Αλέξανδρος Παναγιώτου

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Διπλωματική Εργασία

Easy to Use Java QnA Bot Library

Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Περιεχόμενα

Πρόλογος

Στον ταχέως εξελισσόμενο τομέα των bots ερωτήσεων και απαντήσεων (QnaBots), η ανάγκη για προσαρμόσιμα και φιλικά προς το χρήστη εργαλεία έχει γίνει υψίστης σημασίας. Η παρούσα διατριβή παρουσιάζει μια νέα βιβλιοθήκη Java που έχει σχεδιαστεί για να προσφέρει μια αβίαστη διεπαφή για τη δημιουργία προσαρμοσμένων QnaBots. Το διακριτό στοιχείο της βιβλιοθήκης αυτής είναι η αρθρωτή αρχιτεκτονική της, που επιτρέπει στους χρήστες να προσαρμόζουν διάφορα δομικά στοιχεία με ελάχιστη δυσκολία, συμπεριλαμβανομένης της μεθόδου τεμαχισμού εγγράφων (Chunking Method), το μοντέλο ενσωμάτωσης (Embedding Model), τη μετρική για τους K-Nearest Neighbors (KNN) αλγόριθμο, την τιμή του K στον KNN, και το μοντέλο ολοκλήρωσης (Completion Model) που χρησιμοποιείται για τη δημιουργία της τελικής απάντησης.

Παρουσιάζεται μια ολοκληρωμένη ανασκόπηση του state of the art, η οποία εμβαθύνει σε την ιστορική εξέλιξη των QnABots, τις μεθοδολογίες τεμαχισμού εγγράφων τις εξελίξεις στα μοντέλα ενσωμάτωσης και ολοκλήρωσης και τη συζήτηση μεταξύ της λεπτομερούς προσαρμογής (Fine-Tuning) και της μάθησης εντός πλαισίου (In-Context Learning). Επιπλέον, η διατριβή αντιπαραβάλλει τις διανυσματικές βάσεις δεδομένων (Vector Database) με τις συμβατικές βάσεις δεδομένων, συγκρίνοντας τα αντίστοιχα πλεονεκτημάτα και μειονεκτήματα τους. Ο αλγόριθμος KNN, κομβικής σημασίας στο σχεδιασμό της βιβλιοθήκης, διερευνάται σε βάθος, διευκρινίζοντας το ρόλο του στη στην ενίσχυση των επιδόσεων του QnaBot.

Για να επικυρωθεί η αποτελεσματικότητα της βιβλιοθήκης, διεξήχθη ένα εξαντλητικό πείραμα με τη χρήση κορυφαίων συνόλων δεδομένων αξιολόγησης για μοντέλα γλωσσικής μάθησης (Large Language Models). Τα αποτελέσματα παρέχουν πληροφορίες σχετικά με τις βέλτιστες διαμορφώσεις και αποδεικνύουν την ικανότητα της βιβλιοθήκης να επιτυγχάνει ανταγωνιστικά επίπεδα ακρίβειας. Παρέχεται επίσης ένας οδηγός χρήσης, ο οποίος τονίζει την απλότητα και τον προσανατολισμένο στο χρήστη σχεδιασμό της βιβλιοθήκης.

Συμπερασματικά, η παρούσα διατριβή όχι μόνο συνεισφέρει ένα ευέλικτο εργαλείο στην κοινότητα ανάπτυξης του QnaBot, αλλά προσφέρει επίσης μια συνοπτική αλλά περιεκτική επισκόπηση των υποκείμενων θεωριών και των τρεχουσών τάσεων στον τομέα. Η βιβλιοθήκη αποτελεί απόδειξη των δυνατοτήτων που προσφέρει ο συνδυασμός της σύγχρονης έρευνας με τον φιλικό προς τον χρήστη σχεδιασμό, ανοίγοντας τον δρόμο για μελλοντικές καινοτομίες στον τομέα των QnaBots

Κεφάλαιο 1

Εισαγωγή

Η τεχνητή νοημοσύνη (AI) και η μηχανική μάθηση (ML) έχουν σημειώσει σημαντική πρόοδο τα τελευταία χρόνια, μεταμορφώνοντας τον τρόπο με τον οποίο αντιλαμβανόμαστε και αλληλεπιδρούμε με την τεχνολογία (). Αυτές οι εξελίξεις έχουν ανοίξει το δρόμο για την ανάπτυξη εξελιγμένων συστημάτων που μπορούν να κατανοούν, να μαθαίνουν, να προβλέπουν και ενδεχομένως να λειτουργούν αυτόνομα (). Η ενσωμάτωση της τεχνητής νοημοσύνης και της ML σε διάφορους τομείς έχει οδηγήσει στη δημιουργία πιο διαισθητικών και εξατομικευμένων εμπειριών για τους χρήστες.

Ένα από τα αξιοσημείωτα επιτεύγματα σε αυτόν τον τομέα είναι η ανάπτυξη των ρομπότ ερωτήσεων και απαντήσεων (QnABots), τα οποία έχουν φέρει επανάσταση στον τρόπο αναζήτησης και παροχής πληροφοριών στο διαδίκτυο (). Αυτά τα ρομπότ, που υποστηρίζονται από αλγόριθμους βαθιάς μάθησης και επεξεργασίας φυσικής γλώσσας, μπορούν να κατανοούν και να απαντούν σε ερωτήματα χρηστών σε πραγματικό χρόνο, παρέχοντας ακριβείς και σχετικές απαντήσεις ().

