مجموعه‌ای از ویژگی‌های آماری جدید برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی

محمدمهدی حسینی[۱]، دانشجوی دکترا؛ مرتضی زاهدی[۲]، استادیار؛ حمید حسن‌پور[۳]، استاد

۱- دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران- hosseini_mm@shahroodut.ac.ir
۲- دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران - zahedi@shahroodut.ac.ir
۳- دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات - دانشگاه صنعتی شاهرود - شاهرود - ایران- h.hassanpour@shahroodut.ac.ir

چکیده: ارزیابی نقش مهمی در سیستم‌های پرسش و پاسخ تعاملی ایفا می‌نماید. روش استانداردی وجود ندارد که به ارزیابی کلی این سیستم‌ها پرداخته باشد. مشکل اصلی در طراحی این سیستم‌ها، عدم امکان پیش‌گویی بخش تعاملی است. به همین منظور، باید انسان در فرآیند ارزیابی شرکت داشته باشد. در این مقاله مجموعه‌ای از ویژگی‌های آماری جدید ساخته‌شده بر اساس n-گرم‌ها و بزرگ‌ترین رشته مشترک برای ارزیابی سیستم‌های پرسش و پاسخ تعاملی معرفی‌شده است. چهار سیستم پرسش و پاسخ تعاملی موجود برای ایجاد پایگاه داده‌ای از مکالمات ردوبدل شده بین کاربران و سیستم‌ها استفاده گردید. خروجی‌های تولیدشده، تعداد ۵۴۰ نمونه به‌عنوان داده مناسب در نظر گرفته شد تا مجموعه تست و آموزش بر اساس آن ایجاد گردد. سپس پیش‌پردازش بر روی متن‌ها صورت پذیرفت و ویژگی‌های تعریف‌شده از متن مکالمه‌ها استخراج و بر اساس آن ماتریس ویژگی تشکیل گردید. درنهایت با استفاده از ماشین بردار پشتیبان به دسته‌بندی نظرات به دو گروه با امتیاز خوب و بد پرداخته شد. نتایج حاصل از ضریب همبستگی بین نظرات انسانی و نظرات حاصل از ویژگی‌های پیشنهادی حاکی از دقت بالای مجموعه ویژگی‌های ارائه‌شده، در ارزیابی سیستم‌های پرسش و پاسخ تعاملی است.

واژه‌های کلیدی: ارزیابی، سیستم پرسش و پاسخ تعاملی، ماشین بردار پشتیبان، ویژگی آماری.

A Set Statistical features for Evaluating Interactive Question Answering

Mohammad Mehdi Hosseini[1], PhD Student, Morteza Zahedi[2], Assistant professor, Hamid Hassanpour[3], Professor

1- Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran, Email: hosseini_mm@shahroodut.ac.ir
2- Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran, Email: zahedi@shahroodut.ac.ir
3- Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran, Email: h.hassanpour@shahroodut.ac.ir

Abstract: Evaluation plays an important role in the interactive question answering(IQA) systems. In the context of evaluating IQA systems, there is practically no specific methodology for evaluating these systems in general. The main problem with designing an assessment method for IQA systems lies in the fact that is rarely possible to predict interaction part. To this end, human needs to be involved in the evaluation process. In this paper, an appropriate model is presented by introducing a set of built-in features for evaluating IQA systems. To conduct the evaluation process, four IQA systems were considered, and then a database of conversation was exchanged between users and systems. After performing the preprocessing on the conversation, the statistical characteristics of the conversation was extracted and base on that characteristics matrix was formed. Finally, using SVM, human thinking divided into two groups. The correlation coefficient between human thinking and proposed set features indicated the high accuracy of set features presented in evaluating of IQA systems.

Keywords: Evaluation, Interactive Question Answering Systems, SVM, Statistical Feature.

تاریخ ارسال مقاله: ۱۳۹۶/۹/۷
تاریخ اصلاح مقاله: ۱۳۹۶/۱۱/۲۳
تاریخ پذیرش مقاله: ۱۳۹۷/۳/۶
نام نویسنده مسئول: مرتضی زاهدی
نشانی نویسنده مسئول: ایران – شهرستان شاهرود – میدان هفت تیر - دانشگاه صنعتی شاهرود - دانشکده کامپیوتر و فناوری اطلاعات- آزمایشگاه وب کاوی و شناسایی الگو.

۱ -مقدمه

سیستم پرسش و پاسخ (QAS)[۱] به‌عنوان سیستمی با پتانسیل بالا شناخته می‌شود که کاربران را قادر می‌سازد تا به منابع علمی با استفاده از زبان طبیعی (از طریق پرسش) دسترسی داشته باشند و یک پاسخ مرتبط، مناسب و مختصر را دریافت کنند. بااین‌حال، همچنان مشکلات چالش‌برانگیز فراوانی جهت مرتفع نمودن در این سیستم‌ها موجود است. سیستم‌های QA شکل پیچیده‌تر سیستم‌های بازیابی اطلاعات هستند که در این سیستم‌ها به‌جای ارائه‌ی کل سند، تنها بخش‌های خاصی از اطلاعات سند به‌عنوان پاسخ بازگردانده می‌شود. بنابراین پاسخ ارائه‌شده ممکن است یک کلمه، یک جمله یا یک پاراگراف باشد. یک سیستم QA از سه بخش پردازش پرسش، بازیابی اطلاعات و پردازش پاسخ تشکیل می‌شود.

از سال ۱۹۹۹ هرساله تحقیق پیرامون سیستم‌های پاسخ دامنه باز[۲] که از منابع اطلاعات غیر ساختاری بهره می‌برند، توسط کمپین ارزیابی TREC[۳] مرتباً در حال انجام است[۱-۳]. در کمپین‌های بعدی TREC با توجه به افزایش تعداد و پیچیدگی درخواست‌ها، اسناد مورداستفاده و پیچیدگی سؤالات روش‌های ارزیابی پاسخ نیز پیشرفته‌تر شدند. با گسترش صفحات وب، استفاده از این مجموعه اطلاعات برای سیستم‌های QA موردتوجه قرار گرفت و چندین سیستم QA بر مبنای وب توسعه یافتند[۴-۵]. سیستم‌های QA مبتنی بر وب را می‌توان به QAS دامنه باز و QAS دامنه بسته طبقه‌بندی کرد[۶].

فقدان تعامل دوطرفه بین سیستم و کاربر یکی از مهم‌ترین مشکلات سیستم‌های QA محسوب می‌شود[۱]. این معضل ازآنجا نشئت می‌گیرد که سیستم‌های QA راهکاری برای رفع ابهام درزمانی که پرسش کاربر دارای ابهام بوده، یا اینکه پاسخ سیستم مطلوب کاربر نبوده یا کاربر نیازمند دریافت اطلاعات بیشتری باشد، ارائه ننموده‌اند. بنابراین با اضافه شدن سطح تعامل در سیستم‌های پرسش و پاسخ تعاملی (IQA)[۴] این مشکل رفع شده است. سیستم‌های موجود در زمینه IQA می‌توانند با توجه به شرایط و کاربردهایشان در سه گروه مختلف شامل مدیریت محدودیت، QA ارتقاءیافته و سؤالات متوالی قرار گیرند[۱]. بدیهی است که وجود یک سیستم ارزیابی استاندارد نقش بسیار مهمی در ارتقای این سیستم‌ها ایفا می‌نماید. بااین‌وجود تقریباً هیچ روش استانداری در زمینه ارزیابی سیستم‌های IQA طراحی نشده است و روش‌های ارزیابی فعلی بر مبنای روش‌های مورداستفاده در QA و سیستم‌های دیالوگ بنانهاده شده‌اند. یک سیستم IQA از دو موجودیت سیستم و کاربر تشکیل‌شده است و ممکن است کاربر تحت تأثیر عوامل بسیاری با سیستم کار نماید لذا کار ارزیابی بسیار سخت و پیچیده است. اگرچه 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 است. بنابراین اکثر سیستم‌های ارزیابی موجود، از ارزیابی انسانی بهره می‌گیرند که در نتیجه، عملکرد یک سیستم از کاربری به کاربر دیگر متفاوت خواهد بود[۲]. بر اساس مطالعات صورت گرفته در زمینه ارزیابی سیستم‌های IQA توسط انسان، پارامترهای مختلفی جهت ارزیابی موردتوجه قرار می‌گیرد. بنابراین با توجه به گسترش روز افزون سیستم‌های IQA و کم کردن خطای ارزیابی انسانی، نیاز به‌جایگزینی یک مدل اتوماتیک به‌جای ارزیاب انسانی است. لذا ابتدا باید پارامترهای مؤثر در ارزیابی مشخص و روشی برای اندازه‌گیری اتوماتیک آن‌ها ارائه گردد که این مسئله خود یکی از چالش‌های روبروی محققین این حوزه است. بنابراین در این مقاله مجموعه‌ای از ویژگی‌های آماری جدید، جهت استفاده برای ارزیابی سیستم‌ها، بر اساس خروجی تولیدشده از سؤال و پرسش‌های ردوبدل شده بین کاربر و سیستم (که در این مقاله مکالمه نامیده می‌شود) ارائه گردیده است تا بتوانیم گام مهمی در این راستا برداریم.

