

ارایهٔ یک پیکرهٔ پرسش و پاسخ مذهبی در زبان فارسی



یاسمن برشبان، حامد یوسفی نسب و سید ابوالقاسم میرروشندل*
گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه گیلان، رشت، ایران

چکیده

سامانه‌های پرسش و پاسخ، زیرشاخه‌ای از علوم پردازش زبان طبیعی و بازیابی اطلاعات محسوب می‌شوند که در چند دههٔ اخیر مورد علاقهٔ زیاد پژوهش‌گران قرار گرفته‌اند. با توجه به رشد فزایندهٔ علاقه‌مندی به این زمینهٔ پژوهشی، نیاز به در اختیار داشتن منابع داده‌ای مناسب برای آن، به خوبی احساس می‌شود. تاکنون اغلب پژوهش‌های صورت‌گرفته در رابطه با توسعهٔ پیکرهٔ پرسش و پاسخ در زبان انگلیسی بوده است؛ در صورتی که در زبان‌های دیگر مانند فارسی، نیاز شدیدی به وجود چنین پیکره‌هایی احساس می‌شود. در این مقاله، مراحل کامل توسعه یک پیکرهٔ متنی پرسش و پاسخ با نام رسائل و مسائل در زبان فارسی شرح داده خواهد شد. این پیکره شامل ۲,۱۱۸ سؤال غیرحقیقت و ۲,۰۵۱ سؤال حقیقت بوده که برای هر سؤال، متن سؤال، نوع سؤال، سختی سؤال از نظر پرسشگر و پاسخ‌دهنده، طبقه معنایی پاسخ در سطح درشت‌دانه و ریزدانه، پاسخ دقیق سؤال و شماره صفحه و پاراگراف پاسخ، نشانه‌گذاری شده است. پیکرهٔ پیشنهادی برای یادگیری کلیه مؤلفه‌های سامانه‌های پرسش و پاسخ شامل دسته‌بندی سؤال، بازیابی اطلاعات و استخراج پاسخ، مورد استفاده می‌تواند قرار گیرد و به صورت رایگان در دسترس پژوهش‌گران قرار دارد. در ادامه یک سامانه پرسش و پاسخ بر روی پیکره رسائل و مسائل معرفی می‌شود. نتایج نشان می‌دهد که سامانهٔ پیشنهادی توانسته است به دقت ۸۲/۲۹ و میانگین معکوس رتبه ۵۶/۷۳ درصد دست یابد. می‌توان اظهار کرد که پیکره و سامانهٔ پیشنهادی در نوع خود، نخستین پیکره و سامانهٔ مربوط به پرسش و پاسخ با چنین ویژگی‌هایی برای زبان فارسی است.

واژگان کلیدی: سامانه‌های پرسش و پاسخ، پردازش زبان طبیعی، بازیابی اطلاعات، پیکره رسائل و مسائل

Providing a Religious Corpus of Question Answering System in Persian

Yasaman Boreshban, Hamed Yousefi Nasab & Seyed Abolghasem Mirroshandel*
Computer Engineering Department, faculty of Engineering, University of Guilan, Rasht, Iran

Abstract

Question answering system is a field in natural language processing and information retrieval noticed by researchers in these decades. Due to a growing interest in this field of research, the need to have appropriate data sources is perceived. Most researches about developing question answering corpus area have been done in English so far, but in other languages as Persian, the lack of these corpora is perceived. In this article, the development of a Persian question answering corpus called Rasayel&massayel will be discussed. This corpus consists of 2,118 non-factoid and 2,051 factoid questions that for each question, question text, question type, question difficulty from questioner and responder's perspective, expected answer type in coarse-grained and fine-grained level, exact answer, and page and paragraph number of answer are annotated. The proposed corpus can be applied to learn components of question answering system, including question classification, information retrieval, and answer extraction. This corpus is freely available for the academic purpose as well.

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات

فصلنامه علمی پژوهشی



سال ۱۳۹۷ شماره ۱ پیاپی ۳۵

۸۷

www.SID.ir

In the following, a question answering system is presented on the Rasayel&massayel corpus. Our experimental result represents that the intended proposed system has achieved 82.29 % accuracy and 56.73 % mean reciprocal rank. It could be also claimed that this is the first ever question answering system and corpus with such features in Persian.

