# یک سیستم نوین پرسش و پاسخ مذهبی در زبان فارسی

یاسمن برشبان[1]، سید ابوالقاسم میرروشندل[2]
[1] کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشگاه گیلان،
[2] استادیار گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه گیلان، mirroshandel@guilan.ac.ir

## چکیده

*سیستم‌های پرسش و پاسخ، زیرشاخه‌ای از علوم پردازش زبان طبیعی و بازیابی اطلاعات محسوب می‌شوند که در چند دهه اخیر مورد علاقه زیاد محققین قرار گرفته‌اند. هدف سیستم‌های پرسش و پاسخ این است که تکنیک‌هایی فراتر از سیستم‌های بازیابی اطلاعات امروزی را توسعه دهند تا بتوانند پاسخ دقیق سوالات زبان طبیعی را بازیابی کنند. در این پژوهش، معماری پیشنهادی سیستم پرسش و پاسخ در زبان فارسی بر روی سوالات غیرحقیقت معرفی می‌شود. معماری پیشنهادی از سه مولفه پردازش سوال، بازیابی سند و رتبه‌بندی مجدد تشکیل شده است. در مولفه پردازش سوال، پیش پردازش‌های لازم بر روی سوال انجام شده و عنوان سوال تشخیص داده شده است. در مولفه بازیابی سند، موتورهای جستجوی منبع باز مورد استفاده قرار گرفته است و به منظور رتبه‌بندی مجدد، از روش‌های یادگیری ماشین استفاده شده است. نتایج نشان می‌دهد که سیستم پیشنهادی توانسته است به دقت ۸۲.۲۹ درصد و میانگین معکوس رتبه ۷۱.۸۸ درصد دست یابد. می‌توان اظهار کرد که سیستم پیشنهادی در نوع خود، اولین سیستم پرسش و پاسخ با چنین ویژگی‌هایی در زبان فارسی به خصوص در دامنه مذهبی است.*

## کلیدواژه

*سیستم پرسش و پاسخ، پردازش زبان طبیعی، بازیابی اطلاعات، میانگین معکوس رتبه*

## مقدمه

امروزه با افزایش اطلاعات در فضای اینترنت، کاربران باید زمان زیادی را صرف یافتن اطلاعات مدنظر خود کنند. برای تسهیل این امر، سیستم‌های پرسش و پاسخ[1] کلاسیک در قالب موتورهای جستجو ارائه شدند. در این سیستم‌های بازیابی اطلاعات کلاسیک، کاربران پرسش‌های خود را به صورت زبان طبیعی مطرح می‌کنند، سپس موتورهای جستجو لیستی از صفحات مرتبط با این سوال را برای کاربران برمی‌گردانند. مشکل این سیستم‌ها این است که کاربران باید زمان زیادی را صرف مطالعه و بررسی این صفحات کنند تا بهترین پاسخ مدنظر خود را بیابند [۲،۱].

در مقابل بازیابی اطلاعات کلاسیک که در آن کل سند به عنوان پاسخ برگردانده می‌شود، سیستم‌های پرسش و پاسخ معرفی شده‌اند. در سیستم‌های پرسش و پاسخ که شکل پیچیده‌تری از سیستم‌های بازیابی اطلاعات هستند، به جای بازگرداندن کل سند، بخش خاصی از اطلاعات که مدنظر کاربر است، به عنوان پاسخ برگردانده می‌شود. کاربران سیستم‌های پرسش و پاسخ، علاقمند به دریافت پاسخ مختصر، قابل فهم و صحیح هستند که این پاسخ ممکن است یک کلمه، جمله، پاراگراف، تصویر، قطعه صوتی و یا یک سند کامل باشد [۳].

به منظور طراحی یک سیستم پرسش و پاسخ، مشکلات و چالش‌های مختلفی وجود دارد. یک مشکل عمده این است که راه‌های مختلفی برای بیان یک سوال وجود دارد. این تغییرات در نحوه پرسش سوال می‌تواند ناشی از جابجایی ساده کلمات یا دانش متفاوت پرسشگر باشد. به طور مشابه پاسخ‌ها نیز می‌توانند به طور مختلف مطرح شوند. با توجه به اینکه سوال و جواب می‌توانند به شیوه‌های مختلف مطرح شوند؛ ممکن است تعداد کلمات مشترک میان سوال و جواب کم باشد در نتیجه این مساله تشخیص اسناد مرتبط با یک سوال را دشوار می‌کند. بدین ترتیب لازم است که به منظور تشخیص پاسخ مناسب، پردازش‌های سنگینی بر روی متن انجام شود [۱].

تاکنون در زمینه سیستم‌های پرسش و پاسخ غیرحقیقت[2]

[1] Question Answering System

[2] Non-factoid

تاریخ دریافت : ۱۳۹۵/۰۳/۳۱ تاریخ پذیرش : ۱۳۹۵/۰۹/۲۸

تحقیقات محدودی صورت گرفته است و طراحی سیستم پرسش و پاسخ غیرحقیقت حتی در زبان انگلیسی نیز یک چالش محسوب می‌شود.

طراحی یک سیستم پرسش و پاسخ در زبان فارسی در مقایسه با زبان انگلیسی بسیار پیچیده‌تر است. زیرا در زبان فارسی باید از پیش‌پردازش‌ها و ابزارهای پردازش زبان طبیعی و روش‌های خاص این زبان استفاده نمود. با توجه به اینکه در زبان فارسی در حوزه پردازش زبان طبیعی تحقیقات محدودی صورت گرفته و ابزارهای کمتری در اختیار محققین قرار دارد، این امر موجب می‌شود که طراحی سیستم در این زبان به مراتب سخت‌تر از زبان انگلیسی باشد.

در این مقاله، معماری پیشنهادی سیستم پرسش و پاسخ در زبان فارسی بر روی سوالات غیرحقیقت معرفی می‌شود. می‌توان ادعا کرد که این سیستم، اولین سیستم پرسش و پاسخ با چنین ویژگی‌هایی در زبان فارسی است. سیستم پیشنهادی از سه مولفه پردازش سوال، بازیابی سند و رتبه‌بندی مجدد تشکیل شده است. در مولفه پردازش سوال، پیش‌پردازش‌های لازم بر روی سوال انجام شده و عنوان سوال تشخیص داده می‌شود. در مولفه بازیابی سند، موتورهای جستجوی منبع باز مورد استفاده قرار گرفته است و در مولفه رتبه‌بندی مجدد روش‌های یادگیری ماشین استفاده شده است. به منظور بهبود فرآیند یادگیری، انواع ویژگی‌های ظاهری کلمات، جزء کلام، درخت پارس، تشخیص عنوان سوال و رتبه سند حاصل از مولفه بازیابی سند، به کار گرفته شده است و تاثیر هر یک از ویژگی‌ها بر روی مولفه رتبه‌بندی مجدد مورد ارزیابی قرار گرفته است. به منظور بهبود فرآیند رتبه‌بندی، هسته رتبه‌بندی و درختی نیز بر روی روش یادگیری ماشین اعمال شده است.

ساختار مقاله پیش رو به شرح زیر است: ابتدا به تشریح مفاهیم اولیه سیستم‌های پرسش و پاسخ می‌پردازیم و کارهای مرتبط انجام شده در زمینه سیستم‌های پرسش و پاسخ را بیان می‌کنیم. سپس منبع داده‌ای مورد استفاده معرفی شده و معماری پیشنهادی سیستم پرسش و پاسخ به تفصیل بیان می‌شود. در ادامه آزمایشات انجام شده و نتایج حاصل شده، شرح داده می‌شود و مشکلات و چالش‌های پیاده‌سازی معرفی خواهد شد و در انتها جمع‌بندی و کارهای آتی مطرح می‌شود.

## تعاریف

در این بخش، تعاریف و مفاهیم پرکاربرد در سیستم‌های پرسش و پاسخ بیان می‌شود.

### دسته‌بندی سیستم‌های پرسش و پاسخ از نظر دامنه

سیستم‌های پرسش‌وپاسخ از نظر دامنه به دو دسته دامنه باز[۳] و بسته[۴] تقسیم می‌شوند [۴]. سیستم‌هایی با دامنه باز یا نامحدود، باید انواع مختلف سوالاتی را که توسط کاربران در زمینه‌های مختلف مطرح می‌شوند، پوشش دهند. برای مثال تمامی زمینه‌های ورزشی، مذهبی، سیاسی و هر زمینه‌ی دیگری که ممکن است کاربران درمورد آن سوال بپرسند، باید توسط سیستم پوشش داده شود [۲].

سیستم‌هایی با دامنه بسته، تنها جوابگوی سوالات در یک زمینه‌ی خاص هستند. برای مثال فقط سوالات در زمینه‌ی پزشکی یا سوالات مذهبی را پاسخگو خواهند بود و غالبا بر روی یک سایت خاص و یا یک کتاب خاص، کار می‌کنند. عملکرد این سیستم‌ها در مقایسه با سیستم‌های دامنه باز، ساده‌تر است؛ زیرا سیستم‌های پردازش زبان طبیعی اغلب می‌توانند اطلاعات آن دامنه خاص را استخراج کنند و از اطلاعات و ویژگی‌های خاص دامنه، در فرایند یافتن پاسخ مناسب بهره ببرند. در این سیستم‌ها، اغلب انواع محدودی از سوالات که در آن زمینه‌ی مدنظر پرکاربرد هستند، پوشش داده خواهد شد [۲].

### انواع سوالات سیستم‌های پرسش و پاسخ

در یک سیستم پرسش و پاسخ، سوالات متنوعی می‌تواند توسط کاربران مطرح شود و برای پاسخ‌گویی به هر سوال، لازم است از روش‌ها و تکنیک‌های مناسب آن سوال، استفاده شود. دسته‌بندی‌های مختلفی بر روی سوالات صورت گرفته است اما دسته‌بندی معنایی که بیشتر موردتوجه واقع شده است، سوالات را به ۸ دسته حقیقت[۵]، لیست، تعریفی یا توصیفی[۶]، فرضیه‌ای[۷]، علیّتی[۸]، رابطه‌ای[۹]، رویّه‌ای[۱۰] و تاییدی[۱۱] تقسیم می‌کند [۲].

سوال حقیقت، معمولاً در زبان انگلیسی، با استفاده از کلمات پرسشی WH شروع می‌شود و حاوی کلمات پرسشی چه‌کسی، چه چیزی، چه وقت و کجا است. پاسخ این سوال، یک حقیقت یا واقعیت بیان شده در متن و غالبا یک موجودیت عددی و یا اسمی است. برای مثال "شماره تلفن دانشگاه گیلان چیست؟" که پاسخ آن یک موجودیت عددی است.

سوال لیست، سوالی است که پاسخ آن، لیستی از موجودیت‌های متن است. برای مثال "زکات بر چه چیزهایی واجب است؟ گندم،

---

[۳] Open Domain
[۴] Closed Domain
[۵] Factoid
[۶] Definition or description question
[۷] Hypothetical question
[۸] Causal question
[۹] Relationship question
[۱۰] Procedural question
[۱۱] Confirmation question

جو، خرما، کشمش، طلا، نقره، شتر، گاو و گوسفند. "

سوال تعریفی، به دنبال تعریف یک لغت موجود در صورت سوال است. برای مثال "تعریف روش‌های یادگیری ماشینی نیمه‌مربی چیست؟"

سوال فرضیه‌ای، به اطلاعاتی در مورد یک رویداد فرضی نیاز دارد. برای مثال "اگر دانشجویی در امتحانات پایان‌ترم غیبت کند، چه اتفاقی خواهد افتاد؟"

سوال علیّتی، جویای دانستن اطلاعات و توضیحی از یک رویداد است و معمولاً با چرا آغاز می‌شود. برای مثال "چرا مردم به بیماری قند مبتلا می‌شوند؟"

سوال رابطه‌ای، جویای ارتباط بین دو موجودیت است. برای مثال "ارتباط دانشگاه گیلان با شهر رشت چیست؟ دانشگاه گیلان در شهر رشت واقع است. "

سوال رویه‌ای، سوالی است که پاسخ آن، لیستی از دستورالعمل‌ها برای انجام عملیات ذکر شده در سوال است. برای مثال "مراحل گرفتن وضو چگونه است؟"

سوال تاییدی، برای رویداد مطرح شده درصورت سوال، به جواب بله یا خیر نیاز دارد. برای مثال "آیا پردازش زبان طبیعی، زیرشاخه‌ای از علم کامپیوتر است؟ بله"

از منظری دیگر، سوالات به دو دسته کلی حقیقت و غیرحقیقت تقسیم می‌شوند. سوالات حقیقت پیش‌تر معرفی شدند. سوالات غیرحقیقت[۱۲]، غالبا حاوی پاسخ‌های طولانی هستند و پاسخگویی به آن‌ها دشوارتر است. این سوالات در زبان فارسی معمولا حاوی کلمات پرسشی چرا و چگونه هستند [۲].

