یک مدل آماری جهت ارزیابی سامانه‌های پرسش و پاسخ تعاملی با استفاده از رگرسیون

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

[1] دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد شاهرود، شاهرود، ایران

[2و3] دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات، دانشگاه صنعتی شاهرود، شاهرود، ایران

چکیده

همانند بسیاری از زمینه‌های دیگر زبان‌شناسی محاسباتی، ارزیابی نقش مهمی در سامانه‌های پرسش و پاسخ تعاملی ایفا می‌کند. با این وجود، در زمینه ارزیابی سامانه‌های پرسش و پاسخ تعاملی به‌طورتقریبی هیچ روش خاصی وجود ندارد که به ارزیابی کلی این سامانه‌ها پرداخته و همواره انسان باید در فرآیند ارزیابی مشارکت داشته باشد. ارائه مدلی که بتواند جایگزین انسان در فرآیند ارزیابی شود، یکی از موضوعات مورد توجه در این حوزه است. در این مقاله، یک مدل آماری مناسب برای ارزیابی سامانه‌های پرسش و پاسخ تعاملی جهت جایگزین‌کردن به جای انسان توسط مجموعه‌ای از ویژگی‌های جدید و رگرسیون ارائه شده است. با استفاده از چهار سامانهٔ تعاملی موجود پایگاه داده‌ای مناسب ایجاد شد. تعداد ۵۴۰ نمونه به‌عنوان داده مناسب در نظر گرفته شد تا مجموعه آزمون و آموزش بر اساس آن تشکیل شود. ابتدا پیش‌پردازش بر روی مکالمات صورت پذیرفت و بر اساس روابط تعریف‌شده، ویژگی‌های آماری از متن مکالمه‌ها استخراج و بر اساس آن ماتریس ویژگی تشکیل و سپس با استفاده از انواع رگرسیون سعی شد تا بهترین مدل استخراج شود که درنهایت رگرسیون غیرخطی سری توانی با RMSE به میزان ۰/۱۳ بهترین مدل را ارائه کرد.

واژگان کلیدی: ارزیابی، سامانهٔ پرسش و پاسخ تعاملی، رگرسیون غیرخطی، استخراج ویژگی

A New Statistical Model for Evaluation Interactive Question Answering Systems Using Regression

Mohammad Mehdi Hosseini*[1], Morteza Zahedi[2] & Hamid Hassanpour[3]

[1]Department of Computer Engineering, Islamic Azad University, Shahrood branch, Shahrood, Iran

[2,3]Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

Abstract

The development of computer systems and extensive use of information technology in the everyday life of people have just made it more and more important for them to make quick access to information that has received great importance. Increasing the volume of information makes it difficult to manage or control. Thus, some instruments need to be provided to use this information. The QA system is an automated system for obtaining the correct answers to questions posed by the human in the natural language. In these systems, if the response is found, and if it is not the user's expected response or if it needs more information, there is no possibility of exchanging information between the system and the user to ask more questions and get answers related to it. To solve this problem, interactive Question answering (IQA) systems were created. Interactive question answering (IQA) systems are associated with linguistic ambiguous structures, so these systems are more accurate than QA systems. Regarding the probability of ambiguity (ambiguity in the user question or ambiguity in the answer provided by the system), the repetition is possible in these systems to obtain the clarity. No standard methods have been developed on IQA systems evaluation, and the existing evaluation

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال مقاله: ۱۳۹۶/۹/۳ ● تاریخ پذیرش: ۱۳۹۷/۱۰/۱۹ ● تاریخ انتشار: ۱۳۹۸/۱۰/۰۷ ● نوع مطالعه: پژوهشی

methods have been developed based on the methods used in QA and dialogue systems. In evaluating IQA systems, in addition to quantitative evaluation, a qualitative evaluation is used. It requires users' participation in the evaluation process to determine the success level of interaction between the system and the user. Evaluation plays an important role in the IQA systems. In the context of evaluating IQA systems, there is partially no specific methodology for evaluating these systems in general. The main problem with designing an assessment method for IQA systems lies in the rare possibility to predict the 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 based on the conversation exchanged between users and systems. Moreover, 540 samples were considered as suitable data to create a test and training set. The statistical characteristics of each conversation were extracted after performing the preprocessing on them. Then a feature matrix was formed based on the obtained characteristics. Finally, using linear and nonlinear regression, human thinking was predicted. As a result, the nonlinear power regression with 0.13 Root Mean Square Error (RMSE) was the best model.

Keywords: Evaluation, Interactive Question Answering Systems, Nonlinear Regression, Feature Extraction.

۱- مقدمه

سامانهٔ پرسش و پاسخ (QAS[1]) به‌عنوان سامانه‌ای 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. با این حال، همچنان مشکلات چالش‌برانگیز فراوانی جهت مرتفع‌کردن در این سامانه‌ها موجود است. سامانه‌های QA شکل پیچیده‌تر سامانه‌های بازیابی اطلاعات هستند که در این سامانه‌ها به جای ارائه کل سند، تنها بخش‌های خاصی از اطلاعات سند به‌عنوان پاسخ بازگردانده می‌شود؛ بنابراین پاسخ ارائه‌شده ممکن است یک واژه، یک جمله یا یک پاراگراف باشد. یک سامانه QA از سه بخش پردازش پرسش، بازیابی اطلاعات و پردازش پاسخ تشکیل می‌شود. از سال ۱۹۹۹ هر ساله پژوهش دربارهٔ سامانه‌های پاسخ دامنه‌باز[2] که از منابع اطلاعات غیرساختاری بهره می‌برند، توسط کمپین ارزیابی TREC[3] به‌طورمرتب در حال انجام است [4, 7, 12]. در کمپین‌های بعدی TREC با توجه به افزایش تعداد و پیچیدگی درخواست‌ها، اسناد مورد استفاده و پیچیدگی سؤالات روش‌های ارزیابی پاسخ نیز پیشرفته‌تر شدند. با گسترش صفحات وب، استفاده از این مجموعه اطلاعات برای سامانه‌های QA مورد توجه قرار گرفت و چندین سامانه QA بر مبنای وب توسعه یافتند [5-8]. سامانه‌های QA مبتنی بر وب را می‌توان به QAS دامنه‌باز و QAS دامنه‌بسته طبقه‌بندی کرد [11].