Ωστόσο, η ανάπτυξη τέτοιων εξελιγμένων συστημάτων δεν είναι απλή υπόθεση. Απαιτεί βαθιά κατανόηση διαφόρων δομικών στοιχείων, συμπεριλαμβανομένων μεθόδων ομαδοποίησης εγγράφων, μοντέλων ενσωμάτωσης και μοντέλων συμπλήρωσης (). Επιπλέον, η επιλογή της μετρικής για τον αλγόριθμο K-Nearest Neighbors (KNN) και η τιμή του K μπορούν να επηρεάσουν σημαντικά την απόδοση του QnABot ().

Το κύριο κίνητρο πίσω από αυτή τη διατριβή είναι η ανάπτυξη μιας εύχρηστης βιβλιοθήκης για προσαρμοσμένα QnABots σε Java. Αυτή η βιβλιοθήκη έχει ως στόχο να παρέχει στους χρήστες την ευελιξία να επιλέγουν και να τροποποιούν τα δομικά στοιχεία του QnABot, εξασφαλίζοντας βέλτιστη απόδοση προσαρμοσμένη σε συγκεκριμένες ανάγκες.

Επιπλέον, με την αυξανόμενη εξάρτηση από τα μοντέλα τεχνητής νοημοσύνης και ML, υπάρχει αυξανόμενη ζήτηση για επεξηγηματικότητα και διαφάνεια σε αυτά τα μοντέλα (21). Οι χρήστες και οι προγραμματιστές πρέπει να κατανοούν πώς αυτά τα μοντέλα λαμβάνουν αποφάσεις για να τα εμπιστεύονται και να τα χρησιμοποιούν αποτελεσματικά. Η παρούσα διατριβή εμβαθύνει επίσης στις έννοιες του Fine-tuning έναντι του In-Context Learning και στη σύγκριση μεταξύ διανυσματικών και συμβατικών βάσεων δεδομένων (22).

Κεφάλαιο 2

Ιστορική Αναδρομή

Η εξέλιξη των chatbots είναι ένα συναρπαστικό ταξίδι στα χρονικά της τεχνητής νοημοσύνης (AI) και της επεξεργασίας φυσικής γλώσσας (NLP). Αυτό το κεφάλαιο επιχειρεί να καταγράψει αυτή την εξέλιξη, ρίχνοντας φως στα σημαδιακά ορόσημα, στα τεχνολογικά θεμέλια και στις ευρύτερες επιπτώσεις για την αλληλεπίδραση ανθρώπου-υπολογιστή.

2.1 Η Γένεση: ELIZA

Τα χρονικά της ιστορίας των chatbot συχνά εγκαινιάζονται με την αναφορά του ELIZA, ενός πρωτοποριακού προγράμματος που αναπτύχθηκε στα μέσα της δεκαετίας του 1960 από τον Joseph Weizenbaum στο Ινστιτούτο Τεχνολογίας της Μασαχουσέτης (7). Σχεδιασμένο ως πείραμα για την προσομοίωση ενός ροτζεριανού ψυχοθεραπευτή, το ELIZA βασίστηκε σε μεθοδολογίες αντιστοίχισης προτύπων και υποκατάστασης για την προσομοίωση της συζήτησης. Οι χρήστες εισήγαγαν δηλώσεις και το ELIZA απαντούσε με βάση ένα σύνολο κανόνων σεναρίου, συχνά αντανακλώντας τα λόγια του ίδιου του χρήστη. Παρά την απλότητά του, το ELIZA κατάφερε να πείσει πολλούς χρήστες για την "κατανόησή" του, αναδεικνύοντας τις δυνατότητες της επικοινωνίας μέσω μηχανής. Αυτό το πρώιμο πείραμα υπογράμμισε τις βαθιές επιπτώσεις των μηχανών που μπορούσαν να "συνομιλούν" και έθεσε τις βάσεις για τις μετέπειτα εξελίξεις στον τομέα αυτό.

2.2 Από τα συστήματα βασισμένα σε κανόνες στα συστήματα μάθησης

Στη μετά-ELIZA εποχή εμφανίστηκαν αρκετά συστήματα chatbot, τα περισσότερα από τα οποία είχαν τις ρίζες τους σε παραδείγματα βασισμένα σε κανόνες. Συστήματα όπως το PARRY (8), που αναπτύχθηκε στις αρχές της δεκαετίας του 1970, σχεδιάστηκαν για να προσομοιώνουν συγκεκριμένες προσωπικότητες ή συμπεριφορές, στην περίπτωση του PARRY, έναν ασθενή με παρανοϊκή σχιζοφρένεια. Αυτά τα βασισμένα σε κανόνες συστήματα περιορίζονταν από την εξάρτησή τους από προκαθορισμένα σενάρια, γεγονός που τα καθιστούσε προβλέψιμα και χωρίς προσαρμοστικότητα.

Ωστόσο, καθώς προχωρούσε ο 20ός αιώνας, οι περιορισμοί των συστημάτων που βασίζονται σε κανόνες γίνονταν όλο και πιο εμφανείς. Η δεκαετία του 1990 προανήγγειλε μια νέα εποχή με την εισαγωγή της μηχανικής μάθησης (ML) στις αρχιτεκτονικές chatbot. Αντί να βασίζονται αποκλειστικά σε σκληρά κωδικοποιημένους κανόνες, τα συστήματα αυτά άρχισαν να μαθαίνουν από τα δεδομένα, προσαρμόζοντας και βελτιώνοντας τις απαντήσεις τους με βάση τις αλληλεπιδράσεις (9). Αυτή η αλλαγή σηματοδότησε μια σημαντική απομάκρυνση από τη στατική φύση των προηγούμενων bots, εγκαινιάζοντας μια νέα εποχή δυναμικών, προσανατολισμένων στη μάθηση chatbots.