## معماری سیستم‌های پرسش و پاسخ

به طور کلی، سیستم‌های پرسش‌وپاسخ از سه مولفه اصلی پردازش سوال، بازیابی اطلاعات و استخراج پاسخ تشکیل می‌شوند [۱]. در مولفه پردازش سوال، کلیه‌ی پردازش‌های لازم بر روی سوال اعمال می‌شود و روابط ساختاری و معنایی موجود در کلمات سوال، استخراج می‌شود تا بتوان در مولفه‌های بعدی از این ویژگی‌ها به منظور یافتن پاسخ مناسب استفاده نمود. این مولفه شامل تولید پرس‌وجوی بازیابی اطلاعات[۱۳] و تشخیص نوع پاسخ مورد انتظار[۱۴] است [۱].

معمولاً به منظور تولید پرس‌وجوی مناسب، از شیوه کیسه‌ای از کلمات[۱۵] استفاده می‌شود. به‌این‌صورت که کلیه کلمات موجود در سوال بدون در نظر گرفتن ترتیب قرارگیری، در ساخت پرس‌وجو استفاده می‌شوند [۱]. به منظور تولید پرس‌وجو، پیش‌پردازش‌هایی بر روی سوال انجام می‌شود که می‌توان به حذف کلمات توقف[۱۶] و ریشه‌یابی[۱۷] اشاره کرد. یک مرحله بسیار مهم در مولفه پردازش سوال، تشخیص دسته معنایی پاسخ مورد انتظار است. هدف، تشخیص نوع موجودیت نامداری است که می‌تواند پاسخ یک سوال باشد. برای این منظور، استفاده از روش‌های یادگیری ماشین [۵]بسیار متداول است. برای مثال، ممکن است که در یک سوال به دنبال یک شخص، نام یک شرکت یا یک عدد باشیم [۱].

پس از اعمال پردازش‌های اولیه بر روی سوال، مولفه‌ی دومی که در سیستم‌های پرسش و پاسخ به کار می‌رود، بازیابی اطلاعات است. این مولفه، به عنوان ورودی یک پرس‌وجو و مجموعه‌ای از اسناد را دریافت می‌کند. سپس توسط یک تابع، میزان ارتباط پرس‌وجو با مجموعه‌ی اسناد دریافتی را محاسبه کرده و بر این اساس به هر سند، یک رتبه تعلق می‌گیرد. به این‌ترتیب اسناد بازیابی‌شده، در لیستی براساس امتیازشان مرتب می‌شوند [۶].

در یک سیستم پرسش و پاسخ، مولفه‌ی بازیابی سند وظیفه‌ی فیلترکردن اسناد را بر عهده دارد. به بیان دیگر در کوتاه‌ترین و قابل قبول‌ترین زمان ممکن، اسناد مرتبط به یک پرسش را بازیابی کرده و اسناد نامرتبط را فیلتر می‌کند. معمولاً توابعی که در بخش بازیابی سند به کار می‌روند، نسبت به توابع به کار رفته در بخش پردازش پاسخ، کم‌هزینه‌تر هستند تا این مرحله در سریع‌ترین زمان ممکن انجام شود [۱]. مولفه‌ی بازیابی اطلاعات در سیستم پرسش و پاسخ، از اهمیت ویژه‌ای برخوردار است، زیرا گام‌های بعدی، تنها در صورتی که نتایج بازگردانده شده توسط مولفه‌ی بازیابی اطلاعات صحیح و قابل‌اعتماد باشند، می‌توانند پاسخ صحیحی برای پرسش بیابند [۳،۱].

مولفه‌ی سوم، استخراج پاسخ است که در این بخش، از میان اطلاعاتی که توسط مولفه‌ی دوم برگردانده شده، صحیح‌ترین و خلاصه‌ترین پاسخ، به عنوان جواب نهایی به کاربر برگردانده می‌شود. معمولاً در این مرحله به منظور پاسخ‌گویی به سوالات غیرحقیقت از روش‌ها و الگوریتم‌های پیچیده‌تری استفاده می‌شود و به هر یک از پاسخ‌های برگردانده شده توسط مولفه‌ی بازیابی اطلاعات، مجدد رتبه‌ای انتساب داده می‌شود و سرانجام بهترین پاسخ، برگردانده می‌شود [۷].

## معیارهای ارزیابی

به منظور ارزیابی سیستم‌های پرسش و پاسخ، معیارهای ارزیابی متفاوتی مورد استفاده قرار می‌گیرند. ابتدا فرض کنید که متغیرهای مورد استفاده به شرح زیر باشند:

D: مجموعه اسناد و یا قطعه متن‌ها است.

---

[۱۲] Non-factoid question

[۱۳] Producing an IR Query

[۱۴] Expected Answer Type Detection

[۱۵] Bag of Words (BOW)

[۱۶] Stop words

[۱۷] Stemming

$A_{D,q}$: زیرمجموعه ای از D است که حاوی اسناد مرتبط برای سوال q باشد.

$R_{D,q,n}$: n سند یا قطعه متن برتری است که از مجموعه اسناد D بازیابی شده است.

معیار اولیه‌ای که در سیستم‌های پرسش و پاسخ استفاده می‌شود، میانگین معکوس رتبه[۱۸](MRR) است. این معیار برای رتبه‌بندی سیستم‌هایی مناسب است که چندین پاسخ را برای یک سوال برمی‌گردانند. Q یک مجموعه سوال و $r_i$ رتبه اولین پاسخ درست برای سوال i است که اگر هیچ پاسخ درستی برگردانده نشود، مقدارش برابر با صفر خواهد بود [۲،۱].

$$MRR = \frac{\sum_{i=0}^{|Q|} \frac{1}{r_i}}{|Q|} \quad (۱)$$

هرچه میزان MRR به یک نزدیکتر باشد، عملکرد سیستم قابل‌قبول‌تر است. زیرا نشان‌دهنده‌ی این است که پاسخ صحیح به ابتدای لیست بازیابی‌شده نزدیکتر بوده است؛ در نتیجه کاربر سریع‌تر به پاسخ مورد نظر خود دست می‌یابد [۲]. به عنوان نقطه ضعف این معیار می‌توان به این مورد اشاره کرد که این معیار هیچ اهمیتی به بازیابی چندین پاسخ درست نمی‌دهد و در سیستم‌هایی که برای یک سوال چندین پاسخ درست برگردانده می‌شود، این معیار چندان مفید نخواهد بود [۱].

معیار ساده دیگری که استفاده می‌شود، صحت[۱۹] است. این معیار، درصد سوالاتی را نشان می‌دهد که در میان n پاسخ برگردانده شده، حداقل دارای یک پاسخ درست هستند [۲،۱].

$$accuracy^S(Q,D,n) = \frac{|\{q \epsilon Q | F^S_{D,q,n} \cap A_{D,q} \neq \emptyset\}|}{|Q|} \quad (۲)$$

معمولاً در ارزیابی‌های سیستم از نماد a@n استفاده می‌شود که نشان‌دهنده صحت در n سند اول است.

دو معیار ارزیابی استاندارد برای سیستم‌های بازیابی اطلاعات، معیار دقت[۲۰] و بازخوانی[۲۱] است. معیار بازخوانی برای یک سیستم بازیابی S که برای سوال q، n پاسخ برتر را برمی‌گرداند، نشان‌دهنده این است که سیستم توانسته چه میزان از اسناد مرتبط برای سوال q، چه تعدادی را بازیابی کرده است. لذا بازخوانی یعنی از میان کل پاسخ‌های صحیح موجود در مجموعه‌ی اسناد، چند درصدشان یافت شده‌اند.

$$recall^S(D,q,n) = \frac{|R^S_{D,q,n} \cap A_{D,q}|}{|A_{D,q}|} \quad (۳)$$

معیار دقت برای یک سیستم بازیابی S که برای سوال q، n پاسخ برتر را برمی‌گرداند، نشان‌دهنده این است که از میان اسنادی که سیستم به عنوان پاسخ بازیابی کرده است، چه تعداد درست است. لذا دقت یعنی چند درصد پاسخ‌های یافت‌شده، صحیح هستند [۱].

$$Precision^S(D,q,n) = \frac{|R^S_{D,q,n} \cap A_{D,q}|}{|R^S_{D,q,n}|} \quad (۴)$$

هرچه میزان دقت بیشتر باشد، به این معناست که تعداد اسناد مرتبط بازیابی‌شده بیشتر از تعداد اسناد نامرتبطی است که بازیابی شده‌اند. بالا بودن دقت یکی از اهداف مورد نظر طراحان سیستم‌های بازیابی اطلاعات است. از سوی دیگر، هرچه مقدار بازخوانی سیستم بیشتر باشد یعنی سیستم بیشترِ نتایج مرتبط را بازیابی کرده است. دستیابی به مقدار فراخوانی بالا نیز بسیار حائز اهمیت است. اما مطالعات نشان می‌دهد که فراخوانی و دقت، لزوما همزمان افزایش نمی‌یابند و گاهی اوقات با افزایش یکی، دیگری کاهش می‌یابد. لذا در نظرگرفتن تعادلی بین این دو معیار، براساس اهداف سیستم بسیار ضروری است.

با توجه به ارتباطی که بین دقت و بازخوانی وجود دارد، معیار F معرفی شده که از ترکیب معیار دقت و بازخوانی تشکیل شده است. ممکن است از دیدگاه کاربر، دقت سیستم، $\beta$ برابر اهمیت بیشتری نسبت به بازخوانی سیستم داشته باشد. لذا اگر به معیار دقت، وزن یا ارزش $\beta$ داده شود آنگاه معیار F[۲۲] به‌صورت زیر تعریف می‌شود.

$$F = \frac{(\beta^2+1) \times Precision \times recall}{\beta^2 \times Precision + recall} \quad (۵)$$

## کارهای مرتبط

کارهای مرتبط انجام شده در زمینه سیستم‌های پرسش و پاسخ را می‌توان به ۴ بخش شامل توسعه پیکره، پردازش سوال، بازیابی سند یا قطعه‌متن[۲۳] و رتبه‌بندی مجدد[۲۴] یا استخراج پاسخ تقسیم کرد. در این بخش، پژوهش‌های انجام شده در هر یک از این بخش‌ها به تفصیل بیان خواهد شد.

### توسعه پیکره

اغلب پژوهش‌های صورت گرفته در رابطه با توسعه پیکره پرسش و پاسخ در زبان انگلیسی بوده است، در صورتی که در زبان‌های دیگر مانند فارسی، نیاز شدیدی به وجود چنین پیکره‌هایی

[۱۸] mean reciprocal rank
[۱۹] accuracy
[۲۰] precision
[۲۱] recall
[۲۲] F-measure
[۲۳] Passage Retrieval
[۲۴] Reranking

احساس می‌شود. از جمله پیکره‌های پرسش و پاسخ انگلیسی می‌توان به پیکره TREC اشاره کرد که توسط کنفرانس سالانه‌ی TREC[۲۵] ارایه می‌شود [۸]و این پیکره در اختیار پژوهشگرانی که که در این کنفرانس شرکت می‌کنند، قرار داده می‌شود. در سال ۲۰۰۰ انجمن ارزیابی بین‌زبانی CLEF سیستم‌های پرسش و پاسخ بین‌زبانی را توسعه داد. سیستم‌هایی که درآن زبان سوال با زبان اسناد موجود در مخزن اطلاعات متفاوت است [۹].

