ارائه یک روش بازیابی اطلاعات مبتنی بر شبکه های عصبی عمیق برای سیستم‌های پرسش‌وپاسخ پیچیده

آرش غفوری[1] و حسن نادری*[2]

[1] دانشجوی دکتری، دانشگاه علم و صنعت ، تهران ، ایران aghafuri@comp.iust.ac.ir

[2] دانشیار ، دانشگاه علم و صنعت ، تهران ، ایران naderi@iust.ac.ir

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

DOR: https://dorl.net/dor/20.1001.1.23224347.1402.11.4.8.5

چکیده: سیستم‌های پرسش‌پاسخ به عنوان نسل جدید موتورهای جستجو، توانایی بازیابی پاسخ مرتبط به پرسش‌های ارائه شده به زبان طبیعی را دارند. عموما این سیستم‌ها 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 پاراگراف‌های مرتبط است. امروزه اغلب پرسش های کاربران در سیستم‌های پرسش و پاسخ از نوع پیچیده است. برای پاسخدهی به این گونه 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 مجموعه‌داده‌های باکیفیت در مقیاس بزرگ از سوی دیگر باعث جلب توجه پژوهشگران به این حوزه شده است.

در این پژوهش یک روش برای انتخاب پاراگراف‌های مرتبط برای سیستم‌های پرسش پاسخ پیچیده روی مجموعه‌داده هات پات کیوای معرفی شده است. برای انتخاب پاراگراف‌های 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 پاراگراف‌های مرتبط که شواهدی از پاسخ در آن نمایان باشد انتخاب می‌شوند. برای این‌کار از روش استخراج کلمات کلیدی در پرسش استفاده شده است. نتایج بدست آمده نشان از بهتر بودن نتایج نسبت به روش پایه دارد.

کلید واژه‌ها: بازیابی اطلاعات، پرسش پاسخ پیچیده، انتخاب پاراگراف، شبکه‌های عصبی عمیق

۱- مقدمه

سیستم‌های پرسش‌پاسخ یک ابزار نوین برای پاسخدهی به پرسش‌های ساده و پیچیده کاربران به صورت خودکار مبتنی بر مجموعه‌داده‌های ساخت یافته، نیمه‌ساختاریافته و غیرساخت یافته هستند. این سیستم‌ها با دریافت درخواست‌ها و پرسش‌های متنوع کاربران که از ساختار زبان‌طبیعی پیروی می‌کنند نسبت به پاسخدهی به پرسش‌ها اقدام می‌نمایند [1].

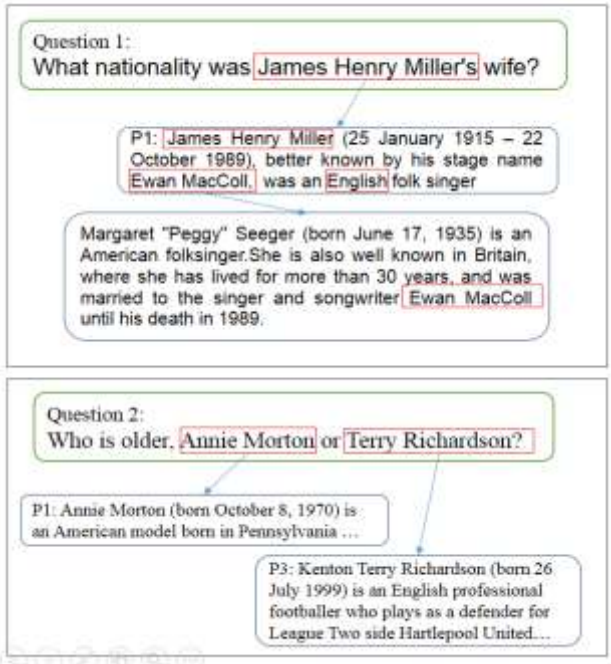
بسیاری از مطالعات اخیر در زمینه پرسش‌پاسخ روی تک پاراگراف‌ها تمرکز کرده‌اند [2]. امروزه پاسخدهی به بسیاری از پرسش‌ها کاربران در سیستم‌های پرسش و پاسخ مبتنی بر یک پاراگراف ممکن نیست. در حال حاضر پاسخدهی به پرسش‌های کاربران که غالبا پیچیده هستند نیاز به شواهد چندگانه دارد که این شواهد معمولا در جاهای مختلف متن یک یا چندین سند پراکنده شده‌اند. در سیستم‌های پرسش و پاسخ بازیابی و انتخاب پاراگراف‌های 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 سیستم‌های پرسش و پاسخ وجود دارد که غالبا