سامانه‌های QA راهکاری برای رفع ابهام در زمانی‌که پرسش کاربر دارای ابهام بوده، یا اینکه پاسخ سامانه مطلوب کاربر نبوده یا کاربر نیازمند دریافت اطلاعات بیشتری باشد، ارائه نکرده‌اند؛ بنابراین نبود تعامل دوطرفه بین سامانه و کاربر یکی از مهم‌ترین مشکلات سامانه‌های QA محسوب می‌شود [7]. با اضافه‌شدن سطح تعامل در سامانه‌های پرسش و پاسخ تعاملی (IQA[4]) این مشکل رفع شده است. سامانه‌های موجود در زمینه IQA می‌توانند با توجه به شرایط و کاربردهایشان در سه گروه مختلف شامل مدیریت محدودیت، QA ارتقایافته و سؤالات متوالی قرارگیرند [5]. وجود یک سامانه ارزیابی استاندارد نقش بسیار مهمی در ارتقای سامانه‌های IQA ایفا می‌کند. با این وجود به‌طورتقریبی هیچ روش استانداری در زمینه ارزیابی این سامانه‌ها وجود ندارد و روش‌های ارزیابی فعلی بر مبنای روش‌های مورد استفاده در QA و سامانه‌های دیالوگ بنا شده‌اند. یک سامانه IQA از دو موجودیت سامانه و کاربر تشکیل شده و ممکن است، کاربر تحت تأثیر عوامل بسیاری با سامانه کار کند؛ لذا کار ارزیابی بسیار سخت و پیچیده است. اگرچه روش‌های استانداردی وجود دارند که می‌توانند اطلاعات مربوط به عملکرد سامانه از قبیل زمان، دقت و یا بازیابی را با استفاده از آنها به‌دست آورد، اما هنوز، نیاز به شناسایی سهم سامانه و کاربران در عملکرد مطلوب یک سامانه است؛ بنابراین بیش‌تر سامانه‌های ارزیابی موجود، از ارزیابی انسانی بهره می‌گیرند که درنتیجه، عملکرد یک سامانه از کاربری به کاربر دیگر متفاوت خواهد بود [4]. بر اساس مطالعات صورت‌گرفته در زمینهٔ ارزیابی سامانه‌های IQA توسط انسان، پارامترهای مختلفی جهت ارزیابی مورد توجه قرار می‌گیرد؛ بنابراین با توجه به گسترش روزافزون سامانه‌های IQA و کم‌کردن خطای ارزیابی انسانی، نیاز به جایگزینی یک مدل خودکار به جای ارزیاب انسانی است؛ لذا ابتدا باید پارامترهای مؤثر در ارزیابی مشخص و روشی برای اندازه‌گیری خودکار آنها ارائه شود که این مسأله خود یکی از

[1] Question Answering System
[2] Open Domain System
[3] Text Retrieval Evaluation conference
[4] Interactive Question Answering system (IQA)

چالش‌های این حوزه است. در این مقاله برای ارزیابی سامانه‌ها، بر اساس خروجی تولیدشده از سؤال و پرسش‌های ردوبدل‌شده بین کاربر و سامانه (که در این مقاله مکالمه نامیده می‌شود) با استفاده از مجموعه ویژگی‌های آماری جدید و رگرسیون، مدلی ارائه شده است تا بتوانیم گام مهمی در این راستا برداریم.

ساختار مقاله بدین صورت است که در بخش نخست مروری بر کارهای انجام‌شده در زمینه ارزیابی سامانه‌های QA، IQA صورت پذیرفته است. در بخش دوم مجموعه ویژگی‌های پیشنهادشده، سامانه IQA پایه تولیدشده و مدل دسته‌بندی تشریح شده است. بخش سوم نتایج به‌دست‌آمده را نشان می‌دهد و در بخش آخر به نتیجه‌گیری و پیشنهادها پرداخته شده است.

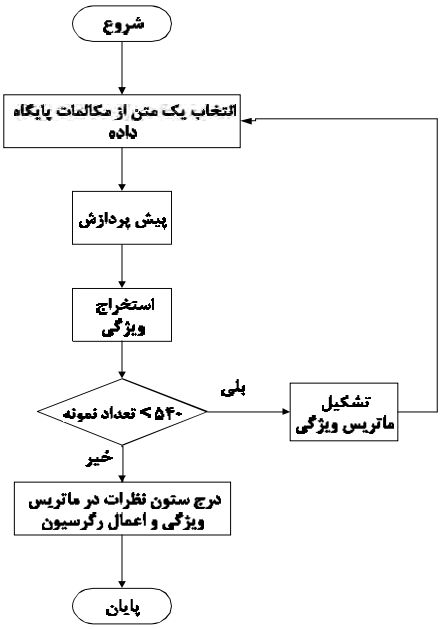
۲- کارهای مرتبط

ارزیابی سامانه‌های QA بسته به ارزیابی سؤالات پیچیده یا ساده متفاوت است. یکی از روش‌های ارزیابی مورد استفاده در سامانه‌های QA استفاده از مجموعه‌ای از سؤالات و پاسخ‌ها به نام «مجموعه استاندارد طلایی» است [7]. در این روش توانایی یک سامانه بر اساس میزان منطبق‌بودن سامانه با این مجموعه استاندارد طلایی، مورد سنجش قرار می‌گیرد. البته این روش برای سؤالات پیچیده و مبهم هنوز تقویت نشده است. در ارزیابی سامانه‌های QA با استفاده از کاربران واقعی پژوهش‌های قابل توجهی وجود دارد. بیشتر ارزیابی‌های صورت‌پذیرفته در این حوزه توسط TREC صورت گرفته و کارهای انجام‌شده در این حوزه، بیشتر در زمینه ارزیابی استخراج پاسخ، نحوه تعامل و استفاده از آن انجام شده است. بیشتر روش‌های پیاده‌سازی‌شده در زمینه ارزیابی سامانه‌های QA از معیارهایی همانند CWS[1], MRR[2], K1 C@1, استفاده کردند که هر کدام از این روش‌ها خود دارای نقاط ضعف بوده و قابلیت تعمیم به همه سامانه‌های مختلف QA را نداشتند [5]. به‌عنوان مثال معیار MRR زمانی به‌کار گرفته می‌شد که سامانه برای پاسخ به سؤال مطرح‌شده، چندین جواب را ارائه می‌کرد؛ اما در سامانه‌هایی که در مجموعه داده‌های خود تنها یک پاسخ برای هر سؤال ارائه می‌کردند، از روش ارزیابی C@1 استفاده می‌شد. بنابراین، این یکی از معضلات استفاده از این روش‌ها در سامانه‌های IQA بود و از طرفی این معیارها بیشتر در جهت انتخاب پاسخ به‌کار گرفته می‌شدند و توانایی سامانه را در این راستا مورد ارزیابی قرار می‌دادند. *سان* [16]، روشی برای ارزیابی دقیق یک سامانه IQA با استفاده از روش ارزیابی تقاطعی[3] معرفی کرد. در روش پیشنهادی، میزان تأثیر متن و وظیفه کاربران بر عملکرد سامانه بررسی شده است که برای حذف این اثرات از روش‌های آماری بهره گرفتند. آن‌ها نشان دادند که روش ارائه‌شده برای مقایسه سامانه‌های QA و IQA بسیارمؤثر و بهره‌وری بالایی دارد. *کلی* [9] به ارزیابی عملکرد چهار سامانه IQA با کاربر واقعی در مقاله خود پرداخته است. آن‌ها به‌دنبال شناسایی معیارهای ارزیابی برای سامانه‌های IQA، با استفاده از تجزیه و تحلیل نظرات ارزیابی ساخته‌شده توسط کاربران، برای چنین سامانه‌هایی بودند. آن‌ها در کار خود از داده‌هایی کیفی که از تحلیل‌گران اطلاعاتی در طول مصاحبه‌ها (که از یک کارگاه سه‌روزه ارزیابی فشرده از یک سامانه تعاملی گردآوری شده بود) جمع‌آوری کرده بودند، بهره گرفتند. *میرروشندل،برشبان و یوسفی‌نسب* [13] برای یک سامانه QA بر روی پیکره رسائل و مسائل معرفی کردند. در سامانه پیشنهادی برای یادگیری کلیه مؤلفه‌های سامانه‌های پرسش و پاسخ شامل دسته‌بندی سؤال، بازیابی اطلاعات و استخراج پاسخ مورداستفاده قرار دادند. نتایج نشان داد که سامانهٔ پیشنهادی توانسته بود به‌دقت ۲۹/۸۲ و میانگین معکوس رتبه ۷۳/۵۶ درصد دست یابد. واکلدر و همکارانش [14] در مقاله خود به توسعهٔ ویژگی‌های موجود در روش ارزیابی، برای سامانه طراحی‌شدهٔ خود پرداختند. در این گزارش دو هدف اساسی پیگیری شد. نخست یک ارزیابی واقع‌بینانه از سودمندی و قابلیت استفاده از سامانه طراحی‌شده به‌عنوان یک سامانه تعاملی ارائه شد و سپس به توسعه معیارهای مقایسه پاسخ به‌دست‌آمده، توسط تحلیل‌گران مختلف و ارزیابی کیفیت پشتیبانی این سامانه صورت پذیرفت. آن‌ها برای به‌دست‌آوردن اطلاعات در مورد راحتی تحلیل‌گر با سامانه طراحی‌شده از ابزار کمی و کیفی استفاده و ویژگی‌های جدیدی در سنجش توانایی یافتن پاسخ به سؤالات پیچیده و گفت‌وگوی تعاملی معرفی کردند. *کوارترونی و ماناندهار* [15] روشی که شامل یک ارزیابی کیفی از سامانه‌های IQA بود، ارائه کردند. آن‌ها در روش خود تعدادی پرسش مطرح کردند و از کاربران خواستند با دادن امتیازی بین یک (کمینه امتیاز) تا پنج (بیشینه امتیاز) کیفیت تعامل را اندازه‌گیری کنند. سؤالات تهیه‌شده در پرسش‌نامه برای ارزیابی، شامل بررسی عملکرد سامانه، مشکلات تعامل، سرعت پاسخ‌گویی و رضایت کلی کاربر از سامانه بود.

[1] Confidence Weighted Score
[2] Mean Reciprocal Rank
[3] Cross Evaluation

۳- روش پیشنهادی

در فرآیند ارزیابی سامانه‌های IQA افراد متخصص و خبره نقش دارند، حدس این‌که این افراد در پس‌زمینهٔ ذهن خود از چه تابع ارزیابی برای نمره‌دهی به یک سامانه استفاده می‌کنند، یکی از چالش‌های موجود در زمینه ارزیابی سامانه‌های IQA است؛ بنابراین تعریف ویژگی‌هایی که بتواند در مدل‌سازی به‌کار گرفته شوند، تا مدل به‌دست‌آمده کمترین خطا را نسبت به نظرات موجود داشته باشد، امری ضروری است. از آن‌جایی‌که ویژگی‌های متعددی در ارزیابی یک سامانه IQA دخالت دارند و اندازه‌گیری خودکار آنها برای ایجاد یک مدل دارای اهمیت است، در این مقاله مجموعه‌ای از ویژگی‌های جدید برای ارزیابی خودکار این سامانه‌ها پیشنهاد شده 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 کند. شکل (۱) روندنمای روش پیشنهادی را نمایش می‌دهد.



(شکل-۱): روند کلی روش پیشنهادی

(Figure-1): Flow chart of Proposed algorithm

۳-۱- سامانهٔ تعاملی پایه

از سامانه تعاملیِ پایهٔ طراحی‌شده در آزمایشگاه پژوهشی فناوری زبان‌های طبیعی[1] دانشگاه صنعتی شاهرود جهت انجام فرآیند مدل‌سازی ارزیابی استفاده شد. از فناوری‌های آماری جهت پاسخ به سؤالات کاربران در این سامانه بهره گرفته شده و مستقل از زبان عمل می‌کرد [2]. جهت آموزش سامانه طراحی‌شده، از سه پایگاه دادگان فارسی با نام‌های WMPR-QA1-2015, WMPR-QA2-2015 و WMPR-QA3-2015 استفاده شده است. پایگاه داده نخست با نام WMPR-QA1-2015 دارای چهار فایل متنی با محتوای آیین‌نامه آموزشی دانشگاه صنعتی شاهرود است که در قالب ۲۹۲ جمله و با فرمت UTF-8 گردآوری شده و به‌عنوان دادهٔ آموزشی شناخته می‌شود. ۸۱ پرسش و پاسخ مطرح‌شده از این آیین‌نامه نیز به‌عنوان مجموعه آزمون پایگاه دادگان بالا در نظر گرفته شده است. پایگاه دادگان دوم با نام WMPR-QA2-2015 دارای یک فایل متنی با محتوای آیین‌نامه مالی شهرداری‌ها است که در قالب ۷۵ جمله و با فرمت UTF-8 گردآوری شده و از آن به‌عنوان مجموعه آموزش استفاده شده است. ۳۳ پرسش و پاسخ مطرح‌شده از این آئین‌نامه نیز به‌عنوان مجموعه آزمون پایگاه دادگان WMPR-QA2-2015 در نظر گرفته شده است. پایگاه دادگان سوم با نام WMPR-QA3-2015 شامل دو مجموعه آموزش و آزمون است. مجموعه آموزش آن دارای یک فایل متنی با محتوای آیین‌نامه استخدام هیأت‌علمی دانشگاه‌ها است که در قالب ۲۵۶ جمله و با فرمت UTF-8 گردآوری شده است و مجموعه آزمون آن در بردارنده ۳۱ پرسش و پاسخ مطرح‌شده از این آیین‌نامه است. سه پایگاه دادگان بالا از تارنمای آزمایشگاه وب‌کاوی و شناسایی الگو[2] قابل دریافت هستند. در جدول (۱) یک نمونه از این پایگاه داده نمایش داده شده است. بررسی نظرات ارائه‌شده توسط کاربران نشان‌دهندهٔ رضایت آنها از کیفیت تعامل برقرارشده با سامانه بود [2]. جهت استفادهٔ بهینه از سامانه و افزایش عملکرد و کارایی، تغییراتی در آن اعمال که منجر به عملکرد بهتر سامانه شد.

(جدول-۱): یک نمونه از پایگاه داده جهت آموزش سامانه پایه

(Table-1): A sample of database to learn base IQA

<Query9>
<Question> غیبتی که دانشجوی کارشناسی ارشد مجازی در هر درس مجاز است چقدر است</Question>
<General-answer> تأیید میزان حضور دانشجو در فضای الکترونیکی بر اساس سامانه آموزشی موسسه و عدم مشاهده به منزله غیبت تلقی می‌گردد. میزان غیبت مجاز در مشاهده دروس و سایر فعالیت‌های یادگیری هر درس حداکثر معادل سه شانزدهم است. نحوه محاسبه حضور و غیاب و ضوابط مرتبط با آن از سوی شورای آموزشی دانشگاه تعیین می‌شود. </General-answer>
<acceptable-answer> میزان غیبت مجاز در مشاهده دروس و سایر فعالیت‌های یادگیری هر درس حداکثر معادل سه شانزدهم است.</acceptable-answer>

[1] www.IIltech.ir

[2] http://IILtech.ir/fa/index/category/53

```
<exact-answer> سه شانزدهم</exact-answer>
</Query9>
.
.
.
.
<Query81>
<Question>  مبلغ وام تحصیلی ممتاز چیست</Question>
<General-answer> وام موارد خاص بیماری‌های خاص و
پرهزینه مبلغ وام ۲۰۰۰۰۰۰۰ ریال. وام تحصیلی ممتاز و نمونه
حداکثر ۲ برابر وام تحصیلی. وام ضروری ممتاز و نمونه حداکثر ۲
برابر وام تحصیلی. </General-answer>
<acceptable-answer> وام تحصیلی ممتاز و نمونه حداکثر ۲
برابر وام تحصیلی.</acceptable-answer>
<exact-answer> حداکثر ۲ برابر وام تحصیلی.</exact-
answer>
</Query81>
```

نتایج حاصل از این بهینه‌سازی در [3] ارائه شد.

۲-۳- پیش‌پردازش

یکی از مراحلی که بر روی متون استخراج‌شده از سامانه‌های IQA صورت گرفت، نرمال‌سازی اطلاعات بود. در این مرحله متن ورودی به ساختاری قابل پردازش برای مراحل بعد تبدیل شد. مراحل نرمال‌سازی به‌ترتیب صورت‌گرفته شامل موارد زیر است:

۱- مشخص‌کردن 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 ما برای متون انگلیسی، علاوه‌بر این علائم از تجزیه‌کنندهٔ استنفورد[1] نیز استفاده کردیم.

۲- ریشه‌یابی: در این حالت یک واژه به شکل عمومی خود کاهش می‌یابد که این شکل عمومی باید برای همه واژگان هم‌ریشه یکسان باشد. برای دادگان انگلیسی از ریشه‌یاب استفورد و مجموعه دادگان فارسی از ابزار پردازش زبان دانشگاه فردوسی[2] استفاده شد.

۳- حذف واژگان و واژه‌های غیر مهم: فهرستی مشتمل‌بر دویست واژهٔ ایست‌واژه[3] آماده (واژگانی که در محتوای اصلی متن تأثیری ندارند) و از مکالمه‌ها حذف شد.

۴- شناسایی مقادیر عددی: بعد از شناسایی اعدادی که به‌صورت حروف در مکالمه‌ها یادشده بودند، این واژگان برچسب مقدار عددی دریافت کردند.

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

تمامی این کارها به‌صورت خودکار انجام و جواب نهایی توسط ناظر انسانی کنترل شد.

[1] https://nlp.stanford.edu/software/lex-parser.html
[2] wtlab.um.ac.ir
[3] Stop-words

۳-۳- استخراج ویژگی

یکی از مهم‌ترین مراحل مربوط به هر سامانه، تشخیص یا مدل‌سازی استخراج ویژگی است. در این مرحله، تعدادی ویژگی آماری بر اساس n-گرم‌ها و بزرگ‌ترین رشته مشترک (LCS[4]) تعریف شد. در ادامه هر کدام از این ویژگی‌ها تشریح شده است. ایده برخی از فرمول‌های تعریف‌شده برای این کار، از روابط تعریف‌شده در زمینه ارزیابی خودکار متن‌های خلاصه‌سازی‌شده اقتباس شده‌اند [10]؛ اما متناسب با کار زمینه ارزیابی سامانه‌های پرسش و پاسخ تعاملی این روابط بازنویسی و به‌روزرسانی شده‌اند. از آنجایی که خروجی هر مکالمه بین کاربر و سامانه به‌صورت مجموعه‌ای از سؤال‌ها و پاسخ‌ها است، بعضی از ویژگی‌های تعریف‌شده، علاوه‌بر این‌که برای مجموعه سؤال- جواب مورد استفاده قرار گرفت، به‌صورت جداگانه برای مجموعه سؤال‌ها و مجموعه جواب‌ها نیز به‌کار گرفته شد که در هر ویژگی توضیح داده شده است.

- ویژگی نخست:

N-گرم‌ها یکی از مشهورترین مدل‌های آماری زبان هستند. در این مدل‌ها ارتباطات زنجیرهٔ واژگان در نظر گرفته می‌شود. به عبارت دیگر، مدل‌های n-گرم براساس هم‌پیوندی و کنار هم قرارگرفتن نویسه‌های لغات در پردازش متن عمل می‌کنند. ابتدا n-گرم‌های مشترک را شمرده با یکدیگر جمع و بر تعداد کل n-گرم‌ها تقسیم می‌کنیم (رابطه ۱):

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

که در این رابطه S_i، i-امین جمله از هر مجموعه مکالمه (conv) و n طول هر n-گرم است. فرض کنید یک مکالمه شامل N سؤال و پاسخ باشد. n-گرم‌های مشترک بین هر سؤال-جواب را شمرده با یکدیگر جمع و بر مجموع تعداد n-گرم‌ها تقسیم می‌کنیم. این کار به‌ازای 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.

[4] Longest Common Substring(LCS)

- 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 [6]. 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 به‌دست آید (رابطه ۲):

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

که در آن M تعداد عضوهای مجموعه برای میانگین‌گرفتن و W_k ضریب تأثیر هر n-گرم و مقدار آن متناسب با عدد n است. این ویژگی نیز برای n=۱،۲،۳ محاسبه شد. نحوهٔ محاسبه این رابطه مانند ویژگی نخست است با این تفاوت که به‌ازای n-گرم‌های مشترک ضریبی برابر با n به هر n-گرم مشترک نسبت داده می‌شود.

- ویژگی سوم:

متناسب با رابطه (۲)۲، معادله (۳) پیشنهاد شد، در این ویژگی، ابتدا به‌ازای هر جفت سؤال- جواب، ابتدا n-گرم‌های مشترک را به‌دست آورده و در ارزش هر n-گرم ضرب، سپس بر مجموع تعداد n-گرم‌ها تقسیم کرده، بیشینهٔ بین آنها را در نظر گرفته و درنهایت پاسخ به‌دست‌آمده از M تا جفت درون یک مکالمه آن را با استفاده از میانگین‌گیری نرمالیزه می‌کنیم:

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

- ویژگی چهارم:

در n-گرم‌ها طول تعریف نمی‌شود؛ زیرا بزرگ‌ترین رشته مشترک در نظر گرفته می‌شود. بنابراین انطباق پشت سر هم در سطح جمله به‌طورمعمول در n-گرم‌ها دیده می‌شود. درحالی‌که در بزرگ‌ترین زیررشتهٔ مشترک، نیاز نیست انطباق پشت سر هم باشد؛ هم‌چنین برای اینکه مسألهٔ هم‌رخدادی در جملات در نظر گرفته شود، از معادلهٔ (۶) استفاده کردیم. در رابطهٔ 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:

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

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

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

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

- ویژگی پنجم:

در این ویژگی اجتماع بزرگترین زیررشتهٔ مشترک بین Qi و مجموعه جواب‌ها را محاسبه کردیم که هر چه این عدد بزرگتر باشد، ارتباط بین جملات در مکالمه بیشتر است. به‌طور مثال فرض کنید جمله Q_1 شامل واژگان $w_1w_2w_3w_4w_5$ باشد و پاسخ A_1 شامل $w_1w_2w_6w_7w_8$ و پاسخ A_2 شامل واژگان $w_1w_3w_8w_9w_5$ باشد، بنابراین LCS در رابطه بین Q_1 و A_1 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 می‌شود. برای کل مجموعه سؤالات یک مکالمه این کار را انجام دادیم. بنابراین روابط به‌صورت زیر پیشنهاد شد:

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

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

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

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 واژه و مجموعه جواب‌ها شامل V جمله با n واژه است.

- ویژگی ششم:

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

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

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

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

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

- ویژگی هفتم:

برای محاسبه امتیاز هر مکالمه معادله (۱۵) پیشنهاد شد. در این ویژگی فرض شد که دو مجموعه S_i و S_j داریم که S_i از N جمله با K واژه و S_j با P جمله با T واژه هستند. بنابراین با به‌روزرسانی روابط قبلی، معادلات جدید به‌صورت زیر معرفی شدند:

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

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

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

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 به‌ازای هر مجموعه مقدار این ویژگی نیز محاسبه شد.

- ویژگی هشتم (محاسبه امتیاز واژگان):

به‌طورمعمول در یک مکالمه درباره موضوع خاصی بین کاربر و سامانه 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[1] واژگان ترسیم می‌شود. ویژگی این گراف در این است که در این حالت واژگان 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 مورد استفاده قرار گیرد. بنابراین در این گراف، تعداد تکرار هر واژه در مکالمه و این که هر واژه با چه واژهٔ دیگری آمده است نمایش داده می‌شود. با توجه به گراف حاصل، از روی آن، فرکانس کلمه، درجه واژه در گراف و نسبت درجه به فرکانس واژه محاسبه می‌شود. نسبت درجه به فرکانس را به‌عنوان امتیاز نهایی هر واژه مشخص کرده مجموع امتیازات هر واژه به‌عنوان امتیاز هر مکالمه در نظر گرفته می‌شود.

[1] Co-occurrence

$$Score_{w_i} = \frac{\deg(w)}{freq(w)} \quad (۱۶)$$

$$Score_{conv} = \frac{1}{N} \times \sum_{j=1}^{N} Score_{w_i} \quad (۱۷)$$

- ویژگی نهم (فاصله همینگ):

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

$$Similarity_{words} = 1 - \frac{\text{Hamming\_Distance}(A,B)}{Max(|A|,|B|)} \quad (۱۸)$$

$$Similarity\_sen = \frac{1}{C(m,n)} \times \sum_{j=1}^{C(m,n)} Similarity\_words_j \quad (۱۹)$$

که در آن A وB واژگان مربوط به جملات، m تعداد واژگان جمله نخست و n تعداد واژگان جمله دوم است.

بر اساس روابط تعریف شده برای هر ویژگی، ۲۲ ویژگی حاصل شد. جدول (۲) شماره رابطه استفاده شده برای استخراج هر ویژگی را نمایش می‌دهد.

(جدول-۲): فهرست ویژگی‌های استخراج‌شده از روابط
(Table-2): List of extracted properties from Equations

شماره ویژگی	شماره رابطه	توضیحات	شماره ویژگی	شماره رابطه	توضیحات
۱	۱	به‌روزرسانی شده	۱۲	۳	پیشنهادی
۲	۱	به‌روزرسانی شده	۱۳	۶	به‌روزرسانی شده
۳	۱	به‌روزرسانی شده	۱۴	۹	پیشنهادی
۴	۱	به‌روزرسانی شده	۱۵	۱۲	پیشنهادی
۵	۱	به‌روزرسانی شده	۱۶	۱۵	پیشنهادی
۶	۱	به‌روزرسانی شده	۱۷	۱۵	پیشنهادی
۷	۲	پیشنهادی	۱۸	۱۵	پیشنهادی
۸	۲	پیشنهادی	۱۹	۱۵	پیشنهادی
۹	۲	پیشنهادی	۲۰	۱۵	پیشنهادی
۱۰	۳	پیشنهادی	۲۱	۱۸	پیشنهادی
۱۱	۳	پیشنهادی	۲۲	۲۲	پیشنهادی

۴-۳- مدل‌سازی

تحلیل رگرسیون، روشی آماری، برای بررسی و مدل‌سازی ارتباط بین متغیر وابسته و متغیر مستقل بوده با هدف پیش‌بینی متغیر وابسته از روی متغیر و یا متغیرهای مستقل است. هدف روش پیشنهادی، ارائه یک مدل آماری جهت ارزیابی سامانه‌های IQA است. بنابراین، با استفاده از رگرسیون به‌دنبال پیش‌بینی امتیاز در نظر گرفته شده توسط ارزیاب‌ها هستیم. در روش پیشنهادی، ویژگی‌های استخراج‌شده به‌عنوان متغییرهای وابسته و نظرات انسانی به‌عنوان متغییر مستقل در نظر گرفته شد. مدل‌های مختلف از رگرسیون اعم از خطی، چندگانه، غیرخطی و غیره مورد بررسی و آزمایش قرار گرفت تا بهترین مدلی که بتواند داده‌ها را توصیف کند، استخراج و بر اساس معیار ارزیابی انتخاب شود.

۱-۴-۳- رگرسیون برای پیش‌بینی نظرات

رگرسیون ساده خطی، برای بررسی رابطه یک متغیر مستقل (پیش‌بین) و یک متغیر وابسته استفاده می‌شود؛ درحالی‌که اگر تعداد متغیرهای مستقل در این رابطه خطی بیش از یک عدد شود، مدل رگرسیون، خطی چندگانه نامیده می‌شود. معادله رگرسیون خطی ساده به شکل $Y=AX+B$ و رگرسیون خطی چندگانه به‌صورت $Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$ است. که در آن X ها، ویژگی‌های استخراج شده و b ضرایب تأثیر هر یک از ویژگی‌ها باشد. مقدار حاصله برای a بیان‌گر مقادیر پیش‌بینی‌شده Y با ثابت‌ماندن مقادیر X است. بنابراین این ضرایب باید طوری محاسبه شدند که معیار کمینه مربعات خطا را تأمین کند. از طرفی با مقایسه اندازه مقادیر ضرایب b با یکدیگر، میزان تأثیر و اولویت هر یک از عوامل مشخص می‌شود. علامت ضرایب هم بر تغییرات متغییر وابسته تأثیرگذارند. رگرسیون غیرخطی مدل‌های مختلفی دارد که از جمله آن می‌توان به مدل‌های درجه دو به بالا، چندجمله‌ای نمایی، توانی و غیره اشاره کرد که متناسب با مدل انتخاب‌شده معادلات آن‌ها متفاوت خواهد بود. تعدادی از معادلات رگرسیون غیرخطی در نظر گرفتیم و بر روی اطلاعات استخراج‌شده، مورد آزمایش قرار گرفته تا بهترین پاسخ انتخاب شود.

۲-۴-۳- معیار ارزیابی

برای ارزیابی نتایج حاصل از مدل به‌دست‌آمده با داده‌های واقعی سه سنجهٔ آماری ضریب تعیین R^2، مجذور میانگین مربعات خطا (RMSE[1]) و درصد میانگین مطلق خطا (MAPE[2]) مورد استفاده قرار گرفت. روابط مربوط به این معیارها در زیر نشان داده شده است.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \qquad (۲۰)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{Y_i}\right| \times 100 \qquad (۲۱)$$

که در این روابط n تعداد پیش‌بینی‌ها و e_i خطای پیش‌بینی است که از تفاوت مقادیر پیش‌بینی‌شده و مقادیر واقعی به‌دست می‌آید و Y_i مقادیر واقعی است.

۵-۳- نتایج آزمایشات

با توجه به وجود انواع مختلف رگرسیون، نیاز به استفاده از نرم‌افزارهای آماری برای استفاده از رگرسیون است؛ لذا ما برای شبیه‌سازی نتایج از نرم‌افزار مطلب استفاده کردیم. همچنین برای رگرسیون غیرخطی از جعبه‌ابزار apm[3] استفاده شد. نتایج حاصل از مدل‌سازی نظرات با استفاده رگرسیون در ادامه آورده شده است.

۱-۵-۳- پایگاه داده

به‌دلیل نبود پایگاه دادهٔ استاندارد در زمینه ارزیابی سامانه‌های IQA، نیاز به ایجاد یک پایگاه دادهٔ مناسب از سؤالات ردوبدل‌شده بین سامانه و کاربر با برچسب‌گذاری مناسب بود. با توجه به این موضوع، علاوه‌بر سامانه تعاملی پایهٔ طراحی‌شده، سه سامانه دیگر تعاملی موجود در نظر گرفته شد. در این راستا تعداد ۱۲۰ کاربر برای پنج موضوع مختلف با سامانه کار کردند و با توجه به موضوع مکالمات هر یک به‌صورت جداگانه ذخیره شد. از این مجموعه ششصدتایی، ۵۴۰ نمونه توسط فرد خبره به‌عنوان نمونه مناسب‌تر انتخاب شد. شرکت‌کنندگان در این دوره ارزیابی شامل دانشجویان دانشگاه و زبان‌آموزان یک مؤسسه بودند. با توجه به اینکه یک سامانه با زبان فارسی و سه سامانه دیگر با زبان انگلیسی عمل می‌کرد، برای یکپارچه‌سازی شرایط کار با این سامانه‌ها و راحتی کاربران، سامانه‌ای تحت وب طراحی شد.

این سامانه متن تبادل‌شده و امتیاز داده‌شده توسط کاربران به سامانه‌ها را به‌صورت خودکار در پایگاه داده‌ای ذخیره می‌کرد. شکل (۲) نمایی یکی از صفحات سامانهٔ طراحی‌شده و جدول (۳)، یک نمونه از خروجی متن ذخیره‌شده (بدون فرمت) را از تعامل با سامانه پایه نشان می‌دهد.

[1] Root Mean Square Error
[2] Mean Absolute Percent Error
[3] http://apmonitor.com/che263/index.php/Main/MatlabDataRegression

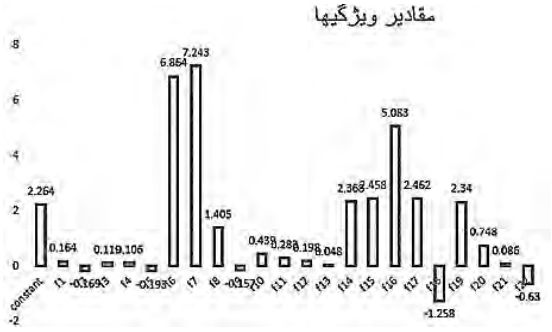
(شکل-۲): نمایی از یک صفحه سامانه تحت وب طراحی شده
(Figure-2): A screenshot of a web page design system

(جدول-۳): نمونهٔ تولیدشده از یک مکالمه در سامانه IQA
(Table-3): Sample of generated conversation in IQA system

کاربر	سامانه
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 کمترشدن معدل از ۱۴ در هر نیمسال دانشجو مشروط تلقی می‌شود.
دانشجوی کارشناسی چطور؟	در صورت کمترشدن معدل از ۱۲ در هر نیمسال دانشجو مشروط تلقی می‌گردد.
چندترم متوالی مشروط‌شدن 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.

۲-۵-۳- نتایج حاصل از پیاده‌سازی رگرسیون خطی

در روش رگرسیون خطی، ویژگی‌های استخراج‌شده به‌عنوان متغییر مستقل و نظرات انسانی به‌عنوان متغییر وابسته در نظر گرفته شدند.



(شکل-۳): ضرایب معادله رگرسیون خطی
(Figure-3): Linear regression equation coefficients

مقدار پارامتر sig برای تمامی ضرایب به‌دست‌آمده در رگرسیون خطی چندگانه کمتر از ۰/۰۵ بود. که این به معنای قابل قبول‌بودن ضرایب به‌دست‌آمده برای ویژگی‌ها است. ضرایب رگرسیونی خطی به‌دست‌آمده با توجه به معادله رگرسیون خطی چندگانه در شکل (۳) نمایش داده شده است، که ضریب تعیین R^2 با توجه به ضرایب به‌دست‌آمده برای معادله برابر ۰/۷۶، MSE برابر ۰/۲۴ و MAPE برابر ۷٪ حاصل شد. همان‌طورکه در شکل (۳) نشان داده شده است؛ مقادیر به‌دست‌آمده دارای علامت و مقدار متفاوتی برای هر متغییر هستند؛ بنابراین می‌توان در مورد نقش و تأثیر هر کدام از ویژگی‌ها در مدل‌سازی نظرات بحث کرد. ویژگی‌هایی که دارای مقدار کمی باشند، دارای تأثیر کمتر در نظرات انسانی و آن‌هایی که دارای مقادیر بزرگ‌تر هستند، حاکی از تأثیر بیشتر این ویژگی‌ها در خروجی هستند؛ همچنین با توجه به ضرایب به‌دست‌آمده، می‌توان تغییرات نظرات انسانی را نسبت به متغیرهای مستقل نیز به‌دست آورد.

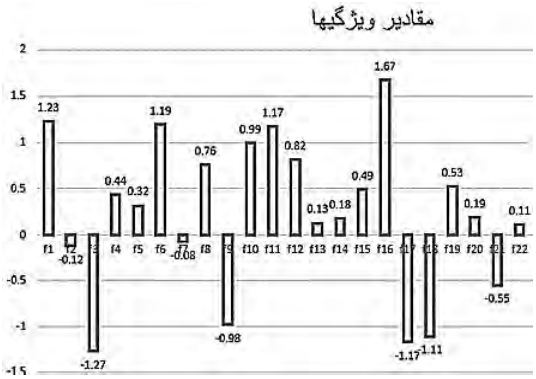
۳-۵-۳- نتایج حاصل از پیاده‌سازی رگرسیون غیرخطی

در رگرسیون غیر خطی از معادلات متعددی برای مدل‌سازی و ضرایبی با مقادیر اولیه متفاوتی آزمایش شد تا بهترین مدل انتخاب شود. با توجه به آزمایش‌های متعدد انجام‌شده، بهترین مدلی که برای داده‌ها استخراج شد، مدل سری توانی رگرسیون غیرخطی بود. معادله آن برای ارزیابی نظرات به‌صورت معادله (۲۲) حاصل شد.

$$Y = 1.6X_1^{1.80} + 0.9X_2^{-0.25} + 0.85X_3^{-1.32} + \ldots\ldots + 1.01X_{22}^{0.11} \quad (۲۲)$$

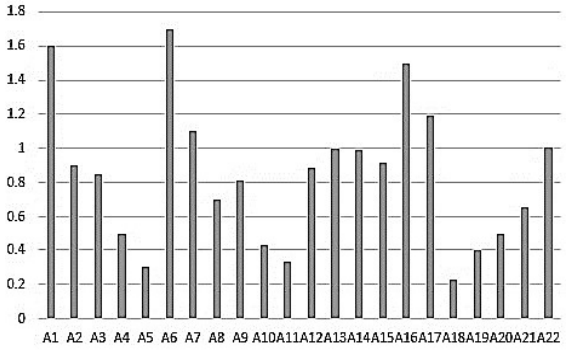
ضرایب مربوط به این معادله در شکل (۴) و (۵) آورده شده است. قابل توجه است که ارزیابی صورت‌گرفته بر اساس سنجه‌های آماری در نظر گرفته، مقدار R2 برابر ۰/۸۳، MSE برابر ۰/۱۳ و MAPE برابر ۶٪ حاصل شد. شکل (۶) مقایسه بین نتایج حاصل از ارزیابی‌ها را برای رگرسیون خطی و غیرخطی نمایش می‌دهد. نتایج حاکی از ارائه دقت بیشتر رگرسیون غیرخطی سری توانی نسبت به بقیه معادلات غیرخطی و برتری آن نسبت به رگرسیون خطی است. در روش پیشنهادی سعی شد ویژگی‌هایی پیشنهاد شود تا بتوانند توصیف بهتری از ارزیابی انسانی بر اساس متن خروجی تولیدشده از سامانه‌های IQA داشته باشد؛ به‌طوری‌که خروجی مدل پیشنهادی بر اساس ویژگی‌های تعریف‌شده بیشترین شباهت را به نظرات انسان داشته باشد. روش پیشنهادی در هیچ یک از کارهای قبلی مشاهده نشد با این حال می‌توان به

مقاله [16] که در زمینه ارزیابی سامانه‌های IQA ارائه شده است، اشاره کرد. در این مقاله به بررسی سؤالات تعاملی کنفرانس TREC و یک سامانه QA مربوط به پزشکی پرداخته و درنهایت بیشتر تمرکز خود را در معرفی عواملی که می‌توانند در ارزیابی یک سامانه پرسش و پاسخ پزشکی تأثیرگذار باشند، پرداخته است. ارزیابی توسط دو گروه از دانشجویان صورت پذیرفته است. نتایج نشان داد که عواملی مانند سن، جنسیت، تجربه در استفاده از رایانه، نگرش نسبت به رایانه و چندین ویژگی دیگر جزء عوامل تعیین‌کننده در موفقیت یک سامانه QA است. درصورتی‌که روش پیشنهادی به معرفی ویژگی‌های آماری که بتواند جایگزین ارزیابی کیفی سامانه‌های پرسش و پاسخ تعاملی شود، پرداخته‌ایم.



(شکل-۴): ضرایب توان معادله رگرسیون غیرخطی

(Figure-4): Non Linear regression equation coefficients

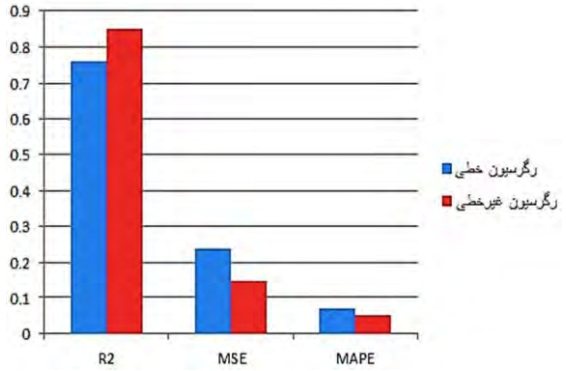


(شکل-۵): ضرایب تأثیر هر ویژگی مربوط به رگرسیون غیرخطی

(Figure-5): The effect coefficients of each characteristic on nonlinear regression

۴- نتیجه‌گیری

در این مقاله، روشی خودکار برای مدل‌سازی نظرات ارزیابی انسانی بر اساس متن خروجی یک سامانه IQA ارائه شد. که مشابه با روش ارائه شده، در هیچ یک از کارهای قبلی مشاهده نشد. در روش پیشنهادی ابتدا مجموعه‌ای از ویژگی‌های آماری جدید پیشنهاد شد که بتوانند به‌خوبی در مدل‌سازی نظرات کارایی داشته باشند. همان‌طورکه در قسمت استخراج ویژگی بیان شد، ویژگی‌های پیشنهادی، بر روی زوج‌های پرسش-پاسخ (Q_i-A_i)، مجموعه جواب‌ها (A_i-A_{i+1}) و مجموعه پرسش‌ها (Q_i-Q_{i+1}) اعمال شد تا تعاملی که بین جملات ردوبدل‌شده در یک مکالمه وجود داشته و نقش به‌سزایی در امتیازدهی توسط کاربر دارد، در نظر گرفته شود؛ سپس از روی متن‌های موجود در پایگاه داده، ویژگی‌ها استخراج و براساس آن ماتریس ویژگی تشکیل شد. در ادامه برای پیش‌بینی امتیاز داده‌شده توسط کاربر (برای مدل‌سازی نظرات) از رگرسیون خطی و غیرخطی بهره گرفته شد.



(شکل-۶): مقایسه نتایج حاصل از مدلسازی با رگرسیون خطی و غیر خطی

(Figure-6): Comparison of the results of modeling with linear and nonlinear regression

نتایج ارائه‌شده بر اساس معیارهای ارزیابی R^2، MSE و MAPE رگرسیون غیرخطی سری توانی دارای دقت بالاتری بود که نشان‌دهندهٔ پایداری مدل پیشنهادی است. برای کارهای آینده، پیشنهاد می‌شود با توجه به ضرایب به‌دست‌آمده و مقادیر آن‌ها، می‌توان تأثیر هر یک از ویژگی‌ها را بر روی خروجی مشخص کرد و با توجه به همبستگی بین ویژگی‌ها، به کاهش ویژگی‌ها پرداخت تا پیچیدگی معادلات به‌دست‌آمده به مراتب کمتر و یا برای رسیدن به معادله رگرسیونی از الگوریتم‌های هوشمند مانند برنامه‌نویسی ژن و یا ژنتیک استفاده شود.

۵- مراجع 5- References

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

[1] S. Shahriini, S. Zahedi, "Interactive Question answering System Using Artificial Intelligence Techniques", Senior Thesis, Shahrood University of Technology, Faculty of Computer and Information Technology, 2015.

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

[2] M.M. Hosseini, M. Zahedi, "Improvement of the response provided in interactive question answering systems using neural network", *Eighth International Conference on Information and Knowledge Technology*, pp. 84-91, 2016.

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

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

[5] 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.

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

[7] 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.

[8] 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, , pp. 119-141, 2009.

[9] 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.

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

[11] 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, pp. 345-361,2016.

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

[۱۳] برشبان، یاسمن، یوسفی نسب، حامد، میرروشندل، سید ابوالقاسم " ارایه یک پیکره پرسش و پاسخ مذهبی در زبان فارسی"، فصل‌نامه پردازش علائم و داده‌ها، دوره ۱۵، شماره ۱، صفحات ۸۷- ۱۰۲، ۱۳۹۷.

[13] Y. Boreshban, H. Yousefinasab, S. A. Mirroshandel, "Providing a Religious Corpus of Question Answering System in Persian", *Journal of Signal and Data Processing (JSDP)*, Vol 15, no 1, pp.87-102, 2018.

[14] 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.

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

[16] 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, pp. 1653-1665, 2011.

محمدمهدی حسینی در حال حاضر هیأت علمی دانشگاه آزاد اسلامی واحد شاهرود هستند. ایشان دوره دکترای خود را در دانشگاه صنعتی شاهرود گذرانده و بر روی سامانه‌های پرسش و پاسخ تعاملی پژوهش کرده است. زمینه‌ها و علایق پژوهشی ایشان پردازش متن، سامانه‌های پرسش و پاسخ تعاملی و پردازش تصویر است.
نشانی رایانامه ایشان عبارت است از:
Hosseini_mm@shahroodut.ac.ir

مرتضی زاهدی در حال حاضر عضو هیات علمی دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود است. وی پروژه‌هایی را در زمینه تعامل انسان و رایانه، شناسایی الگو، پردازش تصویر و ویدئو، و بینایی ماشین در دست اجرا دارد که در آن‌ها به‌طورمعمول از اطلاعات و دانش آماری استفاده می‌شود. وی دارای دکترای تخصصی رایانه از دانشگاه RWTH-Aachen آلمان است. تألیف کتب و مقالات علمی و همچنین سرپرستی پروژه‌های دانشگاهی و صنعتی در ایران و کشورهای اروپایی در کارنامک کاری

ایشان دیده می‌شود. زمینه‌ها و علایق پژوهشی ایشان تعامل انسان و رایانه، پردازش متن، شناسایی الگو، پردازش تصویر و ویدئو، بینایی ماشین است.
نشانی رایانامه ایشان عبارت است از:
Zahedi@ganjineh.co.ir

حمید حسن‌پور مدرک دکترای خود را از دانشگاه صنعتی کوئینزلند استرالیا در گرایش پردازش سیگنال در سال ۱۳۸۳ دریافت کرده‌اند. ایشان مدرک کارشناسی ارشد خود را در گرایش هوش ماشین در سال ۱۳۷۵ از دانشگاه صنعتی امیرکبیر، و مدرک کارشناسی خود را در سال ۱۳۷۲ در گرایش سخت‌افزار از دانشگاه علم و صنعت ایران اخذ کرده‌اند. دکتر حسن‌پور در طی سال‌های ۱۳۸۴ تا ۱۳۸۶ به‌عنوان عضو هیئت علمی در دانشکده مهندسی برق و رایانه دانشگاه صنعتی بابل فعالیت داشتند؛ سپس به دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود منتقل شدند. زمینه‌ها و علایق پژوهشی ایشان پردازش سیگنال، پردازش تصویر، داده‌کاوی، معماری کامپیوتر و پردازش متن است.
نشانی رایانامه ایشان عبارت است از:
h.hassanpour@shahroodut.ac.ir