2.3 Natural Language Processing τη δεκαετία του 2000

Οι αρχές της δεκαετίας του 2000 σηματοδότησαν μια σημαντική περίοδο στην εξέλιξη της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Καθώς ο τομέας μεταπήδησε από συστήματα βασισμένα σε κανόνες σε προσεγγίσεις που βασίζονται περισσότερο σε δεδομένα, προέκυψαν διάφορες εξελίξεις και προκλήσεις. Η διαθεσιμότητα μεγάλων σχολιασμένων σωμάτων δεδομένων και η εμφάνιση αλγορίθμων μηχανικής μάθησης διευκόλυναν αυτή τη μετάβαση. Αντί για χειροκίνητη δημιουργία κανόνων, τα συστήματα εκπαιδεύτηκαν σε δεδομένα για να μαθαίνουν αυτόματα πρότυπα. Αυτή η μετατόπιση ήταν καίριας σημασίας, καθώς επέτρεψε πιο επεκτάσιμες και ισχυρές εφαρμογές NLP.

Η μηχανική μάθηση, ιδίως η μάθηση με επίβλεψη, έγινε η ραχοκοκαλιά πολλών εργασιών NLP. Αλγόριθμοι όπως τα Δέντρα Αποφάσεων (Decision Trees), το Naive Bayes και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) χρησιμοποιήθηκαν συνήθως για εργασίες όπως η ταξινόμηση κειμένου, η ανάλυση συναισθήματος και η επισήμανση μέρους του λόγου.

Μερικά μοντέλα της εποχής που αναδεικνύονται ως τα πλέον σύγχρονα είναι:

1. **Statistical Machine Translation (SMT)**: Κυριαρχώντας στις αρχές της δεκαετίας του 2000, το SMT μετατοπίστηκε από τα συστήματα που βασίζονται σε κανόνες, βασιζόμενη αντίθετα σε τεράστια δίγλωσσα σώματα κειμένων για να διακρίνει τα μεταφραστικά πρότυπα. Το μοντέλο Phrase-Based Machine Translation (PBMT) ήταν ιδιαίτερα αξιοσημείωτο κατά τη διάρκεια αυτής της περιόδου.

2. **Maximum Entropy Models**: Τα μοντέλα αυτά, γνωστά και ως λογαριθμογραμμικά μοντέλα, που χρησιμοποιήθηκαν για εργασίες όπως η επισήμανση μέρους του λόγου και η αναγνώριση ονομαστικών οντοτήτων, ενσωμάτωσαν αυθαίρετα χαρακτηριστικά σε ένα πιθανοτικό πλαίσιο.

3. **Conditional Random Fields (CRFs)**: Εισήχθησαν στις αρχές της δεκαετίας του 2000, τα CRFs έγιναν η πρώτη επιλογή για εργασίες επισήμανσης ακολουθιών. Ως διακριτικά μοντέλα, μπορούσαν να συλλάβουν εξαρτήσεις μεγάλης εμβέλειας και να αφομοιώσουν αυθαίρετα επικαλυπτόμενα χαρακτηριστικά.

4. **Latent Dirichlet Allocation (LDA)**: Ένα παραγωγικό πιθανοτικό μοντέλο που παρουσιάστηκε το 2003, το LDA ήταν καθοριστικό για τη θεματική μοντελοποίηση, εξηγώντας σύνολα παρατηρήσεων με τη χρήση μη παρατηρούμενων ομάδων.

5. **Tree Adjoining Grammars (TAG) and Dependency Parsing**: Αυτά τα μοντέλα ήταν καθοριστικά για τη συντακτική ανάλυση, εγκιβωτίζοντας γλωσσικές δομές με δομημένο τρόπο.

6. **WordNet and Distributional Semantics**: Το WordNet, μια προϋπάρχουσα λεξιλογική βάση δεδομένων, βρήκε εκτεταμένη χρήση στη δεκαετία του 2000 για εργασίες όπως η αποσαφήνιση της σημασίας των λέξεων. Αυτή η δεκαετία σηματοδότησε επίσης την άνοδο των διανεμητικών σημασιολογικών μοντέλων, τα οποία αναπαριστούν τις λέξεις ως διανύσματα για την αποτύπωση των σημασιολογικών αποχρώσεων.

7. **N-gram Language Models**: Θεμελιώδη για πολλές εφαρμογές NLP, ιδίως για την αναγνώριση ομιλίας, τα μοντέλα αυτά προέβλεπαν την επόμενη λέξη σε μια ακολουθία με βάση τις προηγούμενες "n" λέξεις.

Παρά τις προόδους, αρκετές προκλήσεις παρέμειναν:

- **Έλλειψη Annotated Δεδομένων**: Ενώ υπήρχαν μεγάλα σώματα δεδομένων για γλώσσες όπως η αγγλική, πολλές γλώσσες δεν διέθεταν σχολιασμένα σύνολα δεδομένων, γεγονός που εμπόδιζε την ανάπτυξη εργαλείων NLP για αυτές.