Keywords: Question answering system, Natural language processing, Information retrieval, Rasayel&massayel corpus

سؤال غیرحقیقت^۳ و ۲۰۵۱ سؤال حقیقت بوده که برای هر سؤال، متن سؤال، نوع سؤال، سختی سؤال از نظر پرسش‌گر و پاسخ‌دهنده، طبقه معنایی پاسخ در سطح درشت‌دانه و ریزدانه، پاسخ دقیق سؤال، شماره صفحه و پاراگرافی که پاسخ از آن استخراج شده و درجه ارتباط پاسخ استخراج شده با سؤال در دو سطح ارتباط کامل (با مقدار ۲) و ارتباط جزئی (با مقدار ۱) نشانه‌گذاری شده است. این پیکره برای یادگیری کلیه مؤلفه‌های سامانه‌های پرسش و پاسخ شامل دسته‌بندی سؤال، بازیابی اطلاعات و استخراج پاسخ، به‌صورت رایگان مورد استفاده عموم می‌تواند قرار گیرد. در ادامه یک سامانه پرسش و پاسخ بر روی پیکره رسائل و مسائل معرفی می‌شود. نتایج نشان می‌دهد که سامانه پیشنهادی توانسته است به دقت ۸۲/۲۹ و میانگین معکوس رتبه ۵۶/۷۳ درصد دست یابد.

ساختار مقاله پیش رو بدین شرح است: در بخش ۲، به تشریح کارهای مرتبط توسعه پیکره سامانه‌های پرسش و پاسخ می‌پردازیم. مفاهیم اولیه سامانه‌های پرسش و پاسخ در بخش ۳ بیان می‌شود. در بخش ۴، معماری سامانه‌های پرسش و پاسخ تشریح می‌شود. منبع داده‌ای مورد استفاده برای ساخت پیکره و روند توسعه آن در بخش ۵ معرفی می‌شود. همچنین یک طبقه‌بندی سلسله‌مراتبی نوع پاسخ مورد انتظار و اجزا و ساختار پیکره پیشنهادی مطرح می‌شود. ابزار نشانه‌گذاری پیکره و وضعیت در دسترس بودن آن، در بخش ۶ شرح داده می‌شود. در بخش ۷، یک سامانه پرسش و پاسخ بر روی پیکره رسائل و مسائل معرفی و در بخش ۸، جمع‌بندی و کارهای آینده بیان می‌شود.

۲- کارهای مرتبط

در چند سال اخیر در زمینه توسعه پیکره پرسش و پاسخ در زبان انگلیسی پژوهش‌های گسترده‌ای صورت گرفته است. از جمله پیکره‌های پرسش و پاسخ انگلیسی به پیکره TREC می‌توان اشاره کرد که توسط کنفرانس سالانه TREC ارائه می‌شود. فعالیت این کنفرانس در زمینه سامانه‌های پرسش و

۱- مقدمه

امروزه با افزایش اطلاعات در فضای اینترنت، کاربران باید زمان زیادی را صرف یافتن اطلاعات مدنظر خود کنند. برای تسهیل این امر، سامانه‌های بازیابی اطلاعات کلاسیک در قالب موتورهای جستجو ارائه شدند. در این سامانه‌ها، کاربران پرسش‌های خود را به‌صورت زبان طبیعی مطرح می‌کنند؛ سپس موتورهای جستجو، فهرستی را از صفحات مرتبط با این سؤال برای کاربران برمی‌گردانند.

در مقابل بازیابی اطلاعات کلاسیک، سامانه‌های پرسش و پاسخ معرفی شده‌اند [8]. در سامانه‌های پرسش و پاسخ که شکل پیچیده‌تری از سامانه‌های بازیابی اطلاعات هستند، به جای بازگرداندن کل سند، بخش خاصی از اطلاعات که مدنظر کاربر است، به‌عنوان پاسخ برگردانده می‌شود. کاربران سامانه‌های پرسش و پاسخ، علاقه‌مند به دریافت پاسخ مختصر، قابل فهم و صحیح هستند که این پاسخ ممکن است یک کلمه، جمله، پاراگراف، تصویر، قطعه صوتی و یا یک سند کامل باشد [7-8].