پیکره پرکاربرد دیگر، پیکره سوال و جواب مقالات ویکی پدیا است [۱۰]. این پیکره، حاوی سوالات حقیقت استخراج شده از مقالات ویکی‌پدیا، پاسخ سوال، درجه سختی سوال از نظر پرسشگر و پاسخ‌دهنده است که برای استفاده، در دسترس عموم قرار دارد. پیکره‌ای دیگر برای مولفه دسته‌بندی سوال[۲۶] ارائه شده است که حاوی ۱۴،۵۰۰ سوال به همراه برچسب است و برای هر سوال، نوع پاسخ مورد انتظار در سطح درشت‌دانه[۲۷] و ریزدانه[۲۸] ذخیره شده است [۱۲،۱۱]. علاوه بر این پیکره‌ای حاوی ۷۰،۰۰۰ نمونه سوال و جواب موجود است که پاسخ‌ها را در سطوح مختلف جمله، پاراگراف و سند نگهداری می‌کند [۱۲].

در زبان فارسی، پیکره‌ای حاوی ۵،۰۰۰ سوال به منظور دسته‌بندی سوال، ارائه شده است که شامل سوالاتی از مجموعه کتاب‌های درسی است و برای هر سوال، نوع پاسخ مورد انتظار در سطح درشت‌دانه و ریزدانه نگهداری شده است [۱۳]. همچنین پیکره‌ای به منظور مشخص کردن موضوع[۲۹] موجود است که این پیکره حاوی ۱۱۸ سوال است [۱۴]؛ اما این تعداد سوال برای ارزیابی در سیستم‌های پرسش و پاسخ مناسب نیست. در زبان فارسی پیکره‌ای که بتواند برای کلیه مولفه‌های سیستم‌های پرسش و پاسخ، مورد استفاده قرار گیرد، وجود ندارد. از دیگر کارهای مرتبط پرسش و پاسخ در زبان فارسی، می‌توان به پژوهشی اشاره کرد که خاص زندگی‌نامه بوده ولی در دسترس عموم قرار ندارد [۱۵]. اخیرا پیکره و سیستمی در حوزه قرآن در زبان فارسی ارایه شده است[۳۰] اما در مورد جزییات پیاده سازی سیستم، منابع قابل استنادی در دسترس نیست.

## پردازش سوال

پژوهش‌های انجام شده در این زمینه را می‌توان به دو دسته کلی تعیین نوع پاسخ مورد انتظار و ساخت پرس‌وجوی بازیابی اطلاعات تقسیم نمود.

برای تعیین نوع پاسخ مورد انتظار، لازم است که یک مجموعه از پیش تعریف شده از نوع پاسخ مورد انتظار، گردآوری شده باشد تا بتواند به سوالات جدید نسبت داده شود [۱]. در سال‌های اخیر، دسته‌بندی‌های مختلفی بر روی نوع پاسخ مورد انتظار انجام شده است. یک دسته‌بندی بر اساس وردنت[۳۱] ارایه شده که شامل ۱۸ دسته سطح بالا و ۱۵ دسته سطح پایین است [۱۶]. یک دسته‌بندی دو سطحی دیگر شامل ۹۰ گره[۳۲] نیز ارایه شده است و مورد استفاده قرار می‌گیرد [۱۷]. معروف‌ترین این دسته‌بندی‌ها که بسیار مورد توجه پژوهشگران قرارگرفته‌است، یک دسته‌بندی سلسله‌مراتبی ۲ سطحی بوده که شامل ۶ کلاس درشت‌دانه و ۵۰ کلاس ریزدانه است که هر کلاس درشت‌دانه از مجموعه‌ای از کلاس‌های ریزدانه تشکیل شده است [۱۱].

به منظور تشخیص نوع پاسخ مورد انتظار، دو راهکار کلی استفاده از قواعد[۳۳] و روش‌های یادگیری ماشین با مربی[۳۴] مورد استفاده قرار می‌گیرند. به منظور استفاده از قواعد، عموماً از گرامر و دستورات زبان استفاده می‌شود. ویژگی‌هایی که معمولاً در این قوانین لحاظ می‌شوند شامل: عناصر واژگانی مانند کلمات یا عبارات و اطلاعات نحوی حاصل از جزءکلام و تجزیه‌کننده است. تعدادی از سیستم‌هایی که از این شیوه استفاده می‌کنند [۱۹،۱۸]، یک مجموعه‌ی عظیمی از عبارات منظم را ایجاد می‌کنند که هر سوال را به نوع پاسخ مورد انتظار نگاشت می‌دهند. در این روش اغلب نیاز است که تعداد زیادی از سوال‌ها بررسی شوند. همچنین نتیجه حاصل شده وابسته به متن و دامنه است و با تغییر دامنه، نتیجه نیز تغییر خواهد کرد.

علاوه بر این یک دسته‌بندی‌کننده سوال با استفاده از ۶۰ قاعده از پیش تعریف شده، معرفی شده است و برای مواردی که نتوان با استفاده مستقیم از قواعد، سوال را دسته‌بندی نمود، سرواژه[۳۵] سوال استخراج می‌شود. سپس کلمه سرواژه با استفاده از کلماتی با معنای عام‌تر[۳۶] موجود در شبکه واژگانی وردنت، توسعه داده می‌شود تا با استفاده از توسعه سرواژه از طریق وردنت، نتیجه حاصل منطبق با قواعد تهیه شده باشد و بتوان نوع پاسخ این سوال را استخراج کرد [۱۸].

یک راه حل جایگزین برای روش‌های مبتنی بر قاعده، ساخت دسته‌بندی‌کننده با استفاده از داده‌های آموزشی است که غالبا این داده‌های آموزشی به صورت دستی برچسب‌گذاری می‌شوند. از جمله روش‌های یادگیری ماشین که به منظور دسته‌بندی کردن سوال مورد استفاد قرار گرفته است می‌توان به الگوریتم بیزساده[۳۷] (NN)، درخت تصمیم[۳۸] (DT) و ماشین بردار

[۲۵] http://trec.nist.gov/
[۲۶] Question Classification
[۲۷] Coarse-grained
[۲۸] Fine-grained
[۲۹] Topic Detection
[۳۰] http://quranjooy.itrc.ac.ir/
[۳۱] Wordnet
[۳۲] node
[۳۳] Rules
[۳۴] Supervised machine learning
[۳۵] headword
[۳۶] Hypernym
[۳۷] Naïve Bayes
[۳۸] Decision Tree

پشتیبان[۳۹] (SVM) اشاره کرد [۱]. ویژگی‌های مورد استفاده در روش‌های یادگیری اغلب شامل: کلمات، جز کلام، قطعات[۴۰] متن (عباراتی که با هم اشتراک نداشته باشند)، موجودیت‌های نامدار، قسمت اصلی قطعه (عموماً اولین جزء اسم در جمله)، کلمات مرتبط معنایی و ساختارهای نحوی است [۲۰].

به منظور توسعه پرس و جو، کلمات سوال را با استفاده از مترادف‌های[۴۱] آن در وردنت توسعه می‌دهند. اما توسعه تمام کلمات سوال با استفاده از وردنت باعث می‌شود که دقت سیستم کاهش و بازخوانی افزایش یابد؛ زیرا تعداد اسناد غیرمرتبط بیشتری بازیابی می‌شود. بدین جهت با توجه به ویژگی‌های سیستم و دامنه پرسش و پاسخ مدنظر، فقط کلمات خاصی از سوال با استفاده از وردنت توسعه داده می‌شوند [۱].

در روشی دیگر ابتدا یک جدول حاوی کلمات کلیدی متناسب با انواع پاسخ‌های مورد انتظار تعریف می‌شود. پس از تشخیص نوع پاسخ مورد انتظار یک سوال، آن سوال با استفاده از کلمات کلیدی پرکاربرد متناسب با آن نوع پاسخ که در جدول نگهداری شده است، توسعه داده می‌شود. این روش می‌تواند برای سیستم‌هایی با دامنه بسته به کار گرفته شود ولی برای اعمال بر روی سیستم‌های دامنه باز مناسب نیست؛ زیرا در یک دامنه نامحدود تهیه لیستی از کلمات کلیدی متناسب با انواع سوالات، میسر نیست [۱].

در روشی دیگر پرس و جو با استفاده از بازخورد[۴۲] توسعه داده می‌شود. ایده استفاده از روش‌های بازخورد بدین صورت است که ابتدا تعدادی سند اولیه به کاربر برگردانده می‌شود، سپس کاربر هر یک از اسناد برگردانده شده را بررسی کرده و اسناد مناسب و غیرمناسب را نشانه‌گذاری می‌کند و بدین صورت بازخورد کاربر به سیستم داده می‌شود. سپس توسعه پرس‌وجو به شکلی ادامه می‌یابد که تشابه سوال با اسناد مرتبط تعیین‌شده، افزایش و با اسناد غیرمرتبط تعیین‌شده، کاهش یابد. به طورکلی استفاده از روش‌های بازخورد در توسعه پرس‌وجو، نتایج مناسبی را به دنبال داشته است [۲۱].

همان طور که قبلا اشاره شد، وردنت به تنهایی نمی‌تواند در توسعه پرس‌وجو به کار گرفته شود؛ زیرا موجب می‌شود که کلمات نامرتبطی توسعه داده شود و در نتیجه تعداد اسناد بی‌ربط بازیابی شده افزایش می‌یابد. یک راه‌حل، استفاده از فرهنگ لغت[۴۳]‌هایی است که به صورت اتوماتیک ساخته می‌شوند تا در کنار وردنت بتوانند به کار گرفته شوند و کیفیت نتیجه را افزایش دهند [۲۱]. برای این منظور، دو فرهنگ لغت کلمات هم-رخداد[۴۴] و نهاد-گزاره[۴۵] تعریف شده است [۲۲]. ایده کلی استفاده از فرهنگ لغت کلمات هم-رخداد این است که از اطلاعات مربوط به کلماتی که با هم رخ می‌دهند برای ساخت فرهنگ لغت استفاده شود؛ زیرا به احتمال زیاد کلماتی که با هم تکرار می‌شوند، مشابه و مرتبط هستند. در فرهنگ لغت نهاد-گزاره یک فرهنگ لغت یا منبع داده بر اساس ساختار نهاد-گزاره ساخته می‌شود. ایده این روش این است که ساختارهای نحوی می‌توانند نشان‌دهنده کلمات مرتبط باشند. در واقع با هر فعل خاص، تنها کلمات خاصی می‌تواند به کار گرفته شود. همچنین هر اسم می‌تواند بر اساس صفت یا قیدی که معمولاً به همراه آن تکرار می‌شود، مشخص وگروه‌بندی شود. بدین منظور یک منبع داده بر اساس ساختار گرامری ساخته می‌شود. استفاده از این دو فرهنگ لغت در کنار وردنت باعث بهبود نتایج شده است [۲۲].