- **Ασάφεια Γλώσσας**: Η φυσική γλώσσα είναι εγγενώς διφορούμενη. Οι λέξεις μπορεί να έχουν πολλαπλές σημασίες ανάλογα με το περιβάλλον, οδηγώντας σε προκλήσεις σε εργασίες όπως η αποσαφήνιση της σημασίας των λέξεων.

- **Πολυπλοκότητα Γλώσσας**: Οι ιδιωματισμοί, ο σαρκασμός και οι πολιτισμικές αποχρώσεις προσθέτουν επίπεδα πολυπλοκότητας στις εργασίες NLP.

Οι αρχές της δεκαετίας του 2000 έθεσαν τις βάσεις για τις ραγδαίες εξελίξεις στο NLP που θα ακολουθούσαν τις επόμενες δεκαετίες. Η στροφή σε μεθόδους βασισμένες στα δεδομένα, η ενσωμάτωση της μηχανικής μάθησης και η εξερεύνηση νέων γλωσσών και αρχιτεκτονικών ήταν ενδεικτικές της ανάπτυξης του πεδίου και της ετοιμότητάς του να αντιμετωπίσει πιο σύνθετες προκλήσεις.

2.4 Προκάτοχοι των Large Language Models

Τα τέλη της δεκαετίας του 2000 και οι αρχές της δεκαετίας του 2010 χαρακτηρίστηκαν από ραγδαίες εξελίξεις στην επεξεργασία φυσικής γλώσσας. Η εισαγωγή του Word2Vec από τους Mikolov et al. (4) ήταν μια στιγμή καμπής. Αναπαριστώντας τις λέξεις ως διανύσματα σε έναν χώρο υψηλών διαστάσεων, το Word2Vec μπορούσε να συλλάβει τις σημασιολογικές σχέσεις και τις αποχρώσεις της γλώσσας, ένα σημαντικό άλμα σε σχέση με τα προηγούμενα μοντέλα.

Μετά το Word2Vec, ακολούθησε μια σειρά από καινοτομίες. Οι αρχιτεκτονικές μετασχηματιστών, που ενσαρκώθηκαν από μοντέλα όπως το BERT (5), επέφεραν μια βαθύτερη κατανόηση του πλαισίου της γλώσσας. Το BERT, ειδικότερα, έδειξε τη δύναμη της αμφίδρομης εκπαίδευσης, όπου το μοντέλο μαθαίνει τόσο από το αριστερό όσο και από το δεξί πλαίσιο σε όλα τα επίπεδα, επιτρέποντας μια πιο διαφοροποιημένη κατανόηση του κειμένου.

2.5 Η εποχή των Large Language Models

Στο σημερινό τοπίο της τεχνολογίας chatbot κυριαρχούν τα Large Language Models (LLMs). Αυτά τα μοντέλα, με παράδειγμα το GPT-3 της OpenAI (10), αποτελούν το αποκορύφωμα δεκαετιών έρευνας και ανάπτυξης. Με την ικανότητα να επεξεργάζονται τεράστιες ποσότητες δεδομένων και να παράγουν κείμενο που μοιάζει με ανθρώπινο, τα LLMs έχουν επαναπροσδιορίσει τα όρια του τι μπορούν να επιτύχουν τα chatbots. Η ευελιξία τους είναι εμφανής στο ευρύ φάσμα των εφαρμογών τους, από τη δημιουργία πεζού λόγου μέχρι την απάντηση σύνθετων ερωτημάτων και ακόμη και εργασιών κωδικοποίησης.

Το GPT-3, με τις 175 δισεκατομμύρια παραμέτρους του, αποτελεί παράδειγμα της κλίμακας και της πολυπλοκότητας των σύγχρονων LLMs. Η ικανότητά του να εκτελεί εργασίες χωρίς δεδομένα εκπαίδευσης για συγκεκριμένες εργασίες, βασιζόμενο αντ' αυτού σε λίγα παραδείγματα ή ακόμη και σε μάθηση με μηδενικό πλάνο, αποτελεί απόδειξη της ικανότητας του μοντέλου.

Κεφάλαιο 3

Μέθοδοι ομαδοποίησης εγγράφων

3.1 Εισαγωγή

Οι μέθοδοι τμηματοποίησης (Chunking) είναι απαραίτητες για την επεξεργασία μεγάλου μήκους εγγράφων, ειδικά όταν χρησιμοποιούνται μοντέλα ενσωμάτωσης όπως το BERT που έχουν σταθερό μέγεθος εισόδου. Αυτές οι μέθοδοι διασπούν τα έγγραφα σε διαχειρίσιμα κομμάτια ή τεμάχια, επιτρέποντας την αποτελεσματικότερη επεξεργασία και ανάλυση.