در چند دهه اخیر، سامانه‌های پرسش و پاسخ، مورد علاقه زیاد پژوهش‌گران قرار گرفته‌اند. با توجه به رشد فزاینده علاقه‌مندی به این زمینه پژوهشی، نیاز به در اختیار داشتن منابع داده‌ای مناسب برای آن به‌خوبی احساس می‌شود. تاکنون اغلب پژوهش‌های صورت گرفته در رابطه با توسعه پیکره پرسش و پاسخ در زبان انگلیسی بوده است، در صورتی که در زبان‌های دیگر مانند فارسی، نیاز شدیدی به وجود چنین پیکره‌هایی احساس می‌شود.

در این مقاله، مراحل توسعه یک پیکره پرسش و پاسخ در زبان فارسی به‌نام رسائل و مسائل به‌تفصیل شرح داده خواهد شد. می‌توان اظهار کرد که این پیکره در نوع خود، اولین پیکره مربوط به پرسش و پاسخ با چنین ویژگی‌هایی برای زبان فارسی است که حتی با پیکره‌های انگلیسی موجود نیز می‌تواند قابل مقایسه باشد. این پیکره شامل ۲۰۱۱۸

³ Non-factoid

⁴ Text Retrieval Conference

¹ Question Answering system

² Corpus

استخراج رویداد¹¹ [23] و تحلیل احساس¹² [1] که مورد استفاده پژوهش‌گران می‌تواند قرار گیرد.

۳- مفاهیم اولیه سامانه‌های پرسش و پاسخ

در فرآیند نشانه‌گذاری، مفاهیمی به‌کاربرده‌شده مورد استفاده قرار می‌گیرد. در این بخش، دسته‌بندی سامانه‌های پرسش و پاسخ از نظر دامنه و انواع سؤالات سامانه‌های پرسش و پاسخ شرح داده می‌شود.

۳-۱- دسته‌بندی سامانه‌های پرسش و پاسخ از نظر دامنه

سامانه‌های پرسش و پاسخ از نظر دامنه به دودسته دامنه باز¹³ و بسته¹⁴ تقسیم می‌شوند. سامانه‌هایی با دامنه باز یا نامحدود، باید انواع مختلف سؤالاتی را که توسط کاربران در زمینه‌های مختلف مطرح می‌شوند، پوشش دهند. برای مثال تمامی زمینه‌های ورزشی، مذهبی، سیاسی و هر زمینه دیگری که ممکن است کاربران در مورد آن سؤال بپرسند، باید توسط سامانه پوشش داده شود [5].

سامانه‌هایی با دامنه بسته، تنها جواب‌گوی سؤالات در یک زمینه خاص هستند. برای مثال فقط سؤالات در زمینه پزشکی یا سؤالات مذهبی را پاسخ‌گو خواهند بود و اغلب بر روی یک سایت خاص و یا یک کتاب خاص، کار می‌کنند. عملکرد این سامانه‌ها در مقایسه با سامانه‌های دامنه باز، ساده‌تر است؛ زیرا سامانه‌های پردازش زبان طبیعی اغلب می‌توانند اطلاعات آن دامنه خاص را استخراج کنند و از اطلاعات و ویژگی‌های خاص دامنه، در فرایند یافتن پاسخ مناسب بهره‌برند. در این سامانه‌ها، اغلب انواع محدودی از سؤالات که در آن زمینه مدنظر پرکاربرد هستند، پوشش داده خواهد شد [15].

۳-۲- انواع سؤالات سامانه‌های پرسش و پاسخ

در یک سامانه پرسش و پاسخ، سؤالات متنوعی توسط کاربران می‌تواند مطرح شود و برای پاسخ‌گویی به هر سؤال، لازم است از روش‌های مناسب آن سؤال استفاده شود. دسته‌بندی‌های مختلفی بر روی سؤالات صورت گرفته است، اما دسته‌بندی

پاسخ از سال ۱۹۹۹ آغاز شد [22] و این پیکره در اختیار پژوهش‌گرانی که در این کنفرانس شرکت می‌کنند، قرار داده می‌شود [21]. در سال ۲۰۰۰، انجمن ارزیابی بین‌زبانی CLEF¹ تأسیس شد که سامانه‌های پرسش و پاسخ بین‌زبانی را توسعه داد. سامانه‌هایی که در آن، زبان سؤال با زبان اسناد موجود در مخزن اطلاعات، متفاوت است [12].

