Multi-level Persian Dataset for Information Retrieval

Ali Abedzadeh
Master of Software Engineering; Faculty of Computer Engineering; University of Isfahan; Isfahan, Iran;
Email: a.abedzadeh@eng.ui.ac.ir

Reza Ramezani*
Ph.D. in Computer Engineering; Associate Professor; Faculty of Computer Engineering; University of Isfahan; Isfahan, Iran Email: r.ramezani@eng.ui.ac.ir

Afsaneh Fatemi
Ph.D. in Computer Engineering ; Associate Professor; Faculty of Computer Engineering; University of Isfahan; Isfahan, Iran Email: a_fatemi@eng.ui.ac.ir

Iranian Journal of Information Processing and Management

Received: 18, Mar. 2023 **Accepted: 18, Nov. 2023**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 39 | No. 3 | pp. 1109-1138
Spring 2024
https://doi.org/10.22034/jipm.2024.710246

Abstract: Information retrieval systems are an essential part of many smart systems. The applications of this research field include search engines such as Google and Bing, question-answering systems, modern databases, etc. An information retrieval system tries to retrieve documents related to a question/query. The retrieval is done from a large collection of documents, and the size of this collection can be from a few thousand documents to millions of documents. In recent years, a lot of research has been done to develop information retrieval systems using language models. However, in this research field, no research has been done for the Persian language. One of its main reasons is the lack of a suitable Persian dataset for training language models. In this research, first, a Persian dataset for information retrieval is presented. After that, methods for enriching this data set are investigated. This enrichment is done by defining multi-level relationships between a document and a question. In this regard, the new dataset can show the relationship between question and document in four levels (unrelated, related, highly related, completely related) instead of two levels (completely unrelated, completely related). The name of the generated dataset is PersianMLIR. Experiments show that by using multi-level relationships, the performance of the system improves for both Persian and English languages, where the improvement is 1.87% for the Persian language. The results conclude that enriching information retrieval datasets by increasing the number of relations between query and document lead to improving the performance of information retrieval systems.

Keywords: Information Retrieval, Language Models, Information Retrieval Dataset, Persian Dataset

* Corresponding Author

مجموعه‌دادهٔ چندسطحی فارسی برای بازیابی اطلاعات

علی عابدزاده
کارشناسی ارشد مهندسی کامپیوتر؛ دانشکده مهندسی کامپیوتر؛ دانشگاه اصفهان؛ اصفهان، ایران؛
پدیدآور رابط a.abedzadeh@eng.ui.ac.ir

رضا رمضانی
دکتری تخصصی مهندسی کامپیوتر دانشیار؛ دانشکده مهندسی کامپیوتر؛ دانشگاه اصفهان؛ اصفهان، ایران؛
r.ramezani@eng.ui.ac.ir

افسانه فاطمی
دکتری تخصصی مهندسی کامپیوتر دانشیار؛ دانشکده مهندسی کامپیوتر؛ دانشگاه اصفهان؛ اصفهان، ایران؛
a_fatemi@eng.ui.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۱/۱۲/۲۷ | پذیرش: ۱۴۰۲/۰۸/۲۷ | مقاله برای اصلاح به مدت ۵ ماه و ۱۶ روز نزد پدیدآوران بوده است.

چکیده: هر سامانهٔ بازیابی اطلاعات وظیفه دارد با دریافت یک پُرسه، اسناد مرتبط با آن پُرسه را بازیابی کند. این بازیابی از میان مجموعه‌ای بزرگ از هزاران تا میلیون‌ها سند انجام می‌شود. در سال‌های اخیر، پژوهش‌های زیادی برای توسعهٔ سامانه‌های بازیابی اطلاعات با استفاده از مدل‌های زبان انجام شده است؛ اما در این زمینه، پژوهشی برای زبان فارسی یافت نشد. یکی از علت‌های اصلی این امر، نبودِ یک مجموعه‌دادهٔ فارسی مناسب برای آموزش مدل‌های زبان است. در این پژوهش، ابتدا یک مجموعه‌دادهٔ بازیابی اطلاعات فارسی ارائه شده و پس از آن، روش‌هایی برای غنی‌سازی این مجموعه‌داده مورد بحث قرار گرفته است. این غنی‌سازی با کمک چندسطحی کردن ارتباط میان پُرسه و سند انجام می‌شود؛ به نحوی که مجموعه‌دادهٔ جدید می‌تواند رابطه بین پُرسه و سند را به‌جای دو سطح (کاملاً نامرتبط، کاملاً مرتبط) در چهار سطح (نامرتبط، مرتبط، بسیار مرتبط، و کاملاً مرتبط) نشان دهد. مجموعه‌داده ایجادشده PersianMLIR نام دارد. آزمایش‌ها بیانگر بهبود عملکرد سامانه، هم برای زبان فارسی و هم برای زبان انگلیسی است و این میزان بهبود برای زبان فارسی ۱/۸۷ درصد است.

کلیدواژه‌ها: بازیابی اطلاعات، مدل‌های زبان، مجموعه‌داده بازیابی اطلاعات، مجموعه داده فارسی

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۹ | شمارهٔ ۳ | ص‌ص ۱۱۰۹-۱۱۳۸
بهار ۱۴۰۳
https://doi.org/10.22034/jipm.2024.710246

۱. مقدمه

سامانه‌های بازیابی اطلاعات متنی کاربردهای بسیاری دارند. طی دهه‌ها پژوهش در این زمینه، محققان روش‌های متفاوتی برای ساخت سامانه‌های بازیابی اطلاعات ارائه کرده‌اند. از مشهورترین و پراستفاده‌ترین روش‌های بازیابی اطلاعات می‌توان به TF-IDF (Salton and Buckley 1988) و BM25 (Robertson and Zaragoza 2009) اشاره کرد. این دو روش، بر اساس مدل بسته کلمات و شباهت لغوی عمل می‌کنند. بر اساس معیارهای Trec-DL (Craswell et al. 2020; Craswell et al. 2021a)، می‌توان سامانه‌های بازیابی اطلاعات را بر اساس راهکار شباهت‌سنجی به سه دسته تقسیم کرد:

- سامانه‌های سنتی: این سامانه‌ها بر اساس روش‌های قدیمی مانند BM25 و TF-IDF (و به‌طور کلی، شباهت لغوی میان اسناد و پُرسه[1] و [2]) کار می‌کنند. در این سامانه‌ها استفاده از نمایش تُنُک اسناد بسیار رایج است؛
- سامانه‌های شبکه عصبی: در این سامانه‌ها برای بازیابی از شبکه‌های عصبی کمک گرفته می‌شود. تمام سامانه‌هایی که بر اساس شبکه عصبی هستند، اما شبکه عصبی مورد استفاده یک مدل زبان[3] نیست در این دسته قرار می‌گیرند. در این سامانه‌ها استفاده از نمایش متراکم اسناد بسیار رایج است؛
- سامانه‌های شبکه عصبی مدل زبان: اگر شبکه عصبی مورد استفاده در یک سامانه از نوع مدل زبان، مانند BERT (Devlin et al. 2019) باشد، آن سامانه در این دسته قرار می‌گیرد. در این سامانه‌ها نیز استفاده از نمایش متراکم اسناد بسیار رایج است.

از اصلی‌ترین اشکالات روش‌های سنتی مبتنی بر نمایش تُنُک اسناد، به‌وجود آمدن اختلاف واژگانی میان سند و پُرسه کاربر است (برای نمونه «سرخ» و «قرمز») (Mitra and Craswell 2018; Qu et al. 2021). اِشکال دیگر نمایش تُنُک اسناد زمانی پدیدار می‌شود که یک کلمه دو یا چند مفهوم متفاوت داشته باشد؛ مانند «شیر». در نحوه نمایش متراکم اسناد که به‌طور معمول با کمک مدل‌های زبان و شبکه‌های عصبی تولید می‌شوند، معایب روش‌های قبل تا حد زیادی حل شده است. در این روش‌ها به‌جای توجه به

1. query

۲. در این مقاله به‌منظور یکنواختی، از کلمه «پرسه» به‌جای پرسش یا پرسه استفاده می‌شود.

3. language model

تشابه لغوی، تلاش می‌شود که معنا و مفهوم پُرسه و سند درک شود. در این حالت، نه‌تنها مشکل اختلاف واژگانی حل می‌شود، بلکه یک بازیابی معنایی وجود خواهد داشت (Liu et al. 2021b; Qu et al. 2021). اگرچه این روش‌ها دقت بالاتری دارند، اما در رودررویی با واژگان نادر و کمیاب، عملکرد خوبی ندارند و در چنین مواردی، نحوه نمایش تُنُک اسناد می‌تواند نتایج بهتری به‌دست دهد (Mitra and Craswell 2018).

در سال‌های اخیر، پژوهش‌های بسیاری برای ساخت سامانه‌های بازیابی اطلاعات با کمک مدل‌های زبان انجام شده است. از آنجا که مدل‌های زبان به‌طور عمده بر اساس شبکه‌های عصبی عمیق هستند، وجود مجموعه‌داده‌های بزرگ‌مقیاس برای آموزش آن‌ها ضروری است (Zhang, Yates and Lin 2020). در طی سال‌های اخیر، مدل‌های زبان زیادی ساخته شده‌اند و تعداد قابل توجهی از این مدل‌ها قادر به درک و فهم زبان فارسی هستند. با این حال، به‌دلیل نبودِ یک مجموعه‌داده مناسب، پژوهشی روی بازیابی اطلاعات فارسی با کمک مدل‌های زبان انجام نشده است.

در این پژوهش، ابتدا روشی برای ساخت یک مجموعه‌داده بازیابی اطلاعات با کمک مجموعه‌داده‌های درک مطلب ماشینی[1] فارسی موجود ارائه شد[2] و سپس، روشی برای غنی‌سازی این مجموعه‌داده جدید ارائه گردید. در مجموعه‌داده تولیدشده اولیه، همانند مجموعه‌داده‌های بازیابی اطلاعات مرسوم، به ازای هر پُرسه و سند، تنها یکی از دو رابطه «کاملاً نامرتبط» یا «کاملاً مرتبط» وجود دارد. هرچند این مجموعه‌داده تولیدشده اولین مجموعه‌دادهٔ بازیابی اطلاعات برای زبان فارسی است، اما در ادامه، به‌منظور بهبود عملکرد آن، سطح روابط میان هر پُرسه و سند را از دو سطح به چهار سطح افزایش می‌دهیم؛ به نحوی که ارتباط بین پُرسه و سند می‌تواند یکی از چهار حالت «نامرتبط»، «مرتبط»، «بسیار مرتبط» و «کاملاً مرتبط» باشد. مجموعه‌داده ساخته شده PersianMLIR[3] نام دارد[4]. سپس، یک سامانه بازیابی اطلاعات مبتنی بر مدل زبان روی هر دو مجموعه‌داده آموزش داده شده است. نتایج ارزیابی نشان می‌دهد که مجموعه‌داده چهارسطحی می‌تواند به‌طور میانگین

1. machine reading comprehension

۲. این مجموعه‌داده درک زبان ماشینی فارسی در خوشه تحقیقاتی کلان‌داده دانشگاه اصفهان توسعه داده شده است.