3.2 Μέθοδοι κωδικοποίησης και συγκέντρωσης

Οι μέθοδοι κωδικοποίησης και συγκέντρωσης (Encode-and-pool) έχουν γίνει ακρογωνιαίος λίθος στο πεδίο της επεξεργασίας φυσικής γλώσσας, ιδίως όταν πρόκειται για κείμενα μεταβλητού μήκους. Η πρωταρχική πρόκληση έγκειται στη μετατροπή αυτών των κειμένων σε διανύσματα σταθερού μεγέθους χωρίς να χάνονται η ουσία ή οι αποχρώσεις του αρχικού περιεχομένου. Η κωδικοποίηση και της συγκέντρωση αντιμετωπίζει αυτό το πρόβλημα με την πρώτη κωδικοποίηση κάθε λέξης ή συμβόλου στο κείμενο σε ένα διάνυσμα χρησιμοποιώντας ένα μοντέλο ενσωμάτωσης (Embedding). Στη συνέχεια, αυτά τα διανύσματα συγκεντρώνονται, χρησιμοποιώντας λειτουργίες όπως ο μέσος όρος, η μέγιστη συγκέντρωση ή η συγκέντρωση με βάση την προσοχή, για να παραχθεί ένα ενιαίο διάνυσμα σταθερού μεγέθους.

Ιστορικά, για την αναπαράσταση των εγγράφων χρησιμοποιούνταν μέθοδοι όπως ο Σάκος-Λέξεων (Bag-of-Words) ή η Συχνότητα Όρων - Αντίστροφη Συχνότητα Εγγράφων (Term Frequency - Inverse Document Frequency). Ωστόσο, συχνά απέτυχαν να συλλάβουν τις σημασιολογικές σχέσεις μεταξύ των λέξεων. Η εισαγωγή προ-εκπαιδευμένων ενσωματώσεων λέξεων όπως το Word2Vec (4) και το GloVe έφερε επανάσταση σε αυτόν τον χώρο. Αυτές οι ενσωματώσεις, που εκπαιδεύτηκαν σε μαζικά σώματα κειμένων, μπορούσαν να συλλάβουν σημασιολογικές και συντακτικές σχέσεις μεταξύ των λέξεων. Η επακόλουθη λειτουργία συγκέντρωσης αθροίζει στη συνέχεια αυτές τις ενσωματώσεις για να παράγει μια αναπαράσταση σε επίπεδο εγγράφου.

Η εμφάνιση μοντέλων που βασίζονται σε μετασχηματιστές, όπως το BERT (5), βελτίωσε περαιτέρω αυτή την προσέγγιση. Σε αντίθεση με τις παραδοσιακές ενσωματώσεις που προσφέρουν μια στατική αναπαράσταση για κάθε λέξη, οι μετασχηματιστές παρέχουν ενσω-

ματώσεις με βάση τα συμφραζόμενα. Αυτό σημαίνει ότι η αναπαράσταση μιας λέξης αλλάζει με βάση το περιβάλλον της, προσφέροντας μια πιο διαφοροποιημένη κατανόηση του κειμένου. Μόλις ληφθούν αυτές οι δυναμικές ενσωματώσεις, οι μέθοδοι συγκέντρωσης μπορούν να τις αθροίσουν για να παράγουν μια ολιστική αναπαράσταση του εγγράφου, αποτυπώνοντας τόσο τη σημασιολογία των μεμονωμένων λέξεων όσο και τις μεταξύ τους σχέσεις.

3.3 Αλληλεπίδραση μεταξύ ερωτημάτων και εγγράφων

Η κατανόηση της περίπλοκης σχέσης μεταξύ ενός ερωτήματος και ενός εγγράφου (Query-to-Document) είναι θεμελιώδης για την ανάκτηση πληροφοριών. Οι σύγχρονες τεχνικές στοχεύουν στη ρητή μοντελοποίηση αυτής της σχέσης, εξασφαλίζοντας ότι τα πιο σχετικά μέρη του εγγράφου που αφορούν το ερώτημα αναγνωρίζονται και παρουσιάζονται.

3.3.1 Μηχανισμοί προσοχής

Οι μηχανισμοί προσοχής (Attention Mechanisms) έχουν αναδειχθεί ως ένα βασικό εργαλείο στη μοντελοποίηση της αλληλεπίδρασης μεταξύ ερωτημάτων και εγγράφων. Προερχόμενοι από το πεδίο της νευρωνικής μηχανικής μετάφρασης, οι μηχανισμοί προσοχής σταθμίζουν τη σημασία κάθε λέξης ή συμβόλου στο έγγραφο που αφορά το ερώτημα. Αυτό έχει ως αποτέλεσμα μια σταθμισμένη αναπαράσταση του εγγράφου, όπου τονίζονται τα μέρη που είναι πιο σχετικά με το ερώτημα.

Η ομορφιά των μηχανισμών προσοχής έγκειται στην ικανότητά τους να κατανέμουν δυναμικά τη σημασία με βάση το πλαίσιο. Για παράδειγμα, σε ένα έγγραφο που εξετάζει το ηλιακό σύστημα, η λέξη "Άρης" μπορεί να έχει μεγαλύτερη βαρύτητα όταν το ερώτημα αφορά "πλανήτες" σε αντίθεση με τις "σοκολάτες".

3.3.2 Προσεγγίσεις βασισμένες στην κατάτμηση

Οι προσεγγίσεις που βασίζονται στην τμηματοποίηση (Segmentation-Based) εστιάζουν στη διαίρεση του εγγράφου σε σημαντικά τμήματα ή αποσπάσματα με βάση το ερώτημα. Με τον τρόπο αυτό, οι μέθοδοι αυτές μπορούν να επικεντρωθούν στα πιο συναφή τμήματα του εγγράφου, παραμερίζοντας αποτελεσματικά το άσχετο περιεχόμενο. Αυτό είναι ιδιαίτερα επωφελές για έγγραφα μεγάλου μήκους, όπου μόνο συγκεκριμένα τμήματα μπορεί να είναι σχετικά με το ερώτημα.

Για παράδειγμα, σε ένα εκτενές άρθρο για την ιστορία των υπολογιστών, ένα ερώτημα σχετικά με την "κβαντική πληροφορική" μπορεί να βρει μόνο μερικά τμήματα σχετικά. Οι προσεγγίσεις που βασίζονται στην τμηματοποίηση θα αναδείκνυαν αυτά τα τμήματα, διασφαλίζοντας ότι ο χρήστης λαμβάνει τις πιο σχετικές πληροφορίες χωρίς να κατακλύζεται από περιττές λεπτομέρειες.

3.3.3 Μέθοδοι επανακατάταξης

Οι μέθοδοι επανακατάταξης(Re-ranking), όπως η αρθρωτή προσέγγιση επανακατάταξης που προτάθηκε από τους Gao και Callan (26), προσφέρουν μια εκλεπτυσμένη τεχνική για την αλληλεπίδραση μεταξύ ερωτημάτων και εγγράφων. Αρχικά, χρησιμοποιείται μια ευρεία

μέθοδος ανάκτησης για την ανάκτηση ενός καταλόγου δυνητικά σχετικών εγγράφων. Στη συνέχεια, τα έγγραφα αυτά κατατάσσονται εκ νέου με βάση μια πιο λεπτομερή ανάλυση της συνάφειας τους με το ερώτημα.

Με την κατάτμηση των μεγάλων εγγράφων σε τμήματα και την επανακατάταξή τους με βάση τη συνάφεια του ερωτήματος, οι μέθοδοι αυτές εξασφαλίζουν ότι αναδύονται οι πιο συναφείς πληροφορίες. Αυτή η προσέγγιση δύο βημάτων, που περιλαμβάνει αρχική ανάκτηση ακολουθούμενη από λεπτομερή επανακατάταξη, εξασφαλίζει τόσο την αποτελεσματικότητα όσο και την ακρίβεια στην ανάκτηση εγγράφων.

3.4 Μέθοδος RoR

Η μέθοδος Διάβασμα-Ξαναδιάβασμα (Read-over-Read) ή αλλιώς RoR είναι μια αναδυόμενη τεχνική στον τομέα της επεξεργασίας φυσικής γλώσσας και της ανάκτησης πληροφοριών. Ενώ οι ιδιαιτερότητες της μεθόδου μπορεί να ποικίλλουν ανάλογα με την εφαρμογή της, η βασική αρχή περιστρέφεται γύρω από την επαναληπτική ανάγνωση και επεξεργασία πληροφοριών για την εξαγωγή βαθύτερων γνώσεων και κατανόησης.

Η RoR βασίζεται στην ιδέα ότι ένα απλό πέρασμα σε ένα σύνολο δεδομένων ή ένα έγγραφο μπορεί να μην επαρκεί για την εξαγωγή όλων των σχετικών πληροφοριών ή για την κατανόηση των αποχρώσεων που εμπεριέχονται σε αυτά. Επανεξετάζοντας τα δεδομένα πολλές φορές, η μέθοδος στοχεύει στην τελειοποίηση της κατανόησής τους, οδηγώντας σε πιο ακριβή και ολοκληρωμένα αποτελέσματα.

3.4.1 Εφαρμογές στη βαθιά μάθηση

Στη σφαίρα της βαθιάς μάθησης, η RoR μπορεί να εννοηθεί ως πολλαπλά περάσματα πάνω από ένα σύνολο δεδομένων κατά τη διάρκεια της εκπαίδευσης. Κάθε πέρασμα βελτιώνει τα βάρη και τις προκαταλήψεις του μοντέλου, οδηγώντας σε καλύτερη γενίκευση και απόδοση σε αθέατα δεδομένα. Αυτή η επαναληπτική προσέγγιση μπορεί να είναι ιδιαίτερα επωφελής σε σενάρια όπου τα δεδομένα είναι πολύπλοκα ή όπου οι παραδοσιακές μέθοδοι ενός περάσματος δυσκολεύονται να συγκλίνουν.

3.4.2 Οφέλη και προκλήσεις

Το πρωταρχικό πλεονέκτημα της μεθόδου RoR είναι η δυνατότητά της για αυξημένη ακρίβεια και βάθος κατανόησης. Επιτρέποντας πολλαπλές αναγνώσεις, η μέθοδος μπορεί να αποκαλύψει λεπτές αποχρώσεις και σχέσεις που μπορεί να παραβλέπονται σε ένα μόνο πέρασμα. Ωστόσο, αυτή η επαναληπτική προσέγγιση μπορεί επίσης να είναι υπολογιστικά εντατική, απαιτώντας περισσότερους πόρους και χρόνο από τις παραδοσιακές μεθόδους.

3.4.3 RoR και LLMs

Η εφαρμογή της RoR σε μεγάλα γλωσσικά μοντέλα (LLMs) και chatbots έχει δείξει σημαντικές βελτιώσεις στην ποιότητα του παραγόμενου περιεχομένου. Τα LLMs, με τις τεράστιες γνώσεις τους και την ικανότητά τους να κατανοούν το πλαίσιο, μπορούν να επωφεληθούν από τα πολλαπλά περάσματα του παραγόμενου περιεχομένου για να εξασφαλίσουν ακρίβεια

και συνάφεια. Για τα chatbots, αυτό σημαίνει την παροχή πιο ακριβών και κατάλληλων από άποψη πλαισίου απαντήσεων σε ερωτήματα χρηστών. Η επαναληπτική διαδικασία βελτίωσης της RoR επιτρέπει σε αυτά τα μοντέλα να αυτοδιορθώνονται, οδηγώντας σε μια πιο αξιόπιστη εμπειρία χρήστη.

Επιπλέον, η τυπικότητα των απαντήσεων που παράγονται από τα LLM μπορεί να διαφέρει ανάλογα με τη γλώσσα και το πολιτισμικό πλαίσιο. Πρόσφατες μελέτες, όπως αυτή των Ersoy et al. (3), έχουν τονίσει ότι το επίπεδο τυπικότητας που παρουσιάζουν τα πολύγλωσσα LLMs δεν είναι συνεπές σε όλες τις γλώσσες. Αυτή η ασυνέπεια μπορεί να αποδοθεί στις πολιτισμικές προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσης και στις εγγενείς γλωσσικές δομές των διαφόρων γλωσσών. Για παράδειγμα, ορισμένες γλώσσες μπορεί να έχουν πιο επίσημες δομές και λεξιλόγιο, γεγονός που θα μπορούσε να επηρεάσει τα αποτελέσματα του μοντέλου.

Στο πλαίσιο των chatbots, αυτή η διαφοροποίηση στην τυπικότητα μπορεί να επηρεάσει την εμπειρία του χρήστη. Οι χρήστες ενδέχεται να αναμένουν ένα ορισμένο επίπεδο τυπικότητας με βάση το πολιτισμικό και γλωσσικό τους υπόβαθρο. Εάν οι απαντήσεις του chatbot δεν ευθυγραμμίζονται με αυτές τις προσδοκίες, αυτό θα μπορούσε να οδηγήσει σε παρεξηγήσεις ή ακόμη και σε δυσπιστία. Ως εκ τούτου, η ενσωμάτωση του RoR στα chatbots μπορεί να αποτελέσει μια πραγματικά χρήσιμη προσέγγιση για τη βελτίωση των απαντήσεων ώστε να ευθυγραμμίζονται καλύτερα με το αναμενόμενο επίπεδο τυπικότητας, εξασφαλίζοντας μια πιο πολιτισμικά ευαίσθητη και επικεντρωμένη στον χρήστη αλληλεπίδραση.

Επιπλέον, καθώς τα chatbots βρίσκουν εφαρμογές σε διάφορους τομείς, από την υποστήριξη πελατών έως την υγειονομική περίθαλψη, η σημασία των κατάλληλων για το πλαίσιο και πολιτισμικά ευαίσθητων απαντήσεων καθίσταται υψίστης σημασίας. Το RoR, επιτρέποντας στα μοντέλα να βελτιώνουν επαναληπτικά τα αποτελέσματά τους, μπορεί να διαδραματίσει καθοριστικό ρόλο στην επίτευξη αυτού του στόχου, καθιστώντας τα chatbots πιο αποτελεσματικά και φιλικά προς τον χρήστη σε διαφορετικά σενάρια και ομάδες χρηστών.

3.5 Επιβλεπόμενη αντιθετική μάθηση

Η επιβλεπόμενη αντιπαραθετική μάθηση (Supervised Contrastive Learning) αποτελεί μια συγχώνευση των μεθοδολογιών επιβλεπόμενης μάθησης και αντιπαραθετικής μάθησης. Με την ενσωμάτωση πληροφοριών ετικέτας στο πλαίσιο της αντιπαραβολικής μάθησης, η επιβλεπόμενη αντιπαραθετική μάθηση, ή αλλιώς SCL, στοχεύει στην παραγωγή πιο διακριτικών και ισχυρών αναπαραστάσεων, βελτιώνοντας την απόδοση σε διάφορες εργασίες.

3.5.1 Κατανόηση της αντιθετικής μάθησης

Η αντιθετική μάθηση επικεντρώνεται στη διάκριση μεταξύ παρόμοιων και ανόμοιων περιπτώσεων δεδομένων. Ο πρωταρχικός στόχος είναι η ελαχιστοποίηση της απόστασης μεταξύ αναπαραστάσεων παρόμοιων περιπτώσεων στο χώρο ενσωμάτωσης, ενώ μεγιστοποιείται η απόσταση μεταξύ ανόμοιων περιπτώσεων. Αυτή η διαφοροποίηση επιτυγχάνεται μέσω μιας εξειδικευμένης συνάρτησης αντιθετικής απώλειας.

3.5.2 Η εξέλιξη στην επιβλεπόμενη αντιθετική μάθηση

Η παραδοσιακή αντιθετική μάθηση είναι μη επιβλεπόμενη, δημιουργώντας θετικά και αρνητικά ζεύγη κυρίως μέσω τεχνικών επαύξησης δεδομένων. Ωστόσο, η SCL κάνει ένα βήμα παραπέρα, ενσωματώνοντας ετικέτες κλάσεων. Σε αυτό το παράδειγμα, τα θετικά ζεύγη προέρχονται από την ίδια κλάση, ενώ τα αρνητικά ζεύγη προέρχονται από διαφορετικές κλάσεις. Αυτό διασφαλίζει ότι οι αναπαραστάσεις που προκύπτουν δεν είναι μόνο αμετάβλητες σε επαυξήσεις, αλλά φέρουν επίσης διακριτικές πληροφορίες για συγκεκριμένες κλάσεις.

Καθοριστική συμβολή στον τομέα αυτό είχαν οι Khosla et al. (27). Πρότειναν μια επιβλεπόμενη αντιθετική απώλεια που λειτουργεί σε δύο επίπεδα: εντός των επαυξήσεων μιας μεμονωμένης περίπτωσης και μεταξύ διαφορετικών περιπτώσεων της ίδιας κλάσης. Η μεθοδολογία τους έθεσε νέα σημεία αναφοράς απόδοσης σε διάφορα σύνολα δεδομένων.

3.5.3 Οφέλη και Use Cases

Η κύρια δύναμη της SCL έγκειται στην ικανότητά της να παράγει αναπαραστάσεις με υψηλή διακριτική ικανότητα. Χρησιμοποιώντας ετικέτες κλάσεων, εξασφαλίζει στενότερη εγγύτητα μεταξύ των περιπτώσεων της ίδιας κλάσης στο χώρο ενσωμάτωσης, οδηγώντας σε σαφέστερη διάκριση μεταξύ των κλάσεων. Αυτό ενισχύει εγγενώς την απόδοση ταξινόμησης.

Επιπλέον, η προσαρμοστικότητα της SCL σημαίνει ότι μπορεί να συνεργάζεται με άλλες τεχνικές μάθησης, όπως η αυτοεπιβλεπόμενη (Self-Supervised) μάθηση ή η μάθηση μεταφοράς (Transfer Learning). Η δυνατότητα εφαρμογής της καλύπτει ένα ευρύ φάσμα, από εργασίες οπτικής αναγνώρισης έως περίπλοκες προκλήσεις επεξεργασίας φυσικής γλώσσας.

3.5.4 Πιθανοί περιορισμοί και μελλοντικές προοπτικές

Παρά τα πλεονεκτήματά της, η SCL έχει τις δικές της προκλήσεις. Η απαίτηση για επισημειωμένα δεδομένα μπορεί να είναι περιοριστική, ιδίως όταν οι επισημειώσεις είναι περιορισμένες ή δαπανηρές. Επιπλέον, η αποτελεσματικότητα της SCL μπορεί να εξαρτάται από την επιλογή της αντιθετικής απώλειας και τη στρατηγική για τη δημιουργία ζευγών.

Οι μελλοντικές εξερευνήσεις σε αυτόν τον τομέα θα μπορούσαν να εμβαθύνουν στον μετριασμό αυτών των προκλήσεων, στη βελτιστοποίηση του πλαισίου SCL και στην ενσωμάτωσή του με άλλες νέες τεχνικές μάθησης. Τέτοιες εξελίξεις θα μπορούσαν να οδηγήσουν σε ακόμη πιο ισχυρά μοντέλα με ευρύτερες εφαρμογές.

3.6 Εξαγωγή καταλόγων από έγγραφα

Catalog Extraction from Documents (CED)

Βιβλιογραφία

[1] Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. *When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models*. arXiv preprint arXiv:2010.12858, 2020.

[2] Hlib Babii, Andrea Janes, and Romain Robbes. *Modeling Vocabulary for Big Code Machine Learning*. arXiv preprint arXiv:1904.01873, 2019.

[3] Asım Ersoy, Gerson Vizcarra, Tasmiah Tahsin Mayeesha, and Benjamin Muller. *In What Languages are Generative Language Models the Most Formal? Analyzing Formality Distribution across Languages*. arXiv preprint arXiv:2302.12299, 2023.

[4] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Bidirectional encoder representations from transformers. *arXiv preprint arXiv:1810.04805*.

[6] Tunstall et al. "Natural Language Processing with Transformers"

[7] Weizenbaum, Joseph. "ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine." Communications of the ACM, 1966.

[8] Colby, Kenneth M. "Simulation of Behaviour of Psychopathological Patients." ACM Computing Surveys, 1973.

[9] Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (Vol. 33).

[11] Wallace, Richard S. "The Anatomy of ALICE." Minds and Machines, 1999.

[12] Shen, Zho, et al. "A Knowledge-Grounded Neural Conversation Model." AAAI Conference on Artificial Intelligence, 2017.

[13] Minkov, Einat, et al. "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 2020.

[14] Rajpurkar, Pranav, et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." Empirical Methods in Natural Language Processing, 2016.

[15] Devlin, Jacob, et al. "BERT: Bidirectional Encoder Representations from Transformers." arXiv, 2018.

[16] Radford, Alec, et al. "Improving Language Understanding by Generative Pre-training." OpenAI, 2018.

[17] "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Raffel, Colin, et al.

[18] "RoBERTa: A Robustly Optimized BERT Pretraining Approach", Liu, Myle et al.

[19] "GPT3-to-plan: Extracting plans from text using GPT-3", Olmo et al.

[20] "An Overview of Chatbot Technology", Eleni Adamopoulou & Lefteris Moussiades

[21] "Drug discovery with explainable artificial intelligence.", 2020, Jiménez-Luna, J., Grisoni, F., & Schneider, G.

[22] "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", 2009, K. A. Abdul Nazeer, M. P. Sebastian.

[23] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

[24] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the eighteenth international conference on machine learning, ICML*, 1, 282-289.

[25] Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48-54.

[26] Luyu Gao and Jamie Callan. *Long Document Re-ranking with Modular Re-ranker*. arXiv preprint arXiv:2205.04275, 2022.

[27] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). *Supervised contrastive learning*. In *Advances in neural information processing systems* (Vol. 33).