پیکره پرکاربرد دیگر، پیکره سؤال و جواب مقالات ویکی‌پدیا است [18]. این پیکره، حاوی سؤالات حقیقت² استخراج‌شده از مقالات ویکی‌پدیا، پاسخ سؤال، درجه سختی سؤال از نظر پرسش‌گر و پاسخ‌دهنده است که برای استفاده در دسترس عموم قرار دارد. پیکره‌ای دیگر برای مؤلفه دسته‌بندی سؤال³ ارائه شده است که حاوی ۱۴،۵۰۰ سؤال به همراه برجسب است و برای هر سؤال، نوع پاسخ مورد انتظار در سطح درشت‌دانه⁴ و ریزدانه⁵ ذخیره شده است [10-11]. علاوه بر این پیکره‌ای حاوی هفتاد هزار نمونه سؤال و جواب موجود است که پاسخ‌ها را در سطوح مختلف جمله، پاراگراف و سند نگهداری می‌کند [20].

در زبان فارسی، پیکره‌ای حاوی پنج هزار سؤال به‌منظور دسته‌بندی سؤال، ارائه شده که شامل سؤالاتی از مجموعه کتاب‌های درسی و برای هر سؤال، نوع پاسخ مورد انتظار در سطح درشت‌دانه و ریزدانه نگهداری شده است [16]. همچنین پیکره‌ای به‌منظور مشخص کردن موضوع⁶ موجود است که این پیکره حاوی ۱۱۸ سؤال است [3]؛ اما این تعداد سؤال، برای ارزیابی در سامانه‌های پرسش و پاسخ مناسب نیست. در زبان فارسی پیکره‌ای که بتواند برای کلیه مؤلفه‌های سامانه‌های پرسش و پاسخ، مورد استفاده قرار گیرد، وجود ندارد. در همین اواخر پیکره و سامانه‌ای در حوزه قرآن در زبان فارسی ارائه شده است⁷ اما در مورد جزئیات پیاده‌سازی سامانه، منابع قابل استنادی در دسترس نیست.

البته در زبان فارسی تعدادی پیکره مناسب برای سایر زمینه‌های پژوهشی پردازش زبان طبیعی می‌توان نام برد؛ نظیر پیکره‌هایی برای نشانه‌گذاری نقش اجزای جمله⁸ [4]، تجزیه‌کردن وابستگی⁹ [17]، خلاصه‌سازی¹⁰ متن [14]،

¹ Cross language evaluation forum

² Factoid

³ Question Classification

⁴ Coarse-grained

⁵ Fine-grained

⁶ Topic Detection

⁷ <http://quranjooy.itrc.ac.ir/>

⁸ POS Tagging

⁹ Dependency Parsing

¹⁰ Summarization

¹¹ Event Extraction

¹² Sentiment Analysis

¹³ Open Domain

¹⁴ Closed Domain

معرفی شدند. سؤالات غیرحقیقت^۱، اغلب حاوی پاسخ‌های طولانی هستند و پاسخ‌گویی به آن‌ها دشوارتر است. این سؤالات در زبان فارسی به‌طور معمول حاوی کلمات پرسشی چرا و چگونه هستند [8].

۴- معماری سامانه‌های پرسش و پاسخ

به‌طور کلی، سامانه‌های پرسش و پاسخ از سه مؤلفه اصلی پردازش سؤال^۲، بازیابی اطلاعات^۳ و استخراج پاسخ^۴ تشکیل می‌شوند [5-6]. معماری یک سامانه پرسش و پاسخ در شکل (۱) نشان داده شده است.

۴-۱- مؤلفه پردازش سؤال

در مؤلفه پردازش سؤال، کلیه پردازش‌های لازم بر روی سؤال اعمال می‌شود تا روابط ساختاری و معنایی موجود در کلمات سؤال، استخراج شود. این مؤلفه از سه بخش تشخیص نوع سؤال، استخراج کلمات کلیدی و تشخیص نوع پاسخ مورد انتظار تشکیل شده است.