## بازیابی سند و قطعه‌متن

برای بازیابی سند در دامنه بسته، عموماً موتورهای جستجوی منبع باز به منظور نمایه‌گذاری[۴۶] و بازیابی مورد استفاده قرار می‌گیرند. در سال‌های اخیر مطالعات جامعی در زمینه موتورهای جستجوی منبع باز صورت گرفته است و مزایا و معایب موتورهای جستجوی پرکاربرد مطرح شده است [۲۳]. علاوه بر این دو موتور پرکاربرد لوسین و ایندری- لمور[۴۷] از نظر نمایه‌گذاری، ارزیابی پرس‌وجو و کارایی بازیابی، به طور دقیق مورد بررسی قرار گرفته‌اند که نتایج نشان می‌دهد موتور جستجوی ایندری بهتر از لوسین عمل می‌کند [۲۴]. به منظور مقایسه‌ای دیگر در زمینه موتورهای جستجو، تحقیقات نشان داده است که در سیستم‌های پرسش و پاسخ، موتورهای جستجوی بولین مانند لوسین عملکرد بهتری خواهند داشت [۲۵]. تحقیقات نشان می‌دهد که انجام ریشه‌یابی در هنگام نمایه‌گذاری توانایی موتور جستجو را برای بازیابی اسناد مرتبط کاهش می‌دهد [۲۶]. از این رو موتور جستجو باید توسعه داده شود تا تمام انواع مورفولوژیکی یک کلمه را شامل شود. در پژوهشی دیگر، سند به تعدادی قطعه متن تقسیم شده و الگوریتم‌هایی به منظور بازیابی قطعه متن ارایه شده است [۲۵].

## رتبه‌بندی مجدد

پس از مولفه بازیابی سند، لازم است که اسناد بازیابی شده با استفاده از روش‌های پیچیده‌تر و اعمال ویژگی‌های مناسب، مجدد رتبه‌بندی شوند تا پاسخ دقیق‌تری برای کاربر برگردانده

[۳۹] Support Vector Machine
[۴۰] Chunk
[۴۱] Synonym
[۴۲] feedback
[۴۳] thesaurus
[۴۴] Co-occurrence based thesaurus
[۴۵] Predicate-argument thesaurus
[۴۶] Indexing
[۴۷] "http://www.lemurproject.org/."

شود [۷]. برای این منظور استفاده از ویژگی‌های موثر زبانی به منظور بهبود رتبه‌بندی پاسخ‌ها برای سوالات غیرحقیقت مورد ارزیابی قرار گرفته است.

ویژگی‌های شباهت[۴۸]: در این ویژگی‌ها شباهت ظاهری بین سوال Q و پاسخ A محاسبه می‌شود.

مدل‌های ترجمه: نقطه ضعف مدل‌های مبتنی بر شباهت این است که این مدل‌ها نمی‌توانند شکاف لغوی بین سوال و جواب را مدیریت کنند؛ زیرا کلمات به کار رفته در سوال و جواب دقیقا مشابه همدیگر نیستند. برای رفع این مشکل، می‌توان تبدیلات سوال و جواب را با استفاده از مدل ترجمه یاد گرفت.

ویژگی‌های تراکم و تکرار: این ویژگی‌ها، تکرار و تراکم کلمات سوال را در میان پاسخ اندازه‌گیری می‌کنند.

ترکیب ویژگی‌های مطرح شده در دو روش یادگیری پرسپترون و SVM به کار گرفته شده است و نتایج نشان می‌دهد که ویژگی‌های ترجمه در مقایسه با سایر ویژگی‌های مطرح شده نقش بسزایی در بهبود رتبه‌بندی پاسخ‌ها داشته‌اند [۷]. به منظور اعمال روش‌های یادگیری ماشین بر روی مولفه رتبه‌بندی در سیستم‌های پرسش و پاسخ، دو روش رگرسیون و SVM نتیجه بهتری در پی داشته‌اند [۲۷].

علاوه بر ویژگی‌های ذکر شده، ویژگی‌هایی به منظور رتبه‌بندی مجدد بر روی سوالات Why نیز معرفی شده است. برای این منظور می‌توان به ویژگی‌های ساختار نحوی سوال، ساختار معنایی سوال و مترادف‌های سوال اشاره کرد [۲۸].

به منظور یادگیری رتبه‌بندی در سیستم‌های پرسش و پاسخ بر روی سوالات حقیقت، ویژگی‌های کلمات کلیدی، موجودیت‌های نامدار و برچسب نقش معنایی[۴۹] استفاده شده است. نتایج نشان داده است که ویژگی‌های معنایی به نسبت روش پایه‌ای استفاده از کلمات کلیدی، نتایج بهتری را به دنبال داشته است [۲۹]. در پژوهشی دیگر، یک معماری کلی برای سیستم‌های پرسش و پاسخ ارایه شده است که به منظور رتبه‌بندی مجدد اسناد، الگوریتم یادگیری ماشین SVM به همراه ویژگی‌های جزء کلام و موجودیت‌های نامدار به کار گرفته شده است [۳۰].

## منبع داده

در سال‌های اخیر، در زمینه توسعه پیکره در زبان انگلیسی تحقیقات متعددی صورت گرفته است و پیکره‌های مناسبی در این زبان در اختیار محققین قرار گرفته است اما در زبان فارسی منابع محدودی وجود دارد. بدین جهت ما پیکره "رسائل و مسائل" را در زمینه سوالات و احکام دینی در زبان فارسی ارایه کرده‌ایم که در این پژوهش به کار گرفته شده است [۳۱]. این پیکره شامل ۲,۱۱۸ سوال غیرحقیقت و ۲,۰۵۱ سوال حقیقت بوده که برای هر سوال، متن سوال، نوع سوال، سختی سوال از نظر پرسشگر و پاسخ‌دهنده، دسته معنایی پاسخ در سطح درشت‌دانه و ریزدانه، پاسخ دقیق سوال، شماره صفحه و پاراگرافی که پاسخ از آن استخراج شده و درجه‌ی ارتباط پاسخ استخراج شده با سوال در دو سطح ارتباط کامل (با مقدار ۲) و ارتباط جزئی (با مقدار ۱) نشانه‌گذاری شده است و در دسترس عموم قرار دارد[۵۰].

## روش پیشنهادی

شکل (۱) نمای کلی معماری پیشنهادی را نشان می‌دهد که دارای ۳ مرحله پردازش سوال، بازیابی سند و رتبه‌بندی مجدد است.

### مرحله اول، پردازش سوال

در مرحله نخست سوالی که از کاربر دریافت می‌شود، توسط مولفه پردازش سوال، مورد بررسی قرار می‌گیرد. اقداماتی که در مولفه پردازش سوال انجام می‌شود به دو بخش پیش‌پردازش و تشخیص موضوع سوال تقسیم می‌شوند.

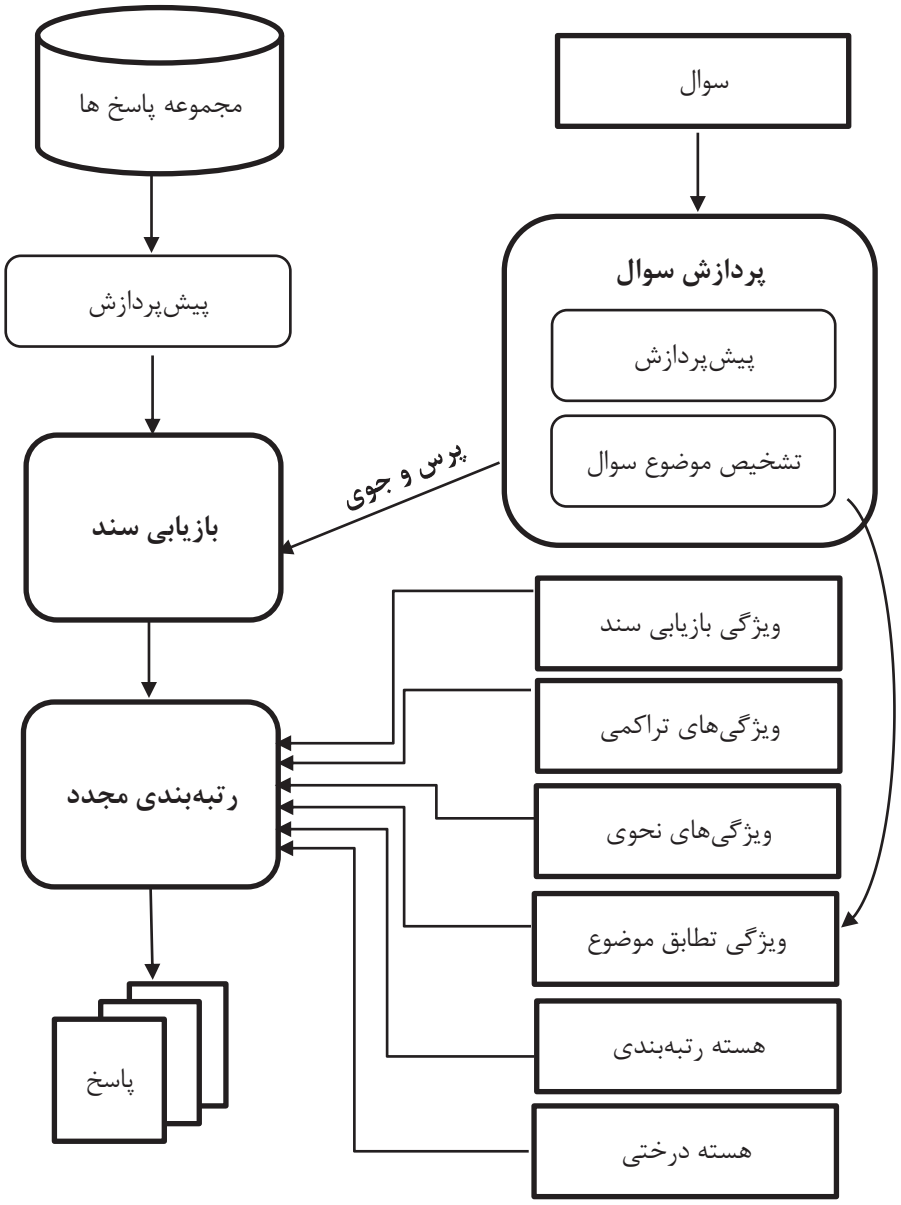
### پیش‌پردازش

در این بخش ابتدا متن ورودی نرمال[۵۱] و تمیز می‌شود. یکی از مهم‌ترین اقداماتی که در نرمال‌سازی انجام می‌شود، اصلاح نیم‌فاصله‌های متن است. برای مثال کلماتی مانند "می شود" و "آن ها"، به ترتیب به شکل "می‌شود" و "آن‌ها" تبدیل می‌شوند. بعد از اعمال نرمال‌سازی، عمل جداسازی جملات[۵۲] بر روی متن نرمال‌شده، انجام می‌گیرد تا متن نرمال شده به تعدادی جمله تبدیل شود. سپس بر روی هر یک از جملات، جداسازی کلمات[۵۳] انجام می‌شود. یکی دیگر از پیش‌پردازش‌هایی که در این مرحله انجام شده، ریشه‌یابی است تا هر کلمه به شکل اصلی و ریشه آن تبدیل شود مثلا کلمه "کتاب‌ها" به "کتاب" تبدیل می‌شود.

به منظور اعمال این پیش‌پردازش‌ها در زبان فارسی، از ابزار هضم استفاده شده است[۵۴]. از ویژگی‌های این برنامه می‌توان به قابلیت تمیز و مرتب کردن متن، تقطیع جمله‌ها و واژه‌ها، ریشه‌یابی واژه‌ها، تحلیل صرفی و تجزیه نحوی جمله اشاره کرد.

