

A Deep Neural Network-Based Information Retrieval Method for Complex Question-Answering Systems

A. Ghafouri¹, H. Naderi^{2*}, M.R.Hassani Ahangar³

*Associate Professor, University of Science and Technology, Tehran, Iran

(Received: 2024/02/20, Revised: 2024/04/14, Accepted: 2024/05/09, Published: 2024/06/02)

DOR: <https://dorl.net/dor/20.1001.1.23224347.1403.12.1.1.9>

ABSTRACT

Question-answering systems, as the next generation of search engines, have the capability to retrieve relevant answers to queries posed in natural language. These systems generally consist of three main components: question processing, information retrieval, and answer extraction, with various methods introduced for each component. One of the most important aspects is information retrieval and selecting relevant paragraphs. Nowadays, most user queries in question-answering systems are complex. To respond to such queries, it initially requires understanding the question and then retrieving various documents that are lexically and semantically related to the query. In recent years, advancements in deep neural network-based learning on one hand and the introduction of high-quality datasets on a large scale on the other have attracted researchers' attention to this field. In this research, a method for selecting relevant paragraphs for complex question-answering systems on the HotpotQA dataset is introduced. To select relevant paragraphs, the type of question is first recognized using a deep neural network. Then, using BERT language models, relevant paragraphs that show evidence of the answer are selected in several steps. This process uses keyword extraction from the question.

The results obtained indicate that the outcomes are better compared to the baseline method

Keywords: Information Retrieval, Complex Question Answering, Paragraph Selection, Deep Neural Networks .

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Publisher: Imam Hussein University

© Authors



*Corresponding Author Email: naderi@just.ac.ir

علمی - پژوهشی

ارائه یک روش بازیابی اطلاعات مبتنی بر شبکه‌های عصبی عمیق برای سامانه‌های پرسش و پاسخ

پیشیده

آرش غفوری^۱، حسن نادری^{۲*}، محمدرضا حسنی آهنگر^۳

۱. دانشجوی دکتری ۲. دانشیار، دانشگاه علم و صنعت ایران، تهران، ایران.

۳. استاد، دانشگاه جامع امام حسین^(ع)، تهران، ایران.

(دریافت: ۱۴۰۲/۱۲/۱۱، بازنگری: ۱۴۰۳/۰۱/۲۶، پذیرش: ۱۴۰۳/۰۲/۲۰، انتشار: ۱۴۰۳/۰۳/۱۳)

DOR: <https://dorl.net/dor/20.1001.1.23224347.1403.12.1.1.9>

* این مقاله یک مقاله با دسترسی آزاد است که تحت شرایط و ضوابط مجوز Creative Commons Attribution (CC BY) توزیع شده است.

نویسندگان

ناشر: دانشگاه جامع امام حسین (ع)

چکیده

سیستم‌های پرسش پاسخ به‌عنوان نسل جدید موتورهای جستجو، توانایی بازیابی پاسخ مرتبط به پرسش‌های ارائه شده به زبان طبیعی را دارند. عموماً این سامانه‌ها از سه بخش اصلی پردازش پرسش، بازیابی اطلاعات و استخراج پاسخ تشکیل شده‌اند و برای هر بخش روش‌های متعددی معرفی شده است. یکی از مهم‌ترین بخش‌ها بازیابی اطلاعات و انتخاب بندهای مرتبط است. امروزه اغلب پرسش‌های کاربران در دستگاه‌های پرسش و پاسخ از نوع پیچیده است. برای پاسخ‌دهی به این‌گونه پرسش‌ها ابتدا نیازمند درک پرسش و سپس بازیابی اسناد مختلف است که از نظر واژگانی و معنایی به پرسش مرتبط هستند. در سال‌های اخیر پیشرفت یادگیری مبتنی بر شبکه‌های عصبی عمیق از یک سو و معرفی مجموعه داده‌های باکیفیت در مقیاس بزرگ از سوی دیگر باعث جلب توجه پژوهشگران به این حوزه شده است. در این پژوهش یک روش برای انتخاب بندهای مرتبط برای دستگاه‌های پرسش پاسخ پیچیده روی مجموعه داده هات پات کیوای معرفی شده است. برای انتخاب بندهای مرتبط ابتدا نوع پرسش را با استفاده از یک شبکه عصبی عمیق تشخیص داده می‌شود. سپس با استفاده از مدل‌های زبانی برت در چند گام بندهای مرتبط که شواهدی از پاسخ در آن نمایان باشد انتخاب می‌شوند. برای این کار از روش استخراج کلمات کلیدی در پرسش استفاده شده است. نتایج به دست آمده نشان از بهتر بودن نتایج نسبت به روش پایه دارد.

کلیدواژه‌ها: بازیابی اطلاعات، پرسش پاسخ پیچیده، انتخاب بند، شبکه‌های عصبی عمیق

۱- مقدمه

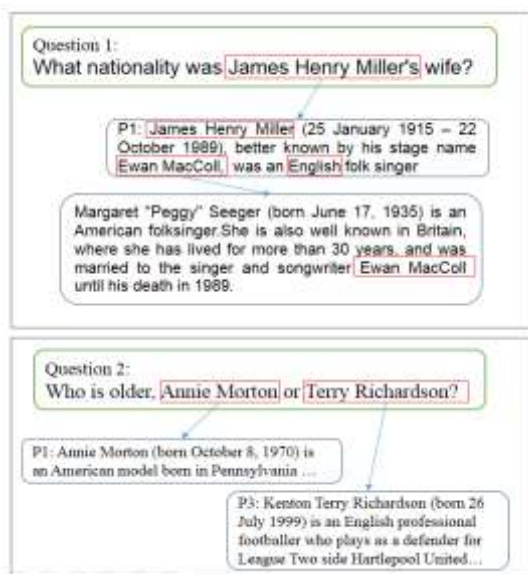
هستند نیاز به شواهد چندگانه دارد که این شواهد معمولاً در جاهای مختلف متن یک یا چندین سند پراکنده شده‌اند. در دستگاه‌های پرسش و پاسخ بازیابی و انتخاب بندهای مرتبط با پرسش از اهمیت ویژه‌ای برخوردار است. این مهم موجب می‌شود این سیستم‌ها بتوانند پاسخ درست را با دقت و در حداقل زمان ممکن با استدلال مناسب استخراج نمایند. دسته‌بندی‌های مختلفی برای دستگاه‌های پرسش و پاسخ وجود دارد که غالباً این دسته‌بندی‌ها براساس محدوده موضوعی، نوع پرسش و نوع پاسخ مورد پشتیبانی توسط این سیستم‌ها است. نوع سیستم پرسش و پاسخ مطرح شده در این مقاله از نوع دامنه‌باز است. در این نوع از دستگاه‌های پرسش و پاسخ بر خلاف دستگاه‌های پرسش و پاسخ