ساختار مقاله بدین‌صورت است که در بخش اول مروری بر کارهای انجام‌شده در زمینه ارزیابی سیستم‌های QA، IQA صورت پذیرفته است. در بخش دوم مجموعه ویژگی‌های پیشنهادشده، سیستم IQA پایه تولیدشده و مدل دسته‌بندی تشریح شده است. بخش سوم نتایج به‌دست‌آمده را نشان می‌دهد و در بخش آخر به نتیجه‌گیری و پیشنهادها پرداخته‌شده است.

۲ -کارهای مرتبط

ارزیابی سیستم‌های QA بسته به ارزیابی سؤالات پیچیده یا ساده (مثل تعریف، روابط و سناریوهای مربوط به سؤالات) متفاوت است. یکی از روش‌های ارزیابی مورداستفاده در سیستم‌های QA استفاده از مجموعه‌ای از سؤالات و پاسخ‌ها به نام «مجموعه استاندارد طلایی» است[۳]. در این روش توانایی یک سیستم بر اساس میزان منطبق بودن سیستم با این مجموعه استاندارد طلایی موردسنجش قرار می‌گیرد. البته این روش برای سؤالات پیچیده و مبهم هنوز تقویت نشده است. در ارزیابی سیستم‌های QA با استفاده از کاربران واقعی تحقیقات قابل‌توجهی وجود دارد. بیشتر ارزیابی‌های صورت پذیرفته در این حوزه توسط TREC صورت پذیرفته و کارهای انجام‌شده در این حوزه، بیشتر در زمینه ارزیابی استخراج پاسخ، نحوه تعامل و استفاده از آن انجام‌شده است. آقای رودریگو و پناس[۷] یک مرور کلی بر روی سیستم‌های QA موجود انجام دادند. سپس به بررسی دلایل عدم پاسخگویی مناسب سیستم‌های QA در مقابل متن‌های دامنه باز پرداختند. برای بررسی این مشکلات، آن‌ها از سیستم‌های درک مطلب متون بهره گرفتند. این سیستم‌ها معمولاً با چالش‌هایی مانند تغییرات مختلف در سؤالات، اسناد منبع و یا اطلاعاتی که به‌طور صریح در اسناد و مدارک ذکر نشده‌اند، مواجه بودند. به همین جهت آن‌ها به دنبال ارائه روشی مناسب، برای حل این معضل در سیستم‌های QA بودند. آن‌ها در روش پیشنهادی خود به اهمیت در نظر گرفتن دانش نهفته در متون اسناد سیستم‌های QA پرداختند و استفاده از استنتاج متنی و پایگاه‌های دانش را به‌عنوان راهی برای بهبود نتایج معرفی نمودند.

بیشتر روش‌های پیاده‌سازی شده در زمینه ارزیابی سیستم‌های QA از معیارهایی همانند K1، MRR [۵]، CWS [۶] و C@1 استفاده نمودند که هرکدام ازاین‌روش‌ها خود دارای نقاط ضعف بوده و قابلیت تعمیم به همه سیستم‌های مختلف QA را نداشتند[۷]. به‌عنوان‌مثال معیار MRR زمانی بکار گرفته می‌شد که سیستم برای پاسخ به سؤال مطرح‌شده، چندین جواب را ارائه می‌نمود. اما در سیستم‌هایی که درمجموعه داده‌های خود تنها یک پاسخ برای هر سؤال ارائه می‌نمودند، از روش ارزیابی C@1 استفاده می‌شد. بنابراین این یکی از معضلات استفاده ازاین‌روش‌ها در سیستم‌های IQA بود و از طرفی این معیارها بیشتر در جهت انتخاب پاسخ بکار گرفته می‌شدند و توانایی سیستم را در این راستا مورد ارزیابی قرار می‌دادند. سان [۸]، روشی برای ارزیابی دقیق یک سیستم IQA با استفاده از روش ارزیابی تقاطعی[۷] معرفی نمود در روش پیشنهادی، میزان تأثیر متن و وظیفه کاربران بر عملکرد سیستم بررسی‌شده است که برای حذف این اثرات از روش‌های آماری بهره گرفتند. آن‌ها نشان دادند که روش ارائه‌شده برای مقایسه سیستم‌های QA و IQA بسیار مؤثر و دارای بهره‌وری بالایی است. کوارترونی و ماناندهار[۹] روشی که شامل یک ارزیابی کیفی از سیستم‌های IQA بود، ارائه نمودند. 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. سؤالات تهیه‌شده در پرسش‌نامه برای ارزیابی، شامل بررسی عملکرد سیستم، مشکلات تعامل، سرعت پاسخگویی و رضایت کلی کاربر از سیستم بود.

واکلدر و همکارانش [۱۰] در مقاله خود به توسعه ویژگی‌های موجود در روش ارزیابی، برای سیستم طراحی‌شده خود پرداختند. در این گزارش دو هدف اساسی پیگیری شد. نخست یک ارزیابی واقع‌بینانه از سودمندی و قابلیت استفاده از سیستم طراحی‌شده به‌عنوان یک سیستم تعاملی ارائه گردید و سپس به توسعه معیارهای مقایسه پاسخ به‌دست‌آمده، توسط تحلیلگران مختلف و ارزیابی کیفیت پشتیبانی این سیستم صورت پذیرفت. آن‌ها برای به دست آوردن اطلاعات در مورد راحتی تحلیلگر با سیستم طراحی‌شده از ابزار کمی و کیفی استفاده کردند و ویژگی‌های جدیدی در سنجش توانایی یافتن پاسخ به سؤالات پیچیده و گفت‌وگو تعاملی معرفی نمودند. منصوری و حسن‌پور [۱۱] برای یک سیستم QA از دانش موجود در سؤالاتی که قبلاً در این سیستم بین کاربران و سیستم ردوبدل شده بود برای پاسخ‌دهی به سؤالات استفاده نمودند. آن‌ها با ارائه یک الگوریتم، مجموعه‌ای از قطعات کارآمد را ایجاد کردند تا با استفاده مجدد از این قطعات برای پاسخ به سؤالات بعدی بتوانند امکان بهبود پاسخ و بازدهی بهتر را برای سیستم فراهم نمایند. کلی [۱۲] به ارزیابی عملکرد چهار سیستم IQA با کاربر واقعی در مقاله خود پرداخته است. آن‌ها به دنبال شناسایی معیارهای ارزیابی برای سیستم‌های IQA، با استفاده از تجزیه‌وتحلیل نظرات ارزیابی ساخته‌شده توسط کاربران، برای چنین سیستم‌هایی بودند. آن‌ها در کار خود از داده‌هایی کیفی جمع‌آوری‌شده از تحلیلگران اطلاعاتی در طول مصاحبه‌ها (از یک کارگاه سه‌روزه ارزیابی) بهره جستند.