۴-۱-۱- تشخیص نوع سؤال

در این مرحله، تعیین می‌شود که سؤال مطرح شده، در کدام دسته از انواع سؤالات قرار می‌گیرد (انواع سؤالات در بخش ۳-۲ معرفی شده است). این مرحله، یکی از گام‌های ضروری و اولیه در سامانه‌های پرسش و پاسخ است. از آنجا که پاسخ‌گویی به سؤالات مختلف، نیازمند به‌کارگیری روش‌های فنی مختلف و مناسب با نوع سؤال است، بنابراین تشخیص درست نوع سؤال مطرح‌شده، بسیار حائز اهمیت است [2,8].

۴-۱-۲- استخراج کلمات کلیدی

یکی دیگر از مراحل مهم در پردازش سؤال، تشخیص کلمات کلیدی موجود در صورت سؤال است. تشخیص درست کلمات کلیدی موجود در سؤال، می‌تواند در بازیابی پاسخ مناسب، بسیار مفید باشد. برای مثال در سامانه پرسش و پاسخ مذهبی، اگر بدانیم که یکی از کلمات کلیدی به‌کاررفته در صورت سؤال، نماز است، می‌توانیم بازه مربوط به بازیابی پاسخ را بسیار محدود کنیم. به‌منظور تشخیص کلمات

معنایی که بیشتر مورد توجه واقع شده است، سؤالات را به هشت دسته حقیقت^۱، فهرست، تعریفی یا توصیفی^۲، فرضیه‌ای^۳، علیتی^۴، رابطه‌ای^۵، رویه‌ای^۶ و تأییدی^۷ تقسیم می‌کند [8].

سؤال حقیقت، به‌طور معمول در زبان انگلیسی، با استفاده از کلمات پرسشی WH شروع می‌شود و حاوی کلمات پرسشی چه کسی، چه چیزی، چه وقت و کجاست. پاسخ این سؤال، یک حقیقت یا واقعیت بیان شده در متن و اغلب یک موجودیت عددی و یا اسمی است. برای مثال "شماره تلفن دانشگاه گیلان چیست؟" که پاسخ آن یک موجودیت عددی است.

سؤال فهرست، سؤالی است که پاسخ آن، فهرستی از موجودیت‌های متن است. برای مثال "زکات بر چه چیزهایی واجب است؟ گندم، جو، خرما، کشمش، طلا، نقره، شتر، گاو و گوسفند."

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

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

سؤال علیتی، جوپای دانستن اطلاعات و توضیحی از یک رویداد است و به‌طور معمول با چرا آغاز می‌شود. برای مثال "چرا مردم به بیماری قند مبتلا می‌شوند؟"

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

سؤال رویه‌ای، سؤالی است که پاسخ آن، فهرستی از دستورالعمل‌ها برای انجام عملیات ذکر شده در سؤال است. برای مثال "مراحل گرفتن وضو چگونه است؟"

سؤال تأییدی، برای رویداد مطرح‌شده در صورت سؤال، به جواب بله یا خیر نیاز دارد. برای مثال "آیا پردازش زبان طبیعی، زیرشاخه‌ای از علم کامپیوتر است؟ بله"

از نگاهی دیگر، سؤالات به دو دسته کلی حقیقت و غیرحقیقت تقسیم می‌شوند. سؤالات حقیقت که پیش‌تر

- 1 Factoid
- 2 Definition or description question
- 3 Hypothetical question
- 4 Causal question
- 5 Relationship question
- 6 Procedural question
- 7 Confirmation question

⁸ Non-factoid question
⁹ Question Processing
¹⁰ Information Retrieval
¹¹ Answer Extraction

کلیدی، از ویژگی‌های زبانی و الگوریتم‌های یادگیری ماشین استفاده می‌شود [8].