سیستم‌های پرسش پاسخ یک ابزار نوین برای پاسخ‌دهی به پرسش‌های ساده و پیچیده کاربران به صورت خودکار مبتنی بر مجموعه داده‌های ساخت یافته، نیمه‌ساختاریافته و غیرساخت یافته هستند. این سیستم‌ها با دریافت درخواست‌ها و پرسش‌های متنوع کاربران که از ساختار زبان طبیعی پیروی می‌کنند نسبت به پاسخ‌دهی به پرسش‌ها اقدام می‌نمایند [۱] بسیاری از مطالعات اخیر در زمینه پرسش پاسخ روی تک پاراگراف‌ها تمرکز کرده‌اند [۲] امروزه پاسخ‌دهی به بسیاری از پرسش‌ها کاربران در

دستگاه‌های پرسش و پاسخ مبتنی بر یک پاراگراف ممکن نیست. در حال حاضر پاسخ‌دهی به پرسش‌های کاربران که غالباً پیچیده

*Corresponding Author E-mail: naderi@iust.ac.ir

نتیجه‌گیری، راجع به پژوهش انجام شده جمع‌بندی صورت می‌گیرد.



شکل (۱): نمونه‌ای از پرسش‌های پیچیده در مجموعه داده هات پات کیوای. پرسش ۱: پرسش پیچیده نوع پل. پرسش ۲: پرسش پیچیده نوع مقایسه‌ای

۲- کارهای مرتبط

برای استدلال از نوع چند گامی در کار آقای کیو و همکاران [۵] از یک ساختار گرافی مبتنی بر موجودیت‌ها برای شناسایی بندهای مرتبط استفاده شده است. در این کار ساخت گراف از موجودیت‌های داخل پرسش آغاز می‌شود و تا رسیدن به جواب مناسب، بندهای مرتبط بازیابی می‌شوند. در این پژوهش یک زیر شبکه آموزش داده می‌شود که بندهای مرتبط را انتخاب کند. این زیر شبکه بر اساس مدل از پیش آموزش داده شده برت [۶] است. در این شبکه با استفاده از لایه سیگموئید پرسش و پاراگراف دریافت شده و مرتبط بودن و نبودن برای هر پاراگراف با نمایش ۱ و ۰ در خروجی اعلام می‌گردد.

در پژوهشی دیگر آقای آسای و همکاران [۷] از پیوندهای موجود در مقاله‌های ویکی‌پدیا به بهترین وجه بهره‌برداری نموده‌اند. برای پیدا کردن پاسخ مربوط به یک پرسش پیچیده ابتدا مقاله‌هایی که شامل قسمتی از اطلاعات پرسش باشند پیدا می‌گردد. این کار توسط روش TF-IDF انجام می‌شود. اگر پاسخ مناسب داخل آن مقاله‌ها پیدا نشود، از طریق پیوندهای موجود در آن مقاله، سراغ مقاله‌های دیگر رفته و این کار تا زمان یافتن اطلاعات کافی برای پاسخ دادن به پرسش کاربر ادامه می‌یابد. علاوه بر روش TF-IDF می‌توان از سایر روش‌های پیشرفته‌تر مانند [۸] و [۹] جهت بالا بردن دقت بازیابی اطلاعات استفاده کرد.

دامنه بسته که تنها از یک موضوع خاص محتوایی به‌عنوان مثال پزشکی پشتیبانی می‌شود، محتوای پرسش و پاسخ به دسته خاصی تعلق نداشته و تنوع موضوعی دارد.

پرسش‌های مطرح شده در دستگاه‌های پرسش و پاسخ نیز شامل دو نوع ساده و پیچیده است. منظور از پرسش‌های ساده غالباً پرسش‌های حقیقت‌محور است و پاسخ آن‌ها معمولاً یک حقیقت یا عبارت کوتاه است. برای پرسش‌های پیچیده انواع مختلفی وجود دارد که از مهم‌ترین آن‌ها می‌توان به پرسش‌های مقایسه‌ای (تصدیقی و یا تاییدی)، چند گزینه‌ای، پل و نوع آزاد (بدون محدودیت) اشاره کرد. پرسش‌های پیچیده مورد بحث در این مقاله از نوع مقایسه‌ای و پل است. در پرسش‌های مقایسه‌ای معمولاً چند موجودیت یا موضوع با یکدیگر مقایسه می‌شوند. همچنین در این پرسش‌ها می‌تواند یک موضوع تصدیق یا رد گردد در این صورت پاسخ پرسش بله یا خیر خواهد بود. پرسش‌های پل از دو یا چند پرسش ساده تو در تو تشکیل شده است به گونه‌ای که به‌عنوان مثال اگر یک پرسش پل قابل تجزیه به دو پرسش ساده باشد، پاسخ پرسش ساده اول بخشی از پرسش ساده دوم خواهد بود. دستگاه‌های پرسش و پاسخ برای بازیابی پاسخ در مواجهه با پرسش‌های پیچیده یا به عبارت دقیق‌تر چندگامی، نیازمند بررسی حداقل دو پاراگراف متفاوت هستند [۳]. این سیستم‌ها در ادامه برای پاسخدهی به پرسش نسبت به استدلال روی پاراگراف‌ها و پرسش از یک پاراگراف به پاراگراف‌های دیگر در همان سند یا اسناد دیگر اقدام می‌نمایند که به این قابلیت استدلال چندگامی می‌گویند [۴]. برای ارزیابی این نوع دستگاه‌های پرسش و پاسخ مجموعه داده‌های متفاوتی ارائه شده است که مجموعه داده هات پات کیوای [۳] از مجموعه داده‌های مطرح در این زمینه است. این مجموعه داده‌ای به استخراج پاسخ مبتنی بر استدلال چندگامی تاکید دارد و شامل حدود ۱۱۳ هزار پرسش دامنه‌باز است. نمونه‌ای از پرسش‌های پیچیده چند گامی که برای استخراج پاسخ نیازمند چند پاراگراف است در شکل (۱) نشان داده شده است. پرسش‌های موجود در این مجموعه داده در دو دسته پرسش‌های پل و پرسش‌های مقایسه‌ای به‌صورت دامنه‌باز برچسب زده شده‌اند.

