

دادگان پرسش و پاسخ زبان فارسی



جواد فروتن راد^{۱*}، مریم حور علی^۲، محمدعلی کیوان راد^۳

^۱ کارشناس ارشد هوش مصنوعی، دانشگاه صنعتی مالک اشتر

^۲ و ^۳ استادیار هوش مصنوعی، مجتمع برق و کامپیوتر، دانشگاه صنعتی مالک اشتر

چکیده:

پاسخ سریع و دقیق به سؤالات مطرح شده به زبان طبیعی یکی از اهداف مهم در توسعه سامانه‌های پرسش و پاسخ است که در آن رایانه یک متن و سؤال را درک و پاسخ دقیق را برای کاربر ارائه می‌کند. با اینکه پیشرفت‌های زیادی در این حوزه صورت گرفته است، اما همچنان جزء مسائلی است که نیاز به ارتقا، به خصوص برای زبان‌های غیر انگلیسی مثل زبان فارسی است. در این مقاله دادگان پرسش و پاسخ زبان فارسی (FarsiQuAD)^۱ که توسط انسان از مقالات وبی پدیدار شده، در دو نسخه منتشر شده است. نسخه یک شامل ۱۰۰۰۰ پرسش و پاسخ و نسخه دوم این مجموعه شامل بیش از ۱۴۵۰۰۰ جفت پرسش و پاسخ است. این دادگان قابلیت تجمیع با نسخه انگلیسی SQuAD و سایر دادگان زبان‌های دیگر را دارد که از این استاندارد استفاده کرده باشند و برای عموم منتشر شده است^۲. این دادگان جهت ساخت مدل‌های هوش مصنوعی مبتنی بر یادگیری عمیق و برای استفاده در سامانه‌های پرسش و پاسخ زبان فارسی است. نتایج این پژوهش نشان می‌دهد دادگان پرسش و پاسخ زبان فارسی ایجاد شده می‌تواند پاسخ به سؤالات مطرح شده به زبان طبیعی فارسی را با معیار تطابق دقیق^۳ ۷۸ درصد و معیار F1^۴ ۸۷ درصد برساند که هنوز نیازمند ارتقا است.

واژگان: دادگان پرسش و پاسخ زبان فارسی، سیستم‌های پرسش و پاسخ^۵، درک مطلب^۵، یادگیری عمیق، پردازش زبان طبیعی

Farsi Question and Answer Dataset (FarsiQuAD)

Javad ForutanRad^{*1}, Maryam Hourali², Mohammad Ali kevanRad³

¹ M.Sc in AI, Malek-Ashtar university of technology,

^{2,3} Assistant professor in AI, electronic & computer Dep, Malek-Ashtar university of technology,

Abstract

A fast and accurate response to questions posed in natural language is a fundamental objective in the advancement of question and answer systems. These systems involve computers comprehending textual content and questions, and subsequently, delivering precise answers to users. Despite significant advancements in this field, there remains room for improvement, particularly when dealing with languages other than English, such as Persian.

In this article, we present the Persian language question and answer dataset, known as FarsiQuAD. This dataset was meticulously crafted by human annotators, drawing from Persian Wikipedia articles. FarsiQuAD is made available in two versions: Version 1 comprises over 10,000 questions and answers, while Version 2 offers an extensive collection of over 145,000 rows. This dataset is designed to seamlessly integrate with the English version of SQuAD and other databases in various languages adhering to this standard, and it is open to the public. These data serve as valuable resources for the development of

¹ Exact match

² <https://github.com/Forutanrad/FarsiQuAD>

³ Exact match

⁴ Question Answering

⁵ Reading comprehension

* Corresponding author

* نویسنده عهده‌دار مکاتبات

artificial intelligence models based on deep learning and for the enhancement of Persian language question and answer systems.

The research findings reveal that the FarsiQuAD dataset is capable of providing answers to questions posed in the natural Persian language with an exact matching accuracy of 78% and an F1 score of 87%. However, there is still room for improvement in achieving even higher accuracy levels.

This project arises from the critical need for non-English languages to have access to more data for training deep learning models, especially in the domain of factoid questions. Hence, the primary objective of this article is to introduce the newly created dataset. Prior to this effort, well-known datasets like SQuAD predominantly focused on English, and similar datasets has been developed in other languages, including French, German, Korean, and Japanese. Nevertheless, the dearth of question datasets in the Persian language was evident. The quality and diversity of questions are pivotal aspects, and as this dataset continues to grow, it will contribute to the broader landscape of research in this domain, allowing for valuable cross-linguistic comparisons and integration with research conducted in other languages.

Keywords: Question And Answer Dataset, Question And Answer systems, Reading omprehension, Deep Learning, Natural Language Processing, Factoid Questions

۱- مقدمه

پرسش و پاسخ یکی از مهم‌ترین زمینه‌های پژوهشی در بازیابی اطلاعات است و ترکیبی از دامنه‌های مختلف پژوهشی همچون پردازش زبان طبیعی، هوش مصنوعی، بازیابی اطلاعات و استخراج اطلاعات است که به‌طور خودکار به پرسش‌های زبان طبیعی با استفاده از منابعی که در پایگاه دانش ذخیره شده است، پاسخ می‌دهد. پاسخ‌گویی به سؤالات یکی از حوزه‌های بسیار پرکاربرد نیز هست که در بازیابی اطلاعات، سامانه‌های تعاملی، ربات‌ها، سامانه‌های کنترلی، نرم‌افزارهای گفتگو و هوش مکالمه‌ای کاربرد گسترده دارد. در این سامانه‌ها بسته به نوع سؤال، اطلاعات از پایگاه دانش استخراج و در اختیار قرار می‌گیرد و این موضوع باعث کاهش چشم‌گیر زمان دریافت پاسخ سؤالات و خودکارسدن فرایندهای مرتبط می‌شود.

با پیشرفت چشم‌گیر یادگیری عمیق در سالیان اخیر و افزایش دقت مدل‌های مبتنی بر این روش، امروزه بیشتر کارهای پرچالش و سخت پردازش زبان طبیعی از این رویکرد استفاده می‌کنند و شرکت‌های بزرگ دنیا مثل گوگل و فیس‌بوک نسبت به جایگزینی روش‌های سنتی با یادگیری عمیق به دلیل نتایج فوق‌العاده آن اقدام کرده‌اند. با اینکه کارهای زیادی به‌منظور توسعه مدل‌های پرسش و پاسخ برای سایر زبان‌های روز دنیا توسعه یافته‌اند، اما کمبود جدی برای زبان فارسی احساس می‌شود تا مشابه مدل‌های زبان انگلیسی توسعه یابند.

استفاده از یادگیری عمیق در مسائلی مانند زبان طبیعی که پیچیدگی‌های زیادی دارد، جزو پرکاربردترین و پیشرو این حوزه است. مهم‌ترین برتری شبکه‌های عصبی یادگیری عمیق این است که مهندسی ویژگی جداگانه را

از بین می‌برند؛ لذا نیاز به برچسب‌گذاری، استخراج ویژگی و نیاز به افراد خبره زبان‌شناس را که جزوی از محدودیت‌های پروژه‌ها است، مرتفع می‌سازند و این کار توسط خود شبکه انجام می‌شود. دومین دلیل، تا حدود زیاد، مرتفع کردن محدودیت‌های معنایی است؛ با توجه به اینکه مدل‌ها به‌طور معمول با چندین زبان مختلف آموزش می‌بینند و مفهوم کلمات و جملات در مدل قرار می‌گیرد؛ در نتیجه نتایج قابل قبولی را ارائه می‌کنند و برتری سوم سرعت مدل‌ها است که نسبت به مدل‌های و روش‌های سنتی سرعت اجرای به‌مراتب بالاتری دارند؛ اما مشکل اصلی در مدل‌های یادگیری عمیق، نیازمندی مدل‌ها به داده‌های اولیه خیلی زیاد است.

۲- تاریخچه

۲-۱- کارهای انجام‌شده در زبان انگلیسی

برای زبان انگلیسی کامل‌ترین دادگان منتشرشده دادگان SQuAD به زبان انگلیسی در دو نسخه با شماره ۱.۰ [۱] و ۲.۰ [۲] است که توسط دانشگاه استنفورد منتشر شده است.

مجموعه داده پاسخ به سؤالات استنفورد (SQuAD)، مجموعه داده‌های درک مطلب بیش از صد هزار سؤال است که توسط اعضای گروه از مقالات ویکی‌پدیا تهیه شد و در آن پاسخ به هر سؤال، بخشی از متن مربوطه است [۱].

