استفاده از رگرسیون غیرخطی و ویژگی‌های آماری جهت ارزیابی سیستم‌های Interactive Question Answering (IQA)

محمد مهدی حسینی[1*]، مرتضی زاهدی[2]، حمید حسن‌پور[3]

[1]استادیار دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد شاهرود، hosseini_mm@yahoo.com

[2]استادیار دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات، دانشگاه صنعتی شاهرود

[3] استاد دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات، دانشگاه صنعتی شاهرود

چکیده

عدم امکان پیش‌گویی بخش تعاملی، مشکل اصلی در ارزیابی سیستم‌های پرسش و پاسخ تعاملی است. به همین منظور، باید انسان در فرآیند ارزیابی شرکت داشته باشد. در این مقاله یک مدل آماری براساس مجموعه‌ای از ویژگی‌های ایجاد شده بر اساس n-گرم‌ها و بزرگترین رشته مشترک،برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی ارائه شده است. ابتدا با استفاده از چهار سیستم پرسش و پاسخ تعاملی ، پایگاه داده‌ای از مکالمات رد و بدل شده بین کاربران و سیستم‌ها ایجاد گردید. در ادامه تعداد ۵۴۰ نمونه به عنوان داده مناسب در نظر گرفته شد تا مجموعه تست و آموزش بر اساس آن ایجاد گردد. سپس بر روی مکالمات، پیش پردازش صورت پذیرفت و بر اساس روابط تعریف شده، تعدادی ویژگی آماری جدید از متن مکالمه‌ها استخراج و بر اساس آن ماتریس ویژگی تشکیل گردید. با توجه به تعداد بالای ویژگی‌ها و برای جلوگیری از برازش خطا، بهترین ویژگی‌ها با استفاده از روش حذف ویژگی به روش بازگشتی انتخاب گردید تا مدل پیشنهادی بر اساس ویژگی های باقیمانده شکل گیرد. در نهایت با استفاده از رگرسیون به پیش بینی نظرات انسانی پرداخته شد که رگرسیون غیرخطی توانی بر اساس معیار مجذور کمترین مربع خطا به میزان ۰/۱۵ بهترین مدل را ارائه نمود.

کلیدواژه

ارزیابی، سیستم پرسش و پاسخ تعاملی، رگرسیون ، استخراج ویژگی.

مقدمه

یافتن پاسخ‌های صحیح و دقیق برای سوالات، در کوتاه‌ترین زمان ممکن به یکی از چالش‌های افراد در دنیای پر از اطلاعات، تبدیل شده است. در این راستا برای پاسخ‌گویی به این نیازمندی اطلاعاتی، سیستم‌هایی طراحی شده‌اند که می‌توانند پاسخ‌هایی را برای سوالات کاربران ارائه نمایند. سیستم پرسش و پاسخ[1] کاربران را قادر می سازد تا به منابع علمی با استفاده از زبان طبیعی (از طریق پرسش) دسترسی داشته باشند و پاسخ مرتبط و مختصر را دریافت کنند. با این حال، همچنان مشکلات چالش برانگیز فراوانی جهت مرتفع نمودن در این سیستم‌ها وجود دارد. سیستم‌های QA شکل پیچیده‌تر سیستم‌های بازیابی اطلاعات هستند که در این سیستم‌ها به جای ارائه‌ی کل سند، تنها بخش‌های خاصی از اطلاعات سند به عنوان پاسخ بازگردانده می شود بنابراین پاسخ ارائه شده ممکن است یک کلمه، یک جمله یا یک پاراگراف باشد. یک سیستم QA از سه بخش اصلی پردازش پرسش، تحلیل متن یا بازیابی اطلاعات و تحلیل یا پردازش پاسخ تشکیل شده است.

بیش از یک دهه است که هر ساله تحقیق پیرامون سیستم‌های پاسخ دامنه باز[2] که از منابع اطلاعات غیرساختاری بهره می‌برند، توسط کمپین ارزیابی TREC[3] مرتباً در حال انجام است[۱-۳]. شرایط ارز یابی سیستم های QA در اولین کمپین ارزیابی TREC شامل ۲۰۰ سوال و یک مجموعه سند بود. پاسخ‌ها در مجموعه‌ای جداگانه، از قبل مشخص شده بودند. حداکثر طول پاسخ می‌بایستی بین ۵۰ یا ۲۵۰ کاراکتر می‌بود و از سیستم‌ها خواسته شده بود که ۵ لیست از پاسخ‌های رتبه‌بندی شده را ارائه نمایند. در کمپین‌های بعدی TREC با توجه به افزایش تعداد و پیچیدگی درخواست‌ها، اسناد مورد استفاده و پیچیدگی سوالات روش‌های ارزیابی پاسخ نیز پیشرفته‌تر شدند. در کمپین‌های اولیه TREC از داده‌های محلی به عنوان منبع اطلاعات برای تولید پاسخ استفاده می شد، اما با گسترش صفحات وب، استفاده از این مجموعه از اطلاعات مورد توجه قرار گرفت. براین اساس چندین QAs بر مبنای وب توسعه یافتند[۴-۵]. QAS های مبتنی بر وب را می‌توان به QAS دامنه باز و QAs دامنه بسته طبقه‌بندی کرد[۶].

[1] Question Answering System (QAs)
[2] Open Domain System
[3] Text Retrieval Evaluation Conference