سازماندهی مقاله در ادامه به‌صورت زیر است. در بخش بعد کارهای مرتبط با پژوهش انجام شده در مقاله بررسی می‌گردد. در بخش سوم روش پیشنهادی، معماری و اجزای آن به‌صورت کامل تشریح می‌گردد. در بخش چهارم مجموعه داده اصلی و مجموعه داده‌های میانی ساخته شده از آن معرفی می‌گردد. در ادامه و در بخش پنجم روش پیشنهادی و مؤلفه‌های آن مورد آزمون قرار گرفته و نتایج آن گزارش می‌گردد. در انتها ضمن بیان

۳- روش پیشنهادی

به‌منظور بازیابی اطلاعات لازم برای پاسخ‌دهی به پرسش‌های پیچیده روشی مبتنی بر یادگیری عمیق پیشنهاد گردید. این روش دارای معماری و متشکل از مؤلفه‌های تشخیص نوع پرسش، استخراج موجودیت و عبارت اسمی و انتخاب بندهای مرتبط است. در ادامه ضمن معرفی هر مؤلفه چگونگی در کنار هم قراردادن مؤلفه‌های مذکور و تعامل آنها با یکدیگر برای بهبود نتایج انتخاب بندهای مرتبط تشریح می‌گردد.

اولین مؤلفه روش پیشنهادی تشخیص نوع پرسش است. این مؤلفه ضمن دریافت پرسش کاربر وظیفه تشخیص نوع پرسش پیچیده (پل یا مقایسه‌ای) را دارد.

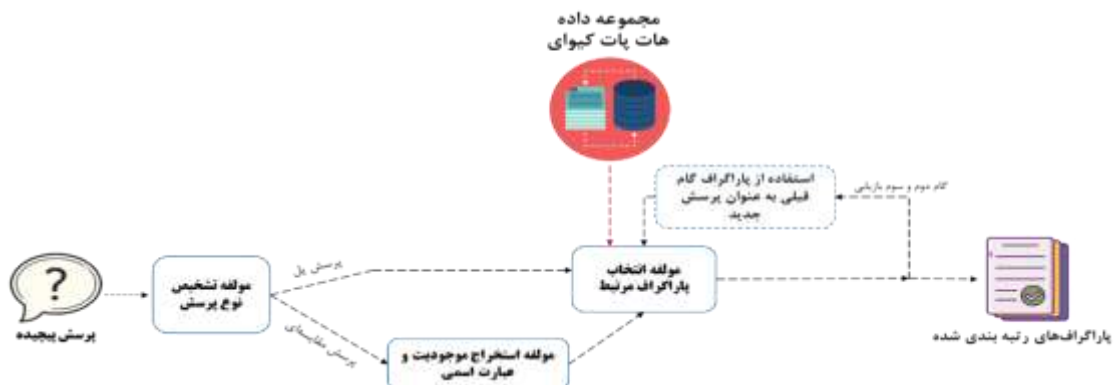
دومین مؤلفه استخراج موجودیت و عبارت اسمی در پرسش است که جهت بازیابی اطلاعات در پرسش‌های مقایسه‌ای مورد استفاده قرار می‌گیرد.

سومین مؤلفه انتخاب بندهای مرتبط است. در این مؤلفه از مدل زبانی برت استفاده می‌شود. به‌منظور شناسایی و انتخاب پاراگراف‌های مرتبط لایه‌های عصبی این مدل زبانی سفارشی شده و بر روی داده‌های مجموعه داده هات پات کیوای تنظیم دقیق گردیده است. این عمل برای رده‌بندی و بهبود عملکرد تشخیص بندهای مرتبط لازم است. در پرسش‌های پل بازیابی اطلاعات به‌صورت تریبی و در چند گام انجام می‌شود که در ادامه به تفصیل راجع به آن بحث می‌گردد. معماری روش پیشنهادی در شکل ۲ نشان داده شده است.

بسیاری از مدل‌های قبلی بر اساس تکنیک «جستجو کن پاسخ رو پیدا کن» کار می‌کنند و نتایج خوبی را روی مجموعه داده‌های معروف نیز گرفته‌اند. اما برای پرسش‌های چندگزینه‌ای این روش‌ها خوب عمل نمی‌کنند. از این رو برای پرسش‌های چندگزینه‌ای آقای نی و همکاران [۱۰] از یک روش مبتنی بر ادغام پرسش و پاسخ‌های موجود برای انتخاب بندهای مرتبط استفاده کرده‌اند. در این پژوهش از فرمول‌سازی مجدد پرسش شامل کلمه‌های ضروری به همراه گزینه‌های پاسخ استفاده شده است. سپس پاراگرافی با بیشترین امتیاز انتخاب می‌شود و پاسخ ادغام شده با پرسش به‌عنوان جواب تعیین می‌شود.

کارهای مرتبط انجام شده معمولاً یا از روش بازیابی یک‌باره (تک گامی) بهره برده‌اند و یا از روش‌های بازیابی مکرر استفاده نموده‌اند. روش‌های بازیابی یک‌باره به دلیل استفاده از تطابق لغوی حجم زیادی از اسناد غیرمرتبط را برای پاسخ‌دهی به پرسش‌های پیچیده بازمی‌گردانند و در بازیابی بندهای مرتبط دوم با مشکل مواجه می‌شوند. در روش‌های بازیابی مکرر با وجود به‌کارگیری تطابق واژگانی دومرحله‌ای به دلیل عدم امکان شناسایی گستره شامل پاسخ، بازیابی اسناد در مرحله دوم دچار رشد تصاعدی می‌گردد و یافتن سند و یا بندهای مرتبط‌تر را دشوار می‌سازد.

برای غلبه بر چالش‌ها و نقاط ضعف موجود در کارهای مرتبط مذکور، در بخش بعد و با معرفی روش پیشنهادی ضمن در نظر گرفتن هم‌زمان تطابق واژگانی و معنایی میان متن اسناد و پرسش و همچنین بهره‌گیری از شبکه‌های عصبی عمیق راه‌حلی جهت بازیابی اطلاعات برای سیستم‌های پرسش‌وپاسخ پیچیده پیشنهاد خواهد گردید.



شکل (۲): معماری روش پیشنهادی. در گام اول تشخیص نوع پرسش توسط مؤلفه اول انجام می‌شود. در صورتی که پرسش از نوع مقایسه‌ای باشد در مؤلفه دوم استخراج موجودیت و عبارت اسمی صورت می‌گیرد و با استفاده از آن بندهای مرتبط توسط مؤلفه سوم استخراج می‌شود. در غیر این صورت و در صورتی که پرسش از نوع پل باشد انتخاب پاراگراف مرتبط به‌صورت مستقیم توسط مؤلفه سوم انجام می‌گردد.

است. طول بردار ورودی به صورت ۵۱۲ در نظر گرفته شده است و از معماری ارائه شده در [۱۳] برای این کار استفاده شده است.