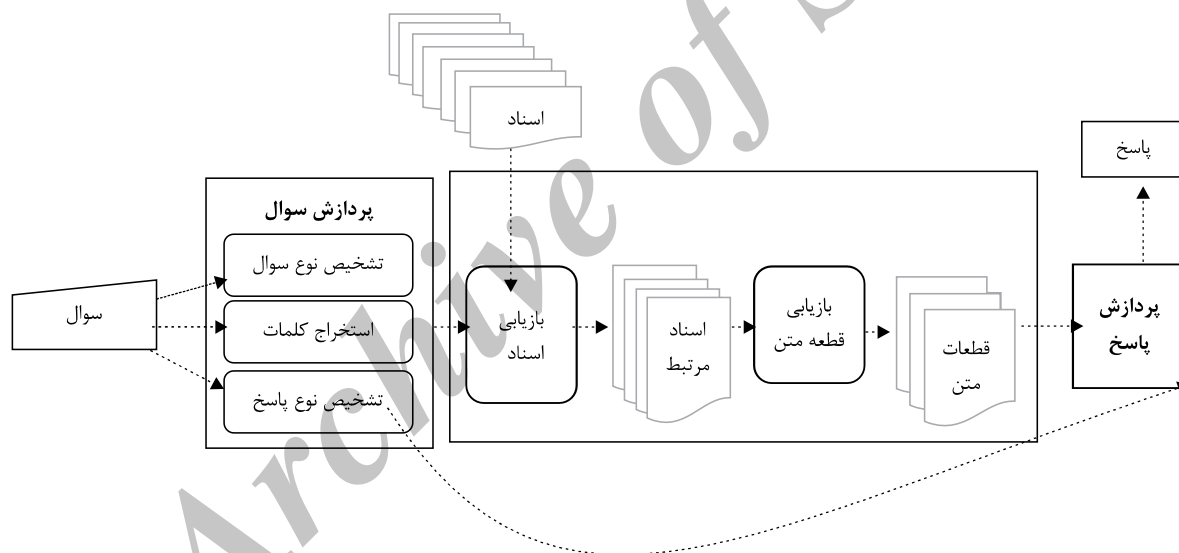
پاسخ مورد انتظار بوده است [8]. بدین جهت، تشخیص مناسب نوع پاسخ مورد انتظار، بسیار حائز اهمیت است.

(جدول-1): طبقه‌بندی نوع پاسخ مورد انتظار
(Table-1): classification of expected answer type

ریزدانه	درشت‌دانه
Abbreviation-expression	ABBREVIATION
Animal - Body - Color - Creative - Currency - Disease/Medicine - Lang - Event - Food - Instrument - Letter - Other - Plant - Product - Religion - Sport - Substance - Symbol - Technique - Term - Vehicle - Word	ENTITY
Definition - Description - Manner - Reason	DESCRIPTION
Group - Individual - Title - Description	HUMAN
City - Country - Mountain - Other - State	LOCATION
Code - Count - Date - Distance - Money - Order - Other - Period - Percent - Speed - Temp - Size - Weight	NUMERIC

۳-۱-۴ - تشخیص نوع پاسخ مورد انتظار

یک مرحله بسیار مهم در سامانه‌های پرسش و پاسخ، تشخیص طبقه معنایی پاسخ مورد انتظار است. در واقع، هدف این است که با توجه به سؤال، دریابیم که پاسخ مدنظر از این سؤال در چه طبقه معنایی شخص، زمان، مکان و یا امثال آن قرار می‌گیرد [8]؛ سپس از میان پاسخ‌های منتخب که از متن استخراج شده است، پاسخی را که مطابق با این طبقه معنایی هستند، شناسایی کنیم. برای مثال، اگر با توجه به سؤال، دریابیم که نوع پاسخ مورد انتظار در مورد یک موجودیت عددی است، بازه‌ای که باید برای جواب جستجو کنیم، بسیار محدود می‌شود. برای این منظور، استفاده از روش‌های یادگیری ماشین [24] بسیار متداول است. پژوهش‌ها نشان داده است که ۳۶,۴ درصد از خطاها در سامانه‌های پرسش و پاسخ، به دلیل تخمین نادرست نوع



(شکل-۱): معماری سامانه پرسش و پاسخ
(Figure-1): Architecture of question answering system

۲-۴ - مؤلفه بازیابی اطلاعات

در سال‌های اخیر، طبقه‌بندی‌های مختلفی بر روی نوع پاسخ مورد انتظار انجام شده است. یکی از این طبقه‌بندی‌ها که مورد توجه زیاد پژوهش‌گران قرار گرفته است (هیرشمن و گایروسکاس، ۲۰۰۲)، یک طبقه‌بندی دوسطحی بوده است که دارای شش طبقه درشت‌دانه و پنجاه طبقه ریزدانه است. این طبقه‌بندی مطابق جدول (۱) است.