این مجموعه برای کمک به چالش‌های پرسش و پاسخ در NLP است. این دادگان بر وظیفه پاسخ به سؤالات زبان طبیعی تمرکز دارد. در شکل (۱) نمونه‌ای از جفت پرسش و پاسخی که توسط سازندگان ارائه شده گذاشته شده است.

در سایر سؤالات مبتنی بر سند که به مجموعه داده‌ها پاسخ می‌دهند و بر استخراج پاسخ تمرکز می‌کنند، پاسخ به یک سؤال داده شده در چندین سند وجود دارد؛ با این حال، در SQuAD، این مدل فقط به یک گذرگاه واحد دسترسی دارد و کار بسیار دشواری را ارائه می‌دهد.

یک نوع مجموعه داده محبوب، مجموعه داده cloze است که از یک مدل می‌خواهد کلمه‌ای را که در یک قسمت وجود ندارد، پیش‌بینی کند. این مجموعه داده‌ها بزرگ هستند و وظیفه‌ای تا حدودی مشابه SQuAD را ارائه می‌دهند. پیشرفت کلیدی که SQuAD در این زمینه انجام می‌دهد این است که پاسخ‌های آن پیچیده‌تر هستند؛ و بنابراین نیاز به استدلال بیشتری دارند؛ بنابراین SQuAD را برای ارزیابی درک و قابلیت مدل بهتر می‌کند. در نسخه ۲.۰ این دادگان [۲] تعداد زیادی سؤال از نوع نداشتن پاسخ افزوده شده است. در این نوع سؤال‌ها سامانه می‌تواند متوجه شود که پاسخ را نمی‌داند. نسخه ۲.۰ دادگان بیش از ۱۵۰۰۰۰ جفت پرسش و پاسخ دارد. [۲].

۲-۲- بررسی دادگان پرسش و پاسخ به سایر

زبان‌ها غیر از انگلیسی

باتوجه به اینکه شبکه‌های آموزش دیده شده با مجموعه سؤالات به زبان مقصد نتایج بهتری نسبت به مدل‌های چندزبانه داشته‌اند در زبان‌های مختلف این مجموعه سؤالات بر مبنای سؤالات دادگان SQuAD تهیه شده است.

برای زبان فرانسوی دادگان FQuAD [3] تهیه شده که دارای بیش از ۶۰,۰۰۰ جفت پرسش و پاسخ است. برای زبان آلمانی دادگان GermanQuAD [4] ارائه شده که شامل 13,722 جفت پرسش و پاسخ است. برای زبان کره‌ای دو کار [5,6] انجام شده است و برای زبان ژاپنی دادگان JaQuAD [7] که مشتمل بر 39,696 هزار جفت پرسش و پاسخ است. در زبان فارسی دادگان PersianQA [8] منتشر شده است که ده‌هزار جفت پرسش و پاسخ دارد که به نسبت سایر زبان‌ها تعداد کمتری است.

۳- مدل پیشنهادی

یکی از مهم‌ترین قسمت‌های شبکه یادگیری عمیق نیازمند بودن به داده‌های زیاد اولیه است که گاهی اوقات

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupe**l and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

(شکل ۱) - نمونه جفت پرسش و پاسخ ارائه شده در

ویکی‌پدیا [۱]

(Figure -1) - A sample question-and-answer pair provided on Wikipedia

در نسخه ۱ تعداد ۵۳۶ مقاله را از بین ۱۰۰۰۰ مقاله ویکی‌پدیا نمونه‌برداری کرده‌اند که در مجموع ۲۳۲۱۵ پاراگراف جداگانه استخراج شده است. در این پاراگراف‌ها مواردی که کمتر از پانصد نویسه و بیشتر از ۱۵۰۰ نویسه داشته‌اند، حذف شده‌اند. در این دادگان هشتاد درصد مقالات را به مجموعه آموزشی، ده درصد را به مجموعه آموزشی و ده درصد را به مجموعه آزمون داده‌ها اختصاص داده‌اند.

در این مجموعه از استخراج‌کنندگان سؤال خواسته شده است که از هر مقاله پنج سؤال و یا حداقل سه سؤال استخراج کند، به‌طوری‌که حتماً جواب سؤال در متن مقاله موجود بوده و از داخل پاراگراف مقاله متن جواب سؤال را برجسته کرده‌اند. برای این منظور نرم‌افزاری در اختیار استخراج‌کنندگان سؤالات گذاشته شده است که می‌توانستند سؤالات را بدون رونوشت از متن، وارد و سپس پاسخ را از روی متن انتخاب کنند.

Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on "Select Answer", and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using the same words/phrases as in the paragraph. Also, you are encouraged to use hard questions.**

Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

(شکل ۲) - نمونه روش مورد استفاده برای جمع‌آوری سؤالات

دادگان SQuAD [۱]

(Figure - 2) - The method used for collecting questions in the SQuAD dataset[1]

به عنوان محدودیت این شبکه ها نیز به دلیلی سختی تهیه این داده ها نیز محسوب می شود. سختی ساخت دادگان باعث شده است که افراد و شرکت های محدودی در این زمینه فعالیت کرده و یا بعضاً نتایج و دادگان خودشان را در اختیار سایرین قرار ندهند.

در این کار تلاش شده است با الگوبرداری از نمونه زبان انگلیسی دادگان دانشگاه استنفورد (SQuAD) که برای پاسخ گویی به پرسش و پاسخ از متون به صورت حقیقت نما و به زبان انگلیسی تهیه شده، برای زبان فارسی نیز توسعه یابد.

۳-۱- بررسی ساختارها و داده های خام اولیه

به منظور تهیه مجموعه نیاز بود مقالات پایه اولیه جهت استخراج سؤال ها و پاسخ ها شناسایی و انتخاب شوند.

مقالات ویکی پدیا [۱۷] این دادگان شامل تمام مقالات خام ویکی پدیای فارسی تا تاریخ ۱۲ مرداد ۱۳۹۹ است این دادگان به صورت ۹ عدد فایل های متنی حجیم ارائه شده است. این دادگان شامل ۷۳۹۸۷۰ مقاله، ۴۰۰۴۷۶۵ جمله و ۹۴۰۰۲۰۹۴ کلمه است.

نظر به بررسی های صورت گرفته در مجموعه های مختلف دادگان های بررسی شده، به دلیل اینکه دادگان ویکی پدیا [۱۷] دارای مقاله هایی در حوزه های مختلف، به روزتر از سایر مجموعه داده ها بود، لذا برای استفاده در سامانه پرسش و پاسخ مناسب تر تشخیص داده شد.

گفتنی است دادگان SQuAD نیز از مقالات ویکی پدیا برای استخراج سؤالات استفاده کرده است که این موضوع باعث اعتبار بیشتر به این مجموعه و انتخاب آن شد.

در این کار سعی شده است که از مقالات عمومی ویکی پدیا استفاده شود؛ لذا از میان ۶۳۲۶۴ به صورت تصادفی و به انتخاب افراد خبره استخراج کننده سؤالات، تعداد ۲۷۰۵ مقاله مورد استفاده قرار گرفته که در سی گروه مختلف اعلام شده در مقاله بوده که تنوع زیادی را شامل شده است.

۳-۱-۱- آماده سازی داده ها و مرتب کردن

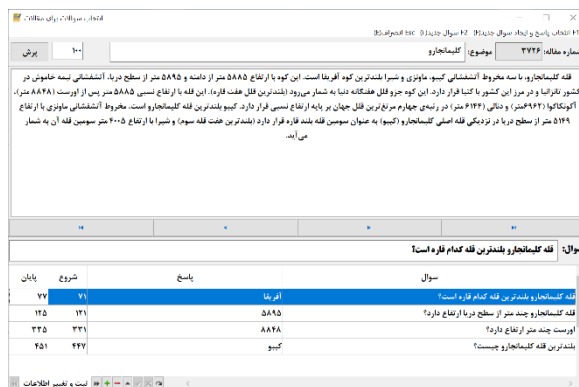
باتوجه به اینکه داده های دادگان ویکی پدیا به صورت فایل های متنی خام سنگین ارائه شده بود، لذا برای آماده سازی جهت استفاده با استفاده از کد زبان برنامه نویسی پایتون هر ۹ فایل متنی دادگان پردازش و

مقالاتی که تعداد نویسه های آن بین ۵۰۰ الی ۱۱۰۰ نویسه بودند، استخراج شدند و سپس با استفاده از کتابخانه هضم عملیات، پیش پردازش مربوط به تمیز و مرتب کردن و نرمال سازی داده صورت گرفت. در این مرحله حرف "ه" که اشکالاتی داشت، جایگزین شد.

۳-۱-۲- ساخت برنامه جهت تسهیل استخراج سؤالات

به منظور جمع آوری و استخراج سؤالات، و همچنین هماهنگی، نیاز به یک واسط کاربری قوی، سریع و کاربر پسند بود تا بتوان به سادگی سؤالات مورد نیاز را استخراج کرد؛ لذا برای این منظور نرم افزاری به شرح زیر ایجاد شد. نام این نرم افزار QAProject (انتخاب سؤالات برای مقالات) نام گذاری شد.

این برنامه با اتصال به بانک اطلاعاتی مقالات را واکنشی و اطلاعات شماره مقاله، موضوع و متن را نمایش داده و کاربر می تواند در قسمت سؤال، متن سؤال را که جواب آن در متن است، وارد کرده، سپس قسمت مرتبط به پاسخ را از متن با انتخاب قسمت درست پاسخ مشخص کرده و با فشردن کلید F1 قسمت انتخاب شده از متن به عنوان پاسخ انتخاب شده و به همراه نویسه شروع و پایان در بانک اطلاعاتی ذخیره شده و یک ردیف جدید برای سؤال بعدی ایجاد می شود.



(شکل - ۳). نمایی از برنامه ساخته شده برای

استخراج سؤالات

(Figure-3). Representation of the program created for question extraction

همان طور که در شکل (۵) نشان داده شده است، مقالات در بالای صفحه و قسمت سؤالات و پاسخ ها در پایین صفحه قرار گرفته است. در قسمت سؤالات امکان ویرایش، حذف، درج جدید و به روزرسانی و در قسمت سمت چپ بالای نرم افزار امکان پرسش به تعداد زیادی مقاله افزوده شده است و در مثال بالا با هر بار زدن روی دکمه پرسش،

نرم افزار صد مقاله جلوتر می رود. در قسمت وسط صفحه امکان رفتن به مقاله بعدی، قبلی، آخرین و اولین نیز گذاشته شده است.

این نرم افزار به همراه تقسیم مساوی از تعداد کل مقالات با کنترل اینکه از قبل در اختیار شخص دیگری نبوده و تضمین عدم دریافت مقاله تکراری، در اختیار کارشناسان استخراج کننده سؤال که در این پروژه هفت نفر بودند، قرار داده است، شخص بعد از مطالعه کامل متن مقاله، در قسمت سؤال، متن سؤالی را که در مقاله جواب آن موجود است، وارد کرده و سپس برای جواب آن نیز باید قسمتی از متن مقاله در اختیار را با موشواره انتخاب و سپس تأیید کند تا به عنوان جواب سؤال درج شود و امکان درج متن تاییپی برای پاسخ مقدور نبوده است. گفتنی است هم زمان با انتخاب جواب، نویسه شروع و پایان جواب از مقاله نیز استخراج و ذخیره شده است. ضریب کاپا ۰.۹۸۳ درصد بوده است.

۳-۱-۳- تهیه سؤالات

به منظور در ادامه نسبت به تهیه پرسش و پاسخ به زبان فارسی با استفاده از نرم افزار ایجاد شده اقدام شد. به منظور استخراج سؤالات، تیمی هفت نفره از افراد خبره با مدرک تحصیلی کارشناسی ارشد فعالیت داشته اند.

• نمونه سؤالات استخراج شده

در ادامه نمونه ای از متن مقاله و سؤال های پرسیده شده به همراه پاسخ استخراج شده به نمایش گذاشته شده است.
متن مقاله: " یغوارد (به ارمنی: Եղվարդ) یک شهر در غرب ارمنستان است که در استان کوتایک واقع شده است. در کیلومتری شمال هرازدان، مرکز استان کوتایک واقع شده است. کلمه یغوارد از دو کلمه ارمنی یغی () به معنی (عطر یا بو) و وارد () به معنی (گل رز) مشتق شده است. به عقیده آرام غانالانیان پژوهشگر و ارمنی شناس نام گذاری یغوارد به دلیل وجود یک جنگل وسیع در این منطقه است که از انواع گل رز و سایر گل های با عطر زیاد پوشیده شده است یغوارد از کهن ترین اقامتگاه های بشری در منطقه است و در آن آثار معماری زیادی یافت می شود. یغوارد ۷ کیلومتر مربع مساحت دارد. یغوارد در جنوب کوه آرابی در ارتفاع ۱۳۳۳ متری از سطح دریا واقع شده است. این شهر در ۱۵ کیلومتری ایروان پایتخت جمهوری ارمنستان قرار گرفته است

کلیسای یغوارد بنا شده در ۱۳۰۱ تنها بنای حفظ شده تاریخی در یغوارد است. اف سی یغوارد یک باشگاه فوتبال بود و نماینده این شهرک در بین سال های ۱۹۸۶ تا ۱۹۹۶ بود و به دلایل مشکلات مالی منحل شد.

سؤال ۱: یغوارد در کدام کشور قرار دارد؟ ارمنستان

سؤال ۲: شهر یغوارد در کدام قسمت از ارمنستان واقع است؟ غرب

سؤال ۳: شهر یغوارد در کدام استان قرار دارد؟ پاسخ: کوتایک

سؤال ۴: یغوارد چقدر مساحت دارد؟ پاسخ: ۷ کیلومتر مربع

سؤال ۵: شهر یغوارد در چه فاصله ای از ایروان واقع است؟ پاسخ: ۱۵ کیلومتری

همان طور که در سؤالات استخراج شده از این مقاله مشخص است در قسمت مساحت چون عنوان نشده بود که به چه واحدی نمایش داده شود پس کلمات بعد از عدد هم جزوی از سؤال است. مثلاً در سؤال اگر گفته شده بود نتیجه سؤال مساحت چند کیلومترمربع است در جواب فقط عدد نمایش داده خواهد شد.

• خلاصه آماری

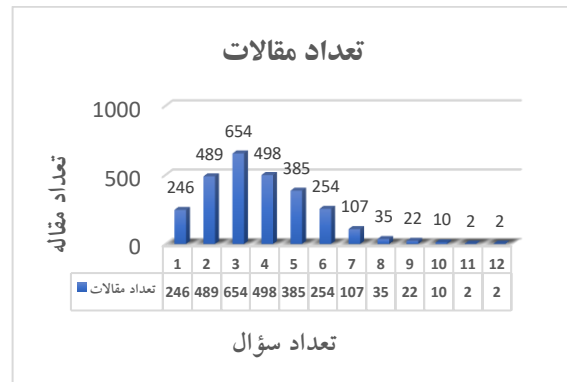
تعداد کل مقالات ۶۳۲۶۴ مقاله بوده است که از این بین، تعداد ۲۷۰۵ مقاله به عنوان مقاله منتخب از حوزه های مختلف موجود انتخاب و سؤالات از این مقالات استخراج شده اند.

تعداد کل سؤالات استخراج شده بیش از ده هزار جفت پرسش و پاسخ است که با درک مطلب و به صورت مفهومی از متن استخراج شده است. در این مجموعه سؤالات دقت شده است که جواب سؤال باید در متن قابل استخراج باشد و برای سؤالاتی که سامانه باید بداند که جواب آن را نمی داند نیز ۱۵۳ سؤال در نسخه ۱ ثبت شده است؛ لذا تعداد سؤالات حاوی پاسخ ۹۸۴۷ جفت هستند.

در این دادگان کمترین تعداد سؤال استخراج شده از مقاله ها عدد ۱، بیشترین تعداد سؤال استخراج شده از مقاله ۱۲ و میانگین سؤال استخراج شده از هر مقاله ۳.۷ است.

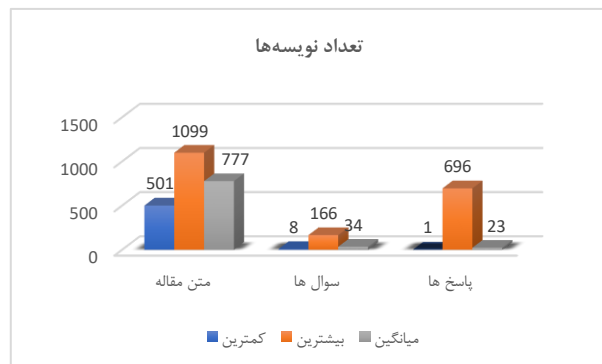
در شکل (۶) فراوانی مقالات بر اساس تعداد سؤال هایی که از هر مقاله استخراج شده، نمایش داده شده است. همان طور که در شکل مشخص است بیشترین استخراج سؤال مربوط به استخراج ۳ سؤال از

مقاله‌ها بوده که ۶۵۴ مقاله را شامل می‌شود و رتبه دوم و سوم به ترتیب در اختیار ۴ و ۲ سؤال از مقاله‌ها قرار دارد. برای استخراج سؤالات این دادگان به صورت متوسط می‌توان ۴۰ تا ۵۰ سؤال را در یک ساعت استخراج کرد که با توجه به تعداد ده‌هزار سؤال استخراج‌شده برابر با ۲۵۰ ساعت زمان صرف استخراج سؤالات از متون مقاله‌ها شده‌است.



(شکل - ۴). فراوانی مقالات در تعداد سؤال استخراج شده
(Figure-4) - Frequency of articles in the number of extracted questions

در شکل (۷) تعداد نویسه‌های مقالات، سؤالات استخراج شده و پاسخ‌ها به صورت گزارش کامل ارائه شده‌است. در این نمودار مشخص می‌شود که مقالات به صورت متوسط ۷۷۷ نویسه داشته‌اند، پاسخ‌ها به صورت میانگین ۲۳ نویسه و سؤال‌های مطرح‌شده نیز به صورت میانگین دارای ۳۴ نویسه بوده‌اند.



(شکل - ۵). نمودار تعداد نویسه‌های مقالات،

سؤالات و پاسخ‌ها

(Figure-5). Chart of the number of characters of articles, questions and answers

موضوع مقالات استفاده شده در تهیه دادگان

موضوعات استفاده شده دارای طیف گوناگونی از موضوعات است که در ذیل تعدادی از این موارد مطرح شده‌است:

اماکن و موزه‌ها، حروف الفبا، اشخاص و شخصیت‌ها، ادبیات، کتاب و رمان، موسیقی، آلات موسیقی و نوازنده‌ها، وسایل جنگی، ریاضیات، جغرافیای کشورها، استان‌ها، شهرستان‌ها، شهرها و روستاها، تاریخی، عناصر شیمیایی، برق، الکترونیک و رایانه، اقتصاد و یکای پولی، فیلم و سینما، قرآن، احزاب، حیوانات، گل، درختان و گیاهان، ورزشی و ورزش کاران، بزرگراه‌ها، آسمان، ستاره، کهکشان و صورت‌های فلکی‌ها

▪ سؤالاتی که باید بدانند که جواب آن را نمی‌داند

خیلی مهم است که سامانه‌ها بدانند، جواب بعضی از سؤالات را با این که تمام یا بخش زیادی از کلمات در متن وجود دارند، نمی‌دانند. به عنوان مثال در متن آمده است "جمعیت شهر بیرجند ۲۱۰۰۰۰ هزار نفر است." در این متن می‌توان به سؤال "جمعیت بیرجند چند نفر است؟" پاسخ دهد؛ اما به سؤال "شهر بیرجند چند نفر مرد جمعیت دارد؟" نمی‌تواند پاسخ دهد چون همچنین چیزی در متن عنوان نشده‌است و فقط کلمه "مرد" در متن سؤال باعث می‌شود که دیگر امکان پاسخ‌گویی به سؤال وجود نداشته باشد.

این حوزه شامل بخش زیادی از سؤالات مختلف می‌شود که انسان این موضوع را می‌داند البته به سختی قابل تشخیص است و یکی از چالش‌های شبکه‌های هوش مصنوعی در پاسخ‌گویی به سؤالات هستند.

▪ مواردی که در تهیه سؤالات رعایت شده‌است:

- در انتخاب پاسخ تلاش شده حداقل‌ترین پاسخ ممکن انتخاب شود.
- در تهیه پاسخ‌هایی که با پیشوندهایی مثل خانم، آقای، حاجی، مهندس، ژنرال و... بوده‌اند به عنوان جزوی از جواب انتخاب شده‌است.
- پیشوندهایی مثل سال، شهر و ... در صورتی که در متن سؤال قرار داشتند در جواب در نظر گرفته نشده‌اند.
- پسوندهایی مانند متر، میلادی، کیلومتر، میلیون، قمری، خورشیدی، میلی‌متر، و... نیز جزوی از جواب در نظر گرفته شده البته در صورتی که در خود متن سؤال این کلمات وجود نداشته‌اند.
- کلماتی که بیان‌گر تقریبی هستند، مثل حدوداً، تقریباً، بیش از نزدیک به و... اگر در صورت

```

{
  "data": [
    {
      "title": "کمال خدیجی",
      "paragraphs": [
        {
          "qas": [
            {
              "answers": [
                {
                  "answer_start": 55,
                  "answer_end": 97,
                  "text": "از عارفان و شاعران پارسی گوی قرن هفتم هجری"
                }
              ],
              "question": "کمال خدیجی که بود؟",
              "is_impossible": false,
              "id": "b9da9a9a-af5d-fefc-aatv-tterf6d-f6d"
            }
          ],
          "answers": [
            {
              "answer_start": 112,
              "answer_end": 122,
              "text": "محمد فرارود"
            }
          ],
          "question": "کمال خدیجی در کجا به دنیا آمد؟",
          "is_impossible": false,
          "id": "dfcf7fab-ac11-f92d-b0da-0c-07e7b7b6f"
        }
      ],
      "context": "آرمگاه او در تبریز است. این بیت بر لوح آرمگاه او ثبت شده است: میوان کمال خدیجی در ونگاه گنجور"
    }
  ],
  "id": 1
}

```

(شکل-۶). نمونه ساختار فایل دادگان FarsiQuAD
(Figure-6). Sample structure of the FarsiQuAD dataset file

۳-۱-۵- نسخه شماره دو برای دادگان

با بررسی‌هایی که در مورد نسخه یک دادگان انجام شده نتایج حاصل برای یک مقاله قابل قبول بود، اما وقتی جستجو در تعداد زیادی از مقالات انجام می‌شد، باعث شده بود نتایج حاصل به دلیل اینکه قبلاً این متون برای سؤال دیده نشده بودند، به عنوان پاسخ نمایش داده شوند؛ لذا برای غلبه بر این مشکل دادگان نسخه شماره ۲ (FarsiQuAD_V2) تهیه شد. در این دادگان به صورت تصادفی پنجاه سؤال را برای هر مقاله که مربوط به این مقاله نیستند انتخاب، و به عنوان سؤالاتی که باید مدل‌ها بدانند جواب آن را نمی‌دانند (جواب غیرممکن) به دادگان افزوده شد. با انجام این کار تعداد ردیف‌های دادگان به بیش از ۱۴۵۰۰۰ سؤال رسید.

۳-۱-۵- مقایسه دو دادگان پرسش و پاسخ زبان فارسی

در این قسمت به شباهت‌ها و تفاوت‌های دو دادگان پرسش و پاسخ زبان فارسی FarsiQuAD و PersianQA پرداخته می‌شود.

- شباهت‌ها در دو دادگان به شرح زیر است:
- بر اساس درک مطلب تهیه شده‌اند.
- بر اساس ساختار دادگان معروف SQuAD تهیه شده‌اند و قابلیت ادغام در نسخه اصلی را دارند.
- به زبان فارسی هستند.
- دارای سؤال‌های فاقد جواب هستند.
- منبع مقالات جهت استخراج پاسخ‌ها ویکی‌پدیای فارسی است.
- تعداد جفت سؤال و پاسخ‌ها تقریباً یک اندازه هستند.

سؤال عنوان نشده‌اند در جواب به عنوان بخشی از پاسخ در نظر گرفته شده‌اند.

○ انعکاس کامل متنی که در نقل قول استفاده شده است.

○ در استخراج تلاش شده است از نوع سؤالات مختلف شامل مالکیتی، زمان، چرایی، کجایی، چپستی، چگونگی، چطورری، چرا، کدام و ... استفاده شود.

○ در تهیه سؤالات از سؤالاتی با چندین ارتباط استفاده شده است؛ برای مثال در سؤال گفته شده است "شغل پدر نویسنده مقاله چیست؟"، در حالی که در متن مقاله اسم شخص آمده و پدر شخص عنوان شده و در ادامه به شغل پدر او پرداخته شده است.

○ سؤالاتی شرطی نیز استخراج شده است مثل سؤال "شرکت ایران خودرو دولتی است یا خصوصی؟"

○ در تهیه سؤالات از کلمات معادل و یا مفهومی که در متن دقیقاً استفاده نشده، به تکرار استفاده شده است.

○ در تهیه سؤالات هم در متن سؤال و هم در جواب سؤال از کلماتی که به غیر از زبان فارسی بوده‌اند نیز استفاده شده است. برای مثال سؤال "سنبل به کدام تیره تعلق دارد؟" جواب: "Hyacinthaceae".

○ در سؤال و جواب آن از معادل نوشتار اعداد فارسی و برعکس آن استفاده شده است؛ مثل (سه، دو، یک و ...)

○ در طراحی سؤال‌ها دقت شده که هر سؤال فقط از یک مقاله قابل استخراج و در مقالات دیگر احتمال داشتن پاسخ خیلی پایین باشد.

۳-۱-۴- آماده سازی دادگان

در این مرحله داده‌های ایجاد شده به دو قسمت آموزش و آزمون تقسیم شدند، در این مرحله برای نسخه یک تعداد نه هزار سؤال برای داده‌های آموزش و هزار سؤال برای داده‌های آزمون در نظر گرفته شدند و فایل‌های JSON مطابق استاندارد SQuAD ایجاد شدند. این فایل‌ها قابلیت ادغام با نسخه اصلی SQuAD را دارند.

به منظور آموزش مدل‌ها دستور ذیل و با پارامترهای نمونه اجرا شده است:

```
python ./src/run_squad.py \
--model_type bert \
--model_name_or_path "${model}" \
--do_train \
--do_eval \
--train_file $DATA_DIR/FarsiQA_V1_Train.json \
--predict_file $DATA_DIR/FarsiQA_V1_Eval.json \
--learning_rate "${learning_rate[@]}" \
--num_train_epochs "${num_train_epoch[@]}" \
--max_seq_length 384 \
--doc_stride 128 \
--output_dir
"Question&Answer_JFmodel/${model}_learning_rate=${learning_rate}_num_t
rain_epoch=${num_train_epoch}" \
--per_gpu_eval_batch_size=256 \
--per_gpu_train_batch_size=4 \
--save_steps 5000
```

(شکل - ۷). نمونه اسکریپت اجرا شده برای آموزش مدل

(Figure-7). Sample script executed for model training

۳-۱-۸- آموزش مدل‌ها با نسخه ۱ دادگان (FarsiQuAD)
برای شروع آموزش‌ها، مدل‌ها با هاپیر پارامترهای ذیل تنظیم و برای داده‌های دادگان ساخته شده آموزش دیدند.

(جدول - ۲) تنظیم هاپیر پارامترها برای آموزش مدل‌ها

(Table - ۲). Hyperparameter Settings for Model Training

۳۸۴	max_seq_length
۱۲۸	doc_stride
۴	Batch_Size
5e-5 و 3e-5	نرخ یادگیری ^۴
۷ و ۳	تعداد مراحل آموزش ^۵

در این مرحله مدل‌های زیر با هاپیر پارامترهای عنوان شده آموزش داده شده و برای وظیفه پرسش و پاسخ به زبان فارسی تنظیم شدند:

- bert-base-multilingual-uncased
- distilbert-base-uncased
- bert-base-multilingual-cased-finetuned-dutch-squad2^۶
- bert-base-parsbert-uncased
- bert-fa-base-uncased
- bert-fa-zwnj-base
- distilbert-fa-zwnj-base

۳-۱-۹- آموزش با دو مجموعه سؤالات زبان فارسی

با توجه اینکه مدل‌های ارائه شده بتوانند نمونه داده‌های بیشتری را ببینند تعدادی از مدل‌هایی که در مرحله قبل نتایج بهتری را به دست آورده بودند با استفاده از تجمیع فایل‌های آموزش و تست برای پیکره ایجاد شده FarsiQuAD و پیکره PersianQA، مجدد آموزش داده شدند.

^۴ Learning Rate

^۵ epoch

^۶ <https://huggingface.co/henryk/bert-base-multilingual-cased-finetuned-dutch-squad2>

این مدل قبلاً توسط پیکره squad2 نیز آموزش داده شده است.

تفاوت‌های عمده دو دادگان به شرح زیر است:

- دادگان PersianQA محاوره‌ای است و FarsiQuAD نوشتار رسمی دارد.
- در PersianQA جواب‌ها از چندین مقاله قابل استخراج است در مقابل FarsiQuAD سؤال‌ها از یک مقاله استخراج می‌شود.
- تعداد سؤالات استخراج شده از هر مقاله با هم متفاوت است.
- میانگین طول نویسه‌های مقاله‌های FarsiQuAD تقریباً سه برابر مقاله‌های PersianQA است.
- میانگین طول نویسه‌های سؤال‌های FarsiQuAD حدوداً چهار برابر است.
- میانگین طول نویسه‌های پاسخ‌های FarsiQuAD تقریباً بیش از دو برابر است.

(جدول - ۱). مقایسه FarsiQuAD با PersianQA

(Table -1) Comparison of FarsiQuAD with PersianQA

Persian QA	Farsi QuAD V2	Farsi QuAD V1	مقایسه نویسه‌ها
۹۹۳۸	۱۴۵۶۲۲	۱۰۰۲۱	تعداد جفت پرسش و پاسخ
2980	۱۳۵۷۵۳	152	بدون پاسخ
۸.۳۹	۳۴	۳۴	میانگین طول سؤالات
۲۲۴.۵۸	۷۷۷	۷۷۷	میانگین طول مقاله‌ها
۹.۶۱	۲۳	۲۳	میانگین طول پاسخ

۳-۱-۷- آموزش مدل یادگیری عمیق

در این بخش مدل‌های مختلف شبکه‌های یادگیری عمیق را با استفاده از داده‌هایی که در گام قبل آماده کرده بودیم آموزش داده شد که نحوه آموزش و ارائه به شرح زیر است: به منظور آموزش از مدل‌های از پیش آموزش دیده شده و روش یادگیری انتقالی استفاده شده تا بتوان مدل‌های که از قبل آموزش دیده شده‌اند برای انجام وظیفه پرسش و پاسخ به زبان فارسی تنظیم د.

برای آموزش مدل‌ها از کتابخانه پایتورچ^۱، کتابخانه مبدل^۲ huggingface، CUDA و اسکریپت ارائه شده برای آموزش مدل‌های SQuAD استاندارد^۳ استفاده شد تا آموزش و نتایج آن استاندارد باشد.

برای آموزش از سخت‌افزار GPU RTX2070 و شبکه‌های مختلف با مدت بیش از ۸۵ ساعت آموزش دیدند.

^۱ <https://pytorch.org>

^۲ <https://github.com/huggingface/transformers>

^۳ https://github.com/huggingface/transformers/blob/main/examples/legacy/question-answering/run_squad.py

گرفته شد که تعداد آن برای نسخه ۱ برابر ۱۰۰۰ جفت پرسش و پاسخ است و هایپرپارامترهای این مرحله متناسب با هایپرپارامترهای زمان آموزش تنظیم شدند. اسکرپت ذیل برای تمام مدل‌های آموزش‌دیده شده اجرا شدند و خروجی بر اساس معیارهای ارزیابی ارائه شد که در قسمت ۵ به تفصیل نتایج ارائه شده است.

```
export model="My Model"
python ./src/run_squad.py \
--model_type bert \
--model_name_or_path "${model}" \
--do_eval \
--predict_file $DATA_DIR/eval.json \
--max_seq_length 512 \
--doc_stride 256 \
--output_dir "${model}/eval" \
--per_gpu_eval_batch_size=256
```

۴- ارزیابی و تجزیه و تحلیل نتایج

در این بخش به نتایج حاصل از کار انجام‌گرفته پرداخته شده و نتایج خروجی‌ها و آزمایش‌های انجام‌گرفته مورد ارزیابی و بررسی قرار خواهد گرفت.

۴-۱- معرفی معیارهای ارزیابی

دو معیار برای بسیاری از مجموعه‌های داده برای پاسخگویی به سؤالات استفاده می‌شود، از جمله این موارد تطابق دقیق (EM) و امتیاز F1 است. این درصدها بر اساس پرسش‌های فردی + جفت پاسخ محاسبه می‌گردند. زمانی که چندین پاسخ صحیح برای یک سؤال امکان‌پذیر می‌شود، حداکثر درصد امتیاز نسبت به همه پاسخ‌های صحیح احتمالی محاسبه می‌شود. به‌طورکلی درصد امتیاز تطابق دقیق و F1 برای یک مدل با میانگین امتیازهای نمونه‌های فردی محاسبه می‌شود.

۴-۱-۱- تطابق دقیق

این معیار خیلی ساده است. برای هر جفت سؤال+پاسخ، اگر نویسه‌های پیش‌بینی مدل، دقیقاً با نویسه‌های (یکی از) پاسخ‌های واقعی مطابقت داشته باشد $EM=1$ و در غیر این صورت $EM=0$ می‌شود. این یک معیار دقیق همه یا هیچ است، یعنی با یک نویسه تغییر نمره این قسمت صفر خواهد بود.

۴-۱-۲- معیار F1

امتیاز F1 یک معیار متداول برای مسائل طبقه‌بندی است و به طور گسترده در QA استفاده می‌شود. موقعی مناسب

آموزش مدل با تنظیم هایپر پارامترهای عنوان شده در جدول (۵) انجام شده است.

(جدول - ۳). تنظیم هایپرپارامترها برای آموزش دو مجموعه سؤالات زبان فارسی

Hyperparameter Settings for Training Two Persian Language Question Sets (Table - ۳)

۵۱۲	max_seq_length
۲۵۶	doc_stride
۴	Batch_Size
5e-5 و 3e-5	نرخ یادگیری ^۱
۷ و ۳	تعداد مراحل آموزش ^۲

۳-۱-۱- آموزش با سه مجموعه دادگان پرسش و پاسخ

به‌منظور اینکه بررسی شود آیا آموزش مدل چندزبانه می‌تواند به بهبود عملکرد نتایج داده‌های تست زبان فارسی کمک کند یا خیر، از ترکیب سه مجموعه داده SQuAD^۳ V2.0، مجموعه دادگان FarsiQuAD و دادگان PersianQA با هم ترکیب شدند تا مجدد شبکه آموزش ببیند. این ترکیب باعث شده شبکه به سنگین شود و مدت ۱۲ ساعت آموزش با GPU زمان برد تا آموزش انجام شود.

مدل پایه انتخاب شده برای آموزش به شرح ذیل است:

- bert-base-multilingual-uncased
- نرخ یادگیری مدل 3e-5 و تعداد مراحل آموزش ۳ تنظیم شدند بقیه موارد تنظیمی مشابه مرحله قبل بوده است.

۳-۱-۱۱- آموزش با نسخه ۲ دادگان FarsiQuAD

به‌منظور افزایش دقت مدل‌ها برای مقابله با یافتن جواب‌های غیرقابل قبول، از تجمیع دو دادگان فارسی نسخه ۲ FarsiQuAD و دادگان PersianQA استفاده شد و با تنظیم هایپرپارامترهای نرخ یادگیری 3e-5 و تعداد مراحل آموزش ۳ مدل ذیل آموزش دید.

- bert-base-multilingual-cased-finetuned-dutch-squad2
- این مدل کمک می‌کند وقتی جواب سؤال در متن وجود ندارد از آن متن برای یافتن پاسخ استفاده نشده و در پاسخ‌ها قرار نگیرند.

۳-۱-۱۲- ارزیابی مدل‌ها

به‌منظور ارزیابی نتایج ۱۰ درصد از سؤالات دادگان به صورت تصادفی به عنوان داده‌های تست در نظر

^۱ Learning Rate

^۲ epoch

^۳ <https://rajpurkar.github.io/SQuAD-explorer/>

است که ما به صحت^۱ و بازخوانی^۲ (حساسیت^۳) اهمیت بدهیم. در این مورد، بر روی کلمات پیش‌بینی‌شده در برابر مواردی که در پاسخ واقعی هستند، محاسبه می‌شود. تعداد کلمات مشترک بین پیش‌بینی و جواب‌های درست اساس نمره F1 است: صحت نسبت تعداد کلمات مشترک به کل کلمات پیش‌بینی‌شده و فراخوانی نسبت تعداد کلمات مشترک به تعداد کل کلمات در جواب‌های درست اصلی است.

$$F1 = \frac{2}{\frac{-1}{\text{صحت}} + \frac{-1}{\text{حساسیت}}} = 2 \cdot \frac{\text{حساسیت} \times \text{صحت}}{\text{حساسیت} + \text{صحت}}$$

۴-۱-۳- اندازه فایل مدل‌ها

باتوجه به اینکه یکی از محدودیت‌های مدل‌های پردازش زبان طبیعی حجم بالای فایل‌های مدل آموزش‌دیده است، لذا در این بررسی اندازه فایل مدل نیز با اندازه مگابایت گزارش می‌شود تا اندازه مدل و دقت مدل‌ها آنها قابل‌اندازه‌گیری باشد و زمانی که محدودیت در استفاده از مدل‌های حجیم داریم، می‌شود با درصد کمی دقت پایین‌تر از این مدل‌ها استفاده کرد.

۴-۲- داده‌های ارزیابی

به‌منظور ارزیابی مدل‌ها ده درصد از سال‌ها به‌عنوان داده آزمون در نظر گرفته شدند که در نسخه ۱ دادگان FarsiQuAD برابر با هزار جفت پرسش‌وپاسخ می‌شود و در نسخه ۲ تعداد جفت سؤال‌وجواب‌ها برابر دوهزار است. در مدل‌هایی که داده‌های آموزشی با سایر مجموعه‌داده‌ها ترکیب شده بودند در موقع تست نیز داده‌های تست آن با هم ترکیب شده و نتایج گزارش شده‌است.

۴-۳- نتایج مدل‌های آموزش‌داده‌شده

باتوجه به اینکه مدل‌های آموزش‌دیده در چندین حالت مختلف بودند، لذا در ادامه به تفکیک نتایج آموزش مدل‌ها ارائه می‌شود. مقادیر ارزیابی استخراج‌شده برای سؤال‌های دارای پاسخ هستند.

۴-۳-۱- نتایج آموزش مدل‌ها بر روی داده‌های

FarsiQuAD نسخه ۱

مدل‌های آموزش‌دیده‌شده بر اساس این دادگان (نسخه ۱) دارای نتایج یادشده است.

¹ precision
² recall
³ sensitivity

همان‌طور که در گزارش مشخص است، در این آموزش‌ها مدل پیش‌آموزش دیده‌شده با زبان فارسی نتایج بهتری را کسب کرده و تعداد مراحل ۳ با نرخ یادگیری 3e-5 نیز نتایج بهتری نسبت به بقیه هابیر پارامترها داشته‌اند.

(جدول - ۴). نتایج معیار F1 برای مدل‌های آموزش دیده با

دادگان FarsiQuAD در نرخ‌های یادگیری و

مراحل آموزش متفاوت

(Table -4) F1 Metric Results for Models Trained on FarsiQuAD Data at Different Learning Rates and Training Stages

نام مدل پیش آموزش دیده	$\eta(3e-5)$ e(3)	$\eta(3e-5)$ e(7)	$\eta(5e-5)$ e(3)	$\eta(5e-5)$ e(7)
bert-m-cased	84.92	83.73	84.83	83.71
bert-m-uncased	83.11	83.62	83.84	82.68
distilbert-uncased	53.99	57.64	47.85	56.10
xlm-roberta-base	81.48	81.78	81.35	80.01
bert-parsbert-uncased	84.94	84.93	83.09	82.43
bert-fa-uncased	85.38	84.88	85.10	84.35
bert-fa-zwnj	79.21	79.29	80.10	80.06
distilbert-fa-zwnj	78.95	77.77	77.83	77.67

(جدول - ۵). معیار تنظیم دقیق برای مدل‌های آموزش دیده با

دادگان FarsiQuAD در نرخ‌های یادگیری و

مراحل آموزش متفاوت

(Table -5). Precision Metric for Models Trained on FarsiQuAD Data at Different Learning Rates and Training Stages

مدل های پیش آموزش دیده	$\eta(3e-5)$ e(3)	$\eta(3e-5)$ e(7)	$\eta(5e-5)$ e(3)	$\eta(5e-5)$ e(7)
bert-m-cased	76.95	76.55	75.53	75.63
bert-m-uncased	75.13	75.23	74.92	73.10
distilbert-uncased	38.98	43.76	32.79	40.10
xlm-roberta-base	72.18	72.69	70.86	70.05
bert-parsbert-uncased	76.55	76.24	73.60	73.30
bert-fa-base-uncased	77.16	76.14	76.04	75.33
bert-fa-zwnj-base	68.53	70.05	71.78	69.85
distilbert-fa-zwnj	68.93	67.82	66.70	67.41

به‌منظور مقایسه حجم فایل مدل‌های آموزش‌دیده

شده در شکل (۱۰) اندازه حجم این مدل‌های آموزش‌دیده

با دادگان FarsiQuAD به مگابایت گزارش شده‌اند.

بر اساس نتایج حاصل در جدول (۷) همچنان مدل DistilBERT دارای حجم بسیاری پائینی نسبت به سایر مدل‌ها بوده، ولی معیار F1 و تنظیم دقیق به‌طور تقریبی ده درصد کمتر را نسبت به سایر مدل‌ها ارائه داده‌است.

(جدول - ۷). حجم فایل مدل آموزش داده شده با دو دادگان و

مقایسه با معیار F1

(Table-7). Size of Trained Model Files with Two Datasets and Comparison Based on F1 Metric

مدل های پایه پیش آموزش دیده	حجم فایل	معیار F1
bert-m-uncased	۶۳۶	۸۰.۵۱
distilbert-uncased	۲۵۳	۴۹.۰۶
bert-parsbert-uncased	۶۱۹	۸۱.۲۱
bert-fa-uncased	۶۱۹	۸۱.۰۹
bert-fa-zwnj-base	۴۴۹	۷۵.۸۱
distilbert-fa-zwnj	۲۸۶	۷۲.۶۹
bert-m-finetuned-dutch-squad2	۶۷۶	۸۳.۲۶

در ادامه مجدد تنها با داده‌های تست دادگان FarsiQuAD عملیات تست انجام شده تا بررسی شود آموزش با دو مجموعه دادگان چقدر می‌تواند نتایج حاصل از مجموعه تست داده‌های دادگان ایجادشده را افزایش دهد.

با بررسی این موضوع در جدول (۷)، آموزش مدل تجمیعی با دو دادگان باعث افزایش کمی در نتایج دادگان شده‌است و نشان‌دهنده تفاوت در سؤالات استخراج شده در دو مجموعه دادگان دارد.

در این قسمت نیز مدل چندزبانه گوگل که در قبل توسط دادگان SQuAD آموزش‌دیده نتایج بهتری را نسبت به سایر مدل‌ها ارائه کرده‌است.

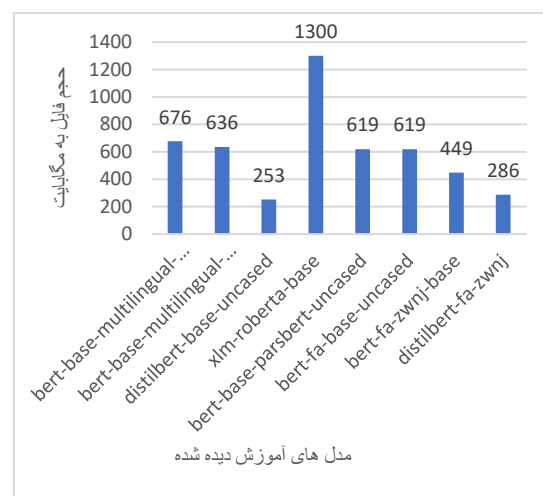
(جدول - ۸). نتایج مدل های تجمیعی بر روی داده های تست

FarsiQuAD

(Table - ۸). Aggregate Model Results on FarsiQuAD Test Data

نام مدل های پیش آموزش دیده	تطابق دقیق	F1
bert-m-uncased	76.04	85.56
distilbert-uncased	39.09	55.86
bert-parsbert-uncased	76.24	86.15
bert-fa-uncased	76.24	86.15
bert-fa-zwnj-base	72.18	82.24
distilbert-fa-zwnj	68.02	79.47
bert-m-cased-finetuned-dutch-squad2	77.97	86.81

با مقایسه حجم‌های مدل‌ها، مشخص است که مدل DistilBERT حجم فایل مدل خیلی کمتری نسبت به سایر مدل‌ها دارد؛ درحالی‌که معیار F1 و تنظیم دقیق برای این مدل اختلاف زیادی را ندارد.



(شکل - ۸). حجم فایل مدل های آموزش داده شده (MB) با

دادگان FarsiQuAD

(Figure-8). Size of trained model files (MB) with the FarsiQuAD dataset

۲-۳-۴- نتایج آموزش مدل‌ها بر روی داده‌های دو

دادگان زبان فارسی

بعد از ترکیب داده‌های تست دادگان سؤال‌وجواب FarsiQuAD و PersianQA نتایج به شرح شکل (۴-۲۱) حاصل شدند. این نتایج برای معیارهای دارای پاسخ‌است.

(جدول - ۶). نتایج مدل تجمیعی با داده تست

دو دادگان زبان فارسی

(Table-6). Aggregate Model Results on Persian Language Test Data

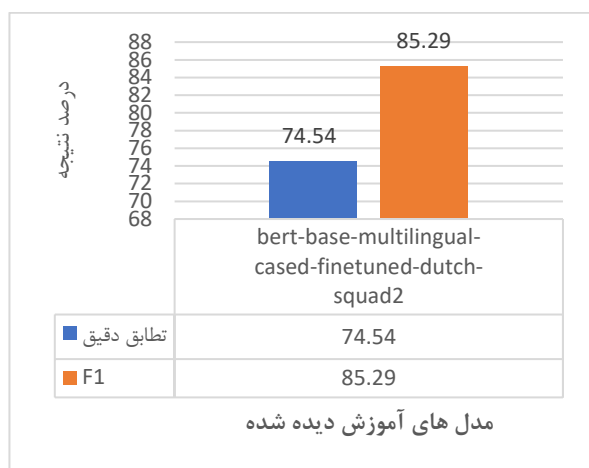
مدل های پیش آموزش دیده شده	تطابق دقیق	F1
bert-m-uncased	64.12	80.51
distilbert-base-uncased	31.48	49.06
bert-parsbert-uncased	64.55	81.21
bert-fa-uncased	65.34	81.09
bert-fa-zwnj-base	59.78	75.81
distilbert-fa-zwnj	56.17	72.69
bert-base-m-cased-dutch-squad2	68.40	83.26

مدل چندزبانه که قبلاً با SQuAD2 آموزش‌دیده در این مرحله نتایج بهتری را نسبت به سایر مدل‌ها داده‌است.

۴-۳-۳- نتایج آموزش مدل‌ها با سه مجموعه

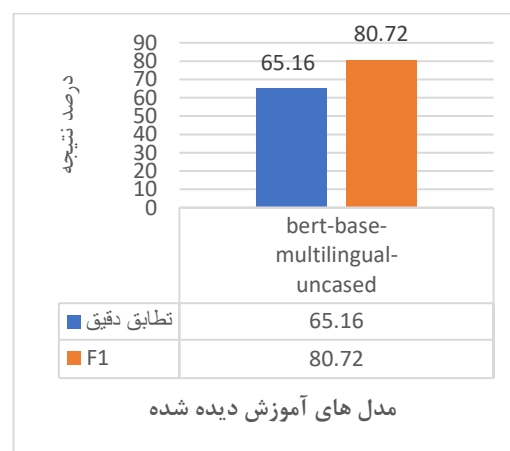
دادگان

با ارزیابی مدل آموزش داده شده برای سه مجموعه دادگان پرسش و پاسخ شامل (SQuADv2، PersianQA، FarsiQuAD) نتایج ذیل حاصل شدند.



(شکل - ۱۱). نتایج حاصل از آموزش مدل با نسخه ۲ دادگان FarsiQuAD

(Figure -11). Results obtained from model training with version 2 of the FarsiQuAD dataset



(شکل - ۹). نتایج آموزش مدل با سه دادگان پرسش و پاسخ و تست

بر روی تجمیع دو دادگان فارسی

(Figure-9) Results of model training with three question-and-answer datasets and testing on the aggregation of two Persian datasets

۴-۳-۵- مقایسه مدل ایجاد شده با کارهای دیگر

باتوجه به اینکه مسئله پرسش و پاسخ جزو مسائل سخت طبقه بندی می شوند، لذا حل این موضوع با روش های عادی به سختی حل می شوند و بیشتر کارهای جدید از طریق مدل های یادگیری عمیق بوده که زبان را یاد گرفته اند، استفاده می شود؛ لذا در این قسمت داده های تستی با تعدادی از مدل های پیش آموزش دیده شده به زبان فارسی و سایر زبان ها بررسی شده و نتایج با بهترین مدل آموزش دیده شده در این مقاله مقایسه شده است.

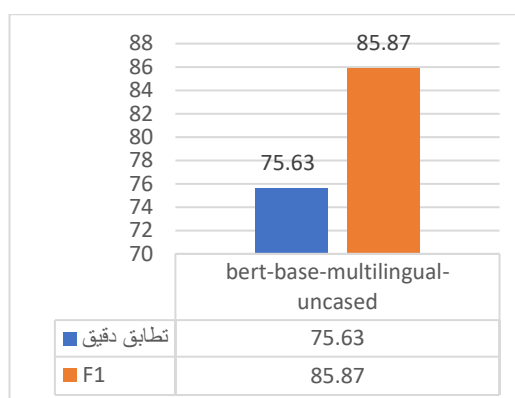
(جدول - ۹). مقایسه بهترین مدل آموزش دیده شده با سایر

مدل ها بر اساس داده های تست دادگان FarsiQuAD

(Table -9) . Comparison of the Best Trained Model with Other Models Based on FarsiQuAD Test Data

نام مدل ها	تطابق دقیق	F1
BERT-m(My Model)	77.97	86.81
BERT-base(FA)	43.00	69.00
BERT-m(EN)	50.61	68.20
XLM-RoBERTa-large(FA)	27.1	66.35
BERT(FA)	37.32	65.64
XLM-RoBERTa-large(EN)	29.41	60.38
RoBERTa-base(EN)	0.91	3.88
RoBERTa-large(EN)	0.91	3.29

نتایج مندرج در جدول (۹) نشان می دهد که مدل آموزش داده شده با داده های دادگان FarsiQuAD توانسته است معیار تطبیق دقیق را ۲۷.۳۶ و معیار F1 را ۱۷.۸۱



(شکل - ۱۰). نتایج آموزش مدل با سه دادگان پرسش و پاسخ و

تست با داده های FarsiQuAD

(Figure -10). Results of model training with three question-and-answer datasets and testing with FarsiQuAD data

در بررسی به دست آمده مشخص شد آموزش مدل چندزبانه به صورت همزمان با سه مجموعه داده های عنوان شده به مقدار کمی نتایج را بهبود می بخشد؛ ولی هنوز نتایج ضعیف تری نسبت به مدل های آموزش داده شده تک زبانه به زبان فارسی دارد.

۴-۳-۴- نتایج آموزش مدل‌ها با نسخه ۲ دادگان

FarsiQuAD

با آموزش مدل با نسخه ۲ دادگان و تجمیع با مجموعه داده دادگان قبلی زبان فارسی، نتایج زیر برای

به منظور مقاومت سازی شبکه برای مقابله با استخراج جواب های نامربوط، مدل منتخب مراحل گذشته با نسخه شماره ۲ دادگان FarsiQuAD آموزش داده شد و به نتایج خوبی رسید.

درصد ارتقا دهد. همچنین در این بررسی مشخص شد مدل هایی که به صورت چندزبانه و یا زبان فارسی آموزش دیده اند، نتایج قابل قبولی دارند و مدل هایی که با زبان های تک زبانه غیر از فارسی آموزش دیده باشند، نتایج مناسبی ندارند.

6-Refrence

۶- مراجع

- [1]. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.
- [2]. Yuanjun Li, Yuzhu Zhang, Question Answering on SQuAD 2.0 Dataset, s. University, Editor, 2018.
- [3]. d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q., & Vidal, M. FQuAD: French question answering dataset. arXiv preprint arXiv:2002.06071, 2020.
- [4]. Möller, T., Risch, J., & Pietsch, M. Germanquad and germanpr: Improving non-english question answering and passage retrieval. arXiv preprint arXiv:2104.12741, 2021.
- [5]. 임승영, 김명지, & 이주열. KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋. 한국정보과학회 학술발표논문집, 539-541, 2018.
- [6]. 김영민, 임승영, 이현정, 박소윤, & 김명지. KorQuAD 2.0: 웹문서 기계독해를 위한 한국어 질의응답 데이터셋. 정보과학회논문지, 47(6), 577-586, 2020.
- [7]. So, B., Byun, K., Kang, K., & Cho, S. Jaquad: Japanese question answering dataset for machine reading comprehension. arXiv preprint arXiv:2202.01764, 2022.
- [8]. Ayoubi MY Sajjad & Davoodeh Persianqa: a dataset for persian question answering. <https://github.com/SajjadAyoubi/PersianQA>, 2021.
- [9]. Mozafari, J., Fatemi, A., & Nematbakhsh, M. A. BAS: an answer selection method using BERT language model. arXiv preprint arXiv:1911.01528, 2019.
- [10]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- [11]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [12]. Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. Parsbert: Transformer-based model for persian language understanding. Neural Processing Letters, 53(6), 3831-3847, 2021.

۵- نتیجه گیری

با گسترش روزافزون استفاده از مدل های یادگیری عمیق، این مدل ها در حال جایگزین شدن با روش های قبلی و یا ترکیب با این مدل های هستند. هم اکنون بیشتر وظایف زبان های طبیعی از طریق یادگیری عمیق حل می شوند. مدل های یادگیری عمیق می توانند مفهوم متن را در خود داشته باشند و جستجوی های مفهومی را به خوبی پاسخ دهند. از قابلیت های خوب مدل های یادگیری عمیق سرعت اجرای مناسب آن ها است.

در این پژوهش تلاش شد وظیفه پرسش و پاسخ به زبان فارسی با یادگیری عمیق را ارتقا دهد. بدین منظور دادگانی از مجموعه جفت پرسش و پاسخ ها به زبان فارسی تهیه و نام FarsiQuAD را برای آن انتخاب شد. این دادگان می تواند در آموزش مدل های یادگیری عمیق که به طور معمول با محدودیت منابع آموزشی دست و پنجه نرم می کند، مفید واقع شود.

برای عملیات استخراج داده ها بیش از ۲۵۰۰ مقاله از مقاله های نمونه استفاده و از آنها سؤال هایی گوناگون در حوزه های مختلف شامل اشخاص، اماکن، حیوانات و ... با انواع مختلف سؤال شامل چرایی، چرایی، چگونگی و ... به تعداد بیش از ده هزار سؤال اقدام شد.

برای آموزش مدل ها از ساختار ترنسفورمر [۱۰، ۹] و تعداد هشت مدل پیش آموزش دیده شده با زبان طبیعی شامل مدل های معروف (BERT [۱۱]، ParsBert [۱۲]، DistilBERT [۱۳]، RoBERTa [۱۴]، XML [۱۵]، XML-RoBERTa [۱۶]) که با متون زیاد زبان فارسی و یا چندزبانه آموزش دیده شده اند، استفاده و عملیات تنظیم آنها با آموزش مجدد بر روی دادگان FarsiQuAD انجام شد.

در ادامه تنظیم مدل های یادگیری عمیق، این مدل ها با تجمیع دو دادگان زبان فارسی FarsiQuAD و PersianQA و سه مجموعه دادگان پرسش و پاسخ (FarsiQuAD، PersianQA، SQuAD انگلیسی) آموزش داده شدند.

محمدعلی کیوان راد استادیار گروه هوش مصنوعی و
عضو هیئت علمی دانشگاه صنعتی مالک اشتر است.
نشانی رایانامه ایشان عبارت است از:
keyvanrad@aut.ac.ir

- [13]. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.
- [14]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- [15]. Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.
- [16]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
- [17]. Persian Wikipedia. Available from: <https://github.com/miladfa7/Persian-Wikipedia-Dataset>



جواد فروتن راد کارشناسی ارشد

خود را در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی از دانشگاه
صنعتی مالک اشتر دریافت کرده و
هم‌اکنون دانشجوی دکترای هوش

مصنوعی است. ایشان علاقه‌مند به استفاده از فناوری‌های
نوین نظیر هوش مصنوعی در سایر حوزه‌های مهندسی و
پیوند آن با صنعت است. زمینه‌های کاری ایشان عبارت
از پردازش زبان طبیعی، فناوری مالی، هوش مصنوعی و
هوش تجاری در حوزه کاری است.
نشانی رایانامه ایشان عبارت است از :

Forutanrad@gmail.com



مریم حور علی کارشناسی ارشد

خود را در رشته مهندسی فناوری
اطلاعات گرایش تجارت الکترونیک
از دانشگاه علم و صنعت ایران و
دکترای خود را در گرایش مهندسی
فناوری اطلاعات از دانشگاه

تربیت‌مدرس دریافت کرده‌است. ایشان در حال حاضر
استادیار و عضو گروه علمی فناوری اطلاعات و فرماندهی
کنترل و هوش مصنوعی دانشگاه صنعتی مالک اشتر است.
زمینه‌های پژوهشی موردعلاقه ایشان عبارت‌اند از: پردازش
زبان طبیعی، مهندسی فناوری اطلاعات و سامانه‌های
فازی.

نشانی رایانامه ایشان عبارت است از:

mhourali@mut.ac.it