این دسته‌بندی‌ها براساس محدوده موضوعی، نوع پرسش و نوع پاسخ مورد پشتیبانی توسط این سیستم‌ها است. نوع سیستم پرسش و پاسخ مطرح‌شده در این مقاله از نوع دامنه‌باز است. در این نوع از سیستم‌های پرسش و پاسخ بر خلاف سیستم‌های پرسش و پاسخ دامنه بسته که تنها از یک موضوع خاص محتوایی به عنوان مثال پزشکی پشتیبانی می‌شود، محتوای پرسش و پاسخ به دسته خاصی تعلق نداشته و تنوع موضوعی دارد.

پرسش‌های مطرح‌شده در سیستم‌های پرسش و پاسخ نیز شامل دو نوع ساده و پیچیده است. منظور از پرسش‌های ساده غالبا پرسش‌های حقیقت‌محور است و پاسخ آن‌ها معمولا 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 معمولا 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 به عنوان مثال اگر یک پرسش پل قابل تجزیه به دو پرسش ساده باشد، پاسخ پرسش ساده اول بخشی از پرسش ساده دوم خواهد بود.

سیستم‌های پرسش و پاسخ برای بازیابی پاسخ در مواجه با پرسش‌های پیچیده یا به عبارت دقیق‌ترچندگامی، نیازمند بررسی حداقل دو پاراگراف متفاوت هستند [3]. این سیستم‌ها در ادامه برای پاسخدهی به پرسش نسبت به استدلال روی پاراگراف‌ها و پرش از یک پاراگراف به پاراگراف های دیگر در همان سند یا اسناد دیگر اقدام می‌نمایند که به این قابلیت استدلال چندگامی‌می‌گویند [4]. برای ارزیابی این نوع سیستم‌های پرسش و پاسخ مجموعه‌داده‌های متفاوتی ارائه شده است که مجموعه‌داده هات پات کیوای [3] از مجموعه‌داده‌های مطرح در این زمینه است. این مجموعه‌داده‌ای 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 مجموعه‌داده در دو دسته پرسش‌های پل و پرسش‌های مقایسه‌ای به صورت دامنه‌باز برچسب زده شده‌اند.

سازماندهی مقاله در ادامه به صورت زیر است. در بخش بعد کارهای مرتبط با پژوهش انجام شده در مقاله بررسی می‌گردد. در بخش سوم روش پیشنهادی، معماری و اجزای آن به صورت کامل تشریح می‌گردد. در بخش چهارم مجموعه‌داده اصلی و مجموعه‌داده های میانی ساخته شده از آن معرفی می گردد. در ادامه و در بخش پنجم روش پیشنهادی و مولفه‌های آن مورد آزمون قرارگرفته و نتایج آن گزارش می‌گردد. در انتها ضمن بیان نتیجه‌گیری، راجع به پژوهش انجام شده جمع‌بندی صورت می‌گیرد.



شکل ۱: نمونه‌ای از پرسش‌های پیچیده در مجموعه‌داده هات پات کیوای. پرسش ۱: پرسش پیچیده نوع پل. پرسش۲: پرسش پیچیده نوع مقایسه‌ای

۲- کارهای مرتبط

برای استدلال از نوع چندگامی در کار آقای کیو و همکاران [5] از یک ساختار گرافی مبتنی بر موجودیت‌ها برای شناسایی پاراگراف‌های مرتبط استفاده شده است. در این کار ساخت گراف از موجودیت‌های داخل پرسش آغاز می‌شود و تا رسیدن به جواب مناسب، پاراگراف‌های مرتبط بازیابی می‌شوند. در این پژوهش یک زیر شبکه آموزش داده می‌شود که پاراگراف‌های مرتبط را انتخاب کند. این زیر شبکه بر

اساس مدل از پیش‌آموزش‌داده شده برت [6] است. در این شبکه با استفاده از لایه سیگموید پرسش و پاراگراف دریافت‌شده و مرتبط بودن و نبودن برای هر پاراگراف با نمایش ۱ و ۰ در خروجی اعلام می‌گردد.

در پژوهشی دیگر آقای آسای و همکارن [7] 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 [8] و [9] جهت بالا بردن دقت بازیابی اطلاعات استفاده کرد.

بسیاری از مدل‌های قبلی بر اساس تکنیک «جستجوکن پاسخ رو پیدا کن» کار می‌کنند و نتایج خوبی را روی مجموعه‌داده‌های معروف نیز گرفته‌اند. اما برای پرسش‌های چندگزینه‌ای این روش‌ها خوب عمل نمی‌کنند. از این رو برای پرسش‌های چندگزینه‌ای آقای نی و همکاران [10] از یک روش مبتنی بر ادغام پرسش و پاسخ‌های موجود برای انتخاب پاراگراف‌های 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 به عنوان جواب تعیین می‌شود.

کارهای مرتبط انجام شده معمولا یا از روش بازیابی یک‌باره (تک‌گامی) 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 پاراگراف‌های مرتبط دوم با مشکل مواجه می‌شوند. در روش‌های بازیابی مکرر با وجود بکارگیری تطابق واژگانی دو مرحله‌ای بدلیل عدم امکان شناسایی گستره‌ی شامل پاسخ، بازیابی اسناد در مرحله دوم دچار رشد تصاعدی می‌گردد و یافتن سند و یا پاراگراف‌های 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 پرسش و پاسخ پیچیده پیشنهاد خواهد گردید.

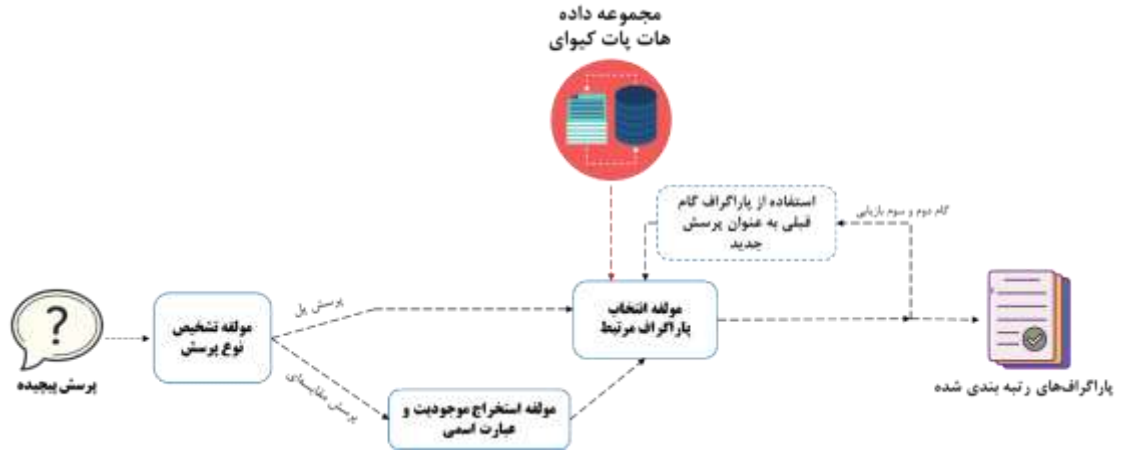
۳- روش پیشنهادی

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

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

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

سومین مولفه انتخاب پاراگراف‌های مرتبط است. در این مولفه از مدل زبانی برت استفاده می‌شود. به منظور شناسایی و انتخاب پاراگراف‌های مرتبط لایه‌های عصبی این مدل زبانی سفارشی شده و بر روی داده‌های مجموعه‌داده هات پات کیوای تنظیم دقیق گردیده است. این عمل برای رده‌بندی و بهبود عملکرد تشخیص پاراگراف‌های مرتبط لازم است. در پرسش‌های پل بازیابی اطلاعات به صورت ترتیبی و در چندگام انجام می‌شود که در ادامه به تفضیل راجع به آن بحث می‌گردد. معماری روش پیشنهادی در شکل ۲ نشان داده شده است.



