Αλέξανδρος Παναγιώτου

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Διπλωματική Εργασία

Easy to Use Java QnA Bot Library

Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Περιεχόμενα

Πρόλογος

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

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

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

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

Κεφάλαιο 1

Εισαγωγή

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

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

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

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

Επιπλέον, με την αυξανόμενη εξάρτηση από τα μοντέλα τεχνητής νοημοσύνης και ML, υπάρχει αυξανόμενη ζήτηση για επεξηγηματικότητα και διαφάνεια σε αυτά τα μοντέλα (16). 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 (17).

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 (2). Σχεδιασμένο ως πείραμα για την προσομοίωση ενός ροτζεριανού ψυχοθεραπευτή, το ELIZA βασίστηκε σε μεθοδολογίες αντιστοίχισης προτύπων και υποκατάστασης για την προσομοίωση της συζήτησης. Οι χρήστες εισήγαγαν δηλώσεις και το ELIZA απαντούσε με βάση ένα σύνολο κανόνων σεναρίου, συχνά αντανακλώντας τα λόγια του ίδιου του χρήστη. Παρά την απλότητά του, το ELIZA κατάφερε να πείσει πολλούς χρήστες για την "κατανόησή" του, αναδεικνύοντας τις δυνατότητες της επικοινωνίας μέσω μηχανής. Αυτό το πρώιμο πείραμα υπογράμμισε τις βαθιές επιπτώσεις των μηχανών που μπορούσαν να "συνομιλούν" και έθεσε τις βάσεις για τις μετέπειτα εξελίξεις στον τομέα αυτό.

2.2 Από τα συστήματα βασισμένα σε κανόνες στα συστήματα μάθησης

Στη μετά-ELIZA εποχή εμφανίστηκαν αρκετά συστήματα chatbot, τα περισσότερα από τα οποία είχαν τις ρίζες τους σε παραδείγματα βασισμένα σε κανόνες. Συστήματα όπως το PARRY (3), 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 (4). 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. (**?**) 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 (**?**), 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 (5), 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.

Βιβλιογραφία

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[19] 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.

[20] 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.