این مؤلفه، به‌عنوان ورودی یک پرس‌وجو و مجموعه‌ای از اسناد را دریافت می‌کند؛ سپس توسط یک تابع، میزان ارتباط پرس‌وجو با مجموعه‌ی اسناد دریافتی را محاسبه کرده و بر این اساس به هر سند، یک رتبه تعلق می‌گیرد. به این ترتیب اسناد بازیابی‌شده، در فهرستی براساس امتیازشان مرتب می‌شوند [6]. در یک سامانه پرسش و پاسخ، مؤلفه بازیابی

سند وظیفه صافی کردن اسناد را بر عهده دارد. به بیان دیگر در کوتاه‌ترین و قابل قبول‌ترین زمان ممکن، اسناد مرتبط به یک پرسش را بازیابی کرده و اسناد نامرتب را جدا می‌کند. به‌طور معمول توابعی که در بخش بازیابی سند به کار می‌روند، نسبت به توابع به‌کاررفته در بخش پردازش پاسخ، کم‌هزینه‌تر هستند تا این مرحله در سریع‌ترین زمان ممکن انجام شود [5]. مؤلفه بازیابی اطلاعات در سامانه پرسش و پاسخ، اهمیت ویژه‌ای دارد؛ زیرا گام‌های بعدی، تنها در صورتی که نتایج بازگردانده‌شده توسط مؤلفه بازیابی اطلاعات صحیح و قابل اعتماد باشند، پاسخ صحیحی برای پرسش می‌توانند بیابند [5-7].

۳-۴ - مؤلفه استخراج پاسخ

مؤلفه سوم، استخراج پاسخ است که در این بخش، از میان اطلاعاتی که توسط مؤلفه دوم برگردانده شده، صحیح‌ترین و خلاصه‌ترین پاسخ، به‌عنوان جواب نهایی به کاربر برگردانده می‌شود. به‌طور معمول در این مرحله به‌منظور پاسخ‌گویی به سؤالات غیرحقیقت از روش‌ها و الگوریتم‌های پیچیده‌تری استفاده می‌شود و به هر یک از پاسخ‌های برگردانده‌شده توسط مؤلفه بازیابی اطلاعات، دوباره رتبه‌ای انتساب داده می‌شود و سرانجام بهترین پاسخ، برگردانده می‌شود [8]. به‌منظور پاسخ‌گویی به سؤالات حقیقت در مؤلفه استخراج پاسخ، از تشخیص‌دهنده موجودیت‌های ناممندا استفاده می‌شود تا از میان پاسخ‌های منتخب، پاسخی را برگزیند که با نوع پاسخ مورد انتظار منطبق باشد [5].

۵- توسعه پیکره

در این بخش منبع داده مورد استفاده در این پژوهش مورد بررسی قرار خواهد گرفت و یک طبقه‌بندی نوع پاسخ مورد انتظار در مقاله مطرح خواهد شد. در ادامه جزئیات مربوط به اجزای پیکره به تفصیل بیان خواهد شد.

۱-۵ - منبع داده و مراحل توسعه پیکره

بدون شک، در روند توسعه هر پیکره، یکی از مهم‌ترین گام‌ها، انتخاب یک منبع داده‌ای مناسب است. با توجه به اینکه تهیه یک پیکره پرسش و پاسخ با دامنه باز بسیار دشوار است، هدف ما تهیه یک پیکره پرسش و پاسخ با دامنه بسته است. یکی از زمینه‌های پرکاربرد پرسش و پاسخ در زبان فارسی،

سؤالات و احکام دینی است. با توجه به ماهیت سؤالات دینی و وجود شرط‌های متعدد در سؤال و جواب، پاسخ‌گویی به این سؤالات، بسیار چالش‌برانگیز است. بدین منظور، برای توسعه پیکره، این دامنه بسته، انتخاب شده است.