۳ -روش پیشنهادی

نظر به اینکه در فرآیند ارزیابی سیستم‌های IQA افراد متخصص و خبره نقش دارند، حدس اینکه این افراد در پس‌زمینه ذهن خود از چه تابع ارزیابی برای نمره دهی به یک سیستم استفاده می‌نمایند یکی از چالش‌های موجود در زمینه ارزیابی سیستم‌های IQA است. بنابراین تعریف ویژگی‌هایی که بتواند در مدل‌سازی بکار گرفته شوند، تا مدل به‌دست‌آمده کمترین خطا را نسبت به نظرات موجود داشته باشد امری ضروری است. ازآنجایی‌که ویژگی‌های متعددی در ارزیابی یک سیستم IQA دخالت دارند و اندازه‌گیری اتوماتیک آن‌ها برای ایجاد یک مدل دارای اهمیت است در این مقاله مجموعه‌ای از ویژگی‌های جدید برای ارزیابی اتوماتیک این سیستم‌ها پیشنهادشده است. اساس روش پیشنهادی بر روی معرفی مجموعه‌ای از این ویژگی‌های جدید و تأثیر هر ویژگی در مدل‌سازی است. هدف از این کار یافتن ویژگی‌هایی بود که خروجی مدل بتواند با کمترین خطا، خروجی نزدیک به نظر داده‌شده، توسط ارزیاب تولید نماید.

۳ ـ۱ -سیستم تعاملی پایه

جهت ارزیابی سیستم‌های IQA نیاز به دسترسی به این سیستم‌ها است. بر این اساس علاوه بر سیستم‌های موجود، برای راحتی و دسترسی آسان‌تر به این سیستم‌ها، از سیستم تعاملی پایه طراحی‌شده در آزمایشگاه تحقیقاتی وب‌کاوی و شناسایی الگو[۸] دانشگاه صنعتی شاهرود استفاده گردید. در این سیستم از تکنیک‌های آماری جهت پاسخ به سؤالات کاربران بهره گرفته‌شده و مستقل از زبان عمل می‌نمود[۱۳]. جهت آموزش سیستم طراحی‌شده، از سه پایگاه دادگان فارسی با نام‌های WMPR-QA1-2015, WMPR-QA2-2015 و WMPR-QA3-2015 استفاده‌شده است. پایگاه داده اول با نام WMPR-QA1-2015 دارای چهار فایل متنی با محتوای آئین‌نامه آموزشی دانشگاه صنعتی شاهرود است که در قالب ۲۹۲ جمله و با فرمت UTF-8 گردآوری‌شده و به‌عنوان داده آموزشی شناخته می‌شود. ۸۱ پرسش و پاسخ مطرح‌شده از این آئین‌نامه نیز به‌عنوان مجموعه تست پایگاه دادگان فوق در نظر گرفته‌شده است. پایگاه دادگان دوم با نام WMPR-QA2-2015 دارای یک فایل متنی با محتوای آئین‌نامه مالی شهرداری‌ها است که در قالب ۷۵ جمله و با فرمت UTF-8 گردآوری‌شده و از آن به‌عنوان مجموعه آموزش استفاده‌شده است. ۳۳ پرسش و پاسخ مطرح‌شده از این آئین‌نامه نیز به‌عنوان مجموعه تست پایگاه دادگان WMPR-QA2-2015در نظر گرفته‌شده است. پایگاه دادگان سوم با نام WMPR-QA3-2015شامل دو مجموعه آموزش و تست است. مجموعه آموزش آن دارای یک فایل متنی با محتوای آئین‌نامه استخدام هیئت‌علمی دانشگاه‌ها است که در قالب ۲۵۶ جمله و با فرمت UTF-8 گردآوری‌شده است و مجموعه تست آن دربردارنده ۳۱ پرسش و پاسخ مطرح‌شده از این آئین‌نامه است. سه پایگاه دادگان فوق از وب‌سایت

آزمایشگاه وب کاوی و شناسایی الگو[۹] قابل دریافت می‌باشند. بررسی نظرات ارائه‌شده توسط کاربران نشان‌دهنده رضایت آن‌ها از کیفیت تعامل برقرارشده با سیستم بود[۱۳]. 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 [۱۴] ارائه شد.

۲ ۳ -پیش پردازش

با توجه به اهمیت پیش‌پردازش اطلاعات، یکی از مراحلی که بر روی متون استخراج‌شده از سیستم‌های IQA صورت پذیرفت، نرمال‌سازی این اطلاعات بود. در این مرحله متن ورودی به ساختاری قابل‌پردازش برای مراحل بعد تبدیل می‌گردد. نرمال‌سازی اطلاعات شامل ۵ گام است:

۱- مشخص کردن مرز جمله‌ها: در بیشتر مواقع، تعیین مرز جمله‌ها از طریق بررسی علائم جداکننده از قبیل فضای خالی، "."، "!"، "؟"، ":" و غیره انجام می‌شود که یافتن این علائم به‌تنهایی کافی نیست لذا برای متون انگلیسی علاوه بر این علائم از تجزیه‌کننده استنفورد[۱۰] استفاده کردیم.

۲- ریشه‌یابی: در این حالت یک کلمه به شکل عمومی خود کاهش می‌یابد که این شکل عمومی باید برای همه کلمات هم‌ریشه یکسان باشد. برای دادگان انگلیسی از ریشه‌یاب استنفورد و مجموعه دادگان فارسی از [۱۵] استفاده گردید.

۳- حذف کلمات و واژه‌های غیر مهم: در این مرحله لیستی مشتمل بر ۲۰۰ کلمه پرتکرار[۱۱] آماده گردید (کلماتی که در محتوای اصلی متن تأثیری ندارند) و از مکالمه‌ها حذف گردید.

۴- شناسایی مقادیر عددی: بعد از شناسایی اعدادی که به‌صورت حروف در مکالمه‌ها ذکرشده بودند، این کلمات برچسب مقدار عددی دریافت می‌کرد.

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

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

۳ ۳ -استخراج ویژگی

استخراج ویژگی یکی از مهم‌ترین قسمت‌های مربوط به هر سیستم تشخیص یا مدل‌سازی محسوب می‌شود. در این مرحله، تعدادی ویژگی آماری بر اساس n-گرم‌ها و بزرگ‌ترین رشته مشترک (LCS)[۱۲] تعریف گردید که در ادامه به معرفی هرکدام از این ویژگی‌ها خواهیم پرداخت.

برخی از فرمول‌های تعریف‌شده برای این ویژگی‌ها، از روابط تعریف‌شده در زمینه ارزیابی اتوماتیک متن‌های خلاصه‌سازی شده اقتباس‌شده‌اند[۱۶] اما متناسب با کار زمینه ارزیابی سیستم‌های پرسش و پاسخ تعاملی این روابط بازنویسی و به‌روزرسانی شده‌اند (جدول ۱). خروجی هر مکالمه بین کاربر و سیستم به‌صورت

$$\text{Count\_Weightmax\_N} = \frac{1}{M} \times \sum_{i=1}^{M} \operatorname{argmax}_{ngram \in S_i} \left(\frac{W_k \times Count_{match}(gram_n)}{\sum_{ngram \in S_i} Count(gram_n)} \right) \quad (۳)$$

مجموعه‌ای از سؤال‌ها و پاسخ‌ها است، بعضی از ویژگی‌های تعریف‌شده علاوه بر اینکه برای مجموعه سؤال- جواب مورداستفاده قرار گرفت، به‌صورت جداگانه برای مجموعه سؤال‌ها و مجموعه جواب‌ها نیز بکار گرفته شد.

- ویژگی اول:

N-گرم‌ها یکی از مشهورترین مدل‌های آماری زبان هستند. در این مدل‌ها ارتباطات زنجیره کلمات در نظر گرفته می‌شود. به‌عبارت‌دیگر، مدل‌های n-گرم بر اساس هم پیوندی و کنار هم قرار گرفتن کاراکترهای لغات در پردازش متن عمل می‌نمایند. ابتدا n-گرم‌های مشترک را شمرده با یکدیگر جمع و بر تعداد کل n-گرم‌ها تقسیم می‌نماییم (رابطه ۱).

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

که در این رابطه S_i، i-امین جمله از هر مجموعه مکالمه (conv) و n طول هر n-گرم است. فرض نمایید یک مکالمه شامل N سؤال و پاسخ باشد. n-گرم‌های مشترک بین هر سوال-جواب را شمرده با یکدیگر جمع و بر مجموع تعداد n-گرم‌ها تقسیم می‌نماییم. این کار به ازای n=۱,۲,۳ برای مجموعه‌های پرسش و پاسخ (Q-A)، مجموعه سؤال‌ها (Q-Q) و مجموعه جواب‌ها (A-A) از هر مکالمه صورت پذیرفت.

- ویژگی دوم:

در یک مکالمه برای nهای بزرگ‌تر، هر چه تعداد n-گرم‌های مشترک بیشتر باشد امتیاز آن مکالمه بیشتر خواهد بود و احتمال پیوستگی متن مکالمه بیشتر خواهد شد [۱۷]. بر این اساس در این ویژگی پیشنهادی، هرکدام از n-گرم‌ها، بر اساس ارزش یک ضریب وزنی برای هر n-گرم به ارزش W_k با یکدیگر جمع می‌شوند تا مقدار این ویژگی به دست آید (رابطه ۲).

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

که در آن M تعداد عضوهای مجموعه برای میانگین گرفتن و W_k ضریب تأثیر هر n-گرم و مقدار آن متناسب با عدد n است که k از ۱ تا

حداکثر مقدار n است. این ویژگی برای n=۱،۲،۳ محاسبه گردید. نحوه محاسبه این رابطه مانند ویژگی اول است با این تفاوت که به ازای n-گرم‌های مشترک ضریبی برابر با n به هر n-گرم مشترک نسبت داده می‌شود.

- ویژگی سوم:

متناسب با رابطه ۲، معادله ۳ پیشنهاد گردید، در این ویژگی، ابتدا به ازای هر جفت سؤال- جواب، ابتدا n-گرم‌های مشترک را به دست آورده و در ارزش هر n-گرم ضرب نموده سپس بر مجموع تعداد n-گرم‌ها تقسیم نموده، ماکزیمم بین آن‌ها را در نظر گرفته و درنهایت پاسخ به‌دست‌آمده از M تا جفت درون یک مکالمه آن را میانگین می‌کنیم.

- ویژگی چهارم:

انطباق پشت سر هم در سطح جمله معمولاً در n-گرم‌ها دیده می‌شود. بنابراین در n-گرم‌ها طول تعریف نمی‌شود. زیرا بزرگ‌ترین رشته مشترک در نظر گرفته می‌شود. درحالی‌که در بزرگ‌ترین زیررشته مشترک نیاز نیست انطباق پشت سر هم باشد. همچنین برای اینکه مسئله هم رخدادی در جملات در نظر گرفته شود، از معادله ۶ استفاده نمودیم. در رابطه تعریف‌شده، برای یک مکالمه ابتدا یک جفت سؤال- پاسخ را در نظر گرفته، سپس برای هر جفت بازیابی و دقت را محاسبه و برای تمامی جفت سؤال- پاسخ این کار را انجام می‌دهیم. درنهایت پاسخ به‌دست‌آمده را در رابطه ۶ قرار داده و امتیاز هر مکالمه را محاسبه می‌کنیم.

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

$$P_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{LCS(Q_i, A_i)}{L_{A_i}} \quad (۵)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۶)$$

که در آن $\beta = \frac{P_{LCS}}{R_{LCS}}$، M تعداد جفت سوال-پاسخ هر مکالمه، $LCS(Q_i, A_i)$ بزرگ‌ترین زیررشته مشترک بین سؤالات و پاسخ‌های یک مکالمه و L طول سؤال یا جواب است.

- ویژگی پنجم:

در این ویژگی اجتماع بزرگ‌ترین زیررشته مشترک بین Q_i و مجموعه جواب‌ها را محاسبه نمودیم که هر چه این عدد بزرگ‌تر باشد ارتباط بین جملات در مکالمه بیشتر است. به‌طور مثال فرض کنید جمله Q_1 شامل کلمات $w_1w_2w_3w_4w_5$ باشد و پاسخ A_1 شامل $w_1w_2w_6w_7w_8$ و پاسخ A_2 شامل کلمات $w_1w_3w_8w_9w_5$ باشد بنابراین LCS در رابطه بین Q_1 و A1 برابر w_1w_2 و در رابطه بین Q_1 و A_2 برابر $w_1w_3w_5$ است لذا اجتماع بین Q_1، A_1 و A_2 برابر با $w_1w_2w_3w_5$ است که $LCS_{\cup}(Q_i, A) = 4/5$ حاصل می‌گردد. برای کل مجموعه سؤالات یک مکالمه این کار را انجام دادیم. بنابراین روابط به‌صورت زیر پیشنهاد گردید.

$$R_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{\sum_{j=1}^{U} LCS_{\cup}(Q_i, A)}{P} \quad (۷)$$

$$P_{LCS} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{\sum_{j=1}^{V} LCS_{\cup}(Q_i, A)}{n} \quad (۸)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۹)$$

که در آن M تعداد سؤالات در یک مکالمه و $\beta = \frac{P_{LCS}}{R_{LCS}}$ خواهد بود. در این ویژگی Q_i شامل U جمله با P کلمه و مجموعه جواب‌ها شامل V جمله با n کلمه است.

- ویژگی ششم:

در این ویژگی، بزرگ‌ترین زیررشته مشترک بین هر سؤال و مجموعه جواب‌ها را یافته و درون یک مجموعه قرار داده، سپس در بین همه اعضای این مجموعه بزرگ‌ترین زیررشته حاصل‌شده را انتخاب می‌نماییم. روابط به‌صورت زیر تعریف گردید:

$$R_{LCS} = \frac{1}{N} \times \sum_{i=1}^{N} \max\left(\sum_{j=1}^{P} \frac{LCS(Q_i, A_j)}{L_{Q_i}}\right) \quad (۱۰)$$

$$P_{LCS} = \frac{1}{N} \times \sum_{i=1}^{N} \max\left(\sum_{j=1}^{P} \frac{LCS(Q_i, A_j)}{L_{A_j}}\right) \quad (۱۱)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۱۲)$$

که در این رابطه β =1، P تعداد جواب‌ها و N تعداد سؤالات است.

- ویژگی هفتم:

برای محاسبه امتیاز هر مکالمه معادله ۱۵ پیشنهاد گردید. در این 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. بنابراین با به‌روزرسانی روابط قبلی، معادلات جدید به‌صورت زیر معرفی شدند.

$$R_{LCS} = \frac{1}{N} \times \sum_{S_i \in S_1} \max_{S_j \in S_2} (LSC(S_i, S_j)) \quad (۱۳)$$

$$P_{LCS} = \frac{1}{P} \times \sum_{S_i \in S_1} \max_{S_j \in S_2} (LSC(S_i, S_j)) \quad (۱۴)$$

$$F_{LCS} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (۱۵)$$

که در رابطه ۱۵ مقدار β =1 در نظر گرفته شد. همچنین مجموعه‌های S_1 و S_2 را به‌صورت (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. بنابراین به ازای هر مجموعه‌مقدار این ویژگی نیز محاسبه گردید.

- ویژگی هشتم:

در این ویژگی پیشنهادشده، n-گرم‌های مشترک بین مجموعه سؤالات و پاسخ‌ها را به دست آورده و بعد از نرمال‌سازی به‌عنوان ارزش یک مکالمه گزارش می‌کنیم. در این ویژگی، با فرض اینکه ما دو مجموعه از سؤالات و جواب‌ها داریم بر اساس روابط زیر میزان امتیاز هر Q_i با مجموعه جواب‌ها محاسبه و درنهایت توسط رابطه ۱۸ امتیاز مکالمه را محاسبه می‌نماییم.

$$R_{skip_n} = \frac{1}{t} \times \frac{1}{k} \times \sum_{i=1}^{t} \sum_{j=1}^{k} \frac{skip_n(Q_i, A_j)}{C(m,n)} \quad (۱۶)$$

$$P_{skip_n} = \frac{1}{t} \times \frac{1}{k} \times \sum_{i=1}^{t} \sum_{j=1}^{k} \frac{skip_n(Q_i, A_j)}{C(L,n)} \quad (۱۷)$$

$$F_{skip_n} = \frac{1+\beta^2 \times R_{skip_n} \times P_{skip_n}}{R_{skip_n} + \beta^2 \times P_{skip_n}} \quad (۱۸)$$

که در آن t تعداد سؤالات، k تعداد پاسخ‌ها، n اندازه n- گرم (در اینجا مقدار ۲ و ۳ در نظر گرفته شد و skip همان n-گرم در رابطه است)، m طول سؤال Q_i، L طول پاسخ A_j، C ترکیب و B=1 در نظر گرفته شد.

- ویژگی نهم (امتیازدهی به جملات):

این ویژگی برای امتیازدهی به جملات استفاده گردید. یک جفت سؤال و جواب به‌صورت یک جمله در نظر گرفته شد. سپس برای هر یک از جملات، امتیاز محاسبه گردید. درنهایت با توجه به امتیاز هر جمله، امتیاز نهایی برای هر مکالمه محاسبه گردید. نحوه امتیازدهی به کلمات و جملات بدین‌صورت است که ابتدا امتیاز مربوط به کلمات را محاسبه و سپس بر اساس امتیاز به‌دست‌آمده برای کلمات با استفاده از رابطه ۲۰، امتیاز هر جمله محاسبه می‌شود.

$$Word\_score = K \times f_{word} \quad (۱۹)$$

$$Sentence\_score = \sum Word\_score \quad (۲۰)$$

که در آن K یک عدد ثابت و f_{word} تعداد تکرار کلمه در متن است. از طرفی با توجه به موقعیت مکانی هر جمله، امتیاز متفاوتی به آن تخصیص داده خواهد شد. نحوه امتیازدهی بدین‌صورت است که معمولاً جملات میانی دارای ارزش اطلاعاتی بالاتری نسبت به جملات ابتدایی و پایانی هر مکالمه هستند (بر اساس مجموعه داده تهیه‌شده این فرض صورت پذیرفت) بنابراین بر اساس موقعیت قرارگیری جملات ارزش‌گذاری برای هر جمله به‌صورت زیر پیشنهاد گردید.

$$P_{score_i} = \begin{cases} 1 - \dfrac{n-i+1}{n} & i <= \dfrac{n}{2} \\ 1 - \dfrac{i-3}{n} & \dfrac{n}{2} < i <= n \end{cases} \quad (۲۱)$$

که در آن i موقعیت هر جمله و n تعداد جملات هر مکالمه است. با توجه به ارزش هر جمله و امتیاز آن امتیاز نهایی یک مکالمه محاسبه می‌شود (رابطه ۲۲).

$$S_{conversation} = \frac{1}{N} \times \sum_{j=1}^{N} (Sentence\_score_j + P_{score_i}) \quad (۲۲)$$

- ویژگی دهم (فاصله همینگ):

استفاده از فاصله همینگ برای اندازه‌گیری تشابه بین جملات یک مکالمه یکی دیگر از ویژگی‌هایی است که مورداستفاده قرار گرفت. هر جمله در مکالمه، از تعدادی کلمه تشکیل‌شده است با به دست آوردن میزان شباهت بین دو کلمه، میزان شباهت بین جملات را محاسبه می‌کنیم. فاصله همینگ دو کلمه تعداد حروف متناظر نامشابه است. بنابراین این معیار میزان تفاوت را نشان می‌دهد. برای محاسبه شباهت عدد حاصل را بر طول کلمات تقسیم و از عدد یک کسر می‌نماییم. این کار را برای کل کلمات جمله انجام داده و درنهایت بر اساس اعداد به‌دست‌آمده میزان شباهت بین دو جمله محاسبه می‌شود. به‌طور مثال فرض کنید جمله Q_i شامل کلمات $w_1w_2w_3w_4w_5$ با طول m و پاسخ Ai شامل $w_1w_2w_6w_7w_8$ با طول n باشد تعداد حالت‌های ممکن ترکیب m و n است معیار شباهت بین دو کلمه و یک جمله به‌صورت زیر محاسبه می‌شود.

$$Similarity_{words} = 1 - \frac{\text{Hamming\_Distance}(A,B)}{Max(|A|,|B|)} \quad (۲۳)$$

$$Similarity\_sen = \frac{1}{C(m,n)} \times \sum_{j=1}^{C(m,n)} Similarity\_words_j \quad (۲۴)$$

- ویژگی یازدهم (محاسبه امتیاز کلمات):

به دلیل اینکه معمولاً، یک مکالمه درباره موضوع خاصی بین کاربر و سیستم صورت می‌پذیرد، این باور وجود دارد که طرفین مکالمه از واژگان معینی برای ادامه بحث یا تشریح دقیق جنبه‌های مختلف موضوع استفاده و یا از تکرار آن‌ها استفاده می‌کنند. بنابراین در این ویژگی برای هر مکالمه گراف هم‌رخداد[۱۳] کلمات ترسیم می‌شود. ویژگی این گراف در این است که در این حالت کلمات هم‌رخداد در متن بدون استفاده از اندازه پنجره مشخص تعیین می‌شوند. با توجه به اینکه تعداد رخداد هر واژه می‌تواند به‌عنوان عامل تعیین درجه اهمیت واژگان مورداستفاده قرار گیرد. بنابراین در این گراف، تعداد تکرار هر کلمه در مکالمه و اینکه هر کلمه با چه کلمه دیگری آمده است نمایش داده می‌شود. با توجه به گراف حاصل، از روی آن، فرکانس کلمه، درجه کلمه در گراف و نسبت درجه به فرکانس کلمه محاسبه می‌گردد. نسبت درجه به فرکانس را به‌عنوان امتیاز نهایی هر کلمه مشخص کرده مجموع امتیازات هر کلمه به‌عنوان امتیاز هر مکالمه در نظر گرفته می‌شود.

$$Score_{w_i} = \frac{\deg(w)}{freq(w)} \quad (۲۵)$$

$$Score_{conv} = \frac{1}{N} \times \sum_{j=1}^{N} Score_{w_i} \quad (۲۶)$$

- ویژگی دوازدهم (Tf-idf):

پارامتر TF-idf یکی از ویژگی‌های آماری است که در سیستم‌های بازیابی اطلاعات، بر اساس آن می‌توان میزان شباهت میان کلمات منتخب با یک سند را محاسبه کرد. در این سیستم‌ها، از این ویژگی

برای محاسبه میزان شباهت هر سند با پرسش مطرح‌شده استفاده می‌شود. در روش پیشنهادی، برای این ویژگی میزان تکرار یک کلمه در مکالمه، در مقابل تعداد تکرار آن در کل مکالمات، محاسبه می‌شود. سپس مجموع این امتیازات برای کلمات یک مکالمه محاسبه، نرمال‌سازی شده و به‌عنوان امتیاز یک مکالمه گزارش می‌شود. رابطه زیر نحوه محاسبه وزن هر کلمه حاصل را نشان می‌دهد.

$$W_{word\,(i)} = \mathrm{F_W} \times \log \frac{\mathrm{N_{sentence}}}{\mathrm{NC_{sentence}}} \qquad (۲۷)$$

که در آن $\mathrm{F_W}$ تعداد تکرار هر کلمه در جملات یک مکالمه، $\mathrm{N_{sentence}}$ تعداد جملات یک مکالمه و $\mathrm{NC_{sentence}}$ تعداد جملاتی است که شامل کلمه iام است.

- ویژگی سیزدهم:

یکی دیگر از ویژگی‌های پیشنهادی استفاده از میزان شباهت بین سؤال و جواب هر مکالمه با استفاده از روابط ۲۸ تا ۳۱ است. در این ویژگی، دو مجموعه S_1 و S_2 تعریف شد. مجموعه S_1 مجموعه N-گرم‌های مربوط به سؤال و مجموعه S_2 مجموعه N-گرم‌های مربوط به جواب است. بنابراین مقدار شباهت بین این دو مجموعه توسط رابطه ۲۸ به‌صورت جداگانه برای هر جفت پرسش و پاسخ محاسبه گردید. سپس میانگین امتیاز به‌دست‌آمده برای هر مکالمه محاسبه و به‌عنوان یک ویژگی در ماتریس ویژگی‌ها ذخیره گردید.

$$Likeness(S_1,S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \qquad (۲۸)$$

نحوه محاسبه ویژگی ۱۴، ۱۵ و ۱۶ همانند ویژگی سیزدهم است. اما معیار در نظر گرفته‌شده برای هریک به ترتیب از روابط زیر استفاده گردید.

$$\mathrm{Cosine\_dis} = \frac{|S_1 \cap S_2|}{\sqrt{|S_1| * |S_2|}} \qquad (۲۹)$$

$$Containment(S_1,S_2) = \frac{|S_1 \cap S_2|}{|S_2|} \qquad (۳۰)$$

$$Overlap(S_1,S_2) = \frac{|S_2|}{|S_1|} \qquad (۳۱)$$

مقدار این ویژگی برای n-گرم، ۲ تا ۳ محاسبه شد. با استفاده از چهار رابطه ۲۸ تا ۳۱، هشت ویژگی استخراج گردید. یعنی یک‌بار به ازای n=۲ مقدار چهار رابطه محاسبه و بار دیگر به ازای n=۳ این کار انجام شد. بر اساس روابط تعریف‌شده در قسمت استخراج ویژگی، درمجموع ۳۱ ویژگی حاصل گردید. جدول ۱ شماره رابطه استفاده‌شده برای استخراج هر ویژگی را نمایش می‌دهد. درصورتی‌که در این مرحله ویژگی‌های معنایی قابل اضافه شدن باشد قطعاً کارایی ویژگی‌ها بالاتر خواهد رفت [۱۸] اما در روش پیشنهادی ما فرض بر این بود که تنها از ویژگی‌های آماری استفاده شود. برخی ویژگی‌های فازی نیز بعضاً کارایی مناسبی در کاربردهای پردازش زبان از خود نشان می‌دهند [۱۹] که در این کار نیازی برای استفاده از آن‌ها احساس نشد.

جدول ۱: لیست ویژگی‌های استخراج‌شده از روابط

شماره ویژگی	شماره رابطه	توضیحات	شماره ویژگی	شماره رابطه	توضیحات
۱	۱	بروز رسانی شده	۱۷	۱۵	پیشنهادی
۲	۱	بروز رسانی شده	۱۸	۱۸	پیشنهادی
۳	۱	بروز رسانی شده	۱۹	۱۸	پیشنهادی
۴	۲	پیشنهادی	۲۰	۲۲	بروز رسانی شده
۵	۲	پیشنهادی	۲۱	۲۴	بروز رسانی شده
۶	۲	پیشنهادی	۲۲	۲۶	بروز رسانی شده
۷	۳	پیشنهادی	۲۳	۲۷	بروز رسانی شده
۸	۳	پیشنهادی	۲۴	۲۸	پیشنهادی
۹	۳	پیشنهادی	۲۵	۲۸	پیشنهادی
۱۰	۶	بروز رسانی شده	۲۶	۲۹	پیشنهادی
۱۱	۹	پیشنهادی	۲۷	۲۹	پیشنهادی
۱۲	۱۲	پیشنهادی	۲۸	۳۰	پیشنهادی
۱۳	۱۵	پیشنهادی	۲۹	۳۰	پیشنهادی
۱۴	۱۵	پیشنهادی	۳۰	۳۱	پیشنهادی
۱۵	۱۵	پیشنهادی	۳۱	۳۱	پیشنهادی
۱۶	۱۵	پیشنهادی			

۳ ۴ -دسته‌بندی متن خروجی سیستم IQA با استفاده از ماشین بردار پشتیبان

دسته‌بند ماشین بردار پشتیبان (SVM)[۱۴] ازجمله روش‌هایی است که در سال‌های اخیر، به‌منظور دسته‌بندی در زمینه‌های مختلف موردتوجه قرارگرفته است و نتایج رضایت‌بخشی را ارائه داده است. ازاین‌رو، از این روش نظارتی به‌عنوان دسته‌بند بهره می‌بریم. با توجه بااینکه مکالمات صورت گرفته توسط کاربران با سیستم‌ها در دو گروه مکالمات با امتیاز خوب و بد در پایگاه داده ذخیره‌شده بودند. لذا مسئله ما یک دسته‌بندی دو کلاسه است. در این مرحله، ابتدا بردارهای ویژگی از مکالمات ذخیره‌شده در پایگاه داده‌ها استخراج و درون یک ماتریس نگهداری، سپس توسط SVM به دسته‌بندی آن‌ها می‌پردازیم. قابل‌ذکر است که مجموعه آموزش به میزان ۷۰٪ و مجموعه تست به میزان ۳۰٪ از مجموعه داده به‌صورت تصادفی انتخاب شدند. بعد از مرحله آموزش، مدل ساخته‌شده باید به‌منظور به دست آوردن برچسب کلاس نمونه‌های تست مورداستفاده قرار می‌گرفت. با توجه به اینکه خروجی الگوریتم SVM به هر بردار ویژگی مقداری در بازه [-1,1] تخصیص داد. حد آستانه‌ای جهت بهینه‌سازی دسته‌بندی برابر ۰/۷ در نظر گرفته شد. یعنی بردارهایی که با مقادیر کمتر از این حد آستانه بودند به‌عنوان مکالمات در گروه نظرات با امتیاز بد و مکالماتی با مقادیر بیشتر به‌عنوان کلاس نظرات با امتیاز خوب، دسته‌بندی شدند.

۳ ۴ ۱ -معیار ارزیابی

دانشی که در مرحله یادگیری مدل تولید می‌شود، می‌بایست در مرحله ارزیابی مورد تحلیل قرار گیرد تا بتوان ارزش آن را تعیین نمود و به دنبال آن کارایی الگوریتم یادگیرنده مدل را مشخص کرد. بنابراین

جدول ۲: نتایج حاصل از دسته‌بندی کننده SVM

معیار ارزیابی	F-measure	
	۰/۹۷	
معیار اندازه‌گیری خطا	MAPE	RMSE
	٪۹۵	۰/۰۳۵

جدول ۳: نمونه تولیدشده 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 یک‌ترم	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.
چندترم متوالی مشروط شدن دانشجو اخراج می‌شود؟	دانشجو کارشناسی یا ارشد؟
دانشجو کارشناسی	درصورتی‌که دانشجو کارشناسی سه ترم متوالی یا چهار ترم غیر متوالی مشروط شود در آستانه اخراج قرار می‌گیرد.

برای ارزیابی نتایج حاصل از دسته‌بندی به‌دست‌آمده با داده‌های واقعی معیار F و برای اندازه‌گیری نرخ خطا از سنجه‌های آماری مجذور میانگین مربعات خطا (RMSE)[15]و درصد میانگین مطلق خطا (MAPE)[16] مورداستفاده قرار گرفت. روابط مربوط به این معیارها در زیر نشان داده‌شده است.

$$\text{F - measure} = \frac{(\beta^2+1)\times P\times R}{R+\beta^2\times P} \quad (۳۲)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} e_i^{\,2}}{n}} \quad (۳۳)$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{Y_i}\right|\times 100 \quad (۳۴)$$

که در رابطه ۳۲، $P=\frac{TP}{TP+FP}$، $R=\frac{TP}{TP+FN}$ و $\beta=\frac{1-\alpha}{\alpha}$ است. n تعداد پیش‌بینی‌ها و e_i خطای پیش‌بینی در دو رابطه دیگر است که از تفاوت مقادیر پیش‌بینی‌شده و مقادیر واقعی به دست می‌آید. Y_i مقادیر واقعی است.