3. Persian Multi-Level Information Retrieval (PersianMLIR)

۴. لینک دانلود مجموعه‌داده: https://github.com/BigData-IsfahanUni/PersianMLIR

عملکرد مجموعه‌داده دو سطحی را ۱/۸۷ درصد بهبود دهد.

۲. پیشینه

از آنجا که نوآوری اصلی این پژوهش ارائه یک مجموعه‌داده بازیابی اطلاعات فارسی چندسطحی است، در این بخش به بررسی مجموعه‌داده‌های بازیابی اطلاعات و درک مطلب ماشینی انگلیسی و فارسی می‌پردازیم. علت بررسی مجموعه‌داده‌های درک مطلب ماشینی این است که این مجموعه‌داده‌ها شباهت زیادی به مجموعه‌داده‌های بازیابی اطلاعات دارند و در مواردی می‌توان با اِعمال تغییراتی در آن‌ها، یک مجموعه‌داده بازیابی اطلاعات ایجاد کرد. البته، تبدیل کردن یک مجموعه‌داده درک مطلب ماشینی به یک مجموعه‌داده بازیابی اطلاعات می‌تواند مشکلاتی چون سوگیری داده‌های آموزشی، بی‌معنا بودن پُرسه‌ها در صورت وجود نداشتن پاراگراف مربوط، همپوشانی لغوی زیاد بین پُرسه و سند مرتبط به‌وجود آورد. با وجود این، تاکنون از روی این مجموعه‌داده‌های درک مطلب ماشینی، مجموعه‌داده‌های بازیابی اطلاعات مختلفی ایجاد شده است.

مجموعه‌داده MS Marco (Bajaj et al. 2016; Craswell et al. 2021b) را می‌توان بزرگ‌ترین مجموعه‌داده بازیابی اطلاعات انگلیسی دانست که حاوی بیش از ۵۰۰ هزار پُرسه و حدود ۸/۸ میلیون سند (پاراگراف) است. این مجموعه‌داده، یک مجموعه‌داده دوسطحی محسوب می‌شود و مدل‌های بازیابی اطلاعات بسیاری در مرحله آموزش از آن استفاده می‌کنند.

Trec-DL یکی دیگر از مجموعه‌داده‌های معروف بازیابی اطلاعات است که در آن برای هر پُرسه، تعداد زیادی سند نشانه‌گذاری‌شده وجود دارد. بر خلاف مجموعه‌داده MS Marco که یک مجموعهٔ دوسطحی است، مجموعه‌داده Trec-DL چندسطحی است و در آن سطوح روابط بین پُرسه‌ها و اسناد به‌صورت زیر است:

- ◇ کاملاً مرتبط (سطح ۳): سند متعلق به پُرسه است و پاسخ دقیق پُرسه در آن قرار دارد؛
- ◇ بسیار مرتبط (سطح ۲): سند تا حدودی پاسخ پُرسه را در خود دارد، اما پاسخ، دقیق واضح نیست و یا در میان اطلاعات اضافه پنهان شده است؛
- ◇ مرتبط (سطح ۱): سند در ارتباط با پُرسه به نظر می‌رسد، اما پاسخ در آن وجود ندارد؛
- ◇ نامرتبط (سطح ۰): سند هیچ ارتباطی با پُرسه ندارد.

مجموعه‌داده SQuAD (Rajpurkar et al. 2016; Rajpurkar, Jia and Liang 2018) یک مجموعه‌داده درک مطلب ماشینی انگلیسی است که شامل بیش از ۱۰۰ هزار پُرسه و

پاراگراف است. مجموعه‌داده‌های درک مطلب ماشینی به‌طور معمول یک مجموعه از سه‌تایی‌ها به فرم (پاراگراف، پُرسه، پاسخ پُرسه) هستند. با استفاده از این سه‌تایی‌ها، می‌توان درک مطلب ماشینی را آموزش داد که با دریافت پُرسه و پاراگراف سعی می‌کند پاسخ پُرسه را از داخل پاراگراف پیدا کند. به‌طور مشابه، مجموعه‌دادهٔ بازیابی اطلاعات ANTIQUE (Hashemi et al. 2020) شامل بیش از ۲۵۰۰ پُرسه پیچیده و دشوار است که از تالارهای گفت‌وگوی آنلاین جمع‌آوری شده‌اند. در این مجموعه‌داده، به ازای هر پُرسه، حدود ۱۳ پاراگراف با سطوح ارتباطی مختلف توسط انسان نشانه‌گذاری شده است. مجموعه‌دادهٔ NaturalQuestions (Kwiatkowski et al. 2019) نیز یک مجموعه‌دادهٔ به‌نسبت بزرگ‌مقیاس با بیش از ۳۰۰ هزار دادهٔ آموزشی است. در این مجموعه‌داده که برای آموزش و ارزیابی درک مطلب ماشینی طراحی شده، هر پُرسه به‌طور معمول با یک پاسخ کوتاه و یک پاسخ بلند و نیز پاراگراف مربوط همراه شده است.

مجموعه‌داده‌هایی که تاکنون مورد بحث قرار گرفتند، مجموعه‌داده‌های انگلیسی بودند. با این بررسی‌ها، مجموعه‌داده بازیابی اطلاعات فارسی که برای آموزش یک مدل زبان مناسب باشد، شناسایی نشد. مجموعه‌داده مورد نظر این پژوهش باید ساختاری شبیه به مجموعه‌داده‌های انگلیسی بازیابی اطلاعات (که در بالا بررسی شدند) داشته باشد. خوشبختانه در سال‌های اخیر، فعالیت‌های زیادی در زمینه پرسش و پاسخ فارسی و درک مطلب ماشینی فارسی صورت گرفته و مجموعه‌داده‌هایی برای این زمینه‌های پژوهشی فراهم آمده که می‌توان از آن‌ها به‌منظور ایجاد مجموعه‌داده بازیابی اطلاعات فارسی استفاده نمود. به‌عنوان مثال، مجموعه ParsiNLU (Khashabi et al. 2021) متشکل از تعدادی مجموعه‌داده است که می‌توان از آن برای وظایف مختلف پردازش زبان طبیعی مانند استنتاج متنی، بازنویسی پُرسه، پرسش پاسخ چندجوابی، ترجمه ماشینی، تحلیل احساسات و درک مطلب ماشینی استفاده کرد. بخش درک مطلب ماشینی این مجموعه شامل ۱۳۰۰ پُرسه است که تمام این پُرسه‌ها با کمک موتور جست‌وجوی «گوگل» جمع‌آوری شده است.

مجموعه‌داده PersianQA (Ayoubi Sajjad & Davoodeh 2021) یک مجموعه‌داده فارسی برای درک مطلب ماشینی است که نحوه ایجاد و ساختار کلی آن مشابه SQuAD است. این مجموعه‌داده شامل حدود ۱۰ هزار داده آموزشی است و اسناد این مجموعه از ویکی‌پدیای فارسی استخراج شده است. به‌طور مشابه، مجموعه‌داده ParSQuAD (Abadani et al. 2021) که یک مجموعه فارسی برای درک مطلب ماشینی است، نسخه ترجمه‌شده SQuAD

است. مجموعه‌داده PersianQuAD (Kazemi, Mozafari and Nematbakhsh 2022) نیز مانند سایر مجموعه‌های فارسی، یک مجموعه‌داده درک مطلب ماشینی است که فرایند ساخت این مجموعه‌داده کاملاً با مجموعه‌داده SQuAD مشابه بوده و دربردارندهٔ حدود ۲۰ هزار داده آموزشی شامل سه‌تایی‌های (پُرسه، پاسخ، سند مرتبط) است. در این مجموعه نیز از اسناد ویکی‌پدیای فارسی برای طراحی مجموعه استفاده شده و برای طراحی پُرسه‌ها و نشانه‌گذاری آن‌ها از افراد فارسی‌زبان کمک گرفته شده است.

از میان مجموعه‌داده‌های فارسی نام برده‌شده، مجموعه PersianQuAD مناسب‌ترین انتخاب برای ساخت یک مجموعه‌داده جدید بازیابی اطلاعات فارسی است. علت این امر آن است که بر خلاف ParsQuAD که یک نسخه ترجمه شده است، اسناد مورد استفاده برای ساخت این مجموعه در دسترس هستند[1] (ویکی‌پدیای فارسی) و این امر برای ساخت یک مجموعه‌داده بازیابی اطلاعات ضروری است. همچنین، این مجموعه‌داده تعداد داده آموزشی بیشتری نسبت به PersianQA و ParsiNLU دارد و تعداد نمونه‌های آموزشی از اهمیت زیادی برخوردار است (Zhang et al. 2020). در مورد ساخت مجموعه‌داده بازیابی اطلاعات با استفاده از این مجموعه‌داده، در بخش بعد صحبت شده است.

۳. ساخت مجموعه‌داده بازیابی اطلاعات

همان‌طور که پیش‌تر بیان شد، یکی از نوآوری‌های این پژوهش ارائه یک مجموعه‌داده بازیابی اطلاعات فارسی است. در این بخش ابتدا نحوه تبدیل مجموعه‌داده PersianQuAD که یک مجموعه‌داده درک مطلب ماشینی است، به یک مجموعه‌داده بازیابی اطلاعات بیان شده است. مجموعه‌داده ساخته‌شده اولیه در این بخش یک مجموعه دوسطحی است. پس از آن، نحوه غنی‌سازی این مجموعه‌داده به کمک افزایش سطح روابط از دو سطح به چهار سطح مورد بحث قرار گرفته است.

۳-۱. ساخت مجموعه‌داده بازیابی اطلاعات فارسی دوسطحی

همان‌طور که اشاره شد، برای زبان فارسی مجموعه‌داده بازیابی اطلاعاتی که برای آموزش و تنظیم دقیق مدل‌های زبان مناسب باشد، وجود ندارد؛ اما با کمک مجموعه‌داده‌های فارسی درک مطلب ماشینی، می‌توان یک مجموعه‌داده جدید تولید

۱. برای ساخت این مجموعه‌داده از اسناد ویکی‌پدیای فارسی استفاده شده است.

کـرد. بـرای ایـن منظـور در ایـن پژوهـش از مجموعـه‌داده PersianQuAD (Kazemi, Mozafari and Nematbakhsh 2022) استفاده شـده اسـت. در ایـن مجموعـه‌داده بـه ازای هـر پُرسـه، چندیـن سـند مرتبـط وجـود دارد. در ایـن نوشـتار، منظـور از «سـند» یـک سـند کامـل، یـک پاراگـراف، یـا حتـی یـک جملـه اسـت.

در ایـن پژوهـش از بازیابـی متـن متراکـم[1] (Karpukhin et al. 2020) به‌عنـوان مـدل بازیابـی اطلاعـات اسـتفاده خواهـد شـد. بنابرایـن، سـاختار مجموعـه‌داده بازیابـی اطلاعـات فارسـی بایسـتی بـه نحـوی باشـد کـه توسـط DPR قابـل اسـتفاده باشـد. DPR یـک مـدل بازیابـی اطلاعـات اسـت کـه بـا کمـک آن می‌تـوان اسـناد مرتبـط بـا یـک پُرسـه را بازیابـی کـرد. در ایـن مـدل از دو شـبکه عصبـی مجـزا بـرای تعبیـه کـردن پُرسـه‌ها و اسـناد اسـتفاده می‌شـود و شـباهت میـان خروجی‌هـای ایـن دو شـبکه نمایانگـر میـزان ارتبـاط میـان پُرسـه و سـند خواهـد بـود. مجموعـه‌داده مـورد اسـتفاده در DPR شـامل سـه فایـل اسـت:

- **فایـل مجموعـه اسـناد:** یـک پیکـره متنـی بـزرگ اسـت کـه بازیابـی بـر روی آن انجـام می‌شـود. در ایـن پیکـره هـر سـند می‌توانـد یـک سـند کامـل، یـک پاراگـراف، یـا یـک جملـه باشـد؛
- **فایـل مجموعـه آموزشـی:** داده‌هـای آموزشـی مـدل کـه هـر نمونـه شـامل یـک پُرسـه، اسـناد مرتبـط و همین‌طـور اسـناد نامرتبـط بـا آن پُرسـه اسـت؛
- **فایـل مجموعـه ارزیابـی:** ایـن فایـل ماننـد مجموعـه آموزشـی اسـت، امـا بـا هـدف ارزیابـی سـامانه اسـتفاده می‌شـود.

نحوه ساخت این سه فایل در ادامه بیان شده است.

۳-۱-۱. ساخت مجموعه‌ی اسناد

در سـاخت PersianQuAD ابتـدا بـا کمـک الگوریتـم PageRank حـدود ۴۴۰۰ سـند «ویکی‌پدیـا» انتخـاب شـده اسـت (Kazemi, Mozafari and Nematbakhsh 2022). سـازندگان ایـن مجموعـه‌داده بـا اسـتفاده از PageRank اطمینـان حاصـل کرده‌انـد کـه صفحـات مهـم و پُرارجـاع «ویکی‌پدیـا» کـه دربردارنـده اطلاعـات داغ و محبوبـی هسـتند، انتخـاب شـده‌اند. سـپس، بـا کمـک ۱۹۰۵ سـند کـه می‌تـوان آن‌هـا را بـه ۲۶هـزار پاراگـراف تبدیـل کـرد، مجموعـه

1. dense passage retrieval (DPR)

داده درک مطلب ماشینی تولید شده است[۱]. در این پژوهش برای ساختن مجموعهٔ اسناد، از این ۲۶هزار پاراگراف استفاده شده است. با وجود این، از آنجا که قصد داریم مجموعهٔ اسناد مورد استفاده در این پژوهش کمی بزرگ‌مقیاس‌تر از این تعداد سند باشد، لازم است اسناد دیگری نیز به این مجموعه اضافه شوند. برای این منظور، ابتدا یک نسخه متنی کامل از «ویکی‌پدیای فارسی» بارگیری شد. پس از آن، متون داخل این فایل استخراج شده و به پاراگراف‌هایی با اندازه حداکثر ۵۰۰ واژه تبدیل شدند. سرانجام، با انتخاب پاراگراف‌ها به‌صورت تصادفی، اندازهٔ مجموعهٔ اسناد به ۲۰۰هزار پاراگراف رسیده است. لازم به ذکر است که برای جلوگیری از وارد شدن پاراگراف‌های تکراری در مجموعهٔ اسناد، از وارد شدن پاراگراف‌هایی که عنوان سند آن‌ها در لیست اسناد مورد استفاده PersianQuAD وجود داشته است، خودداری شده است.

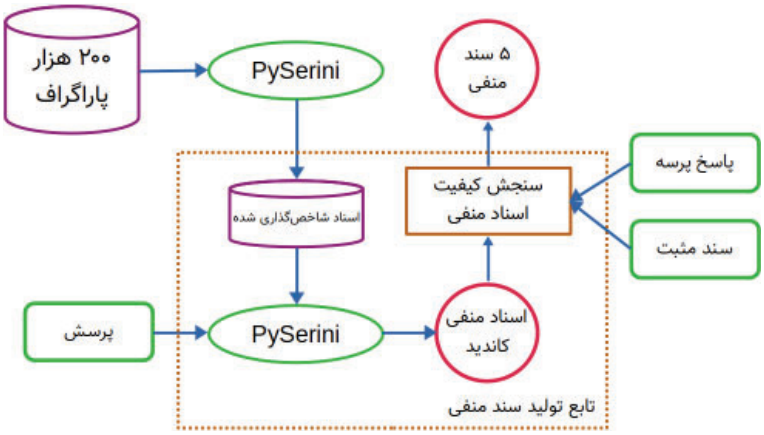
۳-۱-۲. ساخت مجموعه آموزشی

در مجموعه‌دادهٔ آموزشی بازیابی اطلاعات، هر نمونه یک سه‌تایی به شکل پُرسه، مجموعه اسناد مرتبط، و مجموعه اسناد نامرتبط است. با کمک PersianQuAD می‌توان دو عنصر اول این سه‌تایی را تولید کرد، اما برای تولید مجموعه اسناد نامرتبط باید از مجموعه کل اسناد که در مرحله قبل ساخته شده، کمک گرفت. اسناد مرتبط به‌طور معمول نشانه‌گذاری شده‌اند؛ اما برای انتخاب اسناد نامرتبط نیاز به یک مجموعه اسناد بزرگ است. بنابراین، فرض می‌شود که هر سند در این مجموعه، به‌غیر از اسناد مرتبط که نشانه‌گذاری شده‌اند، اسناد نامرتبط و منفی هستند (Karpukhin et al. 2020). روش‌های زیادی برای انتخاب سند منفی وجود دارد (Karpukhin et al. 2020; Liu et al. 2021a; Qu et al. 2021)، اما یکی از رایج‌ترین روش‌ها استفاده از اسناد با رتبه BM25 بالاست که حاوی پاسخ نیستند. در این پژوهش برای انتخاب اسناد منفی، ابتدا اسناد با کمک ابزار PySerini (Lin et al. 2021; Yang, Fang and Lin 2017) شاخص‌گذاری شده‌اند. PySerini ابزاری برای شاخص‌گذاری و بازیابی اطلاعات بر روی اسناد است که می‌توان از آن برای بازیابی اسناد با کمک BM25 استفاده کرد. پس از شاخص‌گذاری اسناد، به ازای هر پُرسه، پنج سند برتر BM25 که شامل پاسخ نیستند، انتخاب می‌شوند. افزون بر شرایط ذکرشده، اطمینان حاصل می‌شود که اسناد منفی انتخاب‌شده، عنوان سند متفاوتی از

۱. این ۱۹۰۵ سند از طریق سازندگان PersianQuAD در اختیار ما قرار گرفته است.

عنوان سند مرتبط داشته باشند. به این ترتیب، می‌توان اطمینان بیشتری حاصل نمود که اسناد منفی انتخاب‌شده ارتباطی با پُرسه مد نظر ندارند. در شکل ۱، کلیات فرایند انتخاب پنج سند منفی به ازای هر پُرسه نشان داده شده است. همان‌طور که در این شکل دیده می‌شود، پس از شاخص‌گذاری اسناد با کمک PySerini، به ازای هر پُرسه ورودی، تعدادی سند بازیابی می‌شود که به‌عنوان اسناد منفی کاندید در نظر گرفته می‌شوند. پس از آن با کمک محدودیت‌هایی که به آن‌ها پرداخته شد (نبودِ پاسخ در سند کاندید و تفاوت عنوان سند کاندید با سند مثبت)، این اسناد منفی فیلتر می‌شوند تا اسناد منفی باکیفیت‌تری تولید شود. اعمال کردن این محدودیت‌ها به عهده بخش «سنجش کیفیت اسناد منفی» است. از میان اسناد منفی فیلترشده، پنج سند با بالاترین امتیاز BM25 به‌عنوان اسناد منفی انتخاب می‌شوند.

با کمک روش بیان شده، می‌توان مجموعه‌داده آموزشی مورد نیاز برای آموزش DPR را ایجاد نمود. چگونگی عملکرد و نحوه آموزش مدل DPR پایه در بخش (۴-۱) توضیح داده شده است.



شکل ۱. فرایند انتخاب اسناد منفی برای هر پُرسه

۳-۱-۳. ساخت مجموعهٔ ارزیابی

روند ساخت مجموعه ارزیابی بازیابی اطلاعات، بسیار مشابه روند ساخت مجموعه آموزشی است؛ با این تفاوت که در این مجموعه آموزشی، نیازی به وجود اسناد نامرتبط و منفی نیست؛ اما در عوض، نیاز به دانستن پاسخ مربوط به پُرسه است. در این حالت

سه‌تایی‌های (پُرسه، پاسخ پُرسه، سند مرتبط) تشکیل می‌شوند. استفاده از پاسخ پُرسه در سامانه‌های بازیابی اطلاعات امری رایجی است، اما DPR از این پاسخ‌ها برای محاسبه سنجه Exact Match و سنجش عملکرد مدل تحت آموزش استفاده می‌کند. در بخش (۵-۱) به شرح سنجه‌های مورد استفاده در DPR و این پژوهش پرداخته شده است.

۳-۲. ساخت مجموعه‌داده بازیابی اطلاعات فارسی چندسطحی

مجموعه‌داده ارائه‌شده در بخش (۳-۱) یک مجموعه‌داده دوسطحی است؛ چرا که برای هر پُرسه، تنها اسناد «مرتبط» و «نامرتبط» تعیین شده است. یک مجموعه‌داده چندسطحی می‌تواند رابطه یک پُرسه و سند را در چند سطح نمایش دهد. در این پژوهش از ساختار مجموعه‌داده Trec-DL که یک مجموعه‌داده بازیابی اطلاعات چندسطحی است (Craswell et al. 2020; Craswell et al. 2021a) استفاده شده است.