داده‌های خام مورد استفاده، شامل دو فایل متنی استفتائات آیت‌الله العظمی خامنه‌ای و رساله آیت‌الله العظمی مکارم شیرازی است که توسط مرجع تحقیقات علوم اسلامی (نور) فراهم شده است. به‌منظور توسعه پیکره، برای هر دو فایل متنی اولیه، اسنادی به شکل XML ایجاد شده است. اجزای این اسناد XML و جزئیات آن، در بخش ۷ شرح داده خواهد شد. در مرحله بعد، ابزاری به‌منظور نشانه‌گذاری طراحی شده است که در بخش ۸ توضیح داده می‌شود. باید خاطر نشان کرد که این پیکره، توسط دو نفر و در طی مدت تقریبی چهار ماه، نشانه‌گذاری شده است.

۲-۵ - ارائه طبقه‌بندی نوع پاسخ مورد انتظار

در رساله

با توجه به اینکه هدف، کار بر روی رساله مراجع تقلید است، لازم است تغییراتی در جدول (۱) اعمال شود تا طبقه‌بندی انجام شده برای نوع پاسخ مورد انتظار، مناسب با رساله باشد. بر این اساس، مطابق با جدول (۲)، انواع پاسخ‌های مورد انتظار، به هشت طبقه درشت‌دانه و ۳۵ طبقه ریزدانه تقسیم می‌شوند.

گروه احکام، حاوی سؤالاتی است که پاسخ آن‌ها می‌تواند، حرام، حلال، مستحب، باطل، واجب، مکروه، مباح، نجس، احتیاط مستحب، احتیاط واجب و مستحب مؤکد باشد. برای مثال "در وضو، شستن صورت برای بار سوم چه حکمی دارد؟ حرام است."

گروه تأییدی، حاوی سؤالاتی است که پاسخ آن‌ها، بله و خیر است. برای مثال "آیا خوردن گوشت خوک، حرام است؟ بله."

گروه توصیفی، شامل سؤالاتی است که پاسخ آن‌ها طولانی است که به سه زیرگروه، تعریفی، توصیفی و شیوه‌ای تقسیم می‌شوند. زیرگروه تعریفی شامل سؤالاتی است که پاسخ آن، تعریف یک عبارت به‌کاررفته در صورت سؤال است. برای مثال "تعریف آب کر چیست؟" که پاسخ آن، تعریف کلمه به‌کاررفته در صورت سؤال است. زیرگروه توصیفی شامل سؤالاتی است که پاسخ آن، اطلاعات تعریفی و یا توصیف یک

²Computer Research Center of Islamic Science(CRCIS)

¹Named Entity Recognizer

شامل هر سؤالی است که در مورد موجودیت باشد؛ ولی در هیچ یک از طبقه‌های فوق جای نگیرد.

گروه زمانی، حاوی سؤالاتی است که پاسخ آن‌ها، در مورد زمان است. این گروه از سؤالات، به سه زیرگروه اسمی، تاریخی و دوره تقسیم می‌شوند. زیرگروه اسمی، سؤالاتی هستند که پاسخ آن‌ها، یک اسم زمانی است. به عنوان مثال نام ماه یا نام اعیاد. برای مثال "مسلمانان در چه ماهی روزه می‌گیرند؟ ماه رمضان" زیرگروه تاریخی، سؤالاتی را شامل می‌شود که پاسخ آن‌ها، یک تاریخ است. زیرگروه دوره، سؤالاتی هستند که پاسخ آن‌ها، یک دوره است. به عنوان مثال پاسخ آن چند روز، چند ماه و یا امثال آن است.

گروه مکانی شامل سؤالاتی است که پاسخ آن‌ها، یک مکان باشد. به عنوان مثال نام مسجد. با توجه به اینکه احتمال می‌رود تعداد این سؤالات در رساله کم باشد، زیرگروهی برای آن در نظر گرفته نشده است. برای مثال "در چه اماکنی، مسافر می‌تواند نماز را کامل بخواند؟ مسجدالحرام، مسجدالنبی و مسجد کوفه."

گروه عددی، شامل سؤالاتی است که پاسخ آن‌ها، یک عدد است که به پنج زیرگروه تعداد، فاصله، پول، نسبت و غیره تقسیم می‌شوند. زیرگروه تعداد، شامل سؤالاتی است که پاسخ آن‌ها، تعداد را نشان می‌دهد. برای مثال "نماز آیات، چند رکعت دارد؟ دو رکعت." زیرگروه فاصله، شامل سؤالاتی است که پاسخ آن‌ها، فاصله باