Αλέξανδρος Παναγιώτου

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Διπλωματική Εργασία

Easy to Use Java QnA Bot Library

Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Περιεχόμενα

Περίληψη

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 Databases) με τις συμβατικές βάσεις δεδομένων, συγκρίνοντας τα αντίστοιχα πλεονεκτημάτα και μειονεκτήματα τους.

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 (42).

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 (13). 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 κανόνων 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 (30), 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 (41). 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. (28) 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 (11), 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 (7), 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.

2.6 Συμπεράσματα

Η ανίχνευση της πορείας των chatbots από το ELIZA μέχρι τα σημερινά μεγαθήρια προσφέρει μια πανοραμική εικόνα των βημάτων που έχουν γίνει στην ΤΝ και το NLP. Κάθε φάση, από τα συστήματα που βασίζονται σε κανόνες έως την εποχή της μηχανικής μάθησης και τη σημερινή κυριαρχία των LLMs, αντανακλά τις ευρύτερες τάσεις στην έρευνα της ΤΝ και τη διαρκώς εξελισσόμενη προσπάθεια για τη δημιουργία μηχανών που κατανοούν και παράγουν ανθρώπινη γλώσσα. Καθώς στεκόμαστε στους ώμους αυτών των γιγάντων, το μέλλον μας καλεί με την υπόσχεση ακόμη πιο εξελιγμένων, ενσυναίσθητων και ευφυών συνομιλητικών πρακτόρων.

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 (28) 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 (11), 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 Query-to-Document Interaction

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 (15), 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 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. (14), 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. (23). 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 (CED)

Η εξαγωγή καταλόγων από έγγραφα (Catalog Extraction from Documents), που συνήθως αναφέρεται ως CED, είναι ένας αναδυόμενος τομέας στην επεξεργασία φυσικής γλώσσας που επικεντρώνεται στην εξαγωγή δομημένων καταλόγων από αδόμητα έγγραφα. Αυτοί οι κατάλογοι, οι οποίοι συχνά χρησιμεύουν ως σκελετός ενός εγγράφου, παρέχουν μια ιεραρχική και οργανωμένη αναπαράσταση του περιεχομένου, διευκολύνοντας την αποτελεσματική ανάκτηση και κατανόηση πληροφοριών.

Οι Chen et al. εισήγαγαν την εργασία CED και πρότειναν ένα πλαίσιο βασισμένο στη μετάβαση (Transition-Based Framework) για την ανάλυση εγγράφων σε δέντρα καταλόγου(9). Η

προσέγγιση αυτή αποσκοπεί στην καταγραφή της ιεραρχικής δομής των εγγράφων, επιτρέποντας την εξαγωγή ουσιαστικών τμημάτων που μπορούν να χρησιμοποιηθούν για διάφορες μεταγενέστερες εργασίες. Οι συγγραφείς πιστεύουν ότι η εργασία CED μπορεί να γεφυρώσει το χάσμα μεταξύ των ακατέργαστων τμημάτων κειμένου και των εργασιών εξαγωγής πληροφοριών, ιδίως για εξαιρετικά μεγάλα έγγραφα.

Ο τομέας της εξαγωγής καταλόγων από έγγραφα υπόσχεται πολλά για τη βελτίωση του τρόπου με τον οποίο επεξεργαζόμαστε και κατανοούμε μεγάλους όγκους κειμένων. Δομώντας τα μη δομημένα δεδομένα, οι τεχνικές CED μπορούν να ανοίξουν το δρόμο για αποδοτικότερη ανάκτηση πληροφοριών, καλύτερη σύνοψη εγγράφων και βελτιωμένη εξαγωγή γνώσης.

Κεφάλαιο 4

Μοντέλα ενσωμάτωσης λέξεων

Τα μοντέλα ενσωμάτωσης (Embedding Models) έχουν φέρει επανάσταση στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP) παρέχοντας πυκνές διανυσματικές αναπαραστάσεις για λέξεις, φράσεις, ακόμη και ολόκληρες προτάσεις. Αυτές οι αναπαραστάσεις, που συχνά αναφέρονται ως ενσωματώσεις, αποτυπώνουν τη σημασιολογική ουσία των στοιχείων κειμένου, επιτρέποντας στις μηχανές να κατανοούν και να επεξεργάζονται την ανθρώπινη γλώσσα πιο αποτελεσματικά (28). Με την πάροδο των ετών, έχει προταθεί πληθώρα μοντέλων ενσωμάτωσης, το καθένα με μοναδική αρχιτεκτονική, μεθοδολογία εκπαίδευσης και φάσμα εφαρμογών. Από το πρωτοποριακό Word2Vec (28) έως το μετασχηματιστικό BERT (11), τα μοντέλα ενσωμάτωσης εξελίσσονται συνεχώς, διευρύνοντας τα όρια της έρευνας και των εφαρμογών του NLP. Αυτό το κεφάλαιο εμβαθύνει στις ιδιαιτερότητες αυτών των μοντέλων, διαφωτίζοντας τις βασικές αρχές, τις μεθοδολογίες και τον βαθύ αντίκτυπό τους στο τοπίο της NLP.

4.1 Word2Vec: Συνεχή μοντέλα Bag-of-Words και Skip-Gram

Το Word2Vec, που παρουσιάστηκε από τους Mikolov et al. το 2013, αποτελεί ένα από τα θεμελιώδη μοντέλα ενσωμάτωσης που έθεσαν τις βάσεις για τις μετέπειτα εξελίξεις στον τομέα (28). Το Word2Vec προσφέρει δύο διαφορετικές αρχιτεκτονικές για τη δημιουργία ενσωμάτωσης λέξεων: Continuous Bag-of-Words (CBOW) και Skip-Gram.

Το μοντέλο CBOW προβλέπει μια λέξη-στόχο με βάση το περιβάλλον της. Με δεδομένο ένα πλαίσιο (ένα σύνολο περιβαλλουσών λέξεων), το CBOW στοχεύει στη μεγιστοποίηση της πιθανότητας εμφάνισης της λέξης-στόχου σε αυτό το πλαίσιο. Στην ουσία, αντιμετωπίζει το πλαίσιο ως είσοδο και προσπαθεί να προβλέψει τη λέξη που ταιριάζει καλύτερα σε αυτό το πλαίσιο (28).

Αντίθετα, το μοντέλο Skip-Gram λειτουργεί προς την αντίθετη κατεύθυνση. Για μια δεδομένη λέξη, στοχεύει στην πρόβλεψη του περιβάλλοντος πλαισίου. Με άλλα λόγια, χρησιμοποιεί μια λέξη ως είσοδο και προσπαθεί να προβλέψει τις λέξεις που είναι πιθανό να εμφανιστούν κοντά της σε ένα κείμενο. Το μοντέλο Skip-Gram έχει αποδειχθεί ιδιαίτερα αποτελεσματικό για μεγάλα σύνολα δεδομένων και για λέξεις με χαμηλότερη συχνότητα, καταγράφοντας ένα ευρύτερο φάσμα σημασιολογικών σχέσεων (28).

Τόσο το CBOW όσο και το Skip-Gram χρησιμοποιούν μια ρηχή αρχιτεκτονική νευρωνικού δικτύου, αλλά οι στόχοι εκπαίδευσής τους διαφέρουν. Η ομορφιά του Word2Vec έγκειται στην ικανότητά του να συλλαμβάνει τόσο τις σημασιολογικές όσο και τις συντακτικές σχέσεις μεταξύ των λέξεων. Λέξεις με παρόμοιες σημασίες τείνουν να βρίσκονται πιο κοντά στο χώρο ενσωμάτωσης, επιτρέποντας εργασίες όπως η αναλογική συλλογιστική (π.χ. ο "άνδρας" είναι με τη "γυναίκα" όπως ο "βασιλιάς" με τη "βασίλισσα") (28).

Το μοντέλο Word2Vec, με την αποδοτικότητα και την αποτελεσματικότητά του, άνοιξε το δρόμο για μια νέα εποχή στο NLP, όπου οι πυκνές διανυσματικές αναπαραστάσεις έγιναν ο κανόνας και οι δυνατότητες των ενσωματώσεων αξιοποιήθηκαν πλήρως.

4.2 GloVe: Διανύσματα για την αναπαράσταση λέξεων

Στον τομέα της ενσωμάτωσης λέξεων, το μοντέλο GloVe (Global Vectors for Word Representation), το οποίο εισήχθη από τους Pennington κ.ά., αποτελεί σημαντική συνεισφορά, γεφυρώνοντας το χάσμα μεταξύ των μεθόδων ενσωμάτωσης που βασίζονται στην καταμέτρηση και των μεθόδων πρόβλεψης (48). Αξιοποιώντας ευφυώς στατιστικές πληροφορίες από μεγάλα σώματα κειμένων, το GloVe προσφέρει έναν ισχυρό και αποτελεσματικό μηχανισμό για την καταγραφή σημασιολογικών και συντακτικών σχέσεων μεταξύ των λέξεων.

Η θεμελιώδης προϋπόθεση του GloVe είναι η σημασία των στατιστικών στοιχείων συνύπαρξης λέξεων για την αποτύπωση του νοήματος. Ενώ οι παραδοσιακές μέθοδοι που βασίζονται στην καταμέτρηση, όπως η Συχνότητα Όρων - Αντίστροφη Συχνότητα Εγγράφων (TF-IDF), χρησιμοποιούν ακατέργαστες μετρήσεις συνύπαρξης, και οι προγνωστικές μέθοδοι, όπως η Word2Vec, προβλέπουν λέξεις συμφραζομένων από λέξεις-στόχους, η GloVe επιτυγχάνει μια ισορροπία. Κατασκευάζει έναν πίνακα συνεμφάνισης από ένα δεδομένο σώμα κειμένων και στη συνέχεια παραγοντοποιεί αυτόν τον πίνακα για να παράγει πυκνά διανύσματα λέξεων (48).

Ένα χαρακτηριστικό γνώρισμα του GloVe είναι η αντικειμενική του συνάρτηση, η οποία έχει σχεδιαστεί για να ελαχιστοποιεί τη διαφορά μεταξύ του τετραγωνικού γινομένου των διανυσμάτων λέξεων και του λογάριθμου των πιθανοτήτων συνύπαρξής τους. Αυτό εξασφαλίζει ότι οι γεωμετρικές σχέσεις μεταξύ των διανυσμάτων λέξεων που προκύπτουν είναι συνεπείς με τις σημασιολογικές τους σχέσεις. Για παράδειγμα, η διαφορά διανύσματος μεταξύ των λέξεων "βασιλιάς" και "άνδρας" θα είναι παρόμοια με τη διαφορά μεταξύ των λέξεων "βασίλισσα" και "γυναίκα", αποτυπώνοντας την ανάλογη σχέση μεταξύ αυτών των ζευγών λέξεων.

Η δύναμη του GloVe έγκειται στην ικανότητά του να συνδυάζει τις συνολικές στατιστικές πληροφορίες (από ολόκληρο το σώμα κειμένων) με τις τοπικές πληροφορίες που σχετίζονται με το πλαίσιο (από συγκεκριμένα ζεύγη λέξεων). Αυτή η διπλή εστίαση επιτρέπει στο GloVe να παράγει ενσωματώσεις που είναι τόσο σημασιολογικά πλούσιες όσο και συναφείς με τα συμφραζόμενα. Επιπλέον, η επεκτασιμότητα του μοντέλου GloVe εξασφαλίζει ότι μπορεί να εκπαιδευτεί σε τεράστια σύνολα δεδομένων, βελτιώνοντας περαιτέρω την ποιότητα των ενσωματώσεων (48).

Συμπερασματικά, το GloVe αποτελεί μια αρμονική συγχώνευση τεχνικών ενσωμάτωσης με βάση την καταμέτρηση και την πρόβλεψη. Αξιοποιώντας τη δύναμη των στατιστικών της συνύπαρξης λέξεων και ενσωματώνοντάς τες σε έναν πυκνό διανυσματικό χώρο, το GloVe προσφέρει ένα ολοκληρωμένο εργαλείο για την καταγραφή των αποχρώσεων και των περιπλοκών της γλώσσας, καθιστώντας το ένα απαραίτητο περιουσιακό στοιχείο στην εργαλειοθήκη

της επεξεργασίας φυσικής γλώσσας.

4.3 FastText: Εξελίξεις στην αναπαράσταση υπολέξεων

Ο τομέας της ενσωμάτωσης λέξεων γνώρισε σημαντική βελτίωση με την εισαγωγή του FastText από τους Bojanowski et al. (**?**). Ενώ τα παραδοσιακά μοντέλα ενσωμάτωσης, όπως το Word2Vec και το GloVe, αναπαριστούν ολόκληρες λέξεις ως ατομικές οντότητες, το FastText εμβαθύνει, αποτυπώνοντας τις μορφολογικές αποχρώσεις των λέξεων λαμβάνοντας υπόψη τα συστατικά των υπολέξεών τους. Αυτή η λεπτομερής προσέγγιση στην αναπαράσταση των λέξεων αντιμετωπίζει διάφορες προκλήσεις που είναι εγγενείς στην επεξεργασία γλωσσών, ιδίως εκείνες που σχετίζονται με τις μορφολογικά πλούσιες γλώσσες και τις λέξεις εκτός λεξιλογίου.

Η θεμελιώδης αρχή του FastText είναι η αναπαράσταση των λέξεων ως σακούλες χαρακτήρων ν-γραμμων(n-grams). Για παράδειγμα, η λέξη "apple" μπορεί να αναλυθεί σε υπολεκτικές μονάδες όπως "<ap", "app", "ppl", "ple", "le>", όπου τα "<" και ">" είναι οριακά σύμβολα που υποδεικνύουν την αρχή και το τέλος μιας λέξης. Αυτή η διάσπαση επιτρέπει στο FastText να καταγράφει την εσωτερική δομή των λέξεων, καθιστώντας το ιδιαίτερα ικανό στην κατανόηση προθημάτων, επιθημάτων και ριζικών λέξεων (**?**).

Ένα από τα πλεονεκτήματα που ξεχωρίζουν αυτής της προσέγγισης των υπολέξεων είναι η ικανότητά της στο χειρισμό λέξεων εκτός λεξιλογίου (out-of-vocabular ή OOV). Οι παραδοσιακές ενσωματώσεις, όταν έρχονται αντιμέτωπες με λέξεις που δεν υπάρχουν στο σώμα εκπαίδευσής τους, συχνά βρίσκονται σε αδυναμία, προεπιλέγοντας γενικές αναπαραστάσεις. Αντίθετα, το FastText, αξιοποιώντας την πληροφορία των υπολέξεων, μπορεί να κατασκευάσει ουσιαστικές αναπαραστάσεις για τις λέξεις OOV με βάση τα n-grams που τις αποτελούν. Αυτή η ικανότητα είναι ιδιαίτερα πολύτιμη για γλώσσες με πλούσια μορφολογία, όπου οι παραλλαγές των λέξεων είναι άφθονες.

Επιπλέον, η έμφαση που δίνει το FastText στις υπολεξικές μονάδες εξασφαλίζει ότι καταγράφει αποτελεσματικότερα τις σημασιολογικές και συντακτικές ομοιότητες. Λέξεις με κοινές ρίζες ή προθέματα, ακόμη και αν διαφέρουν στη συνολική τους μορφή, είναι πιθανό να έχουν ενσωματώσεις που βρίσκονται πιο κοντά στο διανυσματικό χώρο, αντανακλώντας την κοινή τους σημασία ή λειτουργία.

Ο αντίκτυπος του FastText επεκτείνεται πέρα από την ενσωμάτωση λέξεων. Οι αρχές του έχουν προσαρμοστεί για την ενσωμάτωση προτάσεων και εγγράφων, διευρύνοντας περαιτέρω την εφαρμογή του. Λαμβάνοντας υπόψη την ιεραρχική δομή της γλώσσας, από τους χαρακτήρες μέχρι τις υπολέξεις και τις λέξεις, το FastText προσφέρει μια πιο ολιστική και διαφοροποιημένη προσέγγιση στην αναπαράσταση της γλώσσας.

Συνοψίζοντας, το FastText, με την εστίασή του στην πληροφορία των υπολέξεων, έχει εγκαινιάσει μια νέα εποχή στις ενσωματώσεις λέξεων. Αντιμετωπίζοντας τους περιορισμούς των παραδοσιακών μοντέλων και προσφέροντας μια πιο λεπτομερή και ολοκληρωμένη αναπαράσταση, έχει προωθήσει σημαντικά την ικανότητά μας να κατανοούμε και να επεξεργαζόμαστε τη γλώσσα, θέτοντας νέα πρότυπα στον τομέα της επεξεργασίας φυσικής γλώσσας.

4.4 Αρχιτεκτονικές μετασχηματιστών

Το τοπίο της επεξεργασίας φυσικής γλώσσας γνώρισε μια παραδειγματική αλλαγή με την εισαγωγή της αρχιτεκτονικής μετασχηματιστή (Transformer) από τους Vaswani et al. (49). Αποφεύγοντας τα επαναλαμβανόμενα στρώματα που χρησιμοποιούνται παραδοσιακά στα μοντέλα ακολουθίας, η αρχιτεκτονική Transformer βασίζεται σε μηχανισμούς αυτοπροσοχής, επιτρέποντάς της να επεξεργάζεται ακολουθίες εισόδου παράλληλα και όχι διαδοχικά. Αυτός ο καινοτόμος σχεδιασμός όχι μόνο επιταχύνει την εκπαίδευση αλλά και ενισχύει την ικανότητα του μοντέλου να συλλαμβάνει εξαρτήσεις μεγάλης εμβέλειας στα δεδομένα.

Στο επίκεντρο της αρχιτεκτονικής του Transformer βρίσκεται ο μηχανισμός προσοχής (Attention Mechanism), ο οποίος αποδίδει διαφορετικούς βαθμούς σημαντικότητας σε διάφορα τμήματα των δεδομένων εισόδου. Στην ουσία, επιτρέπει στο μοντέλο να "παρακολουθεί" συγκεκριμένα μέρη της εισόδου κατά την παραγωγή μιας εξόδου, διασφαλίζοντας ότι λαμβάνονται υπόψη οι πιο σχετικές πληροφορίες. Ο μηχανισμός αυτοπροσοχής, μια παραλλαγή που χρησιμοποιείται στους Transformers, υπολογίζει βαθμολογίες προσοχής για όλες τις θέσεις σε μια ακολουθία εισόδου σε σχέση μεταξύ τους. Αυτό διασφαλίζει ότι η έξοδος για μια δεδομένη λέξη λαμβάνει υπόψη όλες τις λέξεις της εισόδου, σταθμισμένες με βάση τη σχετικότητά τους (49).

Ένα από τα ιδιαίτερα χαρακτηριστικά της αρχιτεκτονικής Transformer είναι η επεκτασιμότητά της. Με τη στοίβαξη πολλαπλών επιπέδων αυτών των μηχανισμών προσοχής, το μοντέλο μπορεί να συλλάβει περίπλοκα μοτίβα και σχέσεις στα δεδομένα. Επιπλέον, ο σχεδιασμός της προσοχής πολλαπλών κεφαλών επιτρέπει στο μοντέλο να εστιάζει ταυτόχρονα σε διαφορετικές θέσεις, προσφέροντας μια πιο πλούσια αναπαράσταση της εισόδου.

Τα νευρωνικά δίκτυα τροφοδότησης του Transformer, σε συνδυασμό με την κωδικοποίηση κατά θέση, εξασφαλίζουν ότι λαμβάνεται υπόψη η σειρά των λέξεων, αντισταθμίζοντας την έλλειψη επανάληψης στην αρχιτεκτονική. Αυτή η σχεδιαστική επιλογή διασφαλίζει ότι το μοντέλο παραμένει ευαίσθητο στην ακολουθία των λέξεων, μια κρίσιμη πτυχή της γλωσσικής κατανόησης (49).

Ο μετασχηματιστικός αντίκτυπος της αρχιτεκτονικής Transformer είναι εμφανής στην ευρεία υιοθέτησή της και στην πληθώρα μοντέλων που έχει γεννήσει, όπως τα BERT, GPT και T5. Αυτά τα μοντέλα, που βασίζονται στις θεμελιώδεις αρχές των Transformers, έχουν θέσει νέα σημεία αναφοράς σε μια πληθώρα εργασιών NLP, από τη μετάφραση έως τη δημιουργία κειμένου.

Συμπερασματικά, η αρχιτεκτονική Transformer, με την έμφαση που δίνει στους μηχανισμούς προσοχής, αποτελεί ένα μνημειώδες άλμα στον τομέα της επεξεργασίας φυσικής γλώσσας. Δίνοντας προτεραιότητα στον παραλληλισμό και την επεκτασιμότητα και αποτυπώνοντας με δεξιοτεχνία τις περιπλοκές της γλώσσας, άνοιξε το δρόμο για μια νέα γενιά μοντέλων τελευταίας τεχνολογίας.

4.5 BERT: Αμφίδρομοι μετασχηματιστές για την κατανόηση της γλώσσας

Στο διαρκώς εξελισσόμενο τοπίο των μοντέλων ενσωμάτωσης, το BERT (Bidirectional Encoder Representations from Transformers) αναδείχθηκε ως μια μετασχηματιστική δύναμη,

επαναπροσδιορίζοντας τα κριτήρια αναφοράς για μια πληθώρα εργασιών NLP (11). Προτεινόμενη από τους Devlin et al., η αρχιτεκτονική και η μεθοδολογία εκπαίδευσης του BERT τον διαφοροποιούν από τους προκατόχους του, επιτρέποντάς του να καταγράφει τις περιπλοκές της γλώσσας με έναν απαράμιλλο τρόπο.

Σε αντίθεση με τα παραδοσιακά μοντέλα που επεξεργάζονται τις λέξεις είτε από αριστερά προς τα δεξιά είτε από δεξιά προς τα αριστερά, το BERT αξιοποιεί μια αμφίδρομη προσέγγιση. Με την ταυτόχρονη εξέταση τόσο του προηγούμενου όσο και του επόμενου πλαισίου μιας λέξης, η BERT καταγράφει μια πλουσιότερη και πιο ολιστική κατανόηση των κειμενικών πληροφοριών. Αυτή η αμφίδρομη επεξεργασία συμφραζομένων επιτυγχάνεται με τη χρήση της αρχιτεκτονικής Transformer, που αναφέρθηκε στο προηγούμενο υποκεφάλαιο.

Ένα από τα ιδιαίτερα χαρακτηριστικά του BERT είναι η μεθοδολογία προ-εκπαίδευσης. Αντί να εκπαιδεύεται από το μηδέν για συγκεκριμένες εργασίες, η BERT προ-εκπαιδεύεται σε τεράστιες ποσότητες κειμένου χρησιμοποιώντας δύο μη επιβλεπόμενες εργασίες: μοντελοποίηση καλυμμένης γλώσσας και πρόβλεψη επόμενης πρότασης. Στην εργασία μοντελοποίησης καλυμμένης γλώσσας, τυχαίες λέξεις σε μια πρόταση αντικαθίστανται με ένα σύμβολο [MASK] και η BERT εκπαιδεύεται για να προβλέψει την αρχική λέξη. Αυτή η εργασία υποχρεώνει την BERT να κατανοήσει σε βάθος το πλαίσιο. Η εργασία πρόβλεψης της επόμενης πρότασης, από την άλλη πλευρά, εκπαιδεύει την BERT να προσδιορίζει αν δύο προτάσεις είναι διαδοχικές σε ένα κείμενο, ενισχύοντας περαιτέρω την κατανόηση των κειμενικών σχέσεων (11).

Αφού προ-εκπαιδευτεί, η BERT μπορεί να ρυθμιστεί λεπτομερώς για συγκεκριμένες εργασίες με ελάχιστα πρόσθετα δεδομένα εκπαίδευσης. Αυτή η προσέγγιση μάθησης μεταφοράς, όπου η γνώση που αποκτήθηκε κατά την προ-εκπαίδευση μεταφέρεται σε συγκεκριμένες εργασίες, επέτρεψε στην BERT να επιτύχει κορυφαία αποτελέσματα σε ένα ευρύ φάσμα κριτηρίων αναφοράς NLP, από την απάντηση ερωτήσεων έως την ανάλυση συναισθήματος.

Στην ουσία, η BERT αντιπροσωπεύει μια αλλαγή πορείας στο NLP. Αξιοποιώντας τη δύναμη της αμφίδρομης επεξεργασίας συμφραζομένων, των βαθιών μετασχηματιστών και της καινοτόμου προ-εκπαίδευσης, προσφέρει μια ολοκληρωμένη και διαφοροποιημένη κατανόηση της γλώσσας, θέτοντας νέα στάνταρ για τα γλωσσικά μοντέλα και τις τεχνικές ενσωμάτωσης.

4.6 Universal Sentence Encoder: Ενσωματώσεις προτάσεων

Ο Universal Sentence Encoder (USE), που παρουσιάστηκε από τους Cer et al. (51), αντιπροσωπεύει ένα σημαντικό άλμα προς τα εμπρός στον τομέα της ενσωμάτωσης προτάσεων. Ενώ τα παραδοσιακά μοντέλα ενσωμάτωσης επικεντρώνονται κυρίως σε αναπαραστάσεις σε επίπεδο λέξης, ο USE στοχεύει στην παροχή υψηλής ποιότητας ενσωμάτωσης για μεγαλύτερες ακολουθίες κειμένου, όπως προτάσεις και παράγραφοι. Αξιοποιώντας τη δύναμη της μάθησης μεταφοράς, το USE προσφέρει ενσωματώσεις που είναι τόσο ευέλικτες όσο και αποδοτικές σε ένα ευρύ φάσμα εργασιών επεξεργασίας φυσικής γλώσσας.

Στο επίκεντρο του Universal Sentence Encoder βρίσκεται η ιδέα της εκπαίδευσης σε μια ποικιλία πηγών δεδομένων και εργασιών και, στη συνέχεια, της μεταφοράς αυτής της διδαχθείσας γνώσης σε συγκεκριμένες εφαρμογές. Η προσέγγιση αυτή είναι εμπνευσμένη από την επιτυχία της μάθησης μεταφοράς στον τομέα της όρασης υπολογιστών, όπου τα προεκπαιδευμένα μοντέλα σε μεγάλα σύνολα δεδομένων μπορούν να ρυθμιστούν λεπτομερώς για πιο εξειδικευμένες εργασίες με περιορισμένα δεδομένα (51).

Η αρχιτεκτονική του USE έχει σχεδιαστεί για να αποτυπώνει το σημασιολογικό νόημα

των προτάσεων. Χρησιμοποιεί ένα δίκτυο βαθιάς μέσης ανάλυσης (Deep Averaging Network), όπου οι ενσωματώσεις για λέξεις και bi-grams πρώτα υπολογίζονται κατά μέσο όρο και στη συνέχεια περνούν μέσα από ένα βαθύ νευρωνικό δίκτυο τροφοδοσίας. Επιπλέον, χρησιμοποιείται επίσης μια αρχιτεκτονική βασισμένη σε μετασχηματιστές, των Vaswani et al. (49), για να αποτυπώσει τις περίπλοκες σχέσεις μεταξύ των λέξεων σε μια πρόταση.

Ένα από τα χαρακτηριστικά που ξεχωρίζουν στον Universal Sentence Encoder είναι η ικανότητά του να παράγει ενσωματώσεις που έχουν σημασιολογικό νόημα. Προτάσεις με παρόμοια νοήματα, ακόμη και αν έχουν διαφορετικές λεξιλογικές συνθέσεις, απεικονίζονται κοντά η μία στην άλλη στο χώρο ενσωμάτωσης. Αυτή η ιδιότητα είναι ανεκτίμητη για εργασίες όπως η σημασιολογική ομοιότητα κειμένου, η ομαδοποίηση και η ταξινόμηση.

Επιπλέον, η έμφαση του USE στη μάθηση μεταφοράς διασφαλίζει ότι δεν περιορίζεται μόνο σε ένα συγκεκριμένο έργο. Αφού εκπαιδευτεί, ο κωδικοποιητής μπορεί να χρησιμοποιηθεί σε πλήθος εφαρμογών χωρίς την ανάγκη για δεδομένα εκπαίδευσης για συγκεκριμένη εργασία. Αυτή η δυνατότητα γενίκευσης, σε συνδυασμό με την υψηλή απόδοσή του, καθιστά τον USE ένα πολύτιμο εργαλείο για ένα ευρύ φάσμα προκλήσεων NLP.

Εν κατακλείδι, ο Universal Sentence Encoder, με την εστίασή του στη μάθηση μεταφοράς και τις ενσωματώσεις σε επίπεδο πρότασης, έχει επαναπροσδιορίσει το τοπίο της αναπαράστασης κειμένου. Γεφυρώνοντας το χάσμα μεταξύ γενικευσιμότητας και απόδοσης, προσφέρει μια στιβαρή και ευέλικτη λύση για την αποτύπωση της σημασιολογικής ουσίας του κειμένου, θέτοντας ένα νέο σημείο αναφοράς στο πεδίο της κατανόησης της φυσικής γλώσσας.

4.7 ELMo: Ενσωματώσεις από Γλωσσικά Μοντέλα

Η αναζήτηση πλουσιότερων και εκφραστικότερων αναπαραστάσεων λέξεων αποτέλεσε κινητήρια δύναμη στην εξέλιξη της επεξεργασίας φυσικής γλώσσας. Το ELMo, το οποίο σημαίνει "Ενσωματώσεις από γλωσσικά μοντέλα" (Embeddings from Language Models), που παρουσιάστηκε από τους Peters et al. (52), αντιπροσωπεύει μια αλλαγή πορείας σε αυτό το ταξίδι. Ξεπερνώντας τις παραδοσιακές στατικές ενσωματώσεις, το ELMo προσφέρει βαθιές αναπαραστάσεις λέξεων με βάση το περιβάλλον, αποτυπώνοντας τόσο συντακτικές όσο και σημασιολογικές αποχρώσεις με βάση το περιβάλλον των λέξεων.

Οι παραδοσιακές ενσωματώσεις λέξεων, όπως οι Word2Vec και GloVe, παρέχουν μια ενιαία διανυσματική αναπαράσταση για κάθε λέξη, ανεξάρτητα από το περιεχόμενό της. Ενώ αυτές οι ενσωματώσεις έχουν αποδειχθεί πολύτιμες σε πολλές εργασίες NLP, δεν έχουν εγγενώς την ικανότητα να καταγράφουν την πολυσημία, όπου μια λέξη μπορεί να έχει πολλαπλές σημασίες ανάλογα με τη χρήση της. Το ELMo αντιμετωπίζει αυτόν τον περιορισμό δημιουργώντας ενσωματώσεις που είναι συνάρτηση ολόκληρης της πρότασης εισόδου, διασφαλίζοντας ότι η αναπαράσταση είναι ευαίσθητη στο πλαίσιο της λέξης.

Η υποκείμενη αρχιτεκτονική του ELMo έχει τις ρίζες της στα δίκτυα αμφίδρομης μακράς βραχυπρόθεσμης μνήμης (Bidirectional Long Short-Term Memory). Εκπαιδεύοντας ένα βαθύ γλωσσικό μοντέλο BiLSTM σε ένα μεγάλο σώμα κειμένων, το ELMo μαθαίνει να προβλέπει την πιθανότητα μιας λέξης δεδομένου του παρελθόντος και του μελλοντικού της πλαισίου. Στη συνέχεια, οι ενσωματώσεις προκύπτουν από τις εσωτερικές καταστάσεις αυτού του γλωσσικού μοντέλου, με αποτέλεσμα αναπαραστάσεις που ενσωματώνουν πληροφορίες από όλα τα επίπεδα του LSTM, τόσο από αριστερά προς τα δεξιά όσο και από δεξιά προς τα αριστερά.

Ένα ξεχωριστό χαρακτηριστικό του ELMo είναι η ικανότητά του να παράγει ενσωμάτωση

για συγκεκριμένες εργασίες. Αντί να χρησιμοποιεί τις ενσωματώσεις ως σταθερά χαρακτηριστικά, το ELMo συνιστά την ενσωμάτωση του γλωσσικού μοντέλου BiLSTM απευθείας στην εργασία-στόχο και τη λεπτομερή ρύθμισή του, επιτρέποντας την προσαρμογή των ενσωματώσεων σε συγκεκριμένες εφαρμογές (52). Αυτή η προσέγγιση έχει οδηγήσει σε σημαντική αύξηση της απόδοσης σε ένα ευρύ φάσμα εργασιών, από την ανάλυση συναισθήματος (sentiment analysis) έως την αναγνώριση ονομαστικών οντοτήτων (named entity recognition).

Μια άλλη αξιοσημείωτη πτυχή του ELMo είναι η ευελιξία του. Οι ενσωματώσεις μπορούν εύκολα να ενσωματωθούν σε υπάρχοντα μοντέλα, οδηγώντας συχνά σε άμεση βελτίωση της απόδοσης. Αυτός ο plug-and-play χαρακτήρας, σε συνδυασμό με το βάθος των αναπαραστάσεων, έχει καταστήσει το ELMo μια δημοφιλή επιλογή μεταξύ ερευνητών και επαγγελματιών.

Εν κατακλείδι, η εισαγωγή από το ELMo βαθιών αναπαραστάσεων λέξεων με βάση τα συμφραζόμενα έχει εγκαινιάσει μια νέα εποχή στον τομέα των ενσωματώσεων λέξεων. Καταγράφοντας την περίπλοκη αλληλεπίδραση του συντακτικού και της σημασιολογίας στη γλώσσα, το ELMo όχι μόνο προώθησε την κατανόησή μας για τις αναπαραστάσεις λέξεων, αλλά και έθεσε νέα σημεία αναφοράς επιδόσεων σε μια πληθώρα εργασιών NLP, υπογραμμίζοντας τη δυνατότητα των context-aware embeddings να οδηγήσουν το επόμενο κύμα καινοτομιών στον τομέα.

4.8 DistilBERT: Απόσταξη του BERT

Η έλευση του BERT έφερε επανάσταση στο τοπίο της επεξεργασίας φυσικής γλώσσας, θέτοντας νέα σημεία αναφοράς στην απόδοση σε μια πληθώρα εργασιών. Ωστόσο, το τεράστιο μέγεθος και οι υπολογιστικές απαιτήσεις του BERT δημιούργησαν προκλήσεις, ιδίως για εφαρμογές πραγματικού χρόνου και εφαρμογές σε συσκευές με περιορισμένους πόρους. Το DistilBERT, το οποίο παρουσιάστηκε από τους Sanh et al. (53), αναδύθηκε ως λύση σε αυτό το αίνιγμα, προσφέροντας μια πιο λιτή αλλά ιδιαίτερα αποτελεσματική έκδοση του BERT μέσω της διαδικασίας απόσταξης γνώσης.

Η απόσταξη γνώσης, στον πυρήνα της, περιλαμβάνει την εκπαίδευση ενός μικρότερου μοντέλου (του μαθητή) ώστε να μιμείται τη συμπεριφορά ενός μεγαλύτερου, πιο σύνθετου μοντέλου (του δασκάλου). Το μαθητικό μοντέλο μαθαίνει προσπαθώντας να αναπαράγει την κατανομή εξόδου του μοντέλου του δασκάλου, αντί να βελτιστοποιεί άμεσα την βασική αλήθεια. Αυτή η διαδικασία επιτρέπει στο μαθητή να κληρονομήσει τις δυνατότητες γενίκευσης του δασκάλου, οδηγώντας συχνά σε επιδόσεις που διαψεύδουν το μικρότερο μέγεθός του.

Το DistilBERT ενσαρκώνει αυτή την αρχή, έχοντας περίπου το μισό μέγεθος του BERT, αλλά διατηρώντας το 95 % της απόδοσής του σε εργασίες ταξινόμησης προτάσεων (53). Η διαδικασία "απόσταξης" περιλαμβάνει την εκπαίδευση του DistilBERT στα ίδια δεδομένα με το BERT, αλλά χρησιμοποιώντας ως οδηγό τις μαλακές κατανομές στόχων που παράγονται από το BERT. Αυτή η προσέγγιση διασφαλίζει ότι το DistilBERT όχι μόνο συλλαμβάνει τη ρητή γνώση που κωδικοποιείται στα επισημασμένα δεδομένα αλλά επωφελείται επίσης από την έμμεση γνώση που ενσωματώνεται στα περίπλοκα μοτίβα του BERT.

Τα οφέλη του DistilBERT είναι πολύπλευρα. Πρώτον, το μειωμένο μέγεθός του μεταφράζεται σε ταχύτερους χρόνους εξαγωγής συμπερασμάτων, καθιστώντας το πιο κατάλληλο για εφαρμογές πραγματικού χρόνου. Δεύτερον, το μικρότερο αποτύπωμα του μοντέλου σημαίνει ότι απαιτεί λιγότερη μνήμη, διευκολύνοντας την ανάπτυξη σε συσκευές άκρων και κινητές πλατφόρμες. Τέλος, ο μειωμένος αριθμός παραμέτρων σημαίνει ότι το DistilBERT είναι λιγό-

τερο επιρρεπές στην υπερπροσαρμογή, ιδίως όταν γίνεται λεπτομερής ρύθμιση σε μικρότερα σύνολα δεδομένων.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι ενώ το DistilBERT προσφέρει μια συναρπαστική ισορροπία μεταξύ μεγέθους και επιδόσεων, ενδέχεται να υπάρχουν συγκεκριμένες εργασίες ή σενάρια όπου απαιτείται η πλήρης ικανότητα του BERT. Η επιλογή μεταξύ BERT και DistilBERT θα πρέπει να καθοδηγείται από τις συγκεκριμένες απαιτήσεις της εφαρμογής, τους διαθέσιμους υπολογιστικούς πόρους και τον επιθυμητό συμβιβασμό μεταξύ ταχύτητας και ακρίβειας.

Συνοψίζοντας, το DistilBERT αποτελεί απόδειξη των δυνατοτήτων της απόσταξης γνώσης για τη δημιουργία αποδοτικών και αποτελεσματικών νευρωνικών μοντέλων. Ενσωματώνοντας την ουσία του BERT σε μια πιο συμπαγή μορφή, το DistilBERT έχει εκδημοκρατίσει την πρόσβαση στις δυνατότητες επεξεργασίας φυσικής γλώσσας τελευταίας τεχνολογίας, ανοίγοντας το δρόμο για ευρύτερες και πιο ποικίλες εφαρμογές στον πραγματικό κόσμο.

Κεφάλαιο 5

Μοντέλα ολοκλήρωσης

Τα μοντέλα συμπλήρωσης (Completion Models) έχουν αναδειχθεί στον ακρογωνιαίο λίθο στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP), ιδίως σε εργασίες που απαιτούν τη δημιουργία ή τη συμπλήρωση ακολουθιών κειμένου. Αυτά τα μοντέλα εκπαιδεύονται για να προβλέπουν τα επόμενα σημεία σε μια ακολουθία, καθιστώντας τα ανεκτίμητα για μια πληθώρα εφαρμογών, από τη δημιουργία κειμένου μέχρι τις αλληλεπιδράσεις chatbot. Αυτό το κεφάλαιο εμβαθύνει στις θεμελιώδεις έννοιες, τις αρχιτεκτονικές και τις εξελίξεις στα μοντέλα ολοκλήρωσης, θέτοντας τις βάσεις για την εφαρμογή τους σε συστήματα chatbot.

5.1 Οι βάσεις των μοντέλων συμπλήρωσης

Τα μοντέλα ολοκλήρωσης, στον πυρήνα τους, αποσκοπούν στην πρόβλεψη ή την ολοκλήρωση μιας δεδομένης ακολουθίας με βάση προηγούμενες πληροφορίες. Αυτά τα μοντέλα έχουν βρει εκτεταμένες εφαρμογές σε διάφορους τομείς, συμπεριλαμβανομένων των συστημάτων chatbot, όπου προβλέπουν την επόμενη λέξη ή φράση σε μια συνομιλία. Τα θεμέλια αυτών των μοντέλων βρίσκονται στην κατανόηση των υποκείμενων μαθηματικών και στατιστικών αρχών που καθοδηγούν τις προβλέψεις τους.

5.2 GPT: Αυτοπαλίνδρομη Γλωσσική Μοντελοποίηση

Οι Γενετικοί Προεκπαιδευμένοι Μετασχηματιστές (Generative Pre-trained Transformers ή GPT) έχουν φέρει επανάσταση στο τοπίο της επεξεργασίας φυσικής γλώσσας, ιδίως στον τομέα της γλωσσικής μοντελοποίησης. Τα μοντέλα GPT, που εισήχθησαν από το OpenAI, βασίζονται στην αρχιτεκτονική του μετασχηματιστή και χρησιμοποιούν μια αυτοπαλινδρομική (autoregressive) προσέγγιση για τη δημιουργία συνεκτικών και σχετικών με το πλαίσιο ακολουθιών κειμένου (34).

5.2.1 Αυτοπαλινδρομικά Μοντέλα

Τα αυτοπαλινδρομικά μοντέλα (Autoregressive Models) προβλέπουν το επόμενο σύμβολο σε μια ακολουθία με βάση τα προηγούμενα σύμβολα. Μαθηματικά, η πιθανότητα μιας ακο-

λουθίας $x_1, x_2, \ldots, x_T$ παραγοντοποιείται ως εξής:

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t | x_1, \ldots, x_{t-1})$$

Στο πλαίσιο του GPT, αυτή η αυτοπαλινδρομική ιδιότητα αξιοποιείται για τη δημιουργία κειμένου με πρόβλεψη ενός συμβόλου κάθε φορά, υπό την προϋπόθεση των προηγούμενων συμβόλων.

5.2.2 Σημαντικότητα στις αλληλεπιδράσεις chatbot

Η ικανότητα των μοντέλων GPT να παράγουν συνεκτικό και σχετικό με το πλαίσιο κείμενο τα έχει καταστήσει ιδιαίτερα πολύτιμα για αλληλεπιδράσεις chatbot. Μερικά αξιοσημείωτα πλεονεκτήματα περιλαμβάνουν:

- **Contextual Understanding**: Τα μοντέλα GPT μπορούν να διατηρούν το πλαίσιο σε μακροχρόνιες συνομιλίες, επιτρέποντας στα chatbots να παρέχουν σχετικές απαντήσεις με βάση προηγούμενες αλληλεπιδράσεις.
- **Ευελιξία**: Σε αντίθεση με τα chatbots που βασίζονται σε κανόνες, τα chatbots με GPT μπορούν να χειριστούν ένα ευρύ φάσμα ερωτημάτων χρηστών, ακόμη και αυτά που δεν έχουν εμφανιστεί κατά τη διάρκεια της εκπαίδευσης.
- **Μάθηση λίγων στιγμών**: (**?**), τα μοντέλα GPT μπορούν να εκτελούν εργασίες με ελάχιστα παραδείγματα, καθιστώντας τα προσαρμόσιμα σε ποικίλα σενάρια chatbot.

5.2.3 Προκλήσεις και προβληματισμοί

Ενώ τα μοντέλα GPT προσφέρουν εντυπωσιακές δυνατότητες, υπάρχουν προκλήσεις που πρέπει να ληφθούν υπόψη:

- **Επαληθευσιμότητα**: Τα μοντέλα GPT μπορούν να παράγουν πληροφορίες που μπορεί να μην είναι πάντα ακριβείς ή πραγματικές.
- **Ηθικές ανησυχίες**: Τα μοντέλα μπορεί μερικές φορές να παράγουν μεροληπτικό ή ακατάλληλο περιεχόμενο, γεγονός που απαιτεί προσεκτική παρακολούθηση και φιλτράρισμα.
- **Κόστος υπολογισμού**: Η εκπαίδευση μοντέλων GPT μεγάλης κλίμακας απαιτεί σημαντικούς υπολογιστικούς πόρους.

Στις πρόσφατες εξελίξεις, μοντέλα όπως το GPT-NeoX-20B έχουν διευρύνει τα όρια της αυτοπαλινδρομικής γλωσσικής μοντελοποίησης, παρουσιάζοντας τη δυνατότητα για ακόμη πιο ισχυρές αλληλεπιδράσεις chatbot στο μέλλον (**?**).