تاریخ دریافت: ۱۳۹۶/۰۵/۲۲ تاریخ پذیرش: ۱۳۹۷/۰۷/۲۰

سیستم‌های QA در زمانیکه پرسش کاربر دارای ابهام بوده یا پاسخ سیستم مطلوب کاربر نبوده و یا کاربر نیازمند دریافت اطلاعات بیشتری باشد، راهکاری برای رفع ابهام ارائه ننموده‌اند. بنابراین فقدان تعامل دو طرفه بین سیستم و کاربر یکی از بزرگترین معضلات این سیستم‌ها می‌باشد. لذا سیستم‌های پرسش و پاسخ تعاملی[4] مطرح شدند که با اضافه شدن سطح تعامل این مشکل در این سیستم‌ها رفع شد. سیستم‌های موجود در زمینه IQA می‌توانند با توجه به شرایط و کاربردهایشان در سه گروه مدیریت محدودیت، QA ارتقاء یافته و سوالات متوالی دسته‌بندی شوند. بدیهی است که وجود یک روش ارزیابی استاندارد نقش بسیار مهمی در ارتقای این سیستم‌ها ایفا می‌نماید. با این وجود تقریباً روش استانداری در زمینه ارزیابی سیستم‌های IQA پیشنهاد نشده است و روش‌های ارزیابی فعلی بر مبنای روش های مورد استفاده در QA و سیستم‌های دیالوگ بنا نهاده شده اند[۷]. یک سیستم IQA از دو موجودیت سیستم و کاربر تشکیل شده است لذا کار ارزیابی بسیار سخت و پیچیده است. در روش‌های موجود برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی علاوه بر ارزیابی کمی از ارزیابی کیفی نیز استفاده می شود که نیازمند مشارکت کاربران در فرآیند ارزیابی برای تعیین میزان موفقیت تعامل بین سیستم و کاربر می‌باشد. اگر چه روش‌های استانداردی وجود دارند که می‌توانند اطلاعات مربوط به عملکرد سیستم از قبیل زمان، دقت و یا بازیابی را با استفاده از آن‌ها به دست آورد اما هنوز، نیاز به شناسایی سهم سیستم و کاربران در عملکرد مطلوب یک سیستم می‌باشد. بنابراین اکثر سیستم‌های ارزیابی موجود، از ارزیاب انسانی بهره می‌گیرند که در نتیجه، عملکرد یک سیستم از کاربری به کاربر دیگر متفاوت خواهد بود.. بر اساس مطالعات صورت گرفته در فرآیند ارزیابی انسانی، پارامترهای مختلفی مد نظر قرار می‌گیرند[۸]. در نتیجه برای جایگزینی یک مدل به جای انسان، نیازمند استخراج و اندازه‌گیری اتوماتیک این ویژگی‌ها خواهیم بود که این مسئله خود یکی از چالش‌های این حوزه می‌باشد. بنابراین در این مقاله با معرفی یکسری ویژگی جدید و ارائه یک مدل آماری، به ارزیابی اتوماتیک سیستم‌های IQA با استفاده از متن خروجی تولید شده از سوالات رد و بدل شده بین کاربر و سیستم پرداخته‌ایم بطوریکه مدل پیشنهادی می‌تواند جایگزین انسان در ارزیابی این سیستم‌ها شود.

ساختار مقاله پیشنهادی بدین صورت است که در بخش اول به مروری بر تحقیقات صورت گرفته در زمینه ارزیابی سیستم‌های QA و IQA پرداخته شده است. در بخش دوم روش پیشنهادی، سیستم پایه تولید شده در آزمایشگاه و مدل ارزیابی تشریح خواهد شد. در بخش سوم آزمایشات و نتایج بدست آمده تشریح شده است و در بخش آخر به نتیجه گیری و پیشنهادات پرداخته شده است.

کارهای مرتبط

بیشتر ارزیابی‌های صورت پذیرفته در حوزه ارزیابی سیستم‌های QA توسط TREC[5] انجام شده که اکثر این ارزیابی‌ها به جای اینکه مبتنی بر سیستم باشد بر اساس نظرات کاربران صورت گرفته است[۸]. ارزیابی سیستم‌های QA بسته به ارزیابی سوالات پیچیده یا ساده متفاوت است. از روش‌های ارزیابی در سیستم‌های QA می‌توان به استفاده مجموعه‌ای از سوالات و پاسخ‌ها به نام « مجموعه استاندارد طلایی» اشاره کرد. در این روش توانایی یک سیستم بر اساس میزان منطبق بودن پاسخ سیستم با این مجموعه استاندارد طلایی سنجیده می‌شود. البته این روش برای سوالات پیچیده و مبهم هنوز تقویت نشده است[۹]. ساجدی و خانی[۱۰] اولین سیستم پرسش و پاسخ فارسی در دامنه نامحدود و وب مبنا به نام جویا را معرفی کردند. جویا بصورت کلی شامل زیرمولفه‌های پردازش پرسش، بازیابی اطلاعات و استخراج جواب دقیق می‌باشد. به دلیل نبود مجموعه داده ارزیابی در زبان فارسی مجموعه داده ارزیابی برای این سیستم تهیه شده است. مجموعه داده ارزیابی شامل ۴۱۲ پرسش متنوع، و پاسخ متناظر آن است. سیستم پیشنهادی به ۸۰ درصد صحت دست پیدا کرده است. امروزه اکثر روش‌های پیاده‌سازی شده در زمینه ارزیابی سیستم‌های QA، از معیارهایی همانند MRR[6]، C@1، CWS[7]، K1 و دیگر موارد استفاده می‌کنند که هر کدام از این روش‌ها خود دارای نقاط ضعف بوده و قابلیت تعمیم به همه سیستم‌های مختلف QA را ندارند[۱۱]. این یکی از معضلات برای استفاده این روش‌ها، در ارزیابی سیستم‌های IQA است. از طرفی این معیارها بیشتر در جهت انتخاب پاسخ بکار گرفته می‌شدند و توانایی سیستم را در این راستا مورد ارزیابی قرار می‌دادند[۱۱]. بنابراین ایجاد یک روش کلی که بتواند به ارزیابی سیستم‌های IQA کمک نماید امری ضروری به نظر می‌رسد. سان [۱۲]، روشی را برای ارزیابی سیستم‌های IQA معرفی نمود که X-EVAL نامیده می‌شد. این روش ارزیابی، به تعیین میزان مشارکت عوامل موثر در یک سیستمIQA، برای رسیدن کاربر به نتایج نهایی می‌پرداخت. مطالعه صورت گرفته توسط آن‌ها، شامل گزارش تجربی از سه سیستم IQA تعاملی و یک سیستم پایه بود. آن‌ها تاکید داشتند که هدف از این گزارش ارزیابی سیستم نیست بلکه این کار را به منظور ارزیابی اثر X-EVAL در

[4] Interactive Question Answering system (IQA)
[5] Text Retrieval Evaluation Conference
[6] Mean Reciprocal Rank
[7] Confidence Weighted Score