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

۳-۱- تشخیص نوع پرسش

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

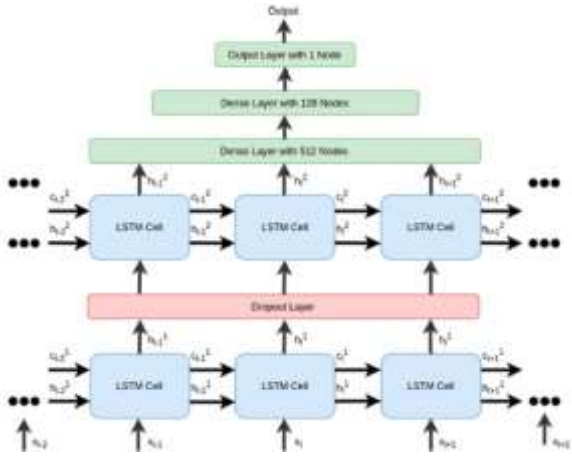
برای این‌کار از مجموعه‌داده هات پات کیوای که پرسش‌ها را در دو دسته پرسش‌های پل و مقایسه‌ای برچسب گذاری کرده است استفاده می‌شود.

برای این کار از یک شبکه عصبی LSTM [11] به همراه شبکه عصبی FFNN [12] جهت تشخیص نوع پرسش استفاده گردید. برای آموزش مدل از یک مجموعه‌داده شامل ۱۷۴۵۶ پرسش که در دو دسته مقایسه‌ای و پل برچسب گذاری شده است استفاده گردید. در جدول ۲ می‌توان نمونه‌ای از پرسش‌های برچسب گذاری 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 مولفه تشخیص نوع پرسش در شکل ۳ نشان داده شده است.



شکل ۳: معماری شبکه عصبی جهت تعیین نوع پرسش [13].

ورودی مدل آموزش دیده یک پرسش است که نمایش برداری آن $Q = [q1, q2, ..., qn]$ است و در آن q بیانگر هر توکن از پرسش است. طول بردار ورودی به صورت ۵۱۲ در نظر گرفته شده است و از معماری ارائه شده در [13] برای این کار استفاده شده است.

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

$$U = LSTM(Q) \in R^{n \times h} \quad (۱)$$

که در آن h بعد خروجی رمزنگار و n اندازه بردار خروجی است. بردار خروجی به یک شبکه عصبی FFNN [12] 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.

$$H_i = \sum W_{ij} X_{jk} + b_i \quad (۲)$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (۳)$$

$$Y = \text{Softmax}(U) \in R^k \quad (۴)$$

در معادله (۴) k خروجی دودویی و بیان‌کننده کلاس پرسش ورودی است. معماری این کار در شکل ۳ نشان داده شده است.

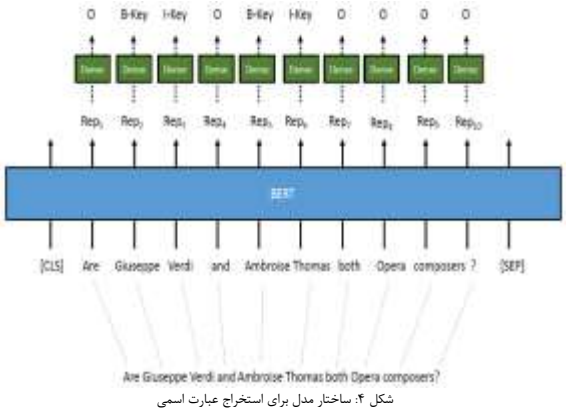
با روش بیان شده می‌توان نوع پرسش را مشخص کرده و آن را در یکی از دسته پرسش‌های پل و مقایسه‌ای رده‌بندی نمود. در نهایت با توجه به مشخص شدن نوع پرسش می‌توان استراتژی دقیق‌تری جهت انتخاب پاراگراف‌های مرتبط برگزید.

۳-۲- استخراج موجودیت و عبارت اسمی

در پژوهش‌های گذشته انتخاب پاراگراف‌های 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 استراتژی بکارگرفته شده جهت انتخاب پاراگراف‌های مرتبط متناسب با پرسش تعیین شده است. از این رو 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 غالبا بله و خیر است. از این رو استخراج موجودیت‌ها یا عبارات اسمی در انتخاب پاراگراف‌های مرتبط که شامل آن عبارت باشند کمک کننده خواهد بود. مدل‌های قبلی که روی استخراج موجودیت‌ها و عبارات اسمی کلیدی تمرکز داشتند به دلیل تعیین نوع موجودیت‌ها نسبت به مولفه 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 [6] در این مولفه 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 شکل ۴ به نمایش در آمده است. برای هر کلمه یا توکن خروجی از مدل برت یک رده‌بند جداگانه از جنس شبکه عصبی کاملا متصل متراکم [12] قرار داده شده است. این شبکه عصبی تمام بردارهای خروجی برای هر کلمه را 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.

$$F(q) = [t_1, t_2, t_3, ..., t_n] \quad (۵)$$

در این روش خروجی موجودیت یا عبارت اسمی است.

یک نمونه از استخراج موجودیت‌ها در جدول ۳ 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 مطرح‌شده از نوع پل لزوما موجودیت و یا عبارت اسمی مشخصی وجود ندارد، بهتر است از همه متن پرسش برای انتخاب پاراگراف مرتبط استفاده شود. به‌عنوان نمونه پرسش جدول ۲ 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 پاراگراف‌های مرتبط

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

- پرسش مقایسه‌ای

در صورتی که پرسش مطرح‌شده 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 آن‌که در این نوع پرسش‌ها عملا دو یا چند نوع موجودیت با هم مقایسه می‌شوند، با استفاده از پرسش می‌توان به صورت همزمان تمام پاراگراف‌های مرتبط را شناسایی کرد. در این نوع پرسش‌ها همان‌طور که گفته شد نیاز به بازیابی‌های ترتیبی پاراگراف‌ها نمی‌باشد.

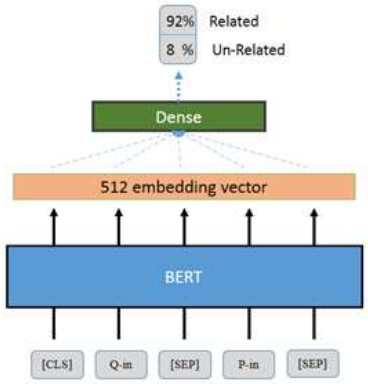
بعد از تشخیص نوع پرسش و استخراج عبارت اسمی داخل پرسش که شامل موجودیت‌ها نیز است عمل انتخاب پاراگراف‌های مرتبط باید صورت گیرد.

برای این‌کار عبارت اسمی استخراج شده در عنوان پاراگراف‌ها جسجو می‌شود. در صورتی که عبارت اسمی در عنوان پاراگراف‌ها به صورت تطابق کامل پیدا شود، آن پاراگراف به عنوان 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 به عنوان پاراگراف مرتبط انتخاب می‌شود. این کار برای تمام عبارات اسمی در متن پرسش انجام شده و تمام پاراگراف‌های مرتبط با استفاده از این روش انتخاب می‌شوند.

- پرسش پل

در صورتی که پرسش مطرح‌شده 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 اسامی موجودیت‌ها ذکر نمی‌شود. به طور کلی در این نوع پرسش‌ها سعی می‌شود که اطلاعات کلی در رابطه با بازیابی پاراگراف اول درج شود. لذا تنها با تاکید 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 ۵: انتخاب پاراگراف‌های مرتبط با مدل زبانی برت

برای تعیین پاراگراف‌ها همان‌طور که در شکل ۵ 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 قرار دادن «[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 $Q = \{q_1, q_2, ..., q_n\}$ و $P = \{p_1, p_2, ..., p_n\}$ 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.

با توجه به ساختار ارائه شده، پرسش و پاراگراف ابتدا وارد کدگشای مدل برت پایه شده، سپس بردار خروجی وارد یک لایه متراکم می‌شود. در مرحله آخر تابع فعال‌ساز سافت‌مکس در خروجی تعیین می‌کند که پرسش و پاراگراف دریافت‌شده به هم تا چه اندازه مرتبط هستند. با استفاده از این روش پاراگرافی که بیشترین امتیاز را داشته باشد به عنوان پاراگراف مرتبط اول انتخاب می‌شود.

در ادامه پاراگراف انتخاب‌شده اول به عنوان پرسش در نظر گرفته می‌شود و به پاراگراف‌های موجود با استفاده از جدا کننده «[SEP]» ادغام می‌شود. با استفاده از روش قبلی که در شکل ۵ نشان داده شده است پاراگرافی که بیشترین امتیاز را داشته باشد به عنوان پاراگراف مرتبط دوم انتخاب می‌شود.

۴- مجموعه‌داده‌ها

مجموعه‌داده هات پات کیوای مجموعه‌ای متشکل از ۱۱۳ هزار پرسش به همراه پاسخ است. این مجموعه‌داده در دو حالت فول ویکی و دیسترکتور آماده‌سازی شده است. در حالت فول‌ویکی پرسش‌ها و عناوین اسناد مرتبط به همراه پاسخ نهایی پرسش قرار داده شده و روش پیشنهادی باید از طریق بازیابی 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 مورد نظر با استدلال روی تعدادی از این پاراگراف‌ها قابل استنتاج است. در این روش انتخاب پاراگراف‌های مرتبط برای مدل استدلال مهم و کاربردی است. دسته‌بندی مجموعه‌داده هات پات کیوای مورد استفاده در روش پیشنهادی به صورت جدول ۴ است که از داده‌های آموزشی برای آموزش مدل و از داده‌های توسعه برای آزمایش مدل بهره گرفته شده است.

جدول ۴: دسته‌بندی مجموعه‌داده هات پات کیوای [3]

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 مولفه‌های مهم در این پژوهش است که نیازمند داده‌های آموزشی و آزمایشی جهت آموزش و ارزیابی است. ازاین‌رو از پرسش‌های مجموعه‌داده هات 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 به صورت جدول ۵ 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 مولفه استفاده شده در این پژوهش که نیازمند مجموعه‌داده‌ای جهت آموزش و ارزیابی مدل است، مولفه استخراج موجودیت و عبارت اسمی است. مجموعه‌داده مورد نظر باید روی هات پات کیوای ایجاد شده باشد تا بتواند عبارت اسمی را در این مجموعه‌داده به خوبی استخراج کند. به علت آن‌که مجموعه‌داده هات پات کیوای با استفاده از اسناد پایگاه دانش ویکی‌پدیا ساخته شده است این روش به خوبی 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 که از طریق مجموعه‌داده هات پات کیوای و صفحات ویکی‌پدیا ساخته شده است به‌صورت جدول ۶ است. همچنین نمونه‌ای از داده‌های مجموعه‌داده در جدول ۷ 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 بدست آمده توسط مولفه‌های مختلف روش پیسنهادی با پژوهش‌های مرتبط مقایسه می‌گردد. سپس نتایج نهایی بازیابی اطلاعات توسط روش پیشنهادی در مقایسه با روش پایه مورد بررسی قرار می‌گیرد.

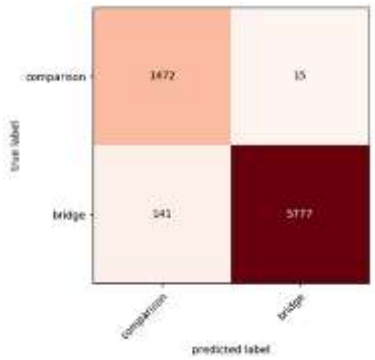
۵-۱- مولفه تشخیص نوع پرسش

همان‌طور که در بخش‌های قبل اشاره گردید، یکی از مولفه‌های مهم در وظیفه بازیابی اطلاعات روی مجموعه‌داده هات‌پات‌کیوای 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 مجموعه‌داده هات‌پات‌کیوای هیچ‌کدام از مولفه تشخیص پرسش به طور 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 مجموعه‌داده هات‌پات‌کیوای که شامل ۷۴۰۵ پرسش است انجام شده است. نتایج بدست آمده در جدول ۸ قابل مشاهده است.

جدول ۸: نتایج بدست آمده برای تشخیص نوع پرسش

	Precesion	Recall	F1
Comparison	99%	91%	95%
Bridge	98%	95%	96.5%



شکل ۶: ماتریس درهم ریختگی تشخیص نوع پرسش

در ماتریس درهم‌ریختگی شکل ۶ تعداد اشتباهات در تشخیص نشان داده شده است. در این ماتریس سطرها بیان‌کننده 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. در حالی که به صورت کلی نتایج خوب و رضایت بخش است.