در نهایت پرسش موردنظر تبدیل به یک بردار شده و وارد شبکه LSTM به صورت معادله (۱) می‌شود.

$$U = \text{LSTM}(Q) \in \mathbb{R}^{n \times h} \quad (1)$$

که در آن h بعد خروجی رمزنگار و n اندازه بردار خروجی است. بردار خروجی به یک شبکه عصبی FFNN [۱۲] با یک لایه ورودی به ابعاد ۵۱۲ و خروجی با ابعاد ۱ با معادله (۲) و تابع فعال ساز سافت‌مکس با معادله (۳) نشان داده می‌شود و بیان کننده نوع پرسش است. خروجی نهایی با معادله (۴) مشخص می‌شود.

$$H_i = \sum W_{ij} X_{jk} + b_i \quad (2)$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}} \quad (3)$$

$$Y = \text{Softmax}(U) \in \mathbb{R}^k \quad (4)$$

در معادله (۴) k خروجی دودویی و بیان‌کننده کلاس پرسش ورودی است. معماری این کار در **Error! Reference source not found.** نشان داده شده است.

با روش بیان شده می‌توان نوع پرسش را مشخص کرده و آن را در یکی از دسته پرسش‌های پل و مقایسه‌ای رده‌بندی نمود. در نهایت با توجه به مشخص شدن نوع پرسش می‌توان راهبرد دقیق‌تری جهت انتخاب بندهای مرتبط برگزید.

۳-۲- استخراج موجودیت و عبارت اسمی

در پژوهش‌های گذشته انتخاب بندهای مرتبط فارغ از نوع پرسش و به صورت مستقیم از طریق محتوای پرسش انجام گرفته است. اما در روش پیشنهادی ارائه شده همان‌طور که پیش‌تر بیان شد راهبرد بکار گرفته شده جهت انتخاب بندهای مرتبط متناسب با پرسش تعیین شده است. از این‌رو در این بخش برای پرسش‌های مقایسه‌ای از استخراج عبارات اسمی کلیدی و برای پرسش‌های پل مجموعه کلمات پرسش به صورت کامل مورد استفاده قرار می‌گیرد.

پرسش مقایسه‌ای

در پرسش‌های مقایسه‌ای دو یا چند موجودیت یا عبارت اسمی مقایسه می‌شوند و جواب این نوع پرسش‌ها غالباً بله و خیر است. از این‌رو استخراج موجودیت‌ها یا عبارات اسمی در انتخاب بندهای مرتبط که شامل آن عبارت باشند کمک‌کننده خواهد بود. مدل‌های قبلی که روی استخراج موجودیت‌ها و عبارات اسمی

۳-۱- تشخیص نوع پرسش

تشخیص نوع پرسش در پرسش‌های پیچیده و بکار بردن تکنیک‌های مختلف متناسب با نوع پرسش دقت انتخاب بندهای مرتبط را بهبود می‌بخشد.

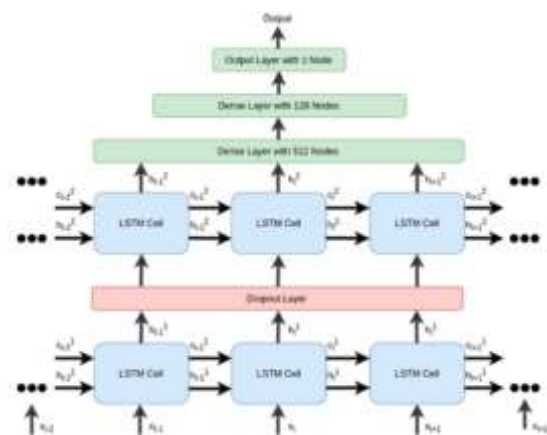
برای این کار از مجموعه داده هات پات کیوای که پرسش‌ها را در دودسته پرسش‌های پل و مقایسه‌ای برچسب‌گذاری کرده است استفاده می‌شود.

برای این کار از یک شبکه عصبی LSTM [۱۱] به همراه شبکه عصبی FFNN [۱۲] جهت تشخیص نوع پرسش استفاده گردید. برای آموزش مدل از یک مجموعه داده شامل ۱۷۴۵۶ پرسش که در دو دسته مقایسه‌ای و پل برچسب‌گذاری شده است استفاده گردید. در جدول ۱ می‌توان نمونه‌ای از پرسش‌های برچسب‌گذاری شده را مشاهده کنید.

جدول (۱): نمونه‌هایی از پرسش‌های پل و مقایسه‌ای

پرسش	برچسب تعیین شده
Who is older, Annie Morton or Terry Richardson?	Comparison
What nationality was James Henry Miller's wife?	Bridge
Brown State Fishing Lake is in a country that has a population of how many inhabitants?	Bridge
What is the middle name of the actress who plays Bobbi Bacha in Suburban Madness?	Bridge
Are Giuseppe Verdi and Ambroise Thomas both Opera composers?	Comparison

معماری مدل عصبی مؤلفه تشخیص نوع پرسش در **Error! Reference source not found.** نشان داده شده است.



شکل (۳): معماری شبکه عصبی جهت تعیین نوع پرسش [۱۳].

ورودی مدل آموزش‌دیده یک پرسش است که نمایش برداری آن $Q = [q_1, q_2, \dots, q_n]$ است و در آن q بیانگر هر توکن از پرسش

می‌تواند در تعداد مشخصی از کلاس‌ها که در این روش سه کلاس است تعیین کند.

سه کلاس خروجی شامل کلاس B-key که نشان‌دهنده آغاز عبارت اسمی شناسایی شده، کلاس I-Key که نشان‌دهنده ادامه عبارت اسمی و O که نشان‌دهنده عبارت اسمی نبودن است. بر همین اساس ورودی جداساز ما به صورت معادله ۵ است.

$$F(q) = [t_1, t_2, t_3, \dots, t_n] \quad (5)$$

در این روش خروجی موجودیت یا عبارت اسمی است.

یک نمونه از استخراج موجودیت‌ها در جدول نشان داده شده است.

پرسش پل

آنچه در بازیابی برای پرسش‌های پل اهمیت دارد ایجاد یک رابطه بین پرسش، پاراگراف اول و پاراگراف دوم است. این یعنی از طریق پرسش بتوان پاراگراف اول و از طریق پاراگراف اول باید پاراگراف دوم را بازیابی نمود. از آنجایی که در پرسش‌های مطرح شده از نوع پل لزوماً موجودیت و یا عبارت اسمی مشخصی وجود ندارد، بهتر است از همه متن پرسش برای انتخاب پاراگراف مرتبط استفاده شود. به عنوان نمونه پرسش جدول شامل موجودیتی نیست که بتوان با استفاده از آن پاراگراف اول را شناسایی کرد و بیشتر کلمات، کلمات مهمی هستند.

جدول (۲): نمونه‌ای از پرسش‌های پل

R	نمونه‌ای از پرسش پل
1	Which genus of moth in the world's seventh-largest country contains only one species?

برای این نوع پرسش‌ها از تمام متن پرسش برای بازیابی در گام اول استفاده می‌شود.

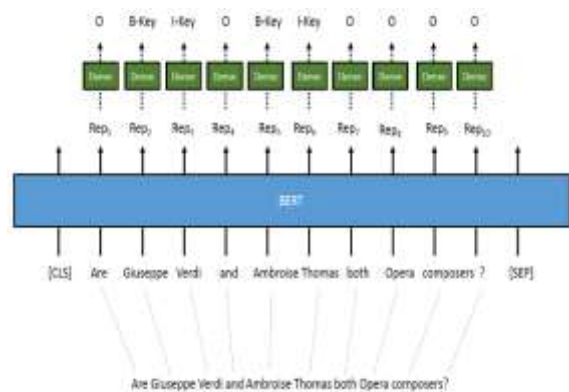
جدول (۳): نمونه استخراج کلمه‌های کلیدی در پرسش

پرسش و موجودیت‌ها، عبارت اسمی استخراج شده
Who is older, Annie Morton or Terry Richardson?
What nationality was James Henry Miller's wife?
Brown State Fishing Lake is in a country that has a population of how many inhabitants ?
What is the middle name of the actress who plays Bobbi Bacha in Suburban Madness?

۳-۳- انتخاب بندهای مرتبط

برای پرسش‌های مقایسه‌ای و پرسش‌های پل به صورت کامل تشریح می‌گردد.

کلیدی تمرکز داشتند به دلیل تعیین نوع موجودیت‌ها نسبت به مؤلفه استخراج موجودیت و کلمات کلیدی روش پیشنهادی عملکرد ضعیف‌تری دارند. این مدل‌ها علاوه بر تشخیص عبارت متنی موجودیت، نوع آن را در چندین دسته مختلف تشخیص می‌دهند و به همین دلیل دقت مدل‌های تشخیص موجودیت کاهش می‌یابد. از این رو برای این کار نیازمند مدلی است که کلمات کلیدی و مهم داخل پرسش را بدون نیازمندی به نوع آن تشخیص دهد.



شکل (۴): ساختار مدل برای استخراج عبارت اسمی

برای این کار از مدل برت پایه [۱۲] در این مؤلفه از روش پیشنهادی استفاده شده است که به حروف بزرگ و کوچک نیز حساس است. این مدل نیز همانند مدل‌های تشخیص موجودیت آموزش داده می‌شود، تفاوت این مدل با روش‌های قبلی در این است که فقط عبارت اسمی را استخراج می‌کند. بدلیل صرف نظر از تعیین نوع عبارت اسمی از پیچیدگی آن کاسته شده است.

معماری ساختار استخراج عبارت اسمی در **Error!** معماری ساختار استخراج عبارت اسمی در **Reference source not found.** برای هر کلمه یا توکن خروجی از مدل برت یک رده‌بند جداگانه از جنس شبکه عصبی کاملاً متصل متراکم [۱۲] قرار داده شده است. این شبکه عصبی تمام بردارهای خروجی برای هر کلمه را

انتخاب بندهای مرتبط با توجه به دسته‌بندی نوع پرسش به دو صورت انتخاب پاراگراف برای پرسش‌های مقایسه‌ای و انتخاب پاراگراف برای پرسش‌های پل انجام می‌شود. در ادامه این روند

پرسش مقایسه‌ای

برای تعیین پاراگراف‌ها همان‌طور که در شکل ۲ نشان داده شده است از ساختاری مبتنی بر مدل زبانی برت استفاده شده است. این ساختار بر اساس پیش‌بینی ادامه‌دار بودن جمله دوم برای جمله اول است. از این روش به همراه پیش‌بینی ماسک برای آموزش مدل برت استفاده می‌شود. به همین علت دقت خوبی برای مسئله مورد بررسی دارد. در این ساختار پرسش و پاراگراف به طور هم‌زمان و با قراردادن «[SEP]» بین آن‌ها جهت تشخیص و جداسازی در کدگشا وارد مدل می‌شود و در نهایت خروجی نشان‌دهنده میزان تشابه دو عبارت ورودی خواهد بود.

در ساختار ارائه شده ورودی به صورت $Q = \{q_1, q_2, \dots, q_n\}$ و $P = \{p_1, p_2, \dots, p_n\}$ است که Q و P به ترتیب پرسش و پاراگراف هستند. q_1 تا q_n توکن‌های پرسش هستند که بعد از استفاده از جداساز مدل برت وارد کدگذار برت می‌شوند. همچنین p_1 تا p_n توکن‌های پاراگراف ورودی هستند که برای جداسازی آن‌ها از جداساز مدل برت پایه استفاده شده است.

باتوجه به ساختار ارائه شده، پرسش و پاراگراف ابتدا وارد کدگشای مدل برت پایه شده، سپس بردار خروجی وارد یک‌لایه متراکم می‌شود. در مرحله آخر تابع فعال‌ساز سافت مکس در خروجی تعیین می‌کند که پرسش و پاراگراف دریافت شده به هم تا چه اندازه مرتبط هستند. با استفاده از این روش پاراگرافی که بیشترین امتیاز را داشته باشد به‌عنوان پاراگراف مرتبط اول انتخاب می‌شود.

در ادامه پاراگراف انتخاب شده اول به‌عنوان پرسش در نظر گرفته می‌شود و به بندهای موجود با استفاده از جداکننده «[SEP]» ادغام می‌شود. با استفاده از روش قبلی که در شکل ۲ نشان داده شده است پاراگرافی که بیشترین امتیاز را داشته باشد به‌عنوان پاراگراف مرتبط دوم انتخاب می‌شود.

۴- مجموعه داده‌ها

مجموعه داده هات پات کیوای مجموعه‌ای متشکل از ۱۱۳ هزار پرسش به همراه پاسخ است. این مجموعه داده در دو حالت فول ویکی و دیسترکتور آماده‌سازی شده است. در حالت فول ویکی پرسش‌ها و عناوین اسناد مرتبط به همراه پاسخ نهایی پرسش قرار داده شده و روش پیشنهادی باید از طریق بازبازی اطلاعات روی کل اسناد ویکی‌پدیا توانایی پیدا کردن پاسخ را داشته باشد. در حالت دیسترکتور در بیشتر موارد برای هر پرسش ۱۰ پاراگراف در نظر گرفته شده است که پاسخ پرسش موردنظر با استدلال روی تعدادی از این پاراگراف‌ها قابل استنتاج است. در این روش انتخاب بندهای مرتبط برای مدل استدلال مهم و کاربردی است. دسته‌بندی مجموعه داده هات پات کیوای

در صورتی که پرسش مطرح شده از نوع مقایسه‌ای باشد، مدل پیشنهادی که در بخش‌های قبل معرفی گردید نوع آن را تشخیص می‌دهد. در این صورت مراحل انتخاب پاراگراف برای پرسش‌های مقایسه‌ای انجام می‌شود. به علت آن که در این نوع پرسش‌ها عملاً دو یا چند نوع موجودیت با هم مقایسه می‌شوند، با استفاده از پرسش می‌توان به‌صورت هم‌زمان تمام بندهای مرتبط را شناسایی کرد. در این نوع پرسش‌ها همان‌طور که گفته شد نیاز به بازیابی‌های ترتیبی پاراگراف‌ها نمی‌باشد.

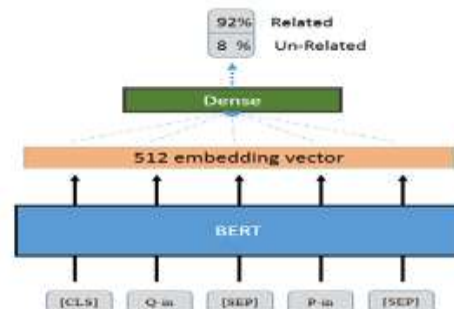
بعد از تشخیص نوع پرسش و استخراج عبارت اسمی داخل پرسش که شامل موجودیت‌ها نیز است عمل انتخاب بندهای مرتبط باید صورت گیرد.

برای این کار عبارت اسمی استخراج شده در عنوان پاراگراف‌ها جستجو می‌شود. در صورتی که عبارت اسمی در عنوان پاراگراف‌ها به‌صورت تطابق کامل پیدا شود، آن پاراگراف به‌عنوان پاراگراف مرتبط بازیابی می‌شود. در غیر این صورت عبارت اسمی استخراج شده در متن پاراگراف‌ها جستجو شده و پاراگرافی که آن عبارت اسمی را در خود داشته باشد به‌عنوان بند مرتبط انتخاب می‌شود. این کار برای تمام عبارات اسمی در متن پرسش انجام شده و تمام بندهای مرتبط با استفاده از این روش انتخاب می‌شوند.

پرسش پل

در صورتی که پرسش مطرح شده از نوع پرسش پل باشد، از طریق پرسش به‌تنهایی نمی‌توان تمام پاراگراف‌ها که شواهدی از پاسخ داخل آن‌ها موجود است را تشخیص داد.

در موارد زیادی از پرسش‌های نوع پل اسمی موجودیت‌ها ذکر نمی‌شود. به‌طور کلی در این نوع پرسش‌ها سعی می‌شود که اطلاعات کلی در رابطه با بازیابی پاراگراف اول درج شود؛ لذا تنها با تأکید بر موجودیت‌ها و عبارت اسمی نمی‌توان انتظار انتخاب پاراگراف خوبی داشت. به همین علت از تمامی متن پرسش استفاده شده و پاراگراف مرتبط اول انتخاب می‌شود.



شکل ۲: انتخاب بندهای مرتبط با مدل زبانی برت

ساخته شده است به‌صورت جدول است. همچنین نمونه‌ای از داده‌های مجموعه‌داده در جدول قابل مشاهده است.

جدول (۶): مجموعه‌داده استخراج موجودیت و عبارت اسمی

نوع داده	داده‌های آموزشی	داده‌های آزمایشی
تعداد نمونه	۱۷۴۵۶ جمله	۱۴۸۵ جمله

جدول (۷): نمونه‌ای از مجموعه‌داده موجودیت و عبارت اسمی

کلمه کلیدی	جمله با پرسش
Arthur's Magazine	Which magazine was started first Arthur's Magazine or First for Women?
First for Women	

۵- آزمایش‌ها

در این بخش نتایج به‌دست‌آمده توسط مؤلفه‌های مختلف روش پیشنهادی با پژوهش‌های مرتبط مقایسه می‌گردد. سپس نتایج نهایی بازیابی اطلاعات توسط روش پیشنهادی در مقایسه با روش پایه مورد بررسی قرار می‌گیرد.

۵-۱- مؤلفه تشخیص نوع پرسش

همان‌طور که در بخش‌های قبل اشاره گردید، یکی از مؤلفه‌های مهم در وظیفه بازیابی اطلاعات روی مجموعه‌داده‌های پات‌کیوای تشخیص نوع پرسش است. در این روش نوع پرسش پل و مقایسه‌ای از هم تشخیص داده می‌شوند. تاکنون در پژوهش‌های انجام شده برای بازیابی اطلاعات روی مجموعه‌داده‌های پات‌کیوای هیچ‌کدام از مؤلفه تشخیص پرسش به‌طور صریح استفاده نکرده‌اند؛ لذا در این بخش فقط به بیان نتایج مدل ارائه شده پرداخته می‌شود.

در این روش مدل روی ۳۴۹۱۲ پرسش آموزش داده شده است. برای آموزش این مدل ۱۷۴۵۶ نمونه پرسش پل و ۱۷۴۵۶ نوع پرسش مقایسه‌ای استفاده شده است. همچنین ارزیابی روی داده‌های توسعه مجموعه‌داده‌های پات‌کیوای که شامل ۷۴۰۵ پرسش است انجام شده است. نتایج بدست آمده در جدول قابل مشاهده است.

جدول (۸): نتایج بدست آمده برای تشخیص نوع پرسش

	Precision	Recall	F1
Comparison	%۹۹	%۹۱	%۹۵
Bridge	%۹۸	%۹۵	%۹۶

مورد استفاده در روش پیشنهادی به‌صورت جدول است که از داده‌های آموزشی برای آموزش مدل و از داده‌های توسعه برای آزمایش مدل بهره گرفته شده است.

جدول (۴): دسته‌بندی مجموعه‌داده‌های پات کیوای [۳]

نام مجموعه‌داده	موارد استفاده	آمار تعداد پرسش‌ها
داده آسان	آموزش مدل	۱۸,۰۸۹
داده متوسط	آموزش مدل	۵۶,۸۱۴
داده سخت	آموزش مدل	۱۵,۶۶۱
داده توسعه	توسعه و بهبود	۷,۴۰۵

۴-۱- مجموعه‌داده تشخیص پرسش

تشخیص نوع پرسش یکی از مؤلفه‌های مهم در این پژوهش است که نیازمند داده‌های آموزشی و آزمایشی جهت آموزش و ارزیابی است. از این‌رو از پرسش‌های مجموعه‌داده‌های پات کیوای که برچسب نوع پرسش را دارند یک مجموعه‌داده جهت آموزش این مدل ایجاد شده است. این مجموعه‌داده شامل دو نوع پرسش پل و مقایسه‌ای است که شرح آن به‌صورت جدول است.