تشخیص پاسخ صحیح در میان این چهار سیستم بکار گرفته‌اند. ویلیام هرش [۱۳] در مقاله خود به بررسی عوامل تاثیرگذار در ارزیابی پرداخت. ایشان ابتدا به تحلیل مجموعه سوالات تعاملی TREC پرداخته، سپس یک سیستم پرسش و پاسخ پزشکی را در نظر گرفتند و توسط دو گروه از دانشجویان به بررسی این سیستم پرداختند و نشان دادند که عواملی مانند سن، جنسیت، تجربه در استفاده از کامپیوتر، نگرش نسبت به کامپیوتر و چندین ویژگی دیگر جزء عوامل تعیین کننده در موفقیت یک سیستم QA می‌باشد. کوارترونی و ماناندهار[۱۴] روشی را ارائه نمودند که شامل یک ارزیابی کیفی از سیستم‌های پرسش و پاسخ تعاملی بود. آن‌ها در روش خود تعدادی پرسش مطرح کردند و از کاربران خواستند با دادن امتیازی بین یک (حداقل امتیاز) تا پنج (حداکثر امتیاز) کیفیت تعامل را اندازه‌گیری نمایند. سوالات تهیه شده در پرسش‌نامه برای ارزیابی، شامل بررسی عملکرد سیستم، مشکلات تعامل، سرعت پاسخگویی و رضایت کلی کاربر از سیستم بود. واکلدر و همکارانش [۱۵] در مقاله خود به توسعه المان‌های موجود در روش ارزیابی، برای سیستم HITIQA پرداختند. در این گزارش دو هدف اساسی پیگیری شد. نخست یک ارزیابی واقع بینانه از سودمندی و قابلیت استفاده از HITIQA به عنوان یک سیستم تعاملی ارائه گردید و سپس به توسعه معیارهای مقایسه پاسخ به دست آمده، توسط تحلیلگران مختلف و ارزیابی کیفیت پشتیبانی این سیستم صورت پذیرفت. آن‌ها از ابزار کمی و کیفی برای به دست آوردن اطلاعات در مورد راحتی تحلیلگر با سیستم HITIQA استفاده کردند خصوصاً آنکه از ویژگی‌های جدیدی در سنجش توانایی یافتن پاسخ به سوالات پیچیده و گفت و گو تعاملی بهره گرفته بودند. منصوری و حسن-پور [۱۶] برای یک سیستم QA، از دانش موجود در سوالاتی که قبلاً در این سیستم بین کاربران و سیستم رد و بدل شده بود، برای پاسخ‌دهی به سوالات جدید استفاده نمودند. آن‌ها با ارائه یک الگوریتم، مجموعه‌ای از قطعات کارآمد ایجاد کردند تا با استفاده مجدد از این قطعات، بهبود بازدهی پاسخ به سوالات بعدی را برای سیستم فراهم نمایند. کلی [۱۷] به ارزیابی عملکرد چهار سیستم IQA با کاربر واقعی در مقاله خود پرداخته است. آن‌ها به دنبال شناسایی پتانسیل معیارهای ارزیابی برای سیستم-های IQA، با استفاده از تجزیه و تحلیل نظرات ارزیابی ساخته شده توسط کاربران، برای چنین سیستم‌هایی بودند. آن‌ها در کار خود از داده‌هایی کیفی که از تحلیلگران اطلاعاتی در طول مصاحبه‌ها جمع‌آوری نموده بودند بهره جستند.

روش پیشنهادی

با توجه به اینکه در فرآیند ارزیابی سیستم‌های پرسش و پاسخ تعاملی، افراد متخصص و خبره نقش دارند. حدس اینکه این افراد در پس زمینه ذهن خود از چه تابع ارزیابی برای نمره دهی به یک سیستم استفاده می‌نمایند یکی از چالش‌های موجود در زمینه ارزیابی سیستم‌های IQA می‌باشد. بنابراین ایجاد یک مدل ارزیابی که کمترین خطا را نسبت به نظرات موجود داشته باشد امری ضروری است. از آنجایی‌که ویژگی‌های متعددی در ارزیابی یک سیستم IQA دخالت دارند و اندازه‌گیری اتوماتیک آن‌ها برای ایجاد یک مدل دارای اهمیت می‌باشد. در این مقاله مدلی مبتنی بر ویژگی‌های جدید برای ارزیابی اتوماتیک این سیستم‌ها پیشنهاد شده است. اساس روش پیشنهادی بر روی معرفی ویژگی‌های جدید برای پیش‌بینی نظرات می‌باشد.

سیستم تعاملی پایه

جهت ارزیابی سیستم‌های IQA نیاز به دسترسی به این سیستم-ها می‌باشد. بر این اساس علاوه بر سیستم‌های موجود، برای راحتی و دسترسی آسان‌تر از سیستم تعاملی پایه طراحی شده در آزمایشگاه وب‌کاوی دانشگاه صنعتی شاهرود استفاده گردید. این سیستم از تکنیک‌های آماری جهت پاسخ به سوالات کاربران بهره گرفته و مستقل از زبان عمل می‌نماید. بنابراین با در اختیار داشتن پایگاه دادگان مناسب هر زبان این سیستم می‌تواند به سوالات مطرح شده به آن زبان پاسخ دهد[۱۸]. جهت آموزش سیستم طراحی شده، از سه پایگاه دادگان فارسی با نام‌های WMPR-QA1-2015، WMPR-QA2-2015 و WMPR-QA3-2015 استفاده شده است. پایگاه دادگان اول با نام WMPR-QA1-2015 دارای چهار فایل متنی با محتوای آئین نامه آموزشی دانشگاه صنعتی شاهرود است که در قالب ۲۹۲ جمله و با فرمت UTF-8گردآوری شده و به عنوان داده آموزشی شناخته می‌شود. ۸۱ پرسش و پاسخ مطرح شده از این آئین نامه نیز به عنوان مجموعه تست پایگاه دادگان فوق در نظر گرفته شد. پایگاه دادگان دوم با نام WMPR-QA2-2015 دارای یک فایل متنی با محتوای آئین نامه مالی شهرداری‌ها می‌باشد که در قالب ۷۵ جمله و با فرمت UTF-8 گردآوری شده و از آن به عنوان مجموعه آموزش استفاده شده است. ۳۳ پرسش و پاسخ مطرح شده از این آئین نامه نیز به عنوان مجموعه تست پایگاه دادگان WMPR-QA2-2015 در نظر گرفته شد. پایگاه دادگان سوم با نام WMPR-QA3-2015 شامل دو مجموعه آموزش و تست می‌باشد. مجموعه آموزش آن دارای یک فایل متنی با محتوای آئین نامه استخدام هیات علمی دانشگاه‌ها می‌باشد که در قالب ۲۵۶ جمله و با فرمت UTF-8 گردآوری شده است و مجموعه تست آن در بردارنده ۳۱ پرسش و پاسخ مطرح شده از این آئین نامه می‌باشد. سه پایگاه دادگان فوق از وب سایت آزمایشگاه وب کاوی و

شناسایی الگو دانشگاه صنعتی شاهرود[8] قابل دریافت می‌باشند. بررسی نظرات ارائه شده توسط کاربران نشان دهنده رضایت آن-ها از کیفیت تعامل برقرار شده با سیستم بود[۱۸]. در راستای استفاده بهینه از سیستم و افزایش عملکرد و کارایی، تغییراتی در آن ایجاد گردید که منجر به عملکرد بهتر سیستم گردید. نتایج حاصل از این بهینه‌سازی در [۱۹] ارائه شد.

پیش‌پردازش

در این مرحله متن ورودی به ساختاری قابل پردازش برای مراحل بعد تبدیل می‌گردد. پیش پردازش متن‌ها شامل ۵ گام می‌باشد:

۱- مشخص کردن مرز جمله‌ها: مرز جمله‌ها از طریق بررسی علائم جدا کننده از قبیل فضای خالی، "."، "!"، "؟"، ":" و دیگر موارد انجام شد.

۲- ریشه‌یابی: در این حالت یک کلمه به شکل عمومی خود کاهش می‌یابد که این شکل عمومی باید برای همه کلمات هم ریشه یکسان باشد. روش استفاده شده برای اینکار مبتنی بر حذف پسوندها و پیشوندها می‌باشد.

۳- حذف کلمات و واژه‌های غیر مهم: در این مرحله کلماتی که در محتوای اصلی متن تاثیری ندارند (stop words) حذف گردید. برای اینکار لیستی مشتمل بر ۲۰۰ کلمه آماده گردید.

۴- شناسایی مقادیر عددی: این کلمات بعد از شناسایی، برچسب مقدار عددی دریافت می‌کنند.

۵- یکسان‌سازی متن‌ها: در متون انگلیسی تمامی کلماتی که با حروف بزرگ بودند به حروف کوچک تبدیل شدند و در متون فارسی یکسان‌سازی حروف (مثل حروف "ی" و "ک") صورت پذیرفت.

تمامی این کارها به صورت اتوماتیک صورت گرفت و جواب نهایی توسط ناظر انسانی کنترل شد.

استخراج ویژگی

یکی از مهمترین قسمت‌های مربوط به هر سیستم تشخیص یا مدل‌سازی، استخراج ویژگی می‌باشد. هر چه این مرحله با دقت بالاتری صورت پذیرد، نتایج حاصل از مراحل بالاتر دارای دقت بالاتری خواهد بود. در این مرحله، تعدادی ویژگی تعریف گردید که در ادامه به معرفی هر یک از این ویژگی‌ها خواهیم پرداخت. با توجه به اینکه خروجی هر مکالمه صورت گرفته بین کاربران و سیستم‌ها به صورت مجموعه‌ای از سوال‌ها و پاسخ‌ها آماده گردید، بعضی از ویژگی‌های تعریف شده علاوه بر اینکه برای مجموعه سوال- جواب مورد استفاده قرار گرفتند، به صورت جداگانه برای مجموعه سوال‌ها و مجموعه جواب‌ها نیز بکار گرفته شدند.

- ویژگی اول:

N-گرم‌ها یکی از مشهورترین مدل‌های آماری زبان می‌باشند. مدل‌های n-گرم براساس هم‌پیوندی و کنار هم قرار گرفتن کاراکترهای لغات در پردازش متن عمل می‌نمایند. به عبارت دیگر، در این مدل‌ها ارتباطات زنجیره‌ای کلمات در نظر گرفته می‌شود. در این ویژگی برای n=1,2,3 ابتدا n-گرم‌های مشترک را شمرده با یکدیگر جمع و بر تعداد کل n-گرم‌ها تقسیم می-نماییم (رابطه ۱). اینکار را برای مجموعه‌های پرسش و پاسخ (Q-A)، مجموعه سوال‌ها (Q-Q) و مجموعه جواب‌ها (A-A) صورت پذیرفت.

$$\text{Count\_N} = \sum_{S_i \in conv} \frac{\sum_{ngram \in S_i} Count_{match}(gram_n)}{\sum_{ngram \in S_i} Count(gram_n)} \qquad (۱)$$

S_i، i-امین جمله از هر مکالمه و n طول هر n-گرم می‌باشد.

- ویژگی دوم:

در یک مکالمه، برای n های بزرگتر، هر چه تعداد n-گرم‌های مشترک بیشتر باشد امتیاز آن مکالمه بیشتر خواهد بود و احتمال پیوستگی متن مکالمه نیز بیشتر خواهد شد [۲۰]. بر این اساس در این ویژگی پیشنهادی، هر کدام از n-گرم‌ها، بر اساس ارزش یک ضریب وزنی برای هر n-گرم به ارزش W_i با یکدیگر جمع می‌شوند تا مقدار این ویژگی بدست آید(رابطه ۲).

$$\text{Count\_Weight\_N} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{\sum_{ngram \in S_i} W_k \times Count_{match}(gram_n)}{\sum_{ngram \in S_i} Count(gram_n)} \qquad (۲)$$

M تعداد عضوهای مجموعه و W_k ضریب تاثیر هر n-گرم و مقدار آن متناسب با عدد n می‌باشد. این ویژگی نیز برای n=۱،۲،۳ محاسبه گردید.

- ویژگی سوم:

انطباق پشت سر هم در سطح جمله معمولاً در n-گرم‌ها دیده می‌شود. بنابراین به دلیل در نظر گرفتن بزرگترین رشته مشترک در n-گرم‌ها، طول تعداد کاراکتر تعریف نمی‌شود. اما در بزرگترین رشته مشترک(LCS[9]) نیاز به انطباق پشت سرهم رشته کلمات نمی‌باشد و بزرگترین رشته مشترک بین دو عبارت حاصل می‌گردد. از طرفی برای اینکه مسئله همرخدادی نیز در جملات در نظر گرفته شود از معیار F استفاده نمودیم. در رابطه تعریف شده، برای یک مکالمه ابتدا یک جفت سوال- پاسخ را در نظر گرفته، سپس برای هر جفت بازیابی و دقت را محاسبه و

[8] http://wmpr.ir/fa/index/category/53

[9] Longest Common Substring

برای تمامی جفت سوال- پاسخ اینکار را انجام می‌دهیم. در نهایت پاسخ بدست آمده را در رابطه معیار F قرار داده و امتیاز هر مکالمه را محاسبه می‌کنیم.

$$R_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{LCS(Q_i, A_i)}{L_{Q_i}} \quad (۳)$$

$$P_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{LCS(Q_i, A_i)}{L_{A_i}} \quad (۴)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۵)$$

$\beta = \frac{P_{LCS}}{R_{LCS}}$ و M تعداد جفت سوال-پاسخ هر مکالمه می‌باشد.

- ویژگی چهارم:

در این ویژگی فرض کردیم که Q_i و مجموعه جواب‌ها به ترتیب شامل U جمله با P کلمه و V جمله با n کلمه باشد. در این ویژگی اجتماع بزرگترین زیر رشته مشترک بین Q_i و مجموعه جواب‌ها را محاسبه نمودیم که هر چه این عدد بزرگتر باشد، ارتباط بین جملات در مکالمه بیشتر است. برای کل مجموعه سوالات یک مکالمه اینکار را انجام دادیم. روابط به صورت زیر پیشنهاد گردید.

$$R_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{\sum_{j=1}^{U} LCS_{\cup}(Q_i, A)}{P} \quad (۶)$$

$$P_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{\sum_{j=1}^{V} LCS_{\cup}(Q_i, A)}{n} \quad (۷)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۸)$$

M تعداد سوالات در یک مکالمه و $\beta = \frac{P_{LCS}}{R_{LCS}}$ خواهد بود.

- ویژگی پنجم:

در این دسته ویژگی فرض گردید که دو مجموعه S_i و S_j داریم که S_i از N جمله با K کلمه و S_j با P جمله با T کلمه می‌باشند. روابط زیر به عنوان روابط جدید در محاسبه امتیاز هر مکالمه مطرح شدند.

$$R_{LCS} = \frac{1}{N} \times \sum_{S_i \in S_1} \max_{S_j \in S_2} (LSC(S_i, S_j)) \quad (۹)$$

$$P_{LCS} = \frac{1}{P} \times \sum_{S_i \in S_1} \max_{S_j \in S_2} (LSC(S_i, S_j)) \quad (۱۰)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۱۱)$$

در رابطه ۱۱ مقدار β =1 در نظر گرفته شد. همچنین مجموعه-های S_1 و S_2 را به صورت (Q_i, Q_{i+1})، (A_i, A_{i+1})، (Q_i, A_i)، (Q_{i+1}, A_i) و (Q_i, A_{i+1}) در نظر گرفتیم. و به ازای هر مجموعه مقدار این ویژگی محاسبه گردید.

- ویژگی ششم:

n-گرم‌های مشترک بین مجموعه سوالات و پاسخ‌ها را بدست آورده و بعد از نرمال‌سازی به عنوان ارزش یک مکالمه گزارش می‌کنیم. در این ویژگی، با فرض اینکه دو مجموعه از سوال‌ها و جواب‌ها داریم براساس روابط زیر میزان امتیاز هر Q_i با مجموعه جواب‌ها محاسبه و در نهایت توسط رابطه ۱۴ امتیاز مکالمه را محاسبه می‌نماییم. اینکار برای n=۳,۲ که بر روی مجموعه داده ما دارای نتایج بهتری بود انجام شد.

$$R_{skip_n} = \frac{1}{t} \times \frac{1}{k} \times \sum_{i=1}^{t} \sum_{j=1}^{k} \frac{skip_n(Q_i, A_j)}{C(m,n)} \quad (۱۲)$$

$$P_{skip_n} = \frac{1}{t} \times \frac{1}{k} \times \sum_{i=1}^{t} \sum_{j=1}^{k} \frac{skip_n(Q_i, A_j)}{C(L,n)} \quad (۱۳)$$

$$F_{skip_n} = \frac{1+\beta^2 \times R_{skip_n} \times P_{skip_n}}{R_{skip_n} + \beta^2 \times P_{skip_n}} \quad (۱۴)$$

که در آن t تعداد سوالات، k تعداد پاسخ‌ها، n اندازه n- گرم ، m طول سوال Q_i، L طول پاسخ A_j و B=1 در نظر گرفته شد.

- ویژگی هفتم (امتیاز دهی به جملات):

در این ویژگی، یک جفت سوال و جواب به صورت یک جمله در نظر گرفته شد. سپس، برای هر یک از جملات مکالمه امتیازی محاسبه گردید. نحوه امتیازدهی به کلمات و جملات بدین صورت است که ابتدا امتیاز مربوط به کلمات را محاسبه و بر اساس امتیاز بدست آمده برای کلمات، با استفاده از رابطه ۱۶، امتیاز هر جمله محاسبه می‌شود. در نهایت با توجه به امتیازات بدست آمده برای هر جمله، امتیاز نهایی برای هر مکالمه محاسبه گردید.

$$Word\_score = K \times f_{word} \quad (۱۵)$$

$$Sentence\_score = \sum Word\_score \quad (۱۶)$$

K یک عدد ثابت و f_{word} تعداد تکرار کلمه در متن می‌باشد. نحوه محاسبه امتیاز هر مکالمه بدین صورت است که به هر جمله با توجه به موقعیت مکانیش ، امتیاز متفاوتی تخصیص داده شد. نحوه امتیاز دهی بدین صورت است که معمولاً جملات میانی دارای ارزش اطلاعاتی بالاتری نسبت به جملات ابتدایی و پایانی هر مکالمه هستند (بر اساس مجموعه داده تهیه شده این فرض صورت پذیرفت). بنابراین براساس موقعیت قرارگیری جملات ارزش‌گذاری برای هر جمله به صورت زیر پیشنهاد گردید.

$$P_{score_i} = \begin{cases} 1 - \frac{n-i+1}{n} & i <= \frac{n}{2} \\ 1 - \frac{i-3}{n} & \frac{n}{2} < i <= n \end{cases} \quad (۱۷)$$

i موقعیت هر جمله و n تعداد جملات هر مکالمه می‌باشد. با توجه به ارزش هر جمله و امتیاز آن امتیاز نهایی یک مکالمه محاسبه می‌شود (رابطه ۱۸).

$$S_{conversation} = \frac{1}{N} \times \sum_{j=1}^{N} (Sentence\_score_j + P_{score_i}) \quad (۱۸)$$

با توجه به مطالعات صورت گرفته در حوزه ارزیابی، معمولاً پارامترهای تطبیق نام موجودیت[10] (ویژگی که مشخص می‌کند که آیا تمام موجودیت های ظاهر شده به طور مثال در سوال Q_{i+1} در سوال Q_i قرار دارند یا نه)، تطبیق هدف[11] (تعیین کننده میزان شباهت بطور مثال بین نوع سوال Q_{i+1} و سوال Q_i می‌باشد)، فهم سوال، پوشش[12]، گستردگی[13]، تمامیت[14]، ارتباط[15]، دقت[16] و حجم[17] جزء پارامترهایی است که در نمره‌دهی توسط ارزیاب‌ها در نظر گرفته می‌شود[۸]. هر کدام از روابط معرفی شده سعی بر پوشش این ابعاد از نظرات ارزیاب‌ها را دارند. بعضی از ویژگی‌های پیشنهادی تنها یک بعد و بعضی از ویژگی‌ها چندین بعد از دیدگاه ارزیاب را پوشش می‌دهد. جدول ۱ اینکه چه پارامتری توسط کدام ویژگی پوشش داده شده را نمایش می‌دهد.

انتخاب ویژگی

مساله انتخاب ویژگی، یکی از مسائلی است که در شناسائی آماری الگو مطرح است. این مساله در طبقه‌بندی، استخراج مدل و دیگر موارد مسائل بینایی ماشین اهمیت به سزائی دارد. در مسائل مختلف، معمولاً تعداد زیادی ویژگی وجود دارد، که بسیاری از آن‌ها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد. در پروسه انتخاب ویژگی، الگوریتم‌های متعددی وجود دارد که متناسب با کار مورد نظر می‌توان از آن‌ها استفاده نمود. از بین روش‌های موجود برای انتخاب ویژگی،

جدول۱. نمایش ابعاد پوشش پارامترهای ارزیابی بوسیله ویژگی‌های معرفی شده

پارامتر	ویژگی
تطبیق نام موجودیت	اول، دوم، سوم
تطبیق هدف	چهارم، پنجم، ششم
فهم سوال	-------
پوشش	اول تا هفتم
گستردگی	سوم تا هفتم
تمامیت	-------
ارتباط	اول تا هفتم
دقت	پنجم، ششم، هفتم

روش انتخاب ویژگی به صورت بازگشتی حذف شونده (RFE[18]) انتخاب گردید. دلیل عمده آن را می‌توان مقیاس پذیری این روش و سادگی استفاده و محاسبات سریع آن دانست. در روش RFEتعداد ویژگی‌ها بصورت بازگشتی کاهش یافته و در هر مرحله دقت طبقه‌بند محاسبه و ویژگی‌هایی که بالاترین دقت را در طبقه‌بندی ایجاد می‌کنند، انتخاب می‌شوند. ملاک انتخاب ویژگی‌ها وزن آن‌ها است. بعد از اعمال این روش، تعداد ۱۹ ویژگی باقی ماندند.

مدل‌سازی

تحلیل رگرسیون، تکنیکی آماری، برای بررسی و مدل سازی ارتباط بین متغیر وابسته و متغیر مستقل بوده و هدف آن پیش‌بینی متغیر وابسته از روی متغیرهای مستقل می‌باشد. با توجه به اینکه هدف ما ارائه یک مدل آماری جهت ارزیابی سیستم‌های IQA می‌باشد، بنابراین با توجه به نظرات موجود کاربران در کار با این سیستم‌ها و استخراج ویژگی‌ها از روی متن-های تولید شده، به دنبال استفاده از رگرسیون برای پیش‌بینی این نظرات هستیم. در روش پیشنهادی ویژگی‌های استخراج شده به عنوان متغییرهای وابسته و نظرات انسانی به عنوان متغییر مستقل در نظر گرفته شد. جهت انتخاب نوع درست از انواع رگرسیون، مدل‌های مختلف از رگرسیون بررسی گردید تا بهترین مدل انتخاب گردد.

رگرسیون خطی برای پیش‌بینی نظرات

رگرسیون خطی به بررسی رابطه یک متغیر مستقل (پیش‌بین) و یک متغیر وابسته می‌پردازد. حال اگر تعداد متغیرهای مستقل در این رابطه خطی بیش از یک شود، مدل رگرسیون، خطی چندگانه نامیده می‌شود. معادله رگرسیون خطی ساده به شکل $Y=AX+B$ و رگرسیون خطی چند گانه به صورت $Y =$

[10] Named entity matching
[11] Target Matching
[12] Coverage
[13] Extensiveness
[14] Completeness
[15] Relevance
[16] Accuracy
[17] Size
[18] Recursive feature elimination

$a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$ می‌باشد که در آن Xها ویژگی‌های استخراج شده و b ضرایب تاثیر هر یک از ویژگی‌ها باشد. بنابراین این ضرایب باید طوری محاسبه گردند که معیار حداقل مربعات خطا را تامین نمایند. مقدار حاصل برای a بیانگر مقادیر پیش‌بینی شده Y با ثابت ماندن مقادیر X است. از طرفی با مقایسه اندازه مقادیر ضرایب b با همدیگر اولویت و میزان تاثیر هر یک از عوامل مشخص می‌شود. همچنین علامت ضرایب هم بر تغییرات متغییر وابسته تاثیرگذارند.

رگرسیون غیرخطی برای پیش‌بینی نظرات

رگرسیون غیرخطی مدل های مختلفی دارد که از جمله آن می‌توان به مدل‌های درجه ۲ به بالا، چندجمله‌ای نمایی، توانی و دیگر موارد، اشاره کرد که متناسب با مدل انتخاب شده معادلات آن ها متفاوت خواهد بود. متناسب با آزمایشات، تعدادی از معادلات رگرسیون غیرخطی در نظر گرفته شده و بر روی ماتریس ویژگی مورد آزمایش قرار گرفت تا بهترین پاسخ انتخاب گردد.

معیار ارزیابی

برای ارزیابی نتایج حاصل از مدل بدست آمده با داده‌های واقعی سه سنجه آماری ضریب تعیین R^2، مجذور میانگین مربعات خطا (RMSE[19]) و درصد میانگین مطلق خطا (MAPE[20]) مورد استفاده قرار گرفت. روابط مربوط به این معیارها در زیر نشان داده شده است.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^{\,2}}{n}} \qquad (۱۹)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{Y_i}\right| \times 100 \qquad (۲۰)$$

که در این روابط n تعداد پیش‌بینی ها و e_i خطای پیش‌بینی است که از تفاوت مقادیر پیش‌بینی شده و مقادیر واقعی بدست می‌آید و Y_i مقادیر واقعی می‌باشد.

نتایج آزمایشات

به جهت انتخاب بهترین مدل در جهت مدلسازی نظرات انسانی با توجه به وجود انواع مختلف رگرسیون، انواع مختلفی از رگرسیون‌ها مورد استفاده قرار گرفت ولی تنها نتایج حاصل از رگرسیون خطی چندگانه و غیرخطی توانی در ادامه آورده شده است.

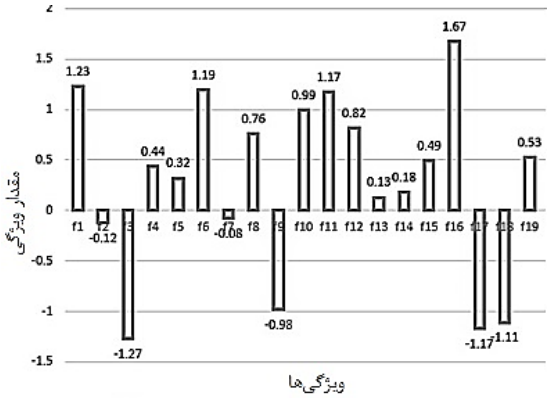
پایگاه داده

بدلیل نبود پایگاه داده ا ستاندارد در زمینه ارزیابی سیستم‌های IQA، نیاز به ایجاد یک پایگاه داده مناسب از سوالات رد و بدل شده بین سیستم و کاربر با برچسب گذاری مناسب بود. بر این اساس، علاوه بر سیستم تعاملی پایه طراحی شده، سه سیستم دیگر تعاملی موجود در نظر گرفته شد. برای یکپارچه‌سازی شرایط کار با این سیستم‌ها و راحتی کاربران، سامانه‌ای تحت وب طراحی گردید که متن تبادل شده و امتیاز داده شده توسط کاربران به سیستم‌ها به صورت اتوماتیک در پایگاه داده ذخیره گردد. برای ایجاد پایگاه داده، تعداد ۱۲۰ کاربر برای ۵ موضوع مختلف با سیستم کار نمودند و با توجه به موضوع مکالمات هر یک به صورت جداگانه ذخیره گردید. از این مجموعه ۶۰۰ تایی، ۵۴۰ نمونه توسط فرد خبره به عنوان نمونه مناسب‌تر انتخاب شد. شرکت‌کنندگان در این دوره ارزیابی شامل دانشجویان دانشگاه و زبان آموزان یک موسسه بودند. با توجه به اینکه یک سیستم با زبان فارسی و سه سیستم دیگر با زبان انگلیسی عمل می‌نمود از افرادی در این راستا بهره گرفته شد که در حد متوسط به زبان انگلیسی مسلط باشند. تعداد ۸۴ نفر از شرکت‌کنندگان دارای مدرک لیسانس، ۳۰ نفر فوق‌لیسانس و ۶ نفر دانشجوی دکتری از رشته‌های مختلف مهندسی بودند. جدول ۲، یک نمونه از خروجی متن ذخیره شده (بدون فرمت) از تعامل با سیستم پایه را نشان می‌دهد.

جدول ۲. نمونه تولید شده از یک مکالمه در سیستم IQA

کاربر	سیستم
اگر معدل دانشجو کمتر از ۱۴ شود چه باید کرد؟	دانشجو مربوط به چه مقطعی است؟
دانشجو کارشناسی ارشد	منظور شما معدل کل هست؟
خیر، معدل یک ترم	در صورت کمتر شدن معدل از ۱۴ در هر نیمسال دانشجو مشروط تلقی می‌گردد.
دانشجوی کارشناسی چطور؟	در صورت کمتر شدن معدل از ۱۲ در هر نیمسال دانشجو مشروط تلقی می‌گردد.
چند ترم متوالی مشروط شدن دانشجو اخراج می‌شود؟	دانشجو کارشناسی یا ارشد؟
دانشجو کارشناسی	در صورتیکه دانشجو کارشناسی سه ترم متوالی یا چهار ترم غیر متوالی مشروط شود در آستانه اخراج قرار می‌گیرد.

[19] Root Mean Square Error
[20] Mean Absolute Percent Error



شکل ۱. ضرایب معادله رگرسیون خطی

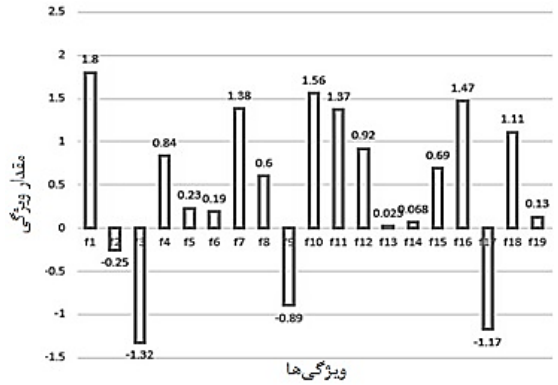
نتایج حاصل از پیاده‌سازی رگرسیون خطی

در روش رگرسیون خطی ویژگی‌های استخراج شده به عنوان متغییر مستقل و نظرات انسانی به عنوان متغییر وابسته در نظر گرفته شدند. مقدار پارامتر sig برای تمامی ضرایب بدست آمده در رگرسیون خطی چندگانه کمتر از ۰/۰۵ بود که این به معنای قابل قبول بودن ضرایب بدست آمده برای ویژگی‌ها می‌باشد. ضرایب رگرسیونی بدست آمده با توجه به معادله رگرسیون خطی چندگانه در شکل ۱ نمایش داده شده است. مقدار ضریب تعیین R^2 برای معادله برابر ۰/۷۶، RMSE برابر ۰/۲۴ و MAPE برابر ۷٪ حاصل گردید. همانطور که در شکل ۱ نشان داده شده، مقادیر بدست آمده دارای علامت و مقدار متفاوتی برای هر متغییر می باشند ویژگی هایی که دارای مقدار کم می‌باشند دارای تاثیر کمتر در نظرات انسانی و آن‌هایی که دارای مقادیر بزرگتر هستند حاکی از تاثیر بیشتر این ویژگی‌ها در خروجی می‌باشند. همچنین با توجه به ضرایب بدست آمده، می‌توان تغییرات نظرات انسانی را نسبت به متغییرهای مستقل نیز بدست آورد.

نتایج حاصل از پیاده‌سازی رگرسیون غیرخطی

در رگرسیون غیر خطی از انواع مختلفی برای مدل‌سازی استفاده گردید و همینطور ضرایب با مقادیر اولیه متفاوتی تست گردید تا بهترین مدل انتخاب گردد. با توجه به آزمایشات متعدد انجام شده، بهترین مدلی که برای داده‌ها استخراج گردید مدل توانی رگرسیون غیرخطی بود که معادله آن برای ارزیابی نظرات به صورت زیر حاصل گردید.

$$Y = X_1^{1.80} X_2^{-0.25} X_3^{-1.32} \ldots\ldots X_{22}^{0.16} \qquad (۲۱)$$

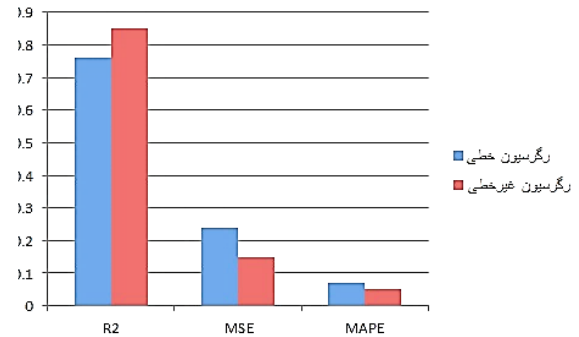


شکل ۲. ضرایب معادله رگرسیون غیرخطی

ضرایب مربوط به این معادله در شکل ۲ آورده شده است. قابل توجه است که ارزیابی صورت گرفته بر اساس سنجه‌های آماری در نظر گرفته، مقدار R^2 برابر ۰/۸۵، RMSE برابر ۰/۱۵ و MAPE برابر ۵٪ حاصل گردید. شکل ۳ مقایسه بین نتایج حاصل از ارزیابی‌ها را برای رگرسیون خطی و غیر خطی نمایش می‌دهد. نتایج حاکی از ارائه دقت بیشتر رگرسیون غیرخطی و برتری آن نسبت به رگرسیون خطی می‌باشد.

نتیجه گیری

در این مقاله روشی اتوماتیک برای مدلسازی نظرات ارزیابی انسانی بر اساس متن خروجی یک سیستم IQA ارائه شد که مشابه با آن در هیچ یک از کارهای قبلی مشاهده نگردید. در روش پیشنهادی ابتدا مجموعه‌ای از ویژگی‌های آماری جدید پیشنهاد گردید که بتوانند به خوبی در مدلسازی نظرات کارایی داشته باشند. سپس از روی متن‌های موجود در پایگاه داده ویژگی‌ها استخراج و براساس آن ماتریس ویژگی تشکیل گردید. برای مدل سازی نظرات از رگرسیون خطی و غیرخطی بهره گرفته شد.



شکل ۳. مقایسه نتایج حاصل از مدلسازی با رگرسیون خطی و غیر خطی

نتایج ارائه شـــده بر اســـاس معیارهای ارزیابی R^2، RMSE و MAPE رگرسیون غیرخطی توانی دارای دقت بالاتری نسبت به رگرســـیون خطی چندگانه بود که نشـــان دهنده پایداری مدل پیشـــنهادی اســـت. همچنین پیشـــنهاد می‌گردد که با توجه به ضـــرایب بدســـت آمده و مقدار آن‌ها می‌توان تاثیر هر یک از ویژگی‌ها را بر روی خروجی م شخص کرد و از طرفی با توجه به همبســـتگی بین ویژگی ها به کاهش ویژگی ها پردا خت تا پیچیدگی معدلات بدست آمده به مراتب کمتر گردد.

مراجع

[1] Voorhees, E.M, the TREC question answering track. Nat.Lang. Eng. 7 (4), 361–378, 2001.

[2] Voorhees, E.M, Overview of the TREC 2003 question answering Trac, Twelth Text REtrieval Conference, Volume 500–255 of NIST Special Publications, Gaithersburg, MD. National Institute of Standards and Technology, 2004.

[3] Burger, John, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin et al. "Issues, tasks and program structures to roadmap research in question & answering (Q&A)." In Document Understanding Conferences Roadmapping Documents, pp. 1-35. 2001.

[4] Li, Xin, and Dan Roth. "Learning question classifiers." In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pp. 1-7. Association for Computational Linguistics, 2002.

[5] Mishra, Amit, and Sanjay Kumar Jain. "A survey on question answering systems with classification." Journal of King Saud University-Computer and Information Sciences 28, no. 3: 345-361, 2016.

[6] Lopez, Vanessa, Victoria Uren, Marta Sabou, and Enrico Motta. "Is question answering fit for the semantic web? A survey." Semantic Web 2, no. 2: 125-155, 2011.

[7] M. Amit, and S. K. Jain. “A survey on question answering systems with classification.” Journal of King Saud University-Computer and Information Sciences 28, no. 3: 345-361, 2016.

[8] Hersh, William. "Evaluating interactive question answering." In Advances in Open Domain Question Answering, pp. 431-455. Springer, Dordrecht, 2008.

[9] Bouziane, Abdelghani, Bouchiha, Doumi, and Malki, Question Answering Systems: Survey and Trends, Procedia Computer Science, pp. 366-375, 2015.

[۱۰] ساجدی، خانی، "جویا: یک سیستم پرسش و پاسخ"، مجله علوم رایانشی، ص ۵۱-۶۶، زمستان ۱۳۹۵.

[11] Rodrigo, A., Penas, A., “A study about the future evaluation of Question-Answering system”, Knowledge-Based Systems, Volume 137, 2017.

[12] Sun, Ying, Paul B. Kantor, and Emile L. Morse. "Using cross-evaluation to evaluate interactive QA systems." Journal of the Association for Information Science and Technology 62, no. 9: 1653-1665, 2011.

[13] Hersh, William. "Evaluating interactive question answering." In Advances in Open Domain Question Answering, pp. 431-455. Springer, Dordrecht, 2008.

[14] Quarteroni, Silvia, and Suresh Manandhar. "Designing an interactive open-domain question answering system." Natural Language Engineering 15, no. 1: 73-95, 2009.

[15] Wacholder, Nina, Sharon G. Small, Bing Bai, Diane Kelly, Robert Rittman, Sean Ryan, Robert Salkin et al. "Designing a Realistic Evaluation of an End-to-end Interactive Question Answering System." In LREC. 2004.

[16] Mansoori, M., and H. Hassanpour. "Boosting passage retrieval through reuse in question answering." International Journal of Engineering 25, no. 3:187-196, 2012.

[17] Kelly, Diane, Paul B. Kantor, Emile L. Morse, Jean Scholtz, and Ying Sun. "Questionnaires for eliciting evaluation data from users of interactive question answering systems." Natural Language Engineering 15, no. 1: 119-141, 2009.

[۱۸] شهرآیینی، زاهدی، "سیستم پاسخگوی تعاملی با استفاده از تکنیک‌های هوش مصنوعی"، دانشگاه صنعتی شاهرود، دانشکده کامپیوتر و فناوری اطلاعات، پایان نامه ارشد، ۱۳۹۴.

[۱۹] محمدمهدی حسینی، مرتضی زاهدی، "بهبود پاسخ ارائه شده در سیستم‌های پرسش و پاسخ تعاملی با استفاده از شبکه عصبی"، هشتمین کنفرانس بین المللی فناوری اطلاعات و دانش، صفحات ۸۴-۹۱، ۱۳۹۵.

[20] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8. 2004.