---

[۴۸] Similarity Features
[۴۹] Semantic Role Labeling
[۵۰] http://nlp.guilan.ac.ir/
[۵۱] Normalize
[۵۲] Sentence Tokenizer
[۵۳] Word Tokenizer
[۵۴] https://github.com/mojtaba-khallash/JHazm




شکل۱. معماری پیشنهادی سیستم پرسش و پاسخ در زبان فارسی

یکی دیگر از اقداماتی که در این بخش انجام شده، حذف کلمات توقف است. در زبان فارسی لیستی حاوی ۳۳۲ کلمه توقف ارایه شده است[۵۵] که در این پژوهش از این مجموعه کلمات توقف استفاده شده است. این مجموعه کلمات توقف برای کلیه سیستم‌های پردازش زبان طبیعی طراحی شده و خاص سیستم‌های پرسش پاسخ نیست. از این رو بسیاری از کلمات پرسشی مانند "آیا"و "چرا" نیز به عنوان کلمه توقف در نظر گرفته شده است. با توجه به اهمیت کلمات پرسشی در سیستم‌های پرسش و پاسخ، در این پژوهش یک مجموعه کلمات توقف خاص سیستم‌های پرسش و پاسخ نیز تهیه شد که حاوی ۳۱۰ کلمه توقف است و در ارزیابی‌ها مورد استفاده قرار گرفت تا بررسی شود که حذف کلمات پرسشی از مجموعه کلمات توقف چه میزان بر نتیجه تاثیر خواهد گذاشت. کلیه نتایج روش‌های اعمال شده در بخش بعدی ارایه می‌شود.

سوال ورودی پس از اعمال پیش‌پردازش‌های بیان شده، به عنوان پرس‌وجوی بازیابی اطلاعات در اختیار مولفه بازیابی سند قرار می‌گیرد. لازم به ذکر است که کلیه پیش‌پردازش‌های انجام شده بر روی سوال، به طور مشابه بر روی مجموعه پاسخ‌ها نیز اعمال می‌شود و مجموعه پاسخ‌ها پس از اعمال پیش‌پردازش‌های بیان شده در اختیار مولفه بازیابی سند قرار می‌گیرد.

[۵۵] http://members.unine.ch/jacques.savoy/clef/index.html

## تشخیص موضوع

یکی از اقداماتی که در مولفه پردازش سوال سیستم پیشنهادی انجام شده است، تشخیص موضوع سوال است. در پیکره "رسائل و مسایل"، موضوع (عنوان فصل) هر سوال و جواب نگهداری شده است، بنابراین تشخیص موضوع سوال مطرح شده می‌تواند موجب شود که فقط پاسخ‌هایی مربوط به همین موضوع بازیابی شوند.

یک روش به کار گرفته شده برای تشخیص نوع موضوع سوال، روش مبتنی بر بازیابی اطلاعات است. در این روش یک سوال جدید به عنوان پرس‌وجوی بازیابی اطلاعات در نظر گرفته می‌شود و موتور جستجوی لوسین، n سوال مشابه با این سوال جدید را بازیابی می‌کند. سپس در میان این n سوال بازیابی شده، پرتکرارترین موضوع سوال را به سوال جدید نسبت می‌دهد. این روش برای n با مقادیر مختلفی مورد آزمایش قرار گرفت و در نهایت n= ۵ به عنوان تعداد اسناد بازیابی شده در نظر گرفته شد. با توجه به اینکه سیستم ارایه شده در یک دامنه بسته کار می‌کند و تعداد سوالات مجموعه نمایه‌گذاری و تست کم است، استفاده از روش‌های بازیابی مانند روش فوق ممکن است در این دامنه بسته چندان موثر واقع نشود.

یک روش پرکاربرد در سیستم‌های دامنه بسته، استفاده از مجموعه قوانین است. بر این اساس به منظور تشخیص موضوع در دامنه بسته، برای هر فصل مجموعه‌ای از کلمات کلیدی به صورت دستی استخراج شد. برای مثال برای فصل "سحر و شعبده" کلمات "شعبده، تردستی، رمل، احضار ارواح، تسخیر ارواح، فالگیری و هیپنوتیزم" به عنوان کلمات کلیدی ذخیره شده است که این کلمات کلیدی به منظور تشخیص موضوع به کار گرفته شده است.

## مرحله دوم، بازیابی سند

این مولفه، پرس‌وجوی تولید شده از مولفه پردازش سوال را دریافت کرده، سپس مشابه‌ترین اسناد موجود با این پرس‌وجو را از میان مجموعه پاسخ‌ها بازیابی می‌کند. به منظور بازیابی اسناد، عموماً از موتورهای جستجوی منبع باز استفاده می‌شود که اخیرا استفاده از موتورهای جستجوی بولین در سیستم‌های پرسش و پاسخ مورد توجه پژوهشگران قرار گرفته است [۲۵]. یکی از معروف‌ترین موتورهای جستجوی بولین، موتور جستجوی لوسین[۵۶] است [۳۲]که به طور رایگان در دسترس عموم قرار دارد[۵۷]. این موتور جستجو بر اساس پرس‌وجوی بولین عمل می‌کند و برپایه مدل فضای بردار TF. IDF[۵۸] است.

۵۶ lucence

۵۷ https://lucene.apache.org/.

۵۸ term frequency–inverse document frequency

موتور جستجوی ایندری نیز یک موتور جستجوی پر کاربرد در زمینه بازیابی اطلاعات است که بر اساس ترکیبی از روش‌های مدل‌های زبانی و شبکه‌های استدلال[۵۹] عمل می‌کند. تحقیقات نشان داده است که موتور جستجوی ایندری در مقاسیه با لوسین نتایج بهتری در پی داشته است [۲۴]. بدین جهت در مولفه بازیابی سند سیستم پیشنهادی، دو موتور جستجوی لوسین و ایندری به کار گرفته شده است تا مقایسه‌ای بین این دو موتور جستجو در زبان فارسی انجام شود و سرانجام بهترین موتور جستجو بر روی سیستم اعمال شود.

نحوه عملکرد این موتورهای جستجو به‌این‌صورت است که ابتدا مجموعه پاسخ‌ها را نمایه‌گذاری می‌کنند، سپس یک سوال را به عنوان ورودی دریافت کرده و با استفاده از مدل و فرمول بازیابی خود، به هر یک از اسناد با توجه به میزان شباهتشان با سوال، یک نمره نسبت می‌دهند. سپس اسناد را بر اساس نمره به ترتیب نزولی مرتب می‌کنند تا بهترین سند در رتبه ۱ قرار گیرد. یکی از چالش‌ها در این زمینه تعیین تعداد اسناد بازیابی برای هر سوال است که عموماً با انجام آزمایش، این تعداد تعیین می‌شود.

## مرحله سوم، رتبه‌بندی مجدد

پس از اعمال مولفه بازیابی سند، اسناد بازیابی شده به عنوان ورودی به مولفه رتبه‌بندی مجدد داده می‌شود. در این مولفه از روش یادگیری ماشین SVM استفاده شده است که در این پژوهش الگوریتم SVM از نرم‌افزار LibSvm [۳۳]به کار گرفته شده است. یکی از ویژگی‌های این الگوریتم، امکان وزن‌دهی آن برای داده‌های ناهمگون است که در سیستم پیشنهادی مورد استفاده قرار گرفته است. در LibSvm به طور پیش‌فرض ۴ هسته خطی[۶۰]، چندجمله‌ای[۶۱]، تابع شعاعی[۶۲] و سیگموید[۶۳] برای استفاده کاربر تعبیه شده است. به علاوه این امکان فراهم شده است که کاربر بتواند هسته جدید نیز تعریف کند. یک گام اساسی در روش‌های یادگیری ماشین، اعمال مجموعه ویژگی‌های مناسب است. در ادامه ویژگی‌های به کار رفته در این مولفه، به تفصیل بیان می‌شود:

## ویژگی بازیابی سند

این ویژگی بر اساس نتیجه مولفه بازیابی سند تعیین می‌شود که می‌تواند نمره خروجی موتور جستجو باشد یا رتبه سند بازیابی شده و یا هر ترکیب دیگری از آن. در این پژوهش، برای این

۵۹ Inference Network

۶۰ Linear

۶۱ polynomial

۶۲ Radial Basis Function

۶۳ Sigmoid

ویژگی هم نمره خروجی موتور جستجو و هم رتبه سند بازیابی شده مورد آزمایش قرار گرفته است.

## ویژگی‌های تراکمی

در این بخش تعدادی ویژگی تراکمی، به کار گرفته می‌شود. لازم به ذکر است که این ویژگی‌ها پس از اعمال پیش‌پردازش‌ها (ریشه‌یابی و حذف کلمات توقف) بر روی متن سوال و جواب‌ها اعمال می‌شود.

ویژگی تعداد کلمات منطبق: تعداد کلماتی که هم در پرس‌وجو و هم در پاسخ به کار رفته است، محاسبه می‌شود.

ویژگی تعداد کلمات غیرمنطبق: تعداد کلماتی از پرس‌وجو که در پاسخ به کار نرفته است، شمارش می‌شود.

به منظور نرمال کردن دو ویژگی تعداد کلمات منطبق و تعداد کلمات غیرمنطبق، می‌توان نتیجه حاصل شده را تقسیم بر مجموع تعداد کلمات پرس‌وجو و پاسخ کرد.

ویژگی مجموع Idf کلمات منطبق سوال: مجموع idf تک تک کلمات پرس‌وجو که در متن پاسخ به کار رفته است، محاسبه می‌شود. به منظور نرمال‌سازی این ویژگی، حاصل بدست آمده تقسیم بر تعداد کلمات منطبق پرس‌وجو و پاسخ می‌شود.

ویژگی دنباله کلمات: تعداد کلماتی از پرس‌وجو که دقیقا با همان ترتیب در پاسخ به کار رفته‌اند، محاسبه می‌شود. برای مثال برای دو کلمه x و y از پرس‌وجو در صورتی می‌گوییم دنباله این دو کلمه در پرس‌وجو و پاسخ یکسان است که اگر در پرس‌وجو کلمه x قبل از y رخ داده باشد، در متن پاسخ نیز ترتیب قرارگیری این دو کلمه به همین صورت باشد. لازم به ذکر است که برای محاسبه دنباله کلمات، برای هر دو کلمه پرس‌وجو باید ترتیب قرارگیری آن در متن پاسخ بررسی شود. برای پرس‌وجویی با n کلمه، تعداد مقایسات n(n-1)/2 خواهد بود که این مقدار می‌تواند به منظور نرمال‌سازی این ویژگی به کار گرفته شود.

ویژگی کیفیت اطلاعات پاسخ: این ویژگی برابر است با تعداد کلمات اسم، فعل و صفت از متن پاسخ که در پرس‌وجو به کار نرفته است.

## ویژگی‌های نحوی

به منظور اعمال ویژگی‌های نحوی ابتدا لازم است که متن سوال و جواب تجزیه[۶۴] شود. در این پژوهش از ابزار mate[۶۵] [۳۴] به منظور تجزیه جملات استفاده شده است که یک تجزیه‌گر مبتنی بر گراف است. در این بخش ۷ ویژگی نحوی به کار گرفته شده است.

فرض کنید متغیرهای مورد استفاده به شرح زیر هستند:

Q: دربرگیرنده کلمات پرس‌وجو

A: دربرگیرنده کلمات پاسخ

$Q_{Subject}$: شامل کلماتی از Q که دارای نقش نحوی subject هستند.

$A_{Subject}$: شامل کلماتی از A که دارای نقش نحوی subject هستند.

اشتراک ویژگی نقش فاعلی[۶۶]: برای محاسبه اشتراک ویژگی نقش فاعلی میان پرس‌وجو و پاسخ از دو رابطه (۶) و (۷) استفاده می‌شود.

$$F1_{subject}=\frac{count(Q_{Subject} IN\ A)+count(A\ IN\ Q_{Subject})}{count(Q_{Subject})+count(A)} \quad (۶)$$

$$F2_{subject}=\frac{count(Q_{Subject}\ IN\ A_{Subject})+count(A_{Subject}\ IN\ Q_{Subject})}{count(Q_{Subject})+count(A_{Subject})} \quad (۷)$$

$Count(Q_{subject}\ IN\ A)$، برابر است با تعداد کلماتی از $Q_{Subject}$ که در A نیز تکرار شده است. تفاوت دو فرمول ارایه شده در این است که در ویژگی $F1_{subject}$ ،کل کلمات پاسخ در نظر گرفته شده است ولی در ویژگی $F2_{subject}$، تنها کلماتی از پاسخ که دارای نقش نحوی subject هستند، لحاظ شده است.