5.2.4 Συμπεράσματα

Η σειρά GPT, με τη δυνατότητα αυτοπαλινδρομικής γλωσσικής μοντελοποίησης, έχει θέσει νέα πρότυπα για τις αλληλεπιδράσεις chatbot. Καθώς η έρευνα εξελίσσεται, αναμένεται ότι αυτά τα μοντέλα θα γίνουν ακόμη πιο αποτελεσματικά, προσαρμόσιμα και αναπόσπαστα για το μέλλον της συνομιλιακής τεχνητής νοημοσύνης.

5.3 Ενσωμάτωση γνώσης σε μοντέλα συμπλήρωσης

Η ολοκλήρωση γνώσης αναφέρεται στη διαδικασία συνδυασμού, αφομοίωσης και σύνθεσης πληροφοριών από διάφορες πηγές για τη δημιουργία μιας ολοκληρωμένης κατανόησης ή αναπαράστασης ενός συγκεκριμένου τομέα. Στο πλαίσιο των μοντέλων συμπλήρωσης, αυτό συνεπάγεται τη συγχώνευση διαφορετικών βάσεων γνώσης, που κυμαίνονται από γλωσσικά πρότυπα έως πληροφορίες ειδικού τομέα, για την παραγωγή πιο ακριβών, συνειδητοποιημένων στο πλαίσιο και ευέλικτων συμπληρωμάτων (**?**).

5.3.1 Θεωρητικές βάσεις

Η θεωρητική θεμελίωση της ολοκλήρωσης της γνώσης μπορεί να αναχθεί στον τομέα της Μηχανικής της Γνώσης, ο οποίος τονίζει τη σημασία της μοντελοποίησης και της αναπαράστασης της γνώσης με δομημένο τρόπο. Οι Studer et al. (**?**) συζήτησαν τη μετατόπιση του παραδείγματος από την άποψη της μεταφοράς στη μοντελοποίηση και τόνισαν τη σημασία των μεθόδων περιορισμού ρόλων και των γενικών εργασιών στη μηχανική της γνώσης. Αυτές οι αρχές μπορούν να επεκταθούν στα μοντέλα συμπλήρωσης, όπου οι δομημένες αναπαραστάσεις γνώσης μπορούν να ενισχύσουν σημαντικά την ικανότητα του μοντέλου να παράγει συμπληρώσεις σχετικές με το πλαίσιο.

5.3.2 Μαθηματική αναπαράσταση

Ας θεωρήσουμε ένα μοντέλο συμπλήρωσης M που στοχεύει στη δημιουργία μιας συμπλήρωσης C για μια δεδομένη είσοδο I. Η διαδικασία ολοκλήρωσης της γνώσης μπορεί να αναπαρασταθεί μαθηματικά ως εξής:

$$C = M(I, K_1, K_2, \ldots, K_n)$$

Όπου: - C είναι η παραγόμενη ολοκλήρωση. - I είναι η είσοδος. - $K_1, K_2, \ldots, K_n$ αντιπροσωπεύουν διάφορες πηγές γνώσης που ενσωματώνονται στο μοντέλο.

Η διαδικασία ενσωμάτωσης μπορεί να θεωρηθεί ως ένα σταθμισμένο άθροισμα των συνεισφορών από κάθε πηγή γνώσης, διαμορφωμένο από τις παραμέτρους του μοντέλου. Η πρόκληση έγκειται στον αποτελεσματικό συντονισμό αυτών των παραμέτρων ώστε να εξασφαλιστεί η βέλτιστη συγχώνευση γνώσεων.

5.3.3 Οφέλη και προκλήσεις

Η ενσωμάτωση διαφορετικών πηγών γνώσης σε μοντέλα ολοκλήρωσης προσφέρει αρκετά πλεονεκτήματα:

- **Contextual Awareness**: Τα μοντέλα μπορούν να παράγουν συμπληρώσεις που δεν είναι μόνο γραμματικά σωστές αλλά και σχετικές με τα συμφραζόμενα.
- **Προσαρμοστικότητα στο πεδίο εφαρμογής**: Με την ενσωμάτωση γνώσεων ειδικού τομέα, τα μοντέλα μπορούν να ρυθμιστούν λεπτομερώς για συγκεκριμένες εφαρμογές, ενισχύοντας την ευελιξία τους.
- **Μειωμένη ασάφεια**: Η ολοκληρωμένη γνώση μπορεί να βοηθήσει τα μοντέλα να αποσαφηνίσουν τις πιθανές συμπληρώσεις, οδηγώντας σε πιο ακριβείς προβλέψεις.

Ωστόσο, η ενσωμάτωση της γνώσης παρουσιάζει επίσης προκλήσεις. Η εξασφάλιση της απρόσκοπτης συγχώνευσης διαφορετικών πηγών γνώσης χωρίς να υπερφορτωθεί το μοντέλο ή να εισαχθούν προκαταλήψεις δεν είναι τετριμμένη. Επιπλέον, ο δυναμικός χαρακτήρας της γνώσης απαιτεί συνεχείς ενημερώσεις για να εξασφαλιστεί ότι το μοντέλο παραμένει επίκαιρο.

5.3.4 Συμπεράσματα

Η ολοκλήρωση της γνώσης διαδραματίζει καθοριστικό ρόλο στην ενίσχυση της απόδοσης και της προσαρμοστικότητας των μοντέλων ολοκλήρωσης. Με την αποτελεσματική ανάμειξη διαφορετικών πηγών γνώσης, τα μοντέλα αυτά μπορούν να παράγουν πιο συναφείς με το πλαίσιο και ακριβείς συμπληρώσεις, ανοίγοντας το δρόμο για πιο διαισθητικές και αποτελεσματικές αλληλεπιδράσεις chatbot.

5.4 Δυσκολίες

Τα μοντέλα συμπλήρωσης, ιδίως εκείνα που βασίζονται σε αρχιτεκτονικές βαθιάς μάθησης όπως οι μετασχηματιστές, έχουν δείξει αξιοσημείωτες ικανότητες στη δημιουργία συνεκτικού και σχετικού με το περιεχόμενο κειμένου. Ωστόσο, δεν στερούνται των προκλήσεών τους. Αυτή η ενότητα εμβαθύνει στις ιδιαιτερότητες αυτών των προκλήσεων, παραπέμποντας σε πρόσφατες έρευνες και ενσωματώνοντας ιδέες από το προηγούμενο κεφάλαιο και την παρεχόμενη διατριβή.

5.4.1 Μέγεθος μοντέλου και υπολογιστικές απαιτήσεις

Μία από τις πρωταρχικές προκλήσεις στην εκπαίδευση μοντέλων ολοκλήρωσης είναι το τεράστιο μέγεθος των μοντέλων τελευταίας τεχνολογίας. Καθώς τα μοντέλα αυξάνονται σε παραμέτρους, οι υπολογιστικές απαιτήσεις για την εκπαίδευση αυξάνονται επίσης εκθετικά.

$$C \propto P^2 \tag{5.1}$$

Όπου C αντιπροσωπεύει τις υπολογιστικές απαιτήσεις και P αντιπροσωπεύει τον αριθμό των παραμέτρων. Η σχέση αυτή δείχνει ότι ακόμη και μια μικρή αύξηση των παραμέτρων μπορεί να οδηγήσει σε σημαντικές υπολογιστικές επιβαρύνσεις (7).

5.4.2 Γενίκευση vs Απομνημόνευση

Όπως συζητήθηκε στο προηγούμενο κεφάλαιο, υπάρχει μια λεπτή ισορροπία μεταξύ της γενίκευσης και της απομνημόνευσης. Ενώ τα μεγαλύτερα μοντέλα τείνουν να απομνημονεύουν περισσότερα από τα δεδομένα εκπαίδευσης, μπορεί να μην γενικεύουν πάντα καλά σε αόρατα δεδομένα. Αυτό το φαινόμενο μπορεί να αναπαρασταθεί ως εξής:

$$G = \frac{1}{1 + e^{-k(M-m)}} \tag{5.2}$$

Όπου G είναι η ικανότητα γενίκευσης, M είναι το μέγεθος του μοντέλου, m είναι ένα μέγεθος κατωφλίου και k είναι μια σταθερά. Αυτή η σιγμοειδής συνάρτηση υποδηλώνει ότι μετά από ένα ορισμένο μέγεθος μοντέλου, τα κέρδη στη γενίκευση μειώνονται (33).

5.4.3 Ηθικές Ανησυχίες

Τα μοντέλα ολοκλήρωσης, ιδίως εκείνα που εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων, μπορούν να παράγουν ακούσια ακατάλληλο ή μεροληπτικό περιεχόμενο. Αυτό οφείλεται στο γεγονός ότι ενδέχεται να μάθουν και να ενισχύσουν τις προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσής τους. Η αντιμετώπιση αυτών των ανησυχιών απαιτεί προσεκτική τελειοποίηση του μοντέλου και post-hoc αξιολόγηση (4).

5.4.4 Data Privacy

Όπως τονίζεται στην παρεχόμενη διατριβή, υπάρχει πιθανός κίνδυνος τα μοντέλα να αποκαλύπτουν ακούσια πληροφορίες σχετικά με τα δεδομένα εκπαίδευσής τους. Αυτό αποτελεί σημαντική ανησυχία για το απόρρητο των δεδομένων, ιδίως όταν τα μοντέλα εκπαιδεύονται σε ευαίσθητα ή ιδιόκτητα σύνολα δεδομένων.

5.4.5 Ευρωστία

Παρά τις δυνατότητές τους, τα μοντέλα ολοκλήρωσης μπορεί να είναι ευαίσθητα στις διαταραχές εισόδου. Οι επιθέσεις των αντιπάλων μπορούν να εκμεταλλευτούν αυτή την ευαισθησία, προκαλώντας τα μοντέλα να παράγουν λανθασμένες ή μη λογικές εξόδους (44).

5.4.6 Συμπεράσματα

Ενώ τα μοντέλα ολοκλήρωσης προσφέρουν πολλά υποσχόμενα αποτελέσματα σε διάφορες εφαρμογές, η αντιμετώπιση των προκλήσεών τους είναι ζωτικής σημασίας για την ασφαλή και αποτελεσματική ανάπτυξή τους. Οι μελλοντικές ερευνητικές κατευθύνσεις θα μπορούσαν να επικεντρωθούν στην ανάπτυξη πιο αποτελεσματικών, στιβαρών και ηθικά ορθών μοντέλων.

5.5 Μελλοντικές κατευθύνσεις στα μοντέλα ολοκλήρωσης

Το πεδίο των μοντέλων ολοκλήρωσης είναι ιδανική κατάσταση για εξερεύνηση. Με την έλευση μοντέλων που μπορούν να ενσωματώσουν δομημένη γνώση, τα όρια του τι μπορούν

να επιτύχουν αυτά τα μοντέλα διευρύνονται συνεχώς. Η μελλοντική έρευνα μπορεί να επικεντρωθεί στο να γίνουν αυτά τα μοντέλα πιο αποτελεσματικά, ερμηνεύσιμα και απαλλαγμένα από προκαταλήψεις.

Κεφάλαιο 6

Fine-Tuning vs In-Context Learning

Στο πεδίο της μηχανικής μάθησης, ιδίως στο πλαίσιο των μοντέλων βαθιάς μάθησης, έχουν αναδειχθεί δύο σημαντικές στρατηγικές για την προσαρμογή προ-εκπαιδευμένων μοντέλων σε συγκεκριμένες εργασίες: η λεπτομερής ρύθμιση (Fine-Tuning) και η μάθηση εντός πλαισίου (In-Context Learning). Οι στρατηγικές αυτές είναι ιδιαίτερα σημαντικές στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP) και έχουν συμβάλει καθοριστικά στην ανάπτυξη σύγχρονων συστημάτων chatbot.

6.1 Εισαγωγή

Η λεπτομερής ρύθμιση περιλαμβάνει τη λήψη ενός προ-εκπαιδευμένου μοντέλου και τη συνέχιση της διαδικασίας εκπαίδευσης σε ένα νέο σύνολο δεδομένων, συνήθως μικρότερο και πιο συγκεκριμένο για την εργασία, για την προσαρμογή των παραμέτρων του μοντέλου στη νέα εργασία (16). Η προσέγγιση αυτή έχει υιοθετηθεί ευρέως σε διάφορους τομείς, από την όραση υπολογιστών (Computer Vision), όπου τα μοντέλα που έχουν προ-εκπαιδευτεί σε μεγάλα σύνολα δεδομένων όπως το ImageNet προσαρμόζονται για συγκεκριμένες εργασίες ανίχνευσης αντικειμένων (36), έως τη NLP, όπου τα μοντέλα που έχουν προ-εκπαιδευτεί σε τεράστια σώματα δεδομένων προσαρμόζονται για εργασίες όπως η ανάλυση συναισθήματος ή η αναγνώριση ονομαστικών οντοτήτων.

Από την άλλη πλευρά, η εκμάθηση εντός πλαισίου αξιοποιεί τη γνώση που περικλείεται σε ένα προ-εκπαιδευμένο μοντέλο χωρίς να μεταβάλλει τις παραμέτρους του. Αντίθετα, παρέχει πλαίσιο, συχνά με τη μορφή προτροπής ή πρόσθετης εισόδου, για να καθοδηγήσει τις προβλέψεις του μοντέλου για μια συγκεκριμένη εργασία. Αυτή η στρατηγική είναι ιδιαίτερα σημαντική για μοντέλα όπως το GPT, όπου το μοντέλο εκπαιδεύεται να παράγει κείμενο ανάλογα με την είσοδό του, επιτρέποντάς του να εκτελεί ένα ευρύ φάσμα εργασιών χωρίς την εκάστοτε εργασία.

Η επιλογή μεταξύ της λεπτομερούς ρύθμισης και της μάθησης εντός πλαισίου δεν είναι τετριμμένη και εξαρτάται από διάφορους παράγοντες, όπως το μέγεθος και η φύση των διαθέσιμων δεδομένων, οι ειδικές απαιτήσεις της εργασίας και η αρχιτεκτονική του υποκείμενου

μοντέλου. Για παράδειγμα, ενώ η λεπτομερής προσαρμογή μπορεί να είναι κατάλληλη για εργασίες με άφθονα επισημασμένα δεδομένα, η μάθηση εντός πλαισίου μπορεί να είναι πλεονεκτική για εργασίες όπου τα επισημασμένα δεδομένα είναι ελάχιστα ή όπου απαιτείται ταχεία προσαρμογή σε νέες εργασίες χωρίς το κόστος της επανεκπαίδευσης.

Στο πλαίσιο της ανάπτυξης chatbot, η κατανόηση των αποχρώσεων, των πλεονεκτημάτων και των περιορισμών τόσο της λεπτομερούς ρύθμισης όσο και της μάθησης εντός πλαισίου είναι ζωτικής σημασίας. Καθώς τα chatbots αναμένεται να χειρίζονται ένα ευρύ φάσμα ερωτημάτων χρηστών και να προσαρμόζονται σε διάφορους τομείς, οι στρατηγικές που χρησιμοποιούνται για την αξιοποίηση των προ-εκπαιδευμένων μοντέλων μπορούν να επηρεάσουν σημαντικά την απόδοση, τη χρηστικότητα και την προσαρμοστικότητά τους.

Σε αυτό το κεφάλαιο, θα εμβαθύνουμε στα θεωρητικά θεμέλια της λεπτομερούς προσαρμογής και της μάθησης εντός πλαισίου, θα διερευνήσουμε τις εφαρμογές τους σε συστήματα NLP και chatbot και θα συζητήσουμε τα αντισταθμιστικά οφέλη που συνεπάγεται η επιλογή της μιας προσέγγισης έναντι της άλλης.

6.1.1 Μία Ιστορική Προοπτική

Η έννοια του fine-tuning έχει τις ρίζες της στη μάθηση μεταφοράς, όπου η γνώση που αποκτάται κατά την επίλυση ενός προβλήματος εφαρμόζεται σε ένα διαφορετικό αλλά συναφές πρόβλημα. Οι πρώιμες εργασίες στην όραση υπολογιστών, όπως το R-CNN (16) και οι διάδοχοί του, κατέδειξαν τη δύναμη της λεπτομερούς ρύθμισης προ-εκπαιδευμένων μοντέλων σε συγκεκριμένες εργασίες, οδηγώντας σε σημαντικές βελτιώσεις στην ακρίβεια και την αποδοτικότητα.

Παράλληλα, η άνοδος των αρχιτεκτονικών μετασχηματιστών στο NLP, με παραδείγματα μοντέλα όπως το BERT και το GPT, έφερε στο προσκήνιο τις δυνατότητες της μάθησης εντός πλαισίου. Αυτά τα μοντέλα, προ-εκπαιδευμένα σε τεράστιες ποσότητες κειμένου, παρουσίασαν την ικανότητα να εκτελούν μια πληθώρα εργασιών με την απλή προετοιμασία τους με το κατάλληλο πλαίσιο, εξαλείφοντας την ανάγκη για λεπτομερή ρύθμιση συγκεκριμένων εργασιών σε πολλές περιπτώσεις.

Καθώς περιηγούμαστε σε αυτό το κεφάλαιο, θα αντιπαραβάλλουμε αυτές τις στρατηγικές, αντλώντας πληροφορίες από θεμελιώδη έργα και πρόσφατες εξελίξεις, για να παρέχουμε μια ολοκληρωμένη κατανόηση της σημασίας τους στα σύγχρονα συστήματα chatbot.

6.1.2 Σχέση με την ανάπτυξη chatbot

Για τους προγραμματιστές chatbot, η απόφαση μεταξύ της λεπτομερούς ρύθμισης και της μάθησης στο πλαίσιο είναι υψίστης σημασίας. Η λεπτομερής ρύθμιση μπορεί να παρέχει στα chatbots έναν πιο προσαρμοσμένο μηχανισμό απόκρισης για συγκεκριμένους τομείς, εξασφαλίζοντας μεγαλύτερη ακρίβεια και συνάφεια. Ωστόσο, συχνά απαιτεί επισημειωμένα δεδομένα, τα οποία μπορεί να είναι σπάνια ή ακριβά για εξειδικευμένους τομείς.

Αντίθετα, η μάθηση εντός πλαισίου προσφέρει ευελιξία, επιτρέποντας στα chatbots να προσαρμόζονται γρήγορα σε νέες εργασίες ή τομείς, με βάση το πλαίσιο που παρέχουν οι χρήστες. Αυτό μπορεί να είναι ιδιαίτερα πολύτιμο για τα chatbots που αναπτύσσονται σε δυναμικά περιβάλλοντα ή για εκείνα που απευθύνονται σε μια διαφορετική βάση χρηστών με ποικίλες απαιτήσεις.

Στις επόμενες ενότητες, θα αναλύσουμε τους μηχανισμούς και των δύο στρατηγικών, θα διευκρινίσουμε τα δυνατά και αδύνατα σημεία τους και θα παράσχουμε κατευθυντήριες γραμμές για την αποτελεσματική αξιοποίησή τους σε συστήματα chatbot.

6.2 Fine-Tuning

Η λεπτή ρύθμιση είναι μια βασική τεχνική στο πεδίο της βαθιάς μάθησης, ειδικά όταν πρόκειται για την προσαρμογή προ-εκπαιδευμένων μοντέλων σε συγκεκριμένες εργασίες. Περιλαμβάνει τη λήψη ενός μοντέλου που έχει ήδη εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων και την περαιτέρω εκπαίδευσή του σε ένα μικρότερο, συγκεκριμένο για την εργασία σύνολο δεδομένων. Αυτή η προσέγγιση αξιοποιεί τα γενικά χαρακτηριστικά που έχει μάθει το μοντέλο από το μεγαλύτερο σύνολο δεδομένων και τα βελτιώνει για τη συγκεκριμένη εργασία (40).

Τα μοντέλα βαθιάς μάθησης, ειδικά εκείνα που έχουν σχεδιαστεί για εργασίες όπως η αναγνώριση εικόνας, συχνά απαιτούν τεράστιες ποσότητες δεδομένων και υπολογιστικών πόρων για να εκπαιδευτούν από το μηδέν. Ωστόσο, πολλές εφαρμογές του πραγματικού κόσμου ενδέχεται να μην έχουν πρόσβαση σε τόσο εκτεταμένα σύνολα δεδομένων. Σε αυτό το σημείο μπαίνει στο παιχνίδι η λεπτομερής ρύθμιση. Χρησιμοποιώντας ένα μοντέλο που έχει προ-εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων, μπορεί κανείς να αξιοποιήσει τη δύναμη της βαθιάς μάθησης χωρίς την ανάγκη για εκτεταμένους πόρους (27).

Στο πλαίσιο των μοντέλων ολοκλήρωσης, η λεπτή ρύθμιση γίνεται ακόμη πιο κρίσιμη. Αυτά τα μοντέλα συχνά προ-εκπαιδεύονται σε τεράστια σώματα δεδομένων κειμένου, αποτυπώνοντας ένα ευρύ φάσμα γλωσσικών προτύπων και δομών. Ωστόσο, οι εφαρμογές chatbot μπορεί να έχουν συγκεκριμένες απαιτήσεις, όπως η κατανόηση ορολογιών συγκεκριμένου τομέα ή η τήρηση ενός συγκεκριμένου στυλ συνομιλίας. Η λεπτομερής ρύθμιση επιτρέπει στους προγραμματιστές να προσαρμόζουν αυτά τα γενικά μοντέλα ώστε να ανταποκρίνονται σε αυτές τις ειδικές ανάγκες.

Μαθηματικά, η λεπτομερής ρύθμιση μπορεί να θεωρηθεί ως συνέχεια της διαδικασίας βελτιστοποίησης. Ας θεωρήσουμε ένα μοντέλο νευρωνικού δικτύου με παραμέτρους θ. Όταν το μοντέλο προ-εκπαιδεύεται σε ένα μεγάλο σύνολο δεδομένων, οι παράμετροι ενημερώνονται ώστε να ελαχιστοποιείται η συνάρτηση απώλειας L_{pretrain}. Κατά τη διάρκεια της λεπτομερούς ρύθμισης, αυτές οι παράμετροι ενημερώνονται περαιτέρω για να ελαχιστοποιήσουν μια νέα συνάρτηση απώλειας $L_{\text{fine-tune}}$ που ορίζεται πάνω στο συγκεκριμένο σύνολο δεδομένων. Ο κανόνας ενημέρωσης μπορεί να αναπαρασταθεί ως εξής:

$$\theta_{t+1} = \theta_t - \alpha \nabla L_{\text{fine-tune}}(\theta_t)$$

όπου α είναι ο ρυθμός μάθησης και $\nabla L_{\text{fine-tune}}$ είναι η κλίση της απώλειας τελειοποίησης ως προς τις παραμέτρους.

Ένα από τα θεμελιώδη έργα που ανέδειξε τη δύναμη της λεπτομερούς ρύθμισης στη βαθιά μάθηση είναι η προσαρμογή των βαθιών δικτύων πεποιθήσεων για την αναγνώριση ψηφίων (20). Οι συγγραφείς απέδειξαν ότι με τη λεπτομερή ρύθμιση των ανώτερων στρωμάτων του δικτύου, μπορούσαν να επιτύχουν κορυφαίες επιδόσεις στην εργασία, ξεπερνώντας άλλους αλγορίθμους διάκρισης.

Συμπερασματικά, η λεπτομερής ρύθμιση είναι μια ισχυρή τεχνική που επιτρέπει στους προγραμματιστές να αξιοποιήσουν τις δυνατότητες των προ-εκπαιδευμένων μοντέλων και να τα προσαρμόσουν σε συγκεκριμένες εργασίες. Στον τομέα της ανάπτυξης chatbot, παρέχει μια

οδό για τη δημιουργία μοντέλων που είναι τόσο γενικά στην κατανόηση της γλώσσας όσο και ειδικά στις δυνατότητες παραγωγής απαντήσεων.

6.3 In-Context Learning

Στο πεδίο της μηχανικής μάθησης, ιδίως στο πλαίσιο των συστημάτων chatbot και της επεξεργασίας φυσικής γλώσσας, η ικανότητα προσαρμογής σε νέες πληροφορίες χωρίς ρητή επανεκπαίδευση είναι υψίστης σημασίας. Αυτή η προσαρμοστικότητα αναφέρεται συχνά ως "μάθηση εντός πλαισίου". Σε αντίθεση με την παραδοσιακή λεπτομερή προσαρμογή, όπου τα μοντέλα επανεκπαιδεύονται ρητά σε νέα δεδομένα, η μάθηση εντός πλαισίου αξιοποιεί την υπάρχουσα γνώση του μοντέλου και προσαρμόζει τις απαντήσεις του με βάση το παρεχόμενο πλαίσιο.

Στον πυρήνα της, η μάθηση εντός πλαισίου αφορά την κατανόηση και τη δημιουργία περιεχομένου με βάση το περιβάλλον. Ένα από τα πρωτοποριακά έργα σε αυτόν τον τομέα είναι η έννοια των "Κωδικοποιητών πλαισίου" (31). Οι Context Encoders, όπως προτάθηκαν από τους Pathak κ.ά., είναι νευρωνικά δίκτυα συνελικτικού τύπου που εκπαιδεύονται για να παράγουν το περιεχόμενο αυθαίρετων περιοχών εικόνας υπό την προϋπόθεση του περιβάλλοντός τους. Για να επιτύχουν σε αυτό το έργο, οι κωδικοποιητές πλαισίου πρέπει να κατανοούν το περιεχόμενο ολόκληρης της εικόνας και να παράγουν μια εύλογη υπόθεση για το(α) τμήμα(τα) που λείπει(ουν). Αν και η πρωταρχική εφαρμογή ήταν στον τομέα της βαφής εικόνων, η υποκείμενη αρχή είναι εξαιρετικά σημαντική για τα συστήματα chatbot. Στη σφαίρα των chatbots, η περιβάλλουσα συνομιλία παρέχει το πλαίσιο και το μοντέλο πρέπει να παράγει τις κατάλληλες απαντήσεις με βάση αυτό το πλαίσιο.

Από μια ευρύτερη προοπτική, η ιδέα της ανταπόκρισης με βάση το πλαίσιο δεν αφορά αποκλειστικά τις μηχανές. Ο ανθρώπινος εγκέφαλος παρουσιάζει μια παρόμοια συμπεριφορά, η οποία συχνά αναφέρεται ως "φαινόμενο placebo" στην ιατρική βιβλιογραφία. Οι Wager και Atlas (43) εμβαθύνουν στη νευροεπιστήμη του φαινομένου placebo, τονίζοντας πώς οι αντιδράσεις του εγκεφάλου-νοητικού συστήματος στο πλαίσιο μπορούν να οδηγήσουν σε ευεργετικά αποτελέσματα. Υποστηρίζουν ότι αυτά τα αποτελέσματα, που διαμεσολαβούνται από ποικίλες διαδικασίες, συμπεριλαμβανομένης της μάθησης και των προσδοκιών, μπορούν να επηρεάσουν διάφορα κλινικά και φυσιολογικά αποτελέσματα. Η υποκείμενη νευροεπιστήμη εμπλέκει πολλαπλά εγκεφαλικά συστήματα και νευροχημικούς μεσολαβητές. Κάνοντας παραλληλισμούς, τα συστήματα chatbot που είναι εξοπλισμένα με μάθηση εντός του πλαισίου μπορούν να θεωρηθούν ότι μιμούνται αυτή την ανθρώπινη ικανότητα προσαρμογής των απαντήσεων με βάση το πλαίσιο, αν και σε ένα υπολογιστικό περιβάλλον.

Για τα συστήματα chatbot, η μάθηση εντός πλαισίου προσφέρει έναν δυναμικό τρόπο προσαρμογής των απαντήσεων με βάση την τρέχουσα συζήτηση. Αντί να βασίζεται αποκλειστικά σε προ-εκπαιδευμένη γνώση, το μοντέλο μπορεί να αξιοποιήσει το πλαίσιο της συνομιλίας για να παράγει πιο σχετικές και συνεκτικές απαντήσεις. Αυτό είναι ιδιαίτερα σημαντικό για τα chatbots που αναπτύσσονται σε δυναμικά περιβάλλοντα όπου τα ερωτήματα των χρηστών μπορεί να καλύπτουν ένα ευρύ φάσμα θεμάτων και προθέσεων.

Η μάθηση εντός πλαισίου, με τις ρίζες της τόσο στα υπολογιστικά μοντέλα όσο και στην ανθρώπινη νευροεπιστήμη, προσφέρει μια πολλά υποσχόμενη οδό για τη βελτίωση της προσαρμοστικότητας και της συνάφειας των συστημάτων chatbot. Καθώς η έρευνα σε αυτόν τον τομέα εξελίσσεται, αναμένεται ότι τα συστήματα chatbot θα αποκτήσουν ακόμη μεγαλύτερη

επίγνωση του πλαισίου, οδηγώντας σε πιο φυσικές και ουσιαστικές αλληλεπιδράσεις.

6.4 Συμπεράσματα

Τόσο η λεπτομερής προσαρμογή όσο και η μάθηση στο πλαίσιο προσφέρουν μοναδικά πλεονεκτήματα στην προσαρμογή μοντέλων μεγάλης κλίμακας σε συγκεκριμένες εργασίες. Η επιλογή μεταξύ τους εξαρτάται συχνά από τις ειδικές απαιτήσεις της εφαρμογής, τους διαθέσιμους πόρους και τα επιθυμητά αποτελέσματα. Καθώς η έρευνα σε αυτόν τον τομέα συνεχίζεται, είναι πιθανό να προκύψουν υβριδικές προσεγγίσεις που θα συνδυάζουν τα πλεονεκτήματα και των δύο παραδειγμάτων.

Κεφάλαιο 7

Vector Databases vs Conventional Databases

Η εξέλιξη των μηχανισμών αποθήκευσης και ανάκτησης δεδομένων έχει συμβάλει καθοριστικά στη διαμόρφωση του τοπίου των σύγχρονων υπολογιστικών συστημάτων. Καθώς τα chatbots και οι πράκτορες συνομιλίας γίνονται όλο και πιο εξελιγμένοι, η υποκείμενη υποδομή που υποστηρίζει την ανάκτηση της γνώσης τους καθίσταται υψίστης σημασίας. Αυτό το κεφάλαιο εμβαθύνει στα αντιθετικά παραδείγματα των διανυσματικών βάσεων δεδομένων και των συμβατικών βάσεων δεδομένων, διευκρινίζοντας τα αντίστοιχα πλεονεκτήματα, τις αδυναμίες και τη δυνατότητα εφαρμογής τους στο πεδίο της ανάπτυξης chatbot.

7.1 Εισαγωγή

Στην προσπάθεια να γίνουν τα chatbots πιο έξυπνα και με επίγνωση του πλαισίου, ο τρόπος με τον οποίο αποθηκεύουν και ανακτούν πληροφορίες παίζει καθοριστικό ρόλο. Οι παραδοσιακές βάσεις δεδομένων, οι οποίες αποτέλεσαν τη ραχοκοκαλιά πολλών εφαρμογών λογισμικού, είναι δομημένες και καθοδηγούμενες από ερωτήματα. Από την άλλη πλευρά, οι διανυσματικές βάσεις δεδομένων, ένας σχετικά νεότερος εισερχόμενος, αξιοποιούν τη δύναμη των ενσωματώσεων και των χώρων υψηλών διαστάσεων για να διευκολύνουν τις αναζητήσεις που βασίζονται στην ομοιότητα. Αυτή η ενότητα παρουσιάζει τις θεμελιώδεις διαφορές μεταξύ αυτών των δύο τύπων βάσεων δεδομένων και θέτει τις βάσεις για μια βαθύτερη διερεύνηση των επιπτώσεών τους στα συστήματα chatbot.

7.2 Συμβατικές βάσεις δεδομένων: Δομή και δυνατά σημεία

Οι συμβατικές βάσεις δεδομένων, συχνά σχεσιακής φύσης, αποθηκεύουν δεδομένα σε δομημένους πίνακες με προκαθορισμένα σχήματα. Είναι βελτιστοποιημένες για συναλλαγές και ακριβείς αναζητήσεις. Τα πλεονεκτήματά τους έγκεινται στα εξής:

- **Ακεραιότητα δεδομένων**: Διασφάλιση ότι τα δεδομένα παραμένουν συνεπή και ακριβή.

- **Επεκτασιμότητα**: Δυνατότητα αποτελεσματικής διαχείρισης μεγάλου όγκου δεδομένων.
- **Ιδιότητες ACID**: Προσκόλληση στην Ατομικότητα, τη Συνέπεια, την Απομόνωση και την Ανθεκτικότητα.

Για τα chatbots που βασίζονται σε δομημένα δεδομένα, όπως συστήματα κρατήσεων ή bots υποστήριξης πελατών που διασυνδέονται με επιχειρησιακές βάσεις δεδομένων, οι συμβατικές βάσεις δεδομένων είναι συχνά η επιλογή που πρέπει να γίνει.

7.3 Διανυσματικές βάσεις δεδομένων: Η δύναμη των ενσωματώσεων

Οι διανυσματικές βάσεις δεδομένων αποθηκεύουν δεδομένα σε χώρους υψηλών διαστάσεων, επιτρέποντας αναζητήσεις βάσει ομοιότητας. Αυτό είναι ιδιαίτερα χρήσιμο για εργασίες όπως η σημασιολογική αναζήτηση, όπου ο στόχος είναι η εύρεση στοιχείων που είναι παρόμοια από άποψη πλαισίου ή σημασιολογίας και όχι ακριβώς όμοια με ένα ερώτημα. Τα βασικά χαρακτηριστικά περιλαμβάνουν:

- **Embedding-based Retrieval**: Τα στοιχεία δεδομένων αναπαρίστανται ως διανύσματα, διευκολύνοντας την αναζήτηση με βάση την εγγύτητα.
- **Ευελιξία**: Σε αντίθεση με τις δομημένες βάσεις δεδομένων, οι διανυσματικές βάσεις δεδομένων μπορούν να χειριστούν μη δομημένα δεδομένα όπως κείμενο, εικόνες και άλλα.
- **Ενσωμάτωση με μοντέλα ML**: Απρόσκοπτη ενσωμάτωση με μοντέλα μηχανικής μάθησης, ειδικά με εκείνα που εξάγουν ενσωματώσεις.

Για τα chatbots που πρέπει να κατανοούν την πρόθεση του χρήστη και να παρέχουν σχετικές με το πλαίσιο απαντήσεις, οι διανυσματικές βάσεις δεδομένων προσφέρουν ένα επιτακτικό πλεονέκτημα.

7.4 Επιλογή της σωστής βάσης δεδομένων για chatbots

Η επιλογή μεταξύ μιας διανυσματικής βάσης δεδομένων και μιας συμβατικής βάσης δεδομένων εξαρτάται συχνά από τις συγκεκριμένες απαιτήσεις της εφαρμογής chatbot. Οι παράγοντες που πρέπει να ληφθούν υπόψη περιλαμβάνουν:

- Η φύση των δεδομένων: δομημένα έναντι μη δομημένων.
- Το είδος των ερωτημάτων: ακριβείς αντιστοιχίες έναντι αναζητήσεων με ομοιότητα.
- Ενσωμάτωση με άλλα συστήματα και ανάγκες επεκτασιμότητας.

Σε πολλά σύγχρονα συστήματα chatbot, υιοθετείται μια υβριδική προσέγγιση, αξιοποιώντας τα πλεονεκτήματα και των δύο τύπων βάσεων δεδομένων για τη δημιουργία ενός πιο ισχυρού και ευέλικτου μηχανισμού ανάκτησης πληροφοριών.

7.5 Συμπεράσματα

Καθώς οι τεχνολογίες chatbot συνεχίζουν να εξελίσσονται, η υποκείμενη υποδομή που τις υποστηρίζει θα διαδραματίσει κρίσιμο ρόλο στον καθορισμό της αποτελεσματικότητας και της ευελιξίας τους. Με την κατανόηση των αποχρώσεων των διανυσματικών βάσεων δεδομένων και των συμβατικών βάσεων δεδομένων, οι προγραμματιστές μπορούν να λαμβάνουν τεκμηριωμένες αποφάσεις που ενισχύουν τις δυνατότητες των συστημάτων chatbot τους.

Κεφάλαιο 8

Αξιολόγηση

8.1 Εισαγωγή

Η αξιολόγηση της απόδοσης του chatbot είναι μια κρίσιμη πτυχή της ανάπτυξης chatbot. Παρέχει πληροφορίες σχετικά με τα δυνατά και αδύνατα σημεία ενός μοντέλου, καθοδηγώντας περαιτέρω βελτιώσεις. Ενώ υπάρχουν πολλά σύνολα δεδομένων διαθέσιμα για την εκπαίδευση και την αξιολόγηση των chatbots, η επιλογή του σωστού συνόλου δεδομένων είναι υψίστης σημασίας. Αυτό το κεφάλαιο εμβαθύνει στα διάφορα σύνολα δεδομένων που είναι διαθέσιμα για την αξιολόγηση chatbot, με στόχο τον προσδιορισμό του βέλτιστου.

8.2 Δεδομένα Αξιολόγησης

Η αξιολόγηση των επιδόσεων των chatbot είναι ένα πολύπλευρο εγχείρημα, το οποίο απαιτεί σύνολα δεδομένων (Datasets) που μπορούν να μετρήσουν αποτελεσματικά την ικανότητα ενός συστήματος να κατανοεί, να παράγει και να συμμετέχει σε ουσιαστικό διάλογο. Με την πάροδο των ετών, έχουν εισαχθεί διάφορα σύνολα δεδομένων για τη συγκριτική αξιολόγηση συστημάτων chatbot, το καθένα με τα μοναδικά του χαρακτηριστικά και προκλήσεις. Η παρούσα ενότητα παρέχει μια επισκόπηση ορισμένων από τα εξέχοντα σύνολα δεδομένων στον τομέα και συζητά τη σημασία τους στο πλαίσιο της αξιολόγησης συστημάτων chatbot.

8.2.1 Visual Dialog

Το σύνολο δεδομένων Visual Dialog που παρουσιάστηκε από τους Das et al. (10), απαιτεί από έναν πράκτορα ΤΝ να συμμετάσχει σε διάλογο σχετικά με οπτικό περιεχόμενο. Το σύνολο δεδομένων βασίζεται στην όραση, επιτρέποντας την αντικειμενική αξιολόγηση των μεμονωμένων απαντήσεων και τη συγκριτική αξιολόγηση της προόδου. Ο στόχος είναι να διεξαχθεί ένας ουσιαστικός διάλογος με ανθρώπους σε φυσική, διαλογική γλώσσα για οπτικό περιεχόμενο. Το σύνολο δεδομένων περιέχει περίπου 1,4 εκατομμύρια ζεύγη ερωτήσεων-απαντήσεων σε περίπου 140 χιλιάδες εικόνες από το σύνολο δεδομένων COCO. Το σύνολο δεδομένων Visual Dialog δοκιμάζει την ικανότητα ενός chatbot να ενσωματώνει οπτικές και κειμενικές πληροφορίες, καθιστώντας το ένα μοναδικό εργαλείο αξιολόγησης για πολυτροπικά συστήματα

chatbot (10).

8.2.2 Complex Sequential Question Answering

Οι Saha et al. (**?**) παρουσίασαν το σύνολο δεδομένων Complex Sequential Question Answering, το οποίο συνδυάζει εργασίες απάντησης ερωτήσεων και διαλόγου. Το σύνολο δεδομένων απαιτεί από τα chatbots να απαντούν σε πραγματικές ερωτήσεις μέσω σύνθετων συμπερασμάτων πάνω σε έναν γράφο γνώσης μεγάλης κλίμακας. Περιέχει περίπου 200K διαλόγους με συνολικά 1,6M στροφές. Οι ερωτήσεις σε αυτό το σύνολο δεδομένων απαιτούν λογικούς, ποσοτικούς, συγκριτικούς συλλογισμούς και συνδυασμούς τους. Αυτό το σύνολο δεδομένων είναι ιδιαίτερα απαιτητικό, καθώς απαιτεί από τα chatbots να αναλύουν τη φυσική γλώσσα, να χρησιμοποιούν το πλαίσιο της συνομιλίας, να επιλύουν πυρηνοπαραπομπές, να αναζητούν διευκρινίσεις για διφορούμενα ερωτήματα και να ανακτούν σχετικά υπογράμματα από τον γράφο γνώσης για να απαντούν σε ερωτήσεις (**?**).

8.2.3 SQuAD: Stanford Question Answering Dataset

Το Stanford Question Answering Dataset (SQuAD) είναι ένα από τα πιο δημοφιλή σύνολα δεδομένων για την αξιολόγηση των ικανοτήτων κατανόησης κειμένου από συστήματα chatbot. Αν και έχει σχεδιαστεί κυρίως για την εξαγωγή ερωτήσεων, η δομή και το περιεχόμενό του το καθιστούν πολύτιμο πόρο για την αξιολόγηση chatbot. Το σύνολο δεδομένων περιέχει ερωτήσεις που θέτουν οι crowdworkers σε ένα σύνολο άρθρων της Wikipedia, όπου η απάντηση σε κάθε ερώτηση είναι ένα τμήμα κειμένου από το αντίστοιχο απόσπασμα ανάγνωσης. Δεδομένης της ευρείας χρήσης και της προκλητικής φύσης του, το SQuAD χρησιμεύει ως σημείο αναφοράς για συστήματα chatbot που στοχεύουν στην κατανόηση και τη δημιουργία απαντήσεων που μοιάζουν με τις ανθρώπινες.

8.2.4 Άλλα σύνολα δεδομένων

Ενώ τα προαναφερθέντα σύνολα δεδομένων είναι από τα πιο αναγνωρισμένα στο τοπίο της αξιολόγησης chatbot, αρκετά άλλα σύνολα δεδομένων καλύπτουν εξειδικευμένους τομείς ή συγκεκριμένες προκλήσεις στην ανάπτυξη chatbot. Αυτά τα σύνολα δεδομένων κυμαίνονται από διαλόγους προσανατολισμένους σε εργασίες έως συνομιλίες ανοικτού πεδίου, προσφέροντας το καθένα μια μοναδική προοπτική για τις δυνατότητες των chatbot.

8.3 Μετρικές αξιολόγησης

Η αξιολόγηση της απόδοσης των chatbots είναι ένα πολύπλευρο έργο, καθώς δεν περιλαμβάνει μόνο την ακρίβεια της απάντησης αλλά και την ποιότητα, τη συνάφεια και την ευχέρεια του παραγόμενου κειμένου. Στη βιβλιογραφία έχουν προταθεί διάφορες μετρικές για την αντιμετώπιση αυτών των πτυχών. Σε αυτή την ενότητα, θα εμβαθύνουμε σε ορισμένες από τις πιο διαδεδομένες μετρικές και τη δυνατότητα εφαρμογής τους στο πλαίσιο της αξιολόγησης chatbot.

8.3.1 Ακρίβεια

Η ακρίβεια είναι μια απλή μετρική που μετρά το κλάσμα των σωστά προβλεπόμενων περιπτώσεων επί του συνόλου των περιπτώσεων. Ωστόσο, στο πλαίσιο των chatbots, μια απλή μετρική ακρίβειας μπορεί να μην είναι επαρκής λόγω της ανοιχτής φύσης των αλληλεπιδράσεων ανθρώπου- chatbot. Η ακρίβεια, από την άλλη πλευρά, επικεντρώνεται στο κλάσμα των σχετικών περιπτώσεων μεταξύ των ανακτημένων περιπτώσεων. Είναι ιδιαίτερα χρήσιμη όταν το κόστος των ψευδώς θετικών αποτελεσμάτων είναι υψηλό (**?**).

8.3.2 Recall και F1-Score

Η ανάκληση μετρά το κλάσμα των σχετικών περιπτώσεων που έχουν ανακτηθεί επί του συνολικού αριθμού των σχετικών περιπτώσεων. Σε σενάρια όπου η απώλεια μιας σχετικής περίπτωσης κοστίζει ακριβά, η ανάκληση γίνεται κρίσιμη μετρική. Το F1-Score είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, παρέχοντας μια ισορροπία μεταξύ των δύο. Είναι ιδιαίτερα χρήσιμο όταν υπάρχει ανομοιόμορφη κατανομή κλάσεων, καθώς λαμβάνει υπόψη τόσο τα ψευδώς θετικά όσο και τα ψευδώς αρνητικά αποτελέσματα (**?**).