جدول (۵): توزیع مجموعه‌داده تشخیص نوع پرسش

نوع داده	پرسش‌های پل	پرسش‌های مقایسه‌ای
آموزشی	۱۷۴۵۶	۱۷۴۵۶
توسعه	۵۹۲۰	۱۴۸۵

۴-۲- مجموعه‌داده استخراج عبارت اسمی

دیگر مؤلفه استفاده شده در این پژوهش که نیازمند مجموعه‌داده‌ای جهت آموزش و ارزیابی مدل است، مؤلفه استخراج موجودیت و عبارت اسمی است. مجموعه‌داده موردنظر باید روی هات پات کیوای ایجاد شده باشد تا بتواند عبارت اسمی را در این مجموعه‌داده به‌خوبی استخراج کند. به علت آن که مجموعه‌داده‌های پات کیوای با استفاده از اسناد پایگاه دانش ویکی‌پدیا ساخته شده است این روش به‌خوبی روی اسناد ویکی‌پدیا نیز عمل می‌کند.

برای ساخت این مجموعه‌داده از پرسش‌ها و عناوین پاراگراف‌ها استفاده شده است. به این صورت که تمامی عناوین پاراگراف‌های تعیین شده برای هر پرسش در متن پرسش جستجو شده و هر عنوان پاراگرافی که در متن پرسش وجود داشته باشد آن عنوان متن به‌صورت یک عبارت اسمی برای آن پرسش در نظر گرفته می‌شود.

در این مجموعه‌داده ایجاد شده ممکن است یک یا چند عبارت اسمی برای هر پرسش تعیین شود. تعداد نمونه‌های مجموعه‌داده از طریق مجموعه‌داده‌های پات کیوای و صفحات ویکی‌پدیا

است که حداقل یک موجودیت یا عبارت اسمی در آن قرار دارد. نتایج این مدل و مقایسه آن با مدل‌هایی نظیر Spacy و Stanza در جدول قابل مشاهده است.

لازم به ذکر است معیار تعیین شده باید تمام عبارت اسمی را به درستی تشخیص دهد. ارزیابی روی ۱۴۸۵ جمله و ۵۲۷۹ عبارت اسمی انجام شده است.

همان‌طور که در جدول نشان داده شده است، دقت روی مدل‌های Stanza و Spacy پایین است و علت مربوط به تشخیص نوع موجودیت‌ها است. در تشخیص موجودیت‌ها باید نوع موجودیت مانند شخص، مکان، زمان و غیره نیز مشخص شود. همین موارد علتی است که باعث دقت پایین این مدل‌ها در استخراج عبارات اسمی و موجودیت‌ها می‌شوند. اما در مدل ارائه شده، آموزش به صورتی انجام شده است که فقط عبارت اسمی استخراج شود و از نوع آن صرف نظر شده است. همین کار باعث بهبود چشمگیر نتایج می‌شود.

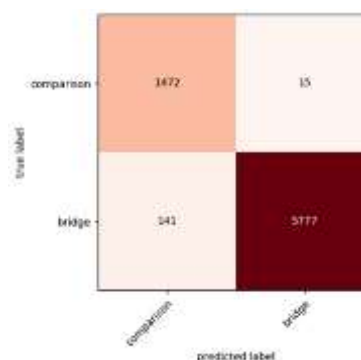
جدول (۹): مقایسه مدل استخراج عبارت اسمی با سایر روش‌ها

	Precesion	Recall	F1
Stanza	%۵۹	%۸۳	%۶۸
Spacy	%۵	%۸۰	%۶۲,۲۹
Our Model	%۹۹	%۹۸	%۹۸,۵

مدل‌های دیگری نظیر کی‌برت [۱۶] و یاکی نیز از روش‌های معروف برای استخراج عبارت اسمی است. مدل کی‌برت یک روش مبتنی بر مدل برت است درحالی که مدل یاکی بیشتر به صورت الگوریتمی سعی در استخراج عبارات اسمی دارد. اما به علت عملکرد ضعیف و عملکرد متفاوت در مجموعه داده هات‌پات کیوای برای استخراج موجودیت‌ها و عبارات اسمی از مقایسه آن با روش ارائه شده خودداری شده است.

۳-۵- ارزیابی

در این بخش به عنوان ارزیابی نهایی روش پیشنهادی، تمام مؤلفه‌های معماری آن در کنار هم قرار گرفته شده و نتایج انتخاب بندهای مرتبط نشان داده شده است. برای آزمون روش پیشنهادی و مقایسه آن با روش پایه از حالت دیسترتور مجموعه داده هات پات کیوای استفاده شده است. در این بخش برای هر پرسش دو پاراگراف به عنوان بندهای مرتبط تعیین شده است که متناسب با ساختار مجموعه داده هات پات کیوای است. روش پیشنهادی با روش پایه HGN [۱۷] مقایسه شده و نتایج مقایسه در جدول نشان داده شده است.



شکل (۶): ماتریس درهم ریختگی تشخیص نوع پرسش

در ماتریس درهم ریختگی شکل تعداد اشتباهات در تشخیص نشان داده شده است. در این ماتریس سطرها بیان کننده برچسب واقعی پرسش‌ها و ستون‌ها برچسب‌های پیش‌بینی شده توسط مدل است. ۱۴۱ پرسش از پرسش‌های مقایسه‌ای را مدل با پرسش پل برچسب زده است و در پرسش‌های پل نیز، مدل روی ۱۵ پرسش برچسب اشتباه پیش‌بینی کرده است.

در ساختار پرسش‌ها به علت آنکه ساختار متفاوتی مانند پرسش‌های متقاطع نیز وجود دارد کمی در پیش‌بینی پرسش نوع مقایسه‌ای مشکل ایجاد می‌گردد. درحالی که به صورت کلی نتایج خوب و رضایت بخش است.

۲-۵- تشخیص عبارت اسمی

دیگر مؤلفه مهم که در انتخاب بندهای مرتبط مهم است، مؤلفه تشخیص عبارت اسمی است. این مؤلفه در پرسش‌های مقایسه‌ای اهمیت ویژه‌ای دارد. در بسیاری از کارهای مرتبط انجام شده برای استخراج موجودیت‌ها از Spacy [۱۴] و یا Stanza [۱۵] استفاده شده است. این مؤلفه‌ها با استفاده از مدل‌های تشخیص موجودیت در بسیاری از کارهای مربوط به پردازش زبان طبیعی بسیار خوب عمل می‌کنند.