شکل ۱: نمایی از یک صفحه سامانه تحت وب طراحی‌شده در راستای فرآیند ارزیابی

۴ -نتایج آزمایش‌ها

به جهت انتخاب بهترین دسته‌بندی کننده، چندین دسته‌بندی کننده مورداستفاده قرار گرفت که نتایج حاصل از SVM بسیار مناسب‌تر نسبت به دیگر دسته‌بندی کننده‌ها بود. نتایج حاصل از دسته‌بندی کننده SVM در جدول ۲ آورده شده است.

۴ ـ۱ مجموعه داده

به دلیل نبود مجموعه داده استاندارد در زمینه ارزیابی سیستم‌های IQA، نیاز به ایجاد یکسری متن خروجی از این سیستم‌ها و تشکیل یک پایگاه داده مناسب از سؤالات ردوبدل شده بین سیستم و کاربر با برچسب‌گذاری مناسب بود. بر این اساس، علاوه بر سیستم تعاملی پایه طراحی‌شده، 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 داده‌شده توسط کاربران به سیستم‌ها را به‌صورت اتوماتیک در پایگاه داده‌ای ذخیره می‌نمود. شکل ۱ نمایی از یکی از صفحات سامانه طراحی‌شده را نمایش می‌دهد. در این راستا تعداد ۱۲۰ کاربر برای ۵ موضوع مختلف با سیستم کار نمودند و با توجه به موضوع مکالمات هر یک به‌صورت جداگانه ذخیره گردید.

از این مجموعه ۶۰۰تایی، ۵۴۰ نمونه توسط فرد خبره به‌عنوان 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. همچنین جهت استفاده از سیستم‌ها و توضیح در پاسخ به سؤالات یک جلسه توجیهی برای کاربران برگزار شد. جدول ۳، یک نمونه از خروجی متن ذخیره‌شده (بدون فرمت) از تعامل با سیستم پایه را نشان می‌دهد. همچنین جهت ارزیابی

مناسب‌تر یک پایگاه داده از مکالمات انگلیسی روزمره با موضوعات مختلف شکل گرفت. تعداد مکالمات در نظر گرفته ۱۰۰۰ مکالمه با ۵ موضوع مختلف بود. همگی مکالمات دارای انسجام متنی بودند تا مکالمات واقعی به‌عنوان مکالمات مناسب در نظر گرفته شد. سپس به‌صورت تصادفی بعضی از خطوط مکالمات با یکدیگر جابجا گردید و به‌عنوان مکالمه نامناسب در پایگاه داده ذخیره گردید تا روابط حاصل نیز بر روی این پایگاه داده مورد آزمون قرار گیرد. جدول ۴ یک نمونه از مکالمه با امتیاز خوب و بد را نمایش می‌دهد.

۴ ۲ -ارزیابی ویژگی‌های پیشنهادی

برای اندازه‌گیری تأثیر ویژگی‌های پیشنهادی در ارزیابی سیستم‌های IQA، همبستگی بین نمرات داده‌شده توسط انسان به مکالمه‌ها و نتایج حاصل محاسبه گردید. همان‌طور که قبلاً ذکر شد، برای محاسبه ضریب همبستگی از دو پایگاه داده که یکی حاصل از تعامل با چهار سیستم IQA و دیگری حاصل از مکالمات روزمره بود استفاده نمودیم. نتایج حاصل از محاسبه ضریب همبستگی پیرسون برای اندازه‌گیری نظرات در مقابل نظرات انسانی هر دو پایگاه داده در جدول ۵ آورده شده است.

جدول ۴: یک نمونه از مکالمه مناسب و نامناسب

مکالمه با امتیاز بد	مکالمه با امتیاز خوب
<**conversation** id="140" subject="Asking about Public Transportation"> <**personA**>Can you give me a little more information about your apartment?</**personA**> <**personB**>Weren't you taught that yellow means slow down, not speed up?</**personB**> <**personA**>What kind of public transportation is near your apartment?</**personA**> <**personB**>You should take a break.</**personB**> <**personA**>Can I get you anything to drink?</**personA**> <**personB**>I'm sorry. If you check online, you can get that kind of information.</**personB**> <**personA**>Okay, I'll just go online.</**personA**> <**personB**>I don't know anything for certain. We always do a double check if there is a question.</**personB**> <Score>Bad</score>	<**conversation** id="2" subject="Viewing Houses with a Realtor"> <**personA**>I have a good feeling about this house.</**personA**> <**personB**>Yes, I liked it the first moment I pulled up to it.</**personB**> <**personA**>Then you're not going to go to sleep.</**personA**> <**personB**>If you like the outside, you are going to really love the inside.</**personB**> <**personA**>What a beautiful home!</**personA**> <**personB**>You'll notice that the window treatments, carpeting, and drapes are all new.</**personB**> <**personA**>I like the way the blinds give you privacy from the street.</**personA**> <**personB**> Follow me into the kitchen. You will love it.</**personB**> <**personA**>I love that they put a wine storage area in the kitchen.</**personA**> <**personB**>The best part is the bedroom and attached bathroom.</**personB**> <**personA**>I love the relaxing colors in the tile and floor covering!</**personA**> <**personB**>Let's take a few pictures so that we can remember what we like about this house.</**personB**> <**Score**>good</**Score**>

جدول ۵: ضرایب همبستگی پیرسون برای پایگاه داده ایجادشده

پایگاه داده مکالمات	پایگاه داده خروجی IQAها	ویژگی‌ها
۰/۹۱	۰/۹۲	ویژگی ۱
۰/۹۱	۰/۹۴	ویژگی ۲
۰/۸۴	۰/۸۵	ویژگی ۳
۰/۷۵	۰/۷۵	ویژگی ۴
۰/۷۳	۰/۷۸	ویژگی ۵
۰/۵۸	۰/۶۸	ویژگی ۶
۰/۹۰	۰/۹۱	ویژگی ۷
۰/۹۱	۰/۸۹	ویژگی ۸
۰/۷۸	۰/۷۳	ویژگی ۹
۰/۵۹	۰/۵۸	ویژگی ۱۰
۰/۷۲	۰/۶۳	ویژگی ۱۱
۰/۷۹	۰/۷۲	ویژگی ۱۲
۰/۵۴	۰/۵۲	ویژگی ۱۳
۰/۶۴	۰/۶۱	ویژگی ۱۴
۰/۸۴	۰/۸۴	ویژگی ۱۵
۰/۷۹	۰/۹۴	ویژگی ۱۶
۰/۸۶	۰/۹۰	ویژگی ۱۷
۰/۹۱	۰/۸۲	ویژگی ۱۸
۰/۷۸	۰/۶۹	ویژگی ۱۹
۰/۶۳	۰/۵۵	ویژگی ۲۰
۰/۸۱	۰/۷۸	ویژگی ۲۱
۰/۶۶	۰/۷۴	ویژگی ۲۲
۰/۷۱	۰/۶۹	ویژگی ۲۳
۰/۷۹	۰/۸۵	ویژگی ۲۴
۰/۸۱	۰/۷۰	ویژگی ۲۵
۰/۷۸	۰/۹۴	ویژگی ۲۶
۰/۶۸	۰/۹۳	ویژگی ۲۷
۰/۵۶	۰/۸۴	ویژگی ۲۸
۰/۸۸	۰/۹۲	ویژگی ۲۹
۰/۹۱	۰/۸۹	ویژگی ۳۰
۰/۷۶	۰/۸۱	ویژگی ۳۱

همان‌طور که در جدول ۵ نشان داده‌شده است، همه ویژگی‌ها همبستگی مثبت و قویی با نظرات انسانی موجود در پایگاه داده دارند. همچنین بهترین مقادیر حاصل در هر ستون با رنگ متفاوتی مشخص‌شده است. همان‌طور که از نتایج پیدا است. همچنین ویژگی‌های ۱، ۲، ۳، ۷، ۸، ۱۵، ۱۷، ۱۸، ۲۱، ۲۵، ۲۹ و ۳۱ دارای مقدار بیشتر از ۰/۸ در مجموعه پایگاه داده مکالمات می‌باشند که نشان از تأثیر بیشتر این ویژگی‌ها در مدل‌سازی برای نظرات انسانی است بنابراین با در نظر گرفتن تنها این ویژگی‌ها به دسته‌بندی مکالمات پرداختیم که با توجه به معیار F به مقدار ۰/۸۹ دست‌یافتیم.