مجموعه‌داده Trec-DL روابط را در چهار سطح نشان می‌دهد: کاملاً مرتبط (سطح ۳)، بسیار مرتبط (سطح ۲)، مرتبط (سطح ۱)، نامرتبط (سطح ۰). برای نمونه، اگر پُرسه مورد بررسی «مصدق در چه سالی متولد شده؟» باشد، پاراگراف «مصدق در سال ۱۲۶۱ چشم به جهان گشود و ...» *کاملاً مرتبط*، پاراگراف «در سال ۱۲۸۰ خورشیدی، مصدق که در آن زمان نوزده سال داشت ...» *بسیار مرتبط*، پاراگراف «... ازدواج این دو ۶۴ سال تا پایان زندگانی‌شان ادامه یافت» *مرتبط* و پاراگراف «پس از مرگ مظفرالدین‌شاه پسرش محمدعلی‌شاه در تاریخ ۲۹ دی ۱۲۸۵ تاج‌گذاری کرد» *نامرتبط* است. در واقع، می‌توان گفت که سطح ۰ همان سند منفی است و سطح ۳ نیز همان سند مثبت است. در این حالت، اطلاعات اضافه‌ای میان این دو سطح اضافه شده است (سطح ۱ و ۲) که مدل می‌تواند با کمک آن‌ها، بازیابی دقیق‌تر و بهتری انجام دهد.

در این پژوهش، برای تبدیل کردن مجموعه‌داده دوسطحی ساخته‌شده در بخش (۳-۱) به یک مجموعه چندسطحی، از دو روش استفاده شده است:

- روش مبتنی بر قوانین دست‌ساز؛
- روش مبتنی بر طبقه‌بندی و ویژگی‌های لغوی.

۳-۲-۱. روش مبتنی بر قوانین دست‌ساز

برای ساخت یک مجموعه‌داده چندسطحی با کمک قوانین دست‌ساز، یکصد سند برتر مرتبط با هر پُرسه با کمک PySerini استخراج شده‌اند تا بتوان از آن‌ها به‌عنوان

اسناد کاندید[1] استفاده کرد. پس از آن، با کمک قوانینی که به‌صورت دستی تعریف شده‌اند، هر سند کاندید به یکی از سطوح ۰ الی ۲ منتصب شده است (نامرتبط، مرتبط، بسیار مرتبط). سطح ۳ (کاملاً مرتبط) تنها شامل یک سند است و آن سند همان سند مثبت است و قوانین مورد بحث نقشی در انتخاب آن ندارند. برای تعیین سطح یک سند کاندید، افزون بر خود سند کاندید و عنوان آن، سند مثبت و عنوان آن و همین‌طور پاسخ کوتاه و ریشه‌های آن مورد نیاز هستند. برای استخراج ریشه‌های پاسخ از ابزار Hazm استفاده شده است[2].

قوانین تعریف‌شده به‌ترتیب زیر هستند:

- قانون ۱: اگر عنوان سند کاندید و عنوان سند مثبت یکی باشد و پاراگراف کاندید نیز شامل بخشی از ریشه لغت‌های پاسخ باشد، سند کاندید، یک سند سطح ۲ خواهد بود؛
- قانون ۲: اگر پاسخ عیناً در سند کاندید وجود داشته باشد، سند کاندید، یک سند سطح ۲ خواهد بود؛
- قانون ۳: اگر عنوان سند کاندید و عنوان سند مثبت یکی باشد یا پاراگراف کاندید شامل بخشی از ریشه لغت‌های پاسخ باشد، سند کاندید، یک سند سطح ۱ خواهد بود؛
- قانون ۴: اگر عنوان سند کاندید و عنوان سند مثبت متفاوت باشد و همین‌طور سند کاندید شامل پاسخ دقیق نباشد، سند کاندید، یک سند سطح ۰ خواهد بود.

از میان قوانین تعریف‌شده، اولین قانونی که برقرار شود، تعیین‌کننده سطح سند است و قوانین بعد از آن اثری نخواهند داشت.

۳-۲-۲. روش مبتنی بر طبقه‌بندی و ویژگی‌های لغوی

ایده اصلی این بخش آن است که با کمک یک مجموعه‌داده چندسطحی انگلیسی موجود، یک مدل آموزش داده شود که بتواند با کمک این مجموعه‌داده، میزان ارتباط اسناد را یاد بگیرد. هدف در واقع، آن است که دانش موجود در یک مجموعه‌داده

---

۱. سند کاندید سندی است که از میان کل اسناد انتخاب شده است، اما میزان ارتباط آن با پرسه مورد بررسی مشخص نیست و می‌تواند به‌صورت بالقوه در یکی از چهار سطح ارتباطی قرار بگیرد.

۲. ابزاری است مانند NLTK پایتون برای تمیز کردن متن، تجزیه نحوی، ریشه‌یابی و ...:https://www.roshan-ai.ir/hazm

چندسطحی به یک مدل منتقل شود و با کمک آن مدل، یک مجموعه‌داده جدید ایجاد شود. برای این منظور در این پژوهش، از مجموعه‌داده Trec-DL که یک مجموعه‌داده چندسطحی انگلیسی است، برای آموزش مدل استفاده گردیده و از مدل ایجادشده برای ساخت یک مجموعه‌داده فارسی استفاده شده است. برای این منظور، ابتدا نیاز به یک تابع برای استخراج ویژگی است. این ویژگی‌ها عبارت‌اند از:

- آیا سند کاندید شامل پاسخ کامل است؟ به گفته دیگر آیا پاسخ ارائه‌شده برای پُرسه، عیناً در سند کاندید وجود دارد؛
- آیا سند کاندید و سند مثبت، عنوان مشترک دارند؟
- لیستی از شباهت‌های لغوی میان سند کاندید و سند مثبت. در این لیست با کمک سه تابع شباهت کسینوسی، جاکارد، و اقلیدوسی شش مقدار مشابهت میان سند کاندید و سند مثبت محاسبه می‌شوند (دو مقدار به ازای هر تابع شباهت)[1]؛
- لیستی از شباهت معنایی میان سند کاندید و سند مثبت. در این لیست از مدل زبان‌های تولیدشده توسط DPR دوسطحی استفاده می‌شود. برای این منظور، DPR به تعداد یک، سه و شش دور تنظیم دقیق می‌شود و از مدل‌های هر اجرا، برای تعبیه‌کردن و به‌دست آوردن شباهت میان بردار سند مثبت و بردار سند کاندید استفاده می‌شود. به‌طور مشابه، میزان شباهت پُرسه و سند کاندید نیز با کمک این مدل‌ها محاسبه می‌شود. افزون بر سه مدل تنظیم دقیق‌شده بالا، یک مدل BERT نیز (بدون تنظیم دقیق شدن) برای تعبیه‌کردن اسناد و پُرسه‌ها استفاده می‌شود. برای به‌دست آوردن شباهت، از توابع ضرب داخلی و همین‌طور شباهت کسینوسی استفاده شده است. حاصل تمام این محاسبات، ۱۶ مقدار مشابهت (دو تابع شباهت‌سنجی اعمال‌شده بر روی هشت مدل زبان) می‌شود که در یک لیست قرار می‌گیرند.

با در دست داشتن این تابع، می‌توان با کمک PySerini یکصد سند کاندید را استخراج و ویژگی‌های هر یک از آن‌ها را استخراج کرد. سپس با وارد کردن این

---

۱. برای محاسبه این شباهت‌ها، ابتدا یک بردار از سند کاندید و یک بردار از سند مثبت ساخته می‌شود. هر عنصر در این دو بردار نشان‌دهنده یک واژه است. مقداردهی این عناصر می‌تواند به‌صورت دودویی (۱ برای وقوع واژه در سند و ۰ برای عدم وقوع آن) و غیر دودویی (تعداد وقوع واژه) باشد. در نتیجه، به ازای هر سند یک بردار دودویی و یک بردار غیردودویی وجود خواهد داشت و با کمک سه تابع شباهت یادشده می‌توان شش مقدار شباهت تولید کرد.

ویژگی‌ها به یک طبقه‌بند می‌توان سطح این اسناد کاندید را تعیین کرد. مانند مرحله قبل، تعیین سطح سند تنها برای سطوح ۰ الی ۲ است و مجموعه اسناد سطح ۳ مرتبط با هر پُرسه تنها شامل یک عضو است که همان سند مثبت است و طبقه‌بند نقشی در تعیین آن ندارد.

مدل‌هایی که در این بخش برای طبقه‌بندی اسناد مورد استفاده قرار گرفتند، عبارت‌اند از: درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی، نزدیک‌ترین همسایه، AdaBoost[1]، و SGD[2].

برای آموزش این مدل‌ها از ویژگی‌های استخراج‌شده در مراحل قبل استفاده شده است و برای سطح ارتباط که بایستی پیش‌بینی شود نیز از سطوح تعیین‌شده در Trec-DL که توسط سازندگان مجموعه داده برای هر پاراگراف مشخص شده، استفاده گردیده است. در میان این ویژگی‌ها مواردی که از پاسخ استفاده می‌کنند، وجود ندارد؛ چرا که مجموعه‌داده Trec-DL چنین اطلاعاتی را در خود ندارد. از آنجا که در میان این ویژگی‌ها، ویژگی‌های مرتبط با پاسخ وجود ندارد، سایر ویژگی‌های باقی‌مانده مستقل از زبان هستند و می‌توان توابع مورد استفاده برای مجموعه انگلیسی را برای استخراج ویژگی از مجموعه فارسی استفاده کرد. پس از آموزش و ارزیابی مدل، می‌توان مانند بخش قبل، اسناد کاندید فارسی را با کمک PySerini و BM25 تولید نمود، ویژگی‌های آن‌ها را استخراج کرد، و سپس آن‌ها را طبقه‌بندی نمود. سرانجام، با کمک طبقه‌بندی‌های انجام‌شده، می‌توان یک مجموعه‌داده چندسطحی فارسی ساخت.

٤. آموزش مدل بازیابی اطلاعات

٤-١. مدل DPR پایه

برای استفاده از مجموعه‌داده بازیابی اطلاعات دوسطحی ساخته‌شده، یک مدل بازیابی اطلاعات به‌عنوان مدل پایه مورد نیاز است. بدین منظور در این پژوهش از مدل DPR استفاده شده است. DPR یک مدل بازیابی اطلاعات است که با کمک مدل‌های زبان قادر است اسناد مرتبط با یک پُرسه را بازیابی کند. DPR از دو رمزگذار BERT مستقل استفاده می‌کند، اما نیازی به پیش‌آموزش آن‌ها ندارد. در عوض، DPR بر یادگیری یک

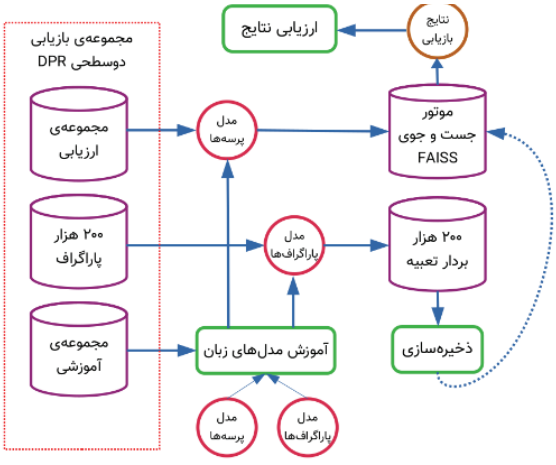
1. adaptive boosting
2. stochastic gradient descent (SGD)

مدل بازیابی اطلاعات قوی با استفاده از پرسش‌ها و پاسخ‌های دوتایی تمرکز می‌کند. DPR در واقع، راه‌هایی را برای انتخاب و استفاده از نمونه‌های منفی برای یک سؤال در نظر گرفته است. نمونه‌های منفی می‌توانند هرگونه سند تصادفی از مجموعه‌داده یا اسناد برتر بازگردانده‌شده توسط BM25 که حاوی پاسخ صحیح نیستند، باشد. از آنجا که مدل DPR پایه تنها دو نوع رابطه «مرتبط» و «نامرتبط» را می‌شناسد، ابتدا ساختار و نحوه آموزش مدل DPR بر روی مجموعه‌داده ارائه‌شده دوسطحی توضیح داده شده است. سپس، در بخش بعد، تغییرات مورد نیاز در DPR برای درک و فهم مجموعه‌داده ارائه‌شده چهارسطحی مورد بحث قرار گرفته است.

آموزش و ارزیابی مدل DPR پایه در سه مرحله انجام می‌شود:

- ◇ آموزش دادن رمزگذارهای دوگانه با کمک مجموعه‌داده آموزشی؛
- ◇ تعبیه کردن مجموعه اسناد با کمک رمزگذار مربوط به اسناد و ذخیره‌سازی مقادیر تعبیه‌شده با کمک ابزار FAISS (Johnson, Douze and Jégou 2019)؛
- ◇ بازیابی اطلاعات برای پُرسه‌های داخل مجموعه ارزیابی با کمک رمزگذار مربوط به پُرسه‌ها و FAISS و سپس، ارزیابی کردن کارایی سامانه بر اساس نتایج به‌دست‌آمده.

در شکل ۲، می‌توان جایگاه این سه مرحله را مشاهده کرد. در این شکل می‌توان دید که افزون بر مجموعه‌داده دوسطحی، دو نمونه از مدل زبان مورد نیاز است تا بتوان آن‌ها را آموزش داد و سپس از آن‌ها برای تعبیه کردن اسناد و پُرسه‌ها استفاده کرد.



شکل ۲. مراحل آموزش، تعبیه و ارزیابی در DPR

برای آموزش مدل بازیابی اطلاعات، یک مدل زبان فارسی و یا یک مدل زبان چندزبانه که از زبان فارسی نیز پشتیبانی کند، مورد نیاز است. در این پژوهش، از مدل BERT چندزبانه[1] استفاده شده است. همان‌طور که پیش‌تر نیز اشاره شد، جهت آموزش مدل، به دو نمونه از این مدل زبان نیاز است. یک نمونه برای تعبیه کردن اسناد (پاراگراف) که آن را با E_P نشان می‌دهیم و نمونه دیگر برای تعبیه کردن پُرسه‌ها که آن را با E_Q نشان می‌دهیم. در زمان آموزش، یک دسته از داده‌های آموزشی انتخاب می‌شوند و همه به‌طور موازی پردازش می‌شوند. در هر دسته، هر پُرسه شامل تعدادی سند مثبت و منفی است که در مجموعه داده تعیین شده‌اند. افزون بر اسناد منفی داخل مجموعه، اسناد مثبت پُرسه‌های دیگر نیز می‌توانند به‌عنوان سند منفی پُرسه جاری استفاده شوند (Karpukhin et al. 2020). پس از تعبیه کردن پُرسه‌ها و اسناد، با استفاده از ضرب داخلی بردارها، شباهت میان هر پُرسه و سند محاسبه می‌شود که آن را به‌صورت $Sim(q, p)$ نشان می‌دهیم که q یک پُرسه و p یک پاراگراف است و برای محاسبه آن از رابطه ۱، استفاده می‌شود.

$$sim(q, p)=E_Q(q)^T E_P(p) \quad (۱)$$

هرچه مقدار ضرب داخلی یک سند و پُرسه بیشتر باشد، ارتباط میان آن دو قوی‌تر است. برای نمونه، اگر فرض شود که برای آموزش مدل از دسته‌های سه‌تایی استفاده می‌شود و هر پُرسه به همراه یک سند مثبت و یک سند منفی باشد، با کمک رابطه ۱، می‌توان ماتریس رسم‌شده در جدول ۱، را به‌دست آورد. در این جدول، سند d_1 یک سند مثبت برای پُرسه q_1 است؛ اما همین سند برای پُرسه‌های q_2 و q_3 یک سند منفی محسوب می‌شود.

جدول ۱. ماتریس فرضی برای محاسبه شباهت سه پُرسه

	سند مثبت	سند منفی	سند مثبت	سند منفی	سند مثبت	سند منفی
	برای پُرسه ۱	برای پُرسه ۱	برای پُرسه ۲	برای پُرسه ۲	برای پُرسه ۳	برای پُرسه ۳
	d_1	d_2	d_3	d_4	d_5	d_6
q_1	$sim(q_1, d_1)$	$sim(q_1, d_2)$	$sim(q_1, d_3)$	$sim(q_1, d_4)$	$sim(q_1, d_5)$	$sim(q_1, d_6)$
q_2	$sim(q_2, d_1)$	$sim(q_2, d_2)$	$sim(q_2, d_3)$	$sim(q_2, d_4)$	$sim(q_2, d_5)$	$sim(q_2, d_6)$

1. BERT-base-multilingual-uncased

q_3	$sim(q_3, d_1)$	$sim(q_3, d_2)$	$sim(q_3, d_3)$	$sim(q_3, d_4)$	$sim(q_3, d_5)$	$sim(q_3, d_6)$

سپس با کمک ماتریس محاسبه‌شده، به ازای هر سطر از ماتریس، تابع زیان محاسبه می‌شوند تا مقادیر زیان را به‌دست آوریم. رابطه ۲، این تابع زیان را نشان می‌دهد که درست‌نمایی لگاریتمی منفی[1] که به اختصار NLL نام دارد (Karpukhin et al. 2020). در این رابطه، $(q_i, p_i^+, p_{i\,1}^-, \ldots, p_{i\,n}^-)$ نشان‌دهنده یک سطر از ماتریس با شاخص i است که شامل یک پُرسه (q_i)، یک سند مثبت (p_i^+) و چند سند منفی $(p_{i\,1}^-, \ldots, p_{i\,n}^-)$ است. بخش کسری این رابطه همواره در بازه $[0, 1]$ است و هرچه مقدار این کسر به ۱ نزدیک‌تر باشد، به این معناست که مدل پیش‌بینی دقیق‌تری انجام داده و شباهت پُرسه و سند مثبت را بالاتر ارزیابی کرده است. به‌طور مشابه، اگر مدل ما شباهت میان یک یا چند سند منفی با پُرسه را بالا محاسبه و ارزیابی کند، مقدار مخرج این کسر افزایش پیدا می‌کند و مقدار نهایی کسر به صفر نزدیک‌تر می‌شود. مقدار شباهت محاسبه‌شده توسط بخش کسری این رابطه، با کمک عملیات لگاریتمی و منفی‌سازی آن، به یک مقدار زیان تبدیل می‌شود. شایان ذکر است که خروجی این رابطه همواره در بازه $[0, \infty]$ است.

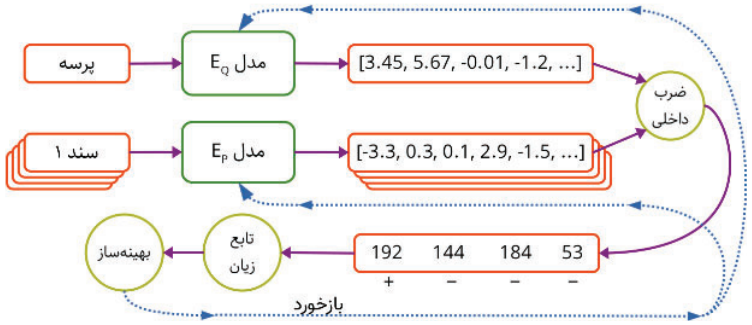
$$L\left(q_i, p_i^+, p_{i\,1}^-, \ldots, p_{i\,n}^-\right) = -Log \frac{e^{sim(q_i, p_i^+)}}{e^{sim(q_i, p_i^+)} + \sum_{i=1}^{n} e^{sim\left(q_i, p_{(i,j)}^-\right)}} \qquad (۲)$$

با در دست داشتن مقادیر زیان، می‌توان به رمزگذار E_P (مربوط به پاراگراف‌ها) و همین‌طور رمزگذار E_Q (مربوط به پُرسه‌ها) یک بازخورد از نحوه عملکرد آن‌ها داد. اگر مقدار تابع زیان نزدیک به صفر باشد، پارامترهای مدل کمتر دستخوش تغییرات می‌شوند و هرچه مقدار تابع زیان بیشتر شود، تغییرات در پارامترهای مدل بیشتر خواهد شد. اعمال این تغییرات تا زمانی که مدل قادر به پیش‌بینی صحیح شباهت میان پُرسه‌ها و اسناد مثبت شود، ادامه می‌یابد. با کمک این تابع زیان می‌توان مدل را تنظیم دقیق کرد و در طی این مرحله، مدل یاد می‌گیرد که برای اسناد و پُرسه‌هایی که از منظر بازیابی اطلاعات مرتبط هستند، بردارهای مشابه تولید کند تا میزان شباهت بردارهای آن‌ها بیشینه شود.

در شکل ۳، می‌توان فرایند کلی آموزش (تنظیم دقیق) را مشاهده کرد. در این شکل که در آن فرایند آموزش ساده‌سازی شده است، می‌توان دید که یک پُرسه، یک سند

1. negative log likelihood

مثبت (سند ۱) و سه سند منفی (اسناد ۲ تا ۴) به رمزگذارهای E_Q و E_P داده شده و به چندین بردار عددی تبدیل شده‌اند. سپس با کمک تابع شباهت ضرب داخلی، شباهت هر یک از بردارهای اسناد با بردار پُرسه محاسبه شده و یک لیست از این شباهت‌ها تولید می‌شود که عنصر اول آن مربوط به سند مثبت و بقیه عناصر مربوط به اسناد منفی است. تابع زیان با دانستن جایگاه سند مثبت می‌تواند عملکرد رمزگذارها را ارزیابی کند. پس از آن با کمک بهینه‌ساز، بازخوردی برای تنظیم پارامترهای دو مدل تولید شده و بدین ترتیب، مدل تنظیم دقیق می‌شود.



شکل ۳. فرایند آموزش DPR

پس از آنکه دو مدل زبان مورد استفاده در مرحله قبل آموزش دیدند، تمامی اسناد، تعبیه و شاخص‌گذاری می‌شوند. برای تعبیه کردن لازم است از مدل مربوط به اسناد E_P استفاده کرده و تمامی اسناد را به بردارهای معادل آن نگاشت کنیم. پس از تعبیه‌شدن اسناد، با کمک ابزار FAISS باید تمام بردارها ذخیره شوند تا بتوان بر روی آن‌ها جست‌وجو انجام داد. این ابزار قادر است پس از ذخیره‌سازی اسناد، یک بردار دریافت کرده و بردارهای شبیه به آن بردار را با سرعت بسیار بالا بازیابی کند. در این مرحله می‌توان با کمک FAISS و مدل زبان E_Q، بازیابی اطلاعات را انجام داد.

برای ارزیابی سامانه، از بخش ارزیابی مجموعه‌داده که شامل سه‌تایی‌های (پُرسه، پاسخ پُرسه، سند مرتبط) است، استفاده می‌شود. با تعبیه کردن هر پُرسه با کمک مدل E_Q یک بردار حاصل می‌شود و با وارد کردن این بردار به FAISS، تعداد مشخصی سند که بردارهای مشابه دارند، به‌دست می‌آید. امتیاز نهایی این سندها همان ضرب داخلی دو بردار است و اسناد بر اساس این مقدار مرتب می‌شوند. حال می‌توان با کمک

سنجه‌های ارزیابی، کیفیت بازیابی اطلاعات این سامانه را سنجید. سنجه‌های ارزیابی در بخش (۵-۱) معرفی شده‌اند.

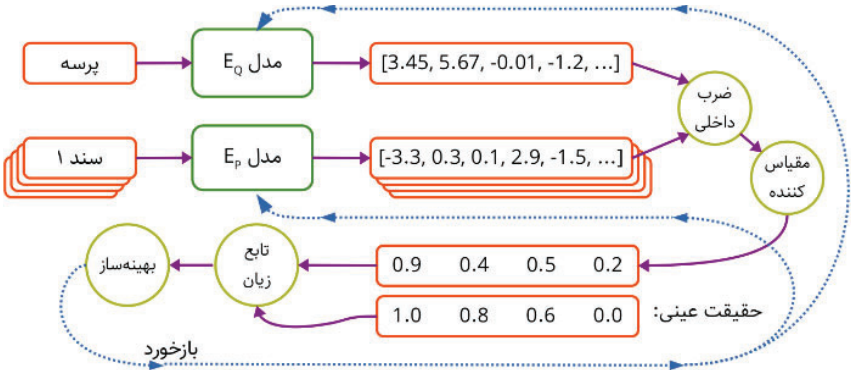
۴-۲. مدل DPR چندسطحی

به‌منظور استفاده از مجموعه‌داده بازیابی اطلاعات چندسطحی، نیاز به تغییر دادن مدل DPR است تا بتواند از سطوح تعریف‌شده در این مجموعه‌داده جدید استفاده کند. مهم‌ترین تغییری که بایستی در مدل DPR اعمال شود، بخش مربوط به تابع زیان است. این بخش باید قادر باشد تا به رتبه‌بندی مدل از اسناد چندسطحی، امتیاز دهد. در بخش ۴-۱، دیده شد که شباهت میان بردار اسناد و پُرسه با کمک ضرب داخلی انجام می‌شود و تابع زیان مدل DPR پایه نیز، تابع NLL بود. در این بخش نیز از ضرب داخلی برای محاسبه شباهت استفاده شده است. اما تابع زیانی که در این بخش استفاده شده، باید قادر باشد یک بردار از شباهت‌ها (به‌صورت نرمال‌شده) دریافت کند و به این بردار امتیاز دهد. بنابراین، برای مدل DPR چندسطحی، از تابع RankCosine (Xia et al. 2008) به‌عنوان تابع زیان استفاده شده است.

برای محاسبه تابع فوق لازم است ابتدا، شباهت‌های تولیدشده توسط ضرب داخلی به بازه ۰ تا ۱ نگاشت شود. برای این منظور از یک مقیاس‌کنندهٔ min-max استفاده شده است. پس از آن با کمک تابع زیان، عملکرد سامانه سنجیده می‌شود. در این بخش، تابع زیان افزون بر دریافت بردار شباهت که حاصل محاسبات دو مدل زبان و ضرب داخلی بردارهاست، یک بردار حقیقت عینی[1] نیز دریافت می‌کند که مقادیر آن به‌طور ضمنی توسط مجموعه‌داده تعیین می‌شود و در واقع، فاصله این دو بردار ورودی محاسبه می‌شود. بردار حقیقت با کمک میزان ارتباط اسناد با پُرسه حاصل می‌شود. برای نمونه، اگر سند اول در بردار مشابهت، یک سند *کاملاً مرتبط* (سطح ۳) باشد، عنصر نظیر آن در بردار حقیقت ۱ خواهد بود که معادل بالاترین مقدار ممکن برای امتیاز یک سند است. به‌طور مشابه، عنصر نظیر یک سند *نامرتبط* برابر ۰ خواهد بود که معادل پایین‌ترین مقدار ممکن برای امتیاز یک سند است. در شکل ۴، می‌توان جایگاه بردار حقیقت در مدل بازیابی اطلاعات چندسطحی را مشاهده کرد. مقادیری که به ازای هر سطح در نظر گرفته

1. ground truth

می‌شوند، یک پارامتر برای مدل محسوب می‌شود و از همین جهت برای یافتن مقدار مناسب، نیاز به تنظیم پارامتر است.



شکل ٤. فرایند آموزش DPR چندسطحی

برای محاسبه RankCosine ابتدا باید شباهت کسینوسی دو بردار ورودی محاسبه شود که این امر با کمک رابطه ۳، امکان‌پذیر است. در این رابطه، V_{Dot} بردار شباهت تولیدشده توسط ضرب داخلی است و هر عنصر از این بردار، شباهت میان پُرسه و یک سند را نشان می‌دهد. همان‌طور که پیش‌تر نیز اشاره شد، مقادیر این بردار به بازه ۰ الی ۱ مقیاس شده‌اند. بردار V_{Ground} همان بردار حقیقت است که هر عنصر آن، نشانگر سطح ارتباط سند با پُرسه است. مقادیر این بردار نیز در بازه ۰ الی ۱ هستند. مقدار محاسبه‌شده توسط رابطه ۳، همواره میان ۱- و ۱ است.

$$Recall_q = \frac{\sum_{(i,d)\in R_q} rel_q(d)}{\sum_{d\in D} rel_q(d)} \quad (۳)$$

برای محاسبه مقدار زیان از رابطه ۴، استفاده می‌شود (Xia et al. 2008). برای محاسبه این مقدار، ابتدا شباهت کسینوسی دو بردار محاسبه شده و سپس، مقدار این شباهت به مقدار زیان تبدیل می‌شود.

$$L(V_{Dot}, V_{Ground}) = \frac{1}{2}\left[1 - Cosine(V_{Dot}, V_{Ground})\right] \quad (۴)$$

با کمک این تابع زیان، می‌توان مدل را با کمک مجموعه‌داده چندسطحی آموزش داد. پس از فرایند آموزش، نحوه استفاده از مدل‌ها دقیقاً مانند یک مدل دوسطحی است. با کمک مدل‌های آموزش‌دیده می‌توان اسناد و پُرسه‌ها را تعبیه کرد، ضرب داخلی

بردارهای به‌دست‌آمده را محاسبه نمود، و اسناد را بر اساس امتیاز ضرب داخلی مرتب کرد. در واقع، در مرحله استفاده، معنا و مفهوم دوسطحی و چندسطحی بودن مدل‌ها از میان می‌رود و تنها رتبه‌بندی اسناد اهمیت خواهد داشت. به گفته دیگر، تفاوت مدل چندسطحی و دوسطحی تنها در بخش آموزش و مجموعه‌داده آموزشی است و در مراحل تعبیه کردن اسناد و ارزیابی، از عملیات و مجموعه‌های ارزیابی کاملاً یکسانی استفاده می‌شود. علت این امر آن است که مدل‌ها (چه دوسطحی و چه چندسطحی) پس از مرحله آموزش، رفتار یکسانی از خود نشان می‌دهند و با دریافت یک پُرسه و سند، یک عدد اعشاری به‌عنوان امتیاز تولید می‌کنند.

۵. ارزیابی

همان‌طور که پیش‌تر گفته شد، تفاوت مدل چندسطحی و دوسطحی، تنها در بخش آموزش و مجموعه‌داده آموزشی است و در مرحله ارزیابی، استفاده از مجموعه‌داده‌های دوسطحی و یا چندسطحی برای هر دو مدل امکان‌پذیر است و این امر هیچ ارتباطی با نحوه آموزش دوسطحی یا چندسطحی مدل‌ها نخواهد داشت. برای نمونه، می‌توان یک مدل را به‌صورت دوسطحی آموزش داد و در ارزیابی آن از یک مجموعه‌داده چندسطحی و یا دوسطحی استفاده کرد. در این پژوهش، برای داشتن یک ارزیابی صحیح و عادلانه از یک مجموعه ارزیابی دوسطحی برای ارزیابی هر دو مدل استفاده شده است تا عملکرد مدل‌ها در شرایط برابر و یکسان سنجیده شود.

۵-۱. سنجه‌های ارزیابی مدل بازیابی اطلاعات

برای سنجش کیفیت بازیابی انجام‌شده در بازیابی اطلاعات، سنجه‌های مختلفی قابل استفاده هستند. از رایج‌ترین این سنجه‌ها می‌توان به سنجه Recall اشاره کرد که در رابطه ۵، نشان داده شده است (Mitra and Craswell 2018). در این رابطه $(i, d) \in R_q$ نشان‌دهنده تمام اسناد بازیابی‌شده توسط سامانه بازیابی اطلاعات است. مقدار $d \in D$ نیز نشان‌دهنده هر سند مرتبط با پُرسه q است. مقادیر $rel_q(d)$ همواره ۱ یا ۰ بوده و نشان‌دهنده ارتباط داشتن یا نداشتن سند مورد بررسی d با پُرسه q است. به‌طور خلاصه، این رابطه نسبت تعداد اسناد مثبت بازیابی‌شده (صورت کسر) را به تعداد کل اسناد مثبت داخل مجموعه‌داده (مخرج کسر) می‌سنجد.

$$Recall_q = \frac{\sum_{(i,d)\in R_q} rel_q(d)}{\sum_{d\in D} rel_q(d)} \quad (۵)$$

از دیگر سنجه‌های مورد استفاده در بازیابی اطلاعات می‌توان به MRR[1] اشاره کرد. برای محاسبه این سنجه به ازای هر پُرسه و نتایج بازیابی‌شدهٔ آن به کمک مدل، مقدار RR[2] که در رابطه ۶، نشان داده شده، محاسبه می‌شود و سپس، میانگین این مقادیر RR محاسبه و به‌عنوان MRR در نظر گرفته می‌شود (Mitra and Craswell 2018). در این رابطه، i را می‌توان به‌عنوان رتبه هر سند بازیابی‌شده تفسیر کرد. به گفته دیگر، برای محاسبه RR هر سند، معکوس رتبه اولین سند مثبت محاسبه می‌شود و سپس با محاسبه میانگین برای این مقادیر RR، مقدار MRR محاسبه می‌شود.

$$RR_q = \max_{(i,d)\in R_q} \frac{red_q(d)}{i} \quad (۶)$$

سنجه دیگری که استفاده از آن در بازیابی اطلاعات رایج است، Precision است که بسیار شبیه به سنجه Recall بوده و در رابطه ۷، نشان داده شده است. در این رابطه $|R_q|$ نشان‌دهنده تعداد کل اسناد بازیابی‌شده توسط سامانه است. به گفته دیگر، این سنجه نسبت تعداد اسناد مثبت بازیابی‌شده به تعداد کل اسناد بازیابی‌شده را می‌سنجد.

$$Precision_q = \frac{\sum_{(i,d)\in R_q} rel_q(d)}{|R_q|} \quad (۷)$$

همچنین، استفاده از گونه تغییریافته‌ای از سنجه Precision که با نام Exact Match شناخته می‌شود در ارزیابی‌های سامانه‌های درک مطلب ماشینی بسیار رایج است. در این سامانه‌ها یافتن پاسخ دقیق یک پُرسه در سند، یک موفقیت محسوب می‌شود و از همین رو، تمامی اسنادی که حاوی پاسخ باشند، مثبت فرض می‌شوند. به گفته دیگر، برای محاسبه سنجه Exact Match کافی است نسبت اسناد بازیابی‌شده که حاوی پاسخ هستند، به کل اسناد بازیابی‌شده محاسبه شود.

در این پژوهش برای ارزیابی از سنجه‌های Recall و MRR استفاده شده است. همین‌طور در صورتی که پاسخ هر پُرسه در دسترس باشد، مقادیر Exact Match نیز محاسبه شده‌اند. لازم به ذکر است که تعداد اسناد بازیابی‌شده به ازای هر سنجه به‌صورت num@ و در

1. mean reciprocal rank
2. reciprocal rank

کنار نام سنجه نشان داده می‌شود. برای نمونه، اگر برای محاسبه MRR از ۱۰ سند و برای محاسبه Recall از ۱۰۰ سند استفاده شده باشد، آن‌ها را به‌ترتیب با MRR@10 و Recall@100 نمایش می‌دهند. همین‌طور شایان ذکر است که مقادیر به‌دست‌آمده از هر یک از این سنجه‌ها یک مقدار اعشاری بین ۰ و ۱ است. در این پژوهش، مقادیر به‌دست‌آمده از این سنجه‌ها در ۱۰۰ ضرب شده‌اند تا تفسیر کردن آن‌ها راحت‌تر و بر حسب درصد باشد.

۵-۲. نتایج آموزش و ارزیابی بر روی Trec-DL

مجموعه‌داده Trec-DL یک مجموعه‌داده چندسطحی انگلیسی است. در این پژوهش با کمک مجموعه‌داده‌های منتشرشده در سال‌های ۲۰۱۹ و ۲۰۲۰، دو مدل آموزش داده می‌شوند. مدل اول، یک مدل DPR دوسطحی است که تنها قادر به استفاده از سطوح ۰ و ۳ به‌عنوان اسناد مثبت و منفی است. مدل دوم، یک DPR چندسطحی است که می‌تواند بر خلاف مدل قبلی، افزون بر سطوح ۰ و ۳، از سطوح ۱ و ۲ نیز استفاده کند. در ادامهٔ این بخش، با کمک سنجه‌های Recall و MRR عملکرد این دو مدل مقایسه و بررسی می‌شود. از آنجا که پاسخ هر پُرسه در مجموعه‌داده Trec-DL وجود ندارد، نمی‌توان سنجه‌هایی چون Exact Match را برای ارزیابی استفاده کرد.

برای ارزیابی دقیق‌تر، فرایند آموزش هر یک از مدل‌های دوسطحی و چندسطحی سه مرتبه و با دانه‌های تصادفی[1] مختلف انجام شده است[2]. پس از آن میانگین این نتایج محاسبه و به‌عنوان نتیجه نهایی ارائه گردیده است. نتایج ارزیابی برای Recall و MRR بر روی مجموعه‌داده Trec-DL را می‌توان در جدول‌های ۲ و ۳، مشاهده نمود. همان‌طور که دیده می‌شود، مدل چندسطحی عملکرد بهتری نسبت به همتای دوسطحی خود دارد (۷/۰۴ درصد برای Recall@100 و ۴/۰۲ درصد برای MRR@100). این بهبود عملکرد نشان می‌دهد که تغییرات اعمال‌شده در مدل اصلی DPR و تبدیل آن به یک مدل چندسطحی، مؤثر واقع شده و مدل جدید چندسطحی DPR قادر بوده به‌درستی از اطلاعات اضافی فراهم‌شده (روابط سطح ۱ و ۲) در جهت افزایش دقت مدل بازیابی اطلاعات استفاده کند.

---

1. random seed

۲. دانه تصادفی عددی تصادفی است که برای مقداردهی پارامترهای مدل و نوع درهم‌سازی داده‌های آموزشی و ... به‌کار می‌رود و از همین رو، تغییر دادن آن می‌تواند بر نحوه آموزش‌دیدن مدل و نتایج به‌دست‌آمده از مدل در مرحله ارزیابی مؤثر باشد. از آنجا که آموزش دادن DPR یک فرایند زمان‌بر است، در این بخش برای ارزیابی، آموزش‌ها تنها با سه دانهٔ تصادفی مختلف انجام شده است.

حال که بر اساس نتایج به‌دست‌آمده، از عملکرد صحیح مدل چندسطحی اطمینان حاصل شد، در بخش بعدی به ارزیابی مجموعه‌داده فارسی و بررسی درستی مجموعه‌داده و نیز ارزیابی کارایی روش‌های تولید آن پرداخته می‌شود.

جدول ۲. مقادیر Recall برای مجموعه‌داده‌های دوسطحی و چندسطحی Trec-DL

مجموعه داده	Recall@1	Recall@10	Recall@20	Recall@100
دوسطحی – اجرای اول	۴/۲۶	۱۴/۲۵	۱۸/۸۶	۳۳/۵۸
دوسطحی – اجرای دوم	۴/۹۵	۱۶/۳۷	۲۱/۶۸	۳۷/۵۸
دوسطحی – اجرای سوم	۳/۲۰	۱۱/۲۵	۱۵/۳۱	۲۸/۶۶
چهارسطحی – اجرای اول	۳/۸۸	۱۳/۷۳	۱۸/۳۱	۳۱/۲۲
چهارسطحی – اجرای دوم	۸/۲۸	۲۲/۳۸	۲۷/۷۲	۴۳/۳۳
چهارسطحی – اجرای سوم	۹/۲۲	۲۴/۹۸	۳۱/۲۳	۴۶/۳۸
دوسطحی – میانگین	۴/۱۴	۱۳/۹۶	۱۸/۶۲	۳۳/۲۷
چهارسطحی – میانگین	۷/۱۳	۲۰/۳۶	۲۵/۷۵	۴۰/۳۱

جدول ۳. مقادیر MRR برای مجموعه‌داده‌های دوسطحی و چندسطحی Trec-DL

مجموعه داده	MRR@1	MRR@10	MRR@20	MRR@100
دوسطحی – اجرای اول	۴/۲۶	۶/۹۷	۷/۲۹	۷/۶۴
دوسطحی – اجرای دوم	۴/۹۵	۸/۰۴	۸/۴۰	۸/۷۸
دوسطحی – اجرای سوم	۳/۲۰	۵/۳۷	۵/۶۵	۵/۹۵
چهارسطحی – اجرای اول	۳/۸۸	۶/۵۵	۶/۸۷	۷/۱۷
چهارسطحی – اجرای دوم	۸/۲۸	۱۲/۱۷	۱۲/۵۵	۱۲/۹۱
چهارسطحی – اجرای سوم	۹/۲۲	۱۳/۵۸	۱۴/۰۱	۱۴/۳۷
دوسطحی – میانگین	۴/۱۴	۶/۷۹	۷/۱۱	۷/۴۶
چهارسطحی – میانگین	۷/۱۳	۱۰/۷۷	۱۱/۱۴	۱۱/۴۸

۵-۳. نتایج آموزش و ارزیابی بر روی مجموعه‌داده فارسی

در بخش ۳-۲، روش‌های مربوط به ساخت یک مجموعه‌داده چندسطحی جدید به کمک قوانین دست‌ساز و طبقه‌بندی خودکار مورد بحث قرار گرفت. در حالت طبقه‌بندی

خودکار، به‌منظور انتخاب بهترین طبقه‌بند جهت ساخت مجموعه‌داده چندسطحی جدید، تعدادی طبقه‌بند مرسوم آزموده شدند که دقت طبقه‌بندی آن‌ها در جدول ۴، ذکر شده است. همان‌طور که مشاهده می‌شود، از میان این طبقه‌بندها، جنگل تصادفی عملکرد بهتری ارائه کرده است. به همین دلیل، در این پژوهش نیز از همین طبقه‌بند جهت ساخت مجموعه‌داده چندسطحی فارسی استفاده شده است.