۵-۲- تشخیص عبارت اسمی

دیگر مولفه مهم که در انتخاب پاراگراف‌های مرتبط مهم است، مولفه تشخیص عبارت اسمی است. این مولفه در پرسش‌های مقایسه‌ای اهمیت ویژه‌ای دارد. در بسیاری از کارهای مرتبط انجام شده برای استخراج موجودیت‌ها از Spacy [14] و یا Stanza [15] استفاده شده است. این مولفه‌ها 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 مجموعه‌داده هات‌پات‌کیوای و ویکی‌پدیا آموزش داده نشده‌اند و یا تنظیم دقیق نشده‌اند دقت خوبی نشان نداده‌اند و در بسیاری از موارد به‌خصوص موجودیت‌ها و عبارات اسمی بلند و چندکلمه‌ای ضعیف عمل می‌کنند. از آنجایی که استخراج عبارت اسمی یکی از راه‌های بهبود عملکرد بازیابی اطلاعات روی مجموعه‌داده هات‌پات‌کیوای 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 داده‌شده است.

مدل ارائه شده روی ۱۷۴۵۶ جمله پرسشی که هرکدام از آن‌ها حداقل ۱ موجود یا عبارت اسمی دارند آموزش داده شده است. همچنین داده آزمایشی ساخته شده، شامل ۱۴۸۵ جمله پرسشی 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 ۹ قابل مشاهده است.

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

همان‌طور که در جدول ۹ نشان داده شده است، دقت روی مدل‌های Spacy و Stanza 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 صرف‌نظر 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	۵۱٪	۸۰٪	۶۲.۲۹٪
Our Model	۹۹٪	۹۸٪	۹۸.۵٪

مدل‌های دیگری نظیر کی‌برت [16] 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 مجموعه‌داده هات‌پات‌کیوای برای استخراج موجودیت‌ها و عبارات اسمی از مقایسه آن با روش ارائه شده خودداری شده است.

۵-۳- ارزیابی

در این بخش به عنوان ارزیابی نهایی روش پیشنهادی، تمام مولفه‌های معماری آن در کنار هم قرار گرفته شده و نتایج انتخاب پاراگراف‌های مرتبط نشان داده شده است. برای آزمون روش پیشنهادی و مقایسه آن با روش پایه از حالت دیسترکتور مجموعه‌داده هات پات کیوای استفاده شده است. در این بخش برای هر پرسش دو پاراگراف به عنوان پاراگراف‌های مرتبط تعیین شده است که متناسب با ساختار مجموعه‌داده هات پات کیوای است. روش پیشنهادی با روش پایه HGN [17] مقایسه شده و نتایج مقایسه در جدول ۱۰ نشان داده شده است.

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 دیسترکتور و مقایسه آن با مدل‌های پایه

	Precision	Recall	F1-score
HGN [17]	94.53	94.53	94.53
Our Model	95.01	94.23	94.61

در روش HGN تمام پاراگراف‌ها به صورت 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 مجموعه‌داده هات‌پات‌کیوای که یک مجموعه‌داده انگلیسی بر اساس مقاله‌های ویکی‌پدیا است پرداخته شد. برای ارزیابی، مولفه‌های مختلف معماری روش پیشنهادی مورد بررسی قرار گرفت. نتایج نشانگر عملکرد بهتر هر مولفه نسبت به سایر روش‌های مشابه مورداستفاده در پژوهش‌های مرتبط گذشته بوده و این امر در نهایت منجر به بهبود نتایج کلی روش پیشنهادی گردیده است. در نهایت انتخاب پاراگراف‌های مرتبط با پرسش پیچیده مبتنی بر روش پیشنهادی روی مجموعه‌داده هات پات کیوای بر اساس دقت، فراخوانی و معیار اف وان با مدل پایه مقایسه گردید. نتایج بدست آمده 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 بکارگیری سازوکار متناسب با نوع سوال در بازیابی پاراگراف‌های مرتبط موثر است و منجر به کسب نتایج بهتر می‌گردد.

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 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 Retriever-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. Urgen 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.