8.3.3 BLEU, ROUGE και METEOR

Πρόκειται για δημοφιλείς μετρικές που χρησιμοποιούνται στη μηχανική μετάφραση και έχουν προσαρμοστεί για την αξιολόγηση chatbot. Η BLEU (Bilingual Evaluation Understudy) μετράει πόσες λέξεις και φράσεις στην απάντηση του chatbot ταιριάζουν με εκείνες σε μια απάντηση αναφοράς. Η ROUGE (Recall-Oriented Understudy for Gisting Evaluation) επικεντρώνεται στο μέγεθος της επικάλυψης μεταξύ των n-grams στο παραγόμενο κείμενο και ενός συνόλου κειμένων αναφοράς. Το METEOR (Metric for Evaluation of Translation with Explicit ORdering) εξετάζει την ακρίβεια, την ανάκληση, τη συνωνυμία και τη σειρά των λέξεων για να αξιολογήσει την ποιότητα του παραγόμενου κειμένου (**?**).

8.3.4 Αμηχανία

Η αμηχανία (Perplexity) είναι μια μετρική που χρησιμοποιείται στη γλωσσική μοντελοποίηση για να μετρήσει πόσο καλά η κατανομή πιθανότητας που προβλέπει το μοντέλο ευθυγραμμίζεται με την πραγματική κατανομή των λέξεων στο κείμενο. Ένα χαμηλότερο perplexity δείχνει ότι το μοντέλο είναι πιο σίγουρο για τις επιλογές λέξεων. Ωστόσο, αξίζει να σημειωθεί ότι ένα μοντέλο με χαμηλότερη αμηχανία μπορεί να μην παράγει πάντα πιο ανθρώπινες ή συνεκτικές απαντήσεις (46).

8.3.5 Ικανοποίηση χρηστών και ανθρώπινη αξιολόγηση

Ενώ οι αυτοματοποιημένες μετρήσεις παρέχουν ένα ποσοτικό μέτρο της απόδοσης του chatbot, ενδέχεται να μην καταγράφουν τις αποχρώσεις των αλληλεπιδράσεων ανθρώπου-chatbot. Ως εκ τούτου, οι έρευνες ικανοποίησης χρηστών και οι ανθρώπινες αξιολογήσεις χρησιμοποιούνται συχνά για τη μέτρηση της συνολικής αποτελεσματικότητας, της συνοχής και της ευχέρειας των απαντήσεων των chatbot. Αυτές οι αξιολογήσεις μπορούν να παράσχουν

πληροφορίες για τομείς βελτίωσης που μπορεί να μην είναι εμφανείς μόνο μέσω αυτοματοποιημένων μετρήσεων (**?**).

Εν κατακλείδι, ενώ υπάρχουν πολυάριθμες μετρικές διαθέσιμες για την αξιολόγηση των chatbots, η επιλογή του σωστού συνδυασμού μετρικών που ευθυγραμμίζονται με τους συγκεκριμένους στόχους και το πλαίσιο του chatbot είναι ζωτικής σημασίας. Επιπλέον, η συνεχής παρακολούθηση και η επαναληπτική ανατροφοδότηση είναι απαραίτητες για να διασφαλιστεί ότι το chatbot εξελίσσεται και παραμένει αποτελεσματικό με την πάροδο του χρόνου.

Κεφάλαιο 9

Βιβλιοθήκη

Στο πεδίο της ανάπτυξης chatbot, οι επιλογές που γίνονται όσον αφορά τα υποκείμενα συστατικά, τις βάσεις δεδομένων, τα μοντέλα και τις μετρικές αξιολόγησης μπορούν να επηρεάσουν σημαντικά τη συνολική απόδοση και χρησιμότητα του συστήματος. Ο πρωταρχικός στόχος της παρούσας διατριβής είναι να σχεδιαστεί ένα σύστημα chatbot που όχι μόνο να υπερέχει στην πρωταρχική του λειτουργία της κατανόησης και της ανταπόκρισης σε ερωτήματα χρηστών, αλλά να είναι επίσης ελαφρύ, επεκτάσιμο και εύκολα ενσωματώσιμο. Αυτό το κεφάλαιο, εμβαθύνει στο σκεπτικό και τις επιπτώσεις των επιλογών που έγιναν κατά τη φάση της ανάπτυξης. Από την απόφαση να χρησιμοποιηθούν λύσεις βασισμένες στο υπολογιστικό νέφος (cloud) έως την επιλογή συγκεκριμένων μοντέλων ενσωμάτωσης και συμπλήρωσης, κάθε επιλογή αναλύεται για να γίνει κατανοητή η συμβολή της στις δυνατότητες του chatbot. Επιπλέον, το κεφάλαιο θα διερευνήσει τις στρατηγικές βελτιστοποίησης που χρησιμοποιήθηκαν για τη λεπτομερή ρύθμιση αυτών των στοιχείων, διασφαλίζοντας ότι το σύστημα chatbot είναι τόσο αποδοτικό όσο και αποτελεσματικό. Στο τέλος αυτού του κεφαλαίου, ο αναγνώστης θα αποκτήσει μια ολοκληρωμένη κατανόηση της πειραματικής διάταξης και της ευθυγράμμισής της με τους στόχους της διατριβής.

9.1 Δομικά Στοιχεία

Η φάση ανάπτυξης του συστήματος chatbot προσεγγίστηκε με σχολαστική προσοχή και ακρίβεια. Αναγνωρίζοντας τη σημασία μιας ελαφριάς και εύκολα ενσωματώσιμης βιβλιοθήκης chatbot, το σύστημα σχεδιάστηκε για να αξιοποιήσει τη δύναμη και την ευελιξία των λύσεων που βασίζονται στο cloud. Αυτή η στρατηγική απόφαση όχι μόνο εξασφαλίζει την επεκτασιμότητα και την ανταπόκριση σε πραγματικό χρόνο, αλλά ευθυγραμμίζεται επίσης με την έμφαση της διατριβής στη δημιουργία ενός συστήματος chatbot plug-and-play. Η προσέγγιση με επίκεντρο το νέφος, σε συνδυασμό με την επιλογή μοντέλων και βάσεων δεδομένων τελευταίας τεχνολογίας, συμπυκνώνει την ουσία της φιλοσοφίας σχεδιασμού του chatbot. Στις επόμενες ενότητες, θα εμβαθύνουμε στο σκεπτικό πίσω από κάθε επιλογή, διευκρινίζοντας τη σημασία τους στο ευρύτερο πλαίσιο των στόχων της διατριβής.

9.1.1 Fine-Tuning vs In-Context Learning

Στο ταξίδι της ανάπτυξης chatbot, δύο εξέχουσες στρατηγικές για την προσαρμογή του μοντέλου είναι η λεπτομερής ρύθμιση και η μάθηση εντός πλαισίου. Η λεπτή ρύθμιση περιλαμβάνει την επανεκπαίδευση ενός προ-εκπαιδευμένου μοντέλου σε ένα συγκεκριμένο σύνολο δεδομένων για την εξειδίκευση των γνώσεών του. Ενώ αυτή η μέθοδος μπορεί να αποδώσει εντυπωσιακά αποτελέσματα, είναι μια χρονοβόρα και αυστηρή διαδικασία. Από την άλλη πλευρά, η μάθηση εντός πλαισίου αξιοποιεί την ικανότητα του μοντέλου να γενικεύει από τα παραδείγματα που παρέχονται στην προτροπή, χωρίς την ανάγκη ρητής επανεκπαίδευσης. Δεδομένων των στόχων της παρούσας διατριβής, οι οποίοι δίνουν έμφαση στην αποτελεσματικότητα και την επεκτασιμότητα, επιλέχθηκε η μάθηση εντός πλαισίου. Η απόφαση αυτή καθοδηγήθηκε από την επιθυμία επίτευξης υψηλών επιδόσεων χωρίς την επιβάρυνση και την πολυπλοκότητα που συνδέεται με τη λεπτομερή ρύθμιση.

9.1.2 Επιλογή βάσης δεδομένων: Διανυσματική βάση δεδομένων

Η απόφαση να χρησιμοποιηθεί μια διανυσματική βάση δεδομένων πηγάζει από την εγγενή ικανότητά της να διευκολύνει τις αναζητήσεις με βάση την ομοιότητα. Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων που βασίζονται σε ακριβείς αντιστοιχίες, οι διανυσματικές βάσεις δεδομένων αξιοποιούν τις ενσωματώσεις για να επιτρέψουν την ανάκτηση δεδομένων με βάση την εγγύτητα. Αυτό είναι ιδιαίτερα επωφελές για συστήματα chatbot που στοχεύουν στην κατανόηση της πρόθεσης του χρήστη και στην παροχή σχετικών με το πλαίσιο απαντήσεων, όπως συζητήθηκε σε προηγούμενα κεφάλαια. Μεταξύ των διαθέσιμων επιλογών, επιλέχθηκε η Pinecone ως διανυσματική βάση δεδομένων για το παρόν σύστημα. Η Pinecone ξεχωρίζει ως το βιομηχανικό πρότυπο επιδόσεων, προσφέροντας απαράμιλλη ταχύτητα και ακρίβεια στις διανυσματικές αναζητήσεις. Επιπλέον, το Pinecone παρέχει την ευελιξία βελτιστοποίησης της μετρικής του αλγορίθμου KNN (Απόσταση Συνημιτόνου και Ευκλείδεια Απόσταση), προσθέτοντας ένα ακόμη επίπεδο τελειοποίησης στον μηχανισμό απάντησης του chatbot. Η επιλογή αυτή ευθυγραμμίζεται με τον στόχο της διατριβής για την αξιοποίηση τεχνολογιών αιχμής για την ενίσχυση της αποτελεσματικότητας και της επίγνωσης του πλαισίου του chatbot.

9.1.3 Chunking Method: Encode-and-Pool

Προτού τα δεδομένα κειμένου περάσουν στα μοντέλα ενσωμάτωσης, υποβάλλονται σε μια διαδικασία τεμαχισμού, ώστε να διασφαλιστεί η εισαγωγή στο σύστημα διαχειρίσιμων και ουσιαστικών τμημάτων κειμένου. Η μέθοδος τεμαχισμού που χρησιμοποιείται ονομάζεται Encode-and-Pool. Δύο κύριες στρατηγικές χρησιμοποιούνται για αυτό το chunking:

Chunking με βάση τις προτάσεις: Η μέθοδος αυτή διαιρεί το κείμενο σε επιμέρους προτάσεις, διασφαλίζοντας ότι κάθε πρόταση αντιμετωπίζεται ως ξεχωριστό κομμάτι. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη για τη διατήρηση της σημασιολογικής ακεραιότητας του κειμένου.

Chunking με βάση τις λέξεις: Σε αυτή τη μέθοδο, το κείμενο διαιρείται με βάση έναν καθορισμένο αριθμό λέξεων, ο οποίος καθορίζεται από το chunkSize. Επιπλέον, διατηρείται μια επικάλυψη λέξεων, που καθορίζεται από το chunkOverlap, μεταξύ διαδοχικών τμημάτων για να διασφαλιστεί η συνέχεια και η διατήρηση του πλαισίου.

9.1.4 Μοντέλα ενσωμάτωσης λέξεων: ADA και BERT

Τα μοντέλα ενσωμάτωσης είναι βασικά εργαλεία για τη μετατροπή κειμενικού περιεχομένου σε διανυσματικές αναπαραστάσεις υψηλής διάστασης. Ένα εξέχον μοντέλο σε αυτόν τον τομέα είναι το Second Generation ADA της OpenAI. Αυτό το μοντέλο έχει τη δυνατότητα να μετατρέψει περίπου 6.000 λέξεις σε διανύσματα που εκτείνονται σε έναν χώρο 1.536 διαστάσεων. Ωστόσο, αξίζει να σημειωθεί ότι ενώ το ADA δεύτερης γενιάς προσφέρει σημαντικά πλεονεκτήματα όσον αφορά την απόδοση, δεν είναι ανοιχτά προσβάσιμο για τροποποίηση, καθώς ο πηγαίος κώδικάς του είναι ιδιόκτητος. Αντ' αυτού, διατίθεται στους χρήστες μέσω ενός ειδικού API. Αντίθετα, το BERT στέκεται ως φάρος στην κοινότητα του NLP, φημισμένο για τον μηχανισμό αμφίδρομης προσοχής και την ικανότητά του να αντιλαμβάνεται τα συμφραζόμενα (5). Τα συνδυασμένα πλεονεκτήματα της ADA δεύτερης γενιάς και του BERT εξασφαλίζουν μια ολιστική και ισχυρή αναπαράσταση του κειμένου, τοποθετώντας το σύστημα chatbot στην κατανόηση και αντιμετώπιση των ερωτημάτων των χρηστών με υψηλή ακρίβεια και επίγνωση του πλαισίου.

9.1.5 Μοντέλα συμπλήρωσης: GPT-4 και LLAMA 2

Η παραγωγή συνεκτικών και κατάλληλων για το πλαίσιο απαντήσεων είναι μια κρίσιμη πτυχή των συστημάτων chatbot. Για να επιτευχθεί αυτό, το σύστημα βασίζεται σε προηγμένα μοντέλα συμπλήρωσης. Το GPT-4 της OpenAI αποτελεί ένα σύγχρονο παραγωγικό μοντέλο, γνωστό για την εκτεταμένη βάση γνώσης του και την ικανότητά του να παράγει κείμενο που μοιάζει με ανθρώπινο (17).

Από την άλλη πλευρά, το LLAMA 2, που αναπτύχθηκε από την Meta AI, αποτελεί σημαντική πρόοδο στο πεδίο των μεγάλων γλωσσικών μοντέλων (LLM). Όπως περιγράφεται λεπτομερώς στην εργασία των Touvron et al., το LLAMA 2 περιλαμβάνει μια συλλογή προεκπαιδευμένων και λεπτομερώς ρυθμισμένων LLM, με κλίμακες που κυμαίνονται από 7 δισεκατομμύρια έως 70 δισεκατομμύρια παραμέτρους (50). Ειδικά βελτιστοποιημένες για περιπτώσεις χρήσης διαλόγων, οι λεπτομερώς ρυθμισμένες εκδόσεις, που ονομάζονται Llama 2-Chat, έχουν επιδείξει ανώτερες επιδόσεις σε σχέση με άλλα μοντέλα συνομιλίας ανοικτού κώδικα σε διάφορα benchmarks. Επιπλέον, με βάση τις ανθρώπινες αξιολογήσεις σχετικά με τη χρησιμότητα και την ασφάλεια, τα μοντέλα LLAMA 2-Chat θα μπορούσαν να χρησιμεύσουν ως βιώσιμες εναλλακτικές λύσεις σε ορισμένα μοντέλα κλειστού κώδικα. Η σχολαστική προσέγγιση στη λεπτομερή ρύθμιση και η έμφαση στις βελτιώσεις της ασφάλειας καθιστούν το LLAMA 2 μια συναρπαστική επιλογή για το σύστημα chatbot.

Τόσο το GPT-4 όσο και το LLAMA 2 φέρνουν τα μοναδικά τους πλεονεκτήματα στο τραπέζι, εξασφαλίζοντας ότι το σύστημα chatbot είναι εξοπλισμένο με τα καλύτερα εργαλεία για να παράγει ποικίλες, ακριβείς και σχετικές με το πλαίσιο απαντήσεις.

9.1.6 Σύνολο δεδομένων αξιολόγησης: SQuAD

Το Stanford Question Answering Dataset (SQuAD) επιλέχθηκε ως το κύριο σύνολο δεδομένων αξιολόγησης λόγω της ευρείας αναγνώρισής του στην κοινότητα NLP και της προκλητικής του φύσης (38). Η δομή του SQuAD, η οποία περιλαμβάνει την εξαγωγή απαντήσεων από αποσπάσματα, ευθυγραμμίζεται καλά με τον στόχο του chatbot να παρέχει ακριβείς και σχετικές με το πλαίσιο απαντήσεις.

9.1.7 Μέτρο αξιολόγησης: Ακρίβεια

Η ακρίβεια, ως μετρική, παρέχει ένα απλό μέτρο της απόδοσης του chatbot. Ποσοτικοποιεί το κλάσμα των σωστά προβλεπόμενων περιπτώσεων, προσφέροντας μια σαφή εικόνα της αποτελεσματικότητας του συστήματος. Δεδομένης της δομημένης φύσης του συνόλου δεδομένων SQuAD, όπου οι απαντήσεις εξάγονται από συγκεκριμένα αποσπάσματα, η ακρίβεια χρησιμεύει ως κατάλληλη μετρική για να μετρηθεί η ακρίβεια του chatbot στην ανάκτηση πληροφοριών.

9.1.8 Συμπεράσματα

Οι επιλογές που έγιναν στην πειραματική διάταξη βασίζονται στον στόχο της ανάπτυξης ενός συστήματος chatbot που δεν είναι μόνο ακριβές αλλά και με επίγνωση του περιεχομένου. Αξιοποιώντας μοντέλα, βάσεις δεδομένων και μετρικές αξιολόγησης τελευταίας τεχνολογίας, το πείραμα στοχεύει να διευρύνει τα όρια των δυνατοτήτων του chatbot, συμβαδίζοντας με τους γενικότερους στόχους της διατριβής.

9.2 Κατασκευή ενός bot και επίτευξη βελτιστοποίησης για την αξιολόγηση SQuAD

Στην περίπλοκη διαδικασία της δημιουργίας chatbot, η βιβλιοθήκη που παρουσιάζεται στην παρούσα εργασία προσφέρει ένα εύκολο, δομημένο αλλά ευέλικτο πλαίσιο για την κατασκευή ενός bot προσαρμοσμένου σε συγκεκριμένες ανάγκες. Σε αυτή την ενότητα δημιουργείται ένα bot από την βιβλιοθήκη αυτή και έπειτα βελτιστοποιείται για να αποδίδει στην αξιολόγηση SQuAD. Αυτή η ενότητα εμβαθύνει στο ταξίδι της κατασκευής του chatbot χρησιμοποιώντας τα δομικά στοιχεία της βιβλιοθήκης και στην διαδικασία εντοπισμού του βέλτιστου συνδυασμού για την αξιολόγηση SQuAD.

9.2.1 Κατασκευή του ChatBot με τη βιβλιοθήκη

Χρησιμοποιώντας τις διατάξεις της βιβλιοθήκης, κατασκευάστηκε σχολαστικά ένα chatbot. Η βιβλιοθήκη, σχεδιασμένη με γνώμονα την αρθρωτότητα, προσφέρει πληθώρα συνδυασμών για τα δομικά στοιχεία του bot. Δεδομένων των επιλογών που είναι διαθέσιμες, αυτή την στιγμή, για καθένα από τα τέσσερα κύρια δομικά στοιχεία, εξαιρώντας τη μεταβλητή K του αλγορίθμου KNN και τις μεταβλητές $chunkSize$ και $chunkOverlap$ της μεθόδου word-based chunking, προκύπτουν συνολικά $2^4 = 16$ πιθανοί συνδυασμοί. Αυτή η ποικιλία συνδυασμών εξασφαλίζει ότι, ενώ η βιβλιοθήκη μπορεί να διευκολύνει τη δημιουργία ενός λειτουργικού chatbot σε οποιαδήποτε διαμόρφωση, η αποδοτικότητα του bot μπορεί να ποικίλλει ανάλογα με την επιδιωκόμενη εργασία που θέτει ο προγραμματιστής.

Για να εξασφαλιστεί μια πιο λεπτή σύγκριση μεταξύ των δύο μεθόδων ομαδοποίησης, υιοθετήθηκε μια αναλυτική προσέγγιση. Ο μέσος αριθμός προτάσεων και το μήκος προτάσεων εξήχθησαν από το σύνολο δεδομένων SQuAD και οι $chunkSize$ και $chunkOverlap$ προσαρμόστηκαν ανάλογα, στρογγυλοποιώντας στον πλησιέστερο ακέραιο αριθμό για λόγους πρακτικότητας.

9.2.2 Βελτιστοποίηση για την Αξιολόγηση SQuAD

Η αξιολόγηση SQuAD, χρησιμεύει ως ένα ισχυρό μέτρο σύγκρισης για την αξιολόγηση της ικανότητας του chatbot να εξάγει ακριβείς απαντήσεις από τα παρεχόμενα κείμενα. Δεδομένης της δομημένης φύσης του συνόλου δεδομένων και της ευθυγράμμισής του με τον στόχο του chatbot να παρέχει ακριβείς και κατάλληλες από άποψη πλαισίου απαντήσεις, η διαδικασία βελτιστοποίησης ήταν αυστηρή. Οι διάφοροι συνδυασμοί των δομικών στοιχείων υποβλήθηκαν σε αυτή την αξιολόγηση, με σκοπό την εύρεση του βέλτιστου.

Τα αποτελέσματα

ΓΡΑΦΗΜΑΤΑ

9.2.3 Συμπεράσματα

Το ταξίδι της κατασκευής και βελτιστοποίησης ενός chatbot για την αξιολόγηση SQuAD υπογραμμίζει την ευελιξία και τη δύναμη της βιβλιοθήκης που παρουσιάστηκε σε αυτή τη διατριβή. Προσφέροντας ένα δομημένο πλαίσιο και ένα πλήθος συνδυασμών, η βιβλιοθήκη δίνει τη δυνατότητα στους προγραμματιστές να δημιουργήσουν chatbots προσαρμοσμένα σε συγκεκριμένες ανάγκες και στη συνέχεια να τα τελειοποιήσουν για βέλτιστη απόδοση. Αυτή η ενότητα, περιγράφοντας λεπτομερώς τη διαδικασία κατασκευής και βελτιστοποίησης, παρέχει ένα σχέδιο για τους προγραμματιστές που στοχεύουν να αξιοποιήσουν στο έπακρο τις δυνατότητες της βιβλιοθήκης.

Κεφάλαιο 10

Συμπεράσματα

Το ταξίδι της ανάπτυξης ενός συστήματος chatbot, όπως περιγράφεται λεπτομερώς σε αυτή τη διατριβή, ήταν τόσο προκλητικό όσο και διαφωτιστικό. Ο πρωταρχικός στόχος ήταν να σχεδιαστεί ένα σύστημα το οποίο όχι μόνο θα είναι ικανό στην κατανόηση και την απάντηση σε ερωτήματα χρηστών, αλλά θα είναι επίσης ελαφρύ, επεκτάσιμο και εύκολα ενσωματώσιμο. Μέσω σχολαστικής έρευνας, πειραματισμού και βελτιστοποίησης, ο στόχος αυτός επιτεύχθηκε.

Οι επιλογές που έγιναν όσον αφορά τα υποκείμενα στοιχεία, τις βάσεις δεδομένων, τα μοντέλα και τις μετρικές αξιολόγησης αναλύθηκαν για να κατανοηθεί η συμβολή τους στις δυνατότητες του chatbot. Η απόφαση να χρησιμοποιηθούν λύσεις βασισμένες στο cloud, σε συνδυασμό με την επιλογή μοντέλων και βάσεων δεδομένων τελευταίας τεχνολογίας, συμπυκνώνει την ουσία της φιλοσοφίας σχεδιασμού του chatbot. Η επαναληπτική διαδικασία κατασκευής και βελτιστοποίησης του chatbot για την αξιολόγηση SQuAD υπογράμμισε περαιτέρω την ευελιξία και τη δύναμη της βιβλιοθήκης που παρουσιάστηκε στην παρούσα διατριβή.

Στην ουσία, η παρούσα διατριβή έθεσε ένα σχέδιο για την ανάπτυξη ενός συστήματος chatbot που είναι τόσο αποδοτικό όσο και αποτελεσματικό, διευρύνοντας τα όρια των δυνατοτήτων του chatbot.

Κεφάλαιο 11

Συζήτηση και μελλοντικές εργασίες

Ενώ τα αποτελέσματα που επιτεύχθηκαν σε αυτή τη διατριβή είναι πολλά υποσχόμενα, υπάρχουν πάντα περιθώρια βελτίωσης και εξερεύνησης στον συνεχώς εξελισσόμενο τομέα της ανάπτυξης chatbot.

Η επιλογή της μάθησης εντός του πλαισίου έναντι της λεπτομερούς ρύθμισης, αν και αποτελεσματική, ανοίγει συζητήσεις σχετικά με τους συμβιβασμούς μεταξύ προσαρμοστικότητας και εξειδίκευσης. Ενώ η μάθηση εντός πλαισίου προσφέρει ευελιξία, η λεπτομερής ρύθμιση μπορεί να παρέχει μια πιο εξειδικευμένη βάση γνώσεων για συγκεκριμένες εφαρμογές.

Η χρήση της αξιολόγησης SQuAD ως πρωτεύον μέτρο σύγκρισης, αν και αυστηρή, είναι μόνο μία από τις πολλές πιθανές μετρικές αξιολόγησης. Διαφορετικά σημεία αναφοράς μπορεί να προσφέρουν διαφορετικές γνώσεις σχετικά με τις δυνατότητες του chatbot.

Μπορούν να διερευνηθούν διάφοροι δρόμοι στη μελλοντική έρευνα:

- **Μοντέλα Συμπλήρωσης**: Ενώ χρησιμοποιήθηκαν μοντέλα όπως το GPT-4 και το LLAMA 2, μπορεί να εξεταστεί η διερεύνηση νεότερων μοντέλων ή ακόμη και υβριδικών μοντέλων.
- **Επέκταση βάσης δεδομένων**: Η επιλογή του Pinecone ως διανυσματικής βάσης δεδομένων βασίστηκε στις τρέχουσες δυνατότητές του. Ωστόσο, καθώς εμφανίζονται νεότερες βάσεις δεδομένων, μπορεί να διερευνηθεί η ενσωμάτωση και η απόδοσή τους.
- **Fine-Tuning**: Μια βαθύτερη εμβάθυνση στις στρατηγικές λεπτής ρύθμισης, διερευνώντας τα πιθανά οφέλη και τα αντισταθμιστικά οφέλη της.
- **Multimodal(Πολυτροπικά) Chatbots**: Με την άνοδο των πολυτροπικών μοντέλων που κατανοούν τόσο το κείμενο όσο και τις εικόνες, η ανάπτυξη ενός chatbot που μπορεί να επεξεργάζεται πολυτροπικά ερωτήματα μπορεί να αποτελέσει μια πιθανή κατεύθυνση.
- **Real-world Deployment**: Ανάπτυξη του chatbot σε σενάρια πραγματικού κόσμου, συλλογή ανατροφοδότησης από τους χρήστες και επανάληψη με βάση την απόδοση στον πραγματικό κόσμο.

Συμπερασματικά, ο τομέας της ανάπτυξης chatbot είναι ευρύς και δυναμικός. Η εργασία που παρουσιάζεται στην παρούσα διατριβή είναι ένα βήμα προς τα εμπρός, αλλά το ταξίδι της εξερεύνησης και της καινοτομίας είναι ατελείωτο.

Βιβλιογραφία

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

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

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

[4] Bender, E. M., et al. (2021). Dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

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

[7] Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

[8] Gully Burns, Nanyun Peng, and Xiangci Li. Scientific Discourse Tagging for Evidence Extraction. *arXiv preprint arXiv:1909.04758*, 2019.

[9] Wenliang Chen, Pingfu Chao, Baoxing Huai, Zhefeng Wang, Mengsong Wu, Junfei Ren, Zijian Yu, Zechang Li, Guoliang Zhang, and Tong Zhu. CED: Catalog Extraction from Documents. *arXiv preprint arXiv:2304.14662*, 2023.

[10] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., ... & Batra, D. (2017). *Visual dialog*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 2, p. 3).

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

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

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

[14] 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.

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

[16] Girshick, R. (2015). *Fast R-CNN*. arXiv preprint arXiv:1504.08083.

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

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

[19] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[20] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). *A fast learning algorithm for deep belief nets*. Neural computation, 18(7), 1527-1554.

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

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

[23] 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).

[24] Lan, Z., Li, A., He, H., Zhang, S., & Qiu, H. (202

[25] Li, Q., Lai, Z., Qiao, Y., Dai, J., Luo, P., Wang, L. M., ... & Zhu, X. (2023). InternGPT: Solving Vision-Centric Tasks by Interacting with ChatGPT Beyond Language.

[26] 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.

[27] Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully convolutional networks for semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

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

[29] 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.

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

[31] Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). *Context Encoders: Feature Learning by Inpainting*. CVPR.

[32] 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.

[33] Raffel, C., et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

[34] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[35] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

[36] Ren, S., He, K., Girshick, R., & Sun, J. (2017). *Faster R-CNN: Towards real-time object detection with region proposal networks*. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.

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

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

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

[40] Too, E. C., Li, Y., Njuki, S., & Liu, Y. (2019). *A comparative study of fine-tuning deep learning models for plant disease identification*. Computers and Electronics in Agriculture, 162, 735-740.

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

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

[43] Wager, T. D., & Atlas, L. Y. (2015). *The neuroscience of placebo effects: connecting context, learning and health*. Nature Reviews Neuroscience, 16(7), 403-418.

[44] Wallace, E., et al. (2019). Universal adversarial triggers for attacking and analyzing NLP. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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

[46] Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... & Dolan, B. (2019). *Dialogpt: Large-scale generative pre-training for conversational response generation*. arXiv preprint arXiv:1911.00536.

[47] Zhang, H., Stoica, I., Gonzalez, J. E., Xing, E. P., Lin, Z., Li, Z., ... & Zheng, L. (2023). LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset.

[48] J. Pennington, R. Socher, C. D. Manning (2014) GloVe: Global Vectors for Word Representation

[49] A. Vaswani and N. Shazeer and N. Parmar and J. Uszkoreit and L. Jones and A. N. Gomez and L. Kaiser and I. Polosukhin (2017) Attention Is All You Need

[50] Touvron, Hugo et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv preprint arXiv:2307.09288, 2023.

[51]

[52]

[53]