جدول ٤. دقت طبقه‌بندهای آزموده‌شده برای طبقه‌بندی اسناد

طبقه‌بند	F1 (micro)	F1 (macro)
جنگل تصادفی	۹۱/۱۰	۹۲/۸۸
درخت تصمیم	۷۸/۰۰	۸۲/۲۹
ماشین بردار پشتیبان	۵۰/۶۰	۵۹/۸۹
SGD	۴۱/۲۰	۴۴/۷۵
نزدیک‌ترین همسایه	۲۹/۶۰	۲۸/۳۴
AdaBoost	۴۰/۷۰	۴۵/۳۸
شبکه عصبی (۴ لایه مخفی)	۴۸/۶۰	۵۸/۴۳

به‌منظور ارزیابی تأثیرات مثبت مجموعه‌داده چندسطحی در مقابل مجموعه‌داده دوسطحی در سامانه‌های بازیابی اطلاعات، عملکرد هر دو مجموعه‌داده چندسطحی فارسی ساخته‌شده (یکی با قوانین دست‌ساز و یکی با طبقه‌بند جنگل تصادفی) با عملکرد مجموعه‌داده دوسطحی فارسی (به‌عنوان مدل پایه) مقایسه شده است. در جدول ۵، می‌توان مقادیر Exact Match@100 برای مدل دوسطحی پایه، مدل چندسطحی آموزش‌دیده بر روی مجموعه‌داده مبتنی بر قوانین دست‌ساز (با عنوان «چندسطحی قانون-محور») و مدل چندسطحی آموزش‌دیده بر روی مجموعه‌داده ایجادشده با کمک جنگل تصادفی (با عنوان «چندسطحی جنگل تصادفی») را مشاهده کرد. هر یک از این مقادیر نشان‌دهنده میانگین هفت اجرای مختلف با دانه‌های متفاوت هستند. همان‌طور که مشاهده می‌شود، هر دو مدل چندسطحی، عملکرد بهتری نسبت به همتای دوسطحی خود دارند. مدل چندسطحی قانون-محور بهبود ۱/۸۷ درصدی Exact Match را حاصل می‌کند و یا به تعبیر دیگر، مدل چندسطحی توانسته به ازای هر پُرسه، حدوداً ۲ سند بیشتر (که شامل پاسخ نیز است) نسبت به مدل دوسطحی بازیابی کند. مشاهده دیگری که در این جدول می‌توان داشت،

برتری روش قانون-محور نسبت به روش جنگل تصادفی است. در واقع، علت این امر در استفاده از ویژگی‌های مرتبط با پاسخ در این روش است.

جدول ۵. مقادیر Exact Match برای مجموعه‌داده‌های دوسطحی و چندسطحی فارسی

مجموعه داده	Exact Match @100
دوسطحی	۶/۲۱
چندسطحی قانون-محور	۸/۰۸
چندسطحی جنگل تصادفی	۷/۳۵

سرانجام، در جدول‌های ۶ و ۷، می‌توان به‌ترتیب ارزیابی Recall و MRR برای سه مجموعه‌داده را مشاهده کرد. در این جداول می‌توان عملکرد بهتر مجموعه‌داده‌ها و مدل‌های چندسطحی (به‌جز Recall@100 برای مدل چندسطحی جنگل تصادفی) را مشاهده کرد. از آنجا که در بخش ۵-۲، عملکرد صحیح مدل‌ها مورد بررسی قرار گرفت، می‌توان ادعا کرد که بهبود به‌دست‌آمده در این بخش مربوط به کارا بودن دو روش ارائه‌شده این پژوهش (روش مبتنی بر قانون و روش مبتنی بر طبقه‌بند، معرفی‌شده در بخش‌های ۳-۲-۱ و ۳-۲-۲) است. به بیان دیگر، با کمک این دو روش می‌توان به‌طور مؤثر یک مجموعه‌داده بازیابی اطلاعات دوسطحی را به یک مجموعه چندسطحی تبدیل کرد.

جدول ٦. مقادیر Recall برای مجموعه‌داده‌های دوسطحی و چندسطحی فارسی

مجموعه داده	Recall@1	Recall@10	Recall@20	Recall@100
دوسطحی	۱۴/۶۲	۳۹/۷۳	۴۸/۰۴	۶۷/۸۱
چندسطحی قانون-محور	۱۷/۹۷	۴۱/۵۸	۴۸/۶۰	۶۸/۶۲
چندسطحی جنگل تصادفی	۱۶/۶۳	۴۲/۵۸	۵۰/۱۵	۶۷/۵۰

جدول ۷. مقادیر MRR برای مجموعه‌داده‌های دوسطحی و چندسطحی فارسی

مجموعه داده	MRR@1	MRR@10	MRR@20	MRR@100
دوسطحی	۱۴/۶۲	۲۱/۹۶	۲۲/۵۳	۲۳/۰۲
چندسطحی قانون-محور	۱۷/۹۷	۲۴/۷۴	۲۵/۲۳	۲۵/۷۱
چندسطحی جنگل تصادفی	۱۶/۶۳	۲۴/۳۶	۲۴/۸۹	۲۵/۳۳

۶. نتیجه‌گیری

در مباحث پردازش زبان طبیعی، زبان فارسی یک زبان کم‌منبع است و نیازمندی‌های زیادی در آن وجود دارد. در چند سال اخیر و با پیشرفت مدل‌های زبان، سامانه‌های بازیابی اطلاعات زیادی توسط محققان بر پایه این ابزارهای جدید ارائه گردیده است. اما برای زبان فارسی، هیچ پژوهشی در این زمینه انجام نشده و در نتیجه، هیچ مجموعه‌داده مناسبی برای آن وجود ندارد.

در این پژوهش، پس از تبدیل یک مجموعه‌داده درک مطلب ماشینی به یک مجموعه‌داده بازیابی اطلاعات دوسطحی، دو روش مختلف برای تبدیل کردن مجموعه ساخته‌شده به یک مجموعه‌داده چندسطحی ارائه شد. سپس، مدل DPR به‌عنوان مدل پایه این پژوهش مورد استفاده قرار گرفت و تغییرات لازم در آن برای درک مجموعه‌داده‌های چندسطحی مورد بحث قرار گرفت.

با کمک دو مجموعه‌داده چندسطحی جدید و مجموعه‌داده دوسطحی و همین‌طور مدل‌های دوسطحی و چندسطحی می‌توان به ارزیابی راهکارهای ارائه‌شده پرداخت. ارزیابی‌های انجام‌شده بر روی این مجموعه‌ها نشان می‌دهد که در صورت فراهم بودن مجموعه‌داده چندسطحی، مدل بازیابی اطلاعات می‌تواند نتایج بهتری نسبت به مدل دوسطحی به‌دست دهد. همین‌طور با کمک روش‌های پیشنهادی در این پژوهش، می‌توان یک مجموعه دوسطحی را به یک مجموعه چندسطحی تبدیل کرد تا بتوان از مزایای مدل چندسطحی بهره برد.

References

Abadani, Negin, Jamshid Mozafari, Afsaneh Fatemi, Mohammd Ali Nematbakhsh, and Arefeh Kazemi. 2021. "*ParSQuAD: Machine Translated SQuAD Dataset for Persian Question Answering*." in 2021 7th International Conference on Web Research (ICWR). IEEE. Tehran, Iran.

Ayoubi Sajjad & Mohammad Yasin. Davoodeh. 2021. *PersianQA: A Dataset for Persian Question Answering*. GitHub Repository.

Bajaj, Payal, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, and Tri Nguyen. 2016. "*Ms Marco: A Human Generated Machine Reading Comprehension Dataset*." ArXiv Preprint ArXiv: 1611.09268.

Craswell, Nick, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021a. "*Overview of the TREC 2020 Deep Learning Track*." ArXiv Preprint ArXiv: 2102.07662.

_____, and Jimmy Lin. 2021b. "*Ms Marco: Benchmarking Ranking Models in the Large-Data Regime*." pp. 1566–76 in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Craswell, Nick, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. "*Overview of the TREC 2019 Deep Learning Track*." ArXiv Preprint ArXiv: 2003.07820.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "*BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*." pp. 4171–86 in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics.

Hashemi, Helia, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2020. "*ANTIQUE: A Non-Factoid Question Answering Benchmark*." Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12036 LNCS: 166–73. doi: 10.1007/978-3-030-45442-5_21.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7 (3): 535–547.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. "*Dense Passage Retrieval for Open-Domain Question Answering*." pp. 6769–81 in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Kazemi, Arefeh, Jamshid Mozafari, and Mohammad Ali Nematbakhsh. 2022. "*PersianQuAD: The Native Question Answering Dataset for the Persian Language*." IEEE Access 10: 26045–57. doi: 10.1109/ACCESS.2022.3157289.

Khashabi, Daniel, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, and Sarik Ghazarian. 2021. ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics* 9: 1163–1178.

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, and Kenton Lee. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7: 453–466.

Lin, Jimmy, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. "*Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations*." pp. 2356–62 in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Liu, Ye, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and S. Yu Philip. 2021a. "*Dense Hierarchical Retrieval for Open-Domain Question Answering*." pp. 188–200 in Findings of the Association for Computational Linguistics: EMNLP 2021.

Liu, Zhenghao, Kaitao Zhang, Chenyan Xiong, Zhiyuan Liu, and Maosong Sun. 2021b. "*OpenMatch: An Open Source Library for NEU-IR Research*." pp. 2531–35 in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Mitra, Bhaskar, and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends® in Information Retrieval* 13 (1): 1–126.

Qu, Yingqi, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. "*RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering*." pp. 5835–47 in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. "*Know What You Don't Know: Unanswerable Questions for SQuAD*." pp. 784–89 in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Texas, USA.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. "*SQuAD: 100,000+ Questions for Machine Comprehension of Text*." pp. 2383–92 in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Melbourne, Australia

Robertson, Stephen, and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3 (4): 333–389.

Salton, Gerard, and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* 24 (5): 513–523.

Xia, Fen, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. "*Listwise Approach to Learning to Rank: Theory and Algorithm*." pp. 1192–99 in Proceedings of the 25th international conference on Machine learning. Tokyo, Japan.

Yang, Peilin, Hui Fang, and Jimmy Lin. 2017. "*Anserini: Enabling the Use of Lucene for Information Retrieval Research*." pp. 1253–56 in Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. Toyo, Japan.

Zhang, Xinyu, Andrew Yates, and Jimmy Lin. 2020. "*A Little Bit is Worse than None: Ranking with Limited Training Data*." pp. 107–12 in Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing.

علی عابدزاده

متولد سال ۱۳۷۵، دارای مدرک تحصیلی کارشناسی ارشد در رشته مهندسی کامپیوتر از دانشگاه اصفهان است.

زمینه تحقیقاتی ایشان پردازش زبان طبیعی است و پایان‌نامه خود را با عنوان «ساخت مجموعه‌داده چندسطحی جهت آموزش مدل‌های زبان برای زبان کم‌منبع فارسی» به انجام رسانده است..

رضا رمضانی

متولد سال ۱۳۶۸، دارای مدرک تحصیلی دکتری تخصصی در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه فردوسی مشهد است. ایشان هم‌اکنون استادیار گروه مهندسی نرم‌افزار دانشکده مهندسی کامپیوتر دانشگاه اصفهان است.

پردازش زبان طبیعی، تحلیل داده و وب معنایی از جمله علایق پژوهشی وی است.

افسانه فاطمی

متولد ۱۳۵۲، دارای مدرک تحصیلی دکتری تخصصی در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه اصفهان است. ایشان هم‌اکنون دانشیار گروه مهندسی نرم‌افزار دانشکده مهندسی کامپیوتر دانشگاه اصفهان است.

سیستم‌های پیچیده، کلان‌داده، سیستم‌های پرسش‌وپاسخ و سیستم‌های گفت‌وگو از جمله علایق پژوهشی وی است..

پژوهش‌نامه
پردازش و
مدیریت
اطلاعات