همان‌طور که در قسمت پیشینه ذکر شد. بیشتر کارهای صورت گرفته در زمینه ارزیابی سیستم‌های IQA از روش مطرح‌شده در سیستم‌های QA و دیالوگ بهره گرفته است و یا اینکه در بیشتر موارد به دلیل پیچیدگی بالای این سیستم‌ها از ارزیاب انسانی استفاده می‌گردد. در مقاله [۲۰] به بررسی سؤالات تعاملی کنفرانس TREC و یک سیستم QA مربوط به پزشکی پرداخته و درنهایت بیشتر تمرکز خود را در معرفی عواملی که می‌توانند در ارزیابی یک سیستم پرسش و پاسخ پزشکی تأثیرگذار باشند پرداخته است. ارزیابی توسط دو گروه از دانشجویان صورت پذیرفته است. نتایج نشان دادد که عواملی مانند سن، جنسیت، تجربه در استفاده از کامپیوتر، نگرش نسبت به کامپیوتر و چندین ویژگی دیگر جزء عوامل تعیین‌کننده در موفقیت یک سیستم QA است. درصورتی‌که مقاله پیشنهادی به معرفی ویژگی‌های آماری که بتواند جایگزین ارزیابی کیفی سیستم‌های پرسش و پاسخ تعاملی شود، پرداخته است. در روش پیشنهادی سعی گردید ویژگی‌هایی پیشنهاد شود تا بتوان یک حدس آگاهانه از ارزیابی انسانی بر اساس متن خروجی تولیدشده از این سیستم‌ها داشت به‌طوری‌که خروجی

مدل پیشنهادی بر اساس ویژگی‌های تعریف‌شده بیشترین شباهت را به نظرات انسان داشته باشد. روش پیشنهادی در هیچ‌یک از کارهای قبلی مشاهده نگردید.

۵ -نتیجه‌گیری

در این مقاله مجموعه‌ای از ویژگی‌های آماری اتوماتیک برای مدل‌سازی نظرات ارزیابی انسانی بر اساس متن خروجی یک سیستم IQA ارائه شد که در هیچ‌یک از کارهای قبلی در حوزه ارزیابی سیستم‌های پرسش و پاسخ تعاملی مشاهده نشده بود. از روی متن‌های موجود در پایگاه داده، ویژگی‌ها استخراج و بر اساس آن ماتریس ویژگی تشکیل گردید. برای دسته‌بندی نظرات از دسته‌بندی کننده SVM بهره گرفته شد. نتایج ارائه‌شده برای ارزیابی دسته‌بندی کننده بر اساس معیارهای ارزیابی F و برای اندازه‌گیری خطا از RMSE و MAPE استفاده گردید که نتایج حاصل حاکی از موفقیت ویژگی‌های پیشنهادی است. بنابراین به محققین این حوزه پیشنهاد می‌گردد که با توجه به ضرایب ست‌آمده و مقدار آن‌ها می‌توان تأثیر هر یک از ویژگی‌ها را بر روی خروجی مشخص کرد و ازآن‌جهت ارائه یک مدل آماری جهت ارزیابی استفاده کرد. از طرفی با توجه به همبستگی بین ویژگی‌ها می‌توان به کاهش ویژگی‌ها پرداخت تا پیچیدگی معادلات به‌دست‌آمده برای مدل‌سازی به‌مراتب کمتر گردد.

مراجع

[1] 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.

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

[3] Hartawan, Andrei, and Derwin Suhartono, *Using Vector Space Model in Question Answering System*, Procedia Computer Science, pp. 305-311, 2015.

[4] Höffner, Konrad, S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A. Ngomo. *Survey on challenges of question answering in the semantic web*, Semantic Web 8, no. 6, pp.895-920, 2017.

[5] Bao, Junwei, Nan Duan, Ming Zhou, and Tiejun Zhao. "Knowledge-based question answering as machine translation." Cell 2, no. 6, 2014.

[6] L. Vanessa, V. Uren, M. Sabou and E. Motta. “Is question answering fit for the semantic web? A survey.” Semantic Web 2, no. 2: 125-155, 2011.

[7] Rodrigo, Alvaro, and Anselmo Penas. "A study about the future evaluation of Question-Answering systems." Knowledge-Based Systems 137, 83-93, 2017.

[8] S. Ying, P. B. Kantor and E. 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.

[9] Quarteroni, Silvia and S. Manandhar. “Designing an interactive open-domain question answering system.” Natural Language Engineering 15, no. 1: 73-95, 2009.

[10] N. Wacholder, S. G. Small, B. Bai, D. Kelly, R. trittman, S. Ryan, R. Salkin, “Designing a Realistic Evaluation of an End-to-end Interactive Question Answering System.” In LREC. 2004.

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

[12] Kelly, Diane, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun. “Questionnaires for eliciting evaluation data from users of interactive question answering systems.” Natural Language Engineering 15, no. 1: 119-141, 2009.

[۱۳] سلیمه شهر آیینی، مرتضی زاهدی،" *سیستم پاسخگوی تعاملی با استفاده از تکنیک‌های هوش مصنوعی*"، دانشگاه صنعتی شاهرود، دانشکده کامپیوتر و فناوری اطلاعات، پایان‌نامه ارشد، ۱۳۹۴.

[۱۴] محمدمهدی حسینی، مرتضی زاهدی، *"بهبود پاسخ ارائه‌شده در سیستم‌های پرسش و پاسخ تعاملی با استفاده از*به‌د *شبکه عصبی"*، هشتمین کنفرانس بین‌المللی فناوری اطلاعات و دانش، صفحات ۸۴-۹۱، ۱۳۹۵.

[15] L. C. Yew. “Rouge: A package for automatic evaluation of summaries.” In Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8. 2004.

[۱۶] ابزارهای پردازش متون زبان فارسی، آزمایشگاه فناوری وب دانشگاه فردوسی مشهد، ۱۳۹۱(wtlab.um.ac.ir) .

[17] C. Guinaudeau, M. Strube, “Graph-based Local Coherence Modeling”, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 93–103, 2013.

[۱۸] فرید قنبری، محسن رحمانی، "ارائه یک روش مبتنی بر گرایش معنایی برای طبقه‌بندی چند برچسبی محتوای فیلم‌ها به کمک متون زیرنویس آن‌ها"، مجله مهندسی برق دانشگاه تبریز، آذرماه ۹۶.

[۱۹] الناز زعفرانی، محمدرضا فیضی درخشانی و آزاده روحانی، " تشخیص هوشمند و خودکار غلط‌های تایپی در پایگاه داده‌های بزرگ بدون استفاده از لغت‌نامه"، مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۱، بهار ۹۶.

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

زیرنویس‌ها

[1] Question Answering System (QAS)
[2] Open Domain System
[3] Text Retrival Evaluation conference
[4] Interactive Question Answering system (IQA)
[5] Confidence Weighted Score
[6] Mean Reciprocal Rank
[7] Cross Evaluation
[8] www.wmpr.com
[9] http://wmpr.ir/fa/index/category/53
[10] https://nlp.stanford.edu/software/lex-parser.shtml
[11] Stop-words
[12] Longest Common Substring(LCS)
[13] Co-occurrence
[14] Support Vector Machine(SVM)
[15] Root Mean Square Error
[16] Mean Percentage Absolute Error