ویژگی نقش مفعولی[۶۷]: برای بدست آوردن مقدار ویژگی نقش مفعولی نیز مانند رابطه (۶) و (۷) عمل می‌شود با این تفاوت که به جای محاسبه $Q_{Subject}$ و $A_{Subject}$، $Q_{Object}$ و $A_{Object}$ محاسبه می‌شود که در نهایت مقادیر دو ویژگی $F1_{object}$ و $F2_{object}$ به دست می‌آید.

ویژگی فعل اصلی: بر اساس تجزیه‌کننده mate، فعل اصلی جمله کلمه‌ای است که حاوی برچسب نحوی Root باشد. به طور مشابه برای این منظور دو ویژگی $F1_{Root}$ و$F2_{Root}$ محاسبه می‌شود.

ویژگی اشتراک سرواژه سوال با موضوع پاسخ: برای این منظور از دو متغییر $Q_{Focus}$ و $A_{Title}$ استفاده می‌شود. $Q_{Focus}$ شامل کلمه سرواژه پرس‌وجو است که همان فاعل جمله است و $A_{Title}$ شامل کلمات عنوان پاسخ است که در پیکره نگهداری شده است. اشتراک سرواژه سوال با موضوع پاسخ مطابق رابطه (۸) محاسبه می‌شود.

$$F_{Focus}=\frac{count(Q_{Focus}\ IN\ A_{Title})+count(A_{Title}\ IN\ Q_{Focus})}{count(Q_{Focus})+count(A_{Title})} \quad (۸)$$

[۶۴] Parse

[۶۵] https://code.google.com/p/mate-tools/

[۶۶] Subject

[۶۷] Object

### ویژگی تطابق موضوع

یکی از بخش‌هایی که در مولفه پردازش سوال توضیح داده شد، تشخیص موضوع سوال است. خروجی مولفه تشخیص موضوع سوال، عنوان موضوع سوال است که به عنوان ورودی مرحله رتبه‌بندی مجدد، به کار گرفته می‌شود. در صورتی که عنوان سوال منطبق با عنوان پاسخ باشد مقدار این ویژگی "۱" و در غیر این صورت مقدار این ویژگی "۰" در نظر گرفته می‌شود.

### هسته رتبه‌بندی

همان طور که پیش‌تر مطرح شد، علاوه بر استفاده از هسته‌های پیش‌فرض LibSvm، می‌توان هسته جدیدی نیز تعریف کرد. در این بخش یک هسته رتبه‌بندی که از ترکیب ویژگی‌های بازیابی سند، تراکمی، نحوی و تطابق موضوع تشکیل شده، اعمال شده است. نحوه کار هسته رتبه‌بندی پیشنهادی بدین‌صورت است که حاصل ضرب داخلی ویژگی‌های دو نمونه را محاسبه می‌کند.

$$K(QA1,QA2)=\sum_i C_i(QA1.F_i \times QA2.F_i) \quad (۹)$$

QA1 و QA2 دو نمونه ورودی هستند و i تعداد ویژگی‌های به کار رفته در هسته رتبه‌بندی و Ci ضریب ثابتی که برای هر ویژگی با آزمایش‌های مکرر بدست خواهد آمد.

### هسته درختی

روش هسته رتبه‌بندی تنها از ویژگی‌های پایه‌ای استفاده کرده است. یکی از ویژگی‌هایی که منبع غنی اطلاعات هستند، درخت تجزیه جملات است [۳۴]که امروزه در زمینه‌های مختلف پردازش زبان طبیعی به کار گرفته می‌شوند. به منظور استفاده بهینه از ویژگی‌های درختی، یک روش پرکاربرد استفاده از روش یادگیری ماشین SVM است. برای این منظور لازم است که هسته مناسب به الگوریتم SVM اعمال شود. در این پژوهش هسته درختی کالینز-دافی به کار گرفته شده است. از این هسته به منظور محاسبه میزان شباهت دو درخت تجزیه استفاده می‌شود که محاسبه تشابه میان درخت T1 وT2 بر اساس ضرب داخلی دو درخت صورت می‌گیرد [۳۴].

به منظور اعمال هسته درختی ابتدا لازم است تبدیلی بر روی خروجی تجزیه‌گر mate صورت گیرد. همان طور که پیش‌تر اشاره شد خروجی تجزیه‌گر mate از نوع درخت وابستگی است در حالی که برای اعمال هسته درختی لازم است که خروجی به صورت درخت مبتنی بر اجزا[۶۸] باشد. به منظور تبدیل درخت وابستگی به درخت مبتنی بر اجزا، دو نوع تبدیل به کار گرفته شده است [۳۵].

## آزمایشات و نتایج

در این بخش نتایج هر یک از روش‌های پیشنهادی به تفصیل بیان می‌شود و هر یک از نتایج به دست آمده مورد تجزیه و تحلیل قرار خواهد گرفت.

### پردازش سوال

یکی از پردازش‌های مهمی که در مولفه پردازش سوال سیستم پیشنهادی بیان شد، تشخیص موضوع سوال است که برای این منظور دو روش مبتنی بر بازیابی اطلاعات و مبتنی بر قاعده مورد استفاده قرار گرفت. در روش مبتنی بر بازیابی اطلاعات، مجموعه دادگان به پنج بخش تقسیم شده و هر بار ۴ بخش به عنوان مجموعه یادگیری و یک بخش به عنوان مجموعه تست در نظر گرفته شده است که در نهایت میانگین آن‌ها به عنوان نتیجه در نظر گرفته می‌شود. با استفاده از این روش، سیستم برای ۶۵.۷۲ درصد از سوالات، موضوع را به درستی تشخیص می‌دهد.

به منظور بهبود نتیجه مولفه تشخیص موضوع سوال، از روش مبتنی بر قاعده استفاده شده است. لازم به ذکر است که استفاده از روش‌های مبتنی بر قاعده در دامنه بسته، بسیار متداول است. برای این منظور مجموعه‌ای از قوانین دستی به کار گرفته شده به گونه‌ای که برای هر فصل کلمات کلیدی آن فصل در نظر گرفته شده است. با به کارگیری کلمات کلیدی هر فصل، نتیجه صحت تشخص موضوع سوال به ۹۱.۹۲ درصد رسید. با توجه به عملکرد بهتر روش مبتنی بر قاعده، در سیستم پیشنهادی این روش به منظور تشخیص موضوع سوال به کار گرفته شده است.

### بازیابی سند

در این بخش، کلیه آزمایشات لازم بر روی مولفه بازیابی سند به همراه نتایج آن‌ها به تفصل بیان می‌شود.

### تاثیر تعداد اسناد بازیابی شده

ابتدا قبل از اعمال پیش‌پردازش بر روی سوال و مجموعه پاسخ‌ها، موتور جستجوی لوسین به کار گرفته شد تا تعداد اسنادی که باید توسط مولفه بازیابی سند بازیابی شود، بررسی شود. مطابق جدول (۱)، با افزایش تعداد اسناد بازیابی شده، میزان صحت افزایش می‌یابد اما اگر تعداد اسناد بازیابی شده بیش از حد افزایش یابد، کار مولفه رتبه‌بندی مجدد دشوار خواهد شد. با توجه به نتایج به دست آمده، تعداد ۳۰ سند مناسب تشخیص داده شده که از این پس کلیه آزمایشات بر روی ۳۰ سند بازیابی شده، انجام می‌شود.

[۶۸] Constituency-based Parse tree

جدول ۱. تاثیر بازیابی تعداد اسناد مختلف بر روی نتیجه موتور جستجوی لوسین

| تعداد اسناد بازیابی شده | معیار صحت ٪ |
|---|---|
| ۱ | ۴۳.۲۷ |
| ۳ | ۵۷.۸۶ |
| ۵ | ۶۱.۹۷ |
| ۱۰ | ۶۷.۹۳ |
| ۱۵ | ۷۱.۳۳ |
| ۲۰ | ۷۳.۲۶ |
| ۳۰ | ۷۶.۰۹ |
| ۴۰ | ۷۷.۸۰ |
| ۵۰ | ۷۸.۹۸ |
| ۷۰ | ۸۱.۰۱ |

## مقایسه موتور جستجوی لوسین و ایندری

به منظور مقایسه موتور جستجوی لوسین و ایندری، موتور جستجوی ایندری بر روی همین پیکره به کار گرفته شد و معیار صحت و MRR آن به ازای ۳۰ سند بازیابی شده محاسبه شد. مقایسه عملکرد موتور جستجوی لوسین و ایندری در جدول (۲) نمایش داده شده است.

همان طور که در جدول (۲) نشان داده شده است، موتور جستجوی لوسین در مقایسه با ایندری عملکرد بهتری دارد به گونه‌ای که معیار صحت آن ۸ درصد و معیار MRR آن ۱۱ درصد بهتر است. بدین جهت در مولفه بازیابی سند سیستم پیشنهادی، موتور جستجوی لوسین به کار گرفته می‌شود.

جدول ۲. مقایسه عملکرد موتور جستجوی لوسین و ایندری بر روی ۳۰ سند بازیابی شده

| موتور جستجو | MRR | صحت ٪ |
|---|---|---|
| لوسین | ۰.۵۱۹۹ | ۷۶.۰۹ |
| ایندری | ۰.۴۰۱۵ | ۶۸.۰۲ |

## تاثیر پیش‌پردازش‌ها بر روی مولفه بازیابی سند

در این بخش تاثیر عمل ریشه‌یابی و حذف کلمات توقف بر روی موتور جستجوی لوسین نشان داده شده است. لازم به ذکر است که قبل از انجام ریشه‌یابی و حذف کلمات توقف، پیش‌پردازش‌های نرمال‌سازی، جداسازی جملات و کلمات انجام شده است.

همان طور که در بخش قبل بیان شد، یک مجموعه کلمات توقف خاص سیستم‌های پرسش و پاسخ طراحی گردید که در این مجموعه، کلمات "چند، چرا، چگونه، چیست، کجاست، کجا، کی، چطور، کدام، آیا، مگر، چندین، چیزی، کسی، هیچ، چیز، بله، بلی، آره، آری، اگر، چه" از مجموعه کلمات توقف فارسی عمومی حذف شده است. مطابق با نتایج بدست آمده اعمال مجموعه کلمات توقف خاص سیستم‌های پرسش و پاسخ در مقایسه با مجموعه کلمات توقف عمومی، نتیجه ضعیف‌تری داشته است. مولفه بازیابی سند در بهترین حالت با انجام ریشه‌یابی و حذف کلمات توقف عمومی توانسته است به بالاترین میزان صحت ۸۲.۲۹ دست یابد. در نتیجه در سیستم طراحی شده به منظور پیش‌پردازش‌های لازم بر روی سوال و مجموعه پاسخ‌ها، حذف کلمات توقف عمومی به همراه ریشه‌یابی انجام می‌شود.