اما این مدل‌ها به علت آنکه روی مجموعه داده هات‌پات کیوای و یکی پدیا آموزش داده نشده‌اند و یا تنظیم دقیق نشده‌اند دقت خوبی نشان ندهاند و در بسیاری از موارد به خصوص موجودیت‌ها و عبارات اسمی بلند و چندکلمه‌ای ضعیف عمل می‌کنند. از آنجایی که استخراج عبارت اسمی یکی از راه‌های بهبود عملکرد بازیابی اطلاعات روی مجموعه داده هات‌پات کیوای است، در این روش پیشنهادی یک مدل زبانی مبتنی بر برت حساس به حروف بزرگ و کوچک برای استخراج عبارت اسمی فارغ از نوع آن‌ها آموزش داده شده است.

مدل ارائه شده روی ۱۷۴۵۶ جمله پرسشی که هر کدام از آن‌ها حداقل ۱ موجود یا عبارت اسمی دارند آموزش داده شده است. همچنین داده آزمایشی ساخته شده، شامل ۱۴۸۵ جمله پرسشی

- Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp. 2369–2380, 2018, doi: 10.18653/v1/d18-1259.
- [4] Y. Feldman and R. El-Yaniv, “Multi-hop Paragraph Retrieval for Open-domain Question Answering,” ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 2296–2309, 2020, doi: 10.18653/v1/p19-1222.
- [5] L. Qiu et al., “Dynamically fused graph network for multi-hop reasoning,” ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 6140–6150, 2020, doi: 10.18653/v1/p19-1617.
- [6] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Naacl-Hlt 2019, no. Mlm, 2018, [Online]. Available: <https://github.com/tensorflow/tensor2tensor>
- [7] A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, and C. Xiong, “Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering,” 2019, [Online]. Available: <http://arxiv.org/abs/1911.10470>
- [8] R. Nogueira and K. Cho, “Passage Re-ranking with BERT,” 2019, [Online]. Available: <http://arxiv.org/abs/1901.04085>
- [9] Y. Nie, S. Wang, and M. Bansal, “Revealing the importance of semantic retrieval for machine reading at scale,” EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 2553–2566, 2019, doi: 10.18653/v1/d19-1258.
- [10] J. Ni, C. Zhu, W. Chen, and J. McAuley, “Learning to Attend On Essential Terms: An Enhanced Reader-Reader Model for Open-domain Question Answering,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 335–344, 2019, doi: 10.18653/v1/n19-1030.
- [11] S. Hochreiter and J. Urgan Schmidhuber, “Long Shortterm Memory,” Neural Comput., vol. 9, no. 8, p. 17351780, 1997, [Online]. Available: <http://www7.informatik.tu-muenchen.de/~hochreit%0Ahttp://www.idsia.ch/~juergen>
- [12] G. Bebis and M. Georgiopoulos, “Feed-forward Neural Networks,” IEEE Potentials, vol. 13, no. 4, pp. 27–31, 2002, doi: 10.1109/45.329294.
- [13] M. F. Rabby, Y. Tu, M. I. Hossen, I. Lee, A. S. Maida, and X. Hei, “Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction,” BMC Med. Inform. Decis. Mak., vol. 21, no. 1, 2021, doi: 10.1186/s12911-021-01462-5.
- [14] M. Neumann, D. King, I. Beltagy, and W. Ammar, “ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing,” BioNLP 2019 - SIGBioMed Work. Biomed. Nat. Lang. Process. Proc. 18th BioNLP Work. Shar. Task, pp. 319–327, 2019, doi: 10.18653/v1/w19-5034.
- [15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” pp. 101–108, 2020, doi: 10.18653/v1/2020.acl-demos.14.
- [16] M. Grootendorst, “KeyBERT: Minimal Keyword Extraction with BERT,” Zenodo, 2020, [Online]. Available: <https://github.com/MaartenGr/KeyBERT>
- [17] Y. Fang, S. Sun, Z. Gan, R. Pillai, S. Wang, and J. Liu, “Hierarchical Graph Network for Multi-hop Question Answering,” EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 8823–8838, 2020, doi: 10.18653/v1/2020.emnlp-main.710.

همان‌طور که مشخص است نتایج روش پیشنهادی در انتخاب ۲ پاراگراف مرتبط نسبت به کار مرتبط بهتر عمل کرده است.

جدول (۱۰): نتایج مدل تشخیص پاراگراف در ساختار دیسترتور و

مقایسه آن با مدل‌های پایه

	Precision	Recall	F1-score
[۱۷] HGN	۹۴٫۵۳	۹۴٫۵۳	۹۴٫۵۳
Our Model	۹۵٫۰۱	۹۴٫۲۳	۹۴٫۶۱

در روش HGN تمام پاراگراف‌ها به‌صورت مستقیم توسط پرسش انتخاب می‌شوند و باعث می‌شود در پرسش‌های پل ارتباط بین پاراگراف‌ها لحاظ نشود. این کار روی پرسش‌های مطرح‌شده جدید به علت عدم استفاده از ارتباط بین پاراگراف‌ها به مشکل می‌خورد. همین مورد باعث شده است که نتایج روش پیشنهادی بهتر از روش پایه عمل کند.

۶- نتیجه‌گیری

در این مقاله به ارزیابی روش پیشنهادی بر روی مجموعه‌داده‌های پات‌کیوای که یک مجموعه‌داده انگلیسی بر اساس مقاله‌های ویکی‌پدیا است پرداخته شد. برای ارزیابی، مؤلفه‌های مختلف معماری روش پیشنهادی مورد بررسی قرار گرفت. نتایج نشانگر عملکرد بهتر هر مؤلفه نسبت به سایر روش‌های مشابه مورد استفاده در پژوهش‌های مرتبط گذشته بوده و این امر در نهایت منجر به بهبود نتایج کلی روش پیشنهادی گردیده است. در نهایت انتخاب بندهای مرتبط با پرسش پیچیده مبتنی بر روش پیشنهادی روی مجموعه‌داده‌های پات‌کیوای بر اساس دقت، فراخوانی و معیار اف وان با مدل پایه مقایسه گردید. نتایج به‌دست‌آمده حاکی از برتری روش پیشنهادی در دقت و معیار اف وان نسبت به روش پایه بود. در این پژوهش نشان داده شد، تشخیص نوع پرسش و به‌کارگیری سازوکار متناسب با نوع سؤال در بازیابی پاراگراف‌های مرتبط مؤثر است و منجر به کسب نتایج بهتر می‌گردد.

۷- مراجع

- [1] M. A. Calijorne Soares and F. S. Parreiras, “A Literature Review on Question Answering Techniques, Paradigms and Systems,” J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 6, pp. 635–646, 2020, doi: 10.1016/j.jksuci.2018.08.005.
- [2] P. Rajpurkar, R. Jia, and P. Liang, “Know what you don’t know: Unanswerable questions for SQuAD,” ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 2, pp. 784–789, 2018, doi: 10.18653/v1/p18-2124.
- [3] Z. Yang et al., “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering,” Proc. 2018

