο 3), η οποία αποτελεί εφαλτήριο για την τεχνική υλοποίηση και αξιολόγηση των προϊόντων της διατριβής (Κεφάλαιο 4). Η εργασία ολοκληρώνεται με ανασκόπηση των κύριων αποτελεσμάτων, περιγραφή των περιορισμών και προτάσεις για επέκταση της έρευνας (Κεφάλαιο 5).

Λέξεις-κλειδιά: επεξεργασία φυσικής γλώσσας, γλώσσες χαμηλών πόρων, εξαγωγή πληροφοριών, έλεγχος ισχυρισμών, έλεγχος σημασιολογικής ομοιότητας, αναγνώριση κειμενικής συνεπαγωγής, σύνδεση οντοτήτων, κατασκευή τεκμηρίων

Abstract

Modern society is characterized by an unprecedented growth in the ways data and information are being produced and shared, as a result of the rapid increase in computing power, of the availability of resources and of the ability to process huge data volumes, mainly derived from Internet sources. The occurring data flood, commonly encountered in the form of natural language, inevitably reduces the recipients' collective attention span, leading more to the stressful and superficial consumption of information, rather than to its actual assimilation and evaluation. Many research groups worldwide are responding to the growing demand for automated management, representation and extraction of valuable knowledge from the continuous data streams that are overwhelming the Web, by exploiting natural language processing methodologies and tools. However, most of today's research is disproportionally focused on around 20 of the world's more than 7000 spoken languages, leaving the vast majority of them under-studied. These languages are characterized as low-resource, since they usually lack the corresponding attention and/or data for the development of meaningful applications. Greek belongs to this language group.

There is a dire need for the development of methods that will distill information from natural language content produced in Greek. This doctoral dissertation represents an attempt to meet the above need, through the design of a modern cognitive engine that enables the detection of latent correlations and patterns between entities, through the exploitation of the information wealth derived from Greek online sources and the combination of previously unrelated data (events, news, opinions etc.). This allows both the capture of information in a structured form, as well as its use for claim validation in natural language.

More specifically, the dissertation utilizes automated crawling and pre-processing techniques on online news sources, in order to extract structured information that can be used for exploratory data analysis purposes and for the formulation of initial claims or hypotheses. In addition, it pertains to the development of advanced cognitive machine learning methods to achieve semantic inference and draw conclusions from the identification and connections between conceptual entities, ultimately aiming at the discovery of correlations between seemingly unrelated events, persons or actions. The final product of this work includes the design and implementation of a set of

methodologies for information extraction and dynamic claim validation based on the accumulated information. All the above are accompanied by the development of corresponding machine learning models to support this work for the Greek use case.

The mechanisms that will result from the development of the aforementioned methodologies allow the transformation of free-text to a structured representation (relational n-tuples), enabling better database management and enrichment with the help of external knowledge bases. Moreover, they render possible the validation or rejection of any textual claim, by aggregating heterogeneous information from multiple sources in real time, via a proposed evidence construction methodology.

The following subsections describe the main objectives and research questions of this dissertation (Chapter 1), provide the necessary theoretical background and present the previous work which constitutes the basis for the development of the aforementioned mechanisms (Chapter 2). A detailed description of the proposed methodology for each research axis (Chapter 3) acts as the spearhead for the technical implementation and evaluation of the derived research products (Chapter 4). The dissertation concludes with a review of the main results, the presentation of its limitations and suggestions for future research extensions and improvements (Chapter 5).

Keywords: natural language processing, low-resource languages, information extraction, claim validation, semantic textual similarity, natural language inference, entity linking, evidence construction

Περιεχόμενα

Δημοσιεύσεις σε επιστημονικά περιοδικά και πρακτικά συνεδρίων

Λίστα άμεσα συσχετιζόμενων δημοσιεύσεων με επιμέρους μεθοδολογίες και εφαρμογές που αναφέρονται στην παρούσα διδακτορική διατριβή:

1. Papadopoulos, D., Papadakis, N., & Litke, A. (2020). A methodology for open information extraction and representation from large scientific corpora: the CORD-19 data exploration use case. Applied Sciences, 10(16), 5630. DOI: 10.3390/app10165630

2. Papadopoulos, D., Papadakis, N., & Matsatsinis, N. (2021). PENELOPIE: Enabling Open Information Extraction for the Greek Language through Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL SRW)* (pp. 23-29). DOI: 10.18653/v1/2021.eacl-srw.4

3. Smith, E., Papadopoulos, D., Braschler, M., & Stockinger, K. (2022). LILLIE: Information extraction and database integration using linguistics and learning-based algorithms. *Information Systems*, *105*, 101938. DOI: 10.1016/j.is.2021.101938

4. Amer-Yahia, S., Koutrika, G., Braschler, M., Calvanese, D., Lanti, D., Lücke-Tieke, H., Mosca, A., Mendez de Farias, T., Papadopoulos, D., Patil, Y., Rull, G., Smith, E., Skoutas, D., Subramanian, S., & Stockinger, K. (2022). INODE: building an end-to-end data exploration system in practice. *ACM SIGMOD Record*, *50*(4), 23-29. DOI: /10.1145/3516431.3516436

5. Papadopoulos, D., Mitropoulou, K., Papadakis, N., & Matsatsinis, N. (2022). FarFetched: Entity-centric Reasoning and Claim Validation for the Greek Language based on Textually Represented Environments. *12th Conference on Artificial Intelligence (SETN 2022)* (Accepted)
    - Poster presentations: i) NAACL 2022: Workshop on Multilingual Information Access (MIA), ii) NAACL 2022: Structured and Unstructured Knowledge Integration (SUKI), iii) NAACL 2022: Deep Learning for Low-Resource NLP (DeepLo)

Λίστα δημοσιεύσεων που αφορούν το γενικότερο ερευνητικό πλαίσιο της μηχανικής μάθησης και εκπονήθηκαν κατά τη διάρκεια των διδακτορικών σπουδών:

1. Attak, H., Combalia, M., Gardikis, G., Gastón, B., Jacquin, L., Katsianis, D., Litke, A., Papadakis, N., Papadopoulos, D., Pastor, A., Roig, M., & Segou, O. (2018). Application of distributed computing and machine learning technologies to cybersecurity. *Computer & Electronics Security Applications Rendez-vous (C&ESAR), 19-21 November 2018*. DOI: 10.5281/zenodo.3266038

2. Papadakis, N., Havenetidis, K., Papadopoulos, D., & Bissas, A. (2020). Employing body-fixed sensors and machine learning to predict physical activity in military personnel. *BMJ Mil Health*. DOI: 10.1136/bmjmilitary-2020-001585

3. Kompougias, O., Papadopoulos, D., Mantas, E., Litke, A., Papadakis, N., Paraschos, D., Kourtis, M., & Xylouris, G. (2021). IoT Botnet Detection on Flow Data using Autoencoders. In *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)* (pp. 506-511). IEEE. DOI: 10.1109/MeditCom49071.2021.9647639.

4. Mantas, E., Papadopoulos, D., Fernández, C., Ortiz, N., Compastié, M., Martínez, A. L., Pérez, M., Kourtis, M., Xylouris, G., Mlakar, I., Tsarsitalidis, S., Klonidis, D., Pedone, I., Canavese, D., Perez, G., Sanvito, D., Logothetis, V., Lopez, D., Pastor, A., Lioy, A., Jacquin, L., Bifulco, R., Kapodistria, A., Priovolos, A., Gardikis, G., Neokosmidis, I., Rokkas, T., Papadakis, N., Paraschos, D., Jeran, P., Litke, A. & Athanasiou, G. (2021). Practical Autonomous Cyberhealth for resilient Micro, Small and Medium-sized Enterprises. In *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)* (pp. 500-505). IEEE. DOI: 10.1109/MeditCom49071.2021.9647609.

5. Kourtis, M. A., Oikonomakis, A., Papadopoulos, D., Xylouris, G., & Chochliouros, I. P. (2021). Leveraging Deep Learning for Network Anomaly Detection. In *2021 Sixth International Conference on Fog and Mobile Edge Computing (FMEC)* (pp. 1-6). IEEE. DOI: 10.1109/FMEC54266.2021.9732556.

Συνεισφορές στην Κοινότητα Ανοιχτού Λογισμικού

Τα παρακάτω γλωσσικά μοντέλα εκπαιδεύτηκαν με σκοπό να υποστηρίζουν τις μεθοδολογίες που αναπτύχθηκαν κατά την εκπόνηση της διατριβής και είναι ελεύθερα διαθέσιμα για ακαδημαϊκή/εμπορική χρήση:

1. Γλωσσικό μοντέλο μηχανικής μετάφρασης από Ελληνικά σε Αγγλικά: https://huggingface.co/lighteternal/SSE-TUC-mt-el-en-cased
2. Γλωσσικό μοντέλο μηχανικής μετάφρασης από Αγγλικά σε Ελληνικά: https://huggingface.co/lighteternal/SSE-TUC-mt-en-el-cased
3. Γλωσσικό μοντέλο μηχανικής μετάφρασης από Ελληνικά σε Αγγλικά (μόνο πεζοί χαρακτήρες): https://huggingface.co/lighteternal/SSE-TUC-mt-el-en-lowercase
4. Γλωσσικό μοντέλο μηχανικής μετάφρασης από Αγγλικά σε Ελληνικά (μόνο πεζοί χαρακτήρες): https://huggingface.co/lighteternal/SSE-TUC-mt-en-el-lowercase
5. Γλωσσικό μοντέλο παραγωγής κειμένου για την ελληνική γλώσσα: https://huggingface.co/lighteternal/gpt2-finetuned-greek
6. Γλωσσικό μοντέλο αυτόματης αναγνώρισης ομιλίας για την ελληνική γλώσσα: https://huggingface.co/lighteternal/wav2vec2-large-xlsr-53-greek
7. Γλωσσικό μοντέλο αναγνώρισης κειμενικής συνεπαγωγής για την ελληνική γλώσσα: https://huggingface.co/lighteternal/nli-xlm-r-greek
8. Γλωσσικό μοντέλο ελέγχου σημασιολογικής ομοιότητας για την ελληνική γλώσσα: https://huggingface.co/lighteternal/stsb-xlm-r-greek-transfer
9. Γλωσσικό μοντέλο αναγνώρισης υποκειμενικότητας/αντικειμενικότητας για την ελληνική γλώσσα: https://huggingface.co/lighteternal/fact-or-opinion-xlmr-el

Λίστα Πινάκων

Λίστα Σχημάτων

Every piece of information in the world has been copied. Backed up.
Except the human mind. The last analog device in a digital world.

-Dr. Robert Ford - Westworld, S2. Ep7: Les Écorchés

1 ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή του ερευνητικού προβλήματος

Η σημερινή εποχή χαρακτηρίζεται από την "έκσταση της επικοινωνίας". Η έκρηξη πληροφοριών που λαμβάνει χώρα κυρίως μέσω του Διαδικτύου τείνει να προσανατολίζει τη διανοητική δραστηριότητα περισσότερο στη διαχείριση της υπερπαραγωγής δεδομένων, παρά στην εμβάθυνση και στην κριτική σκέψη. Παράλληλα, τα δεδομένα αποτελούν πλέον μια νέα μορφή κεφαλαίου, εξίσου σημαντικού με το οικονομικό και το ανθρώπινο κεφάλαιο όσον αφορά τη δημιουργία νέων ψηφιακών προϊόντων και υπηρεσιών. Οι πληροφορίες που αποτελούν το νέο "συνάλλαγμα" βρίσκονται σε αδόμητη μορφή, δηλαδή ως ελεύθερο κείμενο, δυσχεραίνοντας τεχνικές αυτοματοποιημένου συλλογισμού και ερμηνείας.

Η Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ) αναφέρεται στη χρήση υπολογιστικών μεθόδων για την επεξεργασία προφορικής ή γραπτής μορφής τέτοιου ελεύθερου κειμένου, το οποίο αποτελεί το κύριο μέσο ανθρώπινης επικοινωνίας. Τα τελευταία χρόνια, ο συγκεκριμένος κλάδος έχει εξελιχθεί σε δυναμικά αναπτυσσόμενο ερευνητικό πεδίο, ως απόρροια της ραγδαίας αύξησης της υπολογιστικής δύναμης, της δυνατότητας επεξεργασίας τεράστιου όγκου δεδομένων προερχόμενων κυρίως από το Διαδίκτυο και της ανάπτυξης εξαιρετικά επιτυχημένων αλγορίθμων μηχανικής μάθησης. Από τις αρχές του 1990, οι τεχνικές επεξεργασίας φυσικής γλώσσας μετασχηματίστηκαν κατάλληλα ώστε να παράγουν εμπειρικά μοντέλα βασιζόμενα στην εκμάθηση από σύνολα λέξεων (bag of words) ως τμήματα ενός τεράστιου όγκου δεδομένων, χωρίς ωστόσο να εστιάζουν στο πραγματικό νόημα των λέξεων μέσα στην πρόταση. Οι πλέον σύγχρονες προσεγγίσεις περιλαμβάνουν τη χρήση σύνθετων μεθόδων βαθιάς μάθησης και εξόρυξης κειμένου (text mining) για την κατανόηση τόσο της

γλωσσολογικής δομής και της συντακτικής πληροφορίας, όσο και της εννοιολογικής πληροφορίας, η οποία μπορεί να παρουσιάζει ασάφειες ή αμφισημίες ανάλογα με το είδος του λόγου και τα συναισθήματα του συγγραφέα.

Η προαναφερθείσα έκρηξη πληροφοριών και η ανάγκη για πιο εξελιγμένα και αποτελεσματικά εργαλεία διαχείρισής της έχει επηρεάσει κάθε σύγχρονη κοινωνία, συμπεριλαμβανομένης φυσικά και της ελληνικής. Η διεθνής ερευνητική κοινότητα έχει απαντήσει στην ολοένα αυξανόμενη ζήτηση για αντίστοιχες μεθόδους επεξεργασίας, με πλήθος παράλληλων ερευνητικών γραμμών που αφορούν την αυτοματοποιημένη εξαγωγή και ανάκτηση πληροφοριών, με απώτερο στόχο την αποτελεσματική ανάλυση του ελεύθερου κειμένου και την ανακάλυψη πολύτιμης σχετικής γνώσης από αυτό σε δομημένη μορφή. Ωστόσο, το μεγαλύτερο μέρος της σημερινής έρευνας αντίστοιχων τεχνολογιών επικεντρώνεται σε μόλις 20 από τις 7000 γλώσσες του κόσμου, αφήνοντας τη συντριπτική πλειονότητα των γλωσσών υπό-μελετημένη. Αυτές οι γλώσσες, οι οποίες αναφέρονται συχνά ως γλώσσες χαμηλών πόρων, συνήθως στερούνται αντίστοιχης προσοχής, ή/και δεδομένων για την ανάπτυξη αντίστοιχων μεθόδων που συναντώνται στις λεγόμενες γλώσσες υψηλών πόρων, όπως η αγγλική. Από υπολογιστικής σκοπιάς η ελληνική γλώσσα, αν και βρίσκεται σαφώς σε καλύτερη θέση από άποψης πόρων σε σχέση με επαπειλούμενες γλώσσες και ιθαγενείς διαλέκτους, αποτελεί γλώσσα χαμηλότερων πόρων σε σχέση με τις περισσότερες σύγχρονες ευρωπαϊκές γλώσσες. Επομένως, είναι πρόδηλη η ανάγκη ανάπτυξης αντίστοιχων μεθοδολογιών και εργαλείων ΕΦΓ για την ελληνική γλώσσα, ως μέσο διύλισης των δεδομένων που προκύπτουν από τη διάχυση της πληροφορίας στο ευρύ κοινό μέσω του Διαδικτύου.

Αντικείμενο της παρούσας διδακτορικής εργασίας είναι ο σχεδιασμός και η ανάπτυξη μιας σύγχρονης γνωσιακής μηχανής εξαγωγής πληροφοριών από ελεύθερο κείμενο, ανίχνευσης λανθανουσών συσχετίσεων και προτύπων, που θα αξιοποιεί τον πληροφοριακό πλούτο ελληνικών διαδικτυακών πηγών ώστε να αναγνωρίζει, να ακολουθεί και να συνδυάζει την αλληλουχία εμφάνισης προγενέστερα ασυσχέτιστων δεδομένων (γεγονότων, ειδήσεων, απόψεων κτλ.), επιτρέποντας αφενός την αποτύπωση της πληροφορίας σε δομημένη μορφή και αφετέρου την αξιοποίησή της για τον έλεγχο των ισχυρισμών ενός χρήστη.

Οι **επιμέρους στόχοι** της εργασίας συνοψίζονται ως εξής:

i. η ανάπτυξη ενός μηχανισμού συγκέντρωσης και προεπεξεργασίας διαδικτυακών δεδομένων, αξιοποιώντας μεθόδους αυτοματοποιημένης άντλησής τους από ελεύθερες πηγές του Ιστού με χρήση κινητών πρακτόρων,
ii. η ανάπτυξη μεθόδων εξαγωγής πληροφοριών από το συγκεντρωμένο ελεύθερο κείμενο, με απώτερο στόχο τη δομημένη αναπαράσταση, τη διαχείριση και εκμετάλλευσή τους για εργασίες διερευνητικής ανάλυσης και διαμόρφωσης αρχικών υποθέσεων,
iii. η ανάπτυξη και χρήση εξελιγμένων γνωσιακών τεχνικών για την εξαγωγή σημασιολογικών συμπερασμάτων, τον εντοπισμό και τη συσχέτιση εννοιολογικών οντοτήτων, η αξιοποίηση αλγοριθμικών μεθόδων αναγνώρισης προτύπων για την ανακάλυψη συσχετίσεων μεταξύ φαινομενικά ασύνδετων γεγονότων, προσώπων, ειδήσεων και πράξεων,
iv. ο αρχιτεκτονικός σχεδιασμός και η τεχνική υλοποίηση ενός μηχανισμού για την ελληνική γλώσσα με χρήση των παραπάνω τεχνικών, ο οποίος θα επιτρέπει το δυναμικό έλεγχο των ισχυρισμών ενός χρήστη βάσει της δομημένης συγκεντρωθείσας πληροφορίας, καθώς και η αξιολόγησή του μέσω σχετικών περιπτώσεων χρήσης.

1.2 Ερευνητικά ζητήματα και υποθέσεις εργασίας

Τα **ερευνητικά ζητήματα** τα οποία καλείται να απαντήσει η εργασία μπορούν να ενταχθούν σε τρεις διακριτές κατηγορίες βάσει του παραπάνω πλαισίου, και συνοψίζονται ως εξής:

Ζ1. Ανάπτυξη της βέλτιστης μεθοδολογίας συγκέντρωσης, προεπεξεργασίας και αναπαράστασης δεδομένων ελεύθερου κειμένου από διαδικτυακές πηγές. Συγκεκριμένα, μελέτη και συνδυασμός τεχνικών αυτοματοποιημένης ιχνηλάτησης, άντλησης, προεπεξεργασίας και αποθήκευσης δεδομένων από ειδησεογραφικούς ιστοτόπους και κατάλληλη προσαρμογή τους με σκοπό τη δημιουργία ενός δυναμικού συστήματος που θα οδηγεί στην αποδοτικότερη αφομοίωση και αξιοποίηση της αποθηκευμένης πληροφορίας για εργασίες εξαγωγής πληροφοριών και ελέγχου υποθέσεων.

Ζ2. Σχεδιασμός και υλοποίηση μεθοδολογίας για την εξαγωγή πληροφοριών από ελεύθερο κείμενο για την ελληνική γλώσσα. Αποτύπωση ελεύθερου κειμένου σε δομημένη μορφή (πχ. σε μορφή τριπλετών) που θα διευκολύνει την διερευνητική ανάλυσή του (πχ. γραφική σύνοψη γεγονότων). Ανάπτυξη και εφαρμογή μεθόδων επισήμανσης ονοματικών οντοτήτων

και σύνδεσή τους με γνωσιακές βάσεις, αναγνώρισης προτύπων και λανθανουσών συσχετίσεων μεταξύ οντοτήτων. Σχεδιασμός μεθοδολογίας που θα αξιοποιεί συνδυαστικές τεχνικές μηχανικής/βαθιάς μάθησης καθώς και συστήματα που βασίζονται σε γλωσσολογικούς κανόνες για επίτευξη του καλύτερου δυνατού αποτελέσματος και αξιολόγησή της απόδοσης μέσω σχετικών σημείων αναφοράς και ρεαλιστικών περιπτώσεων χρήσης.

Ζ3. Σχεδιασμός μεθοδολογίας ελέγχου υποθέσεων/ισχυρισμών σε ελεύθερο κείμενο για την ελληνική γλώσσα, βάσει της δομημένης αναπαράστασης της αποκτηθείσας πληροφορίας. Ανάπτυξη ενός δυναμικού συστήματος το οποίο θα μπορεί να επικυρώνει ή να απορρίπτει έναν ισχυρισμό, συνδυάζοντας πληροφορίες από ετερογενείς πηγές σε πραγματικό χρόνο, αξιοποιώντας σχετικούς μηχανισμούς (πχ. σημασιολογικής ομοιότητας, κειμενικής συνεπαγωγής), προσφέροντας δυνατότητες αιτιολόγησης του αποτελέσματος και χρησιμοποιώντας αντίστοιχα μοντέλα βαθιάς μάθησης που θα εκπαιδευτούν για την ελληνική γλώσσα. Συγκριτική μελέτη της συνολικής απόδοσης της μεθοδολογίας καθώς και των επιμέρους μηχανισμών που την απαρτίζουν. Αναγνώριση των περιορισμών που διέπουν τη μεθοδολογία και προτάσεις για συνέχιση της έρευνας.

Παράλληλα, η εργασία στοχεύει στον έλεγχο των παρακάτω **υποθέσεων εργασίας**:

Υ1. Οι συνήθεις εργασίες ΕΦΓ όπως η αναγνώριση ονοματικών οντοτήτων, η εξαγωγή πληροφοριών και ο έλεγχος σημασιολογικής ομοιότητας παρουσιάζουν περισσότερες προκλήσεις για μορφολογικά πλούσιες γλώσσες όπως η ελληνική, σε σχέση με τις πιο διαδεδομένες ευρωπαϊκές γλώσσες υψηλών πόρων, δεδομένων των συντακτικών και γραμματικών ιδιαιτεροτήτων που τις χαρακτηρίζουν. Επομένως, ενδέχεται να χρειαστεί κατάλληλη προσαρμογή των αντίστοιχων τεχνικών μηχανικής/βαθιάς μάθησης για την αντιμετώπιση των παραπάνω ιδιαιτεροτήτων.

Υ2. Η έλλειψη επισημασμένων δεδομένων για την ελληνική γλώσσα θα αποτελέσει βασικό εμπόδιο για την εκπαίδευση νευρωνικών μοντέλων που υλοποιούν τις βασικές συνιστώσες των προαναφερθεισών μεθοδολογιών. Δεδομένου ότι στα περισσότερα πολύγλωσσα νευρωνικά μοντέλα τεχνολογίας αιχμής, οι γλώσσες χαμηλότερων πόρων (όπως η ελληνική) υπό-αντιπροσωπεύονται συστηματικά, αναμένεται πως η χρήση τεχνικών μεταφοράς μάθησης ή άλλων ενδιάμεσων τεχνικών θα αποτελέσει μονόδρομο για την αντιμετώπιση του προβλήματος.

Υ3. Αναφορικά με τη μεθοδολογία εξαγωγής πληροφοριών από ελεύθερο κείμενο, αναμένεται πως τόσο οι τεχνικές που βασίζονται σε γλωσσολογικούς κανόνες, όσο και αυτές

που βασίζονται σε υπολογιστικές μεθόδους εμφανίζουν διαφορετικά πλεονεκτήματα και μειονεκτήματα. Επομένως, αξίζει η μελέτη συνδυαστικής χρήσης τους με σκοπό τη βελτίωση του τελικού αποτελέσματος εξαγωγής.

Υ4. Όσον αφορά τη μεθοδολογία ελέγχου ισχυρισμών, είναι προφανής η εξάρτηση της απόδοσής της από πλήθος παραγόντων όπως η ποιότητα και το μέγεθος των εισαχθέντων δεδομένων, η ακρίβεια των επιμέρους μηχανισμών ΕΦΓ, το εκάστοτε θεματικό πλαίσιο κ.α. Ειδικότερα για περιπτώσεις ανάλυσης ειδησεογραφικών πληροφοριών που προϋποθέτουν ένα δυναμικό σύστημα, οι συσχετίσεις μεταξύ οντοτήτων ενδέχεται να μεταβάλλονται συνεχώς, καθιστώντας απαραίτητη την περιοδική τροφοδότηση του μηχανισμού με νέα πληροφορία, ώστε να καταστεί δυνατή η παρακολούθηση τυχόν τάσεων στην πάροδο του χρόνου.

Υ5. Συμπληρωματικά με την παραπάνω υπόθεση, κατά τον έλεγχο ενός ισχυρισμού είναι δυνατή η εμφάνιση συμπληρωματικών τεκμηρίων που προέρχονται από διαφορετικές πηγές, κάτι που πρέπει να ληφθεί υπόψη κατά την ανάπτυξη της μεθοδολογίας. Εξάλλου, ενδέχεται η ύπαρξη ακόμα και αντικρουόμενων τεκμηρίων ως συνιστώσες του ίδιου αποτελέσματος, ως αποτέλεσμα της υποκειμενικότητας της αναλυόμενης πληροφορίας-είδησης κατά την κάλυψη ενός γεγονότος από διαφορετικές πηγές. Επομένως είναι χρήσιμη η ανάπτυξη ενός συμπληρωματικού μηχανισμού που θα διαχωρίζει τα δεδομένα που αφορούν πραγματικά γεγονότα από αυτά που εκφράζουν υποκειμενικές απόψεις.

1.3 Ορισμός και τρόποι εκπλήρωσης τεχνικών απαιτήσεων

Σε πρώιμο στάδιο της διατριβής έγινε αντιστοίχιση των παραπάνω ερευνητικών ζητημάτων και υποθέσεων σε καθορισμένες τεχνικές απαιτήσεις, με σκοπό την καλύτερη παρακολούθηση της υλοποίησης της έρευνας. Αυτές χωρίζονται σε δύο κατηγορίες:

- λειτουργικές απαιτήσεις, οι οποίες περιγράφουν τις αναμενόμενες λειτουργίες που πρέπει να πληρούν οι μηχανισμοί που θα αναπτυχθούν στα πλαίσια της εργασίας
- μη-λειτουργικές απαιτήσεις, οι οποίες περιγράφουν συγκεκριμένες ιδιότητες των επιμέρους στοιχείων (χαρακτηριστικά απόδοσης, χρηστικότητας κτλ.) και περιγράφουν το πόσο καλά οι υλοποιούμενοι μηχανισμοί θα υποστηρίξουν τις λειτουργικές απαιτήσεις.

Στους παρακάτω πίνακες παρατίθενται οι σχετικές απαιτήσεις, ενώ γίνεται αναφορά στο συγκεκριμένο στοιχείο που υλοποιήθηκε για την εκπλήρωσή τους.

Πίνακας 1 Λειτουργικές απαιτήσεις

Λειτουργικές απαιτήσεις		*Τρόπος εκπλήρωσης*
F1	Δυνατότητα πρόσβασης της μηχανής σε διαδικτυακές πηγές δεδομένων (ειδησεογραφικές πηγές).	Στοιχεία ιχνηλάτησης τροφοδοσιών RSS και HTML σελίδων (§4.3.1.1.3, §4.3.1.1.4)
F2	Δυνατότητα συλλογής, προ-επεξεργασίας και αποθήκευσης δεδομένων με αυτοματοποιημένο τρόπο.	Στοιχεία ιχνηλάτησης τροφοδοσιών RSS και HTML σελίδων (§4.3.1.1.3, §4.3.1.1.4) Βάση δεδομένων γράφων (§4.3.1.1.2)
F3	Δυνατότητα μετασχηματισμού ελεύθερου κειμένου σε δομημένη γνώση.	Στοιχεία ανοιχτής εξαγωγής πληροφοριών (§4.2.1.4)
F4	Δυνατότητα αξιοποίησης μεθόδων μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας για εξαγωγή οντοτήτων και σχέσεων από αδόμητα δεδομένα (στα Ελληνικά).	Στοιχεία ανοιχτής εξαγωγής πληροφοριών (§4.2.1.4) Στοιχεία σύνδεσης οντοτήτων (§4.3.1.1.2)
F5	Δυνατότητα ανακάλυψης λανθανουσών συσχετίσεων μεταξύ οντοτήτων ακόμα και αν δεν ανήκουν στην ίδια πηγή.	Στοιχείο κατασκευής τεκμηρίων (§4.3.1.3)
F6	Δυνατότητα αναπαράστασης των εξαχθεισών οντοτήτων.	Στοιχεία σύνδεσης οντοτήτων (§4.3.1.1.2) Βάση δεδομένων γράφων (§4.3.1.1.2)
F7	Δυνατότητα αποτύπωσης της εξέλιξης των συσχετίσεων που χαρακτηρίζουν τις εξαχθείσες οντότητες.	Στοιχεία ανοιχτής εξαγωγής πληροφοριών (§4.2.1.4)
F8	Δυνατότητα αλληλεπίδρασης και εξερεύνησης των ευρημάτων από τον χρήστη.	Βάση δεδομένων γράφων (§4.3.1.1.2)
F9	Δυνατότητα εξαγωγής των προϊόντων ανάλυσης σε δομημένη μορφή (πχ. csv, json, xml).	Βάση δεδομένων γράφων (§4.3.1.1.2)
F10	Δυνατότητα ελέγχου ισχυρισμών/υποθέσεων σε ελεύθερο κείμενο βάσει των αποθηκευμένων πληροφοριών από ειδησεογραφικές πηγές.	Στοιχεία ελέγχου σημασιολογικής ομοιότητας και αναγνώρισης κειμενικής συνεπαγωγής (§4.3.1.4, §4.3.1.5) Στοιχείο κατασκευής τεκμηρίων

		(§4.3.1.3)
F11	Δυνατότητα διάκρισης μεταξύ πληροφοριών που χαρακτηρίζονται από υποκειμενικότητα (άποψη) και αντικειμενικότητα (γεγονός)	Στοιχείο αναγνώρισης υποκειμενικότητας/ αντικειμενικότητας (§4.3.1.6)

Πίνακας 2 Μη-λειτουργικές απαιτήσεις

Μη-Λειτουργικές απαιτήσεις		*Τρόπος εκπλήρωσης*
NF1	Η επεκτασιμότητα των δυνατοτήτων συλλογής και αποθήκευσης δεδομένων θα πρέπει να είναι εξασφαλισμένη (πχ. με υποστήριξη κατανεμημένων συστημάτων).	Η τεχνολογία αποθήκευσης (βάση δεδομένων γράφων) που επιλέχθηκε, επιτρέπει τόσο οριζόντια όσο και κατακόρυφη επεκτασιμότητα.
NF2	Τα υποσυστήματα ανάλυσης δεδομένων της μηχανής θα πρέπει να μπορούν να ανταπεξέλθουν των απαιτήσεων (υπολογιστικές, μνήμης) που αφορούν τη διαχείριση μεγάλων συνόλων δεδομένων.	Τα νευρωνικά μοντέλα που υλοποιούν τα διάφορα στοιχεία των μηχανισμών μπορούν εύκολα να εξαχθούν σε αναπαράσταση βελτιστοποιημένη για εργασίες κατανεμημένης υπολογιστικής (πχ. ONNX runtime).
NF3	Η ανάλυση δεδομένων για κάθε διακριτό στοιχείο του συνόλου (π.χ. ειδησεογραφική πηγή) θα πρέπει να γίνεται σε εύλογο χρονικό διάστημα.	Τα στοιχεία ανοιχτής εξαγωγής πληροφοριών εκτελούν τις εργασίες ανάλυσης σε χρόνο <10 seconds. (βλ. δοκιμές §4.2.2.2.)
NF4	Η αρχιτεκτονική της μηχανής θα πρέπει να εξασφαλίζει την ανεξαρτησία μεταξύ των χρησιμοποιούμενων τεχνολογιών, ακολουθώντας αρθρωτή δομή όπου επιτρέπεται.	Ακολουθήθηκε αρθρωτή αρχιτεκτονική κατά την υλοποίηση τόσο του μηχανισμού εξόρυξης πληροφοριών (§3.3) όσο και του μηχανισμού εξαγωγής σημασιολογικών συμπερασμάτων (§3.4).
NF5	Η αναπαράσταση των δεδομένων ανάλυσης θα πρέπει να είναι φιλική προς το χρήστη, ακολουθώντας κατά το δυνατόν διαδραστικές τεχνικές.	Η βάση δεδομένων γράφων που επιτρέπει την απεικόνιση της αποκτηθείσας πληροφορίας από τον τελικό χρήστη, μέσω διαδραστικών τεχνικών, χωρίς τη χρήση κώδικα.
NF6	Οι παραγόμενοι μηχανισμοί θα πρέπει να είναι	Κανένα από τα επιμέρους στοιχεία που

	αγνωστικοί όσον αφορά το λειτουργικό σύστημα.	απαρτίζουν τους μηχανισμούς δεν εξαρτάται από συγκεκριμένο λειτουργικό σύστημα.
NF7	Η μηχανή θα πρέπει να βασίζεται σε λογισμικό ανοιχτού κώδικα (όπου υπάρχει διαθέσιμο), επιτρέποντας τη μέγιστη δυνατή επεκτασιμότητα και ενισχύοντας τη συνεργατικότητα με την ακαδημαϊκή κοινότητα.	Η ανάπτυξη των επιμέρους στοιχείων βασίστηκε αποκλειστικά σε λογισμικό ανοιχτού κώδικα. Επιπλέον, όλα τα προϊόντα της εργασίας είναι διαθέσιμα στην κοινότητα ανοιχτού λογισμικού.

1.4 Περιγραφή υφιστάμενης κατάστασης

Η ερευνητική προσέγγιση που θα ακολουθηθεί για την εκπλήρωση των στόχων της εργασίας βασίζεται στους πυλώνες της ανοιχτής εξαγωγής πληροφοριών, της χρήσης διανυσματικών αναπαραστάσεων λέξεων/προτάσεων για αποτύπωση σημασιολογικής και συντακτικής πληροφορίας, καθώς και στον έλεγχο ισχυρισμών από ελεύθερο κείμενο. Βασική προϋπόθεση για την ανάπτυξη των παραπάνω μεθόδων αποτελεί η αξιοποίηση τεχνικών συλλογής και ανάκτησης μεγάλου όγκου διαδικτυακών δεδομένων (πχ. μέσω ιχνηλάτησης σελίδων) για την εξαγωγή οντοτήτων από το σώμα κειμένων, την εισαγωγή τους σε βάσεις δεδομένων (σχεσιακές ή γράφων) και την επεξεργασία τους με χρήση τεχνικών βαθιάς μάθησης για την εύρεση λανθανουσών συσχετίσεων. Στην παρούσα ενότητα γίνεται μια εισαγωγή στους παραπάνω βασικούς πυλώνες της εργασίας, αγγίζοντας το υφιστάμενο επίπεδο τεχνολογικής αιχμής καθώς και τις βασικές προκλήσεις που καλείται να καλύψει η παρούσα εργασία. Σημειώνεται ότι μια λεπτομερέστερη βιβλιογραφική ανασκόπηση σε επίπεδο στοιχείου δίνεται στις επιμέρους υποενότητες "Σχετικό έργο" του επόμενου κεφαλαίου.

1.4.1 Εξαγωγή πληροφοριών

Το αντικείμενο της εξαγωγής πληροφοριών αφορά τη συγκέντρωση, αναγνώριση και επισήμανση συγκεκριμένων κλάσεων οντοτήτων, γεγονότων ή σχέσεων από μια συλλογή κειμένων φυσικής γλώσσας καθώς και την εξαγωγή των απαραίτητων πληροφοριών που συνδέουν τα παραπάνω μεταξύ τους (Grishman, 1997). Η διαδικασία περιλαμβάνει συνήθως τη δημιουργία μιας δομημένης αναπαράστασης πληροφοριών που αντλούνται από κείμενα (πχ. με τη βοήθεια μιας βάσης δεδομένων). Η ιδέα πρωτοεισήχθη την δεκαετία του 1950 και θεωρείται έκτοτε αναπόσπαστο προαπαιτούμενο βήμα για την αυτοματοποιημένη κατανόηση κειμένων, καθώς αποσκοπεί στην διύλιση του αχανούς γνωσιακού πλούτου που βρίσκεται συχνά ανεκμετάλλευτος σε κείμενα φυσικής γλώσσας σε δομημένη μορφή, ώστε οι

εξαχθείσες οντότητες να είναι διαθέσιμες για περαιτέρω επεξεργασία. Στη σύγχρονη εποχή που κατακλύζεται από νέα πληροφορία, η δυνατότητα συγκέντρωσης πληροφοριών από πηγές του Ιστού, μέσα κοινωνικής δικτύωσης και ειδησεογραφικούς ιστοτόπους έχει ως επι το πλείστον αυτοματοποιηθεί, με χρήση τεχνικών ιχνηλάτησης που στοχεύουν στην εξαγωγή πληροφοριών από το Διαδίκτυο και τη μετατροπή τους σε μια πιο εκμεταλλεύσιμη μορφή που μπορεί να αποθηκευτεί και αναλυθεί σε μια κεντρική τοπική βάση δεδομένων, σχεσιακή ή μη. Μετά την αποθήκευση της πληροφορίας, ακολουθεί η χρήση μιας σειράς εργαλείων επεξεργασίας φυσικής γλώσσας, όπως μέθοδοι χωρισμού κειμένων σε λέξεις (tokenizers), επισημειωτές μερών του λόγου (PoS taggers), μηχανισμοί στελέχωσης κειμένου (stemmers) για την αποκοπή καταλήξεων, εργαλεία μορφοσυντακτικής λημματοποίησης (lemmatization) και τεχνικές αναγνώρισης ονοματικών οντοτήτων (named entity extraction) για ταξινόμηση των παραπάνω σε προκαθορισμένες κατηγορίες όπως ονόματα ατόμων, οργανισμών, περιοχών, ποσοτήτων, διευθύνσεων κτλ. (Al Omran and Treude, 2017).

Η ανοιχτή εξαγωγή πληροφοριών (Open Information Extraction – OIE) αποτελεί τη βασική ερευνητική γραμμή που ασχολείται με τη μετατροπή της αδόμητης πληροφορίας η οποία εκφράζεται σε ελεύθερο κείμενο, σε μια δομημένη αναπαράσταση με τη μορφή σχεσιακών πλειάδων (τριπλετών) που αποτελούνται από δύο λεκτικά σύνολα (πχ. οντότητες) και μια φράση που δηλώνει σημασιολογική σχέση μεταξύ τους, της μορφής {arg1;rel;arg2} (Martin., 2001). Οι περισσότερες παραδοσιακές προσεγγίσεις επικεντρώνονται στην αντιστοίχιση του κειμένου-εισόδου με προκαθορισμένα γραμματικοσυντακτικά μοτίβα που βασίζονται σε γλωσσολογικούς κανόνες (Tablan et al., 2003; Fader et al., 2011). Νεότερες προσπάθειες βασίζονται στη χρήση μοντέλων μηχανικής/βαθιάς μάθησης και επισημασμένων συνόλων δεδομένων (πχ. ζεύγη προτάσεων με τις τριπλέτες που παρήχθησαν από αυτές), με απώτερο στόχο την εκπαίδευση των μοντέλων στην ανακάλυψη των κανόνων εξαγωγής, στα πλαίσια επιβλεπόμενης μάθησης (supervised learning) (Yates et al., 2007; Pal, 2016).

Παρότι πλησιάζουμε στη δεύτερη δεκαετία έρευνας στον τομέα της ανοιχτής εξαγωγής πληροφοριών, η υφιστάμενη κατάσταση χαρακτηρίζεται από πολλές προκλήσεις και ανοιχτά ερευνητικά ερωτήματα (Niklaus et al., 2018). Αρχικά, η αξιολόγηση και η σύγκριση αποτελεσμάτων μεταξύ διαφορετικών OIE συστημάτων εμφανίζει πολλές δυσκολίες, εξαιτίας της έλλειψης τυποποιημένων συνόλων δεδομένων αναφοράς (benchmark datasets). Ακόμη, η συντριπτική πλειονότητα των προσεγγίσεων είτε περιορίζονται στην αγγλική γλώσσα είτε εκτείνονται σε ελάχιστες ακόμα γλώσσες υψηλών πόρων (πχ. γερμανικά, ισπανικά), στερώντας αντίστοιχες δυνατότητες για το μεγαλύτερο σύνολο των γλωσσών του

κόσμου. Τέλος, ελάχιστη προσπάθεια έχει γίνει στην κανονικοποίηση των σχεσιακών εκφράσεων και των λεκτικών συνόλων που αυτές αφορούν, με αποτέλεσμα να δυσχεραίνεται η αξιοποίηση των εξαγωγών για εφαρμογές τελικού χρήστη που περιλαμβάνουν τον εμπλουτισμό μιας γνωσιακής βάσης, εργασίες σημασιολογικής ομοιότητας, επίλυσης φαινομένων συναναφοράς κτλ. Η παρούσα εργασία στοχεύει στην αντιμετώπιση των παραπάνω προκλήσεων, με την δημιουργία ενός μηχανισμού ανοιχτής εξαγωγής πληροφοριών για την ελληνική γλώσσα, την αξιολόγησή του μέσω κατάλληλων benchmarks που υλοποιήθηκαν για να καλύψουν το αντίστοιχο κενό, καθώς και τη μελέτη μετα-επεξεργασίας και τυποποίησης των αποτελεσμάτων εξαγωγής ώστε να αυξηθεί η χρησιμότητά τους για περαιτέρω εφαρμογές.

1.4.2 Διανυσματικές αναπαραστάσεις λέξεων για εφαρμογές τελικού χρήστη

Η αναπαράσταση γλωσσικών στοιχείων με χρήση διανυσμάτων (embeddings) αποτελεί περιοχή διαχρονικού ερευνητικού ενδιαφέροντος στην ιστορία της ΕΦΓ, καθώς αποτελεί μονόδρομο για την κατανόηση πολύπλοκων γλωσσικών φαινομένων και πληροφοριών, προκειμένου αυτά να αποτυπωθούν σε μορφή αξιοποιήσιμη για εφαρμογές τελικού χρήστη, όπως ο έλεγχος σημασιολογικής ομοιότητας μεταξύ δύο εγγράφων, η αναγνώριση κειμενικής συνεπαγωγής, η μετάφραση από μια γλώσσα σε άλλη, η ανάλυση συναισθήματος, η ταξινόμηση σε θεματικές κατηγορίες κ.α. Οι διανυσματικές αναπαραστάσεις αποτελούν διατάξεις αριθμών που αντιπροσωπεύουν τις σημασιολογικές και συντακτικές πληροφορίες των λεκτικών συνόλων, σε μορφή κατανοητή από τους υπολογιστές. Μέσω αυτών, είναι δυνατή η μετατροπή μεγάλων όγκων κειμένου σε ισοδύναμα διανύσματα που μπορούν να εγκολπώσουν τις ίδιες πληροφορίες, ενώ παράλληλα ξεκλειδώνουν δυνατότητες επεξεργασίας από αλγορίθμους μηχανικής/βαθιάς μάθησης που εφαρμόζονται παραδοσιακά σε αριθμητικά δεδομένα.

Οι κλασσικές μέθοδοι παραγωγής διανυσματικών αναπαραστάσεων αφορούν την μετατροπή κειμένου σε διάνυσμα, όπου κάθε δυνατή λέξη συμβάλλει στην αύξηση του διανυσματικού χώρου και η ύπαρξη της ή μη συμβολίζεται με δυαδικό τρόπο (1 ή 0 αντίστοιχα). Παραδείγματα τέτοιων "αραιών" διανυσματικών αναπαραστάσεων αποτελούν μοντέλα One-hot Encoding, Bag-of-Words (BoW) (Harris, 1970) και Term Frequency-Inverse Document Frequency (TF-IDF) (Jones, 1972) που συναντώνται ακόμα σε απλές εφαρμογές. Ωστόσο, οι προαναφερθείσες κατηγορικές αναπαραστάσεις αποτυγχάνουν να συλλάβουν το νόημα της κάθε λέξης σε μια πρόταση, ενώ παράλληλα υπόκεινται στην "κατάρα τη διαστατικότητας" για λεξιλόγια μεγάλων διαστάσεων. Έτσι, πιο σύγχρονες

μέθοδοι αναπαράστασης υιοθέτησαν τη λογική των "πυκνών" αναπαραστάσεων που αποτελούν προϊόν εκπαίδευσης νευρωνικών δικτύων για να δημιουργήσουν το ισοδύναμο διάνυσμα κάθε στοιχείου, με τη λογική ότι λέξεις με συναφές αντικείμενο θα βρίσκονται κοντά στον πολυδιάστατο διανυσματικό χώρο. Παραδείγματα τέτοιων (στατικών) διανυσματικών αναπαραστάσεων αποτελούν οι υλοποιήσεις Word2Vec (Mikolov et al., 2013a), Continuous Bag of words (CBOW) (Mikolov et al., 2013b), Global Vectors (GloVe) (Manning et al., 2015) και FastText (Bojanowski et al., 2017). Παρά την αδιαμφισβήτητη υπεροχή τους έναντι των κατηγορικών αναπαραστάσεων, οι παραπάνω μέθοδοι επίσης αποτυγχάνουν να συλλάβουν εννοιολογικές πληροφορίες που αφορούν την αναπαράσταση των λέξεων στο εκάστοτε θεματικό πλαίσιο (πχ. περιπτώσεις αμφισημίας οδηγούν λανθασμένα στην ίδια αναπαράσταση). Για το σκοπό αυτό, οι πλέον σύγχρονες μέθοδοι αξιοποιούν γλωσσικά μοντέλα (αρχιτεκτονικές βαθιών νευρωνικών δικτύων) για να λάβουν υπόψη το δυναμικό χαρακτήρα της γλώσσας που επηρεάζει το νόημα των λέξεων στο περιεχόμενο ενός κειμένου. Παραδείγματα τέτοιων γλωσσικών μοντέλων είναι το ELMo (Peters et al., 2018) και το ULMFiT (Howard and Ruder, 2018) που βασίζονται σε ανατροφοδούμενα νευρωνικά δίκτυα, ενώ αξίζει να σημειωθεί ότι από το 2017 και μετά, το ενδιαφέρουν μονοπωλούν διανυσματικές αναπαραστάσεις που βασίζονται σε αρχιτεκτονικές Transformers (Vaswani et al., 2017), όπως το Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019b), το Generative Pre-Training (GPT) (Brown et al., 2020) και οι εκδοχές τους.

Η ανάπτυξη πολύπλοκων γλωσσικών μοντέλων για την παραγωγή διανυσματικών αναπαραστάσεων που βασίζονται στο περιεχόμενο του κειμένου άνοιξε το δρόμο για την ανάπτυξη μιας πληθώρας εφαρμογών που εφαρμόζονται στην ΕΦΓ. Ωστόσο, η αξιοποίηση αντίστοιχων μεθοδολογιών συνοδεύεται από πολλαπλές προκλήσεις και δυσκολίες που αφορούν τόσο το υπολογιστικό κομμάτι, όσο και την ύπαρξη βασικών προαπαιτούμενων για την τεχνική υλοποίησή τους. Συγκεκριμένα, η εκπαίδευση αντίστοιχων μοντέλων προϋποθέτει την ύπαρξη τεράστιου σώματος κειμένων. Αυτό συνεπάγεται τεράστιο υπολογιστικό κόστος και απαιτεί την ύπαρξη υποδομών που κατέχονται από ελάχιστες ερευνητικές ομάδες (πχ. Google, Facebook, Microsoft). Ως αποτέλεσμα, ακολουθείται η τάση χρήσης προεκπαιδευμένων μοντέλων που προσαρμόζονται για εφαρμογές τελικού χρήστη μέσω αντίστοιχων διαδικασιών προσαρμογής (finetuning). Παρότι οι διαδικασίες προσαρμογής οδηγούν σε νέα μοντέλα που ως επί το πλείστον εξυπηρετούν τις αντίστοιχες εφαρμογές τελικού χρήστη, συνήθως εξυπηρετούν μόνο γλώσσες υψηλών πόρων, καθώς αυτές περιλαμβάνονταν στο αρχικό σώμα κειμένων που χρησιμοποιήθηκε κατά την

εκπαίδευση. Επομένως, η χρησιμότητα των παραπάνω πολύγλωσσων γλωσσικών μοντέλων για γλώσσες χαμηλότερων πόρων όπως η ελληνική είναι ιδιαίτερα περιορισμένη. Στα πλαίσια της εργασίας, χρησιμοποιούνται κατάλληλες τεχνικές (πχ. απόσταξης γνώσης) για την αντιμετώπιση των παραπάνω κωλυμάτων. Ακόμη, αξίζει να σημειωθεί ότι τα παραπάνω γλωσσικά μοντέλα βασίζονται συνήθως σε τεχνικές αυτό-επιβλεπόμενης μάθησης, με αποτέλεσμα να ενσωματώνουν στις παραγόμενες διανυσματικές αναπαραστάσεις όλα τα ήδη μεροληψίας που συναντώνται στον γραπτό λόγο (πχ. εθνικές, φυλετικές, θρησκευτικές, πολιτισμικές) (Kurita et al., 2019).

1.4.3 Έλεγχος ισχυρισμών

Η εργασία ελέγχου ισχυρισμών από ελεύθερο κείμενο μπορεί να παρομοιαστεί με τις ανθρώπινες διεργασίες που αξιοποιούνται κατά τη δικονομική σκέψη και αφορούν τη λογική σύγκριση ενός ισχυρισμού με τα αντίστοιχα τεκμήρια (πηγές, μαρτυρίες, δεδομένα) που μπορούν να συμβάλουν στην επικύρωση ή στην απόρριψή του. Καθώς οι διαδικτυακές ειδησεογραφικές πηγές και τα κοινωνικά δίκτυα τείνουν να αντικαταστήσουν τα παραδοσιακά ΜΜΕ στην παροχή ενημερωτικών ειδήσεων, έχει αναδειχθεί η ανάγκη για αυτοματοποιημένο έλεγχο των γεγονότων που περιγράφονται μέσω των παραπάνω πηγών. Αν και συναντώνται πολλές διαφορετικές προσεγγίσεις για το συγκεκριμένο ερευνητικό πεδίο, τα περισσότερα συστήματα ελέγχου ισχυρισμών απαρτίζονται τουλάχιστον από ένα μηχανισμό συγκέντρωσης τεκμηρίων και έναν μηχανισμό επικύρωσης του ισχυρισμού βάσει των τεκμηρίων.

Αναφορικά με τους μηχανισμούς συγκέντρωσης τεκμηρίων, αυτοί συνήθως βασίζονται σε τεχνικές ανάκτησης πληροφοριών, που αντιμετωπίζουν το πρόβλημα ως μια εργασία κατάταξης των κορυφαίων n εγγράφων-τεκμηρίων, βάσει της σύνδεσης ονοματικών οντοτήτων και μετα-δεδομένων που συνοδεύουν τις σχετικές ιστοσελίδες. Ειδικότερα για περιπτώσεις που αφορούν συγκεκριμένες θεματικές ενότητες (πχ. βιολογία) αξιοποιούνται διανυσματικές αναπαραστάσεις λέξεων για την εύρεση του σχετικότερου υποσυνόλου τεκμηρίων. Εναλλακτικές μέθοδοι συγκέντρωσης τεκμηρίων επιστρατεύουν αντιστοιχίσεις λέξεων-κλειδιών και μέτρα ομοιότητας προτάσεων ή συνδυασμό των παραπάνω. Ωστόσο, οι περισσότερες συγκρίσεις γίνονται σε επίπεδο εγγράφου, αγνοώντας τις πληροφορίες που μπορεί να κρύβονται σε επίπεδο προτάσεων. Η επικύρωση ισχυρισμών συνήθως επιτυγχάνεται είτε έναντι κειμενικών παραπομπών όπως λήμματα της Wikipedia, είτε βασίζεται σε υπάρχουσες γνωσιακές βάσεις. Και οι δύο προσεγγίσεις θεωρούν ότι οι πηγές των τεκμηρίων είναι αξιόπιστες. Τέλος, η επικύρωση ισχυρισμών συνήθως αντιμετωπίζεται

ως μια εργασίας αναγνώρισης κειμενικής συνεπαγωγής όπου αποφασίζεται εάν το νόημα ενός κειμένου (ισχυρισμός) μπορεί να συναχθεί από ένα άλλο (τεκμήριο). Για το σκοπό αυτό συνήθως χρησιμοποιούνται γλωσσικά μοντέλα ανάλογα με αυτά που περιεγράφηκαν στην παραπάνω ενότητα.

Όπως γίνεται εύκολα αντιληπτό, οι περισσότερες μεθοδολογίες ελέγχου ισχυρισμών αφορούν κυρίως τη σύγκριση με στατικές πηγές πληροφορίας (πχ. Wikipedia), αγνοώντας το δυναμικό χαρακτήρα της πληροφορίας που παρέχεται από ειδησεογραφικές πηγές. Καθώς η παρούσα εργασία προτείνει μια μεθοδολογία που βασίζεται στην κατασκευή τεκμηρίων που προέρχονται από την περιοδική ιχνηλάτηση διαδικτυακών πηγών, καθίσταται δυνατή η αποδελτίωση της αλληλουχίας των γεγονότων που οδήγησαν στην επικύρωση ή απόρριψη ενός ισχυρισμού. Για παράδειγμα, η ανάλυση οικονομικών ειδήσεων θα μπορούσε να συσχετιστεί με τη μεταβολή της εμπορικής δραστηριότητας ενός προϊόντος, η ανάλυση πολιτικής ειδησεογραφίας θα μπορούσε να χρησιμοποιηθεί για τη συσχέτιση της με μια κοινωνική αναταραχή κ.ο.κ. Τέλος, επισημαίνεται ότι παρά την πληθώρα παράλληλων ερευνητικών γραμμών για τον έλεγχο ισχυρισμών από ελεύθερο κείμενο, καμία προσέγγιση δεν καλύπτει την ελληνική γλώσσα, κάτι που αποτελεί βασικό ερευνητικό ζήτημα της παρούσας διατριβής.

1.5 Πρωτοτυπία και απήχηση της έρευνας

Παρά την αδιαμφισβήτητη άνοδο του ερευνητικού ενδιαφέροντος σε εφαρμογές που άπτονται της επεξεργασίας φυσικής γλώσσας, η διαθεσιμότητα πόρων, συστημάτων και αλγοριθμικών τεχνικών επεξεργασίας φυσικής γλώσσας σχεδόν αποκλειστικά για τις λεγόμενες γλώσσες υψηλών πόρων αποτελεί παγιωμένη πραγματικότητα. Αντίθετα, πολλές γλώσσες που ομιλούνται και γράφονται από εκατομμύρια ανθρώπους -μεταξύ των οποίων και η Ελληνική- δε διαθέτουν παρόμοιους πόρους, ούτε μπορούν να αξιοποιήσουν αντίστοιχα συστήματα στην ίδια κλίμακα. Η ανάπτυξη σχετικών εργαλείων ανοιχτού κώδικα για την επεξεργασία και εξαγωγή συμπερασμάτων από δεδομένα των γλωσσών χαμηλών πόρων αποτελεί σημαντική πρόκληση για την επιστημονική κοινότητα, ενώ παράλληλα ευθυγραμμίζεται πλήρως με τους στόχους της παρούσας διδακτορικής εργασίας που αποσκοπεί στην κάλυψη αυτού του κενού.

Η πρωτοτυπία της παρούσας εργασίας έγκειται στην ανάπτυξη συνδυαστικών μεθοδολογιών και εργαλείων για την ελληνική γλώσσα, βελτιώνοντας σημαντικά την υφιστάμενη κατάσταση

όσον αφορά τα διαθέσιμα εργαλεία ΕΦΓ. Παράλληλα, οι μεθοδολογίες που περιγράφονται μπορούν να επεκταθούν σε άλλες γλώσσες χαμηλών πόρων, καλύπτοντας και τις τρεις κύριες θεματικές περιοχές που περιεγράφηκαν στην προηγούμενη ενότητα, και συγκεκριμένα μέσω της ανάπτυξης:

i. μηχανισμού εξαγωγής πληροφοριών από ελεύθερο κείμενο στην ελληνική γλώσσα, ο οποίος θα επιτρέπει τη δομημένη αναπαράσταση και τον εμπλουτισμό της εξαχθείσας πληροφορίας (σε μορφή τριπλετών της μορφής υποκείμενο-κατηγόρημα-αντικείμενο),
ii. γλωσσικών μοντέλων που θα βασίζονται σε διανυσματικές αναπαραστάσεις λέξεων για εφαρμογές τελικού χρήστη (πχ. σημασιολογική ομοιότητα, κειμενική συνεπαγωγή, σύνδεση οντοτήτων), με χρήση κατάλληλων τεχνικών μεταφοράς μάθησης, προκειμένου να αντιμετωπιστεί η έλλειψη σχετικών πόρων για την εκπαίδευση αντίστοιχων μοντέλων, και
iii. μηχανισμού ελέγχου κειμενικών ισχυρισμών, ο οποίος θα συνδυάζει τα παραπάνω εργαλεία προκειμένου να αποδελτιώνει ειδησεογραφικές πηγές και να επιλέγει τα σχετικότερα αποσπάσματα αυτών, αξιοποιώντας τα ως τεκμήρια για την αξιολόγηση των ισχυρισμών ενός χρήστη σε σημασιολογικό επίπεδο.

Οι δύο κύριοι μηχανισμοί που θα αναπτυχθούν στα επόμενα κεφάλαια της εργασίας, αφορούν α) την εξόρυξη πληροφοριών από ελεύθερο κείμενο και β) τη χρήση τους για την επικύρωση ισχυρισμών τελικού χρήστη από ειδησεογραφικές πηγές. Ο συνδυασμός τους αναμένεται να οδηγήσει σε μια ολοκληρωμένη πλατφόρμα κατανόησης φυσικής γλώσσας, η οποία θα αξιοποιεί μηχανισμούς άντλησης και προεπεξεργασίας κειμένων από ελληνικές πηγές του Ιστού, ανάλυσης και επεξεργασίας τους για εξαγωγή πληροφοριών, εύρεση λανθανουσών συσχετίσεων μεταξύ τους και έλεγχο της σημασιολογικής ομοιότητας των κειμένων που τις περιέχουν με τον ισχυρισμό/υπόθεση προς εξέταση. Όπως θα καταστεί σαφές από την βιβλιογραφική ανασκόπηση του Κεφαλαίου 2, μια μεθοδολογία διερευνητικής ανάλυσης ελεύθερου κειμένου που θα επιτρέπει τη δομημένη αναπαράσταση της πληροφορίας καθώς και τη συσχέτισή τους για την κατασκευή τεκμηρίων που θα επικυρώνουν ή θα αντικρούουν τον ισχυρισμό ενός χρήστη αποτελεί πρωτοποριακή ερευνητική δραστηριότητα τόσο για την ελληνική όσο και για τη διεθνή ερευνητική κοινότητα. Παράλληλα, η παρούσα προσέγγιση εισάγει την καινοτομία της συνεργατικής λειτουργίας ενός αριθμού μεθόδων από διαφορετικά επιστημονικά και τεχνολογικά πεδία (πχ. μηχανική μάθηση, γλωσσολογία), ώστε να επιτύχει την αποτύπωση συσχετίσεων που περιγράφουν τη συμπεριφορά των καταγεγραμμένων οντοτήτων.

Τα μακροπρόθεσμα οφέλη που θα προκύψουν από την καταγραφή και κατανόηση της αλληλουχίας των διασυνδέσεων αυτών, εκτείνονται σε πλήθος διαφορετικών εκφάνσεων της ανθρώπινης δραστηριότητας. Ενδεικτικά, σε πολιτικό-οικονομικό πλαίσιο, η εκπαίδευση ενός τέτοιου μοντέλου θα μπορούσε να παράσχει τη δυνατότητα πρόβλεψης της μεταβολής της κοινής γνώμης, σε σχέση με την ανακοίνωση κυβερνητικών μέτρων που αφορούν την εθνική οικονομία. Σε κοινωνικό-ψυχολογικό επίπεδο, η άντληση δεδομένων από δημόσιες διαδικτυακές συζητήσεις που αφορούν πχ. ένα εμπορικό προϊόν, θα μπορούσε να τροφοδοτήσει ένα μοντέλο μέτρησης της γενικής ικανοποίησης των χρηστών. Αντίστοιχα, στον τομέα της ασφάλειας, χάρη στη συγκέντρωση και ανάλυση σχετικού ενημερωτικού υλικού θα μπορούσε να υπολογιστεί το ενδεχόμενο ξεσπάσματος κινητοποιήσεων, αύξησης της εγκληματικότητας σε μια περιοχή ή ακόμα και της πιθανότητας εκτέλεσης εχθροπραξιών μεταξύ κρατών, με βάση προηγούμενα παρόμοια γεγονότα τα οποία συμμετείχαν στην εκπαίδευση του μοντέλου και οδήγησαν στο ίδιο αποτέλεσμα. Ενδεικτικά παραδείγματα αξιοποίησης της αποκτηθείσας γνώσης για έλεγχο αντίστοιχων υποθέσεων παρατίθενται στο Κεφάλαιο 4.

1.6 Δομή της διατριβής

Τα επόμενα κεφάλαια της διατριβής ακολουθούν την εξής δομή:

- Στο Κεφάλαιο 2 περιγράφεται το απαιτούμενο θεωρητικό υπόβαθρο (ορισμοί, παραδείγματα χρήσης, μαθηματικό/θεωρητικό πλαίσιο, σχετικό έργο) για την κατανόηση των μεθόδων ΕΦΓ που αξιοποιήθηκαν για την υλοποίηση αντίστοιχων μοντέλων βαθιάς μάθησης για την ελληνική γλώσσα.
- Στο Κεφάλαιο 3 αναπτύσσεται η μεθοδολογία, οι αρχιτεκτονικές επιλογές και οι σχεδιαστικές λεπτομέρειες που συνοδεύουν τους δύο μηχανισμούς οι οποίοι εκπληρώνουν τους στόχους της διατριβής.
- Το Κεφάλαιο 4 αναφέρεται στην τεχνική υλοποίηση των μηχανισμών που αναπτύχθηκαν στο Κεφάλαιο 3, περιλαμβάνοντας τις τεχνολογικές επιλογές, τις παραμέτρους διαμόρφωσης και παραθέτοντας αποτελέσματα από τη χρήση και συγκριτική αξιολόγησή τους σε σχετικά benchmarks και περιπτώσεις χρήσης.
- Το Κεφάλαιο 5 ολοκληρώνει την παρούσα διατριβή, με επισκόπηση των βασικών συμπερασμάτων που προέκυψαν κατά την εκπόνησή της και προτείνοντας προτάσεις για μελλοντικές επεκτάσεις.

2 Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή

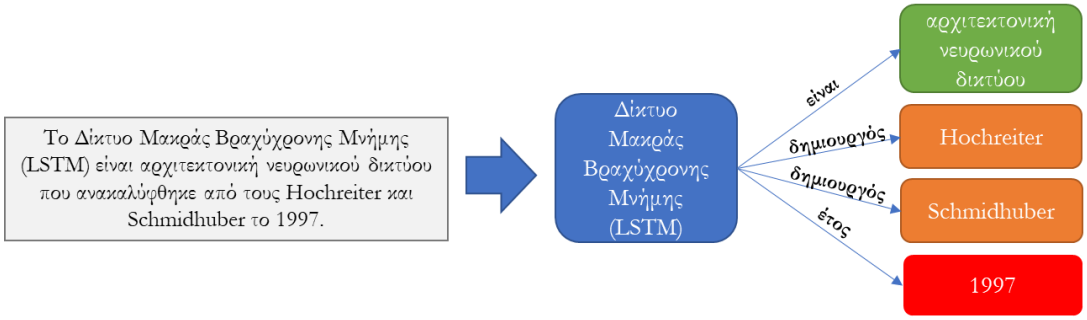
Στο παρόν κεφάλαιο παρατίθεται το απαραίτητο θεωρητικό πλαίσιο για την κατανόηση των τεχνικών επεξεργασίας φυσικής γλώσσας που σχεδιάζονται και υλοποιούνται στα επόμενα κεφάλαια της διδακτορικής εργασίας. Συγκεκριμένα, αφιερώνεται μια ενότητα για καθεμία από τις διακριτές εργασίες που οδήγησαν στην ανάπτυξη αντίστοιχων μεθοδολογιών ή μοντέλων για την ελληνική γλώσσα, οι οποίες άπτονται της εξαγωγής πληροφοριών, της μηχανικής μετάφρασης, της σύνδεσης οντοτήτων, της επίλυσης συναναφορών, της αυτόματης συνόψισης κειμένου, του ελέγχου σημασιολογικής ομοιότητας και κειμενικής συνεπαγωγής, της επικύρωσης ισχυρισμών και της αυτόματης αναγνώριση ομιλίας.

Κάθε ενότητα ακολουθεί πανομοιότυπη δομή και είναι χωρισμένη σε τέσσερις υποενότητες, με σκοπό να διευκολύνει τον αναγνώστη να εισαχθεί στο εννοιολογικό πλαίσιο του εκάστοτε αντικειμένου: Αρχικά παρατίθενται οι απαραίτητοι ορισμοί, δίνονται βασικά παραδείγματα χρήσης και περιγράφονται βασικά στοιχεία της κάθε εργασίας, ενώ στην δεύτερη ενότητα γίνεται αναφορά στη χρησιμότητά της, όπως αυτή αναδεικνύεται σε ποικίλα πεδία εφαρμογής. Η τρίτη υποενότητα είναι αφιερωμένη στην περιγραφή του μαθηματικού και τεχνικού υποβάθρου της στοχευμένης δραστηριότητας που υλοποιείται στα πλαίσια της εργασίας (πχ. στην περιγραφή συγκεκριμένης νευρωνικής αρχιτεκτονικής που χρησιμοποιήθηκε για την εκπαίδευση κάποιου μοντέλου), ενώ η τελευταία υποενότητα αποτελεί επισκόπηση του σχετικού έργου που συναντάται στη διεθνή βιβλιογραφία.

2.2 Εξαγωγή πληροφοριών

2.2.1 Ορισμός και βασικά στοιχεία

Το αντικείμενο της εξαγωγής πληροφοριών (Information Extraction) συνίσταται στην αυτόματη μετατροπή της αδόμητης πληροφορίας σε μια δομημένη αναπαράσταση, μέσω της χρήσης σχεσιακών ν-πλειάδων (n-tuples) της μορφής {ent1 ; rel ; ent2} οι οποίες αποτελούνται από ένα σύνολο οντοτήτων (entities) και από μια φράση που υποδηλώνει μια σημασιολογική σχέση (relationship) μεταξύ τους (Jurafsky and Martin, 2009).



Σχήμα 1 Παράδειγμα εξαγωγής πληροφοριών από φυσική γλώσσα

Στις περισσότερες περιπτώσεις, το εγχείρημα επικεντρώνεται στην επεξεργασία κειμένων φυσικής γλώσσας (Σχήμα 1), ωστόσο οι πρόσφατες εξελίξεις στην επεξεργασία πολυμέσων, όπως ο αυτόματος σχολιασμός και η εξαγωγή περιεχομένου από εικόνες/ήχο/βίντεο θα μπορούσαν να θεωρηθούν ως σχετικοί ερευνητικοί άξονες. Σε κάθε περίπτωση, η εξαγωγή πληροφοριών αποτελεί γενικότερο όρο που περικλείει πιο εστιασμένες δραστηριότητες όπως την αναγνώριση ονοματικών οντοτήτων (named-entity recognition) για την ταυτοποίηση και ταξινόμησή τους σε ένα κείμενο (Nadeau and Sekine, 2007), την εξαγωγή συσχετίσεων (relation extraction) για την εύρεση των σημασιολογικών σχέσεων μεταξύ οντοτήτων (Hobbs and Riloff, 2010), την εξαγωγή γεγονότων (event extraction) για την εύρεση γεγονότων στα οποία συμμετέχουν οι προαναφερθείσες οντότητες (Xiang and Wang, 2019), καθώς την εξαγωγή σχέσεων αιτίου-αιτιατού (cause-effect extraction) για τη συσχέτιση δύο γεγονότων στα οποία το ένα δρα ως αίτιο του δεύτερου και το δεύτερο ως αποτέλεσμα του πρώτου (Gopalan and Devi, 2017).

2.2.2 Χρησιμότητα και εφαρμογές

Η σημασία της εξαγωγής πληροφοριών στη σύγχρονη εποχή σχετίζεται με τον αυξανόμενο όγκο πληροφοριών που διατίθενται σε μη δομημένη μορφή. Ο Tim Berners-Lee, εφευρέτης του παγκόσμιου ιστού, έχει αναφερθεί στο υπάρχον Διαδίκτυο ως "ιστό των εγγράφων"

("web of documents") και υποστηρίζει ότι περισσότερο από το διαδικτυακό περιεχόμενο θα πρέπει να καταστεί διαθέσιμο σε μορφή αξιοποιήσιμων δεδομένων ("web of data"). Μέχρι να συμβεί αυτό, το Διαδίκτυο θα συνεχίσει αποτελείται σε μεγάλο βαθμό από μη δομημένα έγγραφα περιορισμένης χρησιμότητας καθώς αυτά δε διαθέτουν σημασιολογικά μεταδεδομένα (Bizer et al., 2009). Αντιθέτως, μέσω του μετασχηματισμού ενός εγγράφου σε δομημένη αναπαράσταση που ακολουθεί ένα προκαθορισμένο σχεσιακό μοντέλο, η περιεχόμενη σε αυτό γνώση καθίσταται πιο προσιτή για αυτοματοποιημένη επεξεργασία. Χαρακτηριστικά παραδείγματα αξιοποίησης αντίστοιχων τεχνικών αποτελεί η χρήση ευφυών πρακτόρων για την παρακολούθηση ροής δεδομένων ειδήσεων και η σάρωση ενός συνόλου εγγράφων για τη συμπλήρωση μιας βάσης δεδομένων με τις εξαχθείσες πληροφορίες.

Η εξαγωγή πληροφοριών μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα πηγών που εκτείνεται από μηνύματα ηλεκτρονικού ταχυδρομείου και ιστοσελίδες έως τεχνικές αναφορές, κλινικά δεδομένα, νομικά έγγραφα και επιστημονικές εργασίες. Τυπικά παραδείγματα εφαρμογών συναντώνται στους παρακάτω τομείς:

- Συσσώρευση βασικών τμημάτων γνώσης από πολλαπλές πηγές για εφαρμογές επιχειρηματικής ευφυίας όπως την παροχή πληροφοριών σε πελάτες και την τροφοδότηση στατιστικών μοντέλων ή αντίστοιχων εργαλείων (Sage et al., 2021; Saggion et al., 2007; Maynard et al., 2007)
- Αυτοματοποιημένη κατασκευή γνωσιακών βάσεων από επιστημονικές δημοσιεύσεις, εξαγωγή λέξεων-κλειδιών και ομαδοποίηση σχετικών εργασιών (Tshitoyan et al., 2019; Luan et al., 2017; Groth et al., 2018).
- Παρακολούθηση ειδησεογραφικών πηγών για την έγκαιρη πρόβλεψη καταστροφών ή την βελτίωση της επίγνωσης της κατάστασης (Verma et al., 2011; Hernandez-Suarez et al., 2019; Rossi et al., 2018).
- Βιοϊατρική έρευνα για ανακάλυψη φαρμάκων, ανεπιθύμητων ενεργειών και την αυτοματοποιημένη ανάλυση κλινικών δοκιμών (Zhang et al., 2021b; Ramponi et al., 2020; Pingali et al., 2021)

2.2.3 Ανοιχτή εξαγωγή πληροφοριών

Η ανοιχτή εξαγωγή πληροφοριών (Open Information Extraction – OIE), η οποία αποτελεί έναν από τους βασικούς ερευνητικούς άξονες της παρούσας εργασίας, στοχεύει στην εξαγωγή σχεσιακών πλειάδων –συνήθως σχέσεων μεταξύ δύο οντοτήτων– από απλό κείμενο, χωρίς επίβλεψη (Angeli et al., 2015). Η βασική διαφορά της με την κλειστή

εξαγωγή πληροφοριών (Closed Information Extraction), η οποία συναντάται και ως εξαγωγή πληροφοριών βασισμένη σε οντολογίες (Ontology-Based Information Extraction - OBIE) είναι ότι το εννοιολογικό σχήμα (τύποι και χαρακτηριστικά οντοτήτων, είδος συσχετίσεων, πληθικότητα κτλ.) δεν χρειάζεται να καθοριστεί εκ των προτέρων ούτε υπόκειται στους περιορισμούς μιας γνωσιακής βάσης. Αντιθέτως, το όνομα μιας σχέσης είναι απλώς το κείμενο που συνδέει δύο οντότητες. Αυτό έχει ως αποτέλεσμα την μετατροπή φυσικού κειμένου σε μια δομημένη αναπαράσταση, η οποία έχει συνήθως τη μορφή τριπλέτας υποκειμένου-κατηγορήματος-αντικειμένου (subject-predicate-object triple). Ένα τέτοιο παράδειγμα ακολουθεί παρακάτω.

Έστω η πρόταση:

Ο Αριστοτέλης **ήταν** αρχαίος Έλληνας φιλόσοφος και επιστήμονας που **γεννήθηκε** στα Στάγειρα της Χαλκιδικής, στην Μακεδονία.

Μέσω της εισαγωγής της παραπάνω πρότασης σε ένα OIE σύστημα, αναμένεται η παραγωγή τριπλετών subject-predicate-object {S;V;O}, όπως οι ακόλουθες:

{Αριστοτέλης ; ήταν ; αρχαίος Έλληνας φιλόσοφος}

{Αριστοτέλης ; ήταν ; επιστήμονας }

{Αριστοτέλης ; γεννήθηκε ; Στάγειρα της Χαλκιδικής}

{Αριστοτέλης ; γεννήθηκε ; Μακεδονία}

Σημειώνεται ότι το πλήθος και η πολυπλοκότητα των εξαχθεισών τριπλετών εξαρτάται από τους τρόπους με τους οποίους ο μηχανισμός εξαγωγής χειρίζεται τα αντίστοιχα γραμματικά και συντακτικά φαινόμενα (π.χ. σύνδεσμοι, αντωνυμίες, προσδιορισμοί, αμετάβατα ρήματα, επαυξημένες προτάσεις, σύνθετες προτάσεις με παρατακτικές συνδέσεις κτλ.). Για παράδειγμα, εξίσου αποδεκτά αποτελέσματα είναι και τα παρακάτω:

{Αριστοτέλης ; ήταν ; αρχαίος Έλληνας φιλόσοφος και επιστήμονας}

{Αριστοτέλης ; ήταν ; Έλληνας}

Βάσει της παραπάνω αναπαράστασης, η εξαγωγή πληροφορίας από ένα OIE σύστημα μπορεί να θεωρηθεί ως μια αναπαράσταση ενός πιθανού γεγονότος, καθώς τα στοιχεία της δεν συνδέονται με κάποια γνωσιακή βάση, και ως εκ τούτου δεν είναι τεκμηριωμένα (Liu et al., 2020). Για το λόγο αυτό, η ανοιχτή εξαγωγή πληροφοριών αποτελεί συνήθως το πρώτο βήμα για ένα ευρύ φάσμα εργασιών βαθύτερης κατανόησης κειμένου όπως η εξαγωγή συσχετίσεων, ο εμπλουτισμός γνωσιακών βάσεων και η ανάθεση σημασιολογικών ρόλων

(semantic role labelling), οδηγώντας στη δημιουργία σημασιολογικών τριπλετών και διασυνδεδεμένων δεδομένων με χρήση κατάλληλων προτύπων (π.χ. RDF[1]) και γλωσσών οντολογίας (π.χ. OWL [2]). Παράλληλα, οι εξαχθείσες τριπλέτες μπορούν να χρησιμοποιηθούν απευθείας για εφαρμογές τελικού χρήστη όπως η αποδελτίωση εγγράφων και η δομημένη αναζήτηση λέξεων-κλειδιών από φυσικό κείμενο (Prathap Reddy M et al., 2018).

2.2.4 Σχετικό έργο

Οι περισσότερες προσεγγίσεις ανοιχτής εξαγωγής πληροφοριών στοχεύουν είτε στον εντοπισμό γλωσσικών μοτίβων μέσω της εφαρμογής γλωσσολογικών κανόνων (rule-based systems), είτε στην αυτόματη εκμάθηση από επισημασμένα σύνολα δεδομένων (learning-based systems).

Οι προσεγγίσεις που βασίζονται σε κανόνες επικεντρώνονται σε γραμματικοσυντακτικούς περιορισμούς που εκφράζονται ως κανονικές εκφράσεις (regular expressions) βασιζόμενες σε επισήμανση μερών του λόγου (Part-of-Speech tagging) (Fader et al., 2011; Mesquita et al., 2013). Παρόμοια συστήματα βασίζονται στην ανίχνευση βασικών ειδών προτάσεων που συναντώνται συνηθέστερα σε μία γλώσσα, κάνοντας χρήση των συντακτικών ιδιοτήτων της (π.χ. αναφορικές ονοματικές ή επιρρηματικές προτάσεις) για τον προσδιορισμό των επιμέρους συνιστωσών τους, αναδιαρθρώνοντας έτσι μεγαλύτερες σύνθετες προτάσεις σε πολλές απλές (Del Corro and Gemulla, 2013b; Angeli et al., 2015).

Νεότερες προσεγγίσεις που εστιάζουν στην χρήση τεχνικών μηχανικής μάθησης συνήθως αξιοποιούν επισημασμένες πηγές δεδομένων (π.χ. Wikipedia), είτε για να εκπαιδεύσουν ταξινομητές (Banko et al., 2007; Wu and Weld, 2010), είτε ως πηγή ενός μεγάλου σετ επαναδειγματοληψίας (bootstrapping) πάνω στο οποίο θα εκπαιδευτεί ένας αλγόριθμος επισήμανσης μερών του λόγου (Weld et al., 2008). Η δημιουργία και διάχυση επισημασμένων συνόλων κειμένου (annotated corpora) για την εκπαίδευση και αξιολόγηση συστημάτων OIE, άνοιξε το δρόμο για τη χρήση επιβλεπόμενων νευρωνικών μοντέλων, συμβάλλοντας στη γενική εξέλιξη της τεχνολογίας αιχμής μέσω ανάλογων δραστηριοτήτων (Yang et al., 2017; Liu et al., 2018).

[1] https://www.w3.org/RDF/

[2] https://www.w3.org/OWL/

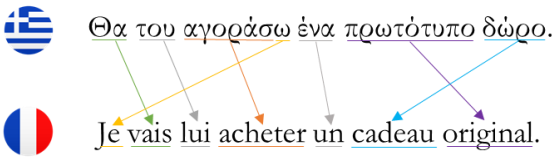
Οι πιο πρόσφατες προσεγγίσεις επεκτείνουν τη χρήση νευρωνικών δικτύων ανάθεσης ετικετών BIO (BIO tagging) που χρησιμοποιούνται για την ανάλυση και κατάτμηση προτάσεων μέσω της επισήμανσης σημασιολογικών ρόλων, όπου το B (beginning) εκχωρείται στο αρχικό λεκτικό στοιχείο (token) μιας ακολουθίας λέξεων που περιλαμβάνει μια οντότητα, το I (interior) στα στοιχεία του εσωτερικού της και το O (outside) στα υπόλοιπα στοιχεία. Τα παραπάνω μοντέλα αξιοποιούν τις διανυσματικές αναπαραστάσεις λέξεων (word embeddings) κάθε πρότασης με χρήση βαθιών νευρωνικών δικτύων (π.χ. bi-LSTMs) προκειμένου να εξάγουν κατανομές πιθανότητας αναφορικά με τις πιθανές ετικέτες BIO για κάθε λέξη (He et al., 2017; Stanovsky et al., 2018). Η χρήση μηχανισμών προσοχής (attention mechanisms) έχει επίσης αναδειχθεί ως μέσο βελτίωσης της ποιότητας των εξαχθεισών τριπλετών (Han and Wang, 2021; Kolluru et al., 2020b), ιδιαίτερα σε περιπτώσεις περίπλοκων, σύνθετων προτάσεων όπου οι συνήθεις τεχνικές OIE είτε αντιμετωπίζουν προβλήματα εντοπισμού των οντοτήτων και των σχέσεων μεταξύ τους, είτε λειτουργούν πλεοναστικά επαναλαμβάνοντας τα ίδια αποτελέσματα.

Από τα παραπάνω, γίνεται εύκολα αντιληπτό ότι η εφαρμογή των περισσότερων συστημάτων εξαγωγής πληροφοριών περιορίζεται στις γλώσσες και στους θεματικούς τομείς για τους οποίους σχεδιάστηκαν αρχικά. Για το λόγο αυτό, οι τελευταίες εξελίξεις στον τομέα εστιάζουν στην εκπαίδευση τεχνικών μεταφοράς μάθησης (transfer learning) (Ro et al., 2020a) ή μηχανικής μετάφρασης (Sheng et al., 2017) για την επέκταση της χρήσης ήδη εκπαιδευμένων μοντέλων από γλώσσες υψηλών πόρων (high-resource languages) σε γλώσσες χαμηλών πόρων (low-resource languages), χωρίς τη ανάγκη επισημασμένων δεδομένων εκπαίδευσης.

2.3 Μηχανική μετάφραση

2.3.1 Ορισμός και βασικά στοιχεία

Η μηχανική μετάφραση (Machine Translation) αποτελεί πεδίο της επεξεργασίας φυσικής γλώσσας που διερευνά τη χρήση αλγορίθμων για την αυτοματοποιημένη μετάφραση κειμένου (ή ομιλίας) από τη μια γλώσσα στην άλλη. Μια τυπική εργασία μηχανικής μετάφρασης λαμβάνει ως είσοδο μια ακολουθία συμβόλων γλώσσας Α και ένας αλγόριθμος αναλαμβάνει να τη μετατρέψει σε μια ακολουθία συμβόλων σε γλώσσα Β, όπως φαίνεται στο Σχήμα 2 (Goodfellow et al., 2016).



Σχήμα 2 Παράδειγμα Μηχανικής Μετάφρασης

Στην απλούστερη υλοποίησή της, η μηχανική μετάφραση πραγματοποιεί κατά βάση μηχανική αντικατάσταση λέξεων, ωστόσο μια τέτοια εργασία από μόνη της σπάνια οδηγεί σε αξιόπιστα αποτελέσματα εξαιτίας της ανάγκης αναγνώρισης του νοήματος ολόκληρων εκφράσεων και της αντιστοίχισής τους σε αντίστοιχες εκφράσεις στη γλώσσα-στόχο. Το έργο αυτό δυσχεραίνεται από το γεγονός ότι δεν έχουν όλες οι λέξεις σε μια γλώσσα ισοδύναμες λέξεις σε μια άλλη γλώσσα, ενώ πολλές λέξεις έχουν περισσότερες από μία σημασίες. Ακόμη, ενώ ορισμένες πτυχές της ανθρώπινης γλώσσας φαίνεται να είναι καθολικές και ισχύουν για τις περισσότερες γλώσσες, όπως η ύπαρξη δομικών γλωσσικών στοιχείων (υποκείμενα, ρήματα, αντικείμενα, επίθετα, αντωνυμίες κτλ.), συχνά διακρίνονται ιδιοσυγκρασιακές και λεξιλογικές διαφορές που πρέπει να αντιμετωπιστούν μία προς μία (π.χ. η λέξη "τιμή" έχει διαφορετικό νόημα και μετάφραση ανάλογα με τη χρήση της στο κείμενο) (Dryer and Haspelmath, 2013). Επιπλέον, η διαρθρωτική ποικιλομορφία των γλωσσών του κόσμου (π.χ. άλλες γλώσσες θέτουν το ρήμα μετά το άμεσο αντικείμενο, σε αντίθεση με την ελληνική) και λοιπές μορφολογικές διαφορές (π.χ. ιδεογραφικά συστήματα γραφής όπου κάθε σύμβολο αντιστοιχεί σε ένα αντικείμενο/ιδέα) καθιστούν αναγκαία την δημιουργία πολύπλοκων αλγορίθμων μετάφρασης που ενθυλακώνουν το νόημα ολόκληρων εκφράσεων και όχι κάθε λέξης χωριστά (Jurafsky and Martin, 2009).

Οι μεθοδολογίες μηχανικής μετάφρασης χωρίζονται σε τρείς κύριες κατηγορίες (Koponen et al., 2019):

- τη μηχανική μετάφραση βασισμένη σε κανόνες (Rule-Based Machine Translation – RBMT) που βασίζεται στην ύπαρξη πλήρους λεξιλογίου και προκαθορισμένων κανόνων για τη σύνδεση και την ορθή αποτύπωση των μορφολογικών χαρακτηριστικών (π.χ. καταλήξεις λέξεων) του μεταφρασμένου κειμένου,
- τη στατιστική μηχανική μετάφραση (Statistical Machine Translation – SMT), η οποία αναλύει δίγλωσσα σύνολα κειμένων προκειμένου να δημιουργήσει στατιστικά μοντέλα που μεταφράζουν ακολουθίες λέξεων από τη μία γλώσσα στην άλλη, και
- τη νευρωνική μηχανική μετάφραση (Neural Machine Translation – NMT), που εκμεταλλεύεται τις δυνατότητες εκπαίδευσης των βαθιών νευρωνικών δικτύων για να

αναγνωρίζει μοτίβα σε ακολουθίες λέξεων τα οποία μετασχηματίζει σε διανυσματικές αναπαραστάσεις (embeddings), προκειμένου να προβλέψει την πιθανότητα μιας αντίστοιχης ακολουθίας στη γλώσσα-στόχο.

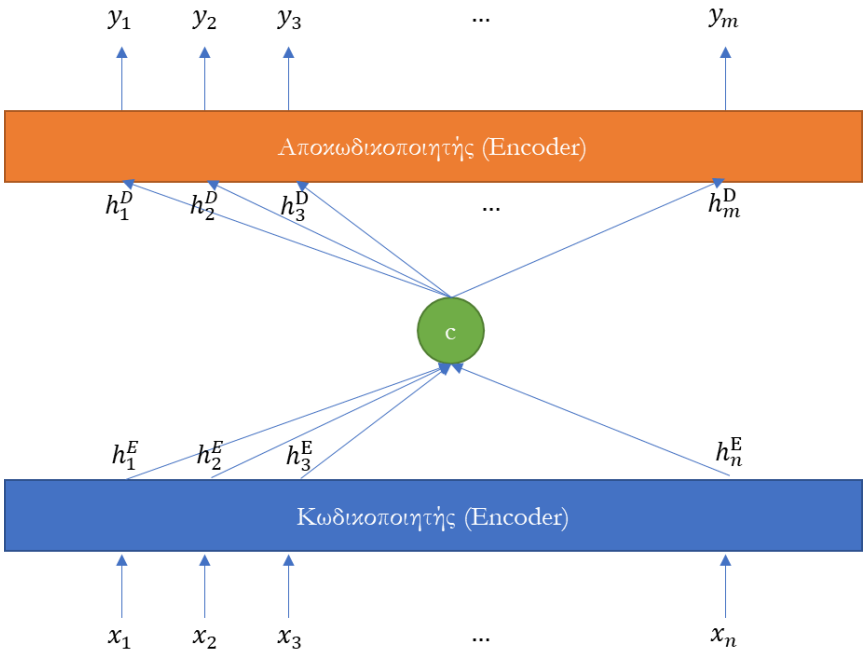
2.3.2 Χρησιμότητα και εφαρμογές

Η χρησιμότητα της μηχανικής μετάφρασης επιβεβαιώνεται από παραδείγματα που προέρχονται από ποικίλα θεματικά πεδία όπως οι νομικές επιστήμες, η οικονομία, οι αμυντικές εφαρμογές κ.λπ. Ενδεικτικά αναφέρονται εφαρμογές που αφορούν επιχειρηματικές δραστηριότητες όπως η μετάφραση τεχνικών εγγράφων και οδηγιών, γραπτής επικοινωνίας και ανακοινώσεων από πολυεθνικές εταιρείες που απασχολούν προσωπικό από διαφορετικές χώρες (Chakravarthi et al., 2019), καθώς και η γλωσσική τοπικοποίηση μεγάλων όγκων διαδικτυακού περιεχομένου, όπως σελίδες και περιγραφές προϊόντων για αύξηση της προσβασιμότητας από το πελατειακό κοινό (Chen et al., 2016). Αντίστοιχες εφαρμογές συναντώνται και στον τομέα της υγείας για την βελτίωση της πολύγλωσσης επικοινωνίας και την παροχή πρόσβασης σε πληροφοριακούς πόρους που διατίθενται σε περιορισμένη ποικιλία γλωσσών (Dew et al., 2018; Kirchhoff et al., 2011). Στον οικονομικό τομέα, συμβάλλει στην μετάφραση οικονομικών κειμένων (π.χ. ισολογισμοί, οικονομικές καταστάσεις) (Nunziatini, 2019), ενώ χρησιμοποιείται και για την μετάφραση νομικών κειμένων που αφορούν την κατοχύρωση πατεντών ή πνευματικών δικαιωμάτων (Elnaggar et al., 2018). Στον αμυντικό τομέα, η χρήση συστημάτων μηχανικής μετάφρασης σε συνδυασμό με τεχνολογίες αναγνώρισης φωνής κατά την ανάπτυξη στρατευμάτων, επιτρέπει την ευκολότερη επικοινωνία τους με τοπικούς πληθυσμούς (Pedtke, 1997). Τέλος, δε μπορούν να αγνοηθούν οι διευκολύνσεις που παρέχονται στους τελικούς χρήστες μέσω της ελεύθερης χρήσης τέτοιων υπηρεσιών (π.χ. Google Translate) όσον αφορά την πρόσβαση σε ξενόγλωσσο περιεχόμενο γενικού ενδιαφέροντος. Σημειώνεται ότι εκτός από τα άμεσα οφέλη που προσφέρει στους παραπάνω τομείς, η μηχανική μετάφραση διαδραματίζει κομβικό ρόλο και ως ενδιάμεσο βήμα για άλλες εφαρμογές επεξεργασίας φυσικής γλώσσας (π.χ. εξαγωγή πληροφοριών) στοχεύοντας γλώσσες για τις οποίες δε διατίθενται αντίστοιχα επισημασμένα δεδομένα που επιτρέπουν την απευθείας εκπαίδευση αντίστοιχων αλγορίθμων (Shterionov et al., 2020; Faruqui and Kumar, 2015).

2.3.3 Μηχανική μετάφραση με αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή

Τα νευρωνικά δίκτυα με αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή (encoder-decoder networks) αποτελούν μοντέλα ικανά να παράξουν εννοιολογικά συναφείς μεταφράσεις ενός κειμένου εισόδου και αποτελούν την τεχνολογία αιχμής στο πεδίο της νευρωνικής μηχανικής μάθησης.

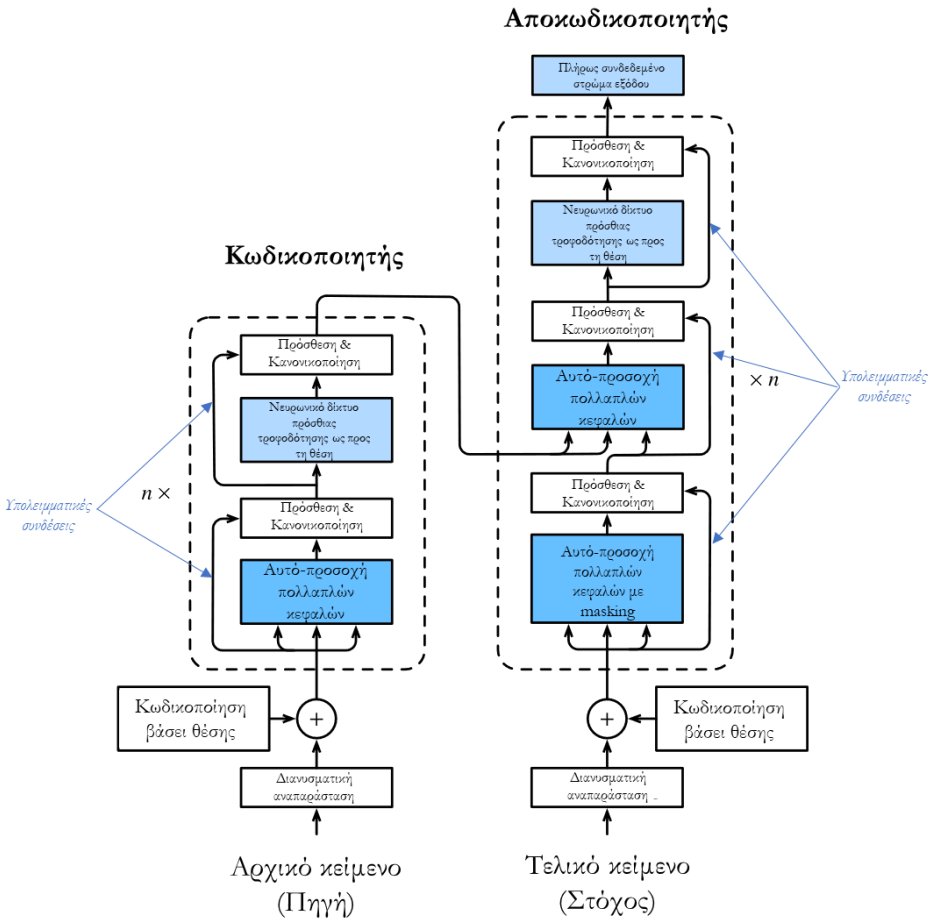
Η βασική τους ιδέα στηρίζεται στην χρήση ενός δικτύου κωδικοποιητή (encoder) που λαμβάνει μια ακολουθία εισόδου $x_1, x_2, \dots, x_n \in X$ και δημιουργεί μια διανυσματική αναπαράστασή της $h_1^E, h_2^E, \dots, h_n^E \in H^E$. Στη συνέχεια, ένα διάνυσμα c που αποτελεί συνάρτηση της προαναφερθείσας διανυσματικής αναπαράστασης $\mathrm{H^E}$ χρησιμοποιείται για να μεταφέρει το νοηματικό πλαίσιο (context) της εισόδου στον αποκωδικοποιητή. Ο αποκωδικοποιητής (decoder) με τη σειρά του δέχεται το διάνυσμα c ως είσοδο και παράγει μια ακολουθία κρυφών καταστάσεων αυθαίρετου μήκους $h_1^D, h_2^D, \dots, h_m^D \in H^D$, από τις οποίες μπορεί να εξαχθεί η τελική ακολουθία εξόδου $y_1, y_2, \dots, y_m \in Y$. Η παραπάνω λογική φαίνεται στο Σχήμα 3. Σημειώνεται ότι τόσο για τον κωδικοποιητή όσο και για τον αποκωδικοποιητή είναι δυνατή η χρήση είτε συνελικτικών (CNN) και ανατροφοδοτούμενων νευρωνικών δικτύων (π.χ. LSTM, GRU), είτε μηχανισμών προσοχής (attention) οι οποίοι περιγράφονται εκτενέστερα παρακάτω.



Σχήμα 3 Τυπική Αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή

2.3.3.1 Μηχανισμοί προσοχής και η αρχιτεκτονική Transformer

Βασική προϋπόθεση για την επιτυχή λειτουργία της προαναφερθείσας αρχιτεκτονικής είναι ότι το διάνυσμα c που μεταφέρει το νοηματικό πλαίσιο (context) της εισόδου στον αποκωδικοποιητή πρέπει να διατηρεί την πληροφορία από όλη την ακολουθία εισόδου X. Η χρήση ανατροφοδοτούμενων νευρωνικών δικτύων (recurrent neural networks – RNNs) που προτάθηκε αρχικά ως μοντέλο του κωδικοποιητή εμφάνισε περιορισμούς (bottlenecks) σε περιπτώσεις μεγάλων ακολουθιών εισόδου, καθώς το διάνυσμα c δε μπορούσε να συμπεριλάβει ολόκληρη την πληροφορία (Bahdanau et al., 2015). Η μετατροπή του c σε διάνυσμα μεταβλητού μήκους σε συνδυασμό με την εισαγωγή μηχανισμών προσοχής (attention mechanisms) έλυσε το παραπάνω πρόβλημα. Το νευρωνικό μοντέλο Transformer (Vaswani et al., 2017) είναι μια αρχιτεκτονική encoder-decoder που αξιοποιεί τους παραπάνω μηχανισμούς και έθεσε τα θεμέλια για τη σύγχρονη εποχή της επεξεργασίας φυσικής γλώσσας, επιτυγχάνοντας σημαντική βελτίωση τόσο στην ποιότητα μετάφρασης, όσο και σε πολλές άλλες εργασίες ΕΦΓ, όπως περιγράφεται στις παρακάτω υποενότητες.



Σχήμα 4 Αρχιτεκτονική Transformer

Αρχικά δίνεται συνοπτική περιγραφή της αρχιτεκτονικής Transformer (Σχήμα 4), η οποία αξιοποιείται για την εκπαίδευση γλωσσικών μοντέλων που περιγράφονται στο Κεφάλαιο 4 της παρούσας εργασίας και στη συνέχεια αναλύονται οι μηχανισμοί προσοχής που χρησιμοποιούν οι παραλλαγές του.

Κωδικοποιητής: Ο Transformer αποτελείται από έναν κωδικοποιητή (encoder) 6 πανομοιότυπων στρωμάτων, καθένα εκ των οποίων περιλαμβάνει 2 υποστρώματα, με διάσταση (embedding size) ίση με 512. Το πρώτο υπόστρωμα είναι ένας μηχανισμός αυτό-προσοχής (self-attention) πολλαπλών (8) κεφαλών (multi-head) και το δεύτερο είναι ένα πλήρως συνδεδεμένο δίκτυο πρόσθιας τροφοδότησης ως προς τη θέση (position-wise fully connected feed-forward neural network) εσωτερικής διάστασης 2048. Για την αντιμετώπιση των εξαφανιζόμενων κλίσεων (vanishing gradients) κατά την εκπαίδευση, χρησιμοποιείται μια υπολειμματική σύνδεση (residual connection) γύρω από το κάθε υπόστρωμα και ακολουθεί η προσθήκη τους (add) στο τελικό στρώμα και κανονικοποίηση του αποτελέσματος (layer normalization) σε ξεχωριστό στρώμα (Add & Norm).

Αποκωδικοποιητής: Έχει τον ίδιο αριθμό στρωμάτων με τον κωδικοποιητή, όμως πέραν των 2 υποστρωμάτων σε κάθε στρώμα εισάγεται και ένα υπόστρωμα προσοχής πολλαπλών κεφαλών στην έξοδο της στοίβας του κωδικοποιητή. Επίσης χρησιμοποιούνται ίδιου τύπου υπολειμματικές συνδέσεις και κανονικοποίηση στρώματος όπως στον κωδικοποιητή. Ειδικά για εργασίες μηχανικής μετάφρασης, η στοίβα αυτό-προσοχής είναι κατάλληλα τροποποιημένη ώστε να παρακολουθεί μόνο τις ήδη γνωστές θέσεις εξόδου με χρήση συμβόλων κάλυψης (masking tokens).

Μηχανισμός προσοχής: Ο μηχανισμός προσοχής (Bahdanau et al., 2015) περιλαμβάνει επιμέρους υπολογισμούς για τη τιμή ευθυγράμμισης (alignment score), τα βάρη (weights) και το νοηματικό πλαίσιο (context vector).

Το μοντέλο ευθυγράμμισης λαμβάνει την κρυφή κατάσταση h_i^E του κωδικοποιητή και την προηγούμενη έξοδο του αποκωδικοποιητή y_{t-1} για να υπολογίσει την τιμή ευθυγράμμισης $e_{t,i}$ που υποδηλώνει πόσο καλά τα στοιχεία της ακολουθίας εισόδου ευθυγραμμίζονται με την τρέχουσα έξοδο. Το μοντέλο ευθυγράμμισης αναπαρίσταται με μια συνάρτηση $a(\cdot)$ που υλοποιείται από ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης.

$$e_{t,i} = a(y_{t-1}, h_i) \tag{2.1}$$

Τα βάρη $\alpha_{(t,i)}$ υπολογίζονται από την εφαρμογή της συνάρτησης softmax στην παραπάνω τιμή ευθυγράμμισης:

$$\alpha_{t,i} = softmax(e_{t,i}) \tag{2.2}$$

Σε κάθε χρονικό βήμα, ένα μοναδικό διάνυσμα νοηματικού πλαισίου (context vector) c_t δίνεται στον αποκωδικοποιητή, το οποίο υπολογίζεται από το σταθμισμένο άθροισμα όλων των T κρυφών καταστάσεων του κωδικοποιητή:

$$c_t = \sum_{i=1}^{T} a_{t,i} h_i \tag{2.3}$$

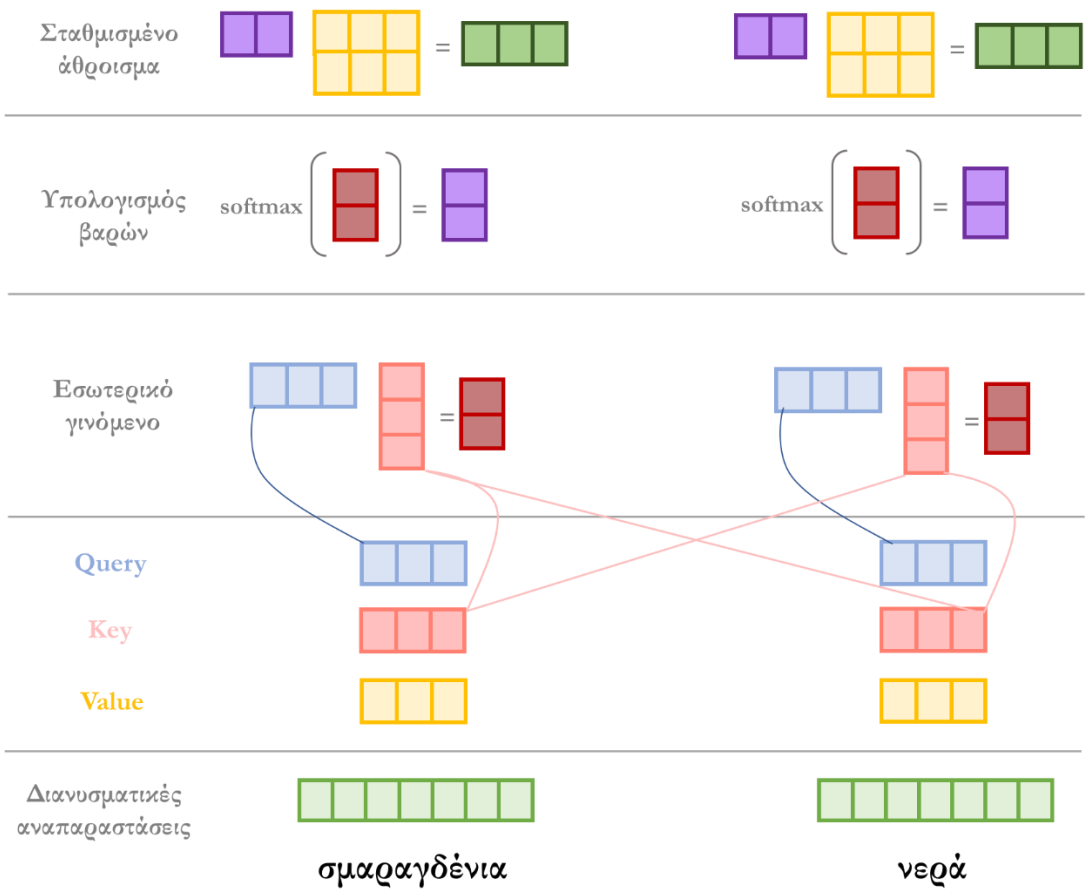
Η γενική συνάρτηση προσοχής βασίζεται στις παραπάνω σχέσεις, ωστόσο χρησιμοποιεί διαφορετική σημειογραφία και βασίζεται σε τρία βασικά στοιχεία: τα "ερωτήματα" (queries – Q), τα "κλειδιά" (keys – K) και τις "τιμές" (values – V). Δεδομένων των τριών παραπάνω διανυσμάτων (query, key, value), η συνάρτηση προσοχής αποτελεί την αντιστοίχιση ενός ερωτήματος και ενός συνόλου ζευγών κλειδιού-τιμής σε μία έξοδο. Σε σχέση με την παραπάνω περιγραφή, το q αναλογεί με την προηγούμενη έξοδο του αποκωδικοποιητή y_{t-1}, το v με την κρυφή κατάσταση h_i^E του κωδικοποιητή, ενώ το k είναι ίδιο με το v. Επομένως προκύπτουν οι παρακάτω ισοδύναμες σχέσεις:

$$e_{q,k_i} = q \cdot k_i \tag{2.4}$$

$$\alpha_{q,k_i} = softmax(e_{q,k_i}) \tag{2.5}$$

$$attention(q, K, V) = \sum_{i=1}^{T} \alpha_{q,k_i} v_{k_i} \tag{2.6}$$

Η σχηματική αναπαράσταση της παραπάνω διαδικασίας απεικονίζεται στο Σχήμα 5 με τη βοήθεια παραδείγματος. Δεδομένης της ακολουθίας "σμαραγδένια νερά", οι αρχικές διανυσματικές αναπαραστάσεις κάθε λέξης μετατρέπονται στα διανύσματα query, key και value. Στη συνέχεια, υπολογίζεται το εσωτερικό γινόμενο μεταξύ του query και όλων των key διανυσμάτων (στην προκειμένη περίπτωση για τις δύο λέξεις της ακολουθίας). Σημειώνεται ότι αυτή η διαδικασία μπορεί να παραλληλοποιηθεί για όλα τα tokens (λέξεις) της πρότασης. Το αποτέλεσμα (τιμές ευθυγράμμισης) τροφοδοτείται στη συνάρτηση softmax, για τον υπολογισμό των βαρών, ως μια κατανομή πιθανότητας των tokens της ακολουθίας. Τέλος, υπολογίζεται το σταθμισμένο άθροισμα των διανυσμάτων value βάσει των βαρών, ώστε να ληφθούν τα διανύσματα νοηματικού πλαισίου. Κατ' αυτό τον τρόπο, αν δυο λέξεις σε μια ακολουθία σχετίζονται νοηματικά μεταξύ τους (πχ. η μία αποτελεί επιθετικό προσδιορισμό της άλλης), η συσχέτιση αυτή θα αποτυπωθεί στις τιμές που περιλαμβάνει το διάνυσμα του κάθε νοηματικού πλαισίου που τις αναπαριστά.



Σχήμα 5 Σχηματική επισκόπηση μηχανισμού αυτο-προσοχής

Σημειώνεται ότι η προσοχή πολλαπλών κεφαλών (multi-head attention) βασίζεται στην παραπάνω λογική, ωστόσο αντί να εκτελείται μία μόνο λειτουργία προσοχής, προβάλλονται γραμμικά τα ερωτήματα, τα κλειδιά και οι τιμές, και εφαρμόζεται παράλληλα η συνάρτηση προσοχής, επιτρέποντας στο μοντέλο να παρακολουθεί τις πληροφορίες από διαφορετικές αναπαραστάσεις σε διαφορετικές θέσεις, κάτι που δεν είναι δυνατό με μία κεφαλή προσοχής. Στα στρώματα προσοχής κωδικοποιητή-αποκωδικοποιητή, τα ερωτήματα προέρχονται από το προηγούμενο στρώμα του αποκωδικοποιητή, ενώ τα κλειδιά και οι τιμές από την έξοδο του κωδικοποιητή. Αντίθετα, στα στρώματα αυτό-προσοχής του κωδικοποιητή (και του αποκωδικοποιητή) όλα τα ερωτήματα, κλειδιά και τιμές προέρχονται από την έξοδο του προηγούμενου στρώματος του κωδικοποιητή (και του αποκωδικοποιητή αντίστοιχα) επιτρέποντας την παρακολούθηση όλων των θέσεων στο προηγούμενο στρώμα.

2.3.4 Σχετικό έργο

Οι νευρωνικές αρχιτεκτονικές μηχανικής μετάφρασης (NMT) αποτελούν σήμερα τις κυρίαρχες προσεγγίσεις στο συγκεκριμένο πεδίο, καθώς επιτυγχάνουν υποσχόμενα αποτελέσματα που ξεπερνούν εκείνα των παραδοσιακών στατιστικών μεθόδων (SMT),

δεδομένου ενός μεγάλου συνόλου δεδομένων εκπαίδευσης (Stahlberg, 2020). Η υλοποίηση αρχιτεκτονικών κωδικοποιητή-αποκωδικοποιητή, από ανατροφοδοτούμενα (Sutskever et al., 2014; Bahdanau et al., 2015) και συνελικτικά νευρωνικά δίκτυα (Kalchbrenner et al., 2014; Gehring et al., 2017) καθώς και η χρήση μηχανισμών αυτό-προσοχής σε αρχιτεκτονικές Transformer (Vaswani et al., 2017; So et al., 2019) έχουν εξελίξει την τεχνολογία αιχμής όσον αφορά την ποιότητα των αποτελεσμάτων καθώς και την αποδοτικότητα αντίστοιχων συστημάτων, ειδικά για περιπτώσεις μορφολογικά πλούσιων γλωσσών.

Παράλληλες ερευνητικές γραμμές εστιάζουν στην μετάφραση της ποιότητας μετάφρασης μέσω της κατάλληλης κωδικοποίησης tokens, με σκοπό τη διαχείριση λέξεων εκτός λεξιλογίου (out-of-vocabulary – OOV) words, στοχεύοντας στην απουσία αμφιμονοσήμαντης (1-1) αντιστοιχίας μεταξύ της γλώσσας-πηγής και της γλώσσας-στόχου. Για το σκοπό αυτό, μέθοδοι κωδικοποίησης ζεύγους byte (byte-pair encoding – BPE) χρησιμοποιούνται για την κατάτμηση λέξεων (word segmentation) ώστε να καταστεί δυνατή η αναπαράσταση σπάνιων ή άγνωστων λέξεων ως ακολουθία από μονάδες υπο-λέξεων (subword units), αποδεικνύοντας έτσι ότι οι NMT μέθοδοι είναι κατάλληλες και σε περιπτώσεις ανοιχτού λεξιλογίου (Sennrich et al., 2016b).

Οι τελευταίες εξελίξεις στο πεδίο περιλαμβάνουν επίσης είτε προ-εκπαίδευση (pretraining) διαγλωσσικών μοντέλων (cross language models) σε πολύγλωσσο σύνολο δεδομένων (δηλαδή παράλληλο κείμενο σε πολλές γλώσσες) (Conneau and Lample, 2019), είτε την αξιοποίηση μονόγλωσσων συνόλων κειμένου για ημι-επιβλεπόμενη μάθηση μέσω αντίστροφης μετάφρασης (back-translation) (Sennrich et al., 2016a). Όσον αφορά την ελληνική γλώσσα, η οποία ανήκει στις γλώσσες χαμηλότερων πόρων, δεν διατίθενται αντίστοιχα NMT μοντέλα με αξιοσημείωτη εξαίρεση το βασιζόμενο σε Transformers EN-to-EL μοντέλο της ομάδας Helsinki NLP[3] το οποίο έχει αξιολογηθεί στο σύνολο δεδομένων αναφοράς (benchmark dataset) Tatoeba (Tiedemann and Thottingal, 2020) και με το οποίο συγκρίνεται το NMT μοντέλο που υλοποιήθηκε στα πλαίσια της παρούσας εργασίας.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-en-el

2.4 Σύνδεση οντοτήτων

2.4.1 Ορισμός και βασικά στοιχεία

Η σύνδεση οντοτήτων (Entity Linking – EL) αποτελεί υποπεδίο της επεξεργασίας φυσικής γλώσσας το οποίο αφορά τη σύνδεση ενός στοιχείου που συναντάται σε φυσικό κείμενο (λέξης ή φράσης) με μια σχετική εγγραφή μεγάλης συλλογής πληροφοριών, συνήθως μιας γνωσιακής βάσης (knowledge base) (Rao et al., 2013). Τα στοιχεία αυτά συνήθως περιλαμβάνουν αναφορές οντοτήτων (entity mentions) όπως ονόματα προσώπων, τοποθεσίες, εταιρείες, τεχνικοί όροι κτλ., ενώ η γνωσιακή βάση εξαρτάται από το πεδίο εφαρμογής (πχ. κλινικά δεδομένα, νομική ορολογία, οικονομικά στοιχεία κ.λπ.). Ειδικότερα για συστήματα σύνδεσης οντοτήτων που επικεντρώνονται σε κείμενα γενικού περιεχομένου (πχ. ειδήσεις) είναι σύνηθες να χρησιμοποιούνται γνωσιακές βάσεις που αντλούν πληροφορία από τη Wikipedia (πχ. WikiData[4], DBpedia[5]). Στην περίπτωση αυτή, κάθε σελίδα της Wikipedia που περιγράφει μια έννοια (concept) αποτελεί ξεχωριστή οντότητα, ενώ η διαδικασία σύνδεσης λέξεων με τις παραπάνω οντότητες είναι γνωστή ως "wikification" (Szymański and Naruszewicz, 2019) (Σχήμα 6).

Η Google είναι μία από τις μεγαλύτερες εταιρείες διαδικτυακών υπηρεσιών με έδρα τις ΗΠΑ. Ιδρύθηκε από τον Λάρρυ Πέιτζ και τον Σεργκέι Μπριν το 1996, όταν αυτοί έκαναν το διδακτορικό τους στο Πανεπιστήμιο Στάνφορντ.

PR	Annotation	Annotation (en)
0.0899	Πανεπιστήμιο Στάνφορντ	Stanford University
0.0825	Λάρρυ Πέιτζ	Larry Page
0.0818	Σεργκέι Μπριν	Sergey Brin
0.0737	Google	Google
0.0711	Ηνωμένες Πολιτείες Αμερικής	United States

Σχήμα 6 Παράδειγμα Σύνδεσης Οντοτήτων με γνωσιακή βάση

Οι τεχνικές σύνδεσης οντοτήτων στοχεύουν στην επίλυση της λεξικολογικής ασάφειας μίας λέξης/φράσης, καθορίζοντας το νόημά της στο κείμενο. Οι προκλήσεις που συναντώνται κατά την διαδικασία αυτή συνοψίζονται παρακάτω:

[4] https://www.wikidata.org/

[5] https://www.dbpedia.org/

- **Παραλλαγές ονομάτων**: Μια οντότητα συνήθως συναντάται με πολλές αναφορές, συμπεριλαμβανομένων συντομογραφιών (πχ. Πολυτεχνείο Κρήτης, ΠΚ, Technical University of Crete, TUC), εναλλακτικής ορθογραφίας (πχ. Μαριάνα, Μαριάννα) και ψευδωνύμων/υποκοριστικών (πχ. Νικόλαος, Νίκος). Οι μορφολογικά πλούσιες γλώσσες που περιέχουν πολλούς κλιτούς τύπους οδηγούν σε παραλλαγές που δυσχεραίνουν επιπλέον τις εργασίες σύνδεσης.
- **Αμφισημία και πολυσημία**: Τα φαινόμενα κατά το οποία μια λέξη/φράση μπορεί να αποκτήσει διφορούμενη σημασία ανάλογα με τα συμφραζόμενα (πχ. "*δε βρέθηκαν εγγραφές στη βάση*" , "*δεν πέρασε τη βάση εισαγωγής*") ή να έχει πολλαπλές και συσχετιζόμενες σημασίες (πχ. γλώσσα – τρόπος επικοινωνίας, γλώσσα – όργανο του στόματος)
- **Απουσία εννοιών**: Κατά την επεξεργασία φυσικού κειμένου είναι σχεδόν βέβαιο ότι πολλές από τις οντότητες δε θα υπάρχουν ως έννοιες στην γνωσιακή βάση αναφοράς.

2.4.2 Χρησιμότητα και εφαρμογές

Η σύνδεση οντοτήτων είναι επωφελής για πεδία που άπτονται της εξαγωγής κειμενικών αναπαραστάσεων, όπως συμβαίνει στην ανάλυση κειμένου, στα συστήματα συστάσεων, στη σημασιολογική αναζήτηση και στα chatbots καθώς διαχωρίζει τις έννοιες σχετικές με την εκάστοτε εφαρμογή από τα υπόλοιπα δεδομένα (Tan et al., 2017). Παράλληλα, αποτελεί σημαντικό εργαλείο σε μεθόδους εξαγωγής ή ανάκτησης πληροφοριών και βασικό στοιχείο για τη σημασιολογική αναζήτηση σε ψηφιακές βιβλιοθήκες (Han et al., 2005; Le and Mikolov, 2014). Ακόμη, χρησιμοποιείται συστηματικά από τις μηχανές αναζήτησης για τη βελτίωση των αποτελεσμάτων καθώς συμβάλει στη μείωση των ψευδώς-αρνητικών και ψευδώς-θετικών συσχετίσεων. Σε γενικές γραμμές, θεωρείται βασικό προαπαιτούμενο για τη γεφύρωση των διαδικτυακών δεδομένων με τις γνωσιακές βάσεις, συμβάλλοντας μαζί με άλλες εργασίες (πχ. σύνδεση γεγονότων) στο όραμα του Σημασιολογικού Ιστού (Shen et al., 2015).

2.4.3 Εκπαίδευση μοντέλου σύνδεσης οντοτήτων

Η εκπαίδευση μοντέλου σύνδεσης οντοτήτων με χρήση γνωσιακής βάσης περιγράφεται εκτενέστερα στην παρούσα υποενότητα καθώς αποτελεί τη συνηθέστερη μέθοδο επισήμανσης εγγράφων και χρησιμοποιήθηκε στα πλαίσια της παρούσας εργασίας για την δημιουργία αντίστοιχου μηχανισμού για την ελληνική γλώσσα.

Η διαδικασία απαιτεί τη δημιουργία ενός συνόλου εκπαίδευσης το οποίο αποτελείται από προτάσεις και τις αντιστοιχίσεις τους με οντότητες της γνωσιακής βάσης. Αρχικά, τα δεδομένα της βάσης μετασχηματίζονται σε κατάλληλη μορφή για συσχέτιση με το κείμενο αναφοράς. Συνήθως χρησιμοποιείται ένα μικρό υποσύνολο της διαθέσιμης πληροφορίας το οποίο περιέχει τα μοναδικά αναγνωριστικά των εννοιών που μας ενδιαφέρουν (πχ. οργανισμοί, τοποθεσίες, άτομα κτλ.). Για την προσθήκη εγγραφών στο σύνολο εκπαίδευσης, η περιγραφή κάθε έννοιας (πχ. από Wikipedia) κωδικοποιείται με χρήση διανυσματικών αναπαραστάσεων. Κάθε εγγραφή-έννοια στη βάση αναπαρίσταται ως ένα διάνυσμα που προκύπτει από τον μέσο όρο των συνιστωσών διανυσμάτων της περιγραφής του. Η τελική μορφή του συνόλου εκπαίδευσης έχει τη μορφή τριπλέτας και περιλαμβάνει την πρόταση, τα αναγνωριστικά των αντιστοιχισμένων οντοτήτων και τις θέσεις τους στο κείμενο (στο παρακάτω παράδειγμα επισημαίνεται μόνο η οντότητα "Κύπρος"):

Η Λευκωσία, γνωστή διεθνώς με το ιταλικό όνομα Nicosia, είναι η πρωτεύουσα της **Κύπρου**.

{"alias": "Κύπρου", "entity": "**Q229**", "start": 80, "end": 85}

Σημειώνεται ότι η διαδικασία μπορεί να παραμετροποιηθεί με τον έλεγχο του μέγιστου αριθμού οντοτήτων που μπορούν να περιλαμβάνονται σε κάθε πρόταση, τη συχνότητα εμφάνισής τους και άλλες παραμέτρους. Ακόμη, η διαθέσιμη πληροφορία συνήθως επιτρέπει την εισαγωγή ψευδωνύμων, συντομεύσεων ή υποκοριστικών για κάθε έννοια, διευκολύνοντας έτσι την αντιστοίχισή της με τις αναφορές στο κείμενο.

Στη συνέχεια ακολουθεί η εκπαίδευση ταξινομητή (πχ. νευρωνικού μοντέλου) στα παραπάνω δεδομένα. Το παραχθέν μοντέλο είναι σε θέση να αντιστοιχίζει λέξεις ή φράσεις με τα αναγνωριστικά των σχετικών οντοτήτων της βάσης, ανάλογα με τη θέση και τη σημασία τους στο κείμενο.

2.4.4 Σχετικό έργο

Οι προσεγγίσεις που αφορούν τη σύνδεση οντοτήτων διαφέρουν τόσο αναφορικά με τις τεχνικές σύνδεσης όσο και με τα εκάστοτε είδη δεδομένων που χρησιμοποιούν. Οι περισσότερες χρησιμοποιούν τη Wikipedia ως τη μόνη πηγή επίβλεψης για την κατασκευή διανυσματικών αναπαραστάσεων (Cucerzan, 2007), μοντέλα bag-of-words (Ratinov et al., 2011) ή χαρακτηριστικά βασισμένα σε κανόνες (πχ. μερική αντιστοιχία λέξεων) (Mcnamee et al., 2009). Στη συνέχεια, αντιμετωπίζουν το πρόβλημα ως μια εργασία ταξινόμησης πολλαπλών κλάσεων, στην οποία οι οντότητες αντιστοιχούν σε κλάσεις (Finkel et al., 2005).

Ο στόχος είναι να προταθεί μια λίστα υποψηφίων οντοτήτων για κάθε πιθανή αναφορά σε οντότητα, κωδικοποιώντας τόσο τις αναφορές όσο και τις υποψήφιες οντότητες σε διανυσματικές αναπαραστάσεις, και στη συνέχεια κατατάσσοντας τους υποψηφίους με βάση την συνάφεια του περιεχομένου. Η κατασκευή και συσχέτιση των embeddings αναφοράς και υποψήφιων οντοτήτων έχει καταστεί δυνατή με χρήση συνελικτικών (Francis-Landau et al., 2016; Sun et al., 2015), ανατροφοδοτούμενων δικτύων (Gupta et al., 2017; Martins et al., 2019) καθώς και μοντέλων αυτό-προσοχής (Wu et al., 2020; Logeswaran et al., 2020).

2.5 Επίλυση Συναναφορών

2.5.1 Ορισμός και βασικά στοιχεία

Η εργασία της επίλυσης συναναφορών (Coreference Resolution) στοχεύει στην εύρεση και ομαδοποίηση όλων των λέξεων ή εκφράσεων που αναφέρονται στην ίδια οντότητα μέσα στο κείμενο (Jurafsky and Martin, 2009). Μια λέξη (πχ. ένα άρθρο) θεωρείται ότι συναναφέρεται σε μια άλλη (πχ. ένα ουσιαστικό) όταν και οι δύο αφορούν την ίδια οντότητα, η οποία καλείται σημείο αναφοράς (antecedent). Στο παρακάτω παράδειγμα, οι λέξεις "του" και "το" αναφέρονται στο "παιδί" και στο "δώρο" αντίστοιχα.

Ο πατέρας υποσχέθηκε **δώρο** στο **παιδί**, αλλά δεν **του το** αγόρασε ποτέ.

Ο όρος "συναναφορά" (coreference) συχνά χρησιμοποιείται για να δηλώσει προβλήματα "πίσω-αναφοράς" (anaphora), η οποία ορίζεται ως το μη συμμετρικό συντακτικό φαινόμενο κατά το οποίο ένα ονοματικό σύνολο (noun phrase) αναφέρεται σε μία οντότητα (σημείο αναφοράς) που εμφανίζεται σε προγενέστερο σημείο του κειμένου (δηλαδή, αυτό που εμφανίζεται πρώτο απαιτείται για την ερμηνεία του δεύτερου). Σπανιότερα, χρησιμοποιείται για να δηλώσει προβλήματα "εμπρός-αναφοράς" ή "καταφοράς" (cataphora), όπου ένα ονοματικό σύνολο αφορά μία οντότητα που βρίσκεται σε μεταγενέστερο σημείο του κειμένου (δηλαδή, αυτό που εμφανίζεται δεύτερο απαιτείται για την ερμηνεία του πρώτου) (Elango, 2006).

2.5.2 Χρησιμότητα και εφαρμογές

Η αναγκαιότητα της επίλυσης συναναφορών στο φυσικό λόγο γίνεται εύκολα αντιληπτή από παραδείγματα ανάλυσης πραγματείας (discourse analysis). Ενώ η σύνταξη και η σημασιολογία λειτουργούν σε επίπεδο πρότασης, η ερμηνεία κειμένων από πολλαπλές συνενωμένες προτάσεις απαιτεί την κατανόηση του συνολικού νοήματος και όχι μεμονωμένη ερμηνεία της καθεμίας. Σε ακραίες περιπτώσεις, η επίλυση τέτοιων προβλημάτων είναι

δύσκολη ακόμα και από τον άνθρωπο και απασχόλησε από νωρίς την επιστημονική κοινότητα, οδηγώντας στην δημιουργία ενός συνόλου προτάσεων που χαρακτηρίζεται από περίπλοκες συναναφορές, γνωστό ως Winograd schema challenge (Levesque, 2011).

Η χρήση σχετικών μεθόδων συναντάται σε συστήματα διαλόγου (chatbots), όπου ο χρήστης συχνά απαντά με σύντομες φράσεις (πχ. "*θέλω να το αγοράσω*") (Stylianou and Vlahavas, 2021). Είναι επίσης διαδεδομένη σε συστήματα ερωταπαντήσεων τα οποία συνδέουν πληροφορίες για την παροχή πληροφοριών στο χρήστη (πχ. "*Η Μαρία Κάλλας ήταν κορυφαία Ελληνίδα υψίφωνος. Το έργο της περιλαμβάνει πάνω από 20 όπερες.*") (Morton, 1999). Παρόμοια είναι η χρήση τους και σε συστήματα εξαγωγής πληροφοριών, για την αυτόματη αντικατάσταση μερών του λόγου με την οντότητα στην οποία αναφέρονται, βελτιώνοντας την απόδοσή τους (Humphreys et al., 1997).

Τέλος, σημαντική είναι η συμβολή της επίλυσης συναναφορών σε συστήματα μηχανικής μετάφρασης, ιδιαίτερα κατά την μετάφραση μεταξύ μιας γλώσσας που δε διαθέτει γραμματικά γένη (πχ. αγγλικά) και μιας που διαθέτει (πχ. ελληνικά) (Stojanovski and Fraser, 2019).

2.5.3 Επίλυση συναναφορών με χρήση νευρωνικών δικτύων

Τα σύγχρονα συστήματα επίλυσης συναναφορών βασίζονται σε μεθόδους επιβλεπόμενης μηχανικής μάθησης. Η δημοφιλέστερη μεθοδολογία, η οποία χρησιμοποιείται και στα πλαίσια της διδακτορικής εργασίας βασίζεται σε νευρωνικό μοντέλο κατάταξης αναφορών (mention-rank model) για τον εντοπισμό αναφορών στο κείμενο και στη σύγκρισή τους με υποψήφιες οντότητες-σημεία αναφοράς (Lee et al., 2017a).

Η χρήση νευρωνικών αρχιτεκτονικών επιτρέπει την αναπαράσταση κάθε διαστήματος κειμένου (text span) i ως ένα διάνυσμα g_i, που προκύπτει από τη συνένωση των διανυσμάτων του αρχικού και του τελικού στοιχείου του διανύσματος h_{start_i}, h_{end_i} σε συνδυασμό με το διάνυσμα h_{att_i} που προκύπτει από εφαρμογή μηχανισμού προσοχής στο διάστημα κειμένου i για εύρεση της πιο σημαντικής λέξης:

$$g_i = [h_{start_i}, h_{end_i}, h_{att_i}] \tag{2.7}$$

Εφόσον κάθε διάστημα κειμένου έχει αναπαρασταθεί με την παραπάνω τεχνική, στόχος του μοντέλου είναι να αντιστοιχίσει σε κάθε i ένα σημείο αναφοράς y_i, που αποτελεί μια τυχαία μεταβλητή που κυμαίνεται μεταξύ των τιμών $[1, \dots, i-1, \varepsilon]$ (όπου το ε συμβολίζει ότι το i δεν αντιστοιχεί σε κανένα σημείο αναφοράς). Έτσι, για κάθε ζεύγος διαστημάτων κειμένου

i, j η μέθοδος υπολογίζει ένα σκορ σύνδεσης $s(i, j)$ βασιζόμενο στο αν καθένα εκ των δύο αποτελεί αναφορά καθώς και στο αν το ένα είναι σημείο αναφοράς του άλλου, χρησιμοποιώντας νευρωνικά δίκτυα πρόσθιας τροφοδότησης. Τέλος υπολογίζεται η κατανομή $P(y_i)$ των σημείων αναφοράς για το διάστημα i ως εξής:

$$P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y' \in Y(i)} e^{s(i,y_i)}} \tag{2.8}$$

Το σύνολο δεδομένων εκπαίδευσης αποτελείται από τις προτάσεις και τα ομαδοποιημένα διαστήματα κειμένου που αποτελούν συναναφορές για κάθε πρόταση. Το εκπαιδευμένο μοντέλο προκύπτει από την μεγιστοποίηση του αθροίσματος της παραπάνω κατανομής για όλα τα σημεία αναφοράς y_i. Καθώς τα νευρωνικά δίκτυα βασίζονται στην ελαχιστοποίηση μιας συνάρτηση κόστους, χρησιμοποιείται το αντίθετο του λογαρίθμου της πιθανότητας που ορίζεται στη σχέση 2.8.

2.5.4 Σχετικό έργο

Οι τεχνικές επίλυσης συναναφορών χωρίζονται στις ακόλουθες κατηγορίες:

- μοντέλα κατάταξης αναφορών (mention rank models), τα οποία αναλύθηκαν παραπάνω και θεωρούν όλα τα δυνατά διαστήματα κειμένου ως υποψήφιες αναφορές και κατατάσσουν την ανά-ζεύγη συνάφειά τους (Lee et al., 2017b; Wiseman et al., 2015),
- μοντέλα ζεύγους αναφορών (mention-pair models), τα οποία καθορίζουν την συναναφορά δύο λέξεων επιλύοντας ένα πρόβλημα δυαδικής ταξινόμησης (Ng and Cardie, 2001; Sapena et al., 2013),
- μοντέλα οντοτήτων-αναφορών (entity-mention models), που στηρίζονται στα mention-pair μοντέλα αλλά λαμβάνουν υπόψη όλες τις αναφορές μιας οντότητας κατά την διαδικασία ταξινόμησης προκειμένου να μειώσουν τις λανθασμένες αντιστοιχίσεις (Luo et al., 2004), και
- μοντέλα κατάταξης ομάδων (cluster rank models), τα οποία αποτελούν παραλλαγή των mention rank μοντέλων καθώς χρησιμοποιούν τις ομάδες συναναφοράς ως είσοδο, προσπαθώντας να ξεπεράσουν προβλήματα λανθασμένων αντιστοιχίσεων λαμβάνοντας υπόψη όλες τις αναφορές μιας οντότητας όπως τα μοντέλα entity-mention (Clark and Manning, 2016).

Καθώς οι παραπάνω τεχνικές αποτελούν εργασίες εποπτευόμενης μηχανικής μάθησης, βασίζονται στην ύπαρξη μεγάλων συνόλων επισημασμένου κειμένου, καθιστώντας δύσκολη

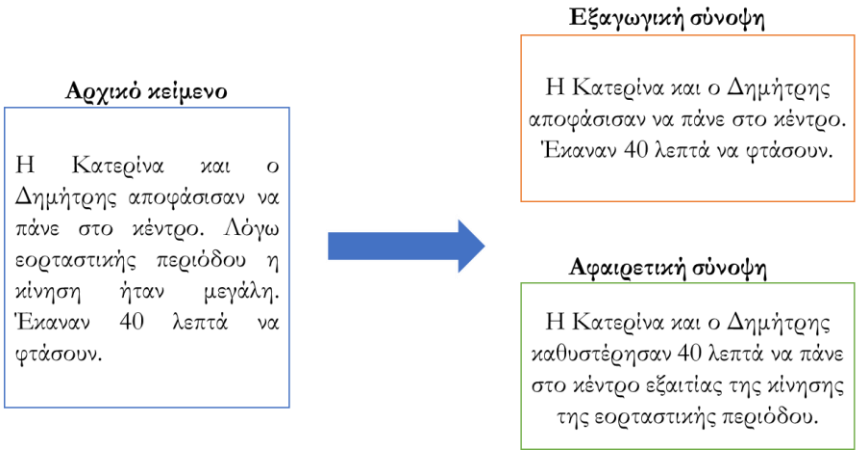
την εκπαίδευση μοντέλων για γλώσσες χαμηλών πόρων. Για το σκοπό αυτό, γίνονται προσπάθειες υλοποίησης διαγλωσσικών μοντέλων επίλυσης συναναφορών, με χρήση κοινών, πολύγλωσσων διανυσματικών αναπαραστάσεων ή παράλληλων κειμένων, που επιτρέπουν τη μεταφορά μάθησης μεταξύ γλωσσών (Kundu et al., 2018; Emelin and Sennrich, 2021).

2.6 Αυτόματη συνόψιση κειμένου

2.6.1 Ορισμός και βασικά στοιχεία

Ως αυτόματη συνόψιση κειμένου (automatic text summarization) ορίζεται η διαδικασία δημιουργίας μιας συντομευμένης, ακριβούς αναπαράστασης, διατηρώντας παράλληλα το βασικό περιεχόμενο και γενικό νόημα του αρχικού κειμένου (Allahyari et al., 2017). Η επιτυχής αυτοματοποίηση αυτού του έργου αποτελεί μη τετριμμένη εργασία καθώς απαιτεί τη δυνατότητα επισήμανσης των κύριων σημείων ενός κειμένου από μηχανές, οι οποίες στερούνται ανθρωπίνων γνώσεων και αντίστοιχων γλωσσικών ικανοτήτων. Ωστόσο, τα τελευταία χρόνια έχουν αναπτυχθεί πολυάριθμες προσεγγίσεις για την εφαρμογή αυτόματης συνόψισης σε διαφορετικούς γνωστικούς τομείς, καθιστώντας την αναζήτηση πληροφοριών ευκολότερη και λιγότερο χρονοβόρα.

Οι βασικές προσεγγίσεις αυτόματης συνόψισης χωρίζονται σε δύο κατηγορίες, όπως φαίνεται στο Σχήμα 7 (Gambhir and Gupta, 2017):



Σχήμα 7 Παράδειγμα εξαγωγικής και αφαιρετικής συνόψισης κειμένου

- Εξαγωγικές (extractive), στις οποίες εξάγεται αυτούσιο το σημαντικότερο περιεχόμενο του αρχικού κειμένου. Κατά τη διαδικασία αυτή, οι προτάσεις που περιέχονται στην περίληψη αποτελούν υποσύνολο αυτών που περιέχονται στο αρχικό κείμενο και μπορεί να παρομοιαστεί με την διαδικασία υπογράμμισης

αξιοσημείωτων μερών από τον άνθρωπο. Για την διάκριση των κατάλληλων προς εξαγωγή προτάσεων συνήθως βαθμονομείται η κάθε λέξη ή φράση βάσει διαφόρων κριτηρίων (πχ. συχνότητα εμφάνισης). Όπως γίνεται εύκολα αντιληπτό, η συντακτική και γραμματική συνοχή της περίληψης από τις εξαγωγικές προσεγγίσεις θεωρείται δεδομένη, ωστόσο η παραγόμενη πληροφορία είναι περιορισμένη καθώς δεν είναι πάντα βέβαιο πως τα κύρια σημεία του κειμένου (που συχνά βρίσκονται διάσπαρτα σε όλο το κείμενο) έχουν εγκολπωθεί στο τελικό αποτέλεσμα.

- Αφαιρετικές (abstractive), οι οποίες δημιουργούν μια εσωτερική σημασιολογική αναπαράσταση του αρχικού περιεχομένου και στη συνέχεια την αξιοποιούν για τη δημιουργία ενός νέου κειμένου-περίληψης, παραφράζοντας τμήματα του εγγράφου για να συμπυκνώσουν το βασικό του νόημα. Αυτού το είδους ο μετασχηματισμός θεωρείται υπολογιστικά δυσκολότερος καθώς απαιτεί τόσο χρήση μηχανισμών παραγωγής κειμένου όσο και βαθύτερη κατανόηση του γνωστικού αντικειμένου στο οποίο εφαρμόζεται (εκπαίδευση σε σχετικά σύνολα δεδομένων). Η διαδικασία προσομοιάζει σε μεγάλο βαθμό την καθιερωμένη ανθρώπινη διαδικασία παραγωγής περίληψης και ως εκ τούτου απαιτεί μεγάλο όγκο δεδομένων εκπαίδευσης, ενώ η γραμματική και συντακτική συνοχή του αποτελέσματος δεν είναι εγγυημένη.

2.6.2 Χρησιμότητα και εφαρμογές

Τα συστήματα αυτόματης συνόψισης κειμένου χρησιμοποιούνται συχνά από υπηρεσίες αποδελτίωσης μέσων ενημέρωσης (πχ. ειδησεογραφικά sites, ιστολόγια) και μέσων κοινωνικής δικτύωσης για την κατασκευή ενημερωτικών δελτίων (newsletters) (Nguyen et al., 2019). Επιπλέον, καθώς το διαδίκτυο αποτελεί αστείρευτη πηγή κειμενικής πληροφορίας, η αξιοποίησή τέτοιων μεθόδων από επιχειρήσεις για λόγους προώθησης προϊόντων, ανακάλυψης νέων αγοραστικών τάσεων και γενικότερης λήψης αποφάσεων οδηγεί στην αποδοτική ανάλυση αποτελεσμάτων αναζήτησης και χρηματοοικονομικών εγγράφων μέσω αντίστοιχων μεθόδων συνόψισης (Ghodratnama et al., 2020). Ως εκ τούτου, συχνά ενσωματώνονται σε μηχανές ανοιχτής εξαγωγής πληροφοριών (OIE) συμβάλλοντας στην μείωση του θορύβου (πχ. άσχετες ή πλεονάζουσες τριπλέτες) (Li, 2015; Adnan and Akbar, 2019).

Στον ακαδημαϊκό τομέα, αξίζει να σημειωθεί η χρήση μηχανισμών συνόψισης για την αυτόματη ομαδοποίηση επιστημονικών δημοσιεύσεων με συναφές περιεχόμενο μέσω της παραγωγής νοηματικά συμπυκνωμένων περιλήψεων καθώς και για εργασίες παρακολούθησης ερευνητικών τάσεων και καινοτομικών ιδεών (Erera et al., 2020).

Αποτελούν ακόμα υποσχόμενη μέθοδο σε περιπτώσεις χρήσης που άπτονται της βιοϊατρικής, συμβάλλοντας στη διαχείριση γνώσης που προέρχεται από πλήθος κλινικών μελετών καθώς εξάγουν αποσπάσματα που περιέχουν τα βασικότερα σημεία του κειμένου (Gulden et al., 2019). Αντίστοιχα, συμβάλλουν στην γρηγορότερη διεκπεραίωση νομικών υποθέσεων, διυλίζοντας εκτενή νομικά έγγραφα για την εύρεση χρήσιμων αποσπασμάτων (Bhattacharya et al., 2021).

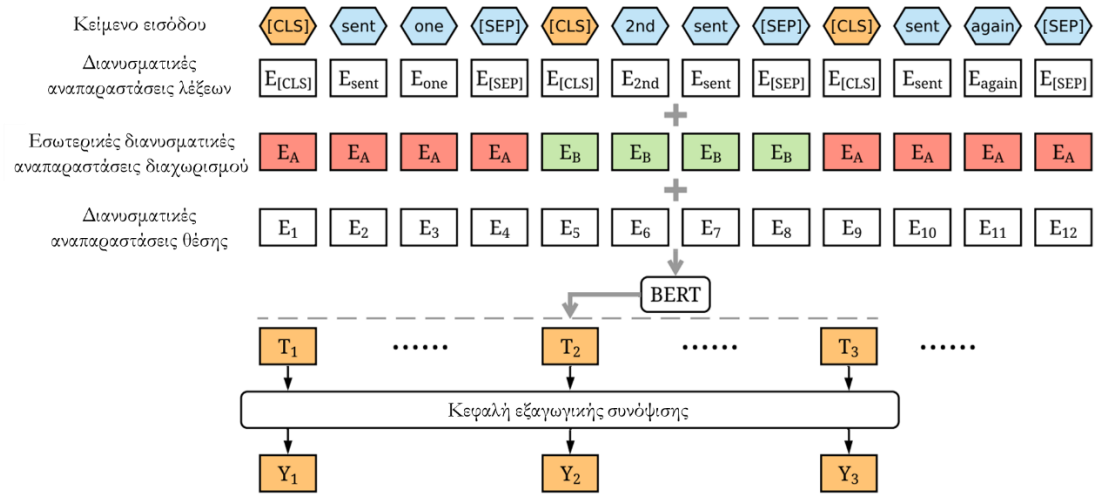
2.6.3 Εξαγωγική συνόψιση με χρήση κωδικοποιητή

Η εκπαίδευση γλωσσικών μοντέλων τεχνολογίας αιχμής, όπως αυτά που βασίζονται σε αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή (encoder-decoder) στοχεύει στη δημιουργία ενός "μαύρου κουτιού" που είναι σε θέση να κατανοεί τη γλώσσα μέσω της τροφοδότησής του με μεγάλο πλήθος μη επισημασμένου κειμένου. Μέσω αυτής της διαδικασίας, η οποία καλείται προεκπαίδευση (pre-training) το μοντέλο μαθαίνει τη χρήση και τη σημασία των λέξεων στο κείμενο από τις διανυσματικές τους αναπαραστάσεις. Έπειτα μπορεί να χρησιμοποιηθεί για πιο συγκεκριμένες εργασίες μέσω της εκπαίδευσής του σε μικρότερο, σχετικό με το αντικείμενο σύνολο δεδομένων, μέσω μιας διαδικασίας που ονομάζεται προσαρμογή (fine-tuning).

Η προσαρμογή προεκπαιδευμένων γλωσσικών μοντέλων για μεταγενέστερες χρήσεις (συστήματα ταξινόμησης ακολουθιών, ερωταποκρίσεων, ανάλυσης συναισθήματος κλπ.) εφαρμόζεται συχνά για την υλοποίηση εξαγωγικών μοντέλων συνόψισης, όπως αυτό που χρησιμοποιήθηκε στα πλαίσια της διδακτορικής διατριβής. Η αρθρωτή αρχιτεκτονική που χαρακτηρίζει τα μοντέλα αρχιτεκτονικής Transformer όπως το BERT το οποίο χρησιμοποιεί μόνο τον κωδικοποιητή (Devlin et al., 2019a) διευκολύνει την προσθήκη ειδικών κεφαλών (heads) στο τελικό στάδιο, δηλαδή στρωμάτων που διαφοροποιούν το αποτέλεσμα εξόδου ανάλογα με την εργασία χρήσης.

Έστω ένα έγγραφο d που περιέχει m προτάσεις $[sent_1, sent_2, \dots, sent_m]$ όπου $sent_i$ είναι η i-οστή πρόταση στο έγγραφο. Το έργο της εξαγωγικής συνόψισης περιλαμβάνει την αντιστοίχιση μιας ετικέτας $y_i \in \{0,1\}$ σε κάθε πρόταση $sent_i$, ανάλογα με το αν αυτή ανήκει στην περίληψη ή όχι. Δεδομένου ότι η προεκπαίδευση περιλαμβάνει διαδικασία masking, δηλαδή την κάλυψη λέξεων (tokens) από το μοντέλο και πρόβλεψής τους αργότερα για εξασφάλιση της αμφιδρομικότητάς του, τα διανύσματα εξόδου αντιστοιχούν σε λέξεις και όχι σε προτάσεις. Επιπλέον, αν και το μοντέλο περιλαμβάνει ειδικές διανυσματικές αναπαραστάσεις διαχωρισμού προτάσεων βασιζόμενο σε ετικέτες που υποδεικνύουν τις διαφορετικές προτάσεις, η πρωτότυπη αρχιτεκτονική σχεδιάστηκε για

εργασίες επεξεργασίας γλώσσας που αφορούν μόνο δύο προτάσεις. Επομένως, τόσο η είσοδος του μοντέλου όσο και οι διανυσματικές του αναπαραστάσεις πρέπει να τροποποιηθούν κατάλληλα για την επεξεργασία πολλαπλών προτάσεων. Αυτό καθίσταται εφικτό με τη χρήση ενός token [CLS] πριν από κάθε πρόταση και ενός token [SEP] μετά από την πρόταση, όπως φαίνεται στο Σχήμα 8 (Liu, 2019).



Σχήμα 8 Προσαρμογή (fine-tuning) μοντέλου αρχιτεκτονικής Transformer για εργασίες εξαγωγικής συνόψισης.

Στη συνέχεια, το κείμενο εισόδου με τα κατάλληλα tokens κωδικοποιείται λέξη-προς-λεξη για τον σχηματισμό της διανυσματικής αναπαράστασης λέξεων. Κατόπιν, οι εσωτερικές διανυσματικές αναπαραστάσεις διαχωρισμού χρησιμοποιούνται για να υποδείξουν την ύπαρξη πολλαπλών προτάσεων στο κείμενο, ενώ οι διανυσματικές αναπαραστάσεις θέσης κωδικοποιούν την απόλυτη θέση των λέξεων στο κείμενο. Κατ' αυτό τον τρόπο, το T_i που ορίζεται ως το διάνυσμα του i -οστού token [CLS] από τα δεδομένα εισόδου χρησιμοποιείται ως η διανυσματική αναπαράσταση της πρότασης $sent_i$. Μετά την εξαγωγή διανυσματικών αναπαραστάσεων προτάσεων T_i, προστίθεται η σχετική κεφαλή συνόψισης, η οποία μπορεί να αποτελείται από ένα ή περισσότερα στρώματα για την εξαγωγή χαρακτηριστικών που αφορούν την ανάθεση της ετικέτας y_i για την προσαρμογή του μοντέλου. Αυτό αποτελεί τυπική διαδικασία εκπαίδευσης ταξινομητή (classifier) επιβλεπόμενης μάθησης, με τη συνάρτηση κόστους να ορίζεται ως η δυαδική εγκάρσια εντροπία (binary cross-entropy) της προβλεφθείσας $\hat{Y}_i$ έναντι της πραγματικής τιμής Y_i. Το τελικό στρώμα εξόδου μπορεί να έχει την μορφή που φαίνεται παρακάτω:

$$\hat{Y}_i = \sigma(W_o T_i + b_o) \tag{2.9}$$

όπου σ η σιγμοειδής συνάρτηση και W_o, b_o τα συναπτικά βάρη και η παράμετρος μεροληψίας του νευρώνα αντίστοιχα.

2.6.4 Σχετικό έργο

Συναντώνται διαφορετικές ερευνητικές γραμμές ανάλογα το είδος των τεχνικών συνόψισης. Όσον αφορά τις εξαγωγικές τεχνικές, συναντάται πληθώρα παραλλαγών που βασίζεται σε νευρωνικές αρχιτεκτονικές ανατροφοδοτούμενων δικτύων ή κωδικοποιητή-αποκωδικοποιητή (Kågebäck et al., 2015; Cao et al., 2015) με τις πιο πρόσφατες να ενσωματώνουν μηχανισμούς προσοχής (Cao et al., 2016; Wang et al., 2019b), όπως αυτή που περιεγράφηκε στην παραπάνω υποενότητα. Σημειώνεται ότι η ύπαρξη επισημασμένου συνόλου δεδομένων με τη μορφή παραγράφων κειμένου και των αντίστοιχων περιλήψεών τους είναι απαραίτητη για την εκπαίδευση αντίστοιχων συστημάτων. Καθώς η εξαγωγική συνόψιση έχει φτάσει σε ώριμο ερευνητικό στάδιο, οι υπάρχουσες τεχνικές καταφέρνουν να συλλάβουν τα βασικά στοιχεία ενός κειμένου, αν και οι αντίστοιχες περιλήψεις ενδέχεται να στερούνται αναγνωσιμότητας και ροής.

Στο πεδίο της αφαιρετικής συνόψισης έχουν επίσης χρησιμοποιηθεί κατά κόρον αντίστοιχες νευρωνικές αρχιτεκτονικές (Nallapati et al., 2016; Paulus et al., 2018; Shi et al., 2021), ωστόσο λόγω της πολυπλοκότητας της εργασίας, οι περισσότερες υλοποιήσεις χαρακτηρίζονται είτε από χαμηλή ανάκληση (αδυναμία να συμπεριλάβουν κομβικά σημεία του κειμένου), είτε χαμηλή ακρίβεια (παραγωγή κειμένου χωρίς λογική συνέπεια – hallucinations) (Kryściński et al., 2020). Πρόσφατες υλοποιήσεις επιστρατεύουν τεχνικές ερωταποκρίσεων (Gunasekara et al., 2021) και εξαγωγής πληροφοριών (Li et al., 2018) σε συνδυασμό με τις προαναφερθείσες υλοποιήσεις, με σκοπό τη μείωση του θορύβου και τη βελτίωση της συνοχής του παραγόμενου κειμένου.

2.7 Σημασιολογική ομοιότητα και κειμενική συνεπαγωγή

2.7.1 Ορισμός και βασικά στοιχεία

Με τον όρο σημασιολογική ομοιότητα κειμένου (semantic textual similarity – STS) ορίζεται το μέτρο της εννοιολογικής απόστασης μεταξύ δύο λεκτικών συνόλων (λέξεων, φράσεων ή προτάσεων) βάσει της σημασίας τους (Corley and Mihalcea, 2005). Η διαδικασία υπολογισμού σημασιολογικής ομοιότητας αξιοποιεί μεθόδους διανυσματικής αναπαράστασης για τη μέτρηση της απόστασης των αντίστοιχων διανυσμάτων τους, η οποία αντιστοιχεί στη νοηματική τους ομοιότητα. Σε αντίθεση με την λεξικογραφική ομοιότητα η

οποία εστιάζει στην εμφάνιση μιας συμβολοσειράς μέσα σε μια άλλη, η σημασιολογική ομοιότητα δύο λεκτικών συνόλων μπορεί να είναι υψηλή ακόμα και αν η μία δεν περιέχει κανένα σύμβολο της άλλης, όπως φαίνεται στο παρακάτω σχήμα (Σχήμα 9).

Πρόταση Α:	Το κινητό έπεσε και έσπασε στα δύο.	**sim(A,B)**
Πρόταση B_1:	Η πτώση κατέστρεψε τη συσκευή.	0.674
Πρόταση B_2:	Το αυτοκίνητο έπεσε σε χαντάκι.	0.499
Πρόταση B_3:	Ο υπουργός έπεσε και έσπασε το πόδι του.	0.455

Σχήμα 9 Παράδειγμα υπολογισμού σημασιολογικής ομοιότητας.

Διακρίνονται δύο βασικές προσεγγίσεις υπολογισμού σημασιολογικής ομοιότητας, αυτές που βασίζονται σε σημασιολογικά δίκτυα ή γνωσιακές βάσεις (πχ. Wordnet[6]) και αυτά που βασίζονται σε μεγάλες συλλογές κειμένων και βασίζονται στη θεωρία του J.R. Firth ότι δύο σύνολα που χρησιμοποιούνται σε παρόμοιο πλαίσιο έχουν παρόμοιο νόημα (Haas, 1958).

Η σημασιολογική ομοιότητα αποτελεί συγγενές αντικείμενο μιας εξίσου βασικής εργασίας επεξεργασίας φυσικής γλώσσας, που ονομάζεται αναγνώριση κειμενικής συνεπαγωγής (recognizing textual entailment – RTE), η οποία συναντάται συχνά και ως εξαγωγή συμπεράσματος (natural language inference – NLI). Αυτή αναπαριστά την ικανότητα εξαγωγής συμπερασμάτων μεταξύ δύο προτάσεων, της προϋπόθεσης (premise) και της υπόθεσης (hypothesis), μέσω μια τιμής-ετικέτας που εκφράζει τη σχέση τους (Tatu and Moldovan, 2005). Η τιμή αυτή μπορεί να προκύψει από εφαρμογή αλγορίθμων σημασιολογικής ομοιότητας μεταξύ των προτάσεων και να αντιστοιχεί σε διαφορετικό αποτέλεσμα συνεπαγωγής ανάλογα με σχετικά κατώφλια που έχουν οριστεί. Έτσι, αν μια υπόθεση μπορεί να συμπεραθεί από την προϋπόθεση δεχόμαστε ότι υπάρχει "συνεπαγωγή", αν αντιφάσκει ως προς την προϋπόθεση δεχόμαστε ότι υπάρχει "αντίφαση", ενώ αν δεν υπάρχει ξεκάθαρη ετυμηγορία με βάση την τιμή, το ζεύγος προτάσεων χαρακτηρίζεται ως "ουδέτερο" (Σχήμα 10).

[6] https://wordnet.princeton.edu/

Προϋπόθεση Α	Υπόθεση Β	rte(A,B)
Δύο άνθρωποι συναντιούνται στο δρόμο.	Ο δρόμος έχει κόσμο.	ΣΥΝΕΠΑΓΩΓΗ
Ένα μαύρο αυτοκίνητο ξεκινάει στη μέση του πλήθους.	Ένας άντρας οδηγεί σε ένα μοναχικό δρόμο.	ΑΝΤΙΦΑΣΗ
Δυο γυναίκες μιλάνε στο κινητό.	Το τραπέζι ήταν πράσινο.	ΟΥΔΕΤΕΡΟ

Σχήμα 10 Παράδειγμα αναγνώρισης κειμενικής συνεπαγωγής

Αξίζει να σημειωθεί ότι, παρά τις ομοιότητες μεταξύ των δύο εργασιών, η σημασιολογική ομοιότητα εκφράζει μια συμμετρική σχέση μεταξύ λεκτικών συνόλων, δηλαδή $sim(A,B) = sim(B,A)$, ενώ η κειμενική συνεπαγωγή είναι μη συμμετρική μεταξύ μιας προϋπόθεσης και μιας υπόθεσης, δηλαδή $rte(A,B) \neq rte(B,A)$ (Ma et al., 2018). Για το σκοπό αυτό, στα πλαίσια εκπαίδευσης αντίστοιχων μοντέλων με χρήση μεθόδων βαθιάς μάθησης, χρησιμοποιούνται διαφορετικά σύνολα εκπαίδευσης ανάλογα με την κάθε εργασία.

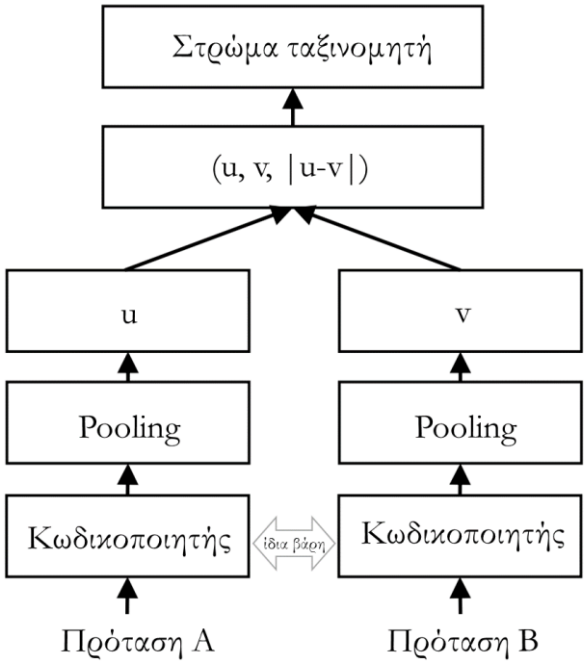
2.7.2 Χρησιμότητα και εφαρμογές

Ο υπολογισμός σημασιολογικής ομοιότητας αποτελεί θεμελιώδη εργασία επεξεργασίας φυσικής γλώσσας υποστηρίζοντας πληθώρα εφαρμογών όπως η αφαίρεση διπλότυπων εγγράφων (Gyawali et al., 2020), η ανίχνευση παραφρασμένων κειμένων (Nighojkar and Licato, 2021), η σημασιολογική αναζήτηση για εξαγωγή πληροφοριών (Ranasinghe et al., 2019) και η απάντηση ερωτήσεων βάσει πληροφοριών που βρίσκονται στο κείμενο (Banerjee et al., 2020). Μπορεί ακόμα να χρησιμοποιηθεί και ως μέτρο αξιολόγησης μεθόδων μηχανικής μετάφρασης, συγκρίνοντας την πιστότητα του παραγόμενου κειμένου σε σχέση με μια άρτια μετάφραση από επαγγελματία μεταφραστή (Magnolini et al., 2016). Ομοίως, οι μέθοδοι κειμενικής συνεπαγωγής συναντώνται σε αντίστοιχες περιπτώσεις και επιπλέον για την αυτοματοποίηση διαδικασιών (πχ. τραπεζικών και λοιπών χρηματοοικονομικών εφαρμογών), όπου απαιτείται ο έλεγχος για το αν το παραγόμενο αποτέλεσμα από τον τελικό χρήστη ακολουθεί τα υπάρχοντα δεδομένα, αντικαθιστώντας σε κάποιες περιπτώσεις τον ανθρώπινο έλεγχο (Sesen et al., 2018). Ακόμη, αποτελούν βασικό στοιχείο εφαρμογών διαλόγου (chatbots) (Wang and Manning, 2010) και μεθόδων επικύρωσης ισχυρισμών (Wang et al., 2019a).

2.7.3 Προσαρμογή κωδικοποιητή για τον εργασίες κειμενικής συνεπαγωγής

Όπως και με τις περισσότερες εργασίες προσαρμογής (finetuning) μοντέλων που βασίζονται σε αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή, η εκπαίδευση ενός αντίστοιχου

μοντέλου για κειμενική συνεπαγωγή προϋποθέτει την ύπαρξη ενός προεκπαιδευμένου μοντέλου – στη συγκεκριμένη περίπτωση κωδικοποιητή (encoder) – που μπορεί να εξάγει διανυσματικές αναπαραστάσεις από τα εισαχθέντα λεκτικά σύνολα. Όπως γίνεται αντιληπτό, πρόκειται για εργασία επιβλεπόμενης μάθησης και επομένως απαιτείται η ύπαρξη επισημασμένου συνόλου δεδομένων εκπαίδευσης, στο οποίο θα βασιστεί η προσαρμογή του μοντέλου.



Σχήμα 11 Παράδειγμα αρχιτεκτονικής διπλού κωδικοποιητή για κειμενική συνεπαγωγή

Για τις ανάγκες της εργασίας, χρησιμοποιούνται δύο κωδικοποιητές σε παράλληλη λειτουργία (με ίδια βάρη προεκπαίδευσης) για την εξαγωγή των διανυσματικών αναπαραστάσεων των δύο προτάσεων που τίθενται προς σύγκριση ακολουθώντας την αρχιτεκτονική του Reimers and Gurevych (2020). Επειδή ο κάθε κωδικοποιητής οδηγεί σε διανυσματική αναπαράσταση διαφορετικής διάστασης ανάλογα με το μήκος της πρότασης, προστίθεται ένα στρώμα υπο-δειγματοληψίας (pooling) που μειώνει τη διάσταση των παραπάνω embeddings, σχηματίζοντας δύο συμπυκνωμένες αναπαραστάσεις $\boldsymbol{u}, \boldsymbol{v}$ σταθερού μήκους $\boldsymbol{n}$ (διανύσματα πρότασης). Για την εφαρμογή της καταλληλότερης μεθόδου pooling χρησιμοποιούνται διάφορες τεχνικές, όπως ο υπολογισμός του μέσου ή του μεγίστου κάθε διανύσματος λέξης που ανήκει στην πρόταση (mean/max-over-time-pooling). Ακολουθεί η προσθήκη του τελικού στρώματος ταξινόμησης, αφού προηγηθεί η σύντμηση των δύο embeddings (πχ. παίρνοντας την διαφορά τους κατά στοιχείο $|\boldsymbol{u} - \boldsymbol{v}|$). Το αποτέλεσμα πολλαπλασιάζεται με το διάνυσμα βάρους $\boldsymbol{W_t} \in \boldsymbol{R^{3nk}}$, όπου $\boldsymbol{n}$ η διάσταση της των

embeddings και k ο αριθμός των ετικετών (στην περίπτωσή μας $k = 3$) που υποδηλώνουν συνεπαγωγή, αντίφαση ή ουδέτερη σχέση. Η εκπαίδευση των βαρών γίνεται μέσω της παρακάτω σχέσης εξόδου:

$$output = softmax(W_t|u - v|) \tag{2.10}$$

Το αποτέλεσμα εκπαίδευσης είναι ένα γλωσσικό μοντέλο που δέχεται δύο προτάσεις και επιστρέφει τα logits κάθε ετικέτας, επιτρέποντας στον χρήστη να εξάγει συμπεράσματα για τη σχέση μεταξύ τους. Σημειώνεται ότι με αντίστοιχο τρόπο μπορεί να γίνει και εκπαίδευση μοντέλου για τον υπολογισμό σημασιολογικής ομοιότητας, απλά αντικαθίσταται η συνάρτηση στο τελικό στρώμα (πχ. με ομοιότητα συνημιτόνου).

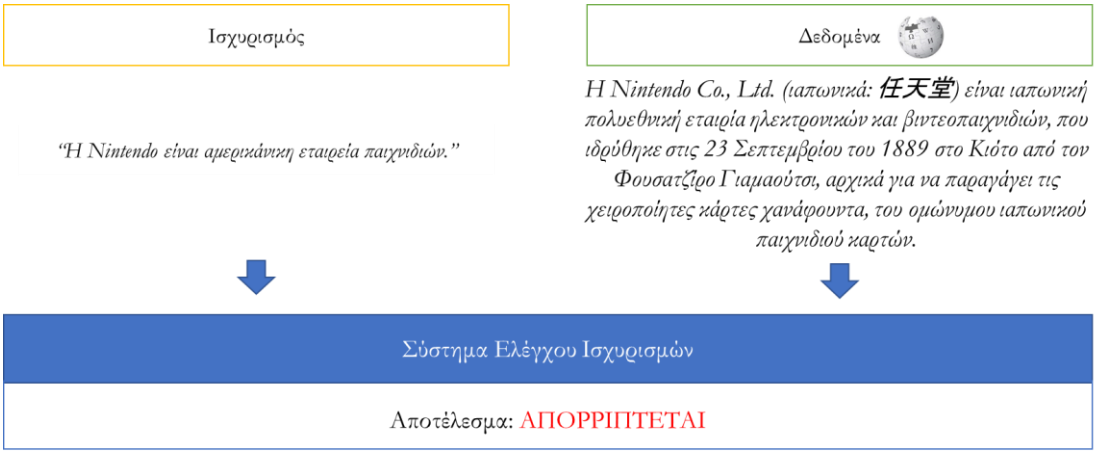
2.7.4 Σχετικό έργο

Στη διεθνή βιβλιογραφία συναντάται πληθώρα μεθόδων για τον υπολογισμό σημασιολογικής ομοιότητας. Συγκεκριμένα, έχουν χρησιμοποιηθεί συνελικτικά (He et al., 2015) και ανατροφοδούμενα (Tai et al., 2015) νευρωνικά δίκτυα, τα οποία προσπαθούν να επεκτείνουν προσεγγίσεις που βασίζονται σε μεμονωμένες λέξεις, ώστε να χρησιμοποιηθούν για την ανάλυση ολόκληρων προτάσεων. Το πρόσφατο ερευνητικό ενδιαφέρον για κωδικοποιητές προτάσεων (sentence encoders) έχει πυροδοτήσει την ανάπτυξη αντίστοιχων συστημάτων επιβλεπόμενης μάθησης που βασίζονται σε μηχανισμούς προσοχής (Raffel et al., 2020), δίκτυα Deep Averaging (Iyyer et al., 2015) και siamese-encoders (Ranasinghe et al., 2019) για την εξαγωγή διανυσματικών αναπαραστάσεων απευθείας από προτάσεις. Ειδικότερα για εργασίες κειμενικής συνεπαγωγής, έχουν αναπτυχθεί τόσο μοντέλα global attention (Beltagy et al., 2020b) που λαμβάνουν υπόψη όλες τις κρυφές καταστάσεις του κωδικοποιητή, όσο και αυτο-παλινδρομικά μοντέλα (Yang et al., 2019) καθώς και κωδικοποιητές με τεχνικές αφαίρεσης θορύβου (Lewis et al., 2020).

2.8 Έλεγχος Ισχυρισμών

2.8.1 Ορισμός και βασικά στοιχεία

Ο έλεγχος ισχυρισμών (claim validation) αποτελεί το σύνθετο έργο της ανάθεσης μιας τιμής αλήθειας (truth value) σε έναν ισχυρισμό, ο οποίος αφορά συγκεκριμένο εννοιολογικό πλαίσιο. Αποτελεί συγγενές αντικείμενο του ελέγχου γεγονότων (fact-checking) που χρησιμοποιείται για την ανίχνευση ψευδών ειδήσεων από διαδικτυακές πηγές (Vlachos and Riedel, 2015).



Σχήμα 12 Επισκόπηση συστήματος ελέγχου ισχυρισμών

Ο αυτοματοποιημένος έλεγχος ισχυρισμών ξεκινάει με την εισαγωγή μιας πρότασης από τον τελικό χρήστη, η οποία επικυρώνεται ή απορρίπτεται βάσει των σχετικών συλλεχθέντων στοιχείων που προκύπτουν από τη συνεχή παρακολούθηση μέσων ενημέρωσης ή αντίστοιχων πηγών (πχ. Wikipedia). Το συγκεκριμένο εγχείρημα συνοδεύεται από προκλήσεις σε όλα τα στάδιά του, καθώς ενδέχεται να συγκεντρώνει πολλές από τις προαναφερθείσες τεχνικές που περιεγράφηκαν στις παραπάνω υποενότητες, από την διαδικασία εξαγωγής αξιόπιστων πληροφοριών, στην επίλυση συναναφορών μεταξύ των αναφερόμενων οντοτήτων για τη συγκέντρωση σχετικών με αυτές περιεχομένου, καθώς και τον έλεγχο της κειμενικής συνεπαγωγής μεταξύ της προϋπόθεσης (συγκεντρωμένη πληροφορία) και της υπόθεσης (ισχυρισμού) (Nie et al., 2019).

2.8.2 Χρησιμότητα και εφαρμογές

Η διάχυση τεχνολογιών Ιστού 2.0 έχει απλοποιήσει τις διαδικασίες διαμοιρασμού πληροφοριών μεταξύ χρηστών, επιτρέποντας παράλληλα την πρόσβαση σε ένα διευρυνόμενο σύμπαν πληροφοριών. Αυτό έχει ως αποτέλεσμα την σημαντική μεταβολή του τρόπου κατανάλωσης πληροφοριών, ωθώντας τα άτομα να βασίζονται σχεδόν ολοκληρωτικά σε διαδικτυακές πηγές για την ενημέρωσή τους. Συνεπώς, η αναγκαιότητα ελέγχου ισχυρισμών βάσει του υπάρχοντος πληροφοριακού πλούτου με αυτοματοποιημένο τρόπο είναι δεδομένη, ειδικά αν ληφθεί υπόψη και η δυναμικότητα του περιβάλλοντος λόγω της συνεχούς τροφοδότησης με νέες πληροφορίες, οι οποίες μπορούν να οδηγήσουν σε ανάκληση του αρχικού αποτελέσματος ελέγχου (Jiang et al., 2020).

Η ύπαρξη ιστοτόπων για την άμεση επαλήθευση ισχυρισμών (Politifact[7], Snopes[8], EllinikaHoaxes[9]) βασιζόμενων σε ανθρώπινη επίβλεψη αναδεικνύει την χρησιμότητα μηχανισμών ελέγχου ισχυρισμών για θέματα γενικού περιεχομένου. Το πεδίο εφαρμογής αντίστοιχων συστημάτων επεκτείνεται, ωστόσο, και σε πιο εξειδικευμένους τομείς όπως τον έλεγχο βιοϊατρικών απόψεων (Wührl and Klinger, 2021), νομικών ισχυρισμών (Surdeanu et al., 2010) και αποτελεσμάτων στατιστικών μελετών (Karagiannis et al., 2020).

2.8.3 Επισκόπηση συστημάτων ελέγχου ισχυρισμών

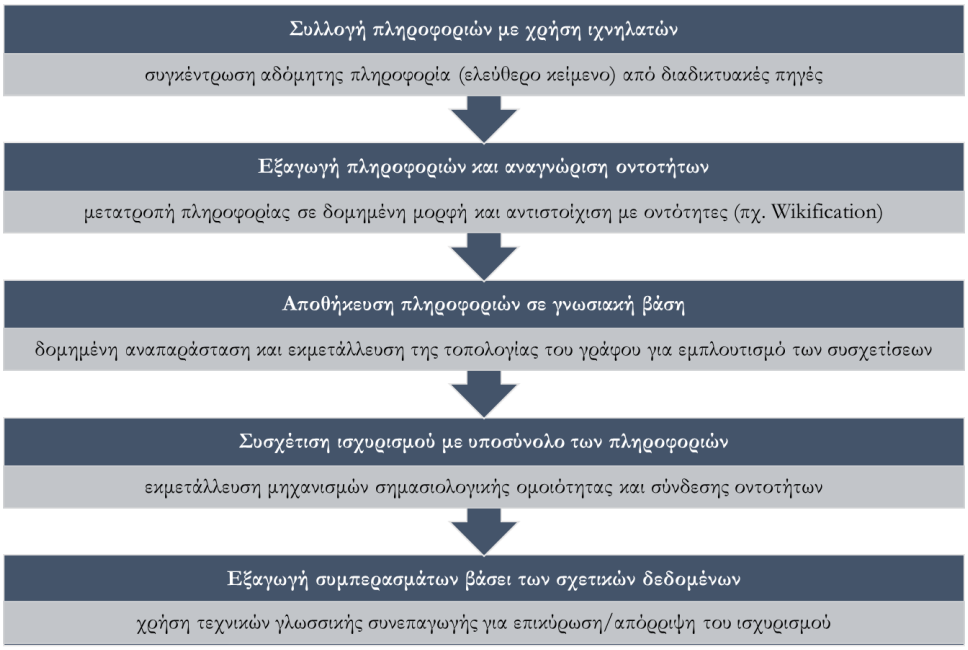
Όπως προαναφέρθηκε στην υποενότητα 2.8.1, η δημιουργία ενός συστήματος που θα ελέγχει έναν ισχυρισμό σε μορφή φυσικού κειμένου και θα επιστρέφει μια τιμή αλήθειας ανάλογα με το αν αυτός επιβεβαιώνεται ή όχι από τα υπάρχοντα δεδομένα, δεν αποτελεί συνήθως διατερματική (end-to-end) διαδικασία, αλλά συνδυασμό βασικών επιμέρους μηχανισμών σε σειριακή μορφή (pipeline) (Soleimani et al., 2020). Οι αρχιτεκτονικές που χαρακτηρίζουν τα περισσότερα συστήματα ελέγχου ισχυρισμών ενσωματώνουν κάποιους από τους κάτωθι μηχανισμούς:

- *συγκέντρωσης περιεχομένου από ιχνηλάτες (crawlers)* που αναλαμβάνουν την συνεχή παρακολούθηση διαδικτυακών πηγών και την αποθήκευση κειμενικού περιεχομένου για περαιτέρω επεξεργασία,
- *εξαγωγής πληροφοριών* για μετατροπή του αδόμητου κειμένου σε δομημένη μορφή καθώς και για την ανίχνευση των οντοτήτων στις οποίες αναφέρεται,
- *αποθήκευσης σε βάσεις δεδομένων* ή *γνωσιακές βάσεις,* οι οποίες θα διατηρούν τη δομημένη αναπαράσταση της συγκεντρωθείσας πληροφορίας,
- *συσχέτισης περιεχομένου,* συνήθως αξιοποιώντας τεχνικές *σημασιολογικής ομοιότητας* ή παρόμοιες ευρετικές μεθόδους που επιτρέπουν την αντιστοίχιση περιεχομένου με τον ισχυρισμό που έθεσε ο χρήστης, και
- *κειμενικής συνεπαγωγής,* ώστε να ελεγχθεί εάν ο ισχυρισμός του χρήστη τεκμαίρεται από τα σχετικά δεδομένα.

[7] https://www.politifact.com/

[8] https://www.snopes.com/

[9] https://www.ellinikahoaxes.gr/



Σχήμα 13 Παράδειγμα μεθοδολογίας ελέγχου ισχυρισμών

Το Σχήμα 13 απεικονίζει συνοπτικά τη ροή των επιμέρους εργασιών οι οποίες απαρτίζουν τη μεθοδολογία ελέγχου ισχυρισμών η οποία αναπτύχθηκε στα πλαίσια της εργασίας για την ελληνική γλώσσα και αναλύεται διεξοδικά στο Κεφάλαιο 3. Η εκπαίδευση των επιμέρους μηχανισμών (πχ. νευρωνικών μοντέλων) που υλοποιούν τις απεικονιζόμενες διαδικασίες έχει περιγραφεί στις προηγούμενες υποενότητες του παρόντος κεφαλαίου.

2.8.4 Σχετικό έργο

Πρόσφατες εξελίξεις στον τομέα της αναπαράστασης γεγονότων σε κειμενικά περιβάλλοντα έχουν οδηγήσει στην ανάπτυξη παράλληλων ερευνητικών γραμμών που περιλαμβάνουν συστήματα κατανόησης κειμένου και πρόβλεψης γεγονότων για την επικύρωση ισχυρισμών. Αυτά στηρίζονται κυρίως στην αξιοποίηση αλυσίδων αφήγησης γεγονότων (Chambers and Jurafsky, 2008), γράφων (Ciampaglia et al., 2015) και συστημάτων ερωταποκρίσεων (Michael et al., 2018) για την διασύνδεση συμβάντων που περιγράφονται σε ελεύθερο κείμενο. Συναντώνται επίσης τεχνικές που βασίζονται στη διασύνδεση οντοτήτων μέσω της πληροφορίας που παρέχεται από γνωσιακές βάσεις για να αναπαραστήσουν συνεκτικές ροές γεγονότων, αξιοποιώντας είτε μεθόδους αιτιολογικής συλλογιστικής (Radinsky et al., 2012), είτε μεθόδους ερωταποκρίσεων (Jin et al., 2021).

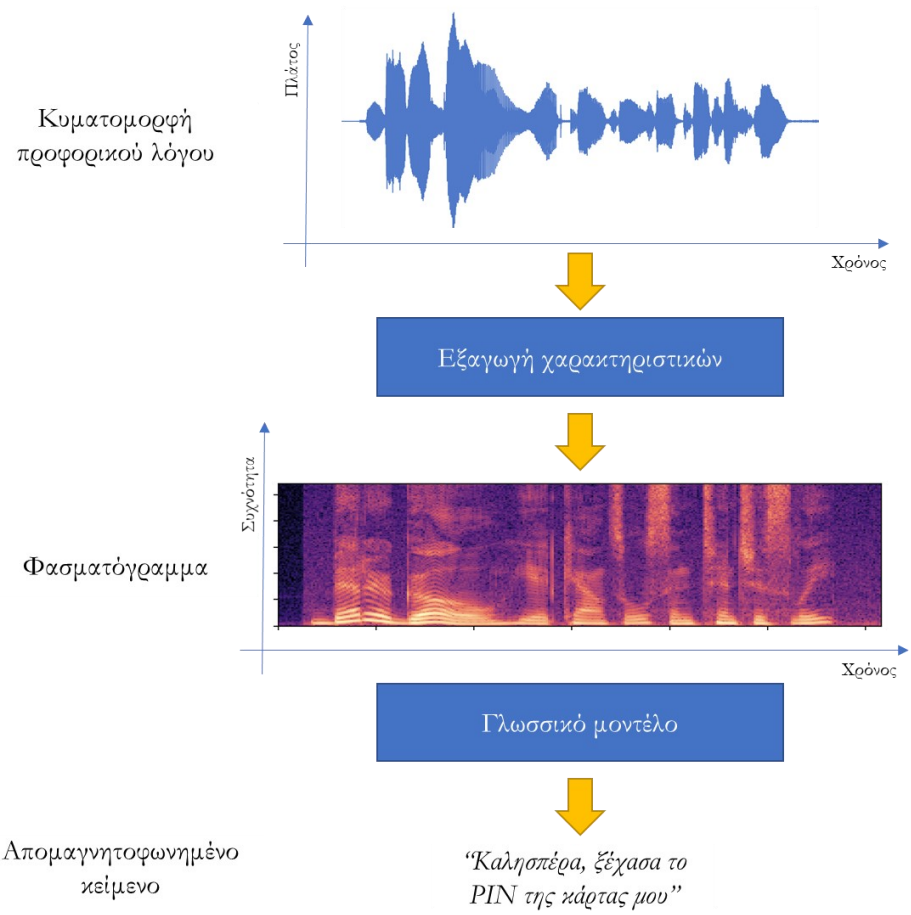
Σημειώνεται ότι οι περισσότερες από τις παραπάνω μεθόδους, είτε βασίζονται στην ύπαρξη μιας αξιόπιστης πηγής που παρέχει επικυρωμένες πληροφορίες (Hassan et al., 2017; Majithia et al., 2019; Zhang et al., 2021a), είτε ζητούν από το χρήστη να εισάγει τις

σχετικές πηγές προκειμένου να προχωρήσουν στον έλεγχο του ισχυρισμού του (Jin et al., 2021). Ακόμη, καμία από τις παραπάνω προσεγγίσεις δεν υποστηρίζει την ελληνική γλώσσα, κυρίως λόγω της έλλειψης επισημασμένων κειμένων για την εκπαίδευση των επιμέρους μηχανισμών μηχανικής μάθησης.

2.9 Αυτόματη αναγνώριση ομιλίας

2.9.1 Ορισμός και βασικά στοιχεία

Η αυτόματη αναγνώριση ομιλίας (automatic speech recognition – ASR), επίσης γνωστή ως μετατροπή ομιλίας σε κείμενο (speech-to-text) αποτελεί διεπιστημονικό υποπεδίο της επεξεργασίας φυσικής γλώσσας και της επιστήμης των υπολογιστών που αφορά την ανάπτυξη μεθόδων που επιτρέπουν την αναγνώριση και μετασχηματισμό του προφορικού λόγου σε κείμενο (Jurafsky and Martin, 2008).



Σχήμα 14 Βασικά στάδια αναγνώρισης ομιλίας

Συνίσταται στην μετατροπή μιας κυματομορφής σε παραμετρικό τύπο αναπαράστασης, συνήθως σε διανύσματα φωνητικών χαρακτηριστικών (πχ. φασματόγραμμα Mel-Frequency

Cepstrum Coefficients – MFCC spectrogram), καθένα από τα οποία αναπαριστά την πληροφορία που αντιστοιχεί σε ένα μικρό χρονικό παράθυρο του σήματος (Davis and Mermelstein, 1980). Τα διανύσματα αυτά αντιστοιχίζονται με τη σειρά τους σε μια ακολουθία λέξεων συνήθως με τη βοήθεια γλωσσικού μοντέλου, παράγοντας το λεκτικό ισοδύναμο της κυματομορφής, το οποίο καλείται απομαγνητοφωνημένο κείμενο (transcript) (Σχήμα 14).

Συναντώνται διάφορες παραλλαγές προβλημάτων αυτόματης αναγνώρισης ομιλίας, μερικές εκ των οποίων λύνονται εύκολα και με μεγάλη ακρίβεια ακόμα και από συστήματα βασισμένα σε κανόνες (πχ. εντοπισμός συγκεκριμένης λέξης, αναγνώριση αριθμών), ωστόσο τα περισσότερα ανοιχτά προβλήματα αφορούν την αναγνώριση μεγάλων αποσπασμάτων προφορικού λόγου ή διαλόγων μεταξύ πολλαπλών ομιλητών και αποτελούν πρόκληση ακόμη και για σύγχρονες μεθόδους βαθιάς μάθησης. Σημειώνεται ότι η αντίστροφη εργασία της αναγνώρισης ομιλίας ονομάζεται σύνθεση λόγου (speech synthesis) και αποτελεί σαφώς ευκολότερη εργασία, καθώς στηρίζεται κατά κύριο λόγο σε προκαθορισμένες διαδικασίες, οδηγώντας σε αποδεκτά αποτελέσματα ακόμα και από συστήματα χαμηλότερης πολυπλοκότητας (Allen, 2003).

2.9.2 Χρησιμότητα και εφαρμογές

Παρόλο που η απομαγνητοφώνηση ομιλιών σε θορυβώδη περιβάλλοντα και από πολλαπλούς ομιλητές βρίσκεται ακόμα σε ερευνητικό στάδιο, οι σχετικές τεχνολογίες έχουν ωριμάσει αρκετά ώστε η χρήση της να είναι βιώσιμη για πολλές πρακτικές εφαρμογές. Δεδομένου ότι η ομιλία αποτελεί την φυσικότερη διεπαφή επικοινωνίας μεταξύ ανθρώπων, τις τελευταίες δεκαετίες έχουν αναπτυχθεί συστήματα αναγνώρισης φωνής για αλληλεπίδραση με έξυπνες οικιακές συσκευές και προσωπικούς βοηθούς (πχ. Google Assistant[10], Alexa[11]), διευκολύνοντας την αναζήτηση πληροφοριών (Yang et al., 2021), την υπαγόρευση κειμένου (Chiu et al., 2018a) και την εκτέλεση απλών εργασιών (Desot et al., 2020). Αντίστοιχες τεχνολογίες αξιοποιούνται για την αυτόματη δρομολόγηση κλήσεων από υπηρεσίες τηλεφωνικών κέντρων (Modipa et al., 2009), την απομαγνητοφώνηση ζωντανών συζητήσεων (πχ. από διαδικτυακές συναντήσεις ή δικαστικές αίθουσες) (Shugrina, 2010) και ως μέσο επαυξητικής επικοινωνίας σε άτομα με αναπηρία (Dabran et al., 2017). Στον

[10] https://assistant.google.com/

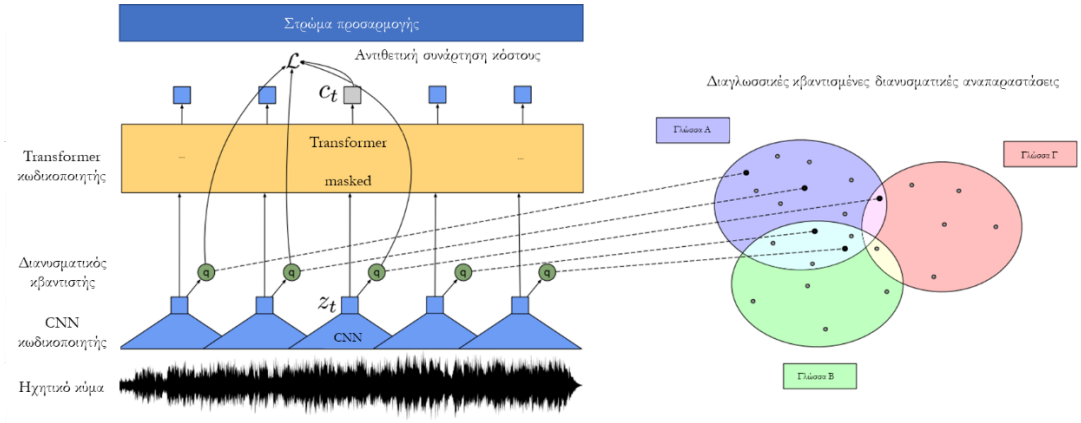
[11] https://developer.amazon.com/en-US/alexa

ιατρικό τομέα, διευκολύνει την αυτόματη καταγραφή συζητήσεων ιατρού-ασθενή (Chiu et al., 2018b) και την καταχώριση σημειώσεων κατά τη διάρκεια εξετάσεων ή συναφών δραστηριοτήτων (Zheng et al., 2011).

2.9.3 Προσαρμογή μοντέλου αναγνώρισης φωνής με Transformer

Στη παρούσα υποενότητα περιγράφεται συνοπτικά η διαδικασία προσαρμογής (finetuning) ενός προεκπαιδευμένου (pretrained) μοντέλου αρχιτεκτονικής XLSR-Wav2Vec2, όπως αυτή αναπτύχθηκε από τους Conneau et al. (2021) και αποτελεί επέκταση της wav2vec 2.0 (Baevski et al., 2020). Η διαδικασία αυτή ακολουθήθηκε για τη δημιουργία αντίστοιχου ελληνικού μοντέλου αναγνώρισης φωνής με αξιοποίηση δημόσια διαθέσιμων απομαγνητοφωνημένων φωνητικών δεδομένων. Το μοντέλο χρησιμοποιήθηκε ως δευτερεύον τρόπος εισαγωγής ισχυρισμών από τον χρήστη, αντικαθιστώντας την πληκτρολόγηση ελεύθερου κειμένου.

Το συγκεκριμένο μοντέλο (Σχήμα 15) μπορεί να προεκπαιδευτεί σε μη επισημασμένα αρχεία ομιλίας διαφόρων γλωσσών, προκειμένου να μάθει διαγλωσσικές, εννοιολογικές διανυσματικές αναπαραστάσεις, ακολουθώντας διαδικασία παρόμοια με αυτή που περιγράφεται στην υποενότητα 2.6 για τις αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή. Στην προκειμένη περίπτωση, βασίζεται σε τεχνικές αυτο-επιβλεπόμενης μάθησης και συγκεκριμένα σε αντιθετική μάθηση (contrastive learning), στοχεύοντας στην εκμάθηση διανυσματικών αναπαραστάσεων όπου τα παρόμοια στοιχεία βρίσκονται κοντά στο διανυσματικό χώρο ενώ τα ανόμοια απομακρύνονται μεταξύ τους. Το μοντέλο περιέχει έναν κωδικοποιητή βασισμένο σε συνελικτικό νευρωνικό δίκτυο (CNN encoder) ο οποίος αντιστοιχίζει δεδομένα ομιλίας Xσε λανθάνουσες διανυσματικές αναπαραστάσεις $z_1 \dots z_T \in Z$. Οι αναπαραστάσεις Z συμπιέζονται σε κβαντισμένες αναπαραστάσεις $q_1 \dots q_T \in Q$, οι οποίες τροφοδοτούνται σε ένα δίκτυο αρχιτεκτονικής Transformer, από το οποίο εξέρχονται οι εννοιολογικές αναπαραστάσεις εξόδου $c_1 \dots c_T \in C$.



Σχήμα 15 Αρχιτεκτονική XLSR-Wav2Vec2

Το μοντέλο προεκπαιδεύεται λύνοντας το παρακάτω πρόβλημα αντιθετικής μάθησης που αφορά τις masked εξόδους του κωδικοποιητή του Transformer: Δεδομένου ενός χρονικού βήματος t και K παρεμβολών Q_t από άλλα masked χρονικά βήματα, σκοπός είναι ταυτοποίηση του σωστού q_T που αντιστοιχεί στο τρέχον χρονικό βήμα. Το παραπάνω εκφράζεται μαθηματικά με την ελαχιστοποίηση της παρακάτω συνάρτησης κόστους:

$$L = \frac{e^{sim(c_t, q_t)}}{\sum_{\tilde{q} \sim Q_t} e^{sim(c_t, \tilde{q})}} \tag{2.11}$$

όπου $sim(\cdot)$ είναι η ομοιότητα συνημιτόνου και c_t η έξοδος του Transformer για το χρονικό βήμα t.

Κατά τη διαδικασία της προσαρμογής, προστίθεται ένα επιπλέον στρώμα (γραμμικό, πρόσθιας τροφοδότησης) στο προεκπαιδευμένο μοντέλο, προκειμένου να καταστεί δυνατή η εκπαίδευσή του σε επισημασμένα δεδομένα (πχ. ηχογραφημένες ομιλίες και τις αντίστοιχες απομαγνητοφωνήσεις τους).

2.9.4 Σχετικό έργο

Το συγκεκριμένο ερευνητικό αντικείμενο έχει συγκεντρώσει μεγάλο ενδιαφέρον τις τελευταίες δεκαετίες, κυρίως λόγω της πρακτικής του χρησιμότητας σε εφαρμογές τελικού χρήστη. Συναντάται πληθώρα προσεγγίσεων που αφορούν το μετασχηματισμό ηχητικών κυμάτων, την εξαγωγή χαρακτηριστικών και τις αρχιτεκτονικές μοντέλων. Η σχετική βιβλιογραφία εκτείνεται από τη χρήση μαρκοβιανών μοντέλων (Baker, 1975; Gales and Young, 2007) και μοντέλων μηχανικής μάθησης (Gales and Young, 2007; Anggraeni et al., 2018) μέχρι την εκμετάλλευση αρχιτεκτονικών βαθιάς μάθησης. Συγκεκριμένα για τις τελευταίες, έχουν χρησιμοποιηθεί αρχιτεκτονικές ανατροφοδοτούμενων δικτύων με ή χωρίς

μηχανισμό προσοχής (Chan et al., 2016), συνελικτικά νευρωνικά δίκτυα (Abdel-Hamid et al., 2014), καθώς και πιο πρόσφατες αρχιτεκτονικές εμπνευσμένες από την βιολογία όπως τα ακμοπυρωδοτούμενα νευρωνικά δίκτυα (spiking neural networks) (Dong and Xu, 2020). Οι τελευταίες προσπάθειες επικεντρώνονται στη χρήση κωδικοποιητών Transformers καθώς και σε διαδικασίες στοίβαξης πολλαπλών στρωμάτων από διάφορες αρχιτεκτονικές για την παραγωγή αποδοτικότερων μεθόδων (Gulati et al., 2020; Conneau et al., 2021). Σημειώνεται ότι, παρά τις καινοτόμες προσεγγίσεις στην κατασκευή νέων μοντέλων, η έλλειψη επισημασμένων δεδομένων εκπαίδευσης για γλώσσες χαμηλότερων πόρων όπως η ελληνική εξακολουθεί να αποτελεί τροχοπέδη για την αποτελεσματική εκπαίδευσή τους, όπως συμβαίνει και στις περισσότερες εργασίες επεξεργασίας φυσικής γλώσσας.

2.10 Σύνοψη κεφαλαίου

Στο κεφάλαιο αυτό έγινε αναφορά στο θεωρητικό πλαίσιο βασικών εργασιών επεξεργασίας φυσικής γλώσσας, οι οποίες συνιστούν τους επιμέρους μηχανισμούς που αξιοποιήθηκαν για την εκπλήρωση των ερευνητικών στόχων που παρουσιάστηκαν στο Κεφάλαιο 1. Για κάθε εργασία παρατέθηκαν οι σχετικοί ορισμοί και βασικές πληροφορίες, εστιάζοντας παράλληλα στο μαθηματικό υπόβαθρο στοχευμένων δραστηριοτήτων που υλοποιούνται στα πλαίσια της διατριβής. Παρουσιάστηκαν ακόμη ενδεικτικές εφαρμογές και παρατέθηκαν βιβλιογραφικές πληροφορίες για τις κυριότερες προσεγγίσεις στο εκάστοτε πεδίο.

Στο επόμενο κεφάλαιο περιγράφεται η διασύνδεση των παραπάνω εργασιών με τους βασικούς ερευνητικούς άξονες και παρουσιάζεται η σχετική μεθοδολογία, η οποία περιλαμβάνει το συνδυασμό επιμέρους μηχανισμών στη συνολική αρχιτεκτονική που υλοποιήθηκε για την εκπλήρωση των στόχων της διατριβής.

3 Μεθοδολογία & Σχεδιασμός

3.1 Εισαγωγή

Το θεωρητικό υπόβαθρο που αναπτύχθηκε στο προηγούμενο κεφάλαιο αποτελεί το εφαλτήριο για την ανάπτυξη αντίστοιχων μεθοδολογιών για την ελληνική γλώσσα, με απώτερο σκοπό τη δημιουργία ενός συστήματος ανίχνευσης λανθανουσών συσχετίσεων και προτύπων μέσω φυσικής γλώσσας, το οποίο θα αξιοποιεί τον πληροφοριακό πλούτο ελληνικών διαδικτυακών πηγών ώστε να αναγνωρίζει και να ακολουθεί την αλληλουχία εμφάνισης προγενέστερα ασυσχέτιστων γεγονότων, παρέχοντας ιδανικά τη δυνατότητα πρόβλεψης παρόμοιων γεγονότων στο μέλλον. Στο παρόν κεφάλαιο, τα ζητήματα και οι υποθέσεις εργασίας του Κεφαλαίου 1 μετασχηματίζονται σε συγκεκριμένους ερευνητικούς άξονες, οι οποίοι χωρίζονται σε επιμέρους ερευνητικά επίκεντρα (διακριτούς μηχανισμούς), καθένα εκ των οποίων συμβάλει στην εκπλήρωση συγκεκριμένων στόχων της διατριβής. Ακολουθεί η αναλυτική περιγραφή της αρχιτεκτονικής καθώς και των μεθοδολογικών επιλογών και παραδοχών που συνοδεύουν κάθε επιμέρους μηχανισμό, με αναφορές στο θεωρητικό υπόβαθρο του προηγούμενου κεφαλαίου, όπου κρίνεται απαραίτητο.

Η δομή του παρόντος κεφαλαίου έχει ως εξής: Η επόμενη ενότητα αποτελεί επισκόπηση των δύο βασικών ερευνητικών αξόνων της διατριβής, αντιστοιχίζοντας τους με τις υποθέσεις εργασίας που παρατέθηκαν στο Κεφάλαιο 1. Ακολουθούν δύο ενότητες (μία για κάθε κύριο ερευνητικό άξονα), στις οποίες περιγράφεται η προτεινόμενη προσέγγιση καθώς και οι επιμέρους μηχανισμοί που σχεδιάστηκαν στα πλαίσια της συνολικής μεθοδολογίας.

Δεδομένου ότι ακολουθήθηκε δομοστοιχειωτός σχεδιασμός για την εκπλήρωση κάθε ερευνητικού άξονα, καθεμία από τις δύο αυτές ενότητες χωρίζεται σε υποενότητες ισάριθμες με τα επιμέρους στοιχεία που τον απαρτίζουν. Προς διευκόλυνση του αναγνώστη, έγινε προσπάθεια η δομή του παρόντος κεφαλαίου να αντιστοιχεί σε συγκεκριμένες υποενότητες του Κεφαλαίου 2, προκειμένου να συσχετίζεται ξεκάθαρα το προαπαιτούμενο θεωρητικό υπόβαθρο που διαδραμάτισε κομβικό ρόλο στην ανάπτυξη της εκάστοτε μεθοδολογικής προσέγγισης.

3.2 Βασικοί ερευνητικοί άξονες

Οι βασικοί ερευνητικοί άξονες που προτείνονται για την εκπλήρωση των υποθέσεων εργασίας του Κεφαλαίου 1, είναι οι εξής:

I. Εξόρυξη πληροφοριών από ελεύθερο κείμενο
II. Εξαγωγή σημασιολογικών συμπερασμάτων από διαδικτυακές πηγές

Ο πρώτος ερευνητικός άξονας αφορά το σχεδιασμό συστήματος ανοιχτής εξαγωγής πληροφοριών από ελεύθερο κείμενο. Η εξαγωγή πληροφοριών αποτελεί αναγκαία προϋπόθεση για τη διύλιση σημασιολογικών συσχετίσεων μεταξύ οντοτήτων από μεγάλο σώμα κειμένων και την αποτύπωσή τους σε δομημένη μορφή. Η σημασία της συγκεκριμένης εργασίας μπορεί να παρομοιαστεί με αυτή των τεχνικών Διερευνητικής Ανάλυσης Δεδομένων (EDA-Exploratory Data Analysis) που εφαρμόζονται συνήθως σε δομημένα δεδομένα για τη διαμόρφωση αρχικών υποθέσεων, τη συνόψιση βασικών χαρακτηριστικών και την εξερεύνηση λανθανουσών συσχετίσεων που δεν είναι διακριτές εκ πρώτης όψεως. Επισημαίνεται ότι, ακόμα και για μια τόσο στοιχειώδη εργασία, παρατηρείται σημαντικό κενό στην ανάπτυξη αντίστοιχης μεθοδολογικής προσέγγισης που να επιτρέπει την εξαγωγή πληροφοριών για γλώσσες χαμηλών πόρων όπως η ελληνική. Ως εκ τούτου, η παρούσα διατριβή επιχειρεί να συμπληρώσει αυτό το κενό με το σχεδιασμό ενός αρθρωτού μηχανισμού που επιτρέπει την δημιουργία σημασιολογικών τριπλετών από ελεύθερο κείμενο. Ο συγκεκριμένος ερευνητικός άξονας απαντά στα ακόλουθα ερευνητικά ζητήματα και υποθέσεις εργασίας του Κεφαλαίου 1: Ζ1, Ζ2, Υ3.

Ο δεύτερος ερευνητικός άξονας αφορά την επικύρωση ισχυρισμών μέσω της αξιοποίησης συσσωρευμένης πληροφορίας από διαδικτυακές ειδησεογραφικές πηγές. Περιλαμβάνει την ανάπτυξη μιας μεθοδολογίας που θα επιτρέπει την αξιολόγηση της εγκυρότητας ενός ισχυρισμού (σε ελεύθερο κείμενο) μέσω της εύρεσης συγκεκριμένων υποσυνόλων της πληροφορίας που σχετίζονται με αυτόν (τεκμήρια) και μπορούν είτε να τον επιβεβαιώσουν

είτε να τον αντικρούσουν. Αξίζει να σημειωθεί ότι η ανάπτυξη ενός αντίστοιχου μηχανισμού περιλαμβάνει επιμέρους στοιχεία συλλογής πληροφοριών και σύνδεσης οντοτήτων, καθώς και μεθόδους ελέγχου σημασιολογικής ομοιότητας και αναγνώρισης κειμενικής συνεπαγωγής, οι οποίες δεν έχουν αναπτυχθεί για την ελληνική γλώσσα. Επομένως, στόχος του συγκεκριμένου ερευνητικού άξονα δεν είναι μόνο ο σχεδιασμός μιας μεθοδολογίας που θα συνδυάζει αποδοτικά τα επιμέρους στοιχεία, αλλά και η ανάπτυξή των αντίστοιχων στοιχείων (νευρωνικών μοντέλων) για την ελληνική γλώσσα. Ο συγκεκριμένος ερευνητικός άξονας απαντά στα ακόλουθα ερευνητικά ζητήματα και υποθέσεις εργασίας: Z3, Y1, Y2, Y4, Y5.

3.3 Εξόρυξη πληροφοριών από ελεύθερο κείμενο

Η εξαγωγή πληροφοριών από ημιδομημένες και αδόμητες πηγές δεδομένων (πχ. επιστημονικές δημοσιεύσεις, ειδησεογραφικά κείμενα κλπ.) αποτελεί βασικό προαπαιτούμενο για την αποτύπωση και εξερεύνηση της περιεχόμενης σε αυτές γνώσης, καθώς επιτρέπει την εξαγωγή συσχετίσεων μεταξύ οντοτήτων και την οργάνωση τους σε δομημένη μορφή. Προκειμένου να αποδελτιωθεί και να καταστεί εκμεταλλεύσιμη η λανθάνουσα πληροφορία που συγκεντρώνεται σε ελεύθερα κείμενα πληροφοριακού χαρακτήρα, συνήθως αξιοποιούνται συστήματα ανοιχτής εξαγωγής πληροφοριών, όπως αυτά που περιεγράφηκαν στην υποενότητα 2.2.3. Ωστόσο, ενώ αντίστοιχες τεχνικές εξαγωγής μπορούν εύκολα να αυτοματοποιηθούν για γλώσσες υψηλών πόρων (πχ. αγγλικά, γερμανικά) για τις οποίες διατίθεται πληθώρα επισημασμένων δεδομένων εκπαίδευσης, η έλλειψη σχετικών δεδομένων για γλώσσες χαμηλότερων πόρων όπως η ελληνική, καθιστά αδύνατη την απευθείας εκπαίδευση αντίστοιχων συστημάτων μηχανικής μάθησης. Γίνεται, λοιπόν φανερή η αναγκαιότητα ανάπτυξης μιας εναλλακτικής μεθοδολογίας που θα επιτρέψει την εξαγωγή πληροφοριών σε δομημένη μορφή για την ελληνική γλώσσα, η οποία δε θα στηρίζεται σε επισημασμένα δεδομένα. Η μεθοδολογία που αναπτύχθηκε στα πλαίσια της εργασίας είναι της μορφής pipeline, καθώς περιλαμβάνει τη σειριακή εκτέλεση διαφόρων σταδίων επεξεργασίας, και περιγράφεται στην επόμενη υποενότητα.

3.3.1 Προτεινόμενη προσέγγιση

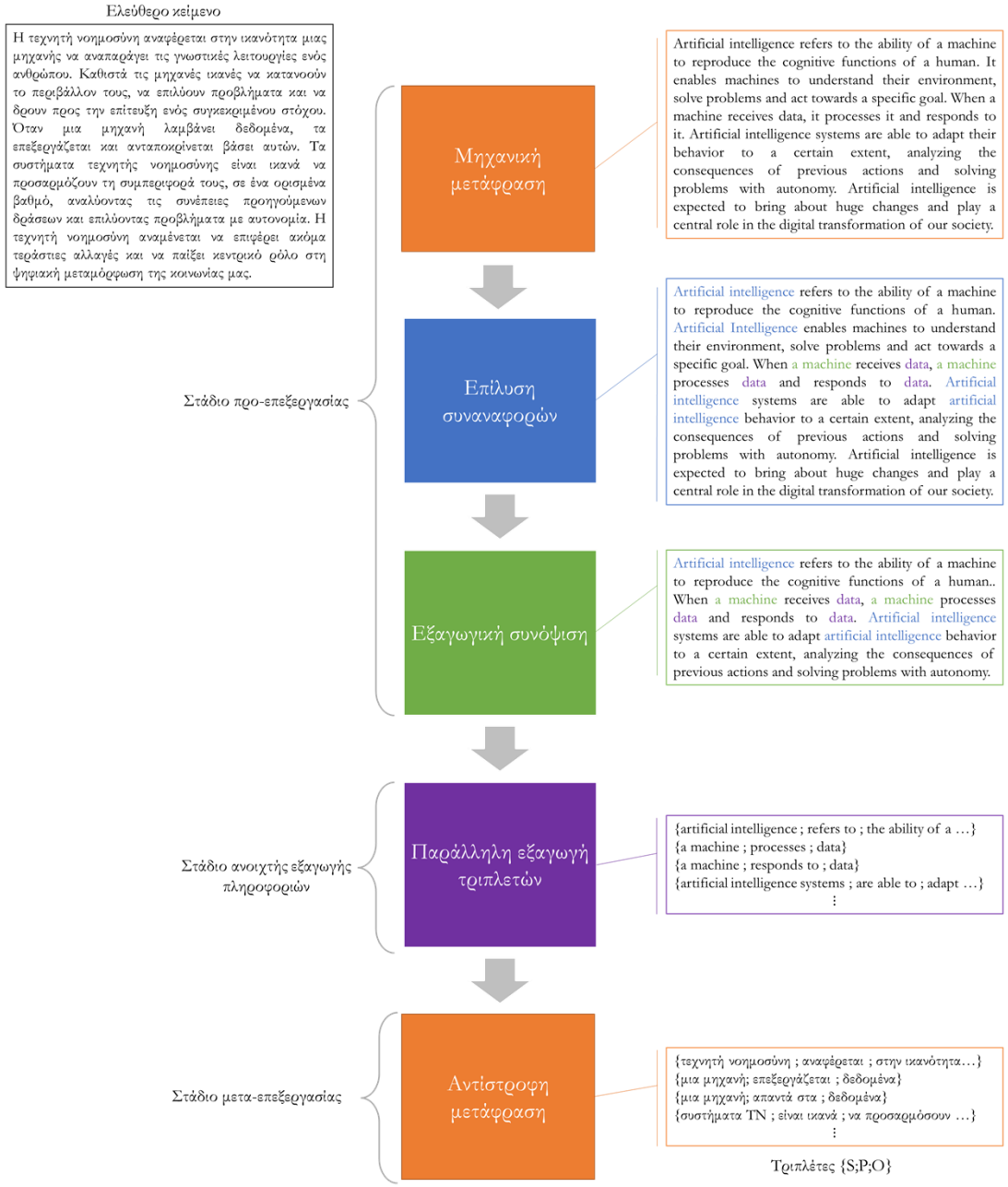
Η προτεινόμενη μεθοδολογία εξαγωγής πληροφοριών από πηγές ελεύθερου κειμένου βασίζεται σε ένα σύνολο εργασιών προ-επεξεργασίας που στοχεύουν στην ενεργοποίηση μηχανισμών ανοιχτής εξαγωγής πληροφοριών (OIE) για την ελληνική γλώσσα. Συγκεκριμένα, εκμεταλλεύεται την ύπαρξη μεγάλου όγκου παράλληλου κειμένου (κείμενο

που διατίθεται σε παραπάνω από μια γλώσσες, πχ. αγγλικά και ελληνικά) ως προϊόν συσσώρευσης από διαγλωσσικούς πόρους (πχ. υπότιτλοι ταινιών, μεταφράσεις διαδικτυακού περιεχομένου, επίσημα έγγραφα της Ευρωπαϊκής Ένωσης μεταφρασμένα σε κάθε γλώσσα της κοινότητας, κτλ.) για την εκπαίδευση μηχανισμού μηχανικής μετάφρασης (machine translation) από τα ελληνικά στα αγγλικά και αντίστροφα. Κατ' αυτόν τον τρόπο, το σύνολο της επεξεργασίας λαμβάνει χώρα στο αγγλικό ισοδύναμο του αρχικού κειμένου. Η διαδικασία υποστηρίζεται από επιπλέον εργασίες προ-επεξεργασίας, που στοχεύουν στην βελτίωση των αποτελεσμάτων εξαγωγής. Συγκεκριμένα, περιλαμβάνει μηχανισμό επίλυσης συναναφορών, για την αντικατάσταση των μερών του λόγου με την οντότητα-σημείο αναφοράς στην οποία αναφέρονται. Ακόμη, στην μεθοδολογία έχει ενσωματωθεί μηχανισμός εξαγωγικής συνόψισης, προκειμένου να εστιαστεί η άντληση πληροφοριών μόνο στα σημαντικότερα τμήματα του αρχικού κειμένου.

Όσον αφορά το κύριο κομμάτι της μεθοδολογίας, αυτό υποστηρίζεται από μηχανισμούς παράλληλης εξαγωγής πληροφοριών που αναπτύχθηκαν στα πλαίσια της εργασίας και αποτελούν συνδυασμό προεκπαιδευμένων μεθόδων μηχανικής μάθησης με αντίστοιχους που βασίζονται σε γλωσσολογικούς κανόνες. Ο λόγος συνύπαρξης δύο φαινομενικά επικαλυπτόμενων μηχανισμών έγκειται στη συμπληρωματικότητα της κάθε προσέγγισης όσον αφορά την εξαγωγή τριπλετών από ελεύθερο κείμενο: Συγκεκριμένα, οι μηχανισμοί εξαγωγής που βασίζονται σε γλωσσολογικούς κανόνες δίνουν έμφαση στην παραγωγή αξιόπιστων αποτελεσμάτων, με αποτέλεσμα να δίνουν έμφαση στην ακρίβεια (precision) των εξαγωγών, παραλείποντας ωστόσο αποτελέσματα που δεν συνεπάγονται από την εφαρμογή των προκαθορισμένων κανόνων. Αντίθετα, οι μηχανισμοί εξαγωγής που βασίζονται σε μηχανική μάθηση προσφέρουν μεγαλύτερη ευελιξία καθώς είναι σε θέση να ανιχνεύσουν πρότυπα παρόμοια με αυτά στα οποία εκπαιδεύτηκαν, οδηγώντας έτσι σε προσεγγίσεις προσανατολισμένες στην ανάκληση (precision). Ωστόσο, η στοχαστικότητα της διαδικασίας που ακολουθείται από τους μηχανισμούς αυτούς ενδέχεται να οδηγήσει σε ψευδώς θετικές εξαγωγές, δηλαδή σε τριπλέτες που δεν συμβάλλουν στην ορθή αναπαράσταση του αρχικού κειμένου. Ο ακρογωνιαίος λίθος της εργασίας εξόρυξης πληροφοριών που περιγράφεται στην παρούσα ενότητα στηρίζεται στο σχεδιασμό μιας γενικευμένης μεθοδολογίας εξαγωγής τριπλετών που εξισορροπεί την ακρίβεια και την ανάκληση των προαναφερθεισών προσεγγίσεων μέσω του συνδυασμού των αποτελεσμάτων τους σε ένα ενοποιημένο σύνολο.

Το αποτέλεσμα της διαδικασίας εξαγωγής αποτελείται από τριπλέτες της μορφής υποκείμενο-κατηγόρημα-αντικείμενο που αντιστοιχούν στο αγγλικό ισοδύναμο του αρχικού κειμένου. Επομένως, απαιτείται ο κατάλληλος μετασχηματισμός τους στα ελληνικά μέσω

μια διαδικασίας μετα-επεξεργασίας, που περιλαμβάνει την αντίστροφη μετάφρασή τους (back-translation). Αυτό καθίσταται δυνατό μέσω της εκπαίδευσης μηχανισμού μηχανικής μετάφρασης από τα αγγλικά στα ελληνικά, παρόμοιο με αυτόν που χρησιμοποιείται στο στάδιο προ-επεξεργασίας. Εναλλακτικά, προτείνεται μηχανισμός ευθυγράμμισης λέξεων (word alignment) που βασίζεται σε διαγλωσσικές διανυσματικές αναπαραστάσεις. Ο μηχανισμός αυτός δέχεται την αρχική (στα ελληνικά) και μεταφρασμένη (στα αγγλικά) πρόταση από το στάδιο προ-επεξεργασίας και αντιστοιχίζει απευθείας τις αγγλικές φράσεις που αποτελούν το υποκείμενο, το κατηγόρημα και το αντικείμενο της τριπλέτας με τις ισοδύναμες ελληνικές που συναντώνται στο αρχικό κείμενο.



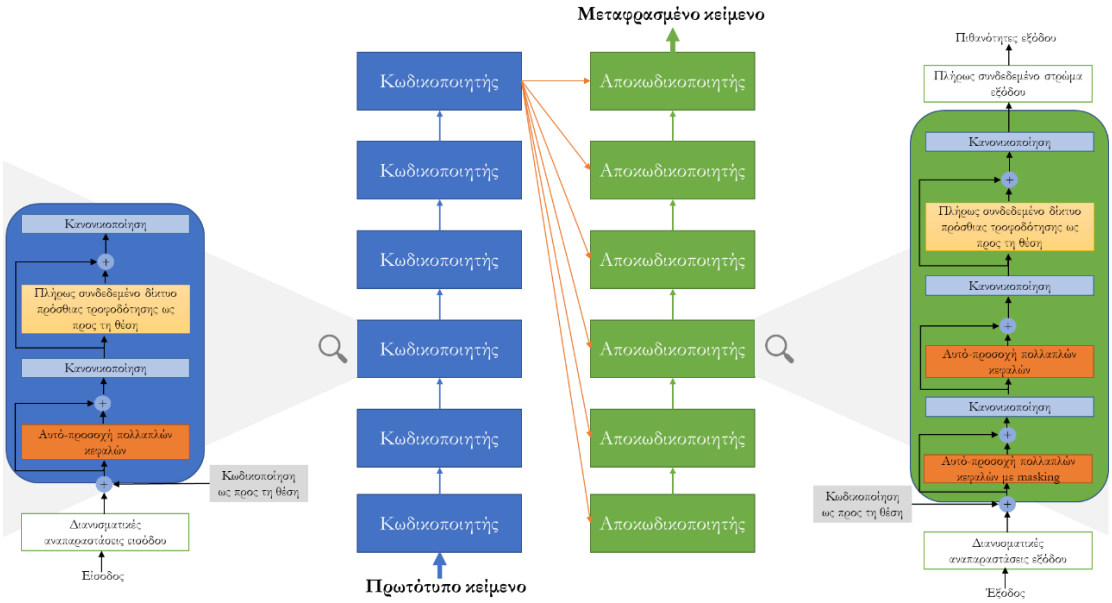
Σχήμα 16 Επισκόπηση προτεινόμενης μεθοδολογίας εξαγωγής πληροφοριών από ελεύθερο κείμενο

Η επισκόπηση της παραπάνω διαδικασίας απεικονίζεται στο Σχήμα 16 με τη βοήθεια ενός παραδείγματος. Στις ακόλουθες υποενότητες περιγράφεται αναλυτικά ο σχεδιασμός και η μεθοδολογία κάθε μηχανισμού που ενσωματώνεται στην προτεινόμενη προσέγγιση εξαγωγής πληροφοριών για ελληνικό ελεύθερο κείμενο, ενώ οι τεχνικές λεπτομέρειες που αφορούν την υλοποίηση και τη δοκιμή της παρουσιάζονται στο Κεφάλαιο 4.

3.3.1.1 Προεπεξεργασία κειμένου

3.3.1.1.1 Μηχανική Μετάφραση

Για την μετάφραση ελληνικού κειμένου σε αγγλικό ισοδύναμο γίνεται εκπαίδευση νευρωνικού μοντέλου μηχανικής μετάφρασης και συγκεκριμένα μια παραλλαγή της αρχιτεκτονικής Transformer που περιεγράφηκε στην υποενότητα 2.3.3.1. Τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής αποτελούνται από 6 στοιβαγμένα στρώματα που ενσωματώνουν υποστρώματα αυτο-προσοχής πολλαπλών κεφαλών συνοδευόμενα από πλήρως συνδεδεμένα δίκτυα πρόσθιας τροφοδότησης, όπως φαίνεται στο Σχήμα 17. Μια βασική διαφορά μεταξύ της πρωτότυπης αρχιτεκτονικής (Vaswani et al., 2017) και της παρούσας είναι ότι η διάσταση του εσωτερικού στρώματος κάθε δικτύου πρόσθιας τροφοδότησης έχει μειωθεί σε $d_{ff,} = 1024$ έναντι της αρχικής $d_{ff} = 2048$, ως αποτέλεσμα της προσπάθειας μείωσης της απαιτούμενης μνήμης, δεδομένων των περιορισμένων διαθέσιμων υπολογιστικών πόρων.



Σχήμα 17 Επισκόπηση μοντέλου μηχανικής μετάφρασης

Επιπλέον, η προτεινόμενη προσέγγιση μηχανικής μετάφρασης ενσωματώνει την ανίχνευση λεκτικών μονάδων του κειμένου εκπαίδευσης σε επίπεδο υπο-λέξεων (sub-word tokenization), με χρήση της μεθόδου byte-pair encoding (BPE) (Sennrich et al., 2016b). Καθώς η ελληνική γλώσσα θεωρείται μορφολογικά πλούσια (περιέχει πολλούς κλιτούς τύπους λέξεων) και χαρακτηρίζεται από πλούσιο λεξιλόγιο, η απευθείας αναπαράσταση κάθε λέξης σε πεπερασμένο διανυσματικό χώρο (το embedding size του Transformer ισούται με 512) αποτελεί πρόκληση για οποιαδήποτε αρχιτεκτονική μηχανικής μετάφρασης. Συγκεκριμένα:

i. αυξάνει δραματικά το μέγεθος του μοντέλου (πχ. για λεξιλόγιο 300000 λέξεων απαιτούνται $300000 \cdot 512 \cdot 4\,bytes \cong 600MB$ αποθηκευτικού χώρου μόνο για τη διανυσματική αναπαράσταση εισόδου)
ii. καθιστά αδύνατη την αναπαράσταση λέξεων εκτός του αρχικού λεξιλογίου (Out-Of-Vocabulary words – OOV), στις οποίες συγκαταλέγονται όλοι οι δυνατοί κλιτοί τύποι κάθε λέξης.

Η μέθοδος BPE πραγματοποιεί συγχώνευση των πιο συχνά εμφανιζόμενων υπο-λέξεων, με τέτοιο τρόπο που η κάθε συμβολοσειρά να αντιστοιχεί σε μοναδική μονάδα μνήμης, οδηγώντας σε σημαντική συμπίεση των διανυσματικών αναπαραστάσεων του μοντέλου. Για παράδειγμα, η συμβολοσειρά "καλ" μπορεί να αποτελέσει μια sub-word μονάδα, η οποία θα χρησιμοποιηθεί για την αναπαράσταση όλων των ακόλουθων λέξεων "καλός, καλύτερος, καλοκαίρι, καλοσύνη" σε συνδυασμό με άλλες υπο-λέξεις. Η μέθοδος BPE επιλέγει αυτόματα τις κατάλληλες συγχωνεύσεις χαρακτήρων για το σχηματισμό υπο-λέξεων ακολουθώντας τα παρακάτω βήματα:

1. Καθορισμός του μέγιστου αριθμού συγχωνεύσεων (ή του μέγιστου αριθμού υπο-λέξεων) ως υπερ-παράμετρο και αρχικοποίηση του λεξιλογίου για κάθε γλώσσα του παράλληλου κειμένου.
2. Αναπαράσταση κάθε λέξης στο σύνολο κειμένου ως συνδυασμό χαρακτήρων. Το σύμβολο <\w> χρησιμοποιείται για να υποδηλώσει το τέλος κάθε λέξης.
3. Επαναληπτική μέτρηση της συχνότητας όλων των n-grams χαρακτήρων (ξεκινώντας από bigrams) που εμφανίζονται σε όλες τις λέξεις του λεξιλογίου.
4. Συγχώνευση των εμφανίσεων του πιο συχνού n-gram και πρόσθεσή του στο λεξιλόγιο υπο-λέξεων.
5. Επανάληψη του παραπάνω βήματος, έως ότου συμπληρωθεί ο μέγιστος αριθμός συγχωνεύσεων ή έως ότου εκπληρωθεί το προκαθορισμένο μέγεθος λεξιλογίου.

Η παραπάνω μέθοδος δημιουργεί ένα λεξιλόγιο υπο-λέξεων, το οποίο χρησιμοποιείται για να αναπαραστήσει το σύνολο δεδομένων που τροφοδοτείται ως παράλληλο κείμενο στο μηχανισμό μηχανικής μετάφρασης. Σημειώνεται ότι η διαδικασία εκτελείται δύο φορές, τόσο για την κωδικοποίηση της αρχικής γλώσσας όσο και της γλώσσας-στόχου. Καθώς οι παραγόμενες διανυσματικές αναπαραστάσεις καταλαμβάνουν σημαντικά λιγότερο χώρο για το ίδιο λεξιλόγιο σε σχέση με τις διανυσματικές αναπαραστάσεις λέξεων, η συγκεκριμένη μέθοδος προτιμάται για την κωδικοποίηση μεγάλων συνόλων κειμένου. Παράλληλα, επιτρέπει την κωδικοποίηση σπάνιων ή άγνωστων λέξεων ως συσσωμάτωση δύο ή περισσότερων υπο-λέξεων, οδηγώντας σε σημαντικά καλύτερη απόδοση μετάφρασης σε σχέση με άλλες μεθόδους διανυσματικής αναπαράστασης.

Η χρήση της παραπάνω μεθοδολογίας για την εκπαίδευση μοντέλου μετάφρασης από τα ελληνικά στα αγγλικά (και αντίστροφα) προϋποθέτει την ύπαρξη παράλληλου κειμένου. Κατά την εκπαίδευση, κάθε ακολουθία (πρόταση) του πρωτότυπου κειμένου τροφοδοτείται στον κωδικοποιητή, ενώ η αντίστοιχη ακολουθία του μεταφρασμένου κειμένου τροφοδοτείται στον αποκωδικοποιητή, μετατοπισμένη κατά μια θέση προς τα δεξιά. Κατ' αυτό τον τρόπο, το μοντέλο εκπαιδεύεται να προβλέπει τη λέξη που αντιστοιχεί στη θέση i του πρωτότυπου κειμένου, λαμβάνοντας υπόψη μόνο τη μεταφρασμένη ακολουθία που αντιστοιχεί στη θέση $i-1$ του αποκωδικοποιητή. Για τον ίδιο λόγο, χρησιμοποιείται masking στο πρώτο υπόστρωμα αυτό-προσοχής πολλαπλών κεφαλών του αποκωδικοποιητή, αποτρέποντας το μοντέλο να έχει πρόσβαση σε "μελλοντικά" στοιχεία της ακολουθίας.

Κατά τη διαδικασία συμπερασμού (inference), το εκπαιδευμένο μοντέλο δέχεται μια πρόταση σε ελεύθερο κείμενο, τη μετασχηματίζει σε υπο-λέξεις βάσει του προκαθορισμένου λεξιλογίου BPE και την εισάγει τμηματικά (ανά token) στα στρώματα του κωδικοποιητή. Ο αποκωδικοποιητής τροφοδοτείται επίσης αρχικά με μια κενή ακολουθία που περιλαμβάνει μόνο ένα σύμβολο το οποίο συμβολίζει την αρχή της (πχ. <BOS>). Καθώς ο κωδικοποιητής τροφοδοτείται με το token που αντιστοιχεί στην πρώτη θέση της ακολουθίας εισόδου, ο αποκωδικοποιητής αναλαμβάνει να το μεταφράσει και να το τοποθετήσει στη δεύτερη θέση της ακολουθίας εξόδου (μετά το σύμβολο αρχής). Στη συνέχεια, η δεύτερη λέξη της ακολουθίας εισόδου περνάει από τα στρώματα του κωδικοποιητή και δίνεται στον αποκωδικοποιητή μαζί με την τρέχουσα ακολουθία εξόδου, ώστε να προβλεφθεί η επόμενη λέξη της ακολουθίας. Η διαδικασία επαναλαμβάνεται έως ότου ο αποκωδικοποιητής προβλέψει ένα σύμβολο τέλους (πχ. <EOS>), το οποίο σηματοδοτεί το τέλος της πρότασης. Για την πρόβλεψη της καλύτερης δυνατής μετάφρασης, το μοντέλο χρησιμοποιεί τον αλγόριθμο ακτινικής αναζήτησης (beam search) (Medress et al., 1977), ο οποίος

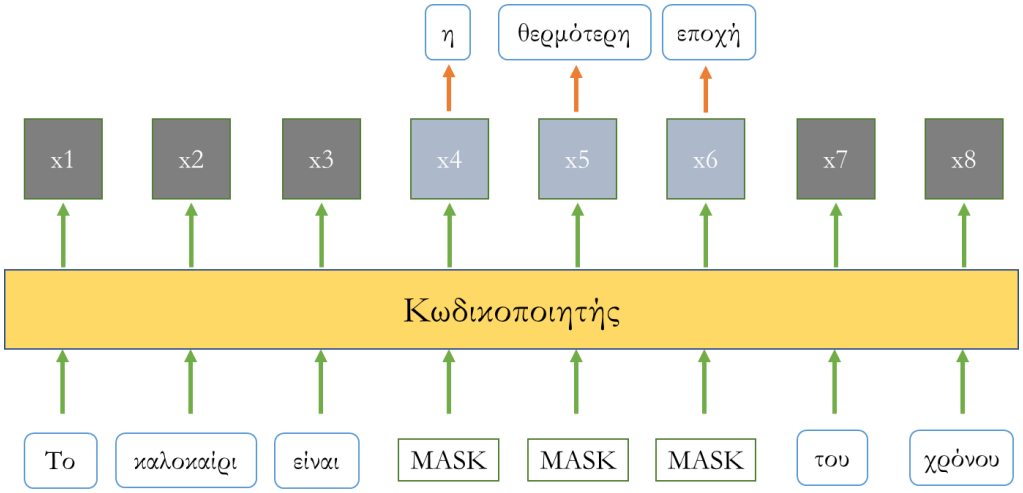
κατασκευάζει διαφορετικές μεταφράσεις για κάθε ακολουθία εισόδου, επιλέγοντας σε κάθε βήμα t την επόμενη λέξη, βάσει της δεσμευμένης πιθανότητας να προβλεφθεί η λέξη αυτή, δεδομένης της εισόδου και της πρόβλεψης στο προηγούμενο χρονικό βήμα $t-1$. Ο αριθμός των δυνατών συνδυασμών μπορεί να καθοριστεί από σχετική υπεπαράμετρο k που ονομάζεται πλάτος αναζήτησης (beam width). Για $k = 1$, επιστρέφεται μόνο η ακολουθία που βασίζεται στις καλύτερες επιμέρους προβλέψεις σε κάθε βήμα, καθώς ο αλγόριθμος εκφυλίζεται σε άπληστη αναζήτηση (greedy search).

3.3.1.1.2 Επίλυση συναναφορών

Δεδομένου ότι το αρχικό κείμενο που πρόκειται να υποβληθεί σε εξαγωγή πληροφοριών έχει πλέον μεταφραστεί στο αγγλικό του ισοδύναμο, είναι δυνατή η χρήση πρόσθετων τεχνικών προ-επεξεργασίας που βασίζονται σε προεκπαιδευμένα γλωσσικά μοντέλα για την αγγλική γλώσσα. Το στάδιο αυτό περιλαμβάνει την τροφοδότηση του κειμένου σε έναν μηχανισμό επι-τόπου επίλυσης συναναφορών, ώστε να αντιμετωπιστούν περιπτώσεις όπου μια οντότητα έχει αντικατασταθεί από μια λέξη ή φράση, η οποία δεν προσφέρει ιδιαίτερη πληροφορία ως μεμονωμένο λεκτικό σύνολο στη διαδικασία εξαγωγής πληροφοριών. Για παράδειγμα, στη φράση "*Η επιστημονική κοινότητα ανακοίνωσε μια νέα επαναστατική θεραπεία κατά του καρκίνου, η οποία μπορεί να γενικευτεί για εκατοντάδες διαφορετικούς τύπους.*", η αντικατάσταση της φράσης "η οποία" με τη φράση αναφοράς "νέα επαναστατική θεραπεία κατά του καρκίνου", θα οδηγήσει στην δημιουργία τριπλέτας SPO της μορφής {νέα επαναστατική θεραπεία κατά του καρκίνου ; γενικεύεται για ; εκατοντάδες διαφορετικούς τύπους}, αντί της λιγότερο χρήσιμης {η οποία ; γενικεύεται για ; εκατοντάδες διαφορετικούς τύπους}.

Γίνεται χρήση προεκπαιδευμένου νευρωνικού μοντέλου επίλυσης συναναφορών που στηρίζεται σε παραλλαγή της αρχιτεκτονικής που προτάθηκε από τους Lee et al. (2017) και περιγράφεται στην υποενότητα 2.5.3. Η διαφορά της πρωτότυπης αρχιτεκτονικής με αυτή που χρησιμοποιείται στην παρούσα εργασία έγκειται στην αντικατάσταση των GloVe embeddings με αυτά του SpanBERT (Joshi et al., 2020), αξιοποιώντας μηχανισμό προσοχής ώστε ο υπολογισμός της διανυσματικής αναπαράστασης κάθε token να βασίζεται στα συμφραζόμενά του. Σημειώνεται ότι οι ιδιαιτερότητες της Transformer αρχιτεκτονικής SpanBERT από τη BERT (Devlin et al., 2019b) είναι οι ακόλουθες: αφενός στο SpanBERT πραγματοποιείται κάλυψη (masking) συνεχόμενων διαστημάτων του κειμένου (spans) αντί μεμονωμένων τυχαίων tokens κατά την εκπαίδευση, και αφετέρου το BERT έχει έναν επιπλέον στόχο εκπαίδευσης πέραν της πρόβλεψης του mask token στην έξοδο,

και συγκεκριμένα την πρόβλεψη για το αν δυο ακολουθίες κειμένου αποτελούν συνέχεια η μία της άλλης (Next Sequence Prediction). Ο στόχος εκπαίδευσης του SpanBERT απεικονίζεται στο Σχήμα 18:



Σχήμα 18 Masking συνεχόμενων διαστημάτων κειμένου κατά την εκπαίδευση αρχιτεκτονικής SpanBERT

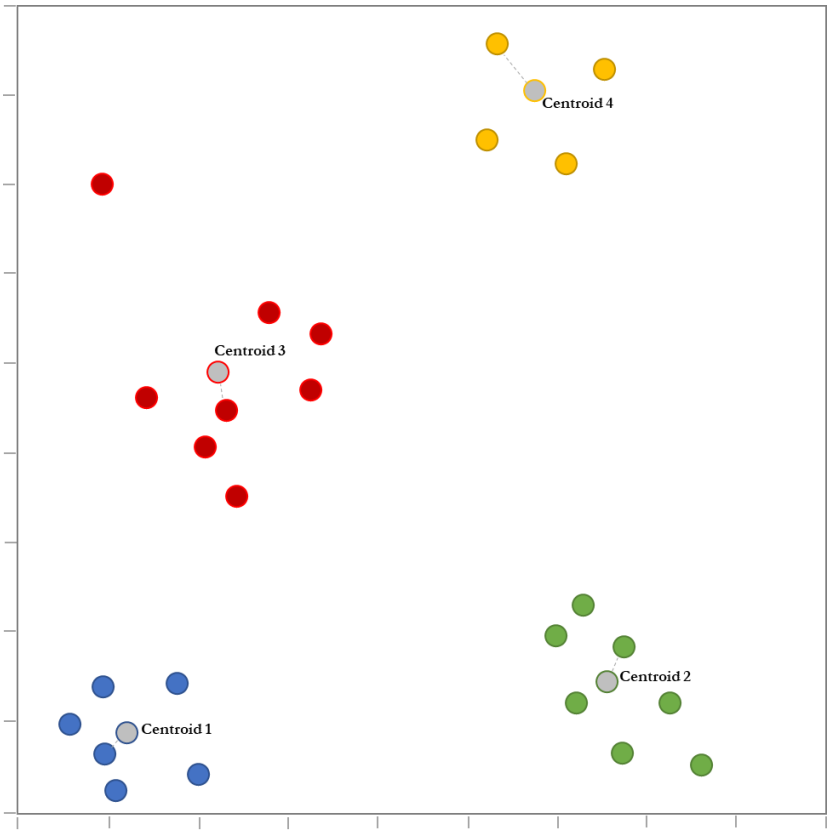
Μέσω του παραπάνω μηχανισμού επίλυσης συναναφορών, το κείμενο καθίσταται πιο αξιοποιήσιμο για εργασίες ανοιχτής εξαγωγής πληροφοριών, καθώς χαρακτηρίζεται από πλουσιότερη παρουσία ονοματικών οντοτήτων, οι οποίες μπορούν να συσχετιστούν με τα δεδομένα μιας γνωσιακής βάσης.

3.3.1.1.3 Εξαγωγική συνόψιση

Ενώ τα προαναφερθέντα στάδια προ-επεξεργασίας εστιάζουν στο περιεχόμενο του κειμένου που θα τροφοδοτηθεί στο μηχανισμό εξαγωγής πληροφοριών, το συγκεκριμένο στάδιο στοχεύει στη μείωση της έκτασής του μέσω της επισήμανσης των σημαντικότερων τμημάτων του. Η επιλογή ενσωμάτωσης προαιρετικού μηχανισμού εξαγωγικής συνόψισης στην προτεινόμενη μεθοδολογία προέκυψε από την παρατήρηση ότι οι περιφερειακές πληροφορίες που συναντώνται σε ελεύθερο κείμενο συχνά οδηγούν σε εξαγωγές περιορισμένης προστιθέμενης αξίας, όπως πλεονάζουσες ή άσχετες με το αντικείμενο τριπλέτες. Στόχος του μηχανισμού σύνοψης είναι να περιορίσει τον συνολικό αριθμό τριπλετών, παράγοντας μια συντομότερη έκδοση του τροφοδοτούμενου κειμένου, η οποία θα διατηρεί παράλληλα τις κομβικές πληροφορίες και το βασικό νοηματικό πλαίσιο του πρωτότυπου.

Ο χρησιμοποιούμενος μηχανισμός εξαγωγικής συνόψισης βασίζεται στη μεθοδολογία του Miller (2019), η οποία στηρίζεται στην παραγωγή διανυσματικών αναπαραστάσεων για κάθε

πρόταση-συνιστώσα ενός κειμένου με χρήση προεκπαιδευμένου κωδικοποιητή (όπως περιεγράφηκε στην υποενότητα 2.6.3). Δεδομένου ότι οι διανυσματικές αναπαραστάσεις μπορούν να θεωρηθούν ως σημεία στον πολυδιάστατο χώρο, ένας αλγόριθμος συσταδοποίησης (k-means) χρησιμοποιείται για να ομαδοποιήσει τα σημεία αυτά σε μη επικαλυπτόμενες ομάδες (συστάδες) και τελικά να επιλέξει ένα ή περισσότερα, ανάλογα με το πόσο κοντά βρίσκονται στο κέντρο βάρους (centroid) της εκάστοτε συστάδας. Κατ' αυτό τον τρόπο, αφενός το κείμενο διαχωρίζεται σε λανθάνουσες θεματικές ενότητες βάσει των συστάδων που δημιουργούνται και αφετέρου επιλέγεται το πιο αντιπροσωπευτικό υποσύνολο κάθε συστάδας βάσει της απόστασής του από το κέντρο βάρους της (Σχήμα 19).



Σχήμα 19 Συσταδοποίηση διανυσματικών αναπαραστάσεων προτάσεων που ανήκουν σε ένα κείμενο, στα πλαίσια εργασίας εξαγωγικής συνόψισης

Σημειώνεται ότι η επιλογή του κατάλληλου κωδικοποιητή για την παραγωγή διανυσματικών αναπαραστάσεων εξαρτάται από το θεματικό πλαίσιο του προς ανάλυση κειμένου. Συγκεκριμένα, για την συνόψιση ειδησεογραφικών άρθρων μπορεί να επιλεγεί μοντέλο που έχει εκπαιδευτεί σε αντίστοιχα σύνολα δεδομένων γενικού περιεχομένου (πχ. sentence-transformers (Reimers and Gurevych, 2020b)), ενώ για την εφαρμογή αντίστοιχου μηχανισμού σε επιστημονικά άρθρα θα προτιμηθεί μοντέλο που έχει εκπαιδευτεί σε ανάλογο περιεχόμενο (πχ. SciBERT (Beltagy et al., 2020a)). Ακόμη, η παραμετροποίηση

της διαδικασίας συσταδοποίησης απαιτεί τον ορισμό τιμής για την αναλογία που θα έχει η τελική περίληψη σε σχέση με το αρχικό κείμενο, καθώς και των ακραίων τιμών για τον ελάχιστο και μέγιστο αριθμό χαρακτήρων που καθιστούν αποδεκτή μια πρόταση (οι υπόλοιπες απορρίπτονται από τη διαδικασία).

Η εξαγωγική συνόψιση αποτελεί το τελευταίο βήμα προ-επεξεργασίας της προτεινόμενης μεθοδολογίας. Το αποτέλεσμα είναι μια συνοπτική απόδοση του αρχικού (ελληνικού) κειμένου, μεταφρασμένη στα αγγλικά και με επιλυμένα τα φαινόμενα συναναφοράς. Στην επόμενη υποενότητα περιγράφεται η διαδικασία εξαγωγής δομημένης πληροφορίας από το προ-επεξεργασμένο ελεύθερο κείμενο μέσω της χρήσης παράλληλων μηχανισμών εξαγωγής πληροφοριών.

3.3.1.2 Ανοιχτή εξαγωγή πληροφοριών

Το στάδιο ανοιχτής εξαγωγής πληροφοριών αξιοποιεί μηχανισμούς OIE βασισμένους τόσο σε μεθόδους μηχανικής μάθησης όσο και σε γλωσσολογικούς κανόνες, στοχεύοντας στην επίτευξη της καλύτερης δυνατής αντιστάθμισης ακρίβειας και ανάκλησης μέσω του βέλτιστου συνδυασμού τους. Επισημαίνεται ότι το συγκεκριμένο στάδιο πραγματοποιείται στο αγγλικό ισοδύναμο του αρχικού κειμένου (για λόγους που περιεγράφηκαν στην υποενότητα 3.3) και επομένως η χρησιμότητα της προτεινόμενης μεθοδολογίας εκτείνεται πέραν του πλαισίου της συγκεκριμένης εργασίας. Οι σχετικοί μηχανισμοί περιγράφονται παρακάτω, με έμφαση σε αυτούς που βασίζονται στη μηχανική μάθηση, καθώς αποτέλεσαν το βασικό στοιχείο μελέτης βάσει των ερευνητικών αξόνων της διατριβής. Σημειώνεται ότι ο συνδυασμός των δύο διαφορετικών προσεγγίσεων εξαγωγής αποτελεί προϊόν συνεργασίας του συγγραφέα με ερευνητική ομάδα του Πανεπιστημίου Εφαρμοσμένων Επιστημών της Ζυρίχης (ZHAW), οδηγώντας σε σημαντικά καλύτερα αποτελέσματα σε σχέση με τις υπάρχουσες τεχνολογίες αιχμής για τη συγκεκριμένη εργασία, όπως αποδεικνύεται στο Κεφάλαιο 4 μέσω της συγκριτικής τους αξιολόγησης.

3.3.1.2.1 Εξαγωγή πληροφοριών βασισμένη στη μηχανική μάθηση και υπολογιστικές μεθόδους

Ο συγκεκριμένος μηχανισμός εξαγωγής ακολουθεί μια προσέγγιση προσανατολισμένη στην ανάκληση (recall), εξασφαλίζοντας το μέγιστο δυνατό αριθμό εξαγωγών μέσω του συνδυασμού τριών συμπληρωματικών στοιχείων εξαγωγής. Ως συνέπεια, ένα ποσοστό των εξαχθέντων τριπλετών αποτελεί ψευδώς θετικά αποτελέσματα, δηλαδή τριπλέτες που είτε συσχετίζουν λάθος οντότητες μεταξύ τους, είτε έχουν εξάγει λάθος κατηγόρημα. Το κάθε στοιχείο του μηχανισμού βασίζεται σε διαφορετική τεχνολογία (περιγράφεται στο Κεφάλαιο 4) και ειδικεύεται σε διαφορετικά είδη προτάσεων. Συγκεκριμένα χρησιμοποιείται:

- ένα στοιχείο εξαγωγής που αποτελεί συνδυασμό 4 εργαλείων, καθένα εκ των οποίων αξιοποιεί απλά γλωσσικά μοντέλα και γλωσσολογικούς περιορισμούς, και ειδικεύεται αντίστοιχα: α) στον διαχωρισμό επαυξημένων προτάσεων σε απλές, β) στην εξαγωγή σχέσεων μεταξύ ουσιαστικών, γ) σε προτάσεις που περιέχουν αριθμητικές πληροφορίες και δ) σε ετικέτες σημασιολογικών ρόλων,
- ένα στοιχείο εξαγωγής που βασίζεται στη μοντελοποίηση των γλωσσικών κανόνων για την αντιστοίχιση συγκεκριμένης στρατηγικής εξαγωγής ανάλογα με το είδος της πρότασης. Το συγκεκριμένο στοιχείο ειδικεύεται στην εξαγωγή πληροφορίας από εμφωλευμένες προτάσεις, καθώς και στο διαχωρισμό σύνθετων προτάσεων σε πολλές επιμέρους τριπλέτες (πχ. η πρόταση "*Ο Κώστας και ο Γιάννης είναι φοιτητές.*", θα οδηγήσει στις τριπλέτες {Κώστας ; είναι ; φοιτητής} και {Γιάννης ; είναι ; φοιτητής},
- ένα στοιχείο εξαγωγής που προσεγγίζει τη συγκεκριμένη εργασία ως μια ανάθεση ετικετών BIO (βλ. υποενότητα 2.2.4) με χρήση νευρωνικού δικτύου LSTM. Το συγκεκριμένο στοιχείο έχει τη δυνατότητα να οδηγήσει στον εντοπισμό πιο σύνθετων συσχετίσεων μεταξύ υποκειμένων, κατηγορημάτων και αντικειμένων, αξιοποιώντας διανυσματικές αναπαραστάσεις λέξεων για την ανάθεση ανεξάρτητων κατανομών πιθανότητας κάθε δυνατής ετικέτας για την κάθε λέξη που απαρτίζει την πρόταση. Στον αντίποδα, ωστόσο, η στοχαστικότητα της διαδικασίας ανάθεσης δεν εγγυάται ότι οι ετικέτες που θα παραχθούν για κάθε πρόταση θα αποτελούνται από ακριβώς τρία ορίσματα της μορφής {υποκείμενο ; κατηγόρημα ; αντικείμενο}, περιπλέκοντας τη διαδικασία εξαγωγής.

3.3.1.2.2 Εξαγωγή πληροφοριών βασισμένη σε γλωσσολογικούς κανόνες

Ο συγκεκριμένος μηχανισμός εξαγωγής έχει σχεδιαστεί με γνώμονα την επίτευξη της μέγιστης δυνατής ακρίβειας (precision), μέσω της αξιοποίησης γλωσσολογικών κανόνων. Ως αποτέλεσμα, οι τριπλέτες που προκύπτουν είναι σχεδόν βέβαιο ότι αποτελούν ορθές εξαγωγές, ωστόσο σε περιπτώσεις περίπλοκων προτάσεων που δεν καλύπτονται από τους προκαθορισμένους κανόνες η εξαγωγή όλων των δυνατών συνδυασμών καθίσταται αδύνατη. Αξίζει να σημειωθεί, ωστόσο, ότι η εφαρμογή αντίστοιχων μηχανισμών σε μορφολογικά απλές γλώσσες (όπως τα αγγλικά) χαρακτηρίζεται από αξιόλογη απόδοση, ιδιαίτερα όσον αφορά κείμενα γενικού περιεχομένου (πχ. ειδησεογραφικά άρθρα).

Βασικό στοιχείο του μηχανισμού αποτελεί η συντακτική ανάλυση κάθε πρότασης και η αποτύπωση της δομής της μέσω της κατασκευής του συντακτικού της δένδρου. Ένας

συντακτικός αναλυτής εξαρτήσεων (dependency parser) χρησιμοποιείται για να αναγνωρίσει σύνολα 3 κόμβων που υπάρχουν σε κάθε δένδρο, με κάθε κόμβο να αντιπροσωπεύει ένα από τα βασικά στοιχεία της τριπλέτας (υποκείμενο, κατηγόρημα, αντικείμενο). Οι κόμβοι αυτοί αποτελούν τους κύριους όρους της πρότασης. Η αναγνώριση των κόμβων γίνεται μέσω ενός συνόλου προκαθορισμένων κανόνων που βασίζονται στην επεξεργασία της γραμματικής δομής της πρότασης και των εξαρτήσεων μεταξύ των λέξεων (πχ. τα επίθετα είναι λέξεις που συνοδεύουν τα ουσιαστικά). Κατόπιν, για κάθε κόμβο εκτελείται αναζήτηση κατά πλάτος (breadth-first), επισημαίνοντας γειτονικούς όρους που ταιριάζουν σε συγκεκριμένους κανόνες ανάλογα με το είδος του κόμβου. Για παράδειγμα, αν ο κόμβος αφορά υποκείμενο η αντικείμενο τριπλέτας, ερευνάται η ύπαρξη επιθέτων καθώς και προσδιοριστικών συνθέσεων (compounds), οι οποίες συναντώνται ιδιαίτερα συχνά στα αγγλικά και αποτελούν σχέσεις ανάμεσα σε ένα ουσιαστικό που προσδιορίζει ένα άλλο κυρίως ουσιαστικό (πχ. *phone* book). Αντίστοιχα, αν ο κόμβος αφορά κατηγόρημα, ο μηχανισμός επισημαίνει τυχόν βοηθητικά ρήματα (πχ. *may* cause) και τα περιλαμβάνει ως επιμέρους κλάδους του συγκεκριμένου κόμβου. Συνολικά χρησιμοποιούνται 40 κανόνες επισήμανσης αντίστοιχων φαινομένων, σε μια προσπάθεια να καλυφθούν όσο δυνατόν περισσότερα γλωσσικά φαινόμενα, αφήνοντας τον χειρισμό ασαφών ή διφορούμενων εξαρτήσεων στην ευχέρεια του μηχανισμού που βασίζεται σε μεθόδους μηχανικής μάθησης και υπολογιστικές μεθόδους.

3.3.1.2.3 Συνδυασμός αποτελεσμάτων εξαγωγής

Μετά την παράλληλη εξαγωγή τριπλετών από τους δύο προαναφερθέντες μηχανισμούς, αυτές συνδυάζονται σε ένα ενιαίο σύνολο με σκοπό τη διατήρηση της υψηλής ακρίβειας του δεύτερου, αυξάνοντας παράλληλα την συνολική ανάκληση του συστήματος με την προσθήκη τριπλετών που προέρχονται από τον πρώτο. Κατά τη διαδικασία συνδυασμού, οι τριπλέτες που προέρχονται από την εφαρμογή μεθόδων μηχανικής μάθησης περνούν από ένα επιπλέον στάδιο βελτίωσης, όπου αντιστοιχίζονται με το συντακτικό δέντρο της εκάστοτε πρότασης. Για κάθε τριπλέτα, επισημαίνονται οι κόμβοι του δένδρου που αποτελούν μέρος του υποκειμένου, κατηγορήματος ή αντικειμένου της τριπλέτας. Σε περίπτωση που οι εξαρτήσεις μεταξύ των επισημασμένων κόμβων δεν είναι έγκυρες (δηλαδή δεν συμφωνούν με το προκαθορισμένο σύνολο κανόνων), τότε οι λεκτικές μονάδες που χαρακτηρίζονται από την πλεονάζουσα αυτή εξάρτηση απορρίπτονται από τον κλάδο του δένδρου. Για παράδειγμα, έστω η παρακάτω πρόταση:

Ο Ερατοσθένης ήταν αρχαίος Έλληνας μαθηματικός που υπολόγισε το μέγεθος της γης.

Μια εκ των τριπλετών που εξήχθησαν από το μηχανισμό της υποενότητας 3.3.1.2.1 είναι:

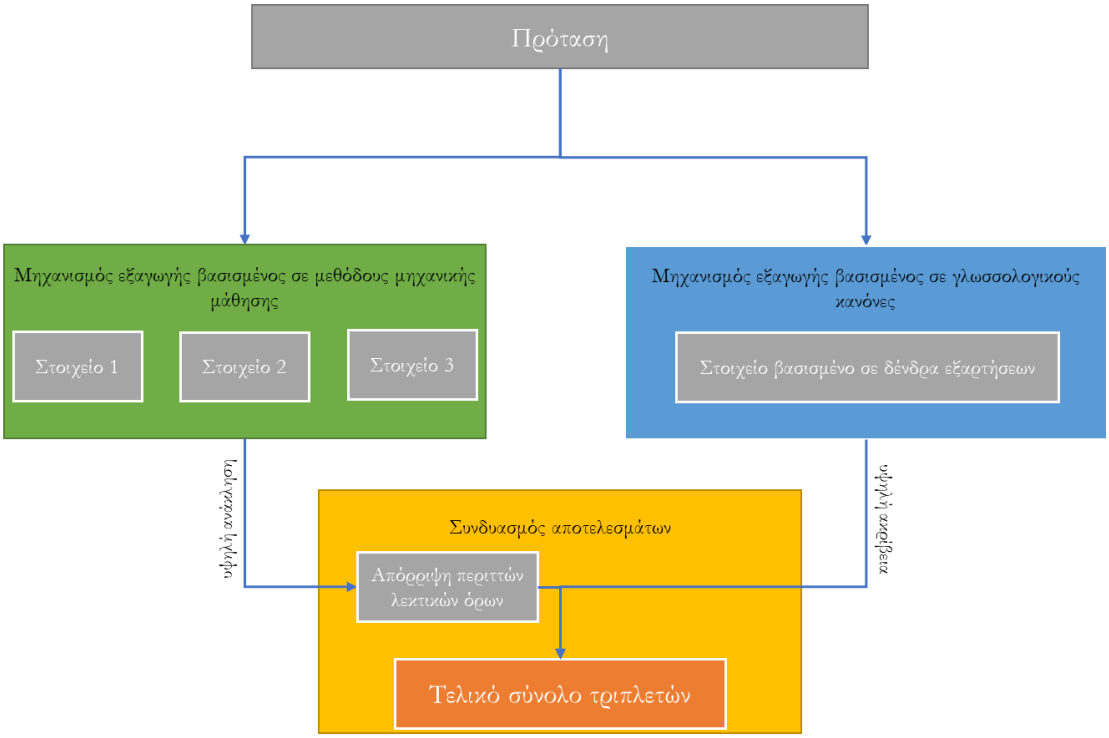
{ Ερατοσθένης ; ήταν ; αρχαίος Έλληνας μαθηματικός που υπολόγισε το μέγεθος της γης}

Η παραπάνω τριπλέτα περιλαμβάνει ένα αρκετά μακροσκελές σύνολο όρων ως αντικείμενο, το οποίο ουσιαστικά ακυρώνει τη χρησιμότητα των μηχανισμών εξαγωγής όσον αφορά την κατασκευή δομημένης πληροφορίας. Βάσει των προκαθορισμένων κανόνων (χρησιμοποιούνται οι ίδιοι με τον μηχανισμό εξαγωγής βασισμένο σε γλωσσολογικούς κανόνες), το αντικείμενο "αρχαίος Έλληνας μαθηματικός που υπολόγισε το μέγεθος της γης" δεν είναι έγκυρο καθώς περιλαμβάνει εκτός των επιθετικών προσδιορισμών "αρχαίος Έλληνας" και την αναφορική φράση "που υπολόγισε το μέγεθος της γης". Έτσι, η φράση αυτή αφαιρείται από την αρχική τριπλέτα, με το τελικό αποτέλεσμα εξαγωγής να είναι το παρακάτω:

{ Ερατοσθένης ; ήταν ; αρχαίος Έλληνας μαθηματικός}

Η βασική διαφορά του παρόντος σταδίου συνδυασμού με τον μηχανισμό εξαγωγής βασισμένο σε κανόνες είναι ότι ο δεύτερος θα απέρριπτε ολόκληρη την τριπλέτα, καθώς ένας από τους τρεις βασικούς της κόμβους περιείχε μη έγκυρα στοιχεία. Χάρη στο στάδιο συνδυασμού όμως, περιλαμβάνονται στα τελικά αποτελέσματα περισσότερες εξαγωγές τις οποίες οι αιτιοκρατικοί μηχανισμοί εξαγωγής αποτυγχάνουν να ανιχνεύσουν ενώ η στοχαστικότητα των μεθόδων μηχανικής μάθησης αδυνατεί να αποτυπώσει με συντακτικά ορθό τρόπο. Σημειώνεται ακόμη ότι το στάδιο συνδυασμού αφορά κυρίως τη βελτίωση αντίστοιχων τριπλετών που δεν προσφέρουν επιπλέον πληροφορία στον τελικό χρήστη και δεν ελαττώνουν την ανάκληση του συνολικού συστήματος. Πιο συγκεκριμένα, οι επιπλέον τριπλέτες που προκύπτουν από το "σπάσιμο" της πρότασης σε επιμέρους απλές προτάσεις (πχ. { Ερατοσθένης ; υπολόγισε ; το μέγεθος της γης}) θα προστεθούν κανονικά στο τελικό σύνολο ως αποτέλεσμα της εφαρμογής μεθόδων μηχανικής μάθησης.

Στο Σχήμα 20 απεικονίζεται η ροή επεξεργασίας μιας πρότασης κατά την εισαγωγή της στο στάδιο ανοιχτής εξαγωγής πληροφοριών με παράλληλη ανάλυση από τους προαναφερθέντες μηχανισμούς, ακολουθούμενη από το στάδιο συνδυασμού των επιμέρους τριπλετών.



Σχήμα 20 Επισκόπηση σταδίου ανοιχτής εξαγωγής πληροφοριών

3.3.1.3 Μετα-επεξεργασία αποτελεσμάτων εξαγωγής

Οι τριπλέτες που προκύπτουν από το παραπάνω στάδιο ανοιχτής εξαγωγής πληροφοριών αφορούν την ανάλυση του αγγλικού ισοδύναμου της εκάστοτε πρότασης και επομένως πρέπει να μετασχηματιστούν στην ελληνική γλώσσα. Το στάδιο μετα-επεξεργασίας περιλαμβάνει δύο εναλλακτικές μεθόδους για την τελική εξαγωγή δομημένης πληροφορίας στα ελληνικά: α) την εφαρμογή μηχανισμού (αντίστροφης) μηχανικής μετάφρασης στις εξαχθείσες τριπλέτες για την μετάφρασή τους στην ελληνική, ή β) την χρήση μηχανισμού ευθυγράμμισης λέξεων (word alignment) για την αντιστοίχιση και αντικατάσταση των αγγλικών λεκτικών μονάδων που περιλαμβάνονται στις τριπλέτες με τις ισοδύναμες ελληνικές που συναντώνται στην αρχική πρόταση. Κάθε προσέγγιση έχει τα δικά της πλεονεκτήματα και μειονεκτήματα, τα οποία αναλύονται στις σχετικές υποενότητες.

3.3.1.3.1 Αντίστροφη μηχανική μετάφραση

Η χρήση μηχανισμού αντίστροφης μετάφρασης για τον μετασχηματισμό των τριπλετών από αγγλικά σε ελληνικά αποτελεί μια λογική επιλογή δεδομένου ότι η ύπαρξη παράλληλου κειμένου επιτρέπει την εκπαίδευση αντίστοιχου νευρωνικού μοντέλου, όπως αυτό που περιεγράφηκε στην υποενότητα 3.3.1.1.1. Στην προκειμένη περίπτωση, το εκπαιδευμένο μοντέλο λαμβάνει ως είσοδο το κάθε στοιχείο (υποκείμενο, κατηγόρημα, αντικείμενο) της τριπλέτας ξεχωριστά και το μεταφράζει στο ελληνικό του ισοδύναμο. Σημειώνεται ότι τα

παραπάνω στοιχεία αποτελούν συνήθως σύντομες φράσεις καθώς περιλαμβάνουν περισσότερες από μια λεκτικές μονάδες, όπως φάνηκε στα παραδείγματα της υποενότητας 3.3.1.2.

Στα πλεονεκτήματα του συγκεκριμένου μηχανισμού συγκαταλέγεται η απλότητα υλοποίησής του και το γεγονός ότι τα τελικά αποτελέσματα για κάθε στοιχείο είναι ανεξάρτητα μεταξύ τους (πχ. λανθασμένη μετάφραση ενός κατηγορήματος δε συνεπάγεται ότι και τα υπόλοιπα στοιχεία της τριπλέτας θα αποδοθούν με λάθος τρόπο στην ελληνική). Ωστόσο, η αποσπασματική τροφοδότηση του μηχανισμού μετάφρασης με στοιχεία της εκάστοτε τριπλέτας ενδέχεται να οδηγήσει σε λανθασμένη απόδοση του κειμένου σε περιπτώσεις που χαρακτηρίζονται από αμφισημία ή ασάφεια (πχ. το κατηγόρημα "lead" μπορεί να αποδοθεί είτε ορθά ως "οδηγώ", είτε λανθασμένα ως "μόλυβδος"). Αυτό οφείλεται στο γενικότερο τρόπο λειτουργίας των ακολουθιακών μοντέλων, των οποίων οι διανυσματικές αναπαραστάσεις για κάθε λέξη εξαρτώνται από το συνολικό περιεχόμενο μιας φράσης. Επιπλέον, η μετάβαση από μια μορφολογικά φτωχότερη σε μια μορφολογικά πλουσιότερη γλώσσα έχει ως αποτέλεσμα τον κίνδυνο απώλειας πληροφορίας που αφορά το γένος ή/και τα κλιτά μέρη του λόγου (πχ. η τριπλέτα {Martha ; is ; smart} μπορεί να αποδοθεί λανθασμένα ως {Μάρθα ; είναι ; έξυπνος}).

3.3.1.3.2 Ευθυγράμμιση λέξεων

Η εναλλακτική μέθοδος που προτείνεται για τον μετασχηματισμό των αγγλικών τριπλετών στην ελληνική γλώσσα βασίζεται σε προεκπαιδευμένο, μη επιβλεπόμενο μοντέλο ευθυγράμμισης λέξεων. Το μοντέλο δέχεται ως είσοδο την αρχική (ελληνική) και τη μεταφρασμένη (αγγλική) πρόταση από το στάδιο προ-επεξεργασίας και επιστρέφει ζεύγη $i-j$ (σε Pharaoh format), καθένα εκ των οποίων υποδηλώνει ότι το i-οστό λεκτικό στοιχείο της αρχικής πρότασης αντιστοιχεί στο j-οστό στοιχείο της μεταφρασμένης, όπως φαίνεται στο παρακάτω παράδειγμα:

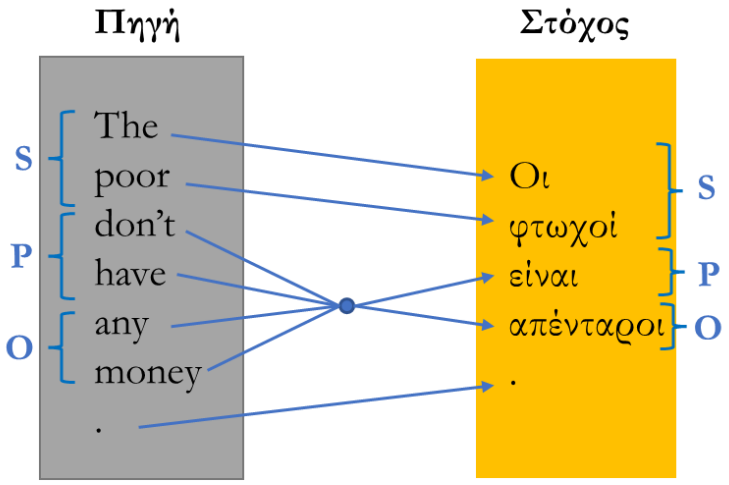
Πηγή: Ο Δημήτρης αγόρασε νέο αυτοκίνητο

Στόχος: Dimitris bought a new car

Ζεύγη ευθυγράμμισης: 0-0, 1-0, 2-1, 3-3, 4-4

Το βασικό πλεονέκτημα της παραπάνω μεθόδου είναι ότι δεν επηρεάζεται από τη νοηματική ρευστότητα που προκαλεί η ύπαρξη πολλαπλών ερμηνειών για μια λέξη ή φράση (σε αντίθεση με το μηχανισμό αντίστροφης μετάφρασης), καθώς το αρχικό κείμενο τροφοδοτείται παράλληλα με το μεταφρασμένο και ο μηχανισμός αναλαμβάνει απλά την

αντιστοίχιση των λεκτικών τους στοιχείων. Ωστόσο, η διαδικασία ενδέχεται να οδηγήσει σε ασύμμετρες ευθυγραμμίσεις (αντιμετωπίζοντας είτε την αρχική γλώσσα είτε την γλώσσα μετάφρασης ως κύρια), με αποτέλεσμα να μην υπάρχει πάντα αμφιμονοσήμαντη (1-1) αντιστοιχία μεταξύ των λεκτικών στοιχείων κάθε γλώσσας. Ως εκ τούτου, υπάρχει πιθανότητα ορισμένα λεκτικά στοιχεία της αρχικής γλώσσας να μην αντιστοιχιστούν σε κανένα στοιχείο του μεταφρασμένου κειμένου, οδηγώντας σε ελλιπείς τριπλέτες (χωρίς υποκείμενο, ρήμα ή κατηγόρημα). Αυτό είναι ιδιαίτερα συχνό σε περιπτώσεις που η μετάφραση αποτελεί σχετικά ελεύθερη απόδοση του αρχικού κειμένου, όπως στο Σχήμα 21. Τέλος, σημειώνεται πως η ορθότητα της ευθυγράμμισης εξαρτάται από τις παραμέτρους εκπαίδευσης του μοντέλου (πχ. αρχιτεκτονική, σύνολο παράλληλου κειμένου που χρησιμοποιήθηκε) και επομένως εμπίπτει σε όλους τους περιορισμούς που χαρακτηρίζουν τις γλώσσες χαμηλότερων πόρων, όπως η ελληνική.



Σχήμα 21 Παράδειγμα ευθυγράμμισης λέξεων μεταξύ αρχικού και μεταφρασμένου κειμένου και συσχέτιση των αντίστοιχων SPO τριπλετών τους

3.4 Εξαγωγή σημασιολογικών συμπερασμάτων από διαδικτυακές πηγές

Στην προηγούμενη ενότητα περιεγράφηκε η προτεινόμενη μεθοδολογία εξόρυξης δεδομένων από ελεύθερο κείμενο για την αποτύπωση της αδόμητης πληροφορίας σε δομημένη μορφή. Κατά τη διάρκεια εκπόνησης της διατριβής διαπιστώθηκε ότι, ενώ η αποθήκευση της πληροφορίας σε μορφή τριπλετών είναι ιδανική για εργασίες εξερεύνησης περιεχομένου, σημασιολογικής αναζήτησης και συσχέτισης οντοτήτων (πχ. με την βοήθεια ενός γνωσιακού γράφου), ο τεμαχισμός του περιεχομένου αυτού σε επίπεδο λέξεων ή σύντομων φράσεων (δηλαδή στις συνιστώσες μια τριπλέτας) δεν εξυπηρετεί πιο πολύπλοκες διαδικασίες συμπερασμού όπως ο έλεγχος σημασιολογικής ομοιότητας δύο πηγών και η

επικύρωση ισχυρισμών ενός χρήστη βάσει της αντληθείσας πληροφορίας. Οι παραπάνω εργασίες προϋποθέτουν επεξεργασία τουλάχιστον σε επίπεδο πρότασης, καθώς το νόημα ενός κειμένου εξαρτάται τόσο από την επιλογή των επιμέρους λεκτικών όρων όσο και από τη διάταξή τους στο κείμενο.

Για τον λόγο αυτό, ο δεύτερος ερευνητικός άξονας της παρούσας διατριβής αφορά την ανάπτυξη ενός συστήματος επαλήθευσης ισχυρισμών και εξαγωγής συμπερασμάτων βάσει της συλλογής, επεξεργασίας και συνδυασμού πληροφοριών από ειδησεογραφικές διαδικτυακές πηγές (πχ. ιστοτόπους ειδήσεων, ηλεκτρονικές εκδόσεις εφημερίδων) σε επίπεδο πρότασης. Η προτεινόμενη μεθοδολογία ενσωματώνει ένα σύνολο μηχανισμών για την συνεχή ανάκτηση περιεχομένου με τη βοήθεια ιχνηλατών (crawlers), την αποθήκευση των δεδομένων σε βάση δεδομένων γράφων, καθώς και για την επισήμανση του περιεχομένου τους μέσω διεργασιών σύνδεσης οντοτήτων, με απώτερο στόχο την οργάνωση της συσσωρευμένης γνώσης για εργασίες σημασιολογικού συμπερασμού. Στις ακόλουθες υποενότητες ακολουθεί η γενική περιγραφή της προτεινόμενης προσέγγισης καθώς και των επιμέρους μηχανισμών και μεθοδολογιών που αναπτύχθηκαν για την υποστήριξή της.

3.4.1 Προτεινόμενη προσέγγιση

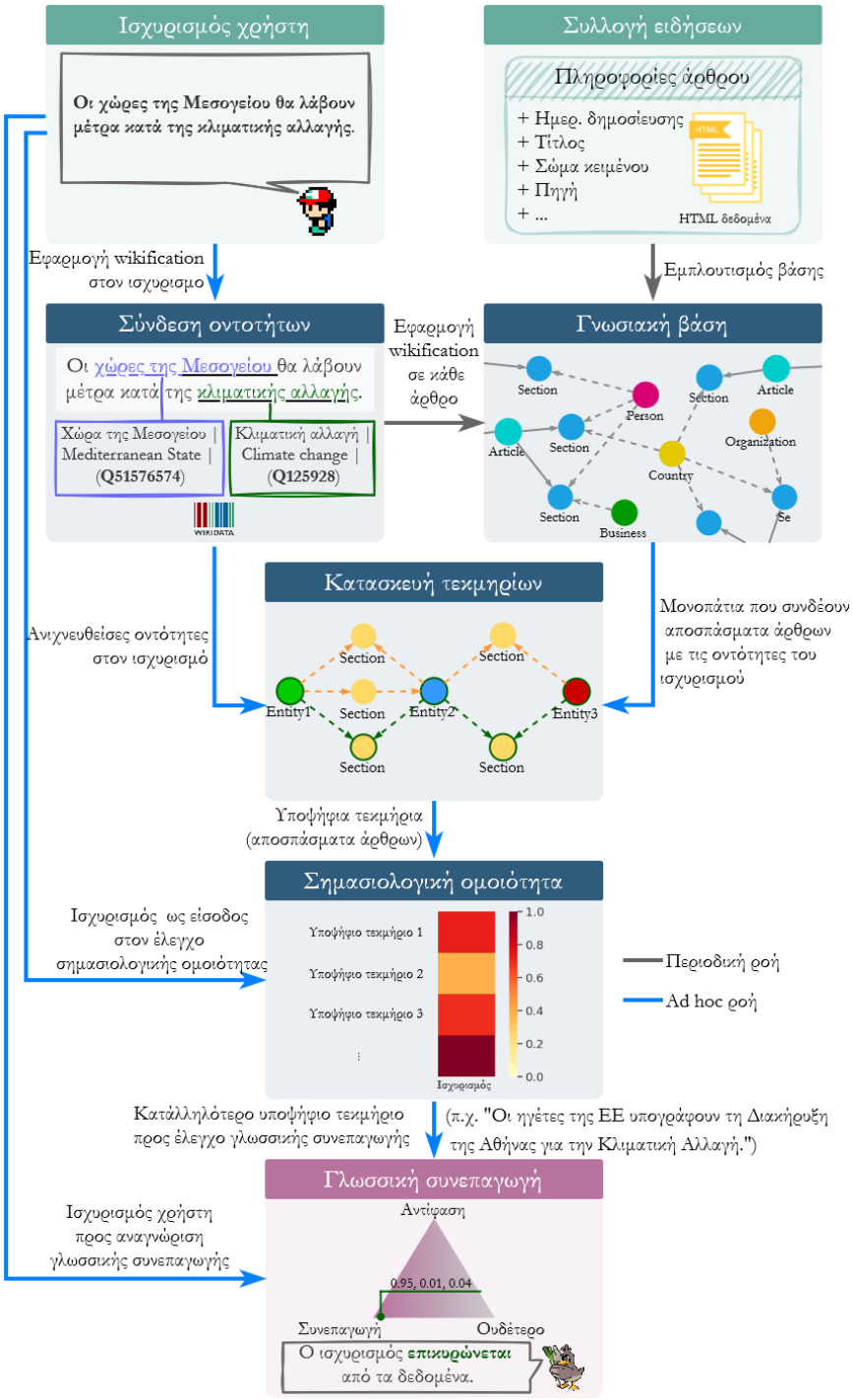
Η προτεινόμενη προσέγγιση αποτελείται από ένα σύνολο μηχανισμών σε διάταξη pipeline για τον έλεγχο ενός ισχυρισμού από τον χρήστη βάσει της συγκεντρωμένης πληροφορίας που προκύπτει από την αποδελτίωση πολλαπλών ειδησεογραφικών πηγών. Σημειώνεται ότι ενώ η συγκεκριμένη προσέγγιση έχει αναπτυχθεί για την ελληνική γλώσσα, η αρθρωτή της σχεδίαση επιτρέπει την τροποποίηση ή αντικατάσταση των επιμέρους μηχανισμών της ώστε να χρησιμοποιηθεί για οποιαδήποτε άλλη γλώσσα. Το σύστημα περιλαμβάνει δύο επιμέρους ροές, τη *ροή εμπλουτισμού βάσης* με νέα πληροφορία και τη *ροή ελέγχου ισχυρισμών* η οποία εκτελείται κάθε φορά που ο χρήστης τροφοδοτεί το σύστημα με έναν ισχυρισμό (πρόταση ή φράση) σε φυσική γλώσσα προκειμένου αυτός να επιβεβαιωθεί ή να απορριφθεί βάσει των δεδομένων. Στη συνέχεια, περιγράφονται οι προαναφερθείσες ροές του συστήματος και ακολουθούν οι σχεδιαστικές λεπτομέρειες των μηχανισμών που χρησιμοποιούνται για την εκτέλεσή τους.

- Ροή εμπλουτισμού γνωσιακής βάσης: Αφορά διεργασία που εκτελείται περιοδικά και συνίσταται στην συλλογή και κατάλληλη επεξεργασία δεδομένων από ειδησεογραφικές πηγές του Διαδικτύου, ώστε να χρησιμοποιηθούν αργότερα από τη ροή ελέγχου ισχυρισμών. Αρχικά, γίνεται συσσώρευση του κειμενικού περιεχομένου από προεπιλεγμένους ειδησεογραφικούς ιστοτόπους σε μια βάση δεδομένων

γράφων, μέσω του μηχανισμού συλλογής και επεξεργασίας δεδομένων. Στη συνέχεια το κάθε στοιχείο της βάσης επισημαίνεται με επιπλέον πληροφορίες που αφορούν τις αναφερόμενες σε αυτό οντότητες, μέσω του μηχανισμού σύνδεσης οντοτήτων. Το αποτέλεσμα από την εκτέλεση της ροής εμπλουτισμού είναι η απεικόνιση ενός μεγάλου σώματος κειμένου σε μορφή γράφου, χωρισμένου σε επίπεδο πρότασης και επισημασμένου με τις αναγνωρισμένες σε αυτό ονοματικές οντότητες. Σημειώνεται ότι, λόγω του ιδιαίτερα δυναμικού χαρακτήρα της πηγής (καταγραφή ειδησεογραφίας σε σχεδόν πραγματικό χρόνο), είναι αναγκαία η συνεχής ανανέωση της βάσης με νέες πληροφορίες, γι' αυτό και η ροή εκτελείται περιοδικά.

- Ροή ελέγχου ισχυρισμών: Αποτελεί την κεντρική διεργασία του συστήματος, η οποία εκτελείται κατόπιν αιτήματος του χρήστη μέσω ενός μηχανισμού συλλογής και επεξεργασίας δεδομένων που αναλαμβάνει την εισαγωγή ενός ισχυρισμού (πρότασης ή φράσης) σε φυσική γλώσσα, είτε με τη μορφή ελεύθερου κειμένου είτε με τη μορφή ομιλίας, η οποία θα μετατραπεί σε κείμενο μέσω ενός στοιχείου αναγνώρισης φωνής που ενσωματώνεται στον μηχανισμό συλλογής και επεξεργασίας δεδομένων. Στη συνέχεια, ο μηχανισμός σύνδεσης οντοτήτων αναλύει τον ισχυρισμό του χρήστη για τον εντοπισμό ονοματικών οντοτήτων (πχ. ονόματα προσώπων, χωρών, εταιρειών, οργανισμών κτλ.). Οι οντότητες αυτές δίνονται ως είσοδος στον μηχανισμό κατασκευής τεκμηρίων, ο οποίος αναλαμβάνει να ανακτήσει και να επιστρέψει πληροφορίες από τη βάση δεδομένων οι οποίες *πιθανώς* να σχετίζονται με τον ισχυρισμό του χρήστη. Βασικό κριτήριο για την κατασκευή ενός τεκμηρίου είναι οι συνιστώσες σε αυτό προτάσεις να συνδέονται με τις οντότητες που ανιχνεύθηκαν στον ισχυρισμό του χρήστη. Ακολουθεί η αξιολόγηση της σημασιολογικής ομοιότητας μεταξύ κάθε πιθανού τεκμηρίου και του ισχυρισμού μέσω σχετικού μηχανισμού, ώστε να αναδειχθεί το πιο σχετικό τεκμήριο που θα προωθηθεί στην τελική φάση ελέγχου. Το ζεύγος ισχυρισμού-τεκμηρίου που θα προκύψει αναλύεται από έναν μηχανισμό αναγνώρισης κειμενικής συνεπαγωγής, από τον οποίο προκύπτει εάν το τεκμήριο επικυρώνει, απορρίπτει ή είναι τελικά ουδέτερο ως προς τον αρχικό ισχυρισμό.

Η αρχιτεκτονική της περιγραφείσας μεθοδολογίας απεικονίζεται στο Σχήμα 22 με διαφορετικό χρώμα για καθεμία από τις δύο ροές. Περισσότερες πληροφορίες για τον σχεδιασμό και τη λειτουργία των επιμέρους μηχανισμών παρατίθενται στις ακόλουθες υποενότητες.



Σχήμα 22 Επισκόπηση μεθοδολογίας εξαγωγής σημασιολογικών συμπερασμάτων από διαδικτυακές πηγές

3.4.1.1 Συλλογή και επεξεργασία δεδομένων

Ο μηχανισμός συλλογής και επεξεργασίας δεδομένων αποτελεί την κύρια διεπαφή του συστήματος με το χρήστη και τις εξωτερικές πηγές δεδομένων. Περιλαμβάνει τα στοιχεία εισαγωγής ισχυρισμών και τα στοιχεία εμπλουτισμού της βάσης δεδομένων γράφων με δεδομένα ειδησεογραφικών πηγών.

Στοιχεία εισαγωγής ισχυρισμών: Ο χρήστης μπορεί να τροφοδοτήσει το σύστημα με τον προς επικύρωση ισχυρισμό με δύο διαφορετικούς τρόπους: είτε μέσω της πληκτρολόγησης μιας πρότασης/φράσης σε μορφή ελεύθερου κειμένου, είτε μέσω της υπαγόρευσης μιας αντίστοιχης φράσης, η οποία θα μετατραπεί σε κείμενο μέσω ενός στοιχείου αναγνώρισης ομιλιας.

3.4.1.1.1 Εισαγωγή ισχυρισμού μέσω πληκτρολόγησης ελεύθερου κειμένου

Σε αυτή την περίπτωση ο χρήστης απλά πληκτρολογεί τον προς επικύρωση ισχυρισμό σε ένα πλαίσιο κειμένου. Δεν υπάρχει περιορισμός ως προς το μέγεθος του κειμένου, σημειώνεται ωστόσο ότι ο ισχυρισμός θα πρέπει να περιλαμβάνει αναφορά σε μία ή περισσότερες ονοματικές οντότητες, προκειμένου να συσχετιστεί με τα κατάλληλα αποσπάσματα της γνωσιακής βάσης στο επόμενο στάδιο.

3.4.1.1.2 Εισαγωγή ισχυρισμού μέσω ομιλίας

Ο χρήστης έχει τη δυνατότητα να υπαγορεύσει έναν ισχυρισμό σε φυσική γλώσσα, αξιοποιώντας ένα στοιχείο αυτόματης αναγνώρισης ομιλίας, το οποίο εκπαιδεύτηκε με τη χρήση νευρωνικού μοντέλου αρχιτεκτονικής XLSR-Wav2Vec2 (Conneau et al., 2021). Ακολουθήθηκε η διαδικασία που περιεγράφηκε στην υποενότητα 2.9.3 για την προσαρμογή (finetuning) του μοντέλου στην ελληνική γλώσσα με χρήση ελεύθερα διαθέσιμων απομαγνητοφωνημένων φωνητικών δεδομένων. Αναλυτικές λεπτομέρειες σχετικά με τις παραμέτρους του μοντέλου και του συνόλου δεδομένων που χρησιμοποιήθηκε παρατίθενται στο Κεφάλαιο 4. Επισημαίνεται ότι, εξαιτίας της ιδιαίτερα χαμηλής διαθεσιμότητας φωνητικών δεδομένων (μικρός αριθμός ομιλητών, περιορισμένο πεδίο εφαρμογής κτλ.) για την ελληνική γλώσσα, το μοντέλο εμφανίζει διακυμάνσεις στην απόδοση και αποτελεί δευτερεύουσα μέθοδο διεπαφής σε σχέση με την απευθείας εισαγωγή ισχυρισμών μέσω πληκτρολόγησης.

Όσον αφορά τα στοιχεία εμπλουτισμού της γνωσιακής βάσης, μελετήθηκαν δύο διαφορετικές μέθοδοι εισαγωγής δεδομένων. Η πρώτη βασίζεται στην χρήση στοιχείου web crawling σε δεδομένα RSS (Rich Site Summary) από ειδησεογραφικές πηγές που διατηρούν και ενημερώνουν αντίστοιχες ροές Ιστού. Η δεύτερη μέθοδος (η οποία και τελικά προτιμήθηκε) βασίζεται στη χρήση crawler απευθείας στις κεντρικές HTML σελίδες προεπιλεγμένων ειδησεογραφικών ιστοτόπων, και την αναδρομική παρακολούθηση όλων των εσωτερικών υπερσυνδέσμων ώστε να εξάγει δομημένη πληροφορία από την εκάστοτε ιστοσελίδα. Στις δύο ακόλουθες υποενότητες δίνονται περισσότερες πληροφορίες για την καθεμία από τις προαναφερθείσες μεθόδους.

3.4.1.1.3 Εισαγωγή δεδομένων μέσω ιχνηλάτησης τροφοδοσιών RSS

Οι τροφοδοσίες RSS έχουν σχεδιαστεί για την ανταλλαγή ψηφιακού πληροφοριακού περιεχομένου, παρέχοντας πρόσβαση εφαρμογών στο περιεχόμενο ιστοτόπων με προγραμματιστικό τρόπο. Συναντώνται κυρίως σε ειδησεογραφικούς ιστοτόπους ή ιστολόγια και αποτελούν μια εύχρηστη μέθοδο για την ενημέρωση των χρηστών με το πιο πρόσφατο περιεχόμενο της εκάστοτε πηγής, το οποίο διατίθεται σε τυποποιημένη μορφή μέσω της γλώσσας σήμανσης XML. Ένας χρήστης του Διαδικτύου μπορεί έτσι να ενημερώνεται αυτομάτως για γεγονότα και νέα από όσες ιστοσελίδες υποστηρίζουν RSS, αρκεί να έχει εγγραφεί ο ίδιος συνδρομητής στην αντίστοιχη υπηρεσία της εκάστοτε ιστοσελίδας. Οι εν λόγω ενημερώσεις (ροές/τροφοδοσίες RSS) είτε περιέχουν τα πλήρη δεδομένα είτε σύνοψη αυτών, συνοδευόμενα από σχετικά μεταδεδομένα, ενώ αποστέλλονται αυτομάτως στον συνδρομητή μέσω Διαδικτύου χωρίς να χρειάζεται ο ίδιος να επισκεφτεί τον σχετικό δικτυακό ιστότοπο.

Για τον εμπλουτισμό της γνωσιακής βάσης αναπτύχθηκε κώδικας ανίχνευσης ροών RSS και αυτόματης ενσωμάτωσης του περιεχομένου τους στη γνωσιακή βάση. Ο συγκεκριμένος crawler λαμβάνει ως παράμετρο μια λίστα από διευθύνσεις URL ενεργών τροφοδοσιών RSS από ελληνικά ειδησεογραφικά sites και συλλέγει το πιο πρόσφατο περιεχόμενο του καθενός με αξιοποίηση εξατομικευμένων κλάσεων spider για κάθε ροή, ώστε να εξασφαλιστεί η σωστή ευρετηρίαση και ανάλυση του εκάστοτε κειμένου. Η διαδικασία εμπλουτισμού εκτελείται περιοδικά με τη βοήθεια λογισμικού προγραμματισμού εργασιών (cron job), μέσω του οποίου ελέγχεται η συχνότητα του crawling. Αν και η αξιοποίηση ροών RSS παρουσιάζει αρκετά οφέλη (απλότητα υλοποίησης, χαμηλό υπολογιστικό κόστος), σημειώνεται ότι η σταδιακή απαξίωση του προτύπου RSS από τα σύγχρονα μέσα διαδικτυακής ενημέρωσης προβληματίζει αναφορικά με τη βιωσιμότητα της μεθοδολογίας. Παράλληλα, παρατηρήθηκε ότι στην πλειονότητά τους οι εν λόγω ροές περιείχαν μόνο συνόψεις του περιεχομένου που περιλαμβάνει η αντίστοιχη HTML σελίδα της εκάστοτε είδησης, περιορίζοντας σημαντικά τη διαδικασία εμπλουτισμού. Τα παραπάνω, οδήγησαν στην ανάπτυξη εναλλακτικής μεθόδου που περιγράφεται στην επόμενη υποενότητα.

3.4.1.1.4 Εισαγωγή δεδομένων μέσω ιχνηλάτησης HTML σελίδων

Η χρήση γενικευμένων ιχνηλατών Ιστού (generic Web crawlers) αποτελεί συνήθη πρακτική σύγχρονων μηχανών αναζήτησης και λοιπών διαδικτυακών εφαρμογών, καθώς επιτρέπει την αυτοματοποιημένη συλλογή και επικαιροποίηση πληροφοριών μέσω της επίσκεψης σχετικών ιστοτόπων (Papadakis et al., 2005). Ένας web crawler αποτελεί είδος πράκτορα λογισμικού

(bot) που επισκέπτεται μια λίστα κύριων διευθύνσεων (root URLs) και εντοπίζει όλους τους υπερσυνδέσμους που περιέχουν ώστε να τους επισκεφτεί στη συνέχεια βάσει συγκεκριμένης πολιτικής, διασχίζοντας έτσι όλο το χάρτη ενός ιστοτόπου (sitemap) και πραγματοποιώντας λήψη του περιεχομένου του. Στην περίπτωση ειδησεογραφικών διαδικτυακών πηγών, ο ιχνηλάτης είναι σε θέση να εξάγει το περιεχόμενο διαφορετικών HTML σελίδων χωρίς την ανάγκη προσαρμογής, αξιοποιώντας ευρετικές μεθόδους που λαμβάνουν υπόψη την πυκνότητα υπερσυνδέσμων και την καταμέτρηση λέξεων προκειμένου να αναγνωρίσει τα τμήματα της ιστοσελίδας που αποτελούν μέρη ενός ειδησεογραφικού άρθρου.

Το συγκεκριμένο στοιχείο ιχνηλάτησης ιστοσελίδων βασίζεται στη μεθοδολογία των Hamborg et al., 2017 και στοχεύει στην πλήρη εξαγωγή των πληροφοριών από έναν ειδησεογραφικό ιστότοπο ανεξαρτήτως της δομής του, βασιζόμενο σε μεθοδολογία τριών βημάτων: Αρχικά, δεδομένου ενός root URL (πχ. "newsbomb.gr") ο ιχνηλάτης κατεβάζει το πλήρες περιεχόμενο της HTML σελίδας με χρήση τεχνολογιών συλλογής (web scraping). Στη συνέχεια, είτε αναλύει το sitemap της ιστοσελίδας (εφόσον υπάρχει) για εύρεση όλων των διαθέσιμων υπερσυνδέσμων που αφορούν ειδησεογραφικά άρθρα, είτε ακολουθεί αναδρομική ιχνηλάτηση ακολουθώντας τους εσωτερικούς υπερσυνδέσμους στην εκάστοτε ιστοσελίδα. Τέλος, για κάθε ιστοσελίδα που επισκέπτεται, ο ιχνηλάτης απομονώνει τις βασικές πληροφορίες της είδησης (τίτλος, σώμα κειμένου, συγγραφέας, ημερομηνία κτλ.) με χρήση κανονικών εκφράσεων (regular expressions) εξάγοντας την πληροφορία για κάθε άρθρο σε μορφή JSON. Τα παραγόμενα αρχεία χρησιμοποιούνται για τον εμπλουτισμό της γνωσιακής βάσης. Η συγκεκριμένη μεθοδολογία ιχνηλάτησης προτιμήθηκε σε σχέση με την προηγούμενη (ιχνηλάτηση RSS), αφενός γιατί επιτρέπει την πρόσβαση τόσο σε πρόσφατες όσο και παλαιότερες ειδήσεις σε αντίθεση με τις τροφοδοσίες RSS που αφορούν μόνο πρόσφατο περιεχόμενο, και αφετέρου γιατί εξασφαλίζει την εξαγωγή ολόκληρου του κειμενικού περιεχομένου μιας είδησης και όχι απλά της σύνοψής του, όπως συμβαίνει με τις τροφοδοσίες RSS.

Μέσω του στοιχείου ιχνηλάτησης περιεχομένου HTML, οι πληροφορίες που εξάγονται από τα ειδησεογραφικά άρθρα προωθούνται σε μια γνωσιακή βάση που βασίζεται σε μοντέλο γράφων ιδιοτήτων με ετικέτες (Labeled Property Graph – LPG) (Webber, 2012). Η συγκεκριμένη δομή αποτελείται από κόμβους (nodes) που αποτελούν τη βασική μορφή δεδομένων και συμβολίζονται με κύκλο, οι οποίοι συνδέονται με άλλους κόμβους μέσω σχέσεων (relationships). Κάθε κόμβος και κάθε σχέση μπορούν να έχουν μία η περισσότερες ιδιότητες (τιμές με όνομα και τιμή), ενώ τα διαφορετικά είδη κόμβων μπορούν να διαχωριστούν με ετικέτες που χαρακτηρίζουν τον ρόλο τους στο γράφο. Οι

σχέσεις μεταξύ κόμβων συμβολίζονται ως ακμές (κατευθυνόμενες ή μη), ενώ δύο κόμβοι επιτρέπεται να έχουν πολλαπλές ή/και αμφίδρομες σχέσεις.

Από την ιχνηλάτηση ειδησεογραφικών δεδομένων προκύπτει αρχικά ένας απλός γράφος που αποτελείται από δύο είδη κόμβων και ένα είδος σχέσης, συγκεκριμένα:

- `Article`: ο κόμβος αυτός αναπαριστά ένα ειδησεογραφικό άρθρο με τις ακόλουθες ιδιότητες (τίτλος, κύρια παράγραφος, σώμα κειμένου, ημερομηνία δημοσίευσης, ημερομηνία ιχνηλάτησης, συγγραφέας, URL πηγής, γλώσσα κειμένου)
- `Section`: ο κόμβος αυτός αναπαριστά ένα τμήμα άρθρου (πχ. πρόταση ή σύντομη φράση) που ανήκει στο εκάστοτε άρθρο και προκύπτει έπειτα από τη χρήση αλγορίθμου τεμαχισμού του σώματος κειμένου κάθε άρθρου σε επιμέρους προτάσεις. Κάθε `Section` κληρονομεί τις ιδιότητες του `Article` στο οποίο ανήκει.
- `HAS_SECTION`: η σχέση αυτή ενώνει κάθε κόμβο `Article` με έναν ή περισσότερους κόμβους `Section` που αντιστοιχούν στο περιεχόμενο του άρθρου.

3.4.1.2 Σύνδεση οντοτήτων

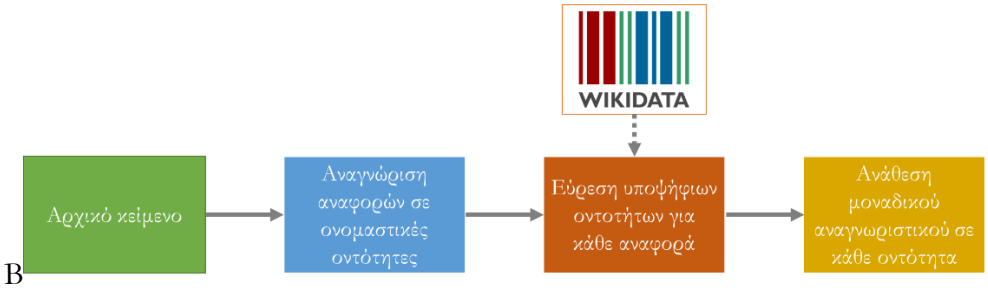
Τα δεδομένα που προέρχονται από τον παραπάνω μηχανισμό συλλογής και επεξεργασίας δεδομένων (ισχυρισμοί χρήστη σε ελεύθερο κείμενο, περιεχόμενο κόμβων `Section` από ειδησεογραφικά άρθρα) τροφοδοτούνται στον μηχανισμό σύνδεσης οντοτήτων προκειμένου να υποβληθούν σε σημασιολογική επισήμανση. Στα πλαίσια της εργασίας μελετήθηκαν δύο στοιχεία σύνδεσης οντοτήτων: το πρώτο περιλαμβάνει την εκπαίδευση νευρωνικού μοντέλου συσχέτισης οντοτήτων με χρήση σημασιολογικού περιεχομένου από γνωσιακές βάσεις δεδομένων (πχ. WikiData, DBpedia) και το δεύτερο τον αυτόματο εντοπισμό λημμάτων της Wikipedia στο κείμενο με χρήση του αλγορίθμου PageRank (Page and Brin, 1998). Τα δύο στοιχεία περιγράφονται παρακάτω.

3.4.1.2.1 Εκπαίδευση μοντέλου σύνδεσης οντοτήτων γνωσιακής βάσης

Έγινε εκπαίδευση στοιχείου σύνδεσης οντοτήτων για την ελληνική γλώσσα (νευρωνικό μοντέλο), με χρήση δεδομένων της ανοιχτής γνωσιακής βάσης WikiData. Η WikiData περιέχει διαγλωσσικές αναφορές για πάνω από 96 εκατομμύρια έννοιες, για οτιδήποτε μπορεί να γίνει αντιληπτό, να παρατηρηθεί ή συζητηθεί (πχ. όνομα ατόμου, τοποθεσίας, οργανισμού, είδους μουσικής κτλ.) συμβάλλοντας ουσιαστικά στην ανάπτυξη του σημασιολογικού Ιστού. Κάθε οντότητα έχει τη δική της ιστοσελίδα και χαρακτηρίζεται από μοναδικό αναγνωριστικό QID της μορφής Q<αριθμός>, το οποίο ανατίθεται κατά τη

δημιουργία της. Όλες οι οντότητες (entities) της γνωσιακής βάσης αποτελούν εκδοχές (instances) συγκεκριμένων κλάσεων (classes), οι οποίες χαρακτηρίζονται από ένα σύνολο κοινών ιδιοτήτων (properties) και αφορούν την περιγραφή μιας συγκεκριμένης κατηγορίας οντοτήτων.

Το πρόβλημα ανίχνευσης οντοτήτων σε ελεύθερο κείμενο μπορεί να θεωρηθεί ως ένα πρόβλημα επιβλεπόμενης μάθησης, όπως περιεγράφηκε στη υποενότητα 2.4.3. Συγκεκριμένα, έστω έγγραφο $\boldsymbol{d}$ (λεκτικό σύνολο) και ένα σημείο θέσης $\boldsymbol{s}$ που περιλαμβάνει ζεύγος δεικτών θέσης $< \boldsymbol{start} >, < \boldsymbol{end} >$ και καθορίζει μια φράση $\boldsymbol{d[s]}$ η οποία αποτελεί υποψήφια ονομαστική οντότητα και πρέπει να αντιστοιχηθεί σε μία εκ των οντοτήτων $\boldsymbol{E[s]}$ της WikiData, της οποίας το $\boldsymbol{d[s]}$ αποτελεί όνομα, παραλλαγή ή ψευδώνυμο. Αρχικά αναγνωρίζονται οι ονομαστικές οντότητες ενός κειμένου και έπειτα εκπαιδεύεται ταξινομητής που αναθέτει για κάθε σημείο $\boldsymbol{s \in d}$ μια οντότητα $\boldsymbol{e \in E[s]}$ που προέρχεται από τη γνωσιακή βάση (Σχήμα 23).



Σχήμα 23 Βήματα εκπαίδευσης στοιχείου σύνδεσης οντοτήτων

Το πρώτο βήμα συνίσταται στην κατάλληλη προσαρμογή της πληροφορίας που βρίσκεται στη γνωσιακή βάση WikiData για την εκπαίδευση του μοντέλου. Αρχικά, κάθε οντότητα αντιστοιχίζεται στη διανυσματική της αναπαράσταση που προκύπτει από την κωδικοποίηση της περιγραφής της (ιδιότητα της WikiData) μέσω της χρήσης προ-εκπαιδευμένου μοντέλου. Ακόμη, καθορίζεται ένα σύνολο βοηθητικών παραμέτρων όπως ο μέγιστος αριθμός υποψήφιων οντοτήτων ανά φράση $\boldsymbol{d[s]}$, και το κατώτερο όριο εμφάνισης μιας οντότητας στο κείμενο (συχνότητα), προκειμένου να ελεγχθεί η προσθήκη της στην προσαρμοσμένη γνωσιακή βάση. Δεδομένου ότι η δεσμευμένη πιθανότητα $\boldsymbol{p(e|d[s])}$ δε μπορεί να υπολογιστεί απευθείας χωρίς την ύπαρξη αντίστοιχης κατανομής πιθανότητας, αυτή προσεγγίζεται μέσω της ανάλυσης Wikipedia άρθρων που περιλαμβάνουν επισημασμένες τις αντίστοιχες οντότητες με τη βοήθεια n-gram μοντέλου. Το τελικό προϊόν της αρχικής φάσης είναι α) η προσαρμοσμένη γνωσιακή βάση με τις διανυσματικές αναπαραστάσεις κάθε οντότητας και β) ένα σύνολο δεδομένων εκπαίδευσης που

περιλαμβάνει προτάσεις προερχόμενες από άρθρα της Wikipedia με επισημασμένες τις ανιχνευμένες σε αυτό οντότητες. Το σύνολο δεδομένων έχει τη μορφή τριπλετών με το κείμενο της πρότασης, τα αναγνωριστικά των αντιστοιχισμένων οντοτήτων και τις θέσεις τους στο κείμενο.

Το δεύτερο βήμα της διαδικασίας περιλαμβάνει την εκπαίδευση ταξινομητή με χρήση νευρωνικού μοντέλου. Αρχικά ένας αλγόριθμος αναγνώρισης ονομαστικών οντοτήτων (NER) απομονώνει τις υποψήφιες προς ταξινόμηση οντότητες σε κάθε πρόταση. Για κάθε ανιχνευθείσα οντότητα, ο ταξινομητής καλείται να την αντιστοιχίσει με ένα μοναδικό αναγνωριστικό QID της προσαρμοσμένης γνωσιακής βάσης, λαμβάνοντας υπόψη τα σημασιολογικά χαρακτηριστικά του κειμένου μέσω των διανυσματικών του αναπαραστάσεων για εξάλειψη φαινομένων αμφισημίας. Η απόδοση του μοντέλου αξιολογείται σε ένα μικρό σύνολο δεδομένων επικύρωσης που παρακρατείται από την αρχική λίστα προτάσεων κατά το στάδιο της εκπαίδευσης.

Μέσω της παραπάνω διαδικασίας εκπαιδεύτηκε νευρωνικό μοντέλο σύνδεσης αναφορών με οντότητες της WikiData. Δεδομένου του μεγέθους της γνωσιακής βάσης (>500GB) και των υπολογιστικών περιορισμών που συνεπάγεται η προεπεξεργασία και εκπαίδευση ενός τέτοιου μοντέλου, χρησιμοποιήθηκε ένα υποσύνολο της βάσης που περιλαμβάνει τις 500.000 πιο συχνά εμφανιζόμενες οντότητες που ανήκουν σε μία εκ των παρακάτω κλάσεων ("EVENT", "GPE", "LOC", "ORG", "PERSON" και "PRODUCT"). Περισσότερες τεχνικές πληροφορίες σχετικά με τα υπολογιστικά εργαλεία που χρησιμοποιήθηκαν, τις παραμέτρους και την απόδοση του ταξινομητή παρατίθενται στο Κεφάλαιο 4.

3.4.1.2.2 Αυτόματη αντιστοίχιση εννοιών με οντότητες της Wikipedia (Wikification)

Η διαδικασία σύνδεσης λεκτικών συνόλων με οντότητες της γνωσιακής βάσης WikiData αποτελείται από επιμέρους αλληλένδετες διαδικασίες. Αρχικά, προσδιορίζονται οι φράσεις/λέξεις που αναφέρονται σε μια έννοια που περιλαμβάνεται στην οντολογία. Στη συνέχεια, τα σύνολα αυτά αντιστοιχίζονται με τις αντίστοιχες οντότητες και τέλος καθορίζεται ποιες από τις οντότητες είναι αρκετά σχετικές με το έγγραφο ώστε να περιληφθούν στην έξοδο του στοιχείου. Το στοιχείο αυτόματης σημασιολογικής επισήμανσης στηρίζεται στο έργο των Brank et al., 2018 που προτείνει τη σύνδεση οντοτήτων βάσει του αλγορίθμου PageRank, αξιοποιώντας τους υπερσυνδέσμους μεταξύ ιστοσελίδων της Wikipedia. Κάθε υπερσύνδεσμος μπορεί να θεωρηθεί ως ένας συνδυασμός τριών αντικειμένων: μιας ιστοσελίδας πηγής (source page), μιας ιστοσελίδας προορισμού (target page) και ενός κειμένου αγκύρωσης (anchor text). Αν μια σελίδα πηγής s περιέχει

σύνδεσμο με το κείμενο αγκύρωσης a που παραπέμπει στην σελίδα προορισμού t που περιγράφει μια έννοια c, αυτό αποτελεί ένδειξη ότι η φράση a μπορεί να αποτελεί αναπαράσταση ή αναφορά της έννοιας που περιγράφεται στην σελίδα προορισμού t. Επομένως, αν το κείμενο που υποβάλλεται σε διαδικασία wikification περιέχει τη φράση a, ενδέχεται αυτή να αποτελεί αναφορά της έννοιας t (που αποτελεί οντότητα στη γνωσιακή βάση) και άρα η έννοια c αποτελεί υποψήφια οντότητα προς επισήμανση της φράσης a. Ωστόσο, τις περισσότερες φορές μια φράση μπορεί να έχει παραπάνω από μια υποψήφιες οντότητες (αμφισημία) και στόχος της μεθόδου είναι η αντιστοίχιση με την πιο σχετική οντότητα, με την προϋπόθεση ότι το κείμενο αφορά συγκεκριμένο αντικείμενο. Για παράδειγμα, η αναφορά "Ελευθέριος Βενιζέλος" σε μια πρόταση που αφορά αναχωρήσεις αεροσκαφών, προφανώς σχετίζεται με το αεροδρόμιο και όχι με τον Έλληνα πολιτικό.

Η αντιστοίχιση της σχετικότερης οντότητας σε κάθε φράση του κειμένου βασίζεται στον σχηματισμό διμερούς γράφου (bipartite graph) μεταξύ των αναφορών a και των εννοιών c της γνωσιακής βάσης. Ορίζεται επίσης κατευθυνόμενη ακμή $a \rightarrow c$, αν και μόνο αν η έννοια c ανήκει στις υποψήφιες προς επισήμανση οντότητες για την φράση a, συνοδευόμενη από την σχετική πιθανότητα μετάβασης $P(a \rightarrow c)$. Η πιθανότητα αυτή υπολογίζεται ως εξής:

$$P(a \rightarrow c) = \frac{\#\ \textit{υπερσυνδέσμων στη Wikipedia με κείμενο αγκύρωσης a και στόχο c}}{\#\ \textit{υπερσυνδέσμων στη Wikipedia με κείμενο αγκύρωσης a}} \qquad 3.1$$

Ο παραπάνω γράφος εμπλουτίζεται επίσης από ακμές μεταξύ των εννοιών $c \rightarrow c'$ βάσει της σημασιολογικής εγγύτητας τους (semantic relatedness – SR), η οποία υπολογίζεται από τη σχέση 3.2:

$$SR(c \rightarrow c') = 1 - \frac{[log(max\{|L_c|, |L_{c'}|\}) - log(|L_c| \cap |L_{c'}|)}{logN - log(min\{|L_c|, |L_{c'}|\}} \qquad 3.2$$

όπου L_c ο αριθμός των σελίδων που περιέχουν υπερσύνδεσμο προς την έννοια c και N ο συνολικός αριθμός των εννοιών που περιλαμβάνονται στη γνωσιακή βάση.

Η πιθανότητα μετάβασης $P(c \rightarrow c')$ ορίζεται ως:

$$P(c \rightarrow c') = \frac{SR(c, c')}{\sum_{c''} SR(c, c'')} \qquad 3.3$$

Ο παραπάνω γράφος αποτελεί τη βάση υπολογισμού ενός διανύσματος που περιλαμβάνει τις PageRank τιμές για κάθε κορυφή, ακολουθώντας μια επαναληπτική μέθοδο όπου σε κάθε επανάληψη κάθε κορυφή v μεταφέρει ένα ποσοστό της PageRank τιμής της στους

άμεσους γείτονές της u στο γράφο, αναλογικό της πιθανότητας μετάβασης της εκάστοτε εξερχόμενης ακμής:

$$PR_{new}(u) = \tau \cdot PR_o(u) + (1-\tau)\sum_{v} PR_{old}(v)P(v \to u) \qquad 3.4$$

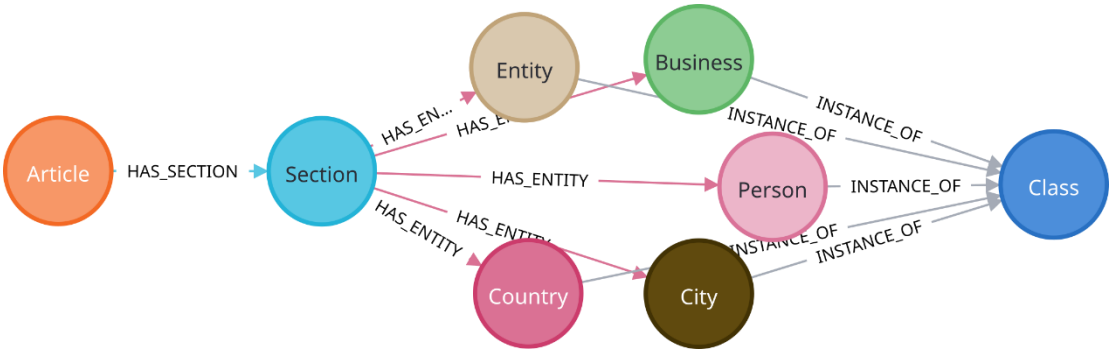
όπου PR_o μια τιμή αναφοράς για κάθε κορυφή (αν η κορυφή είναι έννοια c ισούται με 0, αλλιώς αν η κορυφή αφορά φράση α ισούται με τον λόγο εμφάνισης της σε στο σύνολο των σελίδων της Wikipedia) και $\tau = 0.1$ μια παράμετρος κανονικοποίησης.

Σε κάθε επανάληψη υπολογισμού της τιμής PageRank, αυτή ρέει προς τις κορυφές εννοιών c από κορυφές αναφορών a που σχετίζονται με έννοιες c, καθώς και από άλλες κορυφές εννοιών c' που είναι σημασιολογικά κοντά ως προς τις έννοιες c. Αυτό έχει ως αποτέλεσμα τη συσσώρευση της τιμής σε σύνολα εννοιών που σχετίζονται σημασιολογικά μεταξύ τους, και παράλληλα συσχετίζονται με αναφορές του κειμένου εισαγωγής. Αφού ολοκληρωθεί η διαδικασία, οι τιμές PageRank των εννοιών χρησιμοποιούνται για την αποσαφήνιση των αναφορών ως εξής: Αν μια αναφορά-φράση συνδέεται με πολλαπλές έννοιες, επιλέγεται η έννοια εκείνη με τη μεγαλύτερη τιμή PageRank και η φράση αντιστοιχίζεται με το μοναδικό αναγνωριστικό QID της σχετικής WikiData οντότητας. Δεδομένου ότι κάθε λέξη ενός κειμένου αποτελεί εν δυνάμει αναφορά προς επισήμανση, χρησιμοποιείται ένα κατώφλι που επιτρέπει την αντιστοίχιση οντοτήτων μόνο σε περιπτώσεις που η τιμή ξεπερνά το προκαθορισμένο όριο.

Η παραπάνω μεθοδολογία διατίθεται σε μορφή Web API και επιτρέπει την επισήμανση αναφορών με WikiData οντότητες σε πραγματικό χρόνο. Σε αντίθεση με το στοιχείο σύνδεσης οντοτήτων που περιεγράφηκε στην προηγούμενη υποενότητα, ενσωματώνει μεγαλύτερο μέρος της γνωσιακής βάσης χωρίς να υπόκειται σε υπολογιστικούς περιορισμούς από πλευράς τελικού χρήστη, με το μοναδικό του μειονέκτημα να είναι η ανάγκη συνεχούς σύνδεσης στο Διαδίκτυο. Καθότι η σύνδεση οντοτήτων αποτελεί βασικό προαπαιτούμενο για την μεθοδολογία επικύρωσης ισχυρισμών, η χρήση του στοιχείου αυτού προτιμήθηκε έναντι του προηγούμενου κατά τη διαδικασία ένταξης των διαφορετικών υποσυστημάτων στο τελικό ενοποιημένο σύστημα.

Οι έννοιες που ανιχνεύονται στους κόμβους `Section` των ειδησεογραφικών άρθρων αναπαρίστανται στη γνωσιακή βάση ως κόμβοι τύπου `Entity`, η οποία περιλαμβάνει ως ιδιότητες το μοναδικό αναγνωριστικό QID της αντίστοιχης οντότητας κατά τη WikiData οντολογία και το πλήρες όνομά της. Τα δύο είδη κόμβων συνδέονται μεταξύ τους με τη

σχέση HAS_ENTITY, εφόσον η πρόταση ενός Section περιλαμβάνει αναφορά στη συγκεκριμένη οντότητα. Ο εικονικός γράφος (ή μετα-γράφος) στο Σχήμα 24 αναπαριστά την τελική δομή της γνωσιακής βάσης μετά τη διαδικασία σύνδεσης οντοτήτων. Σημειώνεται ότι ένας κόμβος τύπου Entity μπορεί να περιλαμβάνει επιπλέον ετικέτες (πχ. Person, City, Business) εκτός της γενικής Entity, βάσει της ταξινομίας Wikidata.



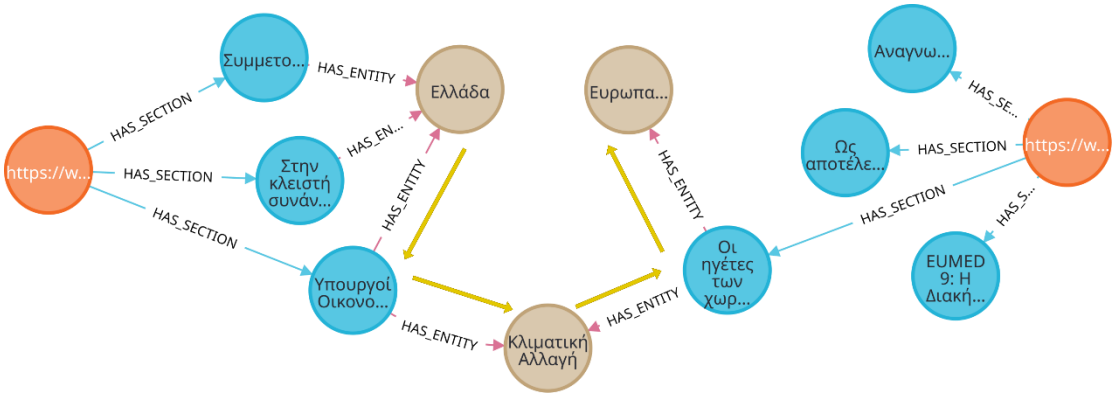
Σχήμα 24 Εικονικός γράφος (μεταγράφος) γνωσιακής βάσης

3.4.1.3 Κατασκευή τεκμηρίων για έλεγχο ισχυρισμών

Σε ένα τυπικό πρόβλημα αναγνώρισης κειμενικής συνεπαγωγής (NLI), το σώμα κειμένου που σχηματίζει την προϋπόθεση (premise) αναπαριστά τη γνώση ή τα διαθέσιμα δεδομένα που σχετίζονται με ένα γεγονός και ως εκ τούτου χρησιμοποιείται για την επαλήθευση μιας υπόθεσης (hypothesis). Στην περίπτωση επικύρωσης ενός ισχυρισμού ωστόσο, τα διαθέσιμα δεδομένα μπορεί βρίσκονται διάσπαρτα σε πολλαπλά τμήματα άρθρων. Παράλληλα, τα τμήματα που αφορούν έναν συγκεκριμένο ισχυρισμό μπορούν να απομονωθούν εύκολα ανιχνεύοντας τις κοινές μεταξύ τους οντότητες, με σκοπό την κατασκευή ενός συνόλου τεκμηρίων που θα οδηγήσει δυνητικά στην επικύρωση ενός ισχυρισμού. Επομένως, το πρόβλημα κατασκευής τεκμηρίων έγκειται στην εργασία αναζήτησης της γνωσιακής βάσης για όλα τα αποσπάσματα (Sections) ειδησεογραφικών άρθρων που συνδέονται με οντότητες (Entities) οι οποίες ανιχνεύονται στον ισχυρισμό του χρήστη. Για το σκοπό αυτό, αναπτύχθηκε αλγόριθμος που αποτελείται από τα παρακάτω βήματα:

1. Ο ισχυρισμός που παρατίθεται από τον χρήστη υποβάλλεται σε διεργασία σύνδεσης οντοτήτων μέσω του μηχανισμού που περιεγράφηκε στην υποενότητα 3.4.1.2. Η λίστα ανιχνευθεισών οντοτήτων (Wikidata QIDs) θα αποτελέσει παράμετρο για το επόμενο βήμα.

2. Υποβάλλεται ερώτημα στη βάση για την εύρεση όλων των συντομότερων διαδρομών που συνδέουν τις ανιχνευθείσες οντότητες (κόμβοι `Entity`) με αποσπάσματα άρθρων (κόμβοι `Section`). Δεδομένης της δομής του γράφου και n κόμβων Entity, αυτό μεταφράζεται σε ελάχιστο μήκος διαδρομής $2(n-1)$ εναλλασσόμενων κόμβων Entity-Section, όπως φαίνεται στο Σχήμα 25. Επισημαίνεται ότι παραπάνω από μια συντομότερες διαδρομές μπορεί να επιστραφούν για ένα σύνολο οντοτήτων. Σημειώνεται ακόμη ότι η παραπάνω σύμβαση ενδέχεται να μην εκπληρώνεται πάντα, καθώς εξαρτάται από το μέγεθος και την ποικιλομορφία της πληροφορίας που έχει αποτυπωθεί στον γράφο μέσω της διαδικασίας εμπλουτισμού. Σε περιπτώσεις μη εύρεσης συντομότερης διαδρομής, ο αλγόριθμος απλά επιστρέφει όλα τα αποσπάσματα άρθρων που συνδέονται με τουλάχιστον μια οντότητα από αυτές που αναφέρθηκαν στον ισχυρισμό.
3. Τα αποσπάσματα άρθρων που ανήκουν σε καθεμία από τις ευρεθείσες συντομότερες διαδρομές συσσωματώνονται σε μια ακολουθία seq η οποία αποτελεί υποψήφιο τεκμήριο για την επικύρωση του ισχυρισμού.
4. Όλα τα υποψήφια τεκμήρια εισάγονται στον μηχανισμό ελέγχου σημασιολογικής ομοιότητας ώστε να επιλεγεί εκείνο (seq^*) που εμφανίζει τη μεγαλύτερη συσχέτιση με τον αρχικό ισχυρισμό.
5. Το καλύτερο υποψήφιο τεκμήριο συγκρίνεται με τον ισχυρισμό μέσω του μηχανισμού συνεπαγωγής και προκύπτουν οι τιμές συνεπαγωγής, αντίφασης και ουδέτερης σχέσης $NLI\_score < c, e, n >$. Κατ' αυτό τον τρόπο είναι δυνατή η ποσοτική αξιολόγηση της εγκυρότητας ενός ισχυρισμού, βάσει των διαθέσιμων δεδομένων.



Σχήμα 25 Συντομότερη διαδρομή μεταξύ οντοτήτων που ανιχνεύθηκαν σε ισχυρισμό

Ο παραπάνω αλγόριθμος παρατίθεται παρακάτω σε μορφή ψευδοκώδικα:

Algorithm 1: Evidence Construction

Input: A claim $\boldsymbol{c}$ provided by the user in natural language

Output: Most relevant evidence (sequence of sentences) $\mathbf{seq}^*$ based on the input claim $\boldsymbol{c}$ along with its $\mathbf{STS\_score}^*$ and $\mathbf{NLI\_score} < \mathbf{c}, \mathbf{e}, \mathbf{n} >$

1: Entity Linking: Find the set of entities $(\boldsymbol{e_1}, \dots, \boldsymbol{e_n}) \in \boldsymbol{E}$ where $\boldsymbol{e} \exists \ \boldsymbol{c}$ and $|\boldsymbol{E}| = \boldsymbol{n}$

2: $\boldsymbol{S} \leftarrow \emptyset$

3: Graph database search: Find all shortest paths between the alternating entities $\boldsymbol{e}$ and sentences $\boldsymbol{s}$:

$$\boldsymbol{p} \leftarrow (\boldsymbol{e_1}, \boldsymbol{s_a}, \boldsymbol{e_2}, \dots, \boldsymbol{e_{n-1}}, \boldsymbol{s_k}, \boldsymbol{e_n}) \in \boldsymbol{P}$$

4: if $\boldsymbol{P} = \emptyset$ then

5: $\boldsymbol{s} \in \boldsymbol{P} \iff \boldsymbol{s}$ has at least 1 entity mention

6: end if

7: for $\boldsymbol{p_i} \in \boldsymbol{P}$ do

8: $\boldsymbol{seq_i} \leftarrow (\boldsymbol{s_a}, \dots, \boldsymbol{s_k})$

9: $\boldsymbol{S} \leftarrow \boldsymbol{S} \cup \boldsymbol{seq_i}$ (sequence $\boldsymbol{seq\_i}$ added to candidate evidence set)

10: end for

11: $\boldsymbol{STS\_Scores} \leftarrow \emptyset$

12: for $\boldsymbol{seq_i} \in \boldsymbol{S}$ do

13: Semantic Textual Similarity: Compare $\boldsymbol{seq_i} \in \boldsymbol{S}$to $\boldsymbol{c}$ (each candidate evidence sequence to the claim) and calculate $\boldsymbol{STS\_score_i}$

$$\boldsymbol{STS\_Scores} \leftarrow \boldsymbol{STS\_Scores} \cup \boldsymbol{STS\_score_i}$$

14: end for

15: Find the candidate $\boldsymbol{seq}^*$ with the highest similarity to the claim:

$$STS\_score^* \leftarrow max(STS\_Scores)$$

$$seq^* \leftarrow argmax(STS\_score^*)$$

16: Natural Language Inference: Compare seq^* to c (the best candidate evidence to the claim) and calculate the scores for contradiction, entailment and neutrality:

$$NLI\_score < c, e, n >$$

3.4.1.4 Έλεγχος σημασιολογικής ομοιότητας

Ο συγκεκριμένος μηχανισμός βασίζεται στην εκπαίδευση νευρωνικού μοντέλου σημασιολογικής ομοιότητας σε επίπεδο διανυσματικών αναπαραστάσεων προτάσεων (sentence embeddings). Στόχος είναι η σύγκριση του ισχυρισμού ενός χρήστη με τα υποψήφια τεκμήρια που κατασκευάστηκαν στην προηγούμενη φάση σύμφωνα με τη διαδικασία που περιεγράφηκε στην υποενότητα 3.4.1.3, προκειμένου να επιλεγεί εκείνο που σχετίζεται περισσότερο με τον ισχυρισμό σε σημασιολογικό επίπεδο.

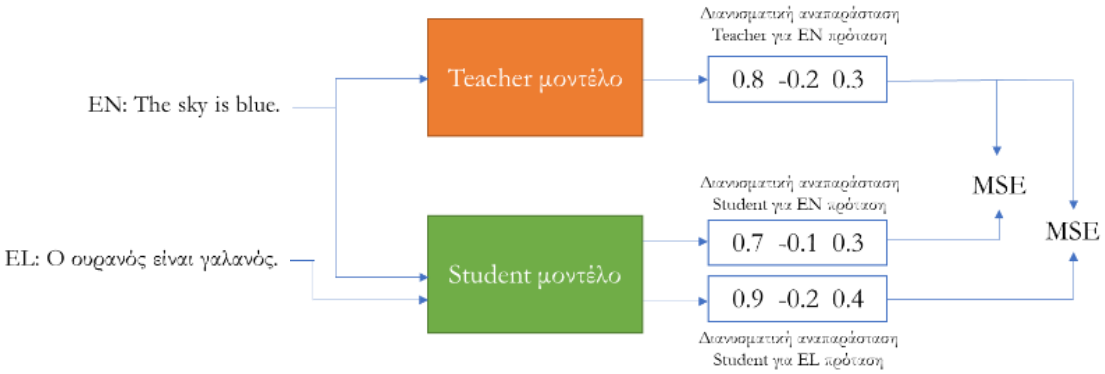
Σημειώνεται ότι, παρά την αφθονία πολύγλωσσων γλωσσικών μοντέλων (πχ. m-BERT, XLM) για την παραγωγή διανυσματικών αναπαραστάσεων από ελεύθερο κείμενο, η απόδοσή τους εμφανίζει διακυμάνσεις μεταξύ γλωσσών υψηλών πόρων (πχ. αγγλικά, γερμανικά κτλ.) σε σύγκριση με γλώσσες χαμηλότερων πόρων όπως η ελληνική (Koutsikakis et al., 2020), ως φυσικό επακόλουθο της υπο-εκπροσώπησής τους στο σύνολο δεδομένων που παρέχεται κατά την εκπαίδευση αντίστοιχων μοντέλων. Επιπροσθέτως, προτάσεις με το ίδιο περιεχόμενο σε διαφορετικές γλώσσες ενδέχεται να αντιστοιχούν σε διαφορετικά σημεία ενός κοινού (πολύγλωσσου) διανυσματικού χώρου, εξαιτίας της έλλειψης αντιστοίχισης μεταξύ των επιμέρους διανυσματικών χώρων κάθε γλώσσας.

Προκειμένου να αντιμετωπιστούν τα παραπάνω κωλύματα, εκπαιδεύτηκε μοντέλο παραγωγής διανυσματικών αναπαραστάσεων με χρήση παράλληλου κειμένου στην ελληνική και στην αγγλική γλώσσα, ακολουθώντας την τεχνική απόσταξης γνώσης (knowledge distillation) που προτάθηκε από τους Reimers and Gurevych, 2020a. Οι τεχνικές απόσταξης αποτελούν παραλλαγή των μεθόδων μεταφοράς μάθησης (transfer learning) καθώς οι πρώτες στοχεύουν στη μεταφορά γνώσης από ένα συνήθως μεγαλύτερο νευρωνικό μοντέλο (teacher) σε ένα μικρότερο (student) χωρίς την απώλεια στην απόδοση, ενώ οι δεύτερες αφορούν την μεταφορά των βαρών (weights) ενός προεκπαιδευμένου δικτύου σε ένα νέο πανομοιότυπης αρχιτεκτονικής αλλά με διαφορετικό στρώμα εξόδου για επαναπροσδιορισμό της τελικής χρήσης του.

Η συγκεκριμένη μεθοδολογία απαιτεί την ύπαρξη ενός προεκπαιδευμένου teacher μοντέλου M που αντιστοιχεί προτάσεις μιας ή περισσότερων γλωσσών σε ένα πυκνό διανυσματικό χώρο. Επιπλέον, απαιτείται η ύπαρξη παράλληλου κειμένου $\left((s_1, t_1), \dots, (s_n, t_n)\right)$ όπου s_i πρόταση σε μια εκτός γλωσσών-πηγής (πχ. αγγλικά) και t_i πρόταση στη γλώσσα-στόχο (πχ. ελληνικά). Η εκπαίδευση του student μοντέλου $\widehat{M}$ στοχεύει στην ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (MSE) για υποσύνολο δεδομένων εκπαίδευσης (minibatch) B, έτσι ώστε $\widehat{M}(s_i) \approx M(s_i)$ και $\widehat{M}(t_i) \approx M(s_i)$, όπως διατυπώνεται στην παρακάτω σχέση:

$$min\left(\frac{1}{|B|}\sum_{i\in B}\left[\left(M(s_i)-\widehat{M}(s_i)\right)^2+\left(M(s_i)-\widehat{M}(t_i)\right)^2\right]\right) \quad 3.5$$

Η διαδικασία απεικονίζεται στο Σχήμα 26. Πρακτικά, η απόσταξη γνώσης στοχεύει στην εκμετάλλευση της καλής γενικευτικής ικανότητας ενός προεκπαιδευμένου μοντέλου, με σκοπό την εξομάλυνση της απόδοσης ενός νέου που θα "κληρονομήσει" τα καλά χαρακτηριστικά του πρώτου επεκτείνοντάς τα για μια νέα γλώσσα, μέσω της χρήσης παράλληλου κειμένου.



Σχήμα 26 Διαδικασία απόσταξης γνώσης για εκπαίδευση ελληνικού μοντέλου σημασιολογικής ομοιότητας

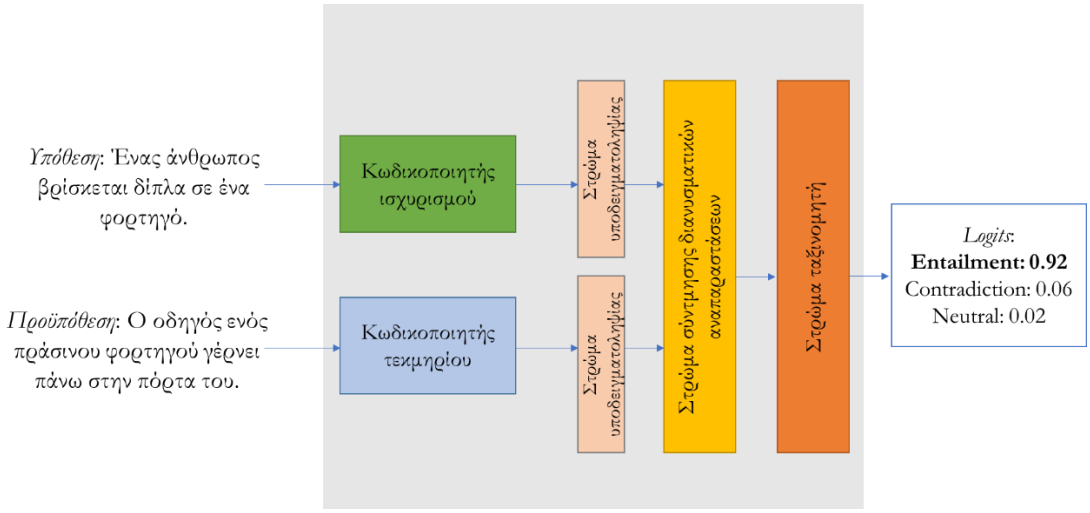
Σημειώνεται ότι δεν υπάρχει περιορισμός στην επιλογή αρχιτεκτονικής για το student μοντέλο $\widehat{M}$. Στην περίπτωσή μας χρησιμοποιήθηκε αρχιτεκτονική XLM-RoBERTa (Conneau et al., 2020) για το $\widehat{M}$, το οποίο εκπαιδεύτηκε βάσει teacher πολύγλωσσου μοντέλου αρχιτεκτονικής DistilRoBERTa (Sanh et al., 2019). Σημειώνεται ότι τα ελληνικά δεν ανήκουν στις γλώσσες που υποστηρίζει το teacher μοντέλο. Οι παραγόμενες διανυσματικές αναπαραστάσεις που προκύπτουν από το εκπαιδευμένο μοντέλο για τον ισχυρισμό χρήστη και τα υποψήφια τεκμήρια συγκρίνονται με χρήση του μέτρου ομοιότητας συνημιτόνου, ώστε να προκύψει το τεκμήριο που θα προωθηθεί στην τελική φάση κειμενικής συνεπαγωγής.

3.4.1.5 Αναγνώριση κειμενικής συνεπαγωγής

Το τελευταίο στάδιο της διαδικασίας επικύρωσης ισχυρισμών περιλαμβάνει τη χρήση μηχανισμού αναγνώρισης κειμενικής συνεπαγωγής προκειμένου να κριθεί εάν ο ισχυρισμός του χρήστη (υπόθεση) επιβεβαιώνεται, απορρίπτεται ή είναι ουδέτερος ως προς το σχετικότερο τεκμήριο (προϋπόθεση) της προηγούμενης φάσης. Προκειμένου να παρακαμφθούν οι περιορισμοί που αναλύθηκαν στην προηγούμενη υποενότητα και αφορούν

τα πολύγλωσσα γλωσσικά μοντέλα, έγινε προσαρμογή (finetuning) μοντέλου NLI για την ελληνική γλώσσα με χρήση δίγλωσσου κειμένου (στην ελληνική και στην αγγλική γλώσσα).

Η εκπαίδευση του σχετικού μοντέλου βασίζεται στην μεθοδολογία των Reimers and Gurevych, 2020b (βλ. υποενότητα 2.7.3) και περιλαμβάνει τη χρήση δύο ενωμένων (σιαμαίων) κωδικοποιητών (Siamese BERT-networks), επιτρέποντας στο μοντέλο να εκτελεί εργασίες που δεν ήταν εφικτές μέσω της εκπαίδευσης κωδικοποιητή πρωτότυπης αρχιτεκτονικής. Το συγκεκριμένο μοντέλο καλείται Bi-Encoder και περιλαμβάνει δύο κωδικοποιητές (encoders), καθένας εκ των οποίων δέχεται μια εκ των δύο προτάσεων (ισχυρισμός, τεκμήριο) για την παραγωγή των διανυσματικών τους αναπαραστάσεων, οι οποίες στη συνέχεια μετατρέπονται σε διανύσματα σταθερού μήκους μέσω στρωμάτων υποδειγματοληψίας. Έπειτα, τα δύο διανύσματα ενώνονται μέσω ενός στρώματος σύντμησης. Τέλος, ένα στρώμα ταξινομητή (3-way-softmax-classifier) εκπαιδεύεται σε επισημασμένο σύνολο δεδομένων εκπαίδευσης, ώστε να υπολογίζει τα logits καθεμίας εκ των τριών ετικετών (e:entailment, c:contradiction, n:neutral). Η αρχιτεκτονική bi-encoder απεικονίζεται στο Σχήμα 27.



Σχήμα 27 Επισκόπηση αρχιτεκτονικής μοντέλου κειμενικής συνεπαγωγής για την επικύρωση ισχυρισμών

Μέσω του παραπάνω μηχανισμού, ελέγχεται εάν το τεκμήριο που εμφανίζει τη μεγαλύτερη σημασιολογική ομοιότητα με τον ισχυρισμό του χρήστη μπορεί να οδηγήσει στην επικύρωσή του. Η διαδικασία κατασκευής τεκμηρίων εξασφαλίζει τη συγκέντρωση πληροφοριών από πολλαπλές ειδησεογραφικές πηγές για τον σχηματισμό κατάλληλων τεκμηρίων, προσομοιάζοντας σε κάποιο βαθμό την ανθρώπινη διεργασία εξερεύνησης και αξιολόγησης αποδεικτικών στοιχείων και σχετικών δεδομένων για την εξαγωγή συμπερασμάτων. Τεχνικές πληροφορίες σχετικά με την υλοποίηση και αξιολόγηση του

μηχανισμού εξαγωγής σημασιολογικών συμπερασμάτων παρατίθενται στο επόμενο Κεφάλαιο, ενώ στο Κεφάλαιο 5 της εργασίας αναλύονται τα γενικότερα συμπεράσματα που προκύπτουν αναφορικά με τους περιορισμούς και τις μελλοντικές ερευνητικές κατευθύνσεις της συγκεκριμένης έρευνας.

3.4.1.6 Αναγνώριση υποκειμενικότητας/αντικειμενικότητας

Όπως αναφέρθηκε στην υποενότητα 3.4.1.3, το στάδιο κατασκευής τεκμηρίων βασίζεται σε μεθόδους ελέγχου σημασιολογικής ομοιότητας και αναγνώρισης γλωσσικής συνεπαγωγής για να ελέγξει εάν το σχετικότερο εννοιολογικά υποσύνολο της διαθέσιμης πληροφορίας μπορεί να επικυρώσει ή να απορρίψει τον ισχυρισμό ενός χρήστη. Ωστόσο, δεδομένου ότι η πληροφορία παρέχεται εξ' ολοκλήρου από αποσπάσματα ειδησεογραφικών άρθρων, ενδέχεται το επιλεγμένο τεκμήριο που θα χρησιμοποιηθεί να αφορά μια υποκειμενική άποψη, οπτική ή πεποίθηση (πχ. απόσπασμα συνέντευξης) και όχι αντικειμενική πληροφορία (Chaturvedi et al., 2018). Η αναγνώριση υποκειμενικότητας/αντικειμενικότητας (subjectivity/objectivity detection) αποτελεί προαιρετικό στάδιο ελέγχου των διαθέσιμων τεκμηρίων και στοχεύει στη διάκριση των αντικειμενικών προτάσεων, οι οποίες εκφράζουν πραγματική πληροφορία από τις υποκειμενικές προτάσεις, οι οποίες εκφράζουν συνήθως μια προσωπική άποψη. Σημειώνεται ότι μια τέτοια διάκριση παρουσιάζει σημαντικές δυσκολίες, αφενός γιατί οι έννοιες της υποκειμενικότητας και της αντικειμενικότητας δεν είναι απόλυτες αλλά εξαρτώνται από την εκάστοτε ανθρώπινη κρίση και ερμηνεύονται διαφορετικά ανά περίπτωση, και αφετέρου γιατί η μοντελοποίηση μιας τέτοιας διαδικασίας (ως εργασία ταξινόμησης) απαιτεί πρόσβαση σε αντιπροσωπευτικό, επισημασμένο σύνολο δεδομένων.

Στα πλαίσια της διατριβής, αναπτύχθηκε μηχανισμός ταξινόμησης των διαθέσιμων τεκμηρίων σε υποκειμενικές και αντικειμενικές προτάσεις, με αξιοποίηση δίγλωσσου επισημασμένου κειμένου (ελληνικό και αγγλικό κείμενο). Συγκεκριμένα, χρησιμοποιήθηκε αρχιτεκτονική XLM-RoBERTa (Conneau et al., 2020) όπως και για τους προηγούμενους μηχανισμούς, με τη διαφορά ότι στο συγκεκριμένο μοντέλο προστέθηκε κεφαλή δυαδικής ταξινόμησης (0: αντικειμενική πρόταση, 1: υποκειμενική πρόταση). Στόχος του συγκεκριμένου προαιρετικού σταδίου είναι να πληροφορεί τον χρήστη για την αντικειμενικότητα του τεκμηρίου που χρησιμοποιείται κατά τον έλεγχο του ισχυρισμού, προκειμένου να αποφεύγεται η επικύρωση ισχυρισμών που βασίζεται σε απόψεις και όχι σε γεγονότα. Τεχνικές λεπτομέρειες για την εκπαίδευση και απόδοση του μοντέλου αναγνώρισης υποκειμενικότητας/αντικειμενικότητας δίνονται στο επόμενο κεφάλαιο.

3.5 Σύνοψη κεφαλαίου

Στο κεφάλαιο αυτό αναπτύχθηκαν διεξοδικά οι δύο κύριοι ερευνητικοί άξονες που απαντούν στις υποθέσεις εργασίας της διδακτορικής διατριβής, δίνοντας έμφαση στις μεθοδολογικές επιλογές και παραδοχές των επιμέρους μηχανισμών. Έγινε αναφορά στην αρχιτεκτονική σχεδίαση κάθε μηχανισμού και περιγραφή των στοιχείων του, με ανάλυση του σχετικού μαθηματικού υποβάθρου και των αλγοριθμικών διεργασιών που τον χαρακτηρίζουν. Σε περιπτώσεις ανάπτυξης αλληλοεπικαλυπτόμενων στοιχείων που μελετήθηκαν για την επίτευξη συγκεκριμένων επιμέρους στόχων, παρατέθηκαν τα πλεονεκτήματα και μειονεκτήματα του καθενός, προκειμένου να αιτιολογηθεί η τελική αρχιτεκτονική επιλογή. Παρόλο που η ανωτέρω περιγραφείσα μεθοδολογία έρχεται να συμπληρώσει το κενό ως προς την ύπαρξη συστημάτων εξαγωγής πληροφοριών και σημασιολογικών συμπερασμάτων συγκεκριμένα για την ελληνική γλώσσα, η επιλογή αρθρωτής σχεδίασης κάθε μηχανισμού επιτρέπει την τροποποίηση ή αντικατάσταση επιμέρους στοιχείων για την υποστήριξη οποιασδήποτε άλλης γλώσσας χαμηλών πόρων.

Στο επόμενο κεφάλαιο, θα περιγραφεί το στάδιο τεχνικής υλοποίησης των παραπάνω μηχανισμών, με αναφορά στην επιλογή συγκεκριμένων τεχνολογιών, στην ενσωμάτωση των επιμέρους στοιχείων σε ένα ενιαίο σύστημα και στη διεξαγωγή πειραμάτων για την εξαγωγή αποτελεσμάτων που αφορούν τόσο τη συνολική απόδοση του συστήματος όσο και των επιμέρους στοιχείων που το απαρτίζουν.

4 Τεχνική Υλοποίηση και Αποτελέσματα

4.1 Εισαγωγή

Ακολουθώντας τη μεθοδολογία που περιεγράφηκε στο Κεφάλαιο 3, το επόμενο βήμα περιλαμβάνει την τεχνική υλοποίηση του μηχανισμού εξόρυξης πληροφοριών από ελεύθερο κείμενο και του μηχανισμού εξαγωγής σημασιολογικών συμπερασμάτων. Οι δύο μηχανισμοί, αναπτύσσονται ακολουθώντας τις κατευθυντήριες γραμμές του προηγούμενου κεφαλαίου, αναδεικνύονται μέσω σχετικών περιπτώσεων χρήσης και αξιολογούνται με τη βοήθεια συνόλων δεδομένων αναφοράς (benchmarks).

Το παρόν κεφάλαιο περιγράφει τις τεχνολογικές επιλογές και τις παραμέτρους διαμόρφωσης των επιμέρους στοιχείων που απαρτίζουν τους δύο μηχανισμούς, παραθέτοντας επίσης πληροφορίες που αφορούν την απόδοσή τους, όπως αυτή μετρήθηκε κατά την δοκιμή τους στα πλαίσια συγκεκριμένων περιπτώσεων χρήσης. Η δομή του κεφαλαίου χαρακτηρίζεται από τον ίδιο βαθμό ανάλυσης με αυτόν του προηγούμενου, ώστε να διευκολυνθεί η αντιστοίχιση μεταξύ της αρχιτεκτονικής των δύο μηχανισμών με τις τεχνικές λεπτομέρειες και τα αποτελέσματα που προέκυψαν από τη χρήση τους. Ειδικότερα όσον αφορά τα εκπαιδευμένα μοντέλα που αφορούν την ελληνική γλώσσα, και τα οποία αποτελούν υποπροϊόντα της παρούσας διατριβής, δίνονται πληροφορίες για την ελεύθερη πρόσβαση και χρήση τους από το κοινό.

4.2 Μηχανισμός εξόρυξης πληροφοριών από ελεύθερο κείμενο

4.2.1 Τεχνική υλοποίηση

Στις ακόλουθες υποενότητες περιγράφονται τα τεχνικά χαρακτηριστικά των στοιχείων που απαρτίζουν τον μηχανισμό εξόρυξης πληροφοριών και συγκεκριμένα των στοιχείων μηχανικής μετάφρασης και ευθυγράμμισης κειμένου, του στοιχείου επίλυσης συναναφορών, του στοιχείου εξαγωγικής συνόψισης και των στοιχείων παράλληλης ανοιχτής εξαγωγής πληροφοριών. Δίνεται έμφαση στην επισκόπηση των τεχνολογιών, δεδομένων και υπερ-παραμέτρων που αφορούν τα στοιχεία τα οποία εκπαιδεύτηκαν στα πλαίσια της εργασίας, ενώ παρατίθενται τεχνικές λεπτομέρειες σχετικά με την παραμετροποίηση των προεκπαιδευμένων μοντέλων και συστημάτων που αξιοποιήθηκαν.

4.2.1.1 Στοιχεία μηχανικής μετάφρασης και ευθυγράμμισης κειμένου

Για την εκπαίδευση μοντέλων μηχανικής μετάφρασης από ελληνικά σε αγγλικά (EL-EN) και αντίστροφα (EN-EL) έγινε χρήση παράλληλου κειμένου από δύο κύριες πηγές:

1) από το αποθετήριο OPUS (Tiedemann, 2012), από το οποίο αντλήθηκαν τα σύνολα δεδομένων ParaCrawl, OpenSubtitles, EUBookshop, DGT και Europarl.
2) από το αποθετήριο CCMatrix (Schwenk et al., 2019) που περιλαμβάνει παράλληλο κείμενο από το πολυγλωσσικό περιεχόμενο της Wikipedia, κατάλληλα αντιστοιχισμένο σε επίπεδο πρότασης.

Το προκύπτον σώμα κειμένου είχε μέγεθος 6.3GB και περιελάμβανε 50.451.352 ζεύγη προτάσεων, καλύπτοντας ποικίλες θεματικές περιοχές. Σε αυτό εφαρμόστηκε ρουτίνα καθαρισμού που απέρριπτε ζεύγη προτάσεων εφόσον οποιαδήποτε πρόταση ξεπερνούσε τους 1.000 χαρακτήρες (καθώς οι κωδικοποιητές του νευρωνικού μοντέλου δέχονται πεπερασμένο μήκος κειμένου εισόδου, αποκόπτοντας τα επιπλέον tokens). Το τελικό σώμα κειμένου απαρτιζόταν από 36.251.157 προτάσεις. Ακολούθησε η αναγνώριση λεκτικών μονάδων (tokenization) με χρήση του Moses tokenizer (Koehn et al., 2007) σε επίπεδο λέξης, κατά την οποία επίσης διαχωρίστηκαν τα σημεία στίξης, διατηρώντας ειδικούς χαρακτήρες (πχ. ημερομηνίες, URLs) και κανονικοποιώντας άλλους (πχ. σύμβολα όπως εισαγωγικά, Unicode χαρακτήρες, κτλ.). Τέλος έγινε διαχωρισμός του σώματος σε σύνολα δεδομένων εκπαίδευσης και επικύρωσης, έτσι ώστε 1 πρόταση ανά 23 να τοποθετείται στο σύνολο επικύρωσης, ενώ οι υπόλοιπες στο σύνολο εκπαίδευσης.

Ακολούθησε η κατασκευή του λεξιλογίου (vocabulary) για καθεμία από τις δύο γλώσσες. Τα λεξιλόγια αυτά περιλαμβάνουν όλες τις δυνατές ακολουθίες χαρακτήρων που μπορούν να

μετασχηματιστούν από τη μία γλώσσα στην άλλη με χρήση του μοντέλου μηχανικής μετάφρασης. Όπως είναι φυσικό, η χρήση ενός στατικού λεξιλογίου με αντιστοιχία 1:1 μεταξύ της γλώσσας πηγής και της γλώσσας στόχου δεν αποτελεί βέλτιστη πρακτική (ειδικά για μορφολογικά πλούσιες γλώσσες), καθώς κάθε λέξη θα πρέπει να περιλαμβάνεται πολλές φορές με όλους τους κλιτούς της τύπους, αυξάνοντας δραματικά το μέγεθος του λεξιλογίου. Όπως περιεγράφηκε στην υποενότητα 3.3.1.1.1, εφαρμόστηκε η μέθοδος ανίχνευσης λεκτικών μονάδων σε επίπεδο υπο-λέξεων byte-pair encoding (BPE), τόσο ώστε να είναι δυνατή η διαχείριση λέξεων εκτός λεξιλογίου, όσο και για να περιοριστεί το μέγεθος του εκάστοτε λεξιλογίου από την αποφυγή νέων εγγραφών για λέξεις με ίδια ρίζα και διαφορετική κατάληξη. Συγκεκριμένα υλοποιήθηκαν δύο κωδικοποιήσεις BPE με χρήση του Python πακέτου subword-nmt[12], με τις ακόλουθες παραμετροποιήσεις:

A. Μετατροπή των χαρακτήρων όλων των tokens των συνόλων εκπαίδευσης και επικύρωσης σε πεζούς για περαιτέρω μείωση του μεγέθους των λεξιλογίων χωρίς απώλειες στην απόδοση του κειμένου. Εφαρμογή της μεθόδου BPE με μέγιστο αριθμό συγχωνεύσεων ίσο με 10.000. Η παραπάνω ρύθμιση οδήγησε σε λεξιλόγιο 12.892 λέξεων για την ελληνική γλώσσα και 9.932 λέξεων για την αγγλική.
B. Απευθείας εφαρμογή της μεθόδου BPE στο σύνολο δεδομένων (πεζοί και κεφαλαίοι χαρακτήρες), με ορισμό του ορίου συγχωνεύσεων στις 20.000 πράξεις. Η παραπάνω ρύθμιση οδήγησε σε λεξιλόγιο 23.220 λέξεων για την ελληνική γλώσσα και 15.284 λέξεων για την αγγλική.

Ακολούθησε η εκπαίδευση μοντέλων μηχανικής μετάφρασης και για τις δύο παραπάνω ρυθμίσεις με χρήση της βιβλιοθήκης ανοιχτού κώδικα Fairseq (Ott et al., 2019) που παρέχεται από την Facebook AI Research για την δημιουργία sequence-to-sequence μοντέλων. Υλοποιήθηκε παραλλαγή της αρχιτεκτονικής Transformer με 4 κεφαλές αυτό-προσοχής, 6 στρώματα κωδικοποιητή και 6 στρώματα αποκωδικοποιητή διάστασης 512 και διάσταση του ενδιάμεσου υποστρώματος στο στρώμα πρόσθιας τροφοδότησης ίση με 1024.

Κατά τη διάρκεια της εκπαίδευσης εφαρμόστηκαν δύο τεχνικές κανονικοποίησης (regularization) για βελτίωση της ευρωστίας των μοντέλων:

[12] https://github.com/rsennrich/subword-nmt

1. τεχνική dropout με παράμετρο ίση με 0.3, κατά την οποία ένα ποσοστό των κόμβων απενεργοποιείται ανά εποχή, ώστε να μειωθεί η πιθανότητα υπερ-προσαρμογής (overfitting) στα δεδομένα,
2. τεχνική label smoothing με παράμετρο ίση με 0.1, κατά την οποία τροποποιούνται τα τελικά διανύσματα πρόβλεψης έτσι ώστε να ομαλοποιηθούν οι διαφορές μεταξύ των logits τους (πχ. σε ένα πρόβλημα ταξινόμησης n κλάσεων, εάν το διάνυσμα της σωστής κλάσης αντιστοιχεί στην τιμή 1 και όλων των υπολοίπων αντιστοιχεί σε 0, η σωστή κλάση αντικαθίσταται με 0.9 και οι υπόλοιπες με $0.1/n$).

Ο διαδεδομένος αλγόριθμος βελτιστοποίησης Adam (Adaptive Movement Estimation) (Kingma and Ba, 2015) επιλέχθηκε για την εκμάθηση των βαρών, καθώς επιτρέπει την προσαρμογή του ρυθμού εκμάθησης κάθε βάρους του νευρωνικού δικτύου βάσει των εκτιμήσεων των ροπών πρώτης και δεύτερης τάξης της κλίσης (gradient). Ο μέγιστος ρυθμός εκμάθησης (max learning rate) ορίστηκε ίσος με 0.0005 και ο αριθμός βημάτων προθέρμανσης (warmup steps) που επιτρέπει στο δίκτυο να ξεκινήσει με μικρό learning rate προκειμένου να προσαρμοστεί σταδιακά στο σύνολο δεδομένων ίσος με 4.000.

Κάθε μοντέλο εκπαιδεύτηκε για 5 εποχές (epochs) και ακολούθησε η επιλογή του καλύτερου checkpoint βάσει της σύγκρισης της περιπλοκής (perplexity) στο σύνολο επικύρωσης. Η περιπλοκή αποτελεί μέτρο ποσοτικής αξιολόγησης παραγωγικών γλωσσικών μοντέλων που ορίζεται ως ο γεωμετρικός μέσος (δηλαδή η ρίζα του προϊόντος ενός συνόλου αυστηρά θετικών αριθμών) των αντίστροφων κατανομών της πιθανότητας εμφάνισης μιας ακολουθίας (πρότασης) που αποτελείται από μια ακολουθία λέξεων $w_1, w_2, \ldots, w_n$:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_n)}} \tag{4.1}$$

Η διαισθητική ερμηνεία του παραπάνω μέτρου έγκειται στο ότι εάν το μοντέλο μεταφράσει με μεγάλη πιθανότητα (και άρα μικρή περιπλοκή) μια ακολουθία που υπάρχει στο σύνολο επικύρωσης, σημαίνει ότι δεν "εκπλήσσεται" από την ακολουθία λέξεων που αποτελούν την μεταφρασμένη πρόταση, κάτι που αποτελεί ένδειξη καλής κατανόησης των εγγενών κανόνων που διέπουν τη γλώσσα.

Η εκπαίδευση υποστηρίχθηκε από μια κάρτα γραφικών NVIDIA GeForce RTX2080 SUPER (8GB VRAM) και διήρκησε περίπου 20 ώρες για κάθε μοντέλο, ενώ έγινε χρήση βιβλιοθήκης μικτής ακρίβειας για αριθμητική κινητής υποδιαστολής (FP16) (Narang et al., 2018) για μείωση των απαιτήσεων σε μνήμη και για επίσπευση της διαδικασίας εκπαίδευσης.

Το σύνολο των υπερ-παραμέτρων που επιλέχθηκε κατά τη διαδικασία εκπαίδευσης συγκεντρώνεται στον παρακάτω πίνακα (τα μεγέθη των λεξιλογίων πηγής και στόχου αφορούν το μοντέλο EN-EL παραμετροποίησης B και διαφέρουν ανάλογα με την επιλεγμένη παραμετροποίηση):

Πίνακας 3 Υπερ-παράμετροι εκπαιδευμένων μοντέλων μηχανικής μετάφρασης


```
{
 "architectures": [
   "FSMTForConditionalGeneration"
 ],
 "model_type": "fsmt",
 "activation_dropout": 0.0,
 "activation_function": "relu",
 "attention_dropout": 0.0,
 "d_model": 512,
 "dropout": 0.3,
 "init_std": 0.02,
 "max_position_embeddings": 1024,
 "num_hidden_layers": 6,
 "src_vocab_size": 15288,
 "tgt_vocab_size": 23224,
 "langs": [
   "en",
   "el"
 ],
 "encoder_attention_heads": 4,
 "encoder_ffn_dim": 1024,
 "encoder_layerdrop": 0,
 "encoder_layers": 6,
 "decoder_attention_heads": 4,
 "decoder_ffn_dim": 1024,
 "decoder_layerdrop": 0,
 "decoder_layers": 6,
 "bos_token_id": 0,
 "pad_token_id": 1,
 "eos_token_id": 2,
 "is_encoder_decoder": true,
 "scale_embedding": true,
 "tie_word_embeddings": false,
 "num_beams": 5,
 "early_stopping": false,
 "length_penalty": 1.0
}
```


Συνολικά παρήχθησαν 4 μοντέλα (από ελληνικά σε αγγλικά και αντίστροφα για τις 2 προαναφερθείσες παραμετροποιήσεις A και B) και συγκεκριμένα:

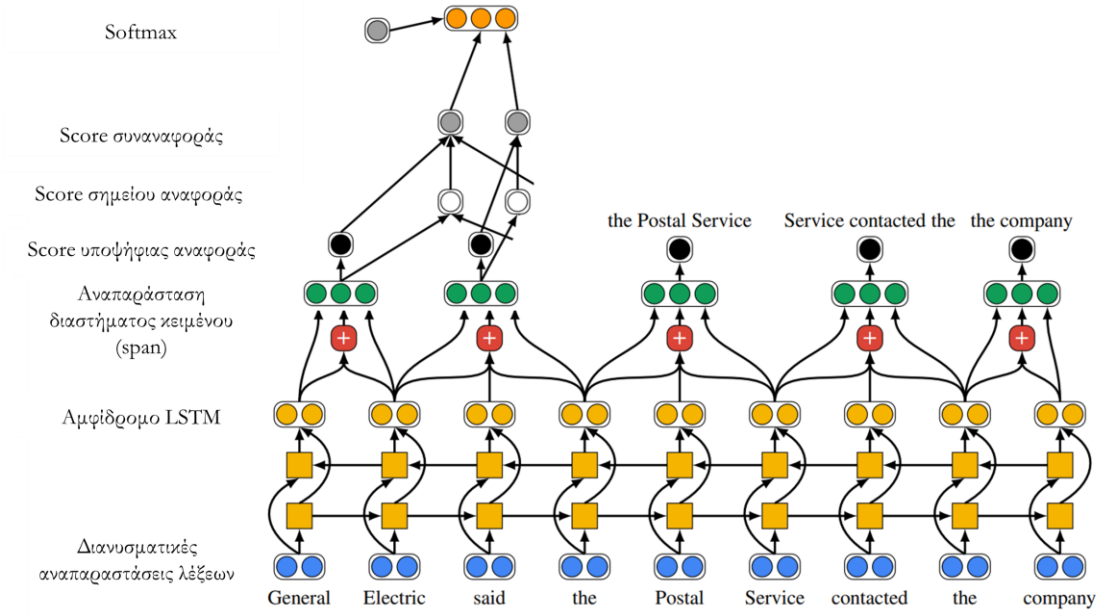
i. lower-case μοντέλο EL-to-EN παραμετροποίησης A
ii. lower-case μοντέλο EN-to-EL παραμετροποίησης A
iii. mixed-case μοντέλο EL-to-EN παραμετροποίησης B
iv. mixed-case μοντέλο EN-to-EL παραμετροποίησης B

Ακολούθησε η μετατροπή (porting) των παραπάνω μοντέλων προκειμένου αυτά να καταστούν συμβατά με τη δημοφιλή βιβλιοθήκη Transformers της Hugging Face[13] και να διατεθούν ελεύθερα μέσω του παρακάτω συνδέσμου: https://huggingface.co/lighteternal.

Αναφορικά με το στοιχείο ευθυγράμμισης κειμένου που μπορεί να αντικαταστήσει την αντίστροφη μετάφραση (back-translation) των εξαχθεισών τριπλετών στα ελληνικά, χρησιμοποιήθηκε υλοποίηση της μεθόδου Fast Align (Dyer et al., 2013) σε περιβάλλον C++[14].

4.2.1.2 Στοιχείο επίλυσης συναναφορών

Έγινε χρήση προεκπαιδευμένου νευρωνικού μοντέλου επίλυσης συναναφορών βασιζόμενο σε αμφίδρομα LSTM δίκτυα και μηχανισμό προσοχής (Lee et al., 2017a) για την αγγλική γλώσσα, που παρέχεται από το Allen Institute for Artificial Intelligence[15].



Σχήμα 28 Χρησιμοποιούμενη αρχιτεκτονική μηχανισμού επίλυσης συναναφορών

Καθώς η αρχική υλοποίηση βασίζεται σε διανυσματικές αναπαραστάσεις GloVe (Pennington et al., 2014) οι οποίες είναι ανεξάρτητες από το εκάστοτε εννοιολογικό πλαίσιο του κειμένου, έγινε παραλλαγή του αρχικού μοντέλου με αντικατάσταση των διανυσματικών αναπαραστάσεων από αντίστοιχες που προέκυψαν από προεκπαίδευση με μέθοδο

[13] https://github.com/huggingface/transformers

[14] https://github.com/clab/fast_align

[15] http://docs.allennlp.org/v0.9.0/api/allennlp.models.coreference_resolution.html

SpanBERT (Joshi et al., 2020) και εξαρτώνται από τα συμφραζόμενα. Χάρη σε αυτή την προσέγγιση, όλα τα διαστήματα του κειμένου (spans) θεωρούνται υποψήφιες συναναφορές, ενώ στόχος του μοντέλου είναι η εκμάθηση της κατανομής πιθανότητας που αφορά το κάθε διάστημα να αποτελεί συναναφορά μιας ονοματικής οντότητας του κειμένου. Το μοντέλο δέχεται το προς ανάλυση κείμενο ανά πρόταση και αντικαθιστά επι τόπου τα διαστήματα που αφορούν ένα συγκεκριμένο σημείο αναφοράς.

4.2.1.3 Στοιχείο εξαγωγικής συνόψισης

Προκειμένου να μειωθεί το μέγεθος του κειμένου που θα τροφοδοτήσει το μηχανισμό ανοιχτής εξαγωγής πληροφοριών, επιλέχθηκε προεκπαιδευμένο μοντέλο εξαγωγικής συνόψισης για την αγγλική γλώσσα, που βασίζεται σε αρχιτεκτονική encoder για την παραγωγή διανυσματικών αναπαραστάσεων σε επίπεδο πρότασης και χρησιμοποιεί συσταδοποίηση k-means για τον εντοπισμό των πιο αντιπροσωπευτικών από αυτές ώστε να συμπεριληφθούν στην τελική σύνοψη (Miller, 2019). Η επιλεχθείσα υλοποίηση βασίζεται στη βιβλιοθήκη Transformers της Hugging Face και είναι διαθέσιμη ως πακέτο της Python[16].

Η χρήση του συγκεκριμένου στοιχείου στον μηχανισμό εξόρυξης πληροφοριών είναι προαιρετική, καθώς πρέπει να ληφθεί υπόψη η αντιστάθμιση μεταξύ πιθανής απώλειας πληροφορίας και ταχύτερης εξαγωγής τριπλετών από το συνοψισμένο κείμενο. Κατά την παραμετροποίηση, το ποσοστό προτάσεων που περιλαμβάνονται στη σύνοψη ως προς το αρχικό σύνολο ορίστηκε ίσο με 0.2. Σημειώνεται επίσης ότι η επιλογή προεκπαιδευμένου μοντέλου (κωδικοποιητή) διανυσματικών αναπαραστάσεων μπορεί να γίνει κατά περίπτωση, ανάλογα με το εννοιολογικό πλαίσιο του εκάστοτε κειμένου. Για την επεξεργασία άρθρων από διαδικτυακές πηγές έγινε χρήση παραλλαγής του κωδικοποιητή MinLM (Wang et al., 2020) που παρέχεται από τη Hugging Face[17].

4.2.1.4 Στοιχεία ανοιχτής εξαγωγής πληροφοριών

Ενσωματώνονται 3 στοιχεία ανοιχτής εξαγωγής πληροφοριών (OIE) που βασίζονται σε υπολογιστικές μεθόδους εξαγωγής (πχ. μηχανική μάθηση) σε συνδυασμό με ένα στοιχείο βασισμένο σε γλωσσολογικούς κανόνες. Καθώς η εξαγωγή τριπλετών {subject, predicate, object} γίνεται σε επίπεδο πρότασης, προηγήθηκε η χρήση NLTK tokenizer[18] σε Python

[16] https://pypi.org/project/bert-extractive-summarizer/

[17] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[18] https://www.nltk.org/api/nltk.tokenize.html

για χωρισμό του κειμένου σε επιμέρους προτάσεις. Παρακάτω παρέχεται μια σύνοψη των τεχνικών λεπτομερειών που αφορούν το κάθε στοιχείο:

- Το OpenIE 5.1[19] από το University of Washington και το IIT Delhi (Mausam, 2016) αποτελεί διάδοχο του OIE συστήματος Ollie. Η παρούσα έκδοση συνδυάζει 4 διαφορετικές μηχανές εξαγωγής και συγκεκριμένα την CALMIE (Saha and Mausam, 2018) που εξειδικεύεται σε επαυξημένες προτάσεις, τη RelNoun (Pal, 2016) για εξαγωγές που αφορούν σχέσεις μεταξύ ουσιαστικών, τη BONIE (Saha et al., 2017) για προτάσεις που περιέχουν αριθμητικές πληροφορίες και τη SRLIE (Christensen et al., 2011) που βασίζεται σε ετικέτες σημασιολογικών ρόλων. Το στοιχείο OpenIE διατίθεται είτε ως προ-μεταγλωττισμένο .jar αρχείο, είτε μέσω Python wrapper, ο οποίος χρησιμοποιήθηκε στα πλαίσια της διατριβής.
- Το ClausIE[20] από το Max Planck Institute (Del Corro and Gemulla, 2013a) ελέγχει κάθε πρόταση για την ύπαρξη προκαθορισμένων συντακτικών προτύπων και την αναπαριστά σε δομημένη μορφή ως σύνολο μιας ή περισσότερων εξαγωγών. Το συγκεκριμένο στοιχείο ειδικεύεται στην ανάλυση εμφωλευμένων/σύνθετων προτάσεων και διατίθεται ως Python wrapper[21].
- Το OIE στοιχείο της AllenNLP (Stanovsky et al., 2018) μοντελοποιεί την εξαγωγή τριπλετών ως ένα πρόβλημα επιβλεπόμενης μηχανικής μάθησης και βασίζεται στη χρήση αμφίδρομου LSTM δικτύου και διανυσματικών αναπαραστάσεων λέξεων για την ανάθεση ετικετών BIO σε λεκτικά σύνολα που αντιστοιχούν σε υποκείμενα ή αντικείμενα ενός κατηγορήματος. Έχει τη δυνατότητα ανίχνευσης πιο πολύπλοκων συσχετίσεων μεταξύ λεκτικών συνόλων που δεν είναι δυνατόν να εντοπιστούν με τα παραπάνω στοιχεία. Διατίθεται ως Python βιβλιοθήκη από τον ίδιο οργανισμό[22].
- Το στοιχείο εξαγωγής που βασίζεται σε γλωσσολογικούς κανόνες αναπτύχθηκε σε περιβάλλον Python (Smith et al., 2022) και αξιοποιεί τον Stanford dependency parser[23] για την αναγνώριση των γραμματικών και συντακτικών ιδιοτήτων κάθε λέξης

[19] https://github.com/dair-iitd/OpenIE-standalone

[20] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/clausie

[21] https://github.com/AnthonyMRios/pyclausie

[22] https://demo.allennlp.org/open-information-extraction

[23] https://nlp.stanford.edu/software/lex-parser.shtml

(πχ. μέρη του λόγου) που περιλαμβάνεται σε μια πρόταση. Στη συνέχεια, αξιοποιεί το συντακτικό δένδρο της πρότασης για την επισήμανση συσχετίσεων σε μορφή τριπλέτας, κάνοντας χρήση 40 γραμματικοσυντακτικών κανόνων που καλύπτουν τα πιο συνήθη γλωσσολογικά φαινόμενα. Καθώς η λειτουργία του διέπεται από κανόνες, η χρήση του στοχεύει στη μείωση των ψευδώς θετικών αποτελεσμάτων και στην βελτίωση της συνδυαστικής ακρίβειας (precision) του συστήματος.

Ο συνδυασμός των τριπλετών που παράγονται από τα παραπάνω στοιχεία και περιγράφεται στην υποενότητα 3.3.1.2.3 υλοποιήθηκε σε περιβάλλον Python, και χαρακτηρίζεται από τις ίδιες τεχνικές απαιτήσεις που διέπουν το στοιχείο εξαγωγής βασισμένο σε γλωσσολογικούς κανόνες.

4.2.2 Εφαρμογές και αποτελέσματα

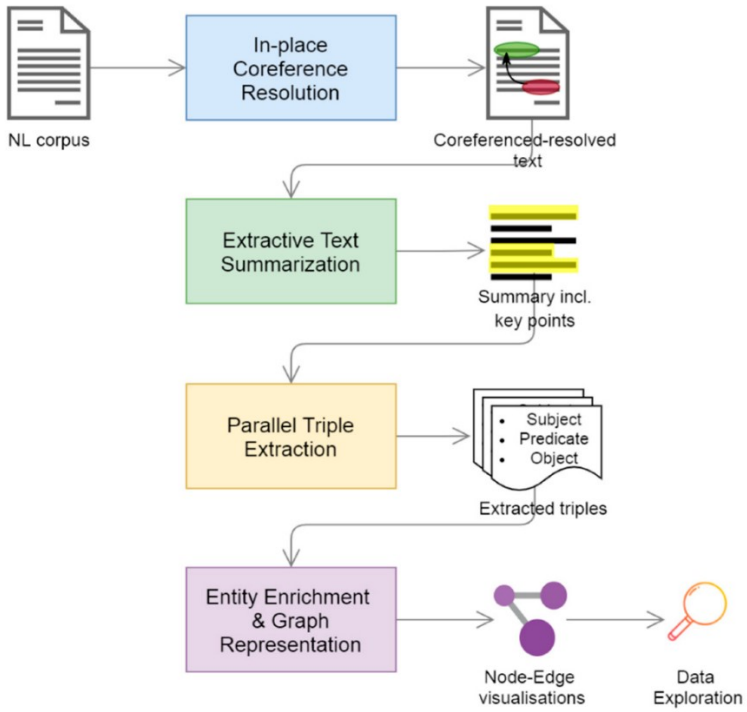
Στις ακόλουθες υποενότητες παρουσιάζονται περιπτώσεις χρήσης του μηχανισμού εξόρυξης πληροφοριών και παρατίθενται σχετικά αποτελέσματα όσον αφορά τη δοκιμή και συγκριτική μελέτη απόδοσης τόσο του συνολικού μηχανισμού όσο και των επιμέρους στοιχείων του. Δεδομένου ότι ο μηχανισμός εξαγωγής υποστηρίζεται από στοιχείο μηχανικής μετάφρασης, οι περιπτώσεις χρήσης που περιγράφονται παρακάτω αφορούν κείμενα τόσο στα αγγλικά όσο και στα ελληνικά. Η χρήση του μηχανισμού σε αγγλικά σώματα κειμένων επέτρεψε την ποσοτική αξιολόγηση της επίδοσής του μέσω υφιστάμενων benchmarks που δεν υπάρχουν σε γλώσσες χαμηλότερων πόρων, όπως η ελληνική.

Αξίζει να σημειωθεί ότι η συνδυαστική προσέγγιση που αναπτύχθηκε στα πλαίσια της εργασίας οδήγησε στην ανάπτυξη πρωτοποριακού (state-of-the-art) εργαλείου ανοιχτής εξαγωγής πληροφοριών (OIE) που υπερτερεί σε σχέση με αντίστοιχα συστήματα σε αγγλικά σύνολα αναφοράς (benchmarks), καλύπτοντας παράλληλα το αντίστοιχο κενό για την ελληνική γλώσσα.

4.2.2.1 Περίπτωση χρήσης 1: Εξαγωγή πληροφοριών από το σώμα κειμένων CORD-19

Μια πρώιμη εκδοχή του μηχανισμού εξόρυξης χρησιμοποιήθηκε για την εξαγωγή πληροφοριών από αγγλικό ελεύθερο κείμενο, αξιοποιώντας το σώμα κειμένων COVID-19 Open Research Dataset (CORD-19) που παρέχεται ελεύθερα από το Allen Institute of AI (Kohlmeier et al., 2020). Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει περισσότερες από 40.000 επιστημονικές δημοσιεύσεις που αφορούν την έρευνα του κορονοϊού και διατίθενται τόσο για εμπορική όσο και για μη-εμπορική χρήση. Η συγκεκριμένη περίπτωση χρήσης εστίασε στο υποσύνολο που διατίθεται ελεύθερα για εμπορική χρήση και κατά την

ημερομηνία πρόσβασης (8 Απριλίου 2020) απαρτιζόταν από 9.365 δημοσιεύσεις. Το κείμενο κάθε δημοσίευσης αναπαρίσταται ως μεμονωμένο JSON αρχείο, του οποίου η δομή περιλαμβάνει ένα μοναδικό αναγνωριστικό (ID), τον τίτλο της δημοσίευσης, τη λίστα των συγγραφέων, την περίληψη του κειμένου, το κυρίως σώμα και τις βιβλιογραφικές παραπομπές.



Σχήμα 29 Αλληλουχία βημάτων για την εξόρυξη δεδομένων κατά την περίπτωση χρήσης στο σώμα κειμένων CORD-19

Στόχος της συγκεκριμένης εργασίας (Papadopoulos et al., 2020) ήταν η ανάδειξη της χρησιμότητας μιας μεθοδολογίας εξόρυξης δεδομένων που εστιάζει στην αποδοτική εξαγωγή αξιοποιήσιμων πληροφοριών από το σώμα κειμένων, στον εμπλουτισμό τους μέσω αναγνώρισης ονοματικών οντοτήτων και στην αναπαράστασή τους με χρήση γνωσιακών γράφων. Η αλληλουχία βημάτων που ακολουθήθηκε απαρτιζόταν από δύο στάδια προεπεξεργασίας (επίλυση συναναφορών και εξαγωγική συνόψιση), ένα στάδιο παράλληλης εξαγωγής πληροφοριών και ένα στάδιο μετα-επεξεργασίας που περιλαμβάνει τον εμπλουτισμό των παραγόμενων τριπλετών μέσω της σύνδεσής τους με σχετική οντολογία, τον καθαρισμό διπλοτύπων και την αναπαράστασή τους με τη βοήθεια γράφου, όπως φαίνεται στο Σχήμα 29.

Κατά το στάδιο επίλυσης συναναφορών έγινε χρήση του στοιχείου που περιεγράφηκε στην υποενότητα 4.2.1.2, οδηγώντας σε επι τόπου αντικατάσταση των ανιχνευθεισών αναφορών με το σημείο αναφοράς τους. Ενδεικτικά παραδείγματα από τη χρήση του στοιχείου

απεικονίζονται παρακάτω (Πίνακας 4). Τα ζεύγη στοιχείων που αποτελούν φαινόμενα συναναφοράς επισημαίνονται με κοινό δείκτη για κάθε πρόταση, εκ των οποίων με πράσινο απεικονίζεται το σημείο αναφοράς και με πορτοκαλί η αναφορά του.

Πίνακας 4 Παράδειγμα επίλυσης συναναφορών για περίπτωση χρήσης στο σώμα κειμένων CORD-19

Article ID	Extract
42e321eedb a25d380ae4 4d16cdf0bb deab83d665	Two articles in the top ten cited articles discussed the emergence of New Delhi metallo-β-lactamase (NDM) gene{1} responsible for carbapenem resistance. ~~This gene~~{1} metallo-β-lactamase (NDM) gene{1} belongs to carbapenemase gene family and bacteria carrying metallo-β-lactamase (NDM) gene{2} are referred to as superbugs because ~~they~~{2} bacteria carrying metallo-β-lactamase (NDM) gene{2} are resistant to most antibiotics.
85eb641e06 b0d6b1a0b2 02275add0c 5d27e53d71	Official health linkages{1} have served to promote good will in some otherwise difficult relationships, as has been the case with Indonesia. ~~They~~{1} Official health linkages{1} have also helped to promote a positive international image for Australia.
4c84dbfd01 f7b2009ebe d54376da8 afcbcf1ec64	However, one should also note that the experiment is based on labelling and quantifying proteins about 4 h post-infection. This relatively early time point{1} allows one to minimize potentially confounding influences of virion particle assembly and production on cytoplasmic levels of viral proteins, but~~ it~~{1}This relatively early time point{1} also represents a point before the majority of viral proteins have been made.
2c5d1ebec4 04ad8061eb 81e94effbe5 2a6dbe809	Purified ALV-A virus particles{1} were incubated with PMB for 30 min at 378C, and infectivity was measured on human 293 cells expressing the ALV receptor Tva (293-Tva){2}. The effect of PMB treatment of ~~these particles~~{1} Purified ALV-A virus particles{1} was comparable to native MLV particles. These findings suggest that ~~Tva~~{2} the ALV receptor Tva (293-Tva){2} binding creates or exposes a functionally important cysteine thiolate target for PMB in ALV-A Env.
d9eb8ffffee8 147c850b00f 613a1978c18 505580	FCoVs and CCoVs{1} are common pathogens and readily evolve. It is necessary to pursue epidemiological surveillance of ~~these viruses~~{2} FCoVs and CCoVs{1}, so as to detect the emergence of new variants, which may have increased pathogenicity and/or a new host range, as early as possible.

Κατά το επόμενο στάδιο επεξεργασίας υλοποιήθηκε εξαγωγική συνόψιση στα τροποποιημένα κείμενα, προκειμένου να απομονωθεί το αντιπροσωπευτικότερο τμήμα τους που θα οδηγήσει σε όσο το δυνατόν πιο υπολογιστικά διαχειρίσιμη εξαγωγή, παράγοντας τριπλέτες με χρήσιμη πληροφορία. Έγινε χρήση του στοιχείου που περιεγράφηκε στην ενότητα 4.2.1.3, με αποτέλεσμα την παραγωγή περιλήψεων για κάθε άρθρο με μέσο αριθμό λέξεων ίσο με 843, σημαντικά μικρότερο από το μέσο αριθμό λέξεων των αρχικών δημοσιεύσεων που ήταν ίσος με 4.482. Στον επόμενο πίνακα (Πίνακας 5) παρατίθεται παράδειγμα εξαγωγικής συνόψισης σε απόσπασμα άρθρου, όπου με επισημαίνονται με κίτρινο τα μέρη που απαρτίζουν την περίληψή του.

Πίνακας 5 Παράδειγμα εξαγωγικής συνόψισης για την περίπτωση χρήσης στο σώμα κειμένων CORD-19

CORD-19 Article ID: 85783a36e7e787302307f42460839435d665f4e7
Article Title: SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat

Article body: […] Subsequently, coronaviruses with high similarity to the human SARS-CoV or civet SARS-CoV-like virus were isolated from **horseshoe bats**{1}, concluding ~~**the bats**{1}~~ **horseshoe bats**{1} as the potential natural reservoir of SARS-CoV whereas masked palm civets are the intermediate host [53] [54] [55] [56] .It is thus reasonable to suspect that bat is the natural host of SARS-CoV-2 considering its similarity with SARS-CoV. The phylogenetic analysis of SARS-CoV-2 against a collection of coronavirus sequences from various sources found that SARS-CoV-2 belonged to the Betacoronavirus genera and was closer to SARS-like coronavirus in bat [19] . By analyzing genome sequence of SARS-CoV-2, it was found that SARS-CoV-2 felled within the subgenus Sarbecovirus of the genus Betacoronavirus and was closely related to two bat-derived SARS-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21, but were relatively distant from SARS-CoV [15, 18, [57] [58] [59] . Meanwhile, Zhou and colleagues showed that SARS-CoV-2 had 96.2% overall genome sequence identity throughout the genome to BatCoV RaTG13, a bat coronavirus detected in Rhinolophus affinis from Yunnan province [14] . Furthermore, the phylogenetic analysis of full-length genome, the receptor binding protein spike (S) gene, and RNA-dependent RNA polymerase (RdRp) gene respectively all demonstrated that RaTG13 was the closest relative of the SARS-CoV-2 [14] . However, despite SARS-CoV-2 showed high similarity to coronavirus from bat, SARS-CoV-2 changed topological position within the subgenus Sarbecovirus when different gene was used for phylogenetic analysis: SARS-CoV-2 was closer to bat-SL-CoVZC45 in the S gene phylogeny but felled in a basal position within the subgenus Sarbecovirus in the ORF1b tree [57] . This finding implies a possible recombination event in this group of viruses. Of note, the receptor-binding domain of SARS-CoV-2 demonstrates a similar structure to that of SARS-CoV by homology modelling but a few variations in the key residues exist at amino acid level [15, 19] .Despite current evidences are pointing to the evolutional origin of SARS-CoV-2 from bat virus [15, 57] , an intermediate host between bats and human might exist. Lu et. al. raised four reasons for such speculation [15] : First, most bat species in Wuhan are hibernating in late December; Second, no bats in Huanan Seafood market were sold or found; Third, the sequence identity between SARS-CoV-2 and bat-SL-CoVZC45 or bat-SL-CoVZXC21, the closest relatives in their analyses, is lower than 90%; Fourth, there is an intermediate host for other humaninfecting coronaviruses that origin from bat. For example, masked palm civet and dromedary camels are the intermediate hosts for SARS-CoV [49] and MERS-CoV respectively [60] . A study of the relative synonymous codon usage (RSCU) found that SARS-CoV-2, bat-SL-CoVZC45, and snakes had similar synonymous codon usage bias, and speculated that snake might be the intermediate host [61] . However, no SARS-CoV-2 has been isolated from snake yet. Pangolin was later found to be a potential intermediate host for SARS-CoV-2. The analysis of samples from Malytan pangolins obtained during anti-smuggling operations from Guangdong and Guangxi Customs of China respectively found novel coronaviruses representing two sub-lineages related to SARS-CoV-2 [62] . The similarity of SARS-CoV-2 to these identified coronaviruses from pangolins is approximately 85.5% to 92.4% in genomes, lower than that to the bat coronavirus RaTG13 (96.2%) [14, 62]. However, the receptor-binding domain of S protein from one sub-lineage of the pangolin coronaviruses shows 97.4% similarity in amino acid sequences to that of SARS-CoV-2, even higher than that to RaTG13 (89.2%) [62]. Interestingly, the pangolin coronavirus and SARS-CoV-2 share identical amino acids at the five critical residues of RBD of S protein, while RaTG13 only possesses one [62] . The discovery of coronavirus close to SARS-CoV-2 from pangolin suggests that pangolin is a potential intermediate host. [..] An EC90 (6.90 μM) against the SARS-CoV-2 in Vero E6 cells is clinically achievable in vivo according to a previous clinical trial [66] . Remdesivir is a drug currently under the development for Ebola virus infection and is effective to a broad range of viruses including SARS-CoV and MERS-CoV [67, 68] . Functioning as an adenosine analogue targeting RdRp, Remdesivir can result in premature termination during the virus transcription [69, 70] . The EC90 of remdesivir against SARS-CoV-2 in Vero E6 cells is 1.76 μM, which is achievable in vivo based on a trial in nonhuman primate experiment [63, 69] [...]

Το τρίτο στάδιο επεξεργασίας περιελάμβανε χρήση του μηχανισμού παράλληλης ανοιχτής εξαγωγής πληροφοριών για την παραγωγή τριπλετών της μορφής {subject, predicate, object}. Αξιοποιήθηκαν τα τρία στοιχεία εξαγωγής που βασίζονται σε υπολογιστικές μεθόδους εξαγωγής και περιγράφονται στην υποενότητα 4.2.1.4. Οι τριπλέτες που εξήχθησαν από τη σύνοψη του κειμένου που φαίνεται στον προηγούμενο πίνακα (Πίνακας 5), παρατίθενται παρακάτω (Πίνακας 6).

Πίνακας 6 Εξαχθείσες τριπλέτες και ανιχνευθείσες οντότητες από σύνοψη για περίπτωση χρήσης σώματος κειμένων CORD-19

Extracted triples using our pipeline:				
subject	**predicate**	**object**	**subj_entity_name**	**obj_entity_name**
coronaviruses with high	were isolated	from horseshoe	('coronaviruses', 'high',	('horseshoe')

similarity to the human SARS - CoV or civet SARS - CoV - like virus		bats	'human', 'SARS-CoV', 'civet', 'SARS-CoV-like')	
horseshoe bats as the potential natural reservoir of SARS-CoV	is	the natural host of SARS-CoV-2	('horseshoe', 'natural', 'SARS-CoV')	('natural', 'SARS-CoV-2')
masked palm civets	are	the intermediate host	('palm')	('intermediate', 'host')
the receptor-binding domain of SARS-CoV-2	demonstrates	a similar structure to that of SARS-CoV	('receptor-binding', 'SARS-CoV-2')	('structure', 'SARS-CoV')
a few variations in the key residues	exist	at amino acid level	('variations', 'residues')	('amino')
Pangolin	to be	a potential intermediate host for SARS-CoV-2	('pangolin')	('potential', 'intermediate', 'host', 'SARS-CoV-2')
The EC90 of remdesivir against SARS - CoV-2 in Vero E6 cells	is	1.76 BμM , which is achievable in vivo	('EC90', 'remdesivir', 'SARS', 'CoV-2')	('in')

Δεδομένου ότι κάθε πρόταση υποβλήθηκε σε τρείς παράλληλες εργασίες εξαγωγής, υπήρξαν περιπτώσεις που η ίδια τριπλέτα ήταν προϊόν εξαγωγής περισσότερων από ενός στοιχείου. Ο Πίνακας 7 απεικονίζει ένα αντίστοιχο παράδειγμα παράλληλης εξαγωγής. Στην στήλη "Engine", υποδεικνύονται τα στοιχεία που εξήγαγαν την συγκεκριμένη τριπλέτα, όπου O: OpenIE, C: ClausIE και A: AllenNLP OIE.

Πίνακας 7 Παράλληλη εξαγωγή τριπλετών για περίπτωση χρήσης σώματος κειμένων CORD-19

Sentence	Extracted Triples (S/P/O)	Engine
"CRP is an acute phase protein that has been linked to the presence and severity of bacterial infection in numerous studies during the past 2 decades [9, 34, 35]", Source: b8e9c45dda9cb8c9c4321a55704ab2a66fb34f7d	CRP / is / an acute phase protein	O
	an acute phase protein / has been linked / to the presence and severity of bacterial infection in numerous studies during the past 2 decades	O, C
	an acute phase protein / has been linked / to the presence and severity of bacterial infection in numerous studies	O
	CRP / is / an acute phase protein that has been linked to the severity of bacterial infection in numerous studies during the past 2 decades	C, A

Το στάδιο μετα-επεξεργασίας ολοκληρώνει την προτεινόμενη μεθοδολογία, ξεκινώντας με την αναγνώριση ονοματικών οντοτήτων και της σύνδεσή τους με στοιχεία της οντολογίας UMLS (Unified Medical Language System) (Bodenreider, 2004), η οποία περιλαμβάνει πάνω από τέσσερα εκατομμύρια οντότητες από κλινικές και βιολογικές γνωσιακές βάσεις

(πχ. CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT) διευκολύνοντας την ψηφιοποίηση αρχείων υγείας (Kormilitzin et al., 2020). Για την αυτοματοποίηση της διαδικασίας χρησιμοποιήθηκε το πακέτο αναγνώρισης οντοτήτων SciSpacy (Neumann et al., 2019) σε υλοποίηση Python, το οποίο πραγματοποίησε αναζήτηση βασιζόμενη σε επικάλυψη συμβολοσειρών (3-grams) συγκρίνοντας τις υποψήφιες ονοματικές οντότητες του κειμένου με αντίστοιχες της οντολογίας UMLS. Συνολικά, αυτή η διαδικασία όχι μόνο αυξάνει την χρησιμότητα των εξαγόμενων τριπλετών, αλλά επιτρέπει την αυτοματοποιημένη άντληση πληροφορίας από ελεύθερο κείμενο μέσω της χαρτογράφησης των σχέσεων μεταξύ των οντοτήτων (και των λεκτικών τους παραλλαγών) που περιέχονται σε αυτό. Ένα αντίστοιχο παράδειγμα σύνδεσης οντοτήτων φαίνεται στον παρακάτω πίνακα (Πίνακας 8). Γίνεται εύκολα κατανοητό ότι η επιλογή της οντολογίας που θα χρησιμοποιηθεί ως βάση πληροφορίας για τη διαδικασία εμπλουτισμού εξαρτάται από την εκάστοτε περίπτωση χρήσης.

Πίνακας 8 Παράδειγμα σύνδεσης οντοτήτων με οντολογία UMLS για περίπτωση χρήσης σώματος κειμένων CORD-19

Subject	Predicate	Object	UMLS in Subj.	UMLS in Obj.
the SARS-CoV-2	might be imported	to the seafood market in a short time	('C5203676')	('C0206208', 'C4526594')
the mortality rate due to 2019-nCoV is comparatively lesser than the earlier outbreaks of SARS and MERS-CoVs, as well as this virus	presents	relatively mild manifestations	('C0026565', 'C5203676', 'C0012652', 'C1175743')	('C1513302')
the initial identification of 2019-NCoV from 7 patients	diagnosed	with unidentified viral pneumonia	('C0020792', 'C5203676', 'C0030705')	('C0032310')
2019-nCoV cases	be detected	outside China	('C5203676', 'C0868928')	('C0008115')
the receptor binding domain of SARS-CoV-2	was	capable of binding ACE2 in the context of the SARS-CoV spike protein	('C0597358', 'C5203676')	('C1167622', 'C1422064', 'C1175743')

Ακολούθησε η υλοποίηση μιας ρουτίνας καθαρισμού των διπλότυπων και λιγότερο χρήσιμων τριπλετών, σύμφωνα με την οποία μόνο οι μοναδικές και "πλήρως συνδεδεμένες" τριπλέτες (δηλαδή τριπλέτες των οποίων τόσο το υποκείμενο όσο και το αντικείμενο περιλαμβάνουν αναφορά σε τουλάχιστον μία UMLS οντότητα) παρέμειναν στην τελική λίστα εξαγωγής. Προέκυψαν συνολικά 411.189 πλήρως συνδεδεμένες τριπλέτες από το

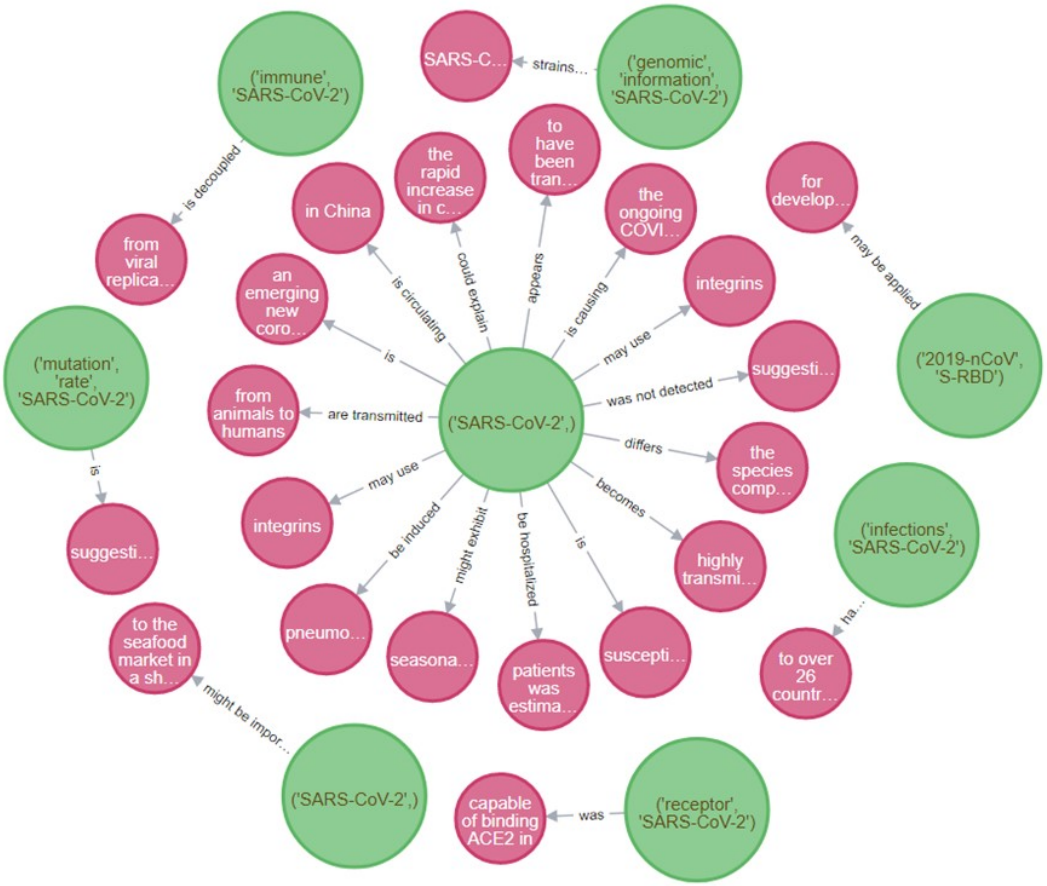
αρχικό σώμα κειμένων, οι οποίες διατίθενται ελεύθερα[24], είτε για εφαρμογές τελικού χρήστη είτε για περαιτέρω επεξεργασία.

Προκειμένου να διευκολυνθούν εργασίες διερευνητικής ανάλυσης των αποτελεσμάτων της περιγραφείσας μεθοδολογίας, τα παραπάνω αποτελέσματα αναπαραστάθηκαν με τη χρήση γράφου, χρησιμοποιώντας την ελεύθερη έκδοση της κατανεμημένης βάσης δεδομένων γράφων Neo4j[25]. Χάρη στην πλούσια εσωτερική δομή που χαρακτηρίζει τα μοντέλα γράφων ιδιοτήτων με ετικέτες (labeled property graphs), κάθε κόμβος του γράφου μπορεί να περιλαμβάνει παραπάνω από μία ιδιότητες. Ακόμη, χάρη στην εκφραστικότητα της γλώσσας Cypher που χρησιμοποιείται για την υποβολή ερωτημάτων σε Neo4J βάσεις, είναι δυνατή η απευθείας αναζήτηση σχέσεων μεταξύ οντοτήτων UMLS που ανακαλύφθηκαν από το προηγούμενο στάδιο εξαγωγής πληροφοριών.

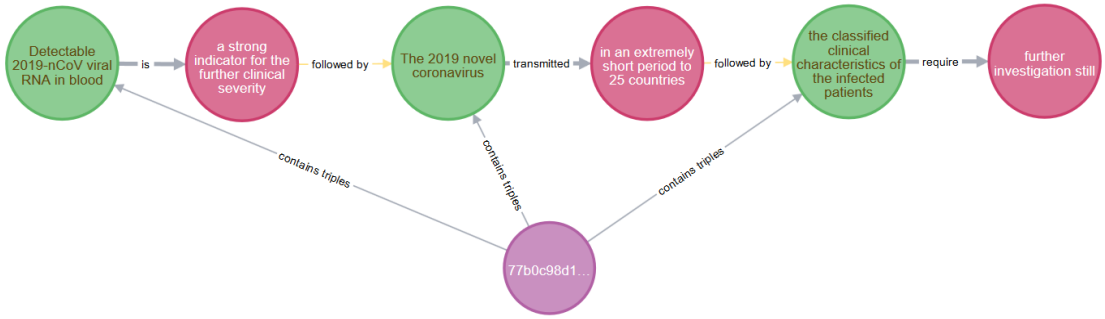
Κατά την αναπαράσταση, κάθε σώμα κειμένου (Corpus) που υποβλήθηκε προς ανάλυση απεικονίζεται με μωβ χρώμα και περιλαμβάνει όλες τις ιδιότητες (πχ. τίτλος, σώμα κειμένου κτλ.) του συνόλου δεδομένων CORD-19. Οι εξαχθείσες τριπλέτες απεικονίζονται με δύο κόμβους διαφορετικού χρώματος που συμβολίζουν το υποκείμενο (Subject, πράσινο χρώμα) και το αντικείμενο (Object, κόκκινο χρώμα) οι οποίοι ενώνονται με μία ή περισσότερες ακμές, καθεμία εκ των οποίων αντιστοιχεί στα κατηγορήματα (Predicate) που συσχετίζουν το εκάστοτε υποκείμενο και αντικείμενο. Οι κόμβοι υποκειμένων και αντικειμένων έχουν ως ιδιότητες τόσο το ελεύθερο κείμενο από τη διαδικασία εξαγωγής, όσο και τις αντιστοιχισμένες UMLS οντότητες που ανιχνεύθηκαν στο κείμενο αυτό, συνοδευόμενες από το βαθμό εμπιστοσύνης της εκάστοτε αντιστοίχισης. Ενδεικτικά παραδείγματα απεικόνισης παρατίθενται στις παρακάτω εικόνες. Στο Σχήμα 30 φαίνονται τα αποτελέσματα αναζήτησης (μεσω Cypher) των τριπλετών που εξορύχθηκαν από τη βιβλιογραφία (CORD-19) και αφορούν τον ιό SARS-CoV-2, χρησιμοποιώντας ως παράμετρο αναζήτησης το μοναδικό UMLS αναγνωριστικό του (C5203676). Στο Σχήμα 31 απεικονίζεται η αλληλουχία όλων των εξαγωγών που αφορούν ένα συγκεκριμένο απόσπασμα κειμένου (επιστημονική δημοσίευση) σε δομημένη μορφή, προσομοιάζοντας ουσιαστικά μια γραφική σύνοψη του άρθρου που αποτελείται από εναλλασόμενα υποκείμενα και αντικείμενα, όπως αυτά εμφανίζονται στο ελεύθερο κείμενο.

[24] https://github.com/lighteternal/CORD-19-OIE-triple-extraction

[25] https://neo4j.com/



Σχήμα 30 Αποτελέσματα αναζήτησης για τριπλέτες που αφορούν την οντότητα SARS-CoV-2



Σχήμα 31 Αναπαράσταση αλληλουχίας εξαχθεισών τριπλετών από απόσπασμα κειμένου

Δεδομένου ότι τα παραπάνω αποτελούν στοιχεία ποιοτικής αξιολόγησης της χρησιμότητας της παραπάνω μεθοδολογίας, στη συνέχεια παρατίθενται επίσης αντίστοιχα αποτελέσματα που εστιάζουν στην ποσοτική αξιολόγηση του μηχανισμού παράλληλης εξαγωγής πληροφοριών. Επισημαίνεται ότι η ποσοτική αξιολόγηση OIE συστημάτων αποτελεί συνήθως περίπλοκη διαδικασία για δύο λόγους: Αφενός είναι δύσκολη η εύρεση επισημασμένων εξαγωγών από εμπειρογνώμονες της μορφής {υποκείμενο, κατηγόρημα,

αντικείμενο} για κάθε περίπτωση χρήσης ώστε να χρησιμοποιηθούν ως σύνολο αναφοράς. Αφετέρου, ακόμα και σε περιπτώσεις που τα αντίστοιχα σύνολα αναφοράς υπάρχουν, δεν αποτελούν πάντα προϊόν συμφωνίας μεταξύ των εμπειρογνωμόνων (Yuan and Yu, 2018). Η συνηθέστερη προσέγγιση για αντιμετώπιση του παραπάνω προβλήματος περιλαμβάνει τη χειρωνακτική επισήμανση ενός μικρού υποσυνόλου του σώματος κειμένου, με βάση την οποία θα γίνει αξιολόγηση ενός συστήματος εξαγωγής πληροφοριών. Στα πλαίσια της εργασίας, επισημάνθηκε ένα υποσύνολο 50 τυχαίων προτάσεων (κάθε πρόταση ενδέχεται να οδηγήσει σε πολλαπλές εξαγωγές, όπως φαίνεται παρακάτω (Πίνακας 7) και ακολούθησε η αξιολόγησή τους ως προς τρία διαφορετικά μέτρα:

- την ακρίβεια (precision) που αφορά το ποσοστό έγκυρων εξαγωγών (τριπλέτες που υπάρχουν επίσης στο σύνολο αναφοράς) έναντι του συνόλου των τριπλετών που εξήχθησαν από το OIE σύστημα,
- την ανάκληση (recall) που μετρά το ποσοστό έγκυρων εξαγωγών ως προς το συνολικό αριθμό επισημασμένων τριπλετών στο σύνολο αναφοράς,
- και το F1-score που αποτελεί τον αρμονικό μέσο ακρίβειας και ανάκλησης, επιτρέποντας την μεταξύ τους αντιστάθμιση.

Τα σχετικά αποτελέσματα παρατίθενται στον επόμενο πίνακα (Πίνακας 9):

Πίνακας 9 Ποσοτική αξιολόγηση OIE συστήματος σε υποσύνολο 50 προτάσεων από το σώμα κειμένων CORD-19

Metric	Value
Precision	0.78
Recall	0.76
F1-score	0.77

Παρότι η άμεση σύγκριση των παραπάνω αποτελεσμάτων με άλλα συστήματα δεν είναι δυνατή καθώς θα απαιτούσε την εφαρμογή τους στο ίδιο υποσύνολο αναφοράς, γίνεται εύκολα κατανοητό ότι η συμπληρωματικότητα των τριών στοιχείων που απαρτίζουν το μηχανισμό παράλληλης εξαγωγής πληροφοριών οδηγεί σε καλύτερη απόδοση έναντι μεμονωμένων στοιχείων. Αυτό μπορεί να συναχθεί επίσης και από τα πειραματικά αποτελέσματα που προέκυψαν από τις συγκριτικές αξιολογήσεις OIE συστημάτων σε διαφορετικά σύνολα δεδομένων, όπου προκύπτει πως τα παραπάνω μέτρα ξεπερνούν οριακά το κατώφλι του 0.7 για μεγάλο αριθμό εξαγωγών (Del Corro and Gemulla, 2013a; Davis et al., 2019).

Η περιγραφείσα περίπτωση χρήσης εστίασε στην παρουσίαση των δυνατοτήτων μιας πρώιμης έκδοσης του μηχανισμού εξόρυξης πληροφοριών και έδωσε έμφαση στην ποιοτική αξιολόγησή του. Οι περιπτώσεις χρήσης που ακολουθούν συνοδεύονται από την ποσοτική αξιολόγηση παραλλαγών του μηχανισμού εξόρυξης σε υφιστάμενα benchmarks, προκειμένου να εξαχθούν ασφαλή συμπεράσματα αναφορικά με την απόδοσή του.

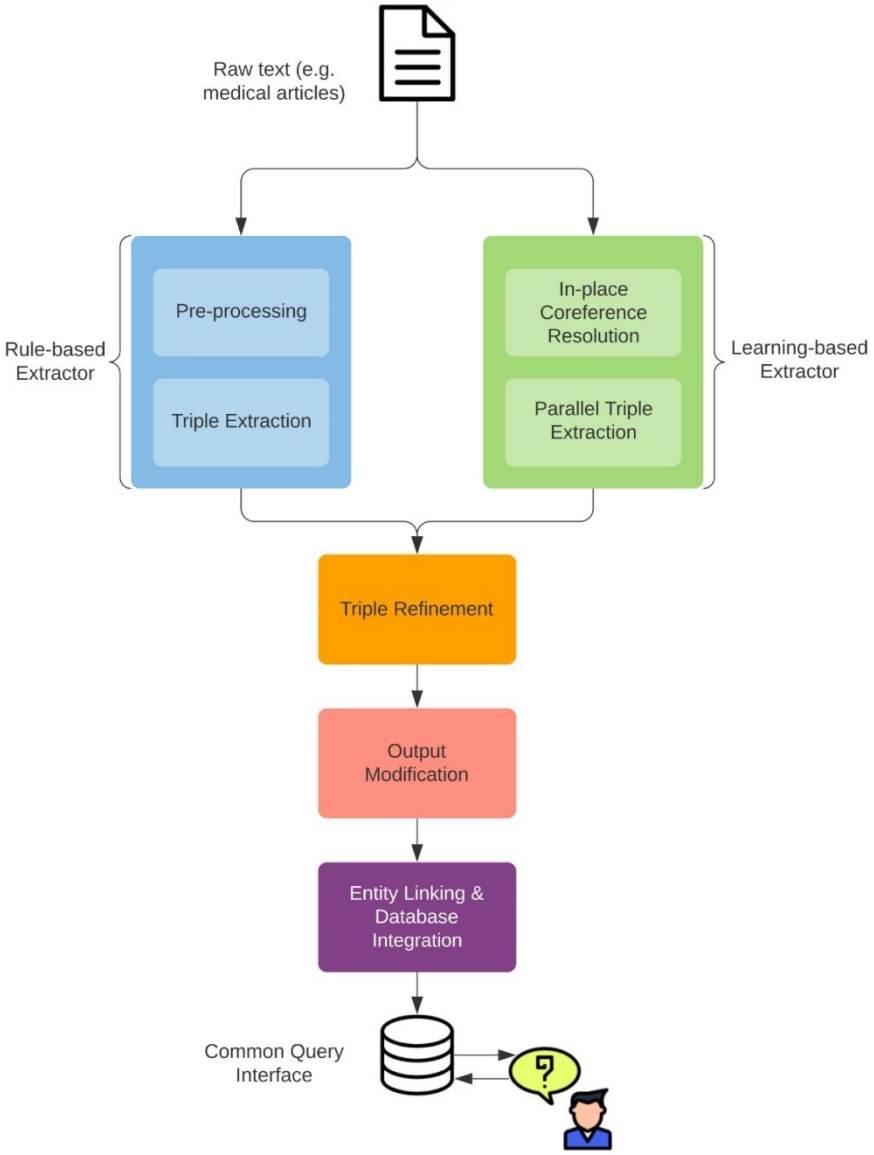
4.2.2.2 Περίπτωση χρήσης 2: Εξαγωγή πληροφοριών από περιλήψεις επιστημονικών δημοσιεύσεων

Η συγκεκριμένη περίπτωση χρήσης αφορά την εξαγωγή πληροφοριών από ελεύθερο κείμενο και την ενσωμάτωσή τους σε σχεσιακή βάση, με απώτερο σκοπό τη δυνατότητα αναζήτησης σε συνδυασμό δομημένων και μη δομημένων δεδομένων, μέσω μιας ενιαίας διεπαφής (πχ. SQL). Για το σκοπό αυτό, παρουσιάζεται η επιτυχής ενσωμάτωση τριπλετών που προκύπτουν από περιλήψεις άρθρων του PubMed [26], εξασφαλίζοντας τις ακόλουθες προϋποθέσεις: α) ότι οι εξαγόμενες πληροφορίες είναι υψηλής ποιότητας και β) ότι διασφαλίζεται η απαραίτητη γενικευσιμότητα για τη σύνδεσή τους με υπάρχουσες οντολογίες που θα διευκολύνουν την αξιοποίησή τους για εφαρμογές τελικού χρήστη.

Η προσέγγιση που ακολουθείται στην συγκεκριμένη εργασία (Smith et al., 2022; Amer-Yahia et al., 2022) περιλαμβάνει τη δημιουργία μιας αρθρωτής αρχιτεκτονικής (LILLIE - Linked Linguistics and Learning-Based Information Extractor) που ενσωματώνει μηχανισμούς παράλληλης εξαγωγής πληροφοριών για την εύρεση οντοτήτων και των μεταξύ τους σχέσεων, στοχεύοντας στη δημιουργία ενός γνωσιακού γράφου. Σε επόμενο βήμα, τα στοιχεία του γράφου συνδέονται με αυτά μιας υφιστάμενης σχεσιακής βάσης μέσω μηχανισμού σύνδεσης οντοτήτων. Το τελικό αποτέλεσμα είναι μια εμπλουτισμένη σχεσιακή βάση που ενσωματώνει πληροφορία η οποία μέχρι πρότινος μπορούσε να προστεθεί μόνο μέσω της χειρωνακτικής επισήμανσης αντίστοιχων κειμένων από εμπειρογνώμονες. Ο κώδικας που υλοποιεί την παραπάνω προσέγγιση είναι διαθέσιμος στον σύνδεσμο: https://github.com/OIELILLIE/LILLIE.

Το σώμα κειμένων που χρησιμοποιήθηκε αποτελείται από περιλήψεις ιατρικών επιστημονικών δημοσιεύσεων στην αγγλική γλώσσα, ενώ η αρθρωτή αρχιτεκτονική που υλοποιήθηκε περιεγράφηκε στην υποενότητα 3.3.1.2 και απεικονίζεται στο Σχήμα 32.

[26] https://pubmed.ncbi.nlm.nih.gov/



Σχήμα 32 Επισκόπηση αρχιτεκτονικής για την εξόρυξη δεδομένων για περίπτωση χρήσης σε περιλήψεις επιστημονικών άρθρων

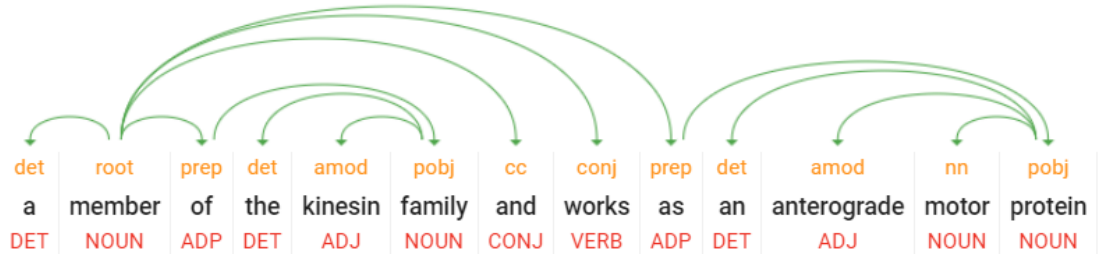
Όπως φαίνεται από το σχήμα, το στάδιο της παράλληλης εξαγωγής πληροφοριών εγκολπώνει έναν συνδυασμό OIE στοιχείων βασισμένο σε γλωσσολογικούς κανόνες (rule-based extractor) καθώς και στοιχεία βασισμένα σε υπολογιστικές μεθόδους (learning-based extractor). Ακόμη, στην παρούσα υλοποίηση αξιοποιείται ο μηχανισμός συνδυασμού αποτελεσμάτων εξαγωγής που αναπτύχθηκε στην υποενότητα 3.3.1.2.3, ενώ διατίθεται και ένα σύνολο επιλογών παραμετροποίησης της εξόδου (output modification) που συμβάλλει στην προσαρμογή του προτεινόμενου συστήματος σε διαφορετικά πεδία εφαρμογής. Τέλος, οι εξαχθείσες τριπλέτες συσχετίζονται με οντότητες δύο υφιστάμενων οντολογιών, προτού ενσωματωθούν σε μια σχεσιακή βάση δεδομένων (entity linking and database integration).

Όσον αφορά το στάδιο παράλληλης εξαγωγής πληροφοριών, χρησιμοποιούνται και τα τέσσερα στοιχεία που αναπτύχθηκαν στην υποενότητα 4.2.1.4, παράγοντας παρόμοια αποτελέσματα με αυτά που παρατέθηκαν στην πρώτη περίπτωση χρήσης και ως εκ τούτου δεν θα αναλυθούν περαιτέρω. Μόλις ολοκληρωθεί η διαδικασία παραγωγής τριπλετών για μια πρόταση, αυτές συνδυάζονται σε ένα ενοποιημένο σύνολο που στοχεύει στην αντιστάθμιση της υψηλής ακρίβειας που προσφέρει το στοιχείο εξαγωγής βασισμένο σε γλωσσολογικούς κανόνες, με την υψηλή ανάκληση που προσφέρει η εξαγωγή με χρήση στοιχείων βασισμένων σε υπολογιστικές μεθόδους. Αυτό επιτυγχάνεται μέσω του σταδίου διύλισης (triple refinement) που αντιστοιχίζει κάθε τριπλέτα με το συντακτικό δένδρο της εκάστοτε πρότασης. Αρχικά, κάθε κόμβος του δένδρου επισημαίνεται εφόσον είναι μέρος του υποκειμένου, του κατηγορήματος ή του αντικειμένου και στη συνέχεια ελέγχεται αν μια ακμή είναι έγκυρη μεταξύ δύο κόμβων βάσει προκαθορισμένων γλωσσολογικών κανόνων, παρόμοιων με αυτούς που χρησιμοποιεί το στοιχείο εξαγωγής βασισμένο σε κανόνες. Για παράδειγμα, έστω η παρακάτω τριπλέτα {S;P;O} που αποτελεί προϊόν στοιχείου εξαγωγής βασισμένου σε μηχανική μάθηση:

{The protein encoded by KIF1A gene ; is ; a member of the kinesin family and works as an anterograde motor protein}

Είναι φανερό ότι τα στοιχεία της παραπάνω τριπλέτας, αν και διαχωρίζονται σωστά, δεν μπορούν να χρησιμοποιηθούν αυτούσια σε μια δομημένη αναπαράσταση της πρότασης, κυρίως λόγω του μεγέθους του αντικειμένου που περιλαμβάνει μια δευτερεύουσα πρόταση.

Ο μηχανισμός αρχικά κατασκευάζει το συντακτικό δένδρο και ελέγχει τις ακμές μεταξύ των κόμβων. Όπως φαίνεται στο Σχήμα 33, οι κόμβοι "member" και "works" συνδέονται μεταξύ τους με συζευκτική σχέση (conjunction) που δεν αποτελεί έγκυρο είδος σχέσης βάσει των προκαθορισμένων κανόνων. Επομένως, ολόκληρος ο κλάδος από το ρήμα "works" και μετά αποκόπτεται, και το αντικείμενο της τριπλέτας γίνεται "a member of the kinesin family".



Σχήμα 33 Παράδειγμα χρήσης συντακτικού δένδρου από μηχανισμό συνδυασμού τριπλετών

Με τον παραπάνω τρόπο, τριπλέτες που ξέφυγαν από το αρχικό στάδιο εξαγωγής μπορούν να προστεθούν στο υπάρχον σύνολο, διατηρώντας την υψηλή ακρίβεια του συνδυαστικού μηχανισμού. Παράλληλα, αναδεικνύονται περιπτώσεις σύνθετων ουσιαστικών (πχ. kinesin family) που μπορούν να αντιστοιχηθούν με στοιχεία οντολογιών σε επόμενο στάδιο.

Η παραπάνω μεθοδολογία εξασφαλίζει και μια σειρά από επιπλέον βελτιώσεις που μπορούν να εφαρμοστούν προαιρετικά, μέσω της ενεργοποίησης βοηθητικών γλωσσολογικών κανόνων στο στάδιο παραμετροποίησης της εξόδου (output modification). Συγκεκριμένα, είναι δυνατό είτε να συμπεριληφθούν είτε να αφαιρεθούν οι επιθετικοί προσδιορισμοί που συνοδεύουν υποκείμενα/αντικείμενα ή αντίστοιχα τα επιρρήματα που συνοδεύουν τα κατηγορήματα της εκάστοτε τριπλέτας, προκειμένου να ελεγχθεί η εκφραστικότητά της ανάλογα με την περίπτωση χρήσης. Ειδικά για συζευκτικές προτάσεις, είναι δυνατός ο διαχωρισμός τους σε πολλές επιμέρους τριπλέτες, πριν την καταχώρισή τους σε σχεσιακή βάση. Ένα συνδυαστικό παράδειγμα των παραπάνω παραμετροποιήσεων φαίνεται παρακάτω:

Πρόταση:
"Long non-coding RNA CCAT2 plays an important role in tumorigenesis, tumor growth and metastasis."

Τριπλέτες:
{Long non-coding RNA CCAT2 ; plays role; tumorigenesis}
{Long non-coding RNA CCAT2 ; plays role; tumor growth}
{Long non-coding RNA CCAT2 ; plays role; metastasis}

Η παραπάνω διαδικασία μπορεί να παρομοιαστεί με εργασίες κανονικοποίησης γνωσιακής βάσης (knowledge base canonicalization) που έγκεινται στην αντικατάσταση οντοτήτων με την κανονική τους μορφή, με σκοπό την ενίσχυση της εννοιολογικής καθαρότητας και την αποφυγή πλεονασμών (Wu et al., 2018). Ωστόσο, η προτεινόμενη προσέγγιση διαφέρει σημαντικά από τις καθιερωμένες ερευνητικές γραμμές (Lin and Chen, 2019), καθώς επιτρέπει τον πλήρη έλεγχο των βοηθητικών πληροφοριών που συλλέγονται από τα στοιχεία εξαγωγής, λαμβάνοντας υπόψη τις προτιμήσεις του χρήστη.

Το τελικό στάδιο της συγκεκριμένης περίπτωσης χρήσης περιλαμβάνει την ενσωμάτωση των εξαχθεισών τριπλετών σε μια σχεσιακή βάση, με χρήση μηχανισμού σύνδεσης οντοτήτων. Συγκεκριμένα, κάθε υποκείμενο και αντικείμενο μιας τριπλέτας συσχετίζεται είτε με συγκεκριμένη ανατομική οντότητα (μέρος του σώματος) της οντολογίας Uberon (Mungall et al., 2012) (πχ. φάρυγγας = UBERON:0006562) είτε με βιομάρτυρα (biomarker) της

γνωσιακής βάσης καρκινικών μεταλλάξεων OncoMX (Dingerdissen et al., 2020) (πχ. Keratin8 = KRT8). Οι διασυνδεδεμένες οντότητες εμπλουτίζουν πίνακα της βάσης OncoMX που αφορά εκφράσεις βιομαρτύρων σε μέρη του σώματος. Ένα παράδειγμα της διαδικασίας εμπλουτισμού φαίνεται στο Σχήμα 34:

pmid integer	gene text	uberon text	uberonname text	subject text	predicate text	object text
24046070	IDH1	UBERON:0002048	lung	idh1	can be used as	a plasma biomarker for the diagnos...
20421816	EGFR	UBERON:0000021	cutaneous appe...	non-small c...	is sensitive to	the small molecule egfr tyrosine kin...
23719586	KRAS	UBERON:0000006	islet of Langerh...	pdx1-driven ...	could induce	islets of langerhans defects
25700797	EGFR	UBERON:0014899	anterolateral lig...	epidermal gr...	are present in	10-20 % of all non-small-cell lung c...

Σχήμα 34 Παράδειγμα εμπλουτισμού γνωσιακής βάσης OncoMX μέσω εξαγωγής πληροφοριών από περιλήψεις άρθρων του PubMed

Η περιγραφείσα μεθοδολογία καθιστά δυνατή την αύξηση του πληροφοριακού περιεχομένου της βάσης OncoMX μέσω της αυτοματοποιημένης εξαγωγής βιβλιογραφικής πληροφορίας, και συγκεκριμένα γονιδίων-βιομαρτύρων που επηρεάζουν την ανάπτυξη καρκίνου σε συγκεκριμένα μέρη του σώματος. Κάθε σειρά του πίνακα περιέχει το μοναδικό αναγνωριστικό (pmid) που αντιστοιχεί σε συγκεκριμένο άρθρο (πηγή της πληροφορίας), ένα όνομα γονιδίου-βιομάρτυρα (gene), μια ανατομική οντότητα (uberon, uberonname), καθώς και τα στοιχεία της τριπλέτας σε μορφή {S;P;O}. Αξίζει να σημειωθεί ότι η δημιουργία και ο εμπλουτισμός αντίστοιχων βάσεων γίνεται κατά κόρον χειροκίνητα από εμπειρογνώμονες-βιολόγους και αποτελεί ιδιαίτερα επίπονη και αργή διαδικασία, εξαιτίας του μεγέθους και της πολυπλοκότητας του προς ανάλυση κειμένου. Μέσω της αυτόματης εξόρυξης πληροφοριών, αφενός απλοποιείται σημαντικά η διαδικασία εμπλουτισμού και αφετέρου διευκολύνεται η αναζήτηση σε συνδυασμό βιβλιογραφικών και δομημένων δεδομένων μέσω μιας ενιαίας κοινής διεπαφής ερωτημάτων, όπως η SQL.

Ακολουθεί η αξιολόγηση της περιγραφείσας μεθοδολογίας, η οποία περιλαμβάνει τα ακόλουθα στάδια: Αρχικά, συγκρίνεται η απόδοση του μηχανισμού παράλληλης εξαγωγής πληροφοριών με δύο OIE συστήματα αιχμής, και συγκεκριμένα το OpenIE6 (Kolluru et al., 2020a) και το IMoJIE (Kolluru et al., 2020b), σε σύνολα δεδομένων αναφοράς γενικού περιεχομένου (benchmark datasets). Στη συνέχεια, τα δύο παραπάνω συστήματα χρησιμοποιούνται στο αρχικό σώμα κειμένων PubMed, προκειμένου να αναδειχθούν τα ποιοτικά πλεονεκτήματα της δικής μας προσέγγισης και να επισημανθεί η δυνατότητα γενίκευσης της σε διαφορετικά πεδία εφαρμογής. Τέλος, παρατίθενται παραδείγματα αξιοποίησης των αποτελεσμάτων εξαγωγής που αφορούν τον εμπλουτισμό σχεσιακής βάσης και την επιτυχή εκτέλεση ερωτημάτων σε αυτήν.

Η συγκριτική αξιολόγηση με τα προαναφερθέντα OIE συστήματα έγινε στα ακόλουθα σύνολα δεδομένων αναφοράς:

- στο σύνολο δεδομένων CaRB (Bhardwaj et al., 2020), το οποίο περιλαμβάνει 1.282 προτάσεις γενικής θεματολογίας, χωρισμένες σε δύο υποσύνολα (για εκπαίδευση και αξιολόγηση αντίστοιχα),
- στο σύνολο δεδομένων Re-OIE16 (Zhan and Zhao, 2020), που αποτελεί βελτιωμένη έκδοση του OIE16 (Stanovsky and Dagan, 2016) και περιλαμβάνει 600 προτάσεις γενικού ενδιαφέροντος.

Για καθένα από τα παραπάνω benchmarks, διατίθενται όλες οι ορθές εξαγωγές τριπλετών επισημασμένες από εμπειρογνώμονες. Στόχος της αξιολόγησης είναι η σύγκριση των εξαγωγών του συστήματός μας (LILLIE) με τις ορθές (gold) εξαγωγές για τον υπολογισμό σχετικών μέτρων (ακρίβεια, ανάκληση κτλ.).

Τα αποτελέσματα που προέκυψαν παρατίθενται στον παρακάτω πίνακα (Πίνακας 10). Η σημαντικότερη βελτίωση παρατηρείται στην τιμή της AUC-PR (περιοχή κάτω από την καμπύλη ακρίβειας-ανάκλησης), με 6% αύξηση έναντι των δύο άλλων OIE συστημάτων. Ειδικότερα για το CaRB benchmark, παρατηρείται καλύτερη επίδοση σε ανάκληση και F1-score, ενώ στο Re-OIE benchmark σημειώθηκε καλύτερη επίδοση σε όλα τα μέτρα. Η ισορροπία ακρίβειας/ανάκλησης, αποτέλεσμα της συνδυαστικής χρήσης τεχνικών εξαγωγής που βασίζονται τόσο στη μηχανική μάθηση όσο και σε γλωσσολογικούς κανόνες, αποτυπώνεται ειδικότερα στο F1-μέτρο. Σημειώνεται ακόμα ότι δε χρησιμοποιήθηκε το υποσύνολο εκπαίδευσης του CaRB benchmark σε προγενέστερο στάδιο (σε αντίθεση με τα άλλα δύο συστήματα IMoJIE και OpenIE6).

Πίνακας 10 Συγκριτική αξιολόγηση του συστήματος εξαγωγής LILLIE με OIE συστήματα αιχμής, σε benchmark σύνολα δεδομένων

	CaRB benchmark				**Re-OIE16 benchmark**			
Σύστημα	AUC	Prec.	Rec.	F1	AUC	Prec.	Rec.	F1
LILLIE	**.391**	.604	**.487**	**.539**	**.543**	**.685**	**.645**	**.664**
IMoJIE	.333	**.647**	.456	.535	.483	.653	.584	.617
OpenIE6	.337	.589	.477	.527	.523	.642	.612	.627

Δεδομένου ότι τα συγκεκριμένα benchmarks είναι γενικής θεματολογίας, τα παραπάνω αποτελέσματα αποτελούν ισχυρή ένδειξη της καλής γενικευσιμότητας της συγκεκριμένης μεθοδολογίας, καθώς και της δυνατότητας χρησιμοποίησής της σε ευρύ θεματικό φάσμα κειμένων. Ακόμα, είναι αξιοσημείωτη η βελτίωση στο μέτρο της ανάκλησης (έως και 10%),

ως αποτέλεσμα του παράλληλου μηχανισμού εξαγωγής βασισμένο σε υπολογιστικές μεθόδους και τεχνικές μηχανικής μάθησης, που αξιοποιήθηκε και στην πρώτη περίπτωση χρήσης.

Προκειμένου να αποσαφηνιστεί η συμβολή των επιμέρους στοιχείων της αρθρωτής αρχιτεκτονικής στα παραπάνω αποτελέσματα, ακολουθεί αξιολόγησή της με αφαίρεση στοιχείων (ablation study) για κάθε μηχανισμό εξαγωγής χωριστά (Πίνακας 11).

Πίνακας 11 Αξιολόγηση μεθοδολογίας LILLIE με αφαίρεση στοιχείων (ablation study)

	CaRB benchmark		**Re-OIE16 benchmark**	
	AUC	F1	AUC	F1
Μηχανισμός εξαγωγής βασισμένος σε γλωσσολογικούς κανόνες				
χωρίς προεπεξεργασία	.337	.503	.459	.624
με προεπεξεργασία	.370	.531	.507	.657
Μηχανισμός εξαγωγής βασισμένος σε υπολογιστικές μεθόδους				
χωρίς επίλυση συναναφορών	.381	.442	.470	.440
με επίλυση συναναφορών	.358	.457	.457	.500
Συνδυασμός παραπάνω μηχανισμών				
χωρίς διύλιση	.365	.424	.478	.448
με διύλιση	**.391**	**.539**	**.543**	**.664**

Με βάση τα παραπάνω αποτελέσματα συμπεραίνεται ότι ο μηχανισμός εξαγωγής βασισμένος σε κανόνες ευνοείται από το στάδιο προεπεξεργασίας (χρήση συντακτικού δένδρου για επισήμανση συντακτικών συσχετίσεων πριν την εφαρμογή των κανόνων). Αντίθετα, το στοιχείο επίλυσης συναναφορών του μηχανισμού που βασίζεται σε υπολογιστικές μεθόδους φαίνεται να οδηγεί σε καλύτερη τιμή του μέτρου F1 αλλά σε χειρότερη τιμή της AUC-PR. Αυτό οφείλεται στο ότι τα δύο benchmarks περιλαμβάνουν μεμονωμένες προτάσεις αντί μεγαλύτερων κειμένων, περιορίζοντας τη θετική επίδραση του στοιχείου επίλυσης συναναφορών. Ωστόσο, η προσθήκη του συγκεκριμένου στοιχείου έχει θετική επίδραση στα συνδυαστικά αποτελέσματα. Τέλος, είναι εμφανή και τα οφέλη του μηχανισμού διύλισης (triple refinement) που συνδυάζει με βέλτιστο τρόπο τις παράλληλες εξαγωγές.

Στη συνέχεια, παρατίθενται συγκριτικά αποτελέσματα εξαγωγής στο σώμα κειμένων PubMed για κάθε σύστημα. Αυτό αποτελείται από 38.703 περιλήψεις άρθρων που περιλαμβάνουν συνολικά 116.049 προτάσεις. Μετά την εξαγωγή τριπλετών από κάθε σύστημα, ακολουθεί η σύνδεση οντοτήτων με τις γνωσιακές βάσεις Uberon και OncoMX, όπως περιεγράφηκε παραπάνω. Τα αποτελέσματα για το σύστημα IMoJIE στο πλήρες σώμα

κειμένου δεν ήταν δυνατόν να αναπαραχθούν λόγω υπερβολικών απαιτήσεων μνήμης του συγκεκριμένου συστήματος. Για το λόγο αυτό, ο Πίνακας 12 περιλαμβάνει τα αποτελέσματα στο πλήρες σώμα κειμένου για τα συστήματα LILLIE και OpenIE6, ενώ η διαδικασία επαναλαμβάνεται (Πίνακας 13) για ένα υποσύνολο 1.000 περιλήψεων (3.035 προτάσεων), όπου απεικονίζεται και η μέση ταχύτητα κάθε συστήματος. Οι παραπάνω μετρήσεις υποστηρίχθηκαν από εξοπλισμό με επεξεργαστή Intel Core i7-7700HQ 2.80 GHz, μνήμη RAM 32 GB και κάρτα γραφικών NVIDIA GeForce GTX 1050.

Πίνακας 12 Συγκριτική αξιολόγηση εξαγωγής τριπλετών στο πλήρες σώμα κειμένων PubMed

	Extracted triples	**Linked triples**
LILLIE (με output modification)	206096	3513
LILLIE (χωρίς output modification)	117290	2448
OpenIE6	247072	3110

Πίνακας 13 Συγκριτική αξιολόγηση εξαγωγής τριπλετών σε υποσύνολο 1000 περιλήψεων του σώματος κειμένων PubMed

	Extracted triples	**Linked triples**	**Time per sentence (sec.)**
LILLIE	5648	**71**	7.54
OpenIE6	**6675**	50	**1.46**
IMoJIE	4565	49	14.19

Τα παραπάνω αποτελέσματα δείχνουν ότι η ποιότητα των εξαγωγών κάθε συστήματος δεν είναι πάντα ανάλογη του πλήθους των παραγόμενων τριπλετών. Συγκεκριμένα, η μεθοδολογία μας επιτυγχάνει υψηλότερο ποσοστό "χρήσιμων" εξαγωγών (δηλαδή τριπλετών που συνδέονται με οντότητες γνωσιακών βάσεων) σε σχέση με αντίστοιχα συστήματα αιχμής. Ακόμη, σημειώνεται ότι η μέση ταχύτητα εξαγωγής κυμαίνεται περίπου στο μέση τιμή των άλλων δύο συστημάτων, καθώς είναι σαφώς μικρότερη του OpenIE6 (που λειτουργεί αποκλειστικά με κανόνες) αλλά διπλάσια του IMoJIE που βασίζεται σε νευρωνικό δίκτυο.

Τέλος, παρατίθενται δύο ενδεικτικά παραδείγματα άντλησης πληροφοριών από τη σχεσιακή βάση OncoMX, μετά τον εμπλουτισμό της με τις εξαγωγές της προτεινόμενης μεθοδολογίας. Στο πρώτο παράδειγμα, αναζητούνται τα γονίδια-βιομάρτυρες που υπερεκφράζονται σε περιπτώσεις καρκίνου του στήθους σύμφωνα με τη βιβλιογραφία. Το αποτέλεσμα που επιστρέφεται είναι 13 γονίδια που αντλήθηκαν μέσω του μηχανισμού εξαγωγής τριπλετών από τις περιλήψεις του σώματος κειμένου PubMed.

SQL Query

```
SELECT distinct gene, uberonname, uberon
FROM triples_fully_linked_v2
WHERE
(predicate like '%overexpress%' or
(predicate like '%express%' and (subject like '%over%'  or
     triples_fully_linked_v2.object like '%over%')) or
     (subject like '%overexpress%' or
     triples_fully_linked_v2.object like '%overexpress%'))
and (subject like '%cancer%' or
     triples_fully_linked_v2.object like '%cancer%') and
polarity='TRUE' and uberonname='breast'
```

Results

	gene text	uberonname text	uberon text
1	EGFR	breast	UBERON:0000310
2	EPCAM	breast	UBERON:0000310
3	CD24	breast	UBERON:0000310
4	NR4A1	breast	UBERON:0000310
5	ECM1	breast	UBERON:0000310
6	RAB5A	breast	UBERON:0000310
7	CXCL13	breast	UBERON:0000310
8	UBQLN1	breast	UBERON:0000310
9	MUC1	breast	UBERON:0000310
10	RET	breast	UBERON:0000310
11	RHOA	breast	UBERON:0000310
12	ERBB2	breast	UBERON:0000310
13	CCND1	breast	UBERON:0000310

Σχήμα 35 Παράδειγμα αναζήτησης γονίδιων που υπερεκφράζονται σε περιπτώσεις καρκίνου του στήθους σύμφωνα με τη βιβλιογραφία

Το δεύτερο παράδειγμα εστιάζει στην εύρεση όλων των βιβλιογραφικών περιπτώσεων που περιλαμβάνουν τις λέξεις-κλειδιά "cancer" και "biomarker" με αυτοματοποιημένο τρόπο, προκειμένου να διευκολυνθεί η ενσωμάτωση αντίστοιχων πληροφοριών σε δομημένη αναπαράσταση. Επιστρέφονται 20 περιπτώσεις γονιδίων-βιομαρτύρων με τη συσχετιζόμενη ανατομική οντότητα καθώς και τα επιμέρους στοιχεία της τριπλέτας.

SQL Query

```
SELECT distinct gene, uberonname, uberon
FROM triples_fully_linked_v2
WHERE
(predicate like '%overexpress%' or
(predicate like '%express%' and (subject like '%over%'  or
     triples_fully_linked_v2.object like '%over%')) or
     (subject like '%overexpress%' or
     triples_fully_linked_v2.object like '%overexpress%'))
and (subject like '%cancer%' or
     triples_fully_linked_v2.object like '%cancer%') and
polarity='TRUE' and uberonname='breast'
```

Results

gene text	uberon text	uberonname text	subject text	predicate text	object text	polarity text
TNF	UBERON:0013756	venous blood	stress variables profile of antioxida...	were biochemically assessed from	venous blood of fifty ovarian c...	TRUE
TNF	UBERON:0000310	breast	tnf polymorphisms	could serve as	useful predictive biomarkers f...	TRUE
CRYAB	UBERON:0002367	prostate gland	lower cryab expression	is a prognostic biomarker for	several types of cancer such ...	TRUE
FABP5	UBERON:0000029	lymph node	fatty acid-binding protein 5 fabp5	was found in previous study to bi...	be a potential for lymph node ...	TRUE
FABP5	UBERON:0002391	lymph	fatty acid-binding protein 5 fabp5	was found in previous study to bi...	be a potential for lymph node ...	TRUE
KLK3	UBERON:0001628	posterior commu...	klk3 gene products like human pro...	are important biomarkers in	the clinical diagnosis of prost...	TRUE
TNF	UBERON:0000178	blood	stress variables profile of inflamma...	were biochemically assessed from	venous blood of fifty ovarian c...	TRUE
CRYAB	UBERON:0002367	prostate gland	lower cryab expression	is a prognostic biomarker for	several types of cancer such ...	TRUE
S100A4	UBERON:0002107	liver	nuclear expression of the calcium-...	is a biomarker of	increased invasiveness in cho...	TRUE
XRCC1	UBERON:0002048	lung	xrcc1 genetic polymorphism	acts	a potential biomarker for lung ...	TRUE
CCND1	UBERON:0000310	breast	ccnd1 mutations	may serve as	biomarkers for early diagnosi...	TRUE
CCND1	UBERON:0000310	breast	ccnd1 and cdk4 mutations	may serve as	biomarkers for early detection...	TRUE
AMACR	UBERON:0002367	prostate gland	expression of the alpha-methylacyl...	has been established as	a specific biomarker for the di...	TRUE
CRYAB	UBERON:0000974	neck	lower cryab expression	is a prognostic biomarker for	several types of cancer such ...	TRUE
QSOX1	UBERON:0000310	breast	qsox1	could be posited as	a new biomarker of good prog...	TRUE
EGFR	UBERON:0000310	breast	egfr expression levels	are	key biomarkers for breast can...	TRUE
MMP1	UBERON:0002048	lung	the mmp1 -1607 1g allele	is a non-significant protective bio...	lung cancer in taiwan	TRUE
TNF	UBERON:0000178	blood	stress variables profile of antioxida...	were biochemically assessed from	venous blood of fifty ovarian c...	TRUE
CRYAB	UBERON:0000974	neck	lower cryab expression	is a prognostic biomarker for	several types of cancer such ...	TRUE
CCND1	UBERON:0000310	breast	ccnd1 and cdk4 mutations	may serve as	biomarkers for early diagnosi...	TRUE

Σχήμα 36 Παράδειγμα αναζήτησης συσχετίσεων μεταξύ περιπτώσεων καρκίνου και γονιδίων σύμφωνα με τη βιβλιογραφία

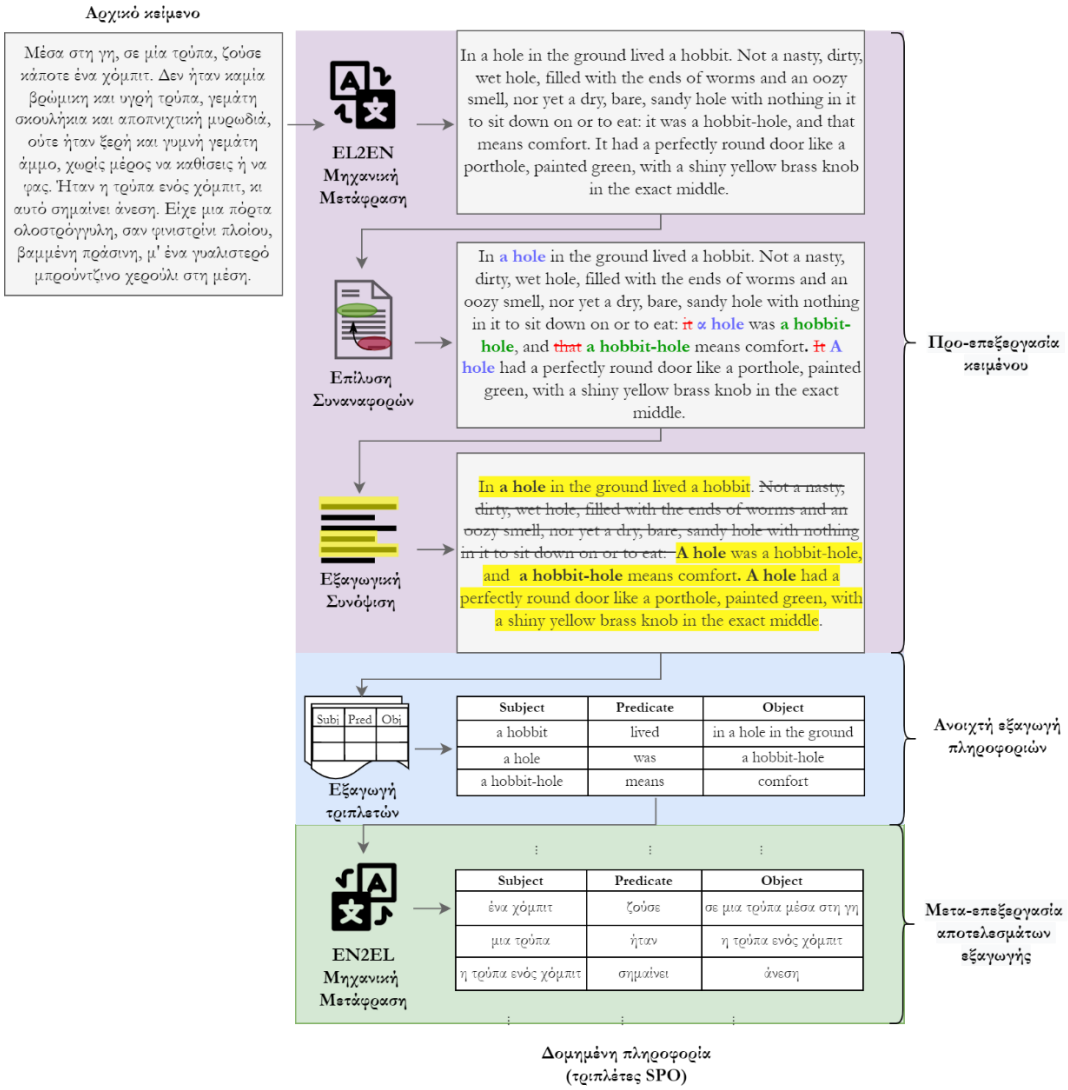
Συμπερασματικά, η δεύτερη περίπτωση χρήσης ανέδειξε μια μεθοδολογία εξαγωγής τριπλετών από ελεύθερο κείμενο που συνδυάζει γλωσσολογικές και υπολογιστικές τεχνικές, με σκοπό των εμπλουτισμό σχεσιακών βάσεων δεδομένων. Αποδείχθηκε ότι ο συνδυασμός συμπληρωματικών μηχανισμών εξαγωγής που βασίζονται σε διαφορετικές τεχνικές οδηγεί σε σημαντική βελτίωση της επίδοσης όσον αφορά τα συνήθη μέτρα (ακρίβεια, ανάκληση, F1, AUC) έναντι αντίστοιχων OIE συστημάτων. Πέραν της ποσοτικής αξιολόγησης, παρατηρήθηκε ότι η προτεινόμενη μεθοδολογία οδηγεί σε τριπλέτες καλύτερης ποιότητας, με την έννοια ότι μπορούν να συσχετιστούν ευκολότερα με οντότητες και άρα να χρησιμοποιηθούν για τον εμπλουτισμό υφιστάμενων βάσεων δεδομένων.

4.2.2.3 Περίπτωση χρήσης 3: Εξαγωγή πληροφοριών από σε ελληνικά κείμενα γενικού περιεχομένου

Οι δύο προηγούμενες περιπτώσεις χρήσης αφορούν την ποιοτική και ποσοτική αξιολόγηση του μηχανισμού εξόρυξης πληροφοριών από ελεύθερο κείμενο για την αγγλική γλώσσα, εστιάζοντας κυρίως στο βέλτιστο συνδυασμό OIE στοιχείων για την εξαγωγή τριπλετών.

Όπως αναπτύχθηκε στην υποενότητα 3.3.1, η γενίκευση της συγκεκριμένης μεθοδολογίας για γλώσσες χαμηλότερων πόρων όπως η ελληνική, μπορεί να επιτευχθεί με τη χρήση στοιχείων μηχανικής μετάφρασης, ώστε το κύριο κομμάτι της επεξεργασίας να γίνεται στο αγγλικό ισοδύναμο του κειμένου και οι εξαχθείσες τριπλέτες να μετασχηματίζονται στα ελληνικά μέσω αντίστροφης μετάφρασης. Η συγκεκριμένη περίπτωση χρήσης ολοκληρώνει την παρουσίαση της συγκεκριμένης μεθοδολογίας, μέσω της εφαρμογής της σε ελληνικά κείμενα γενικού περιεχομένου, αξιολογώντας παράλληλα τόσο την απόδοση των μοντέλων μηχανικής μετάφρασης που εκπαιδεύτηκαν στα πλαίσια της εργασίας, όσο και την απόδοση εξαγωγής πληροφοριών σε end-to-end επίπεδο μέσω αντίστοιχων benchmarks.

Η προσέγγιση της συγκεκριμένης εργασίας (Papadopoulos et al., 2021) στοχεύει στη γεφύρωση του χάσματος όσον αφορά την εξαγωγή πληροφοριών για γλώσσες χαμηλών πόρων, με επίκεντρο την ελληνική γλώσσα. Συγκεκριμένα, προτείνεται ένα ολοκληρωμένο σύστημα εξαγωγής τριπλετών (PENELOPIE - Parallel EN-EL Open Information Extraction) που βασίζεται στη μηχανική μετάφραση από ελληνικά σε αγγλικά και αντίστροφα το οποίο συνδυάζει τα στοιχεία που περιεγράφηκαν στις παραπάνω περιπτώσεις χρήσης. Τα τεχνικά χαρακτηριστικά των στοιχείων που συναποτελούν την προτεινόμενη αρθρωτή αρχιτεκτονική η οποία φαίνεται στο Σχήμα 37, αναλύθηκαν στις υποενότητες 4.2.1.1 έως 4.2.1.4. Ο κώδικας που υλοποιεί τη συγκεκριμένη μεθοδολογία είναι διαθέσιμος στο σύνδεσμο: https://github.com/lighteternal/PENELOPIE.



Σχήμα 37 Επισκόπηση αρχιτεκτονικής PENELOPIE για την εξαγωγή πληροφοριών από ελληνικά κείμενα

Στη συνέχεια, παρατίθενται τα αποτελέσματα της συγκριτικής αξιολόγησης των τεσσάρων μοντέλων μηχανικής μετάφρασης που εκπαιδεύτηκαν στα πλαίσια της εργασίας και περιεγράφηκαν στην υποενότητα 4.2.1.1. Η σύγκριση γίνεται στα πολύγλωσσα σύνολα δεδομένων αναφοράς Tatoeba (Tiedemann, 2020) και XNLI (Conneau et al., 2018), χρησιμοποιώντας το παράλληλο κείμενο που διατίθεται στην ελληνική και στην αγγλική γλώσσα. Αξιοποιούνται δύο δημοφιλή μέτρα, το BLEU (bilingual evaluation understudy) (Papineni et al., 2002) και το chrF (character n-gram F-score) (Popović, 2015).

Το BLEU μετρά την ομοιότητα της πρόβλεψης ενός μοντέλου μηχανικής μετάφρασης (υποψήφια μετάφραση) σε σχέση με την ανθρώπινη μετάφραση (κείμενο αναφοράς), μέσω της ομοιότητας των μεταξύ τους λέξεων (τις οποίες θεωρεί n-grams μέγιστου μήκους N),

ανεξαρτήτως της θέσης τους στο κείμενο και στη συνέχεια υπολογίζει μια τροποποιημένη εκδοχή της ακρίβειας (modified n-gram precision) p_n για κάθε n-gram και τον γεωμετρικό τους μέσο με θετικά βάρη w_n που αθροίζουν στη μονάδα. Παράλληλα, λαμβάνει υπόψη τη διαφορά του μήκους της ακολουθίας της υποψήφιας μετάφρασης c και της μετάφρασης αναφοράς r μέσω του brevity penalty BP, το οποίο ισούται με 1 όταν οι δύο ακολουθίες έχουν ίδιο μήκος. Η σχέση υπολογισμού του BLEU δίνεται παρακάτω:

$$\text{BLEU} = \text{BP} \cdot e^{\sum_{n=1}^{N} w_n \cdot log(p_n)} \quad 4.2$$

όπου:

$$BP = \begin{cases} 1 & \alpha\nu\ r < c \\ e^{1-r/c} & \alpha\nu\ c \leq r \end{cases} \quad 4.3$$

Το συγκεκριμένο μέτρο κυμαίνεται μεταξύ 0 με 1, με τη μονάδα να αντιστοιχεί στην τέλεια μετάφραση.

Το chrF μετρά το ποσοστό επικάλυψης n-grams μεταξύ των συνεχόμενων ακολουθιών από χαρακτήρες (n-grams) που συναντώνται στην υποψήφια μετάφραση σε σχέση με τη μετάφραση αναφοράς, βασιζόμενο στην ακρίβεια $chrP$, η οποία ορίζεται ως το ποσοστό των n-grams στην υποψήφια μετάφραση που υπάρχουν στην μετάφραση αναφοράς, και στην ανάκληση $chrR$, η οποία ορίζεται ως το ποσοστό των n-grams στην μετάφραση αναφοράς που υπάρχουν επίσης στην υποψήφια μετάφραση. Ακόμα, χρησιμοποιείται μια παράμετρος β η οποία συνήθως ισούται με τη μονάδα, αλλά μπορεί να λάβει μεγαλύτερες τιμές ανάλογα με το πόσο σημαντικότερη θεωρείται η ανάκληση σε σχέση με την ακρίβεια. Η σχέση υπολογισμού του chrF δίνεται παρακάτω:

$$chrF = (1 + \beta^2) \frac{chrP \cdot chrR}{chrP + chrR} \quad 4.4$$

Το chrF κυμαίνεται επίσης από το 0 μέχρι το 1, με τη μονάδα να αντιστοιχεί στην τέλεια μετάφραση. Ωστόσο, καθώς βασίζεται σε χαρακτήρες και όχι λέξεις, δεν είναι τόσο ευαίσθητο σε σφάλματα που αφορούν μορφολογικές καταλήξεις, επιβραβεύοντας μερικώς την εύρεση σωστής ρίζας μιας λέξης, σε αντίθεση με το BLEU που απαιτεί πλήρη ομοιότητα. Έτσι, θεωρείται καταλληλότερο για μορφολογικά πλούσιες γλώσσες όπως η ελληνική.

Ο Πίνακας 14 περιλαμβάνει την αξιολόγηση των μοντέλων μηχανικής μετάφρασης βάσει των παραπάνω μέτρων. Για το σύνολο δεδομένων αναφοράς Tatoeba, είναι δυνατή η σύγκριση με τα αντίστοιχα μοντέλα μηχανικής μετάφρασης του Helsinki NLP Language

Technology Research Group[27], τα οποία αποτελούν τις μόνες υλοποιήσεις ανοιχτού κώδικα που αναφέρουν αντίστοιχες μετρήσεις.

Πίνακας 14 Συγκριτική αξιολόγηση μοντέλων μηχανικής μετάφρασης που αναπτύχθηκαν στο PENELOPIE

Αξιολόγηση στο Tatoeba test set (EN-EL)		
	BLEU	**chrF**
Helsinki-2019-12-04-EN2EL	0.527	0.721
Helsinki-2019-12-18-EN2EL	0.564	**0.745**
OURS-lower-case-EN2EL	**0.773**	0.739
OURS-mixed-case-EN2EL	0.769	0.733
Helsinki-2019-12-04-EL2EN	0.694	0.801
OURS-lower-case-EL2EN	**0.799**	**0.802**
OURS-mixed-case-EL2EN	0.793	0.795
Αξιολόγηση στο XNLI test set (EN-EL)		
OURS-lower-case-EN2EL	**0.661**	0.606
OURS-mixed-case-EN2EL	0.654	**0.624**
OURS-lower-case-EL2EN	**0.674**	**0.633**
OURS-mixed-case-EL2EN	0.662	0.623

Από τα παραπάνω αποτελέσματα γίνεται εμφανές ότι τα μοντέλα που εκπαιδεύτηκαν στα πλαίσια του PENELOPIE προσφέρουν σημαντική βελτίωση όσον αφορά το μέτρο BLEU (+0.109 για EN2EL μετάφραση και +0.105 για EL2EN μετάφραση) σε σχέση με τα αντίστοιχα μοντέλα του Helsinki NLP, ενώ όλα έχουν παραπλήσια απόδοση όσον αφορά το μέτρο chrF. Αυτή η διαφορά ανάμεσα στα δύο μέτρα μπορεί να αποδοθεί στις ιδιότυπες μορφολογικές και συντακτικές ιδιότητες της ελληνικής γλώσσας (πχ. κλίσεις ουσιαστικών ή ρημάτων) που οδηγούν σε ελαφρώς παραλλαγμένες υποψήφιες μεταφράσεις σε σχέση με την μετάφραση αναφοράς. Δεδομένου ότι το μέτρο chrF αφορά αντιστοιχίσεις n-grams ενώ το μέτρο BLEU αντιστοιχίσεις ολόκληρων λέξεων, είναι δυνατόν να παράγονται προτάσεις με λανθασμένες καταλήξεις που έχουν χαμηλό BLEU score αλλά αποδεκτό chrF score. Βάσει των παραπάνω, η εξίσου καλή απόδοση των μοντέλων και στα δύο είδη μέτρων, αποτελεί ισχυρή ένδειξη ότι οι παραγόμενες μεταφράσεις προσεγγίζουν αυτές από επαγγελματίες μεταφραστές. Τα αποτελέσματα είναι εξίσου ενθαρρυντικά για το απαιτητικότερο σύνολο δεδομένων αναφοράς XNLI, ωστόσο η έλλειψη αντίστοιχων αποτελεσμάτων από άλλα μοντέλα δεν επιτρέπει την απευθείας σύγκριση. Σημειώνεται ότι και για τα δύο benchmarks,

[27] https://huggingface.co/Helsinki-NLP

τα μοντέλα πεζών χαρακτήρων (lower-case) τείνουν να αποδίδουν ελαφρώς καλύτερα στις παραπάνω δοκιμές, κάτι που θεωρείται αναμενόμενο καθώς λειτουργούν με μικρότερο λεξιλόγιο αγνοώντας μορφολογικές ιδιαιτερότητες που αφορούν τη σωστή κεφαλαιοποίηση ορισμένων λέξεων (πχ. ονοματικών οντοτήτων). Ωστόσο το κέρδος στην απόδοση από τη χρήση μοντέλων πεζών χαρακτήρων δε θεωρείται αρκετό για να αντισταθμίσει τη σωστή κεφαλαιοποίηση που προσφέρουν τα mixed-case μοντέλα. Αξίζει να σημειωθεί ότι σύμφωνα με τα παραπάνω αποτελέσματα, τα υλοποιημένα μοντέλα προσφέρουν την κορυφαία ποιότητα μετάφρασης από ελληνικά σε αγγλικά και αντίστροφα που συναντάται σε υλοποίηση ανοιχτού κώδικα, τουλάχιστον κατά την περίοδο συγγραφής της παρούσας εργασίας.

Ακολουθεί η αξιολόγηση της συνολικής απόδοσης της προτεινόμενης OIE μεθοδολογίας (PENELOPIE), χρησιμοποιώντας τροποποιημένη εκδοχή του CaRB benchmark, το οποίο χρησιμοποιήθηκε στη δεύτερη περίπτωση χρήσης. Δεδομένης της έλλειψης ενός αντίστοιχου συνόλου δεδομένων αναφοράς για την ελληνική γλώσσα, δημιουργήθηκε η ελληνική εκδοχή του CaRB benchmark με χρήση των μοντέλων μηχανικής μετάφρασης που περιεγράφηκαν παραπάνω, για την απόδοση τόσο των προτάσεων όσο και των τριπλετών αναφοράς στα ελληνικά. Έπειτα, οι προτάσεις τροφοδοτήθηκαν στο PENELOPIE, και ακολούθησε η εξαγωγή τριπλετών, η ποιότητα των οποίων μετρήθηκε ως προς την ακρίβεια, την ανάκληση και το F1-score. Όπως προαναφέρθηκε, δεν υπάρχει αντίστοιχο σύστημα εξαγωγής πληροφοριών από αδόμητο κείμενο για την ελληνική γλώσσα, με εξαίρεση ελάχιστες πολύγλωσσες υλοποιήσεις, οι οποίες όμως εξυπηρετούν κυρίως γλώσσες υψηλών πόρων. Το Multi2OIE (Ro et al., 2020b) αποτελεί την κορυφαία αντίστοιχη υλοποίηση καθώς βασίζεται σε πολύγλωσσο BERT μοντέλο, προεκπαιδευμένο σε αγγλικά δεδομένα που ωστόσο μπορεί να διαχειριστεί κείμενα άλλων γλωσσών, χάρη στις δυνατότητες του zero-shot-learning (Huang et al., 2021). Η σύγκριση των δύο υλοποιήσεων παρατίθεται στον παρακάτω πίνακα (Πίνακας 15):

Πίνακας 15 Σύγκριση μοντέλων OIE στο ελληνικό CaRB benchmark dataset

Μοντέλο	Precision	Recall	F1-score
PENELOPIE	**0.231**	**0.284**	**0.255**
Multi2OIE	0.200	0.084	0.118

Είναι εμφανές ότι το PENELOPIE υπερτερεί του πολύγλωσσου Multi2OIE στην εργασία εξαγωγής τριπλετών από ελληνικά κείμενα σε όλα τα μέτρα. Η πιο αξιοσημείωτη διαφορά στην απόδοση παρατηρείται στην ανάκληση, κάτι που μπορεί να αποδοθεί στο ότι το

PENELOPIE συνδυάζει παράλληλα στοιχεία εξαγωγής τριπλετών, οδηγώντας σε μια προσέγγιση προσανατολισμένη στην ανάκληση (recall-oriented). Ως αντιστάθμισμα, πολλές από τις τριπλέτες ενδέχεται να περιέχουν θόρυβο, οδηγώντας σε χαμηλότερη ακρίβεια (precision). Επισημαίνεται ότι, εξαιτίας των εργασιών μετάφρασης του κειμένου στα αγγλικά και της αντίστροφης μετάφρασης των τριπλετών στα ελληνικά, η αντιστοίχιση κάθε μεμονωμένου στοιχείου μιας τριπλέτας με το αρχικό κείμενο δεν είναι πάντα εγγυημένη. Αυτό δικαιολογεί τις σχετικά χαμηλές τιμές σε όλα τα μέτρα σε σχέση με αντίστοιχα συστήματα για γλώσσες υψηλών πόρων (πχ. αγγλικά), όπου οι τιμές για το F1-score για το συγκεκριμένο benchmark κυμαίνονται γύρω στο 0.50. Ωστόσο, η συγκεκριμένη μεθοδολογία επιδεικνύει σαφώς καλύτερη απόδοση (+116% σε F1-score) σε σχέση με τις υπάρχουσες λύσεις και αποτελεί ένα θετικό πρώτο βήμα για την υλοποίηση συστημάτων αντίστοιχης απόδοσης για γλώσσες χαμηλών πόρων. Ακόμη, επιβεβαιώνεται ότι παρά τη δυνατότητα πρόσβασης σε τεράστιο σώμα κειμένων, η χρήση προεκπαιδευμένων πολύγλωσσων μοντέλων Transformer για γλώσσες χαμηλών πόρων συχνά οδηγεί σε χειρότερα αποτελέσματα σε σχέση με αντίστοιχες μονόγλωσσες υλοποιήσεις (Wu and Dredze, 2020).

4.3 Μηχανισμός εξαγωγής σημασιολογικών συμπερασμάτων

4.3.1 Τεχνική υλοποίηση

Στις ακόλουθες υποενότητες περιγράφονται τα τεχνικά χαρακτηριστικά των στοιχείων που απαρτίζουν τον μηχανισμό εξαγωγής σημασιολογικών συμπερασμάτων, και συγκεκριμένα των στοιχείων συλλογής και επεξεργασίας δεδομένων, των στοιχείων σύνδεσης οντοτήτων του στοιχείου κατασκευής τεκμηρίων και των στοιχείων ελέγχου σημασιολογικής ομοιότητας, αναγνώρισης κειμενικής συνεπαγωγής και αναγνώρισης υποκειμενικότητας/αντικειμενικότητας. Δίνεται έμφαση στην επισκόπηση των τεχνολογιών, δεδομένων και υπερ-παραμέτρων που αφορούν τα νευρωνικά μοντέλα τα οποία εκπαιδεύτηκαν στα πλαίσια της εργασίας και αφορούν την ελληνική γλώσσα, ενώ παρατίθενται τεχνικές λεπτομέρειες σχετικά με την παραμετροποίηση των προεκπαιδευμένων μοντέλων και συστημάτων που αξιοποιήθηκαν.

4.3.1.1 Στοιχεία συλλογής και επεξεργασίας δεδομένων

4.3.1.1.1 Στοιχείο αυτόματης αναγνώρισης ομιλίας

Το στοιχείο αυτόματης αναγνώρισης ομιλίας εξυπηρετεί την ευκολότερη εισαγωγή ισχυρισμού από τον τελικό χρήστη. Αναπτύχθηκε νευρωνικό μοντέλο αυτόματης

αναγνώρισης ομιλίας για την ελληνική γλώσσα, αρχιτεκτονικής XLSR-Wav2Vec2, το οποίο εκπαιδεύτηκε στο συνδυαστικό σύνολο απομαγνωτοφωνημένων φωνητικών δεδομένων ανοιχτού κώδικα της Mozilla Common Voice [28] μεγέθους 364MB και του CSS10 (Collection of Single Speaker Speech Datasets for 10 Languages) μεγέθους 1.22GB (Park and Mulc, 2019), στα πλαίσια σχετικού διαγωνισμού που οργανώθηκε από τον οργανισμό Hugging Face[29]. Χρησιμοποιήθηκε η βιβλιοθήκη Transformers[30] για την προσαρμογή (finetuning) του προεκπαιδευμένου μοντέλου Wav2Vec2 με εφαρμογή του αλγορίθμου Connectionist Temporal Classification (CTC)[31]. Καθώς τα μοντέλα αναγνώρισης ομιλίας μετατρέπουν ήχο σε κείμενο, απαιτούν έναν αλγόριθμο εξαγωγής χαρακτηριστικών (feature extractor) που τροφοδοτεί το ηχητικό σήμα σε μορφή συμβατή με την είσοδο του νευρωνικού δικτύου, καθώς και έναν tokenizer που μετατρέπει την έξοδο του νευρωνικού δικτύου σε κείμενο. Χρησιμοποιήθηκε ο Wav2Vec2FeatureExtractor [32] και ο Wav2Vec2CTCTokenizer33 αντίστοιχα.

Όσον αφορά τον Wav2Vec2FeatureExtractor, παραμετροποιήθηκε ώστε να επεξεργάζεται τα φωνητικά δεδομένα με sampling rate ίσο με 16kHz, καθώς αυτή είναι η αναμενόμενη συχνότητα δειγματοληψίας από το μοντέλο Wav2Vec2. Ακόμη, η είσοδος υποβλήθηκε σε κανονικοποίηση, με σκοπό της βελτίωσης της σύγκλισης του μοντέλου κατά την εκπαίδευση. Αντίστοιχα, ο Wav2Vec2CTCTokenizer παραμετροποιήθηκε ώστε να δημιουργεί το δυνατό λεξιλόγιο (σύνολο χαρακτήρων) βάσει του απομαγνητοφωνημένου συνόλου δεδομένων, αφού αφαιρέσει τους περισσότερους ειδικούς χαρακτήρες και μετατρέψει το κείμενο σε πεζό. Δεδομένου ότι σκοπός του τελικού μοντέλου είναι η αντιστοίχιση επιμέρους κομματιών ομιλίας σε χαρακτήρες (φωνήματα), ο συγκεκριμένος tokenizer εξάγει όλα τα διακριτά γράμματα από το σύνολο εκπαίδευσης, τα οποία αποτελούν τις δυνατές κλάσεις τις οποίες μπορεί να ταξινομήσει το εκπαιδευμένο μοντέλο. Στην περίπτωσή μας το λεξιλόγιο αποτελείται από 55 χαρακτήρες (24 γράμματα του ελληνικού αλφαβήτου, συν κάποια τονούμενα φωνήεντα, αγγλικά γράμματα και κάποια ειδικά tokens όπως το κενό).

---

[28] https://commonvoice.mozilla.org/el/datasets

[29] https://discuss.huggingface.co/t/open-to-the-community-xlsr-wav2vec2-fine-tuning-week-for-low-resource-languages/4467

[30] https://huggingface.co/docs/transformers/model_doc/wav2vec2

[31] https://distill.pub/2017/ctc/

[32] https://huggingface.co/transformers/master/model_doc/wav2vec2.html#wav2vec2featureextractor

[33] https://huggingface.co/transformers/master/model_doc/wav2vec2.html#wav2vec2ctctokenizer

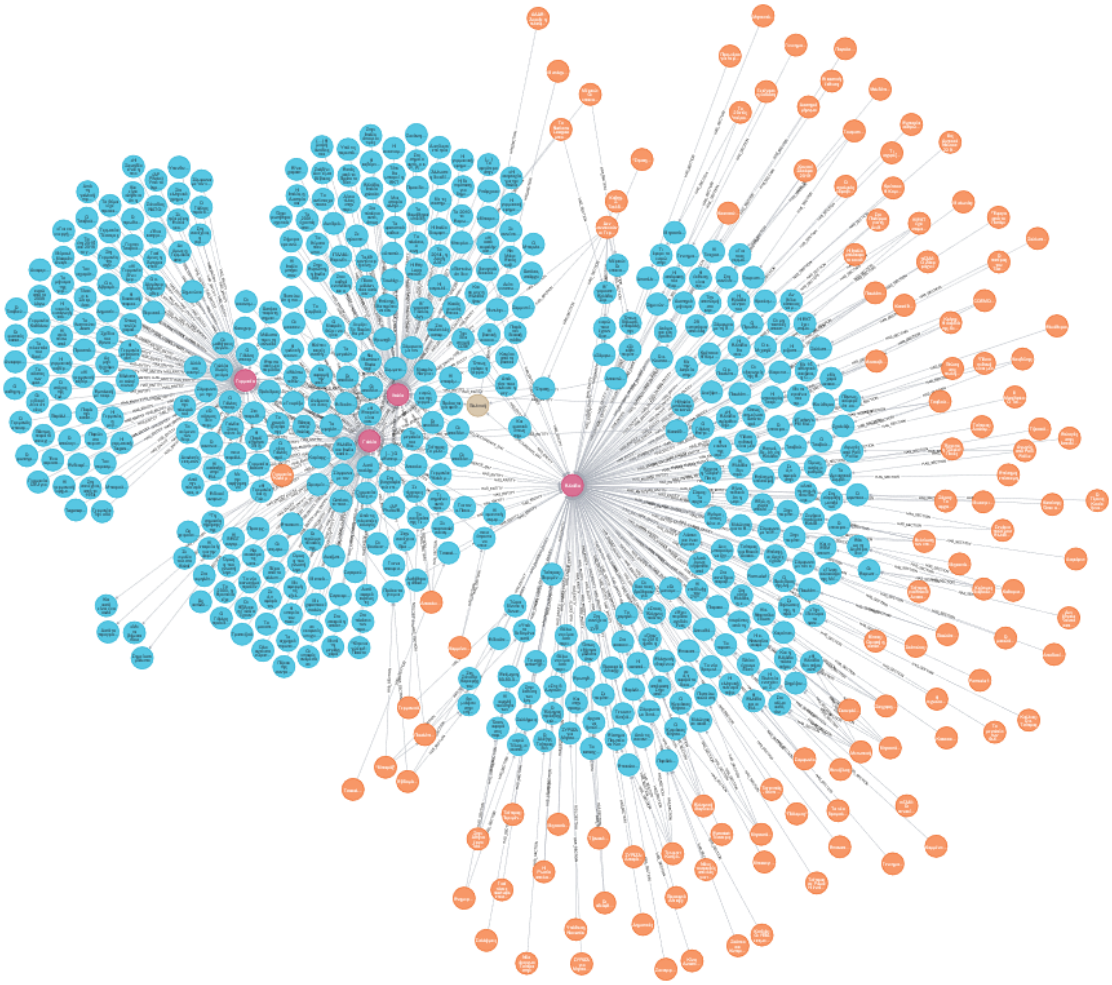
Αφού ολοκληρώθηκε η προεπεξεργασία των φωνητικών δεδομένων, ακολούθησε η εκπαίδευση του μοντέλου για 50 εποχές, διάρκειας περίπου 8 ωρών με χρήση κάρτας γραφικών NVIDIA RTX 3080. Το σύνολο των υπερ-παραμέτρων που χρησιμοποιήθηκε κατά την εκπαίδευση συγκεντρώνεται παρακάτω (Πίνακας 16). Το τελικό μοντέλο διατίθεται προς ελεύθερη χρήση μέσω του παρακάτω συνδέσμου: https://huggingface.co/lighteternal/wav2vec2-large-xlsr-53-greek.

Πίνακας 16 Υπερ-παράμετροι εκπαιδευμένου μοντέλου αναγνώρισης ομιλίας


```
{
  "_name_or_path": "facebook/wav2vec2-large-xlsr-53",
  "activation_dropout": 0.0,
  "apply_spec_augment": true,
  "architectures": ["Wav2Vec2ForCTC"],
  "attention_dropout": 0.1,
  "bos_token_id": 1,
  "conv_bias": true,
  "conv_dim": [512, 512, 512, 512, 512, 512, 512],
  "conv_kernel": [10,3,3,3,3,2,2],
  "conv_stride": [5,2,2,2,2,2,2],
  "ctc_loss_reduction": "mean",
  "ctc_zero_infinity": true,
  "do_stable_layer_norm": true,
  "eos_token_id": 2,
  "feat_extract_activation": "gelu",
  "feat_extract_dropout": 0.0,
  "feat_extract_norm": "layer",
  "feat_proj_dropout": 0.0,
  "final_dropout": 0.0,
  "gradient_checkpointing": true,
  "hidden_act": "gelu",
  "hidden_dropout": 0.1,
  "hidden_size": 1024,
  "initializer_range": 0.02,
  "intermediate_size": 4096,
  "layer_norm_eps": 1e-05,
  "layerdrop": 0.1,
  "mask_channel_length": 10,
  "mask_channel_min_space": 1,
  "mask_channel_other": 0.0,
  "mask_channel_prob": 0.0,
  "mask_channel_selection": "static",
  "mask_feature_length": 10,
  "mask_feature_prob": 0.0,
  "mask_time_length": 10,
  "mask_time_min_space": 1,
  "mask_time_other": 0.0,
  "mask_time_prob": 0.05,
  "mask_time_selection": "static",
  "model_type": "wav2vec2",
  "num_attention_heads": 16,
  "num_conv_pos_embedding_groups": 16,
  "num_conv_pos_embeddings": 128,
  "num_feat_extract_layers": 7,
  "num_hidden_layers": 24,
  "pad_token_id": 54,
  "transformers_version": "4.4.0.dev0",
  "vocab_size": 55
}
```

4.3.1.1.2 Βάση δεδομένων γράφων

Η διαδικασία εξαγωγής σημασιολογικών συμπερασμάτων από προϋποθέτει την ύπαρξη αντίστοιχων πληροφοριών (τεκμηρίων) σε δομημένη μορφή. Για την αποθήκευση και περαιτέρω επεξεργασία των δεδομένων που εξάγονται από την ιχνηλάτηση ειδησεογραφικών πηγών χρησιμοποιήθηκε το σύστημα διαχείρισης βάσεων δεδομένων γράφων Neo4J[34]. Ο εμπλουτισμός της βάσης υλοποιήθηκε μέσω ρουτίνας σε Python και Cypher, η οποία εκτελείται περιοδικά και αναλαμβάνει να εισάγει τα δεδομένα του παραπάνω στοιχείου ιχνηλάτησης στη γνωσιακή βάση.



Σχήμα 38 Παράδειγμα γραφικής απεικόνισης της εξαχθείσας πληροφορίας σε βάση δεδομένων γράφων

Όπως αναφέρθηκε στην υποενότητα 3.4.1.2, ο τελικός γράφος περιλαμβάνει κόμβους για την αναπαράσταση του συνολικού άρθρου (Article nodes) και των επιμέρους αποσπασμάτων του (Section nodes) καθώς και όλων των διαφορετικών τύπων οντοτήτων

[34] https://neo4j.com/

(Entity nodes) με τις οποίες συσχετίζεται κάθε απόσπασμα. Η αποθήκευση πολλαπλών ιδιοτήτων για κάθε κόμβο ή ακμή (που υποδηλώνει σχέση μεταξύ κόμβων) είναι δυνατή, καθώς η Neo4J υλοποιεί μοντέλο γράφων ιδιοτήτων με ετικέτες (Labeled Property Graph). Στο Σχήμα 38 φαίνεται ένα παράδειγμα απεικόνισης της εξαχθείσας πληροφορίας στη βάση, με τα άρθρα να απεικονίζονται ως πορτοκαλί κόμβοι, τα επιμέρους αποσπάσματα ως μπλε κόμβοι και τις ανιχνευθείσες οντότητες ως κόκκινοι κόμβοι.

4.3.1.1.3 Στοιχείο ιχνηλάτησης τροφοδοσιών RSS

Το ένα εκ των δύο στοιχείων ιχνηλάτησης επικεντρώνεται σε τροφοδοσίες RSS ειδησεογραφικών πηγών. Αναπτύχθηκε RSS crawler σε περιβάλλον Python με χρήση της βιβλιοθήκης Scrapy[35]. Ο συγκεκριμένος crawler λαμβάνει ως παράμετρο μια λίστα από URLs που αφορούν τροφοδοσίες RSS γνωστών ελληνικών ειδησεογραφικών sites και εξάγει βασικές πληροφορίες κάθε είδησης (πχ. μοναδικό αναγνωριστικό, τίτλος, ημερομηνία δημοσίευσης, περιεχόμενο, κτλ.) σε μορφή JSON. Στη συνέχεια, κάθε αρχείο εισάγεται στη βάση δεδομένων γράφων Neo4J σχηματίζοντας ξεχωριστό κόμβο τύπου Article, με χρήση της βιβλιοθήκης Py2Neo[36]. Η περιοδικότητα ιχνηλάτησης νέων τροφοδοσιών RSS καθώς και της εισαγωγής των δεδομένων τους στη βάση εξασφαλίζεται μέσω αυτοματοποιημένης cron εργασίας. Ο Πίνακας 17 παρουσιάζει ένα παράδειγμα ιχνηλάτησης τροφοδοσίας RSS από ελληνική ειδησεογραφική πηγή.

Πίνακας 17 Παράδειγμα ιχνηλάτησης τροφοδοσίας RSS


```
{
"id": "ce61d9fd76c92a6609aae204fabe791e34caa562",
"Title": " Θετικό το ισοζύγιο τρεχουσών συναλλαγών λόγω τουρισμού",
"Link": "http://www.tovima.gr/finance/article/?aid=1006884",
"Publication Date": "Mon, 23 Jul 2018 08:39:38 GMT",
"Description": "Πλεόνασμα 192 εκατ. ευρώ εμφάνισε το Μάιο του 2018 το ισοζύγιο τρεχουσών συναλλαγών, έναντι ελλείμματος 658 εκατ. ευρώ τον ίδιο μήνα του 2017, εξέλιξη που αποδίδεται κατά βάση στην έναρξη της τουριστικής σεζόν και την αύξηση των εσόδων από ξένους επισκέπτες.,
"Last Accessed": "2018-07-23 12:00:25.270010"
}
```


4.3.1.1.4 Στοιχείο ιχνηλάτησης HTML σελίδων

Το δεύτερο στοιχείο ιχνηλάτησης επεκτείνει τις δυνατότητες του πρώτου καθώς βασίζεται στην εξαγωγή πληροφοριών από το πλήρες κείμενο μιας ειδησεογραφικής πηγής,

[35] https://scrapy.org/

[36] https://pypi.org/project/py2neo/

εκμεταλλευόμενο τη δομή της HTML σελίδας που το περιέχει. Χρησιμοποιήθηκε η Python βιβλιοθήκη ανοιχτού κώδικα news-please[37] για την ιχνηλάτηση δομημένων πληροφοριών από HTML σελίδες ειδησεογραφικού περιεχομένου. Χάρη στη δυνατότητά της να ακολουθεί αναδρομικά εσωτερικούς υπερ-συνδέσμους ενός ιστοτόπου εκμεταλλευόμενη το sitemap του ως αρχικό σημείο αναζήτησης, η συγκεκριμένη βιβλιοθήκη μπορεί να χρησιμοποιηθεί ως συνεχής διαδικασία παρασκηνίου (daemon) απαιτώντας απλά ως είσοδο τη λίστα κύριων διευθύνσεων (root URLs) των ειδησεογραφικών ιστοτόπων που θα αποτελέσουν την πηγή εξαγωγής πληροφοριών. Κάθε είδηση εξάγεται ως JSON αρχείο, το οποίο περιλαμβάνει βασικές πληροφορίες (πχ. τίτλος, σώμα κειμένου, συγγραφέας, ημερομηνία συγγραφής, ημερομηνία ιχνηλάτησης, URL είδησης κτλ.). ως ξεχωριστά πεδία. Στον παρακάτω πίνακα παρουσιάζεται παράδειγμα ιχνηλάτησης HTML σελίδας από ειδησεογραφικό site, με τα αντίστοιχα πεδία.

Πίνακας 18 Παράδειγμα ιχνηλάτησης HTML σελίδας ειδησεογραφικού περιεχομένου


```
{
  "authors": [],
  "date_download": "2021-10-01 22:07:31",
  "date_modify": "2021-10-01 22:07:31",
  "date_publish": "2018-03-13 00:00:00",
  "description": "Ειδήσεις - Εντολή να ξαναρχίσουν οι ειρηνευτικές διαπραγματεύσεις με την τελευταία οργάνωση ανταρτών που συνεχίζει τη δράση της στην Κολομβία , τον Στρατό Εθνικής...",
  "filename": "kosmos_story_121306_kolomvia-epanenarxi-ton-eirineytikon-diapragmateyseon-me-ton-eln_1633126051.html",
  "image_url": "https://cdn.cnngreece.gr/media/news/2018/03/13/121306/facebook/19014334.jpg",
  "language": "el",
  "localpath": "/home/earendil/news-please-repo//data/2021/10/01/cnn.gr/kosmos_story_121306_kolomvia-epanenarxi-ton-eirineytikon-diapragmateyseon-me-ton-eln_1633126051.html",
  "title": "Κολομβία: Επανέναρξη των ειρηνευτικών διαπραγματεύσεων με τον ELN",
  "title_page": "Κολομβία: Επανέναρξη των ειρηνευτικών διαπραγματεύσεων με τον ELN - CNN.gr",
  "title_rss": "NULL",
  "source_domain": "cnn.gr",
  "maintext": "Εντολή να ξαναρχίσουν οι ειρηνευτικές διαπραγματεύσεις με την τελευταία οργάνωση ανταρτών που συνεχίζει τη δράση της στην Κολομβία , τον Στρατό Εθνικής Απελευθέρωσης (ELN), έδωσε ο πρόεδρος της χώρας των Άνδεων, Χουάν Μανουέλ Σάντος .Οι διαπραγματεύσεις είχαν ανασταλεί για έξι εβδομάδες έπειτα από πολύνεκρες επιθέσεις με στόχο κυρίως αστυνομικούς. «Αφότου τερματίστηκε η αμοιβαία κατάπαυση του πυρός, υπήρξαν και στις δύο πλευρές πάρα πολλοί νεκροί, πάρα πολλοί τραυματίες, πάρα πολλά θύματα - αυτό πρέπει να το σταματήσουμε», τόνισε ο Σάντος σε τηλεοπτικό διάγγελμά του. «Και αυτό μπορεί να γίνει μόνο αν μιλήσουμε», συνέχισε. «Για να σωθούν ζωές, για να υπάρξει απόλυτη ειρήνη στην Κολομβία, αποφάσισα να ξαναρχίσει ο διάλογος με τον ELN», υπογράμμισε ο πρόεδρος Σάντος και διευκρίνισε ότι έδωσε «οδηγίες στον επικεφαλής διαπραγματευτή (της κυβέρνησης) Γκουστάβο Μπελ να μεταβεί στο Κίτο για να επαναληφθούν οι συνομιλίες» στην πρωτεύουσα του Ισημερινού, η λεγόμενη δημόσια φάση των οποίων είχε αρχίσει τον Φεβρουάριο του 2017. Ο Σάντος πήρε αυτή την απόφαση μετά την μονομερή εκεχειρία που ανακήρυξε -και τήρησε απαρέγκλιτα- ο ELN ενόψει των βουλευτικών εκλογών της Κυριακής. Χωρίς να διευκρινίσει πότε ακριβώς θα αρχίσουν εκ νέου οι διαπραγματεύσεις, ο πρόεδρος της Κολομβίας ανέφερε πως καταρχήν τα μέρη θα συζητήσουν «μια νέα (σ.σ. αμοιβαία) κατάπαυση του πυρός και διακοπή των εχθροπραξιών», η οποία θα είναι «επαληθεύσιμη», προκειμένου «να αποφευχθεί ο κίνδυνος να χαθούν κι άλλες ζωές». «Απαντάμε θετικά στο κάλεσμα του προέδρου Σάντος να ξαναρχίσουν οι συνομιλίες, με την πεποίθηση ότι είναι καλύτερο να γίνει ο διάλογος στη βάση μιας αμοιβαίας εκεχειρίας», ανέφερε σε ανακοίνωσή της λίγη ώρα αργότερα η οργάνωση των ανταρτών. Ο Σάντος, ο οποίος εξελέγη στην προεδρία της Κολομβίας το 2010 και θα αποχωρήσει από την εξουσία τον Αύγουστο, έπειτα από δύο συναπτές τετραετείς θητείες, αποπειράται να συνάψει με τον ELN μια συμφωνία ειρήνης παρόμοια με εκείνη που υπέγραψε
```


[37] https://github.com/fhamborg/news-please

το 2016 με την ισχυρότερη οργάνωση ανταρτών της λατινοαμερικάνικης χώρας, τη FARC, η οποία έκτοτε αφοπλίστηκε και μετασχηματίστηκε σε πολιτικό κόμμα. Η πρώτη συμφωνία αμοιβαίας κατάπαυσης του πυρός που συνήψαν τα δύο μέρη στην ιστορία διήρκεσε 101 ημέρες, από την 1η Οκτωβρίου 2017 ως την 9η Ιανουαρίου 2018, αλλά δεν ανανεώθηκε εξαιτίας μιας σειράς κυρίως βομβιστικών επιθέσεων των ανταρτών που είχαν αποτέλεσμα να σκοτωθούν οκτώ αστυνομικοί και 19 στρατιώτες και να τραυματιστούν δεκάδες άλλα μέλη των σωμάτων ασφαλείας και των ένοπλων δυνάμεων. Στις τάξεις των ανταρτών οι απώλειες ήταν τουλάχιστον 34 την ίδια περίοδο. Οι προεδρικές εκλογές, που θα διεξαχθούν σε δύο γύρους, την 27η Μαΐου και τη 17η Ιουνίου, ενδέχεται να κρίνουν την μοίρα των ειρηνευτικών διαπραγματεύσεων ανάμεσα στην Μπογοτά και τους αντάρτες του ELN. Η Κολομβία σπαράσσεται επί μισό αιώνα και πλέον από τον πιο μακρόχρονο εμφύλιο πόλεμο στο δυτικό ημισφαίριο, στον οποίο ενεπλάκησαν περίπου τριάντα οργανώσεις ανταρτών που ανήκαν στην αριστερά, παραστρατιωτικές οργανώσεις της άκρας δεξιάς, συμμορίες του οργανωμένου εγκλήματος, ο στρατός και η αστυνομία. Στον πόλεμο έχασαν τη ζωή τους 260.000 άνθρωποι, πάνω από 60.000 θεωρούνται ως και σήμερα επισήμως αγνοούμενοι, ενώ οι εσωτερικά εκτοπισμένοι ξεπερνούν τα 7,4 εκατομμύρια ανθρώπους.",
"url": "https://www.cnn.gr/kosmos/story/121306/kolomvia-epanenarxi-ton-eirineytikon-diapragmateyseon-me-ton-eln"
}

4.3.1.2 Στοιχεία σύνδεσης οντοτήτων

4.3.1.2.1 Στοιχείο σύνδεσης οντοτήτων βασισμένο σε νευρωνικό μοντέλο

Το στοιχείο σύνδεσης οντοτήτων που εκπαιδεύτηκε σε δεδομένα της WikiData μέσω αρθρωτής (pipeline) διαδικασίας, βασίζεται στη βιβλιοθήκη spaCy [38]. Η διαδικασία περιελάμβανε τη δημιουργία γνωσιακής βάσης από ένα WikiData dump[39] που διατίθεται σε μορφή JSON, και την εκπαίδευση νευρωνικού μοντέλου που θα αντιστοιχεί αναγνωρισμένες ονοματικές οντότητες με συγκεκριμένες έννοιες της WikiData. Η εκπαίδευση του μοντέλου έγινε για 10 εποχές, με dropout 0.5, learning rate 0.005 και σταθερά L2 regularization $10^{(-6)}$. Η διαδικασία διήρκεσε 38 ώρες και υποστηρίχθηκε από GPU NVIDIA GeForce RTX2080 SUPER. Όπως αναφέρθηκε στην υποενότητα 3.4.1.2, λόγω υπολογιστικών περιορισμών χρησιμοποιήθηκε ένα υποσύνολο της γνωσιακής βάσης, με αποτέλεσμα ο ταξινομητής να περιλαμβάνει ως δυνατές κλάσεις μόνο τις 500.000 πιο συχνά εμφανιζόμενες οντότητες που ανήκουν στις κατηγορίες "EVENT", "GPE", "LOC", "ORG", "PERSON" και "PRODUCT".

4.3.1.2.2 Στοιχείο αυτόματης αντιστοίχισης εννοιών με οντότητες της Wikipedia (Wikification)

Ως εναλλακτική μέθοδος σύνδεσης οντοτήτων του προαναφερθέντος νευρωνικού μοντέλου, υλοποιήθηκε κώδικας σε Python που καλούσε το Web API του JSI Wikifier [40], μια ελεύθερα διαθέσιμη διαδικτυακή υπηρεσία που δέχεται ως είσοδο ένα οποιοδήποτε σώμα κειμένου και επισημαίνει τις ανιχνευθείσες σε αυτό έννοιες από τη Wikipedia, η οποία

[38] https://spacy.io/

[39] https://dumps.wikimedia.org/wikidatawiki/entities/

[40] https://wikifier.org/info.html

περιλαμβάνει τη μεγαλύτερη, ενημερωμένη γνωσιακή βάση γενικού περιεχομένου στον κόσμο (WikiData).

Η κλήση του συγκεκριμένου API έγινε σε επίπεδο πρότασης (για κάθε Section ενός άρθρου) και προσαρμόστηκε μέσω της ρύθμισης μιας σειράς παραμέτρων που αφορούν κυρίως την ευαισθησία του στοιχείου βάσει PageRank score (pageRankSqThreshold), την αποσαφήνιση όρων (maxMentionEntropy), καθώς το πλήθος των υποψήφιων εννοιών που λαμβάνονται υπόψη για την αντιστοίχιση μιας αναφοράς (maxTargetsPerMention). Η πλήρης λίστα παραμέτρων που χρησιμοποιήθηκε για την προσαρμογή του στοιχείου παρατίθεται παρακάτω:

Πίνακας 19 Λίστα παραμέτρων JSI Wikifier

```
lang=auto
pageRankSqThreshold=0.80
applyPageRankSqThreshold=true
nTopDfValuesToIgnore=200
nWordsToIgnoreFromList=200
minLinkFrequency=100
maxMentionEntropy=10
wikiDataClasses=false
wikiDataClassIds=false
```

4.3.1.3 Στοιχείο κατασκευής τεκμηρίων

Η διαδικασία συγκέντρωσης αποσπασμάτων από ειδησεογραφικά άρθρα, προκειμένου αυτά να χρησιμοποιηθούν ως υποψήφια τεκμήρια για τον έλεγχο ενός ισχυρισμού περιεγράφηκε διεξοδικά στην υποενότητα 3.4.1.3. Ο σχετικός ψευδοκώδικας που αναπτύχθηκε στα πλαίσια της διατριβής υλοποιήθηκε σε περιβάλλον Python με τα ερωτήματα που τίθενται στη βάση δεδομένων γράφων να γίνονται απευθείας μέσω της γλώσσας Cypher, αξιοποιώντας τον σχετικό Neo4J Bolt Driver[41] που διατίθεται ως PyPI βιβλιοθήκη.

4.3.1.4 Στοιχείο ελέγχου σημασιολογικής ομοιότητας

Η εκπαίδευση του δίγλωσσου (EN-EL) νευρωνικού δικτύου αρχιτεκτονικής XLM-RoBERTa, το οποίο αξιοποιήθηκε για τη σύγκριση κάθε υποψήφιου τεκμηρίου με τον ισχυρισμό του χρήστη σε επίπεδο σημασιολογικής ομοιότητας (βλ. υποενότητα 3.4.1.4), υλοποιήθηκε με χρήση της Python βιβλιοθήκης sentence-transformers[42]. Το παράλληλο

[41] https://github.com/neo4j/neo4j-python-driver

[42] https://sbert.net/

κείμενο που χρησιμοποιήθηκε για την εκπαίδευση του student μοντέλου ακολουθώντας την τεχνική απόσταξης γνώσης, προήλθε από τις ίδιες πηγές που χρησιμοποιήθηκαν για την εκπαίδευση μοντέλου μηχανικής μετάφρασης (OPUS, Wikimatrix, Tatoeba) και είχε συνολικό μέγεθος 340MB. Το μοντέλο εκπαιδεύτηκε για 5 εποχές με batch size ίσο με 16 και με υποστήριξη κάρτας γραφικών NVIDIA GeForce RTX 3080 για 28 ώρες. Το σύνολο των υπερ-παραμέτρων και τα λεπτομερή στοιχεία της αρχιτεκτονικής δίνονται παρακάτω (Πίνακας 29). Το μοντέλο είναι ελεύθερα διαθέσιμο για χρήση μέσω του συνδέσμου: https://huggingface.co/lighteternal/stsb-xlm-r-greek-transfer

Πίνακας 20 Υπερ-παράμετροι εκπαιδευμένου μοντέλου σημασιολογικής ομοιότητας


```
{
  "_name_or_path": "xlm-roberta-base",
  "architectures": [ "XLMRobertaModel"],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "classifier_dropout": null,
  "eos_token_id": 2,
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "xlm-roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "torch_dtype": "float32",
  "transformers_version": "4.10.0",
  "type_vocab_size": 1,
  "use_cache": true,
  "vocab_size": 250002
}
```


4.3.1.5 Στοιχείο αναγνώρισης κειμενικής συνεπαγωγής

Η εκπαίδευση του μοντέλου Bi-Encoder που βασίζεται σε αρχιτεκτονική XLM-RoBERTa για την ταξινόμηση του είδους της σχέσης μεταξύ δύο προτάσεων (ισχυρισμού και τεκμηρίου) σε μία εκ των τριών δυνατών κλάσεων (συνεπαγωγή, αντίθεση, ουδετερότητα) υποστηρίχθηκε από την βιβλιοθήκη sentence-transformers, όπως παραπάνω. Ως σύνολο εκπαίδευσης, χρησιμοποιήθηκε το κείμενο που προέκυψε από την αγγλική και την ελληνική

έκδοση των επισημασμένων συνόλων SNLI[43] και MultiNLI[44], με την ελληνική έκδοση να προκύπτει με χρήση του μοντέλου μηχανικής μετάφρασης που περιεγράφηκε στην υποενότητα 4.2.1.1. Το συνδυαστικό σώμα κειμένων είχε μέγεθος 100MB. Η εκπαίδευση έγινε με batch size ίσο με 6 για μία εποχή και διήρκεσε 22 ώρες, με χρήση κάρτας γραφικών NVIDIA GeForce RTX 3080. Το σύνολο των υπερ-παραμέτρων και τα λεπτομερή στοιχεία της αρχιτεκτονικής δίνονται παρακάτω (Πίνακας 21). Το μοντέλο είναι ελεύθερα διαθέσιμο για χρήση μέσω του συνδέσμου: https://huggingface.co/lighteternal/nli-xlm-r-greek

Πίνακας 21 Υπερ-παράμετροι εκπαιδευμένου μοντέλου αναγνώρισης κειμενικής συνεπαγωγής


```
{
  "_name_or_path": "lighteternal/nli-xlm-r-greek/",
  "architectures": [    "XLMRobertaForSequenceClassification"  ],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "classifier_dropout": null,
  "eos_token_id": 2,
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "id2label": {
    "0": "contradiction",
    "1": "entailment",
    "2": "neutral"
  },
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "label2id": {
    "contradiction": 0,
    "entailment": 1,
    "neutral": 2
  },
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "xlm-roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "torch_dtype": "float32",
  "transformers_version": "4.10.0",
  "type_vocab_size": 1,
  "use_cache": true,
  "vocab_size": 250002
}
```


[43] https://nlp.stanford.edu/projects/snli/

[44] https://cims.nyu.edu/~sbowman/multinli/

4.3.1.6 Στοιχείο αναγνώρισης υποκειμενικότητας/αντικειμενικότητας

Έγινε εκπαίδευση δίγλωσσου (EN-EL) μοντέλου αρχιτεκτονικής XLM-RoBERTa με κεφαλή δυαδικής ταξινόμησης, με στόχο την αξιολόγηση της αντικειμενικότητας των διαθέσιμων τεκμηρίων ως προαιρετικό βήμα του σταδίου κατασκευής τεκμηρίων. Χρησιμοποιήθηκε επισημασμένο σύνολο δεδομένων 9.000 προτάσεων[45] διαθέσιμο στην αγγλική γλώσσα, το οποίο μεταφράστηκε στα ελληνικά. Το μοντέλο υλοποιήθηκε με χρήση της βιβλιοθήκης Transformers της Hugging Face, εκπαιδεύτηκε για 5 εποχές με batch size ίσο με 8 και με υποστήριξη κάρτας γραφικών NVIDIA GeForce RTX 3080 για 30 λεπτά. Ο Πίνακας 22 περιλαμβάνει το σύνολο των υπερ-παραμέτρων και τα λεπτομερή στοιχεία της αρχιτεκτονικής. Το μοντέλο είναι ελεύθερα διαθέσιμο για χρήση μέσω του συνδέσμου: https://huggingface.co/lighteternal/fact-or-opinion-xlmr-el.

Πίνακας 22 Υπερ-παράμετροι εκπαιδευμένου μοντέλου υποκειμενικότητας/αντικειμενικότητας


```
{
  "_name_or_path": "results/checkpoint-8500/",
  "architectures": [ "XLMRobertaForSequenceClassification"],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "classifier_dropout": null,
  "eos_token_id": 2,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "xlm-roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "problem_type": "single_label_classification",
  "torch_dtype": "float32",
  "transformers_version": "4.11.3",
  "type_vocab_size": 1,
  "use_cache": true,
  "vocab_size": 250002
}
```


4.3.2 Εφαρμογές και αποτελέσματα

Στις ακόλουθες υποενότητες παρουσιάζεται η περίπτωση χρήσης του μηχανισμού εξαγωγής σημασιολογικών συμπερασμάτων σε ελληνικές ειδησεογραφικές πηγές, καθώς και

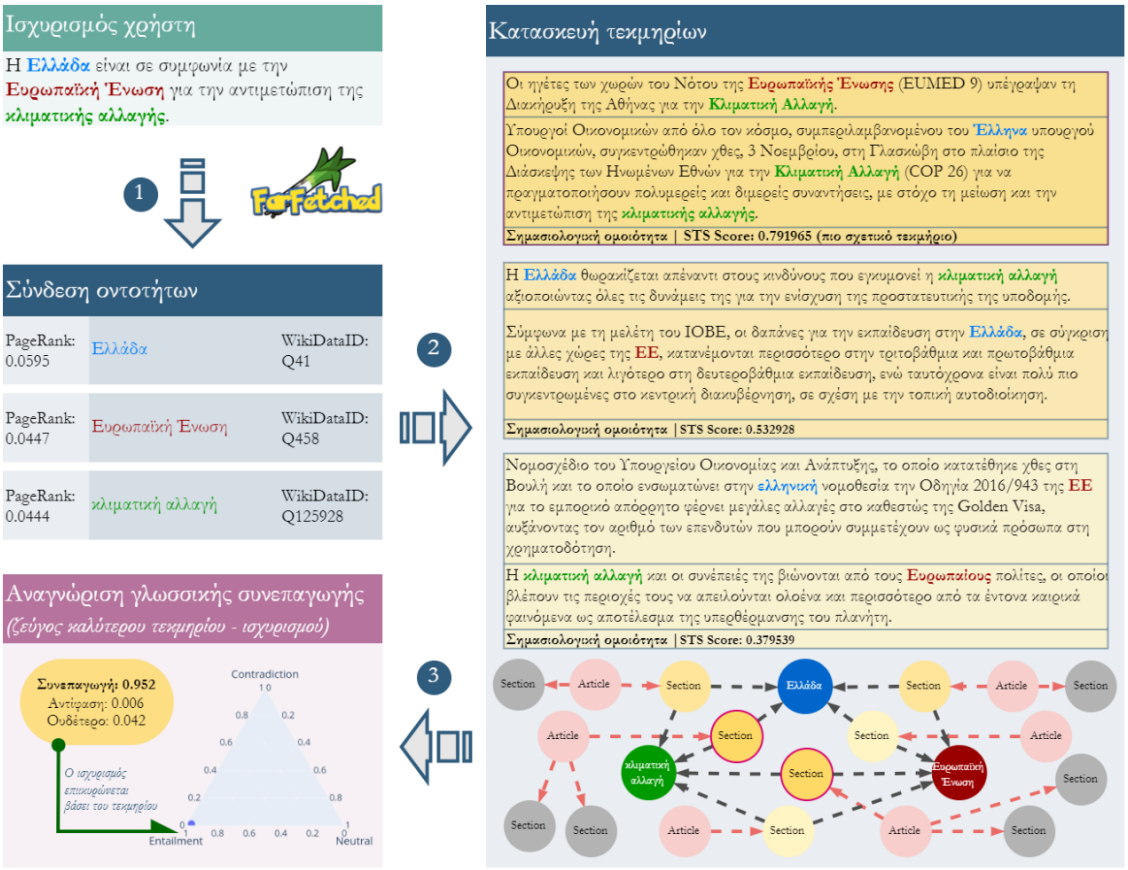
[45] https://github.com/1024er/cbert_aug/tree/crayon/datasets/subj

αποτελέσματα που αφορούν τόσο την απόδοση του συνολικού μηχανισμού όσο και των επιμέρους στοιχείων του. Παρατίθενται επίσης μεμονωμένα αποτελέσματα από την αξιολόγηση βοηθητικών στοιχείων που δεν συμπεριλήφθηκαν στη συγκεκριμένη περίπτωση χρήσης, βάσει σχετικών συνόλων δεδομένων αναφοράς (benchmark datasets).

4.3.2.1 Περίπτωση χρήσης: Επικύρωση ισχυρισμών από ελληνικές ειδησεογραφικές πηγές

Η συγκεκριμένη περίπτωση χρήσης αποτελεί το κύριο επιστέγασμα της παρούσας διδακτορικής εργασίας και επικεντρώνεται στην παρουσίαση ενός pipeline συστήματος για την επαλήθευση οποιουδήποτε κειμενικού ισχυρισμού, βάσει δεδομένων που αντλήθηκαν από ειδησεογραφικές πηγές. Συνδυάζει μια σειρά εργασιών για την περιοδική ιχνηλάτηση ειδησεογραφικών άρθρων, τη σύνδεσή τους με ονοματικές οντότητες και τη χρήση ενός κατάλληλου υποσυνόλου τους για κατασκευή τεκμηρίων που είτε επικυρώνουν είτε καταρρίπτουν τον ισχυρισμό ενός χρήστη. Καθώς η μεθοδολογία στοχεύει στην καλύτερη διαχείριση της πληροφοριακής "πλημμύρας" που μειώνει το συλλογικό εύρος προσοχής, προτείνεται ένα αυτοματοποιημένο συλλογιστικό πλαίσιο που βασίζεται στην ενοποίηση σχετικών αποσπασμάτων από πολλές διαφορετικές πηγές, έχοντας ως επίκεντρο τις ανιχνευθείσες ονοματικές οντότητες για την αποκάλυψη λανθανουσών συσχετίσεων μεταξύ γεγονότων, ενεργειών ή δηλώσεων. Απώτερος στόχος της μεθοδολογίας είναι η κάλυψη του κενού όσον αφορά την επικύρωση ισχυρισμών για γλώσσες χαμηλότερων πόρων και συγκεκριμένα για την ελληνική, επομένως η παρούσα περίπτωση χρήσης αφορά αμιγώς ελληνικά κείμενα. Ωστόσο, μπορεί να χρησιμοποιηθεί για οποιαδήποτε άλλη γλώσσα, με την προϋπόθεση ότι τα συνιστώντα μέρη της αρχιτεκτονικής (κυρίως τα αντίστοιχα μοντέλα βαθιάς μάθησης) είναι διαθέσιμα ή υπάρχουν πόροι για την εκπαίδευσή τους.

Η προσέγγιση που ακολουθείται στη συγκεκριμένη εργασία παρουσιάζει μια αρχιτεκτονική (FarFetched: An Entity-centric Reasoning and Claim Validation Framework based on Textually Represented Environments), η οποία ενσωματώνει τα περισσότερα από τα στοιχεία που αναπτύχθηκαν στην υποενότητα 4.3.1 και συγκεκριμένα: το στοιχείο ιχνηλάτησης HTML σελίδων και τη βάση δεδομένων γράφων, το στοιχείο αυτόματης αντιστοίχισης οντοτήτων με έννοιες της Wikipedia, το στοιχείο κατασκευής τεκμηρίων και τα στοιχεία ελέγχου σημασιολογικής ομοιότητας και αναγνώρισης κειμενικής συνεπαγωγής. Ο κώδικας που χρησιμοποιήθηκε για την υλοποίηση της συγκεκριμένης περίπτωσης χρήσης είναι διαθέσιμος στον παρακάτω σύνδεσμο:
https://github.com/lighteternal/FarFetched_ACL/



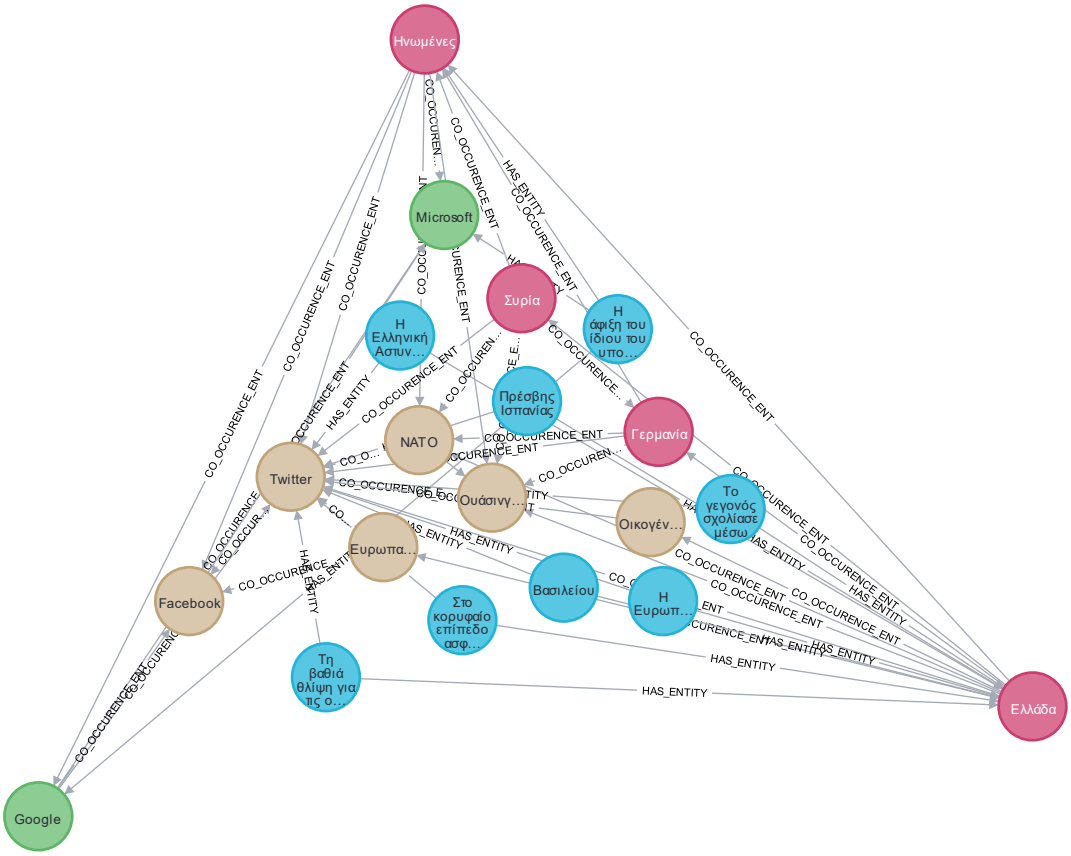
Σχήμα 39 Παράδειγμα ελέγχου ισχυρισμού χρήστη

Η μεθοδολογία απεικονίζεται συνοπτικά στο Σχήμα 39 και περιλαμβάνει τους μηχανισμούς και τις ροές που περιεγράφηκαν στην υποενότητα 3.4.1. Δεδομένου ενός ισχυρισμού σε ελεύθερο κείμενο, το FarFetched χρησιμοποιεί το στοιχείο σύνδεσης οντοτήτων για να ανιχνεύσει τις ονοματικές οντότητες σε αυτόν. Έπειτα μέσω του μηχανισμού κατασκευής τεκμηρίων, ανατρέχει στη βάση και εξάγει ένα υποσύνολο αποσπασμάτων (υποψήφιων τεκμηρίων) από το αποθηκευμένο ειδησεογραφικό περιεχόμενο που αναφέρεται στις ίδιες ονοματικές οντότητες που υπάρχουν στον ισχυρισμό. Επισημαίνεται ότι κάθε υποψήφιο τεκμήριο μπορεί να αποτελείται από συνδυασμό πολλών διαφορετικών αποσπασμάτων, αρκεί να περιέχονται σε αυτό όλες οι οντότητες που συναντώνται στον ισχυρισμό. Στη συνέχεια, επιλέγεται το πιο σημασιολογικά συναφές τεκμήριο ως προς τον ισχυρισμό, με χρήση του στοιχείου σημασιολογικής ομοιότητας και τέλος το στοιχείο αναγνώρισης κειμενικής συνεπαγωγής αναλαμβάνει να κρίνει εάν το σχετικότερο τεκμήριο επικυρώνει, απορρίπτει ή είναι ουδέτερο προς τον ισχυρισμό.

Αξίζει να επισημανθεί ότι η προτεινόμενη προσέγγιση διαφέρει σε αρκετά σημεία από τα περισσότερα συστήματα ελέγχου ισχυρισμών, επισκόπηση των οποίων έγινε στην υποενότητα 2.8.4. Αρχικά, η συλλογή τεκμηρίων στο FarFetched είναι πλήρως απεμπλεγμένη από την ανθρώπινη εμπειρογνωμοσύνη, αλλά βασίζεται στη συνεχώς ανανεούμενη ροή ειδήσεων. Επίσης, η επικύρωση ισχυρισμών μέσω των επιλεχθέντων τεκμηρίων βασίζεται στον αποτελεσματικό συνδυασμό στοιχείων σύνδεσης οντοτήτων και νευρωνικών μοντέλων σημασιολογικής ομοιότητας και κειμενικής συνεπαγωγής για την ελληνική γλώσσα που χρησιμοποιούν μηχανισμούς προσοχής. Πέραν του τελικού αποτελέσματος, η μέθοδος προσφέρει τη δυνατότητα ερμηνείας της συλλογιστικής που ακολουθήθηκε, επιστρέφοντας τα τεκμήρια που οδήγησαν στην επαλήθευση ή απόρριψη του ισχυρισμού ενός χρήστη, χωρίς ωστόσο να ελέγχει την ακεραιότητά τους (τα δεδομένα που εξάγονται από ειδησεογραφικές πηγές θεωρούνται αληθή). Τέλος, το αποτέλεσμα της διαδικασίας είναι δυναμικό: η συνεχής ενσωμάτωση νέων πληροφοριών που αφορά συγκεκριμένες οντότητες μπορεί να οδηγήσει στην αλλαγή της ετυμηγορίας ενός προηγουμένως επικυρωμένου ισχυρισμού.

Στα πλαίσια της συγκεκριμένης περίπτωσης χρήσης, αντλήθηκαν 13.326 ειδήσεις μέσω του στοιχείου ιχνηλάτησης HTML σελίδων από δύο γνωστούς ειδησεογραφικούς ιστοτόπους, οι οποίες καλύπτουν ένα ευρύ θεματικό φάσμα από το 2018 μέχρι το 2021. Στη συνέχεια, μέσω του στοιχείου σύνδεσης οντοτήτων με έννοιες της Wikipedia, ανιχνεύθηκαν 2.516 έννοιες διαφορετικών τύπων (κράτη, πόλεις, άτομα, οργανισμοί, εταιρείες κτλ.). Μέσω της παραπάνω διαδικασίας, οι οντότητες λειτουργούν ως "συνδετικοί κρίκοι" μεταξύ διαφορετικών ειδήσεων, επιτρέποντας το συνδυασμό τους από το στοιχείο κατασκευής τεκμηρίων. Ένα παράδειγμα επισημασμένων αποσπασμάτων στη γνωσιακή βάση φαίνεται στο Σχήμα 40, όπου είναι εμφανή (ως μπλε κόμβοι) τα αποσπάσματα ειδήσεων που ενώνουν τις οντότητες "Ελλάδα", "Ηνωμένες Πολιτείες της Αμερικής" και "Google".

Ακολουθεί η ποσοτική και ποιοτική αξιολόγηση της προτεινόμενης προσέγγισης συνολικά, παρέχοντας παράλληλα επιμέρους αποτελέσματα για την απόδοση των δύο μοντέλων που εκπαιδεύτηκαν στα πλαίσια της εργασίας ώστε να υποστηρίξουν τις διαδικασίες ελέγχου σημασιολογικής ομοιότητας (STS) και αναγνώρισης κειμενικής συνεπαγωγής (NLI).



Σχήμα 40 Παράδειγμα σύνδεσης αποσπασμάτων ειδήσεων μέσω των οντοτήτων στις οποίες αναφέρονται

Δεδομένης της ιδιαιτερότητας της περιγραφείσας μεθοδολογίας όσον αφορά τη συλλογή τεκμηρίων (δυναμική ενημέρωση περιεχομένου μέσω διαδικτυακών πηγών), η συγκριτική αξιολόγησή της με άλλα συστήματα ελέγχου ισχυρισμών που βασίζονται σε στατική γνώση (πχ. από Wikipedia) παρουσιάζει αρκετές προκλήσεις. Παράλληλα, όπως και σε προηγούμενες εργασίες αξιολόγησης, είναι χαρακτηριστική η απουσία συνόλων δεδομένων αναφοράς (benchmark datasets) για την ελληνική γλώσσα. Προκειμένου να καταπολεμηθούν οι παραπάνω δυσκολίες, η ποσοτική αξιολόγηση της συνολικής μεθοδολογίας έγινε με χρήση του συνόλου δεδομένων αναφοράς FEVER (Thorne et al., 2018), το οποίο μοντελοποιεί την αξιολόγηση της εγκυρότητας κειμενικών ισχυρισμών (για την αγγλική γλώσσα) ως τη συνδυαστική εργασία ανάκτησης πληροφοριών από τη Wikipedia με σκοπό την αναγνώριση της κειμενικής συνεπαγωγής τους από τα δεδομένα. Κάθε σειρά του benchmark περιλαμβάνει έναν ισχυρισμό σε ελεύθερο κείμενο, μια λίστα σχετικών τεκμηρίων συμπεριλαμβανομένου και του Wikipedia URL που περιλαμβάνει τις αντίστοιχες πληροφορίες, καθώς και το αποτέλεσμα ελέγχου του ισχυρισμού, το οποίο ανήκει σε μια από τις ακόλουθες κλάσεις: {SUPPORTS, REFUTES, NOT ENOUGH INFO}. Στα πλαίσια της εργασίας, μεταφράστηκε ένα υποσύνολο 150 ισχυρισμών από το

υποσύνολο επικύρωσης του FEVER, ενώ παράλληλα η γνωσιακή βάση εμπλουτίστηκε με το σύνολο των αποσπασμάτων που περιείχαν οι αντίστοιχοι σύνδεσμοι της Wikipedia, προσομοιάζοντας τη διαδικασία ιχνηλάτησης από ειδησεογραφικές πηγές. Για τη μετάφραση του Wikipedia περιεχομένου χρησιμοποιήθηκε το EN-EL μοντέλο μηχανικής μετάφρασης που περιεγράφηκε στην υποενότητα 4.2.1.1. Η απόδοση υπολογίζεται ως προς την ακρίβεια, την ανάκληση και το F1-score, τόσο συνδυαστικά για όλες τις κλάσεις, όσο και για κάθε κλάση χωριστά (Πίνακας 23).

Πίνακας 23 Απόδοση μηχανισμού επικύρωσης ισχυρισμών (FarFetched) στο FEVER benchmark

Κλάση	Precision	Recall	F1
NOT ENOUGH INFO	.36	.80	.49
REFUTES	.91	.72	.80
SUPPORTS	.84	.70	.76
Σταθμισμένος μέσος	**.82**	**.73**	**.75**
Accuracy		**.73**	

Τα παραπάνω αποτελέσματα δείχνουν ότι η ακρίβεια και η ανάκληση των κλάσεων απόρριψης (REFUTES) και επικύρωσης (SUPPORTS) ενός ισχυρισμού βρίσκονται σε ισορροπία μεταξύ τους, σε αντίθεση με την ουδέτερη κλάση (NOT ENOUGH INFO) η οποία παρουσιάζει εμφανώς χαμηλότερη ακρίβεια. Αυτό μπορεί να αποδοθεί εν μέρει στις προκλήσεις που εμφανίζει η διαδικασία σύνδεσης οντοτήτων σε κείμενα που αποτελούν προϊόν μηχανικής μετάφρασης, με αποτέλεσμα μερικοί ισχυρισμοί να μην μπορούν να αντιστοιχηθούν με σχετικά τεκμήρια, μέσω του μηχανισμού κατασκευής τεκμηρίων. Όσον αφορά τη συγκριτική αξιολόγηση της μεθόδου, η απευθείας σύγκριση με άλλα συστήματα θα προϋπέθετε τη χρήση τους στο ελληνικό ισοδύναμο του πρωτότυπου FEVER dataset. Δεδομένου ότι κάτι τέτοιο δεν είναι εφικτό, αξιοποιούνται τα αποτελέσματα μεγάλης συγκριτικής μελέτης που αφορά την αξιολόγηση αντίστοιχων συστημάτων ελέγχου ισχυρισμών στο αγγλικό FEVER dataset (Bekoulis et al., 2020) σε επίπεδο ορθότητας (accuracy). Από αυτήν προκύπτει σημαντικό κέρδος στην απόδοση έναντι της βασικής (baseline) προσέγγισης των Thorne et al. (accuracy = 0.45), ενώ παράλληλα το FarFetched με accuracy ίσο με 0.73 κατατάσσεται στο καλύτερο 30% μεταξύ 31 συστημάτων (με τις τιμές accuracy των οποίων να κυμαίνονται από 0.45 έως 0.84). Επισημαίνεται ωστόσο, ότι κανένα από τα συγκρινόμενα συστήματα επικύρωσης ισχυρισμών δεν καλύπτει την ελληνική γλώσσα.

Συμπληρωματικά με τα παραπάνω αποτελέσματα, πραγματοποιείται μια ποιοτική επισκόπηση των αποτελεσμάτων της εργασίας μέσω τριών διαφορετικών σεναρίων, καθένα εκ των οποίων αποτελείται από δύο μέρη. Στόχος της επισκόπησης είναι η ανάδειξη των δυνατοτήτων που προσφέρει η προτεινόμενη μεθοδολογία, αναγνωρίζοντας παράλληλα τη δυναμικότητα της διαδικασίας συγκέντρωσης και κατασκευής τεκμηρίων, η οποία δεν είναι δυνατόν να αποτιμηθεί μέσω στατικών benchmarks όπως το παραπάνω.

Στο Σενάριο 1 (Πίνακας 24), ο χρήστης εισάγει δύο αντικρουόμενους ισχυρισμούς (1a, 1b) οι οποίοι αφορούν τις ίδιες οντότητες (Δανία, Αυστρία, ΕΕ). Επομένως, το στοιχείο κατασκευής τεκμηρίων θα επιστρέψει τις ίδιες ακολουθίες υποψήφιων τεκμηρίων και στις δύο περιπτώσεις. Τα υποψήφια τεκμήρια (υποσύνολο των οποίων φαίνεται στον σχετικό πίνακα) κατατάσσονται ως προς τη φθίνουσα σημασιολογική τους ομοιότητα με τον εκάστοτε ισχυρισμό. Τέλος επιλέγεται το πιο σχετικό τεκμήριο (STS Score: 0.8505), το οποίο συγκρίνεται με τον κάθε ισχυρισμό, μέσω του στοιχείου κειμενικής συνεπαγωγής. Από τα αντίστοιχα scores για e:entailment, c:contradiction και n:neutrality, είναι εμφανές ότι το σχετικότερο τεκμήριο ορθώς επικυρώνει τον ισχυρισμό 1a και απορρίπτει των ισχυρισμό 1b.

Πίνακας 24 Πρώτο σενάριο επικύρωσης ισχυρισμών μέσω της μεθοδολογίας FarFetched

Ισχυρισμός χρήστη (Σενάριο 1)	**NLI score**
Η Δανία και η Αυστρία πιστεύουν ότι η Ευρωπαϊκή Ένωση πρέπει να αυξήσει τη βοήθεια προς τους πρόσφυγες. **(1a)**	c: 0.014 **e: 0.958** n: 0.028
Η Δανία διαφωνεί με την Αυστρία σχετικά με τη διαχείριση των μεταναστευτικών θεμάτων στην Ευρωπαϊκή Ένωση. (**1b)**	**c: 0.951** e: 0.002 n: 0.047
Υποψήφια τεκμήρια από διαδικτυακές πηγές (↓ σημασιολογική ομοιότητα)	
Η Αυστρία και η Δανία θέλουν να ενισχυθεί επίσης η υποστήριξη της ΕΕ προς κράτη που υποδέχονται πρόσφυγες κοντά σε εστίες κρίσεις, ώστε οι πρόσφυγες αυτοί να μην ταξιδεύουν προς την Ευρώπη. • **STS Score: 0.8505**	
Σε έλεγχο από αστυνομικούς των Αστυνομικών Τμημάτων Αερολιμένων ... οι αλλοδαποί επέδειξαν πλαστά ταξιδιωτικά έγγραφα προκειμένου να αναχωρήσουν από τη χώρα για άλλες χώρες της ΕΕ όπως η Γαλλία, Γερμανία, Ιταλία, Αυστρία, Ολλανδία, Δανία, Ισπανία και Νορβηγία. • **STS Score: 0.2283**	

Στο Σενάριο 2 (Πίνακας 25) διερευνάται η ευαισθησία της μεθοδολογίας κατά την τροφοδοσία του συστήματος με νέα δεδομένα που ενδέχεται να επηρεάσουν την τελική ετυμηγορία. Συγκεκριμένα, οι οντότητες που ανιχνεύονται στον ισχυρισμό (ΗΠΑ, Ιράν) πυροδοτούν το στοιχείο κατασκευής τεκμηρίων, το οποίο επιστρέφει μια σειρά από υποψήφια τεκμήρια (πορτοκαλί χρώμα) και τα κατατάσσει κατά φθίνουσα σημασιολογική

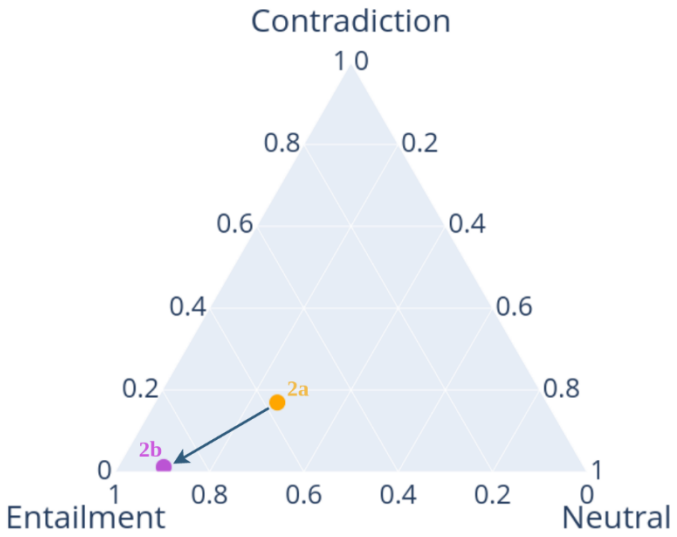
ομοιότητα. Βάσει του σχετικότερου τεκμηρίου (με STS Score: 0.6665), ο ισχυρισμός επικυρώνεται μεν από τα δεδομένα, αλλά με σχετικά χαμηλή πιθανότητα ταξινόμησης για την κλάση e:entailment (2a). Ο ίδιος ισχυρισμός ελέγχεται πάλι στο Σενάριο 2b, μετά την προσθήκη νέας πληροφορίας (με μπλε χρώμα) η οποία προσαρτάται στα υπάρχοντα τεκμήρια. Καθώς η σημασιολογική ομοιότητα του νέου τεκμηρίου είναι μεγαλύτερη των υπαρχόντων, επιλέγεται ως σχετικότερο και τροφοδοτείται στο στοιχείο κειμενικής συνεπαγωγής μαζί με τον ισχυρισμό, οδηγώντας σε επικύρωσή του αλλά με σαφώς μεγαλύτερη βεβαιότητα σε σχέση με πριν. Αυτή η μετατόπιση στην ετυμηγορία, η οποία εκφράζεται μέσω των πιθανοτήτων ταξινόμησης για καθεμία από τις τρεις κλάσεις απεικονίζεται στο Σχήμα 41.

Πίνακας 25 Δεύτερο σενάριο επικύρωσης ισχυρισμών μέσω της μεθοδολογίας FarFetched

Ισχυρισμός χρήστη (Σενάριο 2)	NLI score (2a)	NLI score (2b)
Οι Ηνωμένες Πολιτείες σχεδιάζουν να επιβάλλουν κυρώσεις στο Ιράν.	c: 0.170 **e: 0.571** n: 0.259	c: 0.012 **e: 0.891** n: 0.097
Υποψήφια τεκμήρια από διαδικτυακές πηγές (↓ σημασιολογική ομοιότητα)		
Το Ιράν μπροστά στο δίλημμα αν θα συμμορφωθεί προς τις υποδείξεις της Ουάσιγκτον ή θα οδηγηθεί σε κατάρρευση Οι κυρώσεις που επανήλθαν σε ισχύ σήμερα, θα αναγκάσουν την κυβέρνηση της Ισλαμικής Δημοκρατίας να δεχθεί τις αξιώσεις των ΗΠΑ όσον αφορά το ιρανικό πυρηνικό πρόγραμμα και τις ιρανικές δραστηριότητες στην περιοχή της Μέσης Ανατολής διότι, σε διαφορετική περίπτωση, το καθεστώς θα κινδυνεύσει να καταρρεύσει, υποστήριξε ο Ισραέλ Κατς, ο ισραηλινός υπουργός αρμόδιος για τις Υπηρεσίες Πληροφοριών. **• STS Score: 0.6665**		
Γιατί εξαιρέθηκε η Ελλάδα από τις αμερικανικές κυρώσεις στο Ιράν. Από τις 5 Νοεμβρίου βρίσκονται σε ισχύ οι νέες κυρώσεις των ΗΠΑ για εξαγωγές πετρελαίου από το Ιράν. **• STS Score: 0.6324**		
«Είμαστε πάντα υπέρ της διπλωματίας και των συνομιλιών ... Όμως οι συνομιλίες χρειάζονται εντιμότητα ... Οι ΗΠΑ επιβάλλουν εκ νέου κυρώσεις στο Ιράν και αποσύρονται από την πυρηνική συμφωνία (του 2015) και μετά θέλουν να κάνουν συνομιλίες μαζί μας», δήλωσε ο Ροχανί σε ομιλία του που μεταδόθηκε ζωντανά από την τηλεόραση. **• STS Score: 0.5151**		
Δυναμιτίζει το κλίμα μεταξύ ΗΠΑ - Τουρκίας η καταδίκη του τραπεζίτη Μεχμέτ Ατίλα. Τις ήδη τεταμένες σχέσεις μεταξύ Τουρκίας και ΗΠΑ δυναμιτίζει η απόφαση του ομοσπονδιακού δικαστηρίου του Μανχάταν, το οποίο έκρινε την Τετάρτη ένοχο τον Τούρκο τραπεζίτη Μεχμέτ Χακάν Ατίλα για συμμετοχή σε συνωμοσία με στόχο να προσφερθεί βοήθεια στο Ιράν για να παρακάμψει τις αμερικανικές οικονομικές κυρώσεις. **• STS Score: 0.4018**		
Μετά το ναυάγιο των τελευταίων συνομιλιών μεταξύ ΗΠΑ και Ιράν αναμένεται η ανακοίνωση επιπλέον κυρώσεων τις επόμενες ημέρες. **• STS Score: 0.7195**		

Δεδομένου ότι η προτεινόμενη αρχιτεκτονική καλύπτει τον περιοδικό εμπλουτισμό της γνωσιακής βάσης με νέες πληροφορίες, η παρακολούθηση αντίστοιχων μεταβολών θα μπορούσε να αποδειχθεί χρήσιμη για τον εντοπισμό και την πρόβλεψη τάσεων, ειδικότερα

για περιπτώσεις που επωφελούνται από μακροπρόθεσμο σχεδιασμό (πχ. επιχειρηματικές αποφάσεις, αγοραστικές προτιμήσεις, δημοφιλία, κοινωνική/πολιτική επιρροή κτλ.)



Σχήμα 41 Μεταβολή αποτελεσμάτων αναγνώρισης κειμενικής συνεπαγωγής στο Σενάριο 2

Το Σενάριο 3 (Πίνακας 26) είναι παρόμοιο με το Σενάριο 2, καθώς ο ίδιος ισχυρισμός ελέγχεται βάσει ενός αρχικού συνόλου πληροφοριών (3a), ενώ οι νέες πληροφορίες που εισάγονται μετέπειτα στη βάση μεταβάλουν την ετυμηγορία του στοιχείου κειμενικής συνεπαγωγής (3b). Ωστόσο, σε αυτή την περίπτωση η νέα πληροφορία αποτελεί απόσπασμα συνέντευξης ατόμου και ενδέχεται να χαρακτηρίζεται από μεγαλύτερη υποκειμενικότητα σε σχέση με τα υπόλοιπα τεκμήρια. Ενώ η παρούσα περίπτωση εργασίας δε διακρίνει μεταξύ απόψεων και γεγονότων, το στοιχείο αναγνώρισης υποκειμενικότητας/αντικειμενικότητας που αξιολογείται χωριστά στην υποενότητα 4.3.2.3 μπορεί να χρησιμοποιηθεί για το διαχωρισμό υποκειμενικών και αντικειμενικών προτάσεων.

Πίνακας 26 Τρίτο σενάριο επικύρωσης ισχυρισμών μέσω της μεθοδολογίας FarFetched

Ισχυρισμός χρήστη (Σενάριο 3)	**NLI score (3a)**	**NLI score (3b)**
Η Apple προσπαθεί να ανταγωνιστεί την Netflix στην παραγωγή τηλεοπτικού περιεχομένου.	c: 0.004 **e: 0.967** n: 0.029	**c: 0.982** e: 0.008 n: 0.010
Υποψήφια τεκμήρια από διαδικτυακές πηγές (↓ σημασιολογική ομοιότητα)		
Η Apple αναμένεται να δαπανήσει φέτος περίπου 2 δισεκατομμύρια δολάρια με σκοπό τη δημιουργία πρωτότυπου περιεχομένου που ελπίζει ότι θα ανταγωνιστεί τις ήδη εδραιωμένες στο τηλεοπτικό κοινό υπηρεσίες των Netflix, Hulu και Amazon. • **STS Score: 0.7107**		
«Δεν προσπαθούμε να ανταγωνιστούμε το Netflix στην τηλεόραση», δήλωσε εκπρόσωπος της Apple σε συνέντευξή του. • **STS Score: 0.7134**		

Η απόδοση του εκπαιδευμένου μοντέλου σημασιολογικής ομοιότητας αξιολογείται χωριστά στο υποσύνολο ελέγχου του συνόλου δεδομένων αναφοράς STS2017 (Cer et al., 2018), το οποίο περιλαμβάνει 1.379 ζεύγη προτάσεων στην αγγλική γλώσσα που συνοδεύονται από τη μέση τιμή του similarity score (0-5), όπως αυτό αξιολογήθηκε από πέντε ανεξάρτητους βαθμολογητές. Καθώς το συγκεκριμένο benchmark δεν παρέχεται για την ελληνική γλώσσα, δημιουργήθηκε η δίγλωσση (EN-EL) εκδοχή του με χρήση του εκπαιδευμένου μοντέλου μηχανικής μετάφρασης. Η απόδοση μετράται με χρήση των συντελεστών συσχέτισης Pearson r και Spearman ρ, συγκρίνοντας τα similarity scores των βαθμολογητών και του μοντέλου. Ακόμη, δίνονται αποτελέσματα που αφορούν την ορθότητα μεταφραστικής αντιστοίχισης μέσω της ομοιότητας συνημιτόνου των διανυσματικών αναπαραστάσεων πηγής (EN) και στόχου (EL) που αφορούν το teacher και student μοντέλο αντίστοιχα. Τα αποτελέσματα φαίνονται παρακάτω (Πίνακας 27). Το παραχθέν μοντέλο έχει ελαφρώς καλύτερη απόδοση από το κορυφαίο (πολύγλωσσο) μοντέλο των (Reimers and Gurevych, 2020b) σε όλα τα μέτρα αξιολόγησης.

Πίνακας 27 Συγκριτική αξιολόγηση μοντέλων σημασιολογικής ομοιότητας στο STS2017 benchmark test set (EN-EL version)

Μοντέλο	STS 2017		Translation matching acc.	
	r	ρ	cos(en2el)	cos(el2en)
STS-XLM-R-Greek (Ours)	**83.30**	**84.32**	**98.05**	**97.80**
Paraphrase- multilingual-mpnet- base-v2 (UKP-TUDA)	82.71	82.70	97.50	97.35

Η διαδικασία επαναλαμβάνεται και για το εκπαιδευμένο μοντέλο αναγνώρισης κειμενικής συνεπαγωγής, με χρήση του υποσυνόλου ελέγχου του πολύγλωσσου XNLI benchmark dataset (Conneau et al., 2018). Το benchmark περιλαμβάνει 5.010 ζεύγη υποθέσεων-προϋποθέσεων, συνοδευόμενα από την αντίστοιχη κλάση (entailment, contradiction ή neutral) μετά από αξιολόγηση εμπειρογνωμόνων. Τα αποτελέσματα παρατίθενται στον επόμενο πίνακα (Πίνακας 28) ως προς το F1-score της ταξινόμησης. Παρότι το μοντέλο δεν εκπαιδεύτηκε στο υποσύνολο εκπαίδευσης του XNLI, έχει καλύτερες επιδόσεις σε σχέση με τα αντίστοιχα πολύγλωσσα μοντέλα των Conneau et al. (Facebook) και Hu et al., 2020 (Google) και είναι στο ίδιο επίπεδο με το μονόγλωσσο ελληνικό μοντέλο των Koutsikakis et al. (ΟΠΑ). Ωστόσο, καθώς το μοντέλο μας εκπαιδεύτηκε σε συνδυασμό

αγγλικών και ελληνικών προτάσεων, μπορεί να προτιμηθεί σε περιπτώσεις ανάλυσης δίγλωσσου περιεχομένου (πχ. τεχνολογία, επιστήμη) χωρίς να πάσχει από την υπο-αντιπροσωπευτικότητα της ελληνικής γλώσσας σε πολύγλωσσα μοντέλα.

Πίνακας 28 Συγκριτική αξιολόγηση μοντέλων αναγνώρισης κειμενικής συνεπαγωγής στο XNLI benchmark test set (EL version)

Μοντέλο	F1-score
NLI-XLM-R Greek (Ours)	**78.3**
Greek-BERT (AUEB)	78.6 $\pm$ 0.62
XLM-RoBERTa-base (Facebook)	77.3 $\pm$ 0.41
M-BERT (Google AI Language)	73.5 $\pm$ 0.49

Η παραπάνω αξιολόγηση ολοκληρώνει την επισκόπηση της τεχνικής υλοποίησης και των αποτελεσμάτων για τη συγκεκριμένη περίπτωση χρήσης. Ακολουθεί η αξιολόγηση μεμονωμένων στοιχείων, τα οποία λειτουργούν συμπληρωματικά στην υφιστάμενη αρχιτεκτονική και προστέθηκαν στη συνολική μεθοδολογία σε μεταγενέστερο στάδιο. Οι περιορισμοί και τα συμπεράσματα που προέκυψαν από την περιγραφείσα υλοποίηση συγκεντρώνονται στο Κεφάλαιο 5.

4.3.2.2 Αξιολόγηση στοιχείου αναγνώρισης ομιλίας

Η αξιολόγηση της απόδοσης του στοιχείου αναγνώρισης ομιλίας έγινε μεμονωμένα, με χρήση των μέτρων Word Error Rate (WER) και Character Error Rate (CER) τα οποία βασίζονται στην απόσταση Levenstein (Levenshtein, 1966) και υποδεικνύουν το ποσοστό λανθασμένων λέξεων και χαρακτήρων αντίστοιχα μιας απομαγνητοφώνησης (πρότασης-υπόθεσης) σε σχέση με μια πρόταση αναφοράς. Για τη μέτρηση της απόκλισης μεταξύ των δύο προτάσεων, λαμβάνονται υπόψη τριών ειδών σφάλματα, τα οποία αφορούν είτε ολόκληρες λέξεις (για το WER) είτε μεμονωμένους χαρακτήρες (για το CER), και συγκεκριμένα:

- εισαγωγές (insertions – I): λέξεις/χαρακτήρες που προέκυψαν από το στοιχείο αναγνώρισης ομιλίας και βρίσκονται στην πρόταση-υπόθεση χωρίς να έχουν ειπωθεί (δεν βρίσκονται στην πρόταση αναφοράς),
- αφαιρέσεις (deletions – D): λέξεις/χαρακτήρες που δεν αναγνωρίστηκαν από το στοιχείο αναγνώρισης ομιλίας και ως εκ τούτου δεν περιλαμβάνονται στην πρόταση-υπόθεση αλλά υπάρχουν στην πρόταση αναφοράς,

- αντικαταστάσεις (substitutions – S): λέξεις/χαρακτήρες που αναγνωρίστηκαν λανθασμένα από το στοιχείο αναγνώρισης ομιλίας με αποτέλεσμα να είναι αντικαθιστούν τις ορθές στην πρόταση αναφοράς.

Βάσει των παραπάνω, οι σχέσεις υπολογισμού για τα WER, CER βασίζονται στο πλήθος καθενός από τα παραπάνω σφάλματα σε σχέση με τον συνολικό αριθμό λέξεων ή χαρακτήρων N αντίστοιχα:

$$WER = 100 \cdot \frac{count(I_{word}) + count(D_{word}) + count(S_{word})}{N_{word}}\% \qquad 4.5$$

$$CER = 100 \cdot \frac{count(I_{char}) + count(D_{char}) + count(S_{char})}{N_{char}}\% \qquad 4.6$$

Για την αξιολόγηση χρησιμοποιήθηκε το απομαγνητοφωνημένο υποσύνολο δεδομένων ελέγχου (test set) της CommonVoice. Τα αποτελέσματα παρατίθενται στον παρακάτω πίνακα:

Πίνακας 29 Απόδοση εκπαιδευμένου μοντέλου αναγνώρισης φωνής για την ελληνική γλώσσα στο Common Voice test set

Μέτρο	Τιμή
WER (%)	10.498
CER (%)	2.875

Από τα αποτελέσματα γίνεται εύκολα κατανοητό ότι το μοντέλο έχει ιδιαίτερα ικανοποιητική απόδοση στο συγκεκριμένο σύνολο δεδομένων ελέγχου, καθώς περίπου 9 στις 10 λέξεις αναγνωρίζονται επιτυχώς. Αυτό επιβεβαιώθηκε και στα πλαίσια σχετικού διαγωνισμού, καθώς το μοντέλο κατέκτησε την πρώτη θέση μεταξύ 6 συμμετοχών, εμφανίζοντας το μικρότερο ποσοστό λανθασμένων λέξεων και χαρακτήρων. Ωστόσο, αξίζει να σημειωθεί ότι το σύνολο εκπαίδευσης που χρησιμοποιήθηκε (και κατ' επέκταση το σύνολο ελέγχου) βασίζονται σε σχετικά μικρό εύρος φωνητικών δεδομένων (<2 GB) με περιορισμένη ποικιλομορφία, σε σύγκριση με αντίστοιχα μοντέλα για γλώσσες υψηλών πόρων (>30 GB).

4.3.2.3 Αξιολόγηση στοιχείου αναγνώρισης υποκειμενικότητας/αντικειμενικότητας

Το στοιχείο που υλοποιήθηκε σύμφωνα με τις τεχνικές προδιαγραφές της υποενότητας 4.3.1.6, εξετάστηκε ως προς τα συνήθη μέτρα απόδοσης που συνοδεύουν μοντέλα ταξινόμησης, και συγκεκριμένα ως προς την ορθότητα (accuracy), την ακρίβεια (precision), την ανάκληση (recall) και το F1-score. Για το σκοπό αυτό χρησιμοποιήθηκε υποσύνολο ελέγχου (test set) 2.000 παραδειγμάτων (στην ελληνική και στην αγγλική γλώσσα), το οποίο

προήλθε από την ίδια πηγή με τα δεδομένα εκπαίδευσης. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα (Πίνακας 30) και κρίνονται ως ιδιαίτερα ικανοποιητικά, καθώς οι υψηλές τιμές συνοδεύονται από ισορροπία μεταξύ ακρίβειας και ανάκλησης. Ωστόσο, σημειώνεται ότι η αξιολόγηση αφορά σχετικά μικρό σύνολο δεδομένων περιορισμένου θεματικού εύρους (οι προτάσεις αφορούσαν κυρίως πληροφορίες ή κριτικές ταινιών), επομένως δεν αποκλείεται η περίπτωση υπερπροσαρμογής (overfitting), που ενδέχεται να οδηγήσει σε χειρότερη απόδοση κατά την ταξινόμηση προτάσεων διαφορετικής θεματολογίας (πχ. πολιτική). Παρόλα αυτά, η ελεύθερη διάθεση του τελικού checkpoint του μοντέλου επιτρέπει τη συνέχιση της εκπαίδευσής του με νέα δεδομένα, εφόσον καταστούν διαθέσιμα.

Πίνακας 30 Απόδοση εκπαιδευμένου μοντέλου αναγνώρισης υποκειμενικότητας/αντικειμενικότητας

Μέτρο	**Τιμή**
Accuracy	0.952
Precision	0.945
Recall	0.960
F1-score	0.952

4.4 Σύνοψη κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκαν οι τεχνολογικές επιλογές (βιβλιοθήκες, υπερπαράμετροι. σύνολα δεδομένων εκπαίδευσης κτλ.) που αφορούν τα επιμέρους στοιχεία των δύο μηχανισμών που υλοποιήθηκαν στα πλαίσια της διατριβής, σύμφωνα με τη μεθοδολογία που αναπτύχθηκε στο Κεφάλαιο 3. Παράλληλα, αναδείχθηκαν οι δυνατότητες των δύο μηχανισμών στα πλαίσια μιας σειράς περιπτώσεων χρήσης, δίνοντας έμφαση τόσο στην ποσοτική όσο και στην ποιοτική αξιολόγηση της απόδοσης των επιμέρους στοιχείων. Τα νευρωνικά μοντέλα που αναπτύχθηκαν για την ελληνική γλώσσα και τα οποία διατίθενται προς ελεύθερη χρήση, αποτέλεσαν καταλυτικούς παράγοντες για την πραγμάτωση της αρχιτεκτονικής, γεφυρώνοντας το χάσμα μεταξύ γλωσσών χαμηλών και υψηλών πόρων.

Το επόμενο κεφάλαιο συνοψίζει τη συμβολή της παρούσας διδακτορικής εργασίας, παρουσιάζοντας τα συμπεράσματα που προέκυψαν τόσο κατά τη φάση σχεδιασμού όσο και κατά την τεχνική υλοποίηση των παραπάνω μηχανισμών. Παράλληλα, συζητούνται πιθανές κατευθύνσεις για μελλοντική έρευνα και επεκτάσεις των προσεγγίσεων που παρουσιάστηκαν στα προηγούμενα κεφάλαια της διατριβής.

5 ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Ανασκόπηση βασικών αποτελεσμάτων και συμπερασμάτων

Η παρούσα διδακτορική διατριβή επικεντρώθηκε στο σχεδιασμό και στην ανάπτυξη μεθόδων επεξεργασίας φυσικής γλώσσας (ΕΦΓ) για την αποτελεσματικότερη διαχείριση της πληροφορίας που συναντάται σε ελεύθερο κείμενο, καλύπτοντας το αντίστοιχο κενό για την ελληνική γλώσσα. Συγκεκριμένα, μελετήθηκαν μεθοδολογίες και αναπτύχθηκαν συστήματα για τη συλλογή, ιχνηλάτηση, προεπεξεργασία και αναπαράσταση δεδομένων που προέρχονται από πηγές του Ιστού, για την εξαγωγή πληροφοριών, τον εμπλουτισμό και την αποτύπωσή τους σε δομημένη μορφή, καθώς και για την αξιοποίηση των παραπάνω μέσων με στόχο το δυναμικό έλεγχο ισχυρισμών/υποθέσεων από πλευράς χρήστη.

Οι στόχοι που παρατέθηκαν στο Κεφάλαιο 1 εκπληρώνονται πλήρως μέσω της υλοποίησης δύο βασικών ερευνητικών αξόνων που αναπτύχθηκαν εκτενώς στο Κεφάλαιο 3: τον άξονα εξόρυξης πληροφοριών από ελεύθερο κείμενο και τον άξονα εξαγωγής σημασιολογικών συμπερασμάτων από διαδικτυακές πηγές. Οι δύο αυτοί άξονες λειτουργούν συμπληρωματικά στην απάντηση των ερευνητικών ζητημάτων της διατριβής, υλοποιώντας αρθρωτές αρχιτεκτονικές επεξεργασίας κειμένου. Τα επιμέρους στοιχεία που απαρτίζουν τις αρχιτεκτονικές αυτές επιτρέπουν την παράλληλη εξαγωγή πληροφοριών από ελεύθερο κείμενο σε δομημένη μορφή (σχεσιακών ν-πλειάδων), τη σύνδεσή τους με οντολογίες εξωτερικών γνωσιακών βάσεων για εμπλουτισμό της αποκτηθείσας πληροφορίας και τη διευκόλυνση εργασιών διερευνητικής ανάλυσης, καθώς και την επικύρωση ισχυρισμών με βάση την κατασκευή τεκμηρίων από διαδικτυακές πηγές σε πραγματικό χρόνο.

Συνολικά, τα κύρια αποτελέσματα που προέκυψαν από την παρούσα διατριβή περιλαμβάνουν:

1. Την ανάπτυξη μεθοδολογίας και αντίστοιχου συστήματος ανοιχτής εξαγωγής πληροφοριών από πηγές ελεύθερου κειμένου για την εξαγωγή τριπλετών. Η μεθοδολογία βασίζεται στη χρήση μοντέλων μηχανικής μετάφρασης για το μετασχηματισμό του κειμένου στο αγγλικό του ισοδύναμο, προκειμένου να εφαρμοστούν εργασίες προεπεξεργασίας που δεν υποστηρίζονται από γλώσσες χαμηλότερων πόρων, και συγκεκριμένα η επίλυση φαινομένων συναναφοράς, η σύνδεση οντοτήτων και η δημιουργία εξαγωγικής σύνοψης του κειμένου. Τα παραπάνω βήματα ευνοούν την εστίαση της εξαγωγής πληροφοριών σε μια βελτιωμένη μορφή του αρχικού κειμένου (πχ. χωρίς αμφισημίες, συναναφορές, περιττές πληροφορίες). Ακολουθεί η παράλληλη χρήση μηχανισμών ανοιχτής εξαγωγής πληροφοριών (τριπλέτες της μορφής {υποκείμενο, κατηγόρημα, αντικείμενο}) που βασίζονται τόσο σε υπολογιστικές μεθόδους και τεχνικές μηχανικής/βαθιάς μάθησης, όσο και σε γλωσσολογικούς κανόνες, δρώντας συμπληρωματικά με στόχο την επίτευξη της καλύτερης δυνατής ισορροπίας ακρίβειας-ανάκλησης. Τα τελικά προϊόντα της εξαγωγής μεταφράζονται στην ελληνική γλώσσα μέσω μηχανισμού αντίστροφης μετάφρασης.
2. Την ανάπτυξη μεθοδολογίας και αντίστοιχου συστήματος επαλήθευσης ισχυρισμών που επιτρέπει στον χρήστη να εισάγει μια οποιαδήποτε υπόθεση σε φυσική γλώσσα (μέσω ομιλίας ή κειμένου), η οποία επικυρώνεται ή απορρίπτεται βάσει των συγκεντρωμένων τεκμηρίων από ελληνικές διαδικτυακές ειδησεογραφικές πηγές. Για την επίτευξη του παραπάνω αξιοποιείται μια σειρά στοιχείων για την περιοδική ιχνηλάτηση ειδησεογραφικού περιεχομένου, την επισήμανσή του μέσω σύνδεσης οντοτήτων και την κατασκευή τεκμηρίων που προκύπτουν από τον έλεγχο και συνδυασμό αποσπασμάτων μέσω μηχανισμών ελέγχου σημασιολογικής ομοιότητας. Τέλος, το σημασιολογικά σχετικότερο τεκμήριο τροφοδοτεί (παράλληλα με τον ισχυρισμό προς έλεγχο) ένα στοιχείο ελέγχου κειμενικής συνεπαγωγής, μέσω του οποίου ελέγχεται η εγκυρότητα του ισχυρισμού.

Πέραν των προαναφερθεισών μεθοδολογιών και συστημάτων, αξίζει να σημειωθεί και η αξία των επιμέρους στοιχείων που τα απαρτίζουν, ως μεμονωμένα εργαλεία (πχ. μοντέλα ελέγχου σημασιολογικής ομοιότητας, αναγνώρισης κειμενικής συνεπαγωγής, ανοιχτής εξαγωγής πληροφοριών, μηχανικής μετάφρασης κτλ.) για εργασίες τελικού χρήστη, αποτελώντας

σημαντική συνεισφορά στον εμπλουτισμό της εργαλειοθήκης της ελληνικής κοινότητας ΕΦΓ.

Τα συμπεράσματα που προέκυψαν από την ανάπτυξη των παραπάνω μεθοδολογιών και στοιχείων, συνοψίζονται παρακάτω:

Σ1. Η υλοποίηση μεθόδων ανοιχτής εξαγωγής πληροφοριών για γλώσσες χαμηλότερων πόρων (όπως η ελληνική) με χρήση τεχνικών μηχανικής μάθησης, προϋποθέτει είτε την ύπαρξη διαθέσιμων πόρων (επισημασμένων συνόλων δεδομένων) για την εκπαίδευση αντίστοιχων μοντέλων, είτε την προσαρμογή προεκπαιδευμένων πολύγλωσσων γλωσσικών μοντέλων (νευρωνικών αρχιτεκτονικών), τα οποία όμως πρέπει να εξασφαλίζουν την ισόρροπη αντιπροσώπευση των υποστηριζόμενων γλωσσών. Στα πολύγλωσσα γλωσσικά μοντέλα που συναντώνται σήμερα, η ελληνική γλώσσα υπο-αντιπροσωπεύεται συστηματικά, με αποτέλεσμα να μην μπορούν συνήθως να εκπληρώσουν το στόχο για τον οποίο δημιουργήθηκαν.

Σ2. Ως έμμεση λύση για το παραπάνω πρόβλημα, η χρήση τεχνικών μηχανικής μετάφρασης έχει καταλυτική επίδραση στην ενεργοποίηση δυνατοτήτων που δε διατίθενται για γλώσσες χαμηλότερων πόρων. Όπως αποδείχθηκε στις περιπτώσεις χρήσης του Κεφαλαίου 4, η πλαισίωση του μηχανισμού ανοιχτής εξαγωγής πληροφοριών από στοιχεία μηχανικής (και αντίστροφης) μετάφρασης καθιστά δυνατή την εξαγωγή δομημένης πληροφορίας στα ελληνικά, με ικανοποιητική απόδοση, σημαντικά καλύτερης από αυτή που παρουσιάζουν οι τεχνικές 0-shot-learning που βασίζονται σε πολύγλωσσα γλωσσικά μοντέλα.

Σ3. Η εργασία της ανοιχτής εξαγωγής πληροφοριών (OIE), επωφελείται γενικότερα από την ύπαρξη ενός συνόλου τεχνικών προεπεξεργασίας (επίλυση συναναφορών, σύνδεση οντοτήτων, εξαγωγική συνόψιση) καθώς αυτές συμβάλλουν στην αύξηση της ειδικής πυκνότητας της πληροφορίας που περιέχεται σε ένα σώμα κειμένων. Οι περισσότερες τεχνικές που συναντώνται τείνουν να αγνοούν το παραπάνω, εφαρμόζοντας τεχνικές OIE απευθείας στο αρχικό κείμενο, με αποτέλεσμα την εξαγωγή μεγάλου αριθμού περιττών τριπλετών μη μικρή πληροφοριακή αξία.

Σ4. Η συνδυαστική προσέγγιση στην ανάπτυξη μεθοδολογίας εξαγωγής πληροφοριών, η οποία υλοποιείται μέσω της συμπληρωματικής χρήσης πολλαπλών OIE μηχανισμών που βασίζονται τόσο σε γλωσσολογικούς κανόνες όσο και σε τεχνικές μηχανικής μάθησης, οδηγεί σε βελτίωση της συνολικής απόδοσης, εφόσον ακολουθείται από τον αποδοτικό συνδυασμό των επιμέρους αποτελεσμάτων εξαγωγής. Όπως διαπιστώνεται από τα αποτελέσματα του Κεφαλαίου 4, τα στοιχεία

που βασίζονται σε γλωσσολογικούς κανόνες αποτελούν συνήθως συστήματα υψηλής ακρίβειας (precision), ενώ αντίστοιχα τα στοιχεία που ενσωματώνουν τεχνικές μηχανικής μάθησης χαρακτηρίζονται από υψηλή ανάκληση (recall). Μέσω του κατάλληλου συνδυασμού τους, είναι δυνατή η δημιουργία ενός μηχανισμού ανοιχτής εξαγωγής πληροφοριών με επιδόσεις αιχμής.

Σ5. Παρά την αδιαμφισβήτητη συνεισφορά της ανοιχτής εξαγωγής πληροφοριών σε εργασίες διερευνητικής ανάλυσης και δομημένης αναπαράστασης πληροφοριών από ελεύθερο κείμενο, η χρησιμότητά της σε εργασίες ελέγχου ισχυρισμών κρίθηκε περιορισμένη. Συγκεκριμένα, ο τεμαχισμός του περιεχομένου σε επίπεδο λέξεων ή σύντομων φράσεων (δηλαδή στις συνιστώσες μια τριπλέτας) δεν εξυπηρετεί πιο πολύπλοκες διαδικασίες συμπερασμού όπως ο έλεγχος σημασιολογικής ομοιότητας και η αναγνώριση κειμενικής συνεπαγωγής, οι οποίες γίνονται (τουλάχιστον) σε επίπεδο πρότασης.

Σ6. Ο έλεγχος ισχυρισμών βάσει διαθέσιμων δεδομένων αποτελεί πολύπλοκη εργασία που απαιτεί το συνδυασμό πολλών διαφορετικών εργαλείων ΕΦΓ (πχ. μοντέλα STS, NLI), τα οποία συνήθως δε διατίθενται για γλώσσες χαμηλών πόρων. Ωστόσο, η χρήση παράλληλου κειμένου ή τεχνικών μηχανικής μετάφρασης για τη δημιουργία κατάλληλων συνόλων δεδομένων, σε συνδυασμό με τεχνικές μεταφοράς μάθησης (transfer learning) και απόσταξης γνώσης (knowledge distillation) μπορούν να οδηγήσουν στην δημιουργία αντίστοιχων εργαλείων για οποιαδήποτε γλώσσα.

Σ7. Είναι απαραίτητη η περιοδική ιχνηλάτηση διαδικτυακού περιεχομένου για την συγκέντρωση απαραίτητων πληροφοριών που θα χρησιμοποιηθούν ως βάση ελέγχου των εκάστοτε ισχυρισμών, προκειμένου να εξασφαλίζεται ο δυναμικός χαρακτήρας της διαδικασίας ελέγχου ισχυρισμών. Αυτό δικαιολογείται από το γεγονός ότι η συνεχής ενσωμάτωση νέων πληροφοριών ενδέχεται να οδηγήσει σε αλλαγή της ετυμηγορίας ενός πχ. αρχικά επικυρωμένου ισχυρισμού. Βάσει των μεθόδων που περιεγράφηκαν στο Κεφάλαιο 3, φαίνεται πως η ιχνηλάτηση HTML σελίδων οδηγεί σε μεγαλύτερη κάλυψη της επικαιρότητας, σε σχέση με την ιχνηλάτηση τροφοδοσιών RSS, κυρίως λόγω του συγκριτικά περιορισμένου περιεχομένου που συγκεντρώνει η τελευταία, όπως άλλωστε αναμενόταν.

Σ8. Η χρήση βάσεων δεδομένων γράφων για την αποθήκευση των σωμάτων κειμένου που προέρχονται από ειδησεογραφικές πηγές, σε συνδυασμό με τη σύνδεση οντοτήτων που περιέχονται σε αυτά με εξωτερικές οντολογίες (πχ. Wikification) εξυπηρετεί ιδιαίτερα εργασίες κατασκευής τεκμηρίων για τον έλεγχο ισχυρισμών.

Ουσιαστικά, οι οντότητες λειτουργούν ως "κόλλα" που ενώνει αποσπάσματα φαινομενικά ασύνδετων ειδήσεων, ειδικά για περιπτώσεις ισχυρισμών που περιέχουν πολλές εννοιολογικές οντότητες.

Σ9. Βασικό προαπαιτούμενο για τον αποδοτικό έλεγχο υποθέσεων/ισχυρισμών αποτελεί η καλή ποιότητα των δεδομένων που χρησιμοποιούνται για την κατασκευή τεκμηρίων. Παρ'ότι η διασταύρωση της αλήθειας των αξιοποιούμενων πληροφοριών είναι εκτός των στόχων της εργασίας (όλες οι πηγές θεωρούνται αξιόπιστες), τα αποσπάσματα που απαρτίζουν την εκάστοτε είδηση ενδέχεται να περιλαμβάνουν υποκειμενικές γνώμες/απόψεις (πχ. απόσπασμα συνέντευξης), που δεν περιγράφουν αντικειμενικά ένα γεγονός ή κατάσταση. Για το λόγο αυτό, όπως περιεγράφηκε στο Κεφάλαιο 3, κρίθηκε απαραίτητη η εφαρμογή μεθόδων για την ανίχνευση αντικειμενικού/υποκειμενικού λόγου, τα αποτελέσματα της οποίας να αντανακλώνται (πχ. ως δείκτης) στο εκάστοτε τεκμήριο.

5.2 Περιορισμοί

Τόσο κατά τη φάση σχεδιασμού όσο και κατά τη φάση τεχνικής υλοποίησης των προαναφερθέντων μηχανισμών εντοπίστηκαν ορισμένοι περιορισμοί, οι οποίοι παρατίθενται στην παρούσα υποενότητα:

Π1. Η εξάρτηση της μεθοδολογίας ανοιχτής εξαγωγής πληροφοριών από μηχανισμούς μηχανικής μετάφρασης για το μετασχηματισμό του κειμένου στο αγγλικό του ισοδύναμο και για τη μετάφραση των (αγγλικών) τριπλετών στην ελληνική γλώσσα παρουσιάζει μειωμένη απόδοση σε περιπτώσεις ύπαρξης ασαφών, αμφίσημων ή διφορούμενων όρων (πχ. η λέξη "bank" μπορεί να μεταφραστεί είτε ως "τράπεζα" είτε ως "όχθη"). Αυτό οφείλεται στο ότι οι διανυσματικές αναπαραστάσεις που αποδίδονται στην εκάστοτε λέξη εξαρτώνται από το περιεχόμενο της ακολουθίας που τροφοδοτείται στο στοιχείο μηχανικής μετάφρασης. Εφόσον τα μέρη μιας τριπλέτας αποτελούν υποσύνολα πρότασης, το τροφοδοτούμενο περιεχόμενο ενδέχεται να μην είναι επαρκές για την αντιστοίχιση των σωστών διανυσματικών αναπαραστάσεων, οδηγώντας σε λανθασμένη απόδοση της λέξης στην ελληνική. Σημειώνεται ότι η εναλλακτική μέθοδος ευθυγράμμισης λέξεων που εξετάστηκε, εμπίπτει σε όλους τους περιορισμούς που χαρακτηρίζουν τις γλώσσες χαμηλότερων πόρων όπως η ελληνική, και οι οποίες συζητήθηκαν παραπάνω.

Π2. Οι τεχνικές σύνδεσης οντοτήτων με εξωτερικές γνωσιακές βάσεις (πχ. WikiData) αποτελούν κομβικό στοιχείο για την ευρωστία της μεθοδολογίας εξαγωγής

σημασιολογικών συμπερασμάτων. Ωστόσο, η ύπαρξη ομωνύμων ή παραλλαγών ονοματικών οντοτήτων (πχ. η οντότητα "Ελ.Βενιζέλος" μπορεί να αναφέρεται στον Έλληνα πολιτικό ή στο αεροδρόμιο της Αθήνας) θέτουν προκλήσεις σε αντίστοιχα συστήματα. Δεδομένης της σειριακής λογικής που ακολουθεί η προτεινόμενη μεθοδολογία ελέγχου ισχυρισμών, μια ψευδώς-αληθής αντιστοίχιση αναφοράς με οντότητα θα επιφέρει τη λανθασμένη συσχέτιση ισχυρισμού με το τεκμήριο που την περιέχει, οδηγώντας σε περισσότερες "ουδέτερες" ετυμηγορίες του μηχανισμού αναγνώρισης κειμενικής συνεπαγωγής. Με άλλα λόγια, ο μηχανισμός δε θα είναι σε θέση ούτε να επικυρώσει ούτε να απορρίψει τον ισχυρισμό καθώς τα δεδομένα από το στάδιο κατασκευής τεκμηρίων δε θα σχετίζονται πραγματικά με αυτόν, οδηγώντας σε ουδέτερο αποτέλεσμα.

Π3. Η προτεινόμενη μεθοδολογία που ακολουθείται για την κατασκευή τεκμηρίων βασίζεται στη σύγκριση της σημασιολογικής ομοιότητας αποσπασμάτων που συγκεντρώνονται από ετερογενείς ειδησεογραφικές πηγές με τον προς εξέταση ισχυρισμό. Για τη συγκέντρωση των αποσπασμάτων αυτών ακολουθείται αλγόριθμος εύρεσης συντομότερης διαδρομής, ο οποίος διατρέχει όλα τα πιθανά μονοπάτια εναλλασσόμενων κόμβων οντοτήτων και αποσπασμάτων (οι οντότητες έχουν ήδη επισημανθεί σε κάθε απόσπασμα από το μηχανισμό σύνδεσης οντοτήτων). Σε περιπτώσεις που ο αριθμός οντοτήτων που περιέχονται στον αρχικό ισχυρισμό (και άρα πρέπει να περιλαμβάνονται στο μονοπάτι) είναι μεγάλος (πχ. >5) ο υπολογισμός της συντομότερης διαδρομής είναι υπολογιστικά κοστοβόρος, επιβραδύνοντας την εξαγωγή της ετυμηγορίας.

Π4. Συμπληρωματικά του παραπάνω περιορισμού, η υπόθεση ότι το συντομότερο μονοπάτι έχει εγκολπώσει τα σχετικότερα αποτελέσματα (πχ. έναντι άλλων διαδρομών) για την κατασκευή του καλύτερου τεκμηρίου, αποτελεί εμπειρική θεώρηση, η οποία δε μπορεί να αποδειχθεί. Για το σκοπό αυτό, μπορούν να αξιοποιηθούν εναλλακτικά τα n καλύτερα τεκμήρια, είτε εξάγοντας το μέσο όρο από το μηχανισμό αναγνώρισης συνεπαγωγής, είτε δίνοντας τη δυνατότητα στο χρήστη να αξιολογήσει το κάθε αποτέλεσμα χωριστά.

Π5. Τέλος, η αναγνώριση κειμενικής συνεπαγωγής μεταξύ ενός ζεύγους ακολουθιών (ισχυρισμού, τεκμηρίου) εξαρτάται εν μέρει από τα μορφοσυντακτικά στοιχεία των κατηγορημάτων. Πιο συγκεκριμένα, ο χρόνος και η έγκλιση των ρημάτων διαδραματίζουν σημαντικό ρόλο στην ερμηνεία του νοήματος μια πρότασης (πχ. οι προτάσεις "το κράτος ενέκρινε μέτρα…" και "το κράτος θα ενέκρινε μέτρα…"

έχουν τελείως διαφορετικό νόημα, καθώς η δεύτερη χρησιμοποιεί υποθετικό λόγο για να εκφράσει το αντίθετο του πραγματικού). Αν και η εργασία του Kober et al. (2019) έχει αποδείξει πως τα γλωσσικά μοντέλα είναι εν μέρει σε θέση να κωδικοποιήσουν αντίστοιχες μορφοσυντακτικές ιδιότητες, η απόδοσή τους επηρεάζεται από περιπτώσεις όπου η παρουσία αντίστοιχων φαινομένων στο κείμενο είναι ισχυρή. Επομένως προτείνεται η αποφυγή ισχυρισμών με πολύπλοκες μορφοσυντακτικές ιδιαιτερότητες.

5.3 Προτάσεις για μελλοντική έρευνα

Η παρούσα διδακτορική διατριβή παρουσίασε μια σειρά μεθοδολογιών για την εξαγωγή πληροφοριών από ελεύθερο κείμενο, την προεπεξεργασία, εμπλουτισμό και αναπαράστασή τους σε δομημένη μορφή, καθώς και τη χρήση τους για τον έλεγχο ισχυρισμών σε φυσική γλώσσα. Οι μηχανισμοί που υλοποιήθηκαν αποτελούν μια προσπάθεια εκδημοκρατισμού της επεξεργασίας φυσικής γλώσσας για περιπτώσεις γλωσσών χαμηλότερων πόρων και συγκεκριμένα για την ελληνική, στοχεύοντας στη περαιτέρω διάδοση και ευρύτερη χρήση αντίστοιχων τεχνικών για εφαρμογές τελικού χρήστη. Το μεθοδολογικό πλαίσιο που ακολουθήθηκε ανέδειξε μια σειρά περιορισμών (οι οποίοι αναλύθηκαν στην προηγούμενη υποενότητα), όλοι εκ των οποίων αποτελούν πιθανές ερευνητικές γραμμές που χρήζουν περαιτέρω διερεύνησης καθώς ενδέχεται να οδηγήσουν στη βελτίωση των αρχικών αποτελεσμάτων.

Ενδεικτικά, προτείνεται η τροποποίηση του μηχανισμού αντίστροφης μετάφρασης κατά την μετα-επεξεργασία των εξαχθεισών τριπλετών ώστε να συνδυάζει μηχανισμό ευθυγράμμισης εξασφαλίζοντας την αντιστοίχισή τους με μέρη του αρχικού κειμένου. Εναλλακτικά, κρίνεται ιδιαίτερα βοηθητική η δημιουργία επισημασμένων συνόλων δεδομένων εκπαίδευσης και αντίστοιχων benchmarks για την απευθείας εκπαίδευση μοντέλων επίλυσης συναναφορών, εξαγωγικής συνόψισης και ανοιχτής εξαγωγής πληροφοριών για την ελληνική γλώσσα, προκειμένου να διευκολυνθούν οι αντίστοιχες εφαρμογές και να μην απαιτείται η πλαισίωσή τους από μηχανισμούς μηχανικής μετάφρασης. Τέλος, αξίζει να μελετηθεί η βελτιστοποίηση του μηχανισμού κατασκευής τεκμηρίων, παραλλάσσοντας τη διαδικασία εύρεσης συντομότερης διαδρομής μεταξύ αποσπασμάτων, με απώτερο στόχο τη μείωση του υπολογιστικού κόστους κατά τη διαδικασία κατασκευής τεκμηρίων.

6 Βιβλιογραφια

Ossama Abdel-Hamid, Abdel Rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing. 22*(10), 1533-1545

Kiran Adnan and Rehan Akbar. 2019. An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*. 6.1 (2019): 1-38

Fouad Nasser A. Al Omran and Christoph Treude. 2017. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. In *2017 IEEE/ACM 14th international conference on mining software repositories (MSR)* (pp. 187-197). IEEE.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D., Juan B., and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications. arXiv preprint arXiv:1707.02268*

Jonathan Allen. 2003. Speech Recognition and Synthesis. In *Encyclopedia of Computer Science*, John Wiley and Sons Ltd., GBR (pp. 1664–1667).

Sihem Amer-Yahia, Georgia Koutrika, Martin Braschler, Diego Calvanese, Davide Lanti, Hendrik Lücke-Tieke, Alessandro Mosca, Tarcisio de Farias, Dimitris Papadopoulos, Yogendra Patil, and others. 2022. INODE: building an end-to-end data exploration system in practice. *ACM SIGMOD Record*, 50(4):23–29.

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of*

the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 344-354).

D. Anggraeni, W. S.M. Sanjaya, M. Y.S. Nurasyidiek, and M. Munawwaroh. 2018. The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm. In *IOP Conference Series: Materials Science and Engineering* (Vol. 288, No. 1, p. 012042). IOP Publishing.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*. 33, 12449-12460.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

James K. Baker. 1975. The DRAGON System-An Overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing. 23*(1), 24-29.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2020. Careful selection of knowledge to solve open book question answering. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.* (pp. 6120-6129)

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. *Communications of the ACM*, *51*(12), 68-74.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2020. A Review on Fact Extraction and VERification: The FEVER case. *ACM Computing Surveys (CSUR), 55(1), 1-35*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2020a. SCIBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference.* (pp. 3615-3620).

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020b. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam. 2020. CARB: A crowdsourced benchmark for open IE. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. (EMNLP-IJCNLP)* (pp. 6262-6267).

Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*. (pp. 22-31).

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data - The story so far. *International Journal on Semantic Web and Information Systems*. (pp. 205-227). IGI global

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267-D270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics. 5*, 135-146

Janez Brank, Gregor Leban, and Marko Grobelnik. 2018. Semantic annotation of documents based on wikipedia concepts. *Informatica (Slovenia). 42*(1)

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems. 33*, 1877-1901.

Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. AttSum: Joint learning of focusing and summarization with neural attention. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*. (pp. 547-556)

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 29, No. 1).

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2018. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-*

2017), pages 1–14, Vancouver, Canada. Association for Computational Linguistics

Bharathi Raja Chakravarthi, Mihael Arcan, and John P McCrae. 2019. WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation. *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*. (pp. 1-7).

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. (pp. 789-797).

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. (pp. 4960-4964). IEEE.

Iti Chaturvedi, Erik Cambria, Roy E. Welsch, and Francisco Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion, 44, 65-77*.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan Thorsten Peter. 2016. Guided Alignment Training for Topic-Aware Neural Machine Translation. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track* (pp. 121-134).

Chung Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018a. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. (pp. 4774-4778). IEEE.

Chung Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2018b. Speech recognition for medical conversations. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. (pp. 2972-2976)

Janara Christensen, Stephen Soderland, and Oren Etzioni. 2011. An Analysis of Open Information Extraction Based on Semantic Role Labeling Categories and Subject Descriptors. *Proceeding of K-CAP '11 Proceedings of the sixth international conference on Knowledge capture*. (pp. 113-120).

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLoS ONE*. *10*(6), e0128193.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*. (pp. 643-653).

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 2426-2430).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440-8451).

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems, 32*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. (pp. 2475-2485).

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *EMSEE 2005 - Empirical Modeling of Semantic Equivalence and Entailment@ACL 2005, Proceedings of the Workshop*. (pp. 13-18).

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. (EMNLP-CoNLL)* (pp. 708-716).

Itai Dabran, Tzoof Avny, Eytan Singher, and Haim Ben Danan. 2017. Augmented reality speech recognition for the hearing impaired. In *2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems, COMCAS 2017*. (pp. 1-4). IEEE.

Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. 2019. The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Research*. *47*(D1), D948-D954.

Steven B. Davis and Paul Mermelstein. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.

Luciano Del Corro and Rainer Gemulla. 2013a. ClausIE: Clause-based open information extraction. In *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*. (pp. 355-366).

Luciano Del Corro and Rainer Gemulla. 2013b. ClausIE: Clause-Based Open Information Extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 355–366, New York, NY, USA. Association for Computing Machinery.

Thierry Desot, François Portet, and Michel Vacher. 2020. Corpus generation for voice command in smart home and the effect of speech synthesis on end-to-end slu. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings (LREC 2020)* (pp. 6395-6404).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. {BERT}: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Volume 1 (Long and Short Papers)* (pp. 4171-4186).

Kristin N. Dew, Anne M. Turner, Yong K. Choi, Alyssa Bosold, and Katrin Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics*, *85*, 56-67.

Hayley M. Dingerdissen, Frederic Bastian, K. Vijay-Shanker, Marc Robinson-Rechavi,

Amanda Bell, Nikhita Gogate, Samir Gupta, Evan Holmes, Robel Kahsay, Jonathon Keeney, Heather Kincaid, Charles Hadley King, David Liu, Daniel J. Crichton, and Raja Mazumder. 2020. OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data. *JCO clinical cancer informatics, 4,* 210-220.

Linhao Dong and Bo Xu. 2020. CIF: Continuous Integrate-And-Fire for End-To-End Speech Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. (pp. 6079-6083). IEEE.

Matthew S. Dryer and Martin Haspelmath. 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. *Online: http://wals. info*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference*. (pp. 644-648).

Pradheep Elango. 2006. Coreference Resolution : A Survey. *University of Wisconsin, Madison, WI*, 12.

Ahmed Elnaggar, Christoph Gebendorfer, Ingo Glaser, and Florian Matthes. 2018. Multi-task deep learning for legal document translation, summarization and multi-label classification. In *ACM International Conference Proceeding Series*. (pp. 9-15).

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. (pp. 8517-8532).

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Debasis Ganguly, and David Konopnicki. 2020. A summarization system for scientific documents. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*. (pp. 211-216).

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., July. Association

for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference. (Volume 1: Long Papers)* (pp. 34-43).

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.* (pp. 363-370).

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference.* (pp. 1256-1261).

Mark Gales and Steve Young. 2007. The application of hidden Markov Models in speech recognition. *Signal Processing, 1(3), 195-304.*

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*. *47*(1), 1-66.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *34th International Conference on Machine Learning, ICML 2017*. (pp. 1243-1252). PMLR.

Samira Ghodratnama, Amin Beheshti, Mehrdad Zakershahrak, and Fariborz Sobhanmanesh. 2020. Extractive Document Summarization Based on Dynamic Feature Space Mapping. *IEEE Access, 8, 139084-139095.*

Ian J Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA. url{http://www.deeplearningbook.org}.

Sindhuja Gopalan and Sobha Lalitha Devi. 2017. Cause and Effect Extraction from Biomedical Corpus. *Computacion y Sistemas*. *21*(4), 749-757.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*(Vol 1299, pp: 11-27)

Paul Groth, Mike Lauruhn, Antony Scerri, and Ron Daniel. 2018. Open information

extraction on scientific text: An evaluation. In *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*. (pp. 3414-3423).

Anmol Gulati, James Qin, Chung Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, Proc. Interspeech 2020, 5036-5040.*.

Christian Gulden, Melanie Kirchner, Christina Schüttler, Marc Hinderer, Marvin Kampf, Hans Ulrich Prokosch, and Dennis Toddenroth. 2019. Extractive summarization of clinical trial descriptions. *International Journal of Medical Informatics*. *129*, 114-121.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. 2021. Using Question Answering Rewards to Improve Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. (pp. 518-526).

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. (pp. 2681-2690).

Bikash Gyawali, Lucas Anastasiou, and Petr Knoth. 2020. Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*. (pp. 901-910).

W. Haas. 1958. J.R. Firth: Papers in linguistics, 1934–1951. xii, 233 pp., 11 plates. London, etc.: Oxford University Press, 1957. 35s. *Bulletin of the School of Oriental and African Studies*, *21*(3), 668-671.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. News-Please: A Generic News Crawler and Extractor. In *15th International Symposium of Information Science (ISI 2017)* (pp. 218-223).

Hui Han, Hongyuan Zha, and C. Lee Giles. 2005. Name disambiguation in author citations using a K-way spectral clustering method. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. (JCDL'05)* (pp. 334-343). IEEE.

Jiabao Han and Hongzhi Wang. 2021. Transformer based network for Open Information Extraction. *Engineering Applications of Artificial Intelligence*. 102, 104262.

Zellig S. Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*. (pp. 390-457). Springer, Dordrecht.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (pp. 1803-1812).

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. (pp. 1576-1586).

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Volume 1: Long Papers) (pp. 473-483).*

Aldo Hernandez-Suarez, Gabriel Sanchez-Perez, Karina Toscano-Medina, Hector Perez-Meana, Jose Portillo-Portillo, Victor Sanchez, and Luis Javier García Villalba. 2019. Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors, 19(7), 1746 (Switzerland).*

Jerry R. Hobbs and Ellen Riloff. 2010. Information extraction. In *Handbook of Natural Language Processing, Second Edition*. *15*, 16

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Volume 1: Long Papers) (pp. 328-339).*

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *37th International Conference on Machine Learning, ICML 2020*. (pp. 4411-4421). PMLR.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. (pp. 1684-1697).

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for

information extraction. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé. 2015. Deep unordered composition rivals syntactic methods for text classification. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference. (volume 1: Long papers)* (pp. 1681-1691).

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HOVER: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. (pp. 3441-3460).

Woojeong Jin, Rahul Khanna, Suji Kim, Dong Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. FORECASTQA: A question answering challenge for event forecasting with temporal text data. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference. (Volume 1: Long Papers)* (pp. 4636-4650).

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28 No. 1, pp. 11-21.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dan Jurafsky and James H Martin. 2009. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. *Computational Linguistics*, *26*(4), 638-641.

Daniel Jurafsky and James H Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (second edition).

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2015. Extractive Summarization using Continuous Vector Space Models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)* (pp. 31-39).

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *52nd Annual Meeting of the Association for*

Computational Linguistics, ACL 2014 - Proceedings of the Conference. (Volume 1: Long Papers) (pp. 655-665).

Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: fact checking statistical claims. *Proceedings of the VLDB Endowment. 13*(12), 2965-2968.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Katrin Kirchhoff, Anne M. Turner, Amittai Axelrod, and Francisco Saavedra. 2011. Application of statistical machine translation to public health information: A feasibility study. *Journal of the American Medical Informatics Association. 18*(4), 473-478.

Thomas Kober, Sander de Vroe, and Mark Steedman. 2019. Temporal and Aspectual Entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Philipp Koehn, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, and Christine Moran. 2007. Moses: open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07* (pp. 177-180).

Sebastian Kohlmeier, Kyle Lo, Lucy Lu Wang, and J J Yang. 2020. COVID-19 Open Research Dataset (CORD-19). In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. OpenIE6: Iterative grid labeling and coordination analysis for open information extraction. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 3748-3761).

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative Memory-Based Joint Open Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5871-5886).

Maarit Koponen, Leena Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine

translation output. *Machine Translation, 33(1), 61-90*.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2020. Med7: a transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine, 118, 102086*.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence* (pp. 110-117).

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. (pp. 9332-9346).

Gourab Kundu, Avirup Sil, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Volume 2: Short Papers) (pp. 395-400).*

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 166-172).

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *31st International Conference on Machine Learning, ICML 2014*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017a. End-to-end neural coreference resolution. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. (pp. 188-197).

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017b. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September. Association for Computational Linguistics.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. (Vol. 10, No. 8, pp. 707-710).

Hector J. Levesque. 2011. The Winograd schema challenge. In *AAAI Spring Symposium - Technical Report*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871-7880).

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*. (pp. 1430-1441).

Wei Li. 2015. Abstractive multi-document summarization with semantic information extraction. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. (pp. 1908-1913).

Xueling Lin and Lei Chen. 2019. Canonicalization of open knowledge bases with side information from the source text. In *Proceedings - International Conference on Data Engineering (ICDE) (pp. 950-961). IEEE.*

Guiliang Liu, Xu Li, Jiakang Wang, Mingming Sun, and Ping Li. 2020. Extracting Knowledge from Web Text with Monte Carlo Tree Search. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020* (pp. 2585-2591).

Yang Liu. 2019. Fine-tune BERT for Extractive Summarization. *CoRR*, abs/1903.1.

Yue Liu, Tongtao Zhang, Zhicheng Liang, Heng Ji, and Deborah L McGuinness. 2018. Seq2RDF: An End-to-end Application for Deriving Triples from Natural Language Text. *ArXiv*, abs/1807.0.

Lajanugen Logeswaran, Ming Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2020. Zero-shot entity linking by reading entity descriptions. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (pp. 3449-3460).

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (pp. 2641-2651).

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In

Tengfei Ma, Chiamin Wu, Cao Xiao, and Jimeng Sun. 2018. AWE: Asymmetric Word Embedding for Textual Entailment. *Journal of Environmental Sciences (China) English Ed.*

Simone Magnolini, Ngoc Phuoc An Vo, and Octavian Popescu. 2016. Analysis of the impact of Machine Translation evaluation metrics for Semantic Textual Similarity. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

Sarthak Majithia, Fatma Arslan, Sumeet Lubal, Damian Jimenez, Priyank Arora, Josue Caraballo, and Chengkai Li. 2019. ClaimPortal: Integrated monitoring, searching, checking, and analytics of factual claims on twitter. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations* (pp. 153-158).

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2015. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Daniel Jurafsky & James H. Martin. 2001. Speech and Language Processing: An introduction to natural language processing,. *SPEECH and LANGUAGE PROCESSING An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition.*

Pedro Henrique Martins, Zita Marinho, and Andre F.T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop* (pp. 190-196).

Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 4074-4077).

Diana Maynard, Horacio Saggion, Milena Yankova, Kalina Bontcheva, and Wim Peters. 2007. Natural language technology for information integration in business intelligence. In *International Conference on Business Information Systems* (pp. 366-380). Springer, Berlin, Heidelberg.

Paul Mcnamee, Mark Dredze, Adam Gerber, Nikesh Garera, Tim Finin, James Mayfield, Christine Piatko, Delip Rao, David Yarowsky, and Markus Dreyer. 2009. HLTCOE Approaches to Knowledge Base Population. In *TAC 2009. UMBC Faculty Collection.*

M. F. Medress, F. S. Cooper, J. W. Forgie, C. C. Green, D. H. Klatt, M. H. O'Malley, E. P. Neuburg, A. Newell, D. R. Reddy, B. Ritea, J. E. Shoup-Hummel, D. E. Walker, and W. A. Woods. 1977. Speech understanding systems. Report of a steering committee. *Artificial Intelligence. 9*(3), 307-316.

Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and efficiency of open relation extraction. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 447-457).

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 560-568).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality *Advances in neural information processing systems*, *26*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *Advances in Neural Information Processing Systems, 26*.

Derek Miller. 2019. Leveraging BERT for Extractive Text Summarization on Lectures. *arXiv preprint arXiv:1906.04165*.

Thipe Modipa, Febe de Wet, and Marelie H. Davel. 2009. ASR Performance analysis of an experimental call routing system. In *Proceedings of the 20th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2009)* (pp. 127-130).

Thomas S. Morton. 1999. Using coreference for question answering. In *Coreference and Its Applications*.

Christopher J. Mungall, Carlo Torniai, Georgios V. Gkoutos, Suzanna E. Lewis, and Melissa A. Haendel. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biology 13(1), 1-20*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes. 30*(1), 3-26.

Ramesh Nallapati, Bing Xiang, Bowen Zhou, Watson Question, Answering Algorithms, and Yorktown Heights. 2016. Sequence-To-Sequence Rnns For Text Summarization.

International Conference on Learning Representations

Sharan Narang, Gregory Diamos, Erich Elsen, Paulius Micikevicius, Jonah Alben, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 319-327).

Vincent Ng and Claire Cardie. 2001. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 104-111).

Minh Tien Nguyen, Viet Cuong Tran, Xuan Hoai Nguyen, and Le Minh Nguyen. 2019. Web document summarization by exploiting social context with matrix co-factorization. *Information Processing and Management 56*(3), 495-515.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 6859-6866).

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7106-7116).

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A Survey on Open Information Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Mara Nunziatini. 2019. Machine Translation in the Financial Services Industry: A Case Study. *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks* (pp. 57-63).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the*

Demonstrations Session (pp. 48-53).

Lawrence Page and Sergey Brin. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks 30(1-7), 107-117*.

Harinder Pal and Mausam -. 2016. Demonyms and Compound Relational Nouns in Nominal Open IE. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction* (pp. 35-39)

Nikolaos K. Papadakis, Dimitrios Skoutas, Konstantinos Raftopoulos, and Theodora A. Varvarigou. 2005. STAVIES: A system for information extraction from unknown Web data sources through automatic Web wrapper generation using clustering techniques. *IEEE Transactions on Knowledge and Data Engineering. 17*(12), 1638-1652

Dimitris Papadopoulos, Nikolaos Papadakis, and Antonis Litke. 2020. A methodology for open information extraction and representation from large scientific corpora: The CORD-19 data exploration use case. *Applied Sciences (Switzerland) 10(16), 5630*.

Dimitris Papadopoulos, Nikolaos Papadakis, and Nikolaos Matsatsinis. 2021. {PENELOPIE}: Enabling Open Information Extraction for the {G}reek Language through Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 23-29).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. {B}leu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kyubyong Park and Thomas Mulc. 2019. CSS10: A collection of single speaker speech datasets for 10 languages. In *Proceedings of the Annual Conference of the International Speech Communication Association, Proc. Interspeech 2019, 1566-1570*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Thomas R Pedtke. 1997. U.S. Government Support and Use of Machine Translation: Current Status. In *Proceedings of Machine Translation Summit VI: Plenaries* (pp. 3-13).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in*

Natural Language Processing, Proceedings of the Conference (pp. 1532-1543).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 55-61).

Sriram Pingali, Shweta Yadav, Pratik Dutta, and Sriparna Saha. 2021. Multimodal Graph-based Transformer Framework for Biomedical Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3741-3747).

Maja Popović. 2015. Chrf: Character n-gram f-score for automatic mt evaluation. In *10th Workshop on Statistical Machine Translation, WMT 2015 at the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 - Proceedings* (pp. 392-395).

Veera Prathap Reddy M, Prasad P.V.R.D, Manjunath Chikkamath, and Karthikeyan Ponnalagu. 2018. Extracting Conjunction Patterns in Relation Triplets from Complex Requirement Sentence. *International Journal of Computer Trends and Technology 57, 320-332*.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning to predict from textual data. *Journal of Artificial Intelligence Research 45, 641-684*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21, 1-67*.

Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 5357-5367).

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2019. Semantic textual similarity with Siamese neural networks. In *International Conference Recent Advances in Natural Language Processing, RANLP* (pp. 1004-1011).

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-source, multilingual information extraction and summarization* (pp. 93-115). Springer, Berlin, Heidelberg.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global

algorithms for disambiguation to Wikipedia. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 4512-4525).

Nils Reimers and Iryna Gurevych. 2020b. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 3982-3992).

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020a. {M}ulti^2{OIE}: Multilingual Open Information Extraction based on Multi-Head Attention with {BERT}. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1107–1117, Online. Association for Computational Linguistics.

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020b. Multi2OIE: Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020* (pp. 1107-1117).

C. Rossi, F. S. Acerbo, K. Ylinen, I. Juga, P. Nurmi, A. Bosca, F. Tarasconi, M. Cristoforetti, and A. Alikadic. 2018. Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction 30, 145-157*.

Clément Sage, Thibault Douzon, Alex Aussem, Véronique Eglin, Haytham Elghazel, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. 2021. Data-Efficient Information Extraction from Documents with Pre-trained Language Models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (pp. 455-469). Springer, Cham.

Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. 2007. Ontology-based information extraction for business intelligence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 843-856), Springer, Berlin, Heidelberg.

Swarnadeep Saha and Mausam. 2018. Open Information Extraction from Conjunctive Sentences. In *Proceedings ofthe 27th International Conference on Computational Linguistics* (pp. 2288-2299).

Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for numerical open IE. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* (pp. 317-323).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2013. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics 39(4), 847-884*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. CCMatrix: Mining billions of high-quality parallel sentences on the WEB In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 6490-6500).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Paper*. (pp. 86-96).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers* (pp. 1715-1725).

M. Berkan Sesen, Yazann Romahi, and Victor Li. 2018. Natural Language Processing of Financial News. In *Big Data and Machine Learning in Quantitative Investment*, 185.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering 27(2), 443-460*.

Zhang Sheng, Kevin Duh, and Benjamin Van Durme. 2017. MT/IE: Cross-lingual open information extraction with neural sequence-to-sequence models. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference* (pp. 64-70).

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2021. Neural Abstractive Text Summarization with Sequence-to-Sequence Models. *ACM/IMS Transactions on Data Science 2(1), 1-37*.

Dimitar Shterionov, Félix do Carmo, Joss Moorkens, Murhaf Hossari, Joachim Wagner, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2020. A roadmap to neural automatic post-editing: an empirical approach. *Machine Translation 34(2), 67-96*.

Maria Shugrina. 2010. Formatting time-aligned ASR transcripts for readability. In *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference* (pp. 198-206).

Ellery Smith, Dimitris Papadopoulos, Martin Braschler, and Kurt Stockinger. 2022. LILLIE: Information extraction and database integration using linguistics and learning-based algorithms. *Information Systems*, 105:101938.

David R. So, Chen Liang, and Quoc V. Le. 2019. The evolved transformer. In *36th International Conference on Machine Learning, In International Conference on Machine Learning* (pp. 5877-5886). PMLR.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for evidence retrieval and claim verification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 359-366). Springer, Cham.

Felix Stahlberg. 2020. Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research 69, 343-418*.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Dario Stojanovski and Alexander Fraser. 2019. Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2300-2305).

Nikolaos Stylianou and Ioannis Vlahavas. 2021. A neural Entity Coreference Resolution review. *Expert Systems with Applications*, *168*, 114466.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015.

Modeling mention, context and entity with neural networks for entity disambiguation. In *Proceedings of the 24th International Conference on Artificial Intelligence* (pp. 1333-1339).

Mihai Surdeanu, Ramesh Nallapati, and Christopher Manning. 2010. Legal Claim Identification: Information Extraction with Hierarchically Labeled Data. *LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT-2010)* (p. 22).

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems, 27*.

Julian Szymański and MacIej Naruszewicz. 2019. Review on Wikification methods. *AI Communications*, *32*(3), 235-251.

Valentin Tablan, Kalina Bontcheva, Diana Maynard, and Hamish Cunningham. 2003. {OLLIE}: On-Line Learning for Information Extraction. In *Proceedings of the {HLT}-{NAACL} 2003 Workshop on Software Engineering and Architecture of Language Technology Systems ({SEALTS})*, pages 17–24.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-Term memory networks. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference (*pp. 1556-1566).

Chuanqi Tan, Furu Wei, Pengjie Ren, Weifeng Lv, and Ming Zhou. 2017. Entity linking for queries by searching wikipedia sentences. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (pp. 68-77).

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 371-378).

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 809-819).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge {--} Realistic Data Sets for Low Resource and Multilingual {MT}. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. {OPUS-MT} — {B}uilding open translation services for the {W}orld. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal (p. 479).

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature. 571*(7763), 95-98

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems, 30*.

Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth M Anderson. 2011. Natural Language Processing to the Rescue? Extracting ``Situational Awareness" Tweets During Mass Emergency. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (Vol. 5, No. 1, pp. 385-392).*

Andreas Vlachos and Sebastian Riedel. 2015. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science* (pp. 18-22).

Mengqiu Wang and Christopher D. Manning. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference* (pp. 1164-1172).

Peiyi Wang, Lixia Deng, and Xiujun Wu. 2019a. An Automated Fact Checking System Using Deep Learning Through Word Embedding. In *2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019* (pp. 3246-3250). IEEE.

Qicai Wang, Peiyu Liu, Zhenfang Zhu, Hongxia Yin, Qiuyue Zhang, and Lindong Zhang. 2019b. A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning *Applied Sciences*, *9*(21), 4701.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020.

MINILM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems, 33, 5776-5788*.

Jim Webber. 2012. A programmatic introduction to Neo4J. In *SPLASH'12 - Proceedings of the 2012 ACM Conference on Systems, Programming, and Applications: Software for Humanity* (pp. 217-218).

Daniel S. Weld, Raphael Hoffmann, and Fei Wu. 2008. Using wikipedia to bootstrap open information extraction. *ACM Sigmod Record, 37(4), 62-68*.

Sam Wiseman, Alexander M. Rush, Stuart M. Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference* (pp. 1416-1426).

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. (pp. 118-127).

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 6397-6407).

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 120-130).

Tien Hsuan Wu, Zhiyong Wu, Ben Kao, and Pengcheng Yin. 2018. Towards practical open knowledge base canonicalization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 883-892).

Amelie Wührl and Roman Klinger. 2021. Claim Detection in Biomedical Twitter Posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 131-142).

Wei Xiang and Bang Wang. 2019. A Survey of Event Extraction from Text. *IEEE Access, 7, 173111-173137*.

Runyan Yang, Gaofeng Cheng, Haoran Miao, Ta Li, Pengyuan Zhang, and Yonghong Yan. 2021. Keyword Search Using Attention-Based End-to-End ASR and Frame-Synchronous Phoneme Alignments. *IEEE/ACM Transactions on Audio, Speech, and*

Language Processing, 29, 3202-3215.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems, 32*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks.In *International Conference on Learning Representations ICLR2017 (Poster).*

Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. 2007. {T}ext{R}unner: Open Information Extraction on the Web. In *Proceedings of Human Language Technologies: The Annual Conference of the North {A}merican Chapter of the Association for Computational Linguistics ({NAACL}-{HLT})*, pages 25–26, Rochester, New York, USA. Association for Computational Linguistics.

Shi Yuan and Bei Yu. 2018. An Evaluation of Information Extraction Tools for Identifying Health Claims in News Headlines. In *Proceedings of the Workshop Events and Stories in the News 2018* (pp. 34-43).

Junlang Zhan and Hai Zhao. 2020. Span model for open information extraction on accurate corpus. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 9523-9530).

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021a. FaxPlainAC: A Fact-Checking Tool Based on EXPLAINable Models with HumAn Correction in the Loop. In *International Conference on Information and Knowledge Management, Proceedings* (pp. 4823-4827).

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed S. Elsayed, Skatje Myers, and Martha Palmer. 2021b. Fine-grained information extraction from biomedical literature based on knowledge-enriched abstract meaning representation. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 4823-4827).

Kai Zheng, Qiaozhu Mei, Lei Yang, Frank J. Manion, Ulysses J. Balis, and David A. Hanauer. 2011. Voice-dictated versus typed-in clinician notes: linguistic properties and the potential implications on natural language processing. *In AMIA Annual Symposium Proceedings (Vol. 2011, p. 1630). American Medical Informatics Association.*

7 Ευρετηριο Τεχνικων Όρων

L

M

N

O

P

Q

S

T

V

W

A

Γ

E

Κ

Μ

Σ

Τ