جدول ۳. تاثیر پیش‌پردازش‌ها بر روی مولفه بازیابی سند

| پیش‌پردازش | MRR | صحت ٪ |
|---|---|---|
| **بدون پیش‌پردازش** | ۰.۵۱۹۹ | ۷۶.۰۹ |
| **حذف کلمات توقف عمومی** | ۰.۵۵۶۱ | ۸۱.۴۳ |
| **حذف کلمات توقف عمومی + ریشه‌یابی** | ۰.۵۶۷۳ | ۸۲.۲۹ |
| **حذف کلمات توقف پرسش و پاسخ+ ریشه‌یابی** | ۰.۵۶۰۸ | ۸۰.۸۷ |

مولفه بازیابی سند در بهترین حالت میزان صحت آن ۸۲.۲۹ و معیار MRR آن برابر با ۰.۵۶ است. به عبارت دیگر مولفه بازیابی سند برای ۸۲.۲۹ درصد از سوالات توانسته است در میان ۳۰ سند بازیابی شده، پاسخ درست را برگرداند و برای ۱۷.۷۱ درصد از سوالات پاسخ درستی بازیابی نشده است. می‌توان نتیجه گرفت میزان صحت مولفه رتبه‌بندی مجدد نیز ۸۲.۲۹ خواهد بود؛ زیرا مولفه رتبه‌بندی مجدد بر روی ۱۷.۷۱ درصد سوالات دیگر هیچ تاثیری نمی‌تواند داشته باشد. بدین جهت در آزمایشات انجام شده بر روی مولفه رتبه‌بندی مجدد، تنها مجموعه ۸۲.۲۹ درصد از سوالات که برای آن‌ها حداقل یک پاسخ درست بازیابی شده است در نظر گرفته می‌شود. مقدار MRR برای ۸۲.۲۹ درصد از سوالات باقی مانده، در حالت پایه لوسین قبل از انجام پیش‌پردازش پیشنهادی، برابر با ۰.۶۳۱۷ است که با اعمال پیش‌پردازهای پیشنهادی، این مقدار به ۰.۶۸۹۳ رسیده است.

## رتبه‌بندی مجدد

در این قسمت تاثیر هر یک از ویژگی‌های به کار رفته در مولفه رتبه‌بندی مجدد بررسی می‌شود. برای هر سوال ورودی، مولفه بازیابی سند، ۳۰ سند مشابه با سوال را بازیابی می‌کند و نتیجه را برای مولفه رتبه‌بندی مجدد ارسال می‌کند. به منظور اعمال روش‌های یادگیری ماشین، هر نمونه دربرگیرنده یک جفت سوال و جواب و ویژگی‌های آن‌ها به همراه برچسب کلاس نتیجه است. در صورتی که سوال و جواب مرتبط باشند، برچسب کلاس "۱"

و در غیر این صورت برچسب کلاس "۰" خواهد بود.

## مساله عدم توازن دادگان

در میان مجموعه دادگان مورد آزمایش، به ازای ۲۹ نمونه غلط، تنها یک نمونه درست وجود دارد. اولین اقدامی که باید در این مرحله صورت گیرد رفع مشکل عدم توازن داده‌ها است. بدین منظور از قابلیت وزن‌دهی نرم افزار libSvm استفاده شده است و با اعمال وزن‌های مختلف میان دو دسته کلاس ۰ و ۱، در نهایت بهترین وزن‌دهی انجام شده است. پس از انجام آزمایشات لازم، وزن کلاس "۰" ها ۰.۱۵ و وزن کلاس "۱" ها ۰.۸۵ در نظر گرفته شده است.

## محاسبه MRR

برای محاسبه معیار MRR، لازم است که اسناد بازیابی شده برای یک سوال به ترتیب رتبه‌بندی شوند تا بتوان تعیین کرد که سیستم، پاسخ درست را در چه رتبه‌ای قرار داده است. خروجی الگوریتم SVM به‌این‌صورت است که بر اساس مجموعه آموزش، یک ابر سطح برای جداسازی دسته‌های مثبت و منفی در نظر می‌گیرد و وقتی یک نمونه تست دریافت می‌کند، تعیین می‌کند که این نمونه تست در کدام دسته قرار دارد. به منظور محاسبه MRR، نمونه‌ای از کلاس مثبت‌ها که بیشترین فاصله را از ابرصفحه داشته باشد، به عنوان بهترین نمونه و نمونه‌ای از کلاس منفی‌ها که بیشترین فاصله را از ابرصفحه داشته باشد، به عنوان بدترین نمونه در نظر گرفته می‌شود. بدین ترتیب پس از اعمال الگوریتم یادگیری SVM، به ازای هر سوال، ۳۰ نمونه مربوط به آن بر اساس مقدار فاصله از ابرصفحه مرتب شده و سپس MRR برای آن‌ها محاسبه می‌شود.

## تاثیر ویژگی‌های رتبه‌بندی مجدد

دراین بخش تاثیر هر یک از ویژگی‌های به کار رفته در مولفه رتبه‌بندی مجدد، بیان می‌شود.

پس از حذف سوالاتی که برای آن‌ها هیچ پاسخ درستی بازیابی نشده است، میزان MRR موتور بازیابی لوسین ۶۳.۱۷ درصد بدست آمده است و پس از انجام پیش‌پردازش‌های پیشنهادی، معیار MRR مولفه بازیابی سند به مقدار ۶۸.۹۳ درصد رسیده است.

ویژگی بازیابی سند: به منظور اعمال روش یادگیری ماشین SVM، اولین ویژگی به کار گرفته شده رتبه بازیابی موتور لوسین است. با اعمال این ویژگی، مقدار MRR به ۶۸.۹۵ رسیده است. لازم به ذکر است که در این آزمایش از هسته پیش فرض RBF استفاده شده است. ویژگی رتبه بازیابی لوسین بر روی ۴ هسته پیش فرض SVM اعمال شد و در نهایت هسته RBF در این مجموعه دادگان بهترین نتیجه را در پی داشت. از این رو کلیه آزمایشات آتی بر روی هسته RBF انجام می‌پذیرد.

جدول ۴. تاثیر ویژگی‌های رتبه‌بندی مجدد بر روی MRR

| روش به‌کاررفته | MRR ٪ |
|---|---|
| مقدار پایه لوسین | ۶۳.۱۷ |
| مقدار پایه لوسین با استفاده از پیش‌پردازش پیشنهادی | ۶۸.۹۳ |
| ویژگی بازیای سند | ۶۸.۹۵ |
| ویژگی‌های تراکمی | ۶۸.۹۵ |
| ویژگی‌های نحوی | ۶۶.۰۵ |
| ویژگی تطابق موضوع | ۷۰.۷۹ |
| هسته رتبه‌بندی | ۷۱.۸۸ |
| هسته درختی | ۳۰.۵۱ |
| ترکیب هسته رتبه‌بندی و درختی | ۶۸.۸۸ |

ویژگی‌های تراکمی: در مرحله بعد ویژگی‌های تراکمی نیز به مجموعه ویژگی‌ها اضافه شده که بر روی نتیجه هیچ تاثیری نداشته است. دلیل این امر را می‌توان تشابه ویژگی‌های تراکمی با عملکرد موتور جستجوی لوسین دانست. زیرا این مجموعه ویژگی‌ها نیز مانند لوسین عموماً از ویژگی‌های ظاهری و تعدادی میان کلمات پرس‌وجو و پاسخ استفاده می‌کنند. با توجه به عدم تاثیر ویژگی‌های تراکمی، این ویژگی‌ها از بردار ویژگی حذف می‌شوند.

ویژگی‌های نحوی: اضافه کردن مجموعه ویژگی‌های نحوی بر خلاف انتظار نه تنها بهبودی بر روی نتیجه کل نداشته است بلکه باعث ضعیف شدن نتیجه نیز شده است. بدین جهت این ویژگی‌ها نیز از بردار ویژگی حذف می‌شوند.

ویژگی تطابق موضوع: اعمال ویژگی تطابق موضوع به بردار ویژگی‌ها باعث بهبود نتیجه شده و مقدار MRR را به ۷۰.۷۹ رسانده است.

هسته رتبه‌بندی: تاکنون تمام آزمایشات فوق بر روی هسته پیش فرض RBF صورت گرفته است. اکنون رابطه (۹) که به منظور اعمال هسته رتبه‌بندی ارایه شده است، به کار گرفته می‌شود که در این حالت بردار ویژگی‌ها شامل ویژگی بازیابی سند و ویژگی تطابق موضوع است. با انجام آزمایشات مختلف، در نظر گرفتن ضریب ویژگی بازیابی سند برابر با ۱ و ضریب ویژگی تطابق موضوع برابر با ۰.۵ بهترین نتیجه را در پی داشته است و میزان MRR را به ۷۱.۸۸ رسانده است.

هسته درختی: در این بخش هسته درختی بر روی روش یادگیری ماشین SVM اعمال می‌شود. برای این منظور طبق رابطه (۱۰) عمل شده است.

$$K(QA_1, QA_2) = K(Q_1, Q_2)/MaxQuestionskernel \times K(A_1, A_2) \quad (10)$$

$K(Q_1, Q_2)$ برابر است با مقدار هسته درختی میان سوال $Q_1$ و $Q_2$. به طور مشابه $K(A_1, A_2)$ نیز مقدار هسته درختی میان جواب‌های $A_1$ و $A_2$ را نشان می‌دهد. به ازای هر جفت سوال موجود در پیکره، هسته درختی میان آن‌ها محاسبه شده و حداکثر مقدار بدست آمده به عنوان MaxQuestionsKernel در نظر گرفته می‌شود.

ایده به‌کارگیری هسته درختی در این پژوهش این است که سوال‌های مشابه، احتمالا دارای پاسخ‌های مشابه‌ای هستند. طبق رابطه فوق تاثیر تشابه درختی دو پاسخ، وابسته به میزان تشابه دو سوال است و در صورتی که تشابه درختی دو سوال بیشتر باشد، تشابه پاسخ‌های آن‌ها بیشتر دخالت داده می‌شود.

با وجود اینکه اعمال روش پیچیده هسته درختی در زبان انگلیسی بر روی مولفه‌های دیگر سیستم‌های پرسش و پاسخ مثل دسته‌بندی سوال، نتایج بسیار خوبی را به دنبال داشته، ولی در این پژوهش در زمینه رتبه‌بندی مجدد نتایج ضعیفی را در پی داشته است. یکی از دلایل مهم این است که در مولفه رتبه‌بندی مجدد تعداد نمونه‌های غلط در مقایسه با نمونه‌های درست بسیار زیاد است. در این پژوهش برای هر سوال، ۱ نمونه درست و ۲۹ نمونه غلط نگهداری شده است که این امر موجب گمراهی سیستم یادگیری می‌شود. علاوه بر این با توجه به ماهیت سوال و جواب‌های موجود در رساله، ممکن است سوال و جواب‌های متفاوت دارای میزان تشابه درختی بالا باشند که این امر نیز سیستم یادگیری را گمراه می‌کند.

ترکیب هسته درختی و هسته رتبه‌بندی: با توجه به عملکرد ضعیف هسته درختی در پیکره مورد استفاده، ترکیب هسته درختی با هسته رتبه‌بندی نیز نتوانست منجر به بهبود نتیجه شود. نتایج نشان می‌دهد که با اعمال ترکیب هسته درختی و هسته رتبه‌بندی، همواره هسته ترکیبی ضعیف‌تر از هسته رتبه‌بندی عمل می‌کند.

## مشکلات و چالش‌های پیاده‌سازی

پیاده‌سازی یک سیستم پرسش و پاسخ در زبان فارسی، با مشکلات و موانع زیادی رو به رو خواهد بود. از مهم‌ترین مشکلات موجود که در این پژوهش تاثیرات منفی بر روی نتایج داشته است، می‌توان به موارد زیر اشاره کرد:

### عدم وجود پیکره متنی مناسب

با وجود تلاش‌هایی که در راستای تهیه پیکره رسائل و مسائل صورت گرفته است، این پیکره حاوی ۲۱۱۸ سوال غیرحقیقت و ۲۰۵۱ سوال حقیقت است که این پیکره با این تعداد نمونه به منظور انجام روش‌های یادگیری ماشین مناسب،کافی نخواهد بود. از آن جایی که این پیکره تنها پیکره موجود در زمینه پرسش و پاسخ در زبان فارسی است، در حال حاضر امکان پیاده‌سازی سیستم پرسش و پاسخ در زبان فارسی در دامنه دیگری وجود ندارد و تنها بر روی این دامنه پرسش و پاسخ مذهبی می‌توان کار کرد.

### عدم وجود ابزارهای معنایی

در پژوهش‌های کنونی که در زبان انگلیسی بر روی سیستم‌های پرسش و پاسخ صورت می‌گیرد، ابزارهای معنایی نقش بسزایی ایفا می‌کنند. در زبان انگلیسی استفاده از ابزارهایی چون تشخیص دهنده موجودیت نامدار و تجزیه‌گر معنایی بسیار مورد توجه پژوهشگران قرار گرفته است. اعمال ویژگی‌های معنایی در زبان انگلیسی، نتایج بسزایی در بهبود سیستم‌های پرسش و پاسخ داشته است. به دلیل نبود ابزارهای معنایی در زبان فارسی، به کار گیری ویژگی‌های معنایی در زبان فارسی امکان پذیر نیست و این یکی از مهم‌ترین محدودیت‌هایی است که در زبان فارسی پیش روی پژوهشگران قرار دارد.

### عدم امکان استفاده از فارس‌نت

در زبان فارسی شبکه واژگان فارس‌نت در دسترس عموم قرار دارد. این شبکه واژگان مشابه با شبکه واژگان انگلیسی وردنت است که در آن انواع ارتباط میان کلمات از جمله تضاد و تشابه در نظر گرفته شده است. با وجود در دسترس بودن فارس‌نت، امکان استفاده از آن در این پژوهش میسر نیست؛ زیرا فارس‌نت یک شبکه واژه‌گانی عمومی است در حالی که سیستم طراحی شده یک سیستم در دامنه بسته بر روی سوالات مذهبی است و بسیاری از کلمات کلیدی این دامنه بسته در فارس‌نت وجود ندارد، در نتیجه استفاده از فارس‌نت نیز در این پژوهش میسر نیست.

### دقت پایین ابزارهای فارسی

ابزارهای موجود در زبان فارسی مانند نرمال‌ساز، جداکننده جملات، جداکننده کلمات، ریشه‌یاب و تجزیه‌گر که در این پژوهش مورد استفاده قرار گرفته‌اند در مقایسه با ابزارهای موجود در زبان انگلیسی دارای دقت پایین‌تری هستند. ضعیف‌تر بودن ابزارهای پردازش متن فارسی از جمله دلایل دیگری است که می‌تواند تاثیر منفی بر روی نتایج بگذارد.

## نتیجه‌گیری

در این مقاله معماری پیشنهادی سیستم پرسش و پاسخ در زبان فارسی بر روی سوالات غیرحقیقت معرفی شد. معماری پیشنهادی از سه مولفه پردازش سوال، بازیابی سند و رتبه‌بندی مجدد تشکیل شده است. در مولفه پردازش سوال، پیش‌پردازش‌های لازم بر روی سوال اعمال شد. لازم به ذکر است که پیش‌پردازش‌های پیشنهادی، خاص زبان فارسی است. همچنین به منظور تشخیص عنوان سوال، دو روش مبتنی بر بازیابی و مبتنی بر قاعده ارایه شد که روش مبتنی بر قاعده نتایج بهتری را در پی داشت. تشخیص عنوان سوال تاکنون در سیستم‌های پرسش و پاسخ زبان انگلیسی استفاده نشده است در نتیجه می‌تواند به عنوان نوآوری سیستم پیشنهادی در نظر گرفته شود.

در مولفه بازیابی سند، دو موتور جستجوی لوسین و ایندری مورد استفاده قرار گرفت که نتایج نشان داد که استفاده از موتور جستجوی لوسین در سیستم پرسش و پاسخ در زبان فارسی نتایج بهتری را در پی خواهد داشت.

در مولفه رتبه‌بندی مجدد از روش‌های یادگیری ماشین استفاده شده است. به منظور بهبود رتبه‌بندی، ویژگی‌های متنوعی از جمله رتبه موتور جستجو، ویژگی‌های تراکمی، ویژگی‌های نحوی و ویژگی تشخیص عنوان سوال به کار گرفته شده است. در میان ویژگی‌های به کار رفته در مولفه رتبه‌بندی مجدد، ویژگی موتور جستجو و تشخیص عنوان سوال در مقایسه با سایر ویژگی‌ها نقش موثرتری در نتیجه داشته‌اند.

در نهایت هسته رتبه‌بندی و درختی نیز به طور مجزا و ترکیبی در الگوریتم یادگیری SVM به کار گرفته شده است. نتایج نشان می‌دهد که در بهترین حالت سیستم توانست به دقت ۸۲.۲۹ درصد و میانگین معکوس رتبه ۷۱.۸۸ درصد دست یابد. سیستم پیشنهادی ما توانست معیار میانگین معکوس رتبه را نسبت به حالت پایه لوسین حدود ۸ درصد بهبود دهد. در حال حاضر در زبان فارسی سیستم پرسش و پاسخی ارایه نشده است تا بتوان سیستم پیشنهادی را با آن مقایسه کرد. همچنین در زبان انگلیسی سیستم پرسش و پاسخی در زمینه داده‌های مذهبی وجود ندارد تا امکان مقایسه دقیق سیستم پیشنهادی با سیستم‌های موجود در زبان انگلیسی میسر شود. از طرفی اغلب سیستم‌های ارایه شده در زبان انگلیسی بر روی سوالات حقیقت هستند که پاسخ‌گویی به سوالات حقیقت در مقایسه با سوالات غیرحقیقت بسیار ساده‌تر است.

راهکارهای زیر می‌تواند در آینده به منظور بهبود پژوهش صورت گیرد. ۱) تهیه یک پیکره پرسش و پاسخ جامع در زبان فارسی که حاوی تعداد نمونه‌های بیشتری باشد. ۲) به کار گیری نمایه‌گذاری معنایی نهفته[۶۹] می‌تواند در سیستم‌های پرسش و پاسخ فارسی مورد توجه پژوهشگران قرار گیرد. ۳) سیستم پرسش و پاسخ تعاملی نیز می‌تواند در آینده به کار گرفته شود به گونه‌ای که از نظرات کاربران به منظور بهبود فرآیند یادگیری استفاده شود.

## مراجع


[1] M. A. Greenwood, “Open-domain question answering,” Found. Trends Inf. Retr., no. September, 2005.

[2] O. Kolomiyets and M.-F. Moens, “A survey on question answering technology from an information retrieval perspective,” Inf. Sci. (Ny)., vol. 181, no. 24, pp. 5412–5434, 2011.

[3] L. Hirschman and R. Gaizauskas, “Natural language question answering: the view from here,” Nat. Lang. Eng., vol. 7, no. 04, pp. 275–300, Feb. 2002.

[4] D. Moll and L. Vicedo, “Special Section on Restricted-Domain Question Answering Question Answering in Restricted Domains : An Overview,” no. October 2006, 2007.

[5] D. Zhang and W. S. Lee, “Question classification using support vector machines,” in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, pp. 26–32.

[6] P. Gupta and V. Gupta, “A survey of text question answering techniques,” Int. J. Comput. Appl., vol. 53, no. 4, pp. 1–8, 2012.

[7] M. Surdeanu, M. Ciaramita, and H. Zaragoza, “Learning to rank answers to non-factoid questions from web collections,” Comput. Linguist., vol. 37, no. 2, pp. 351–383, 2011.

[8] E. M. Voorhees, “The TREC-8 Question Answering Track Report.,” in TREC, 1999, vol. 99, pp. 77–82.

[9] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke, “Creating the DISEQuA corpus: a test set for multilingual question answering,” in Comparative Evaluation of Multilingual Information Access Systems, vol. 1994, Springer, 2004, pp. 487–500.

[10] N. A. Smith, M. Heilman, and R. Hwa, “Question generation as a competitive undergraduate course project,” in Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge, 2008, pp. 4–6.

[11] X. Li and D. Roth, “Learning question


[۶۹] Latent Semantic Indexing

classifiers," in Proceedings of the 19th international conference on Computational linguistics-Volume 1, 2002, pp. 1–7.

[12] D. Tom, D. Tomás, J. L. Vicedo, E. Bisbal, and L. Moreno, "TrainQA: a Training Corpus for Corpus-Based Question Answering Systems," in Proc. 8th Int. Conf. on Computational Linguistics and Intelligent Text Processing, CICLing-2007, IEEE Computer Society, 2007, pp. 1–7.

[13] A. Mollaei, S. Rahati-Quchani, and A. Estaji, "Question classification in Persian language based on conditional random fields," 2012 2nd Int. eConference Comput. Knowl. Eng., pp. 295–300, Oct. 2012.

[14] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," Knowledge-Based Syst., vol. 22, no. 5, pp. 382–387, Jul. 2009.

[15] آنالویی، مرتضی، جنقرا، مسلم محمدی،"ارایه یک سیستم پرسش و پاسخ با رده‌بندی سوالات و جملات کاندید"، چهاردهمین کنفرانس انجمن کامپیوتر ایران، ۱۳۸۷.

[16] S. M. Harabagiu, D. I. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. C. Bunescu, R. Girju, V. Rus, and P. Morarescu, "FALCON: Boosting Knowledge for Answer Engines.," in TREC, 2000, vol. 9, pp. 479–488.

[17] E. H. Hovy, L. Gerber, U. Hermjakob, M. Junk, C. Lin, and M. Rey, "Question Answering in Webclopedia.," in TREC, 2000, vol. 52, pp. 53–56.

[18] J. J. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," Artif. Intell. Rev., vol. 35, no. 2, pp. 137–154, Nov. 2011.

[19] E. Breck, J. D. Burger, L. Ferro, D. House, M. Light, and I. Mani, "A Sys Called Qanda.," in TREC, 1999.

[20] X. Li and D. Roth, "Learning question classifiers: the role of semantic information," Nat. Lang. Eng., vol. 12, no. 03, pp. 229–249, 2006.

[21] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval, vol. 1, no. c. Cambridge university press Cambridge, 2008.

[22] R. Mandala, T. Takenobu, and T. Hozumi, "The use of WordNet in information retrieval," in Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference, 1998, pp. 31–37.

[23] E. Eckard and J.-C. Chappelier, "Free Software for research in Information Retrieval and Textual Clustering," 2007.

[24] H. Turtle, S. A. Rowe, and Y. Hegde, "Yet another comparison of lucene and indri performance," in SIGIR 2012 Workshop on Open Source Information Retrieval, 2012, pp. 64–67.

[25] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton, "Quantitative evaluation of passage retrieval algorithms for question answering," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, pp. 41–47.

[26] M. W. Bilotti, B. Katz, and J. Lin, "What works better for question answering: Stemming or morphological query expansion," in Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR, 2004, vol. 2004, no. 1.2, pp. 1–3.

[27] S. Verberne, H. Van Halteren, D. Theijssen, S. Raaijmakers, L. Boves, and H. Van Halteren, "Learning to rank qa data," in Proceedings of the Learning to Rank Workshop at SIGIR, 2009, pp. 41–48.

[28] S. Verberne, L. Boves, P. Coppen, and N. Oostdijk, "What is not in the Bag of Words for Why-QA?," Comput. Linguist., vol. 36, no. 2, pp. 229–245, 2010.

[29] M. W. Bilotti, J. Elsas, J. Carbonell, and E. Nyberg, "Rank learning for factoid question answering with linguistic and semantic constraints," in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 459–468.

[30] S.-J. Yen, Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, and J.-J. Liu, "A support vector machine-based context-ranking model for question answering," Inf. Sci. (Ny)., vol. 224, pp. 77–87, Mar. 2013.

[31] برشبان، یاسمن، یوسفی نسب، حامد و میرروشندل، سید ابوالقاسم،"رسائل و مسائل: توسعه یک پیکره متنی فارسی پرسش و پاسخ"، بیستمین کنفرانس انجمن کامپیوتر ایران-مشهد، ۱۳۹۳، صفحه ۱-۶

[32] M. Mccandless, E. Hatcher, and O. Gospodnetic, "Lucene in action," 2004.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27, 2011.

[34] B. Bohnet, "Very high accuracy and fast dependency parsing is not a contradiction," in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, no. August, pp. 89–97.

[35] S. A. Mirroshandel and G. Ghassem-Sani, "Temporal Relation Classification Using Dependency Convolution Tree Kernels," pp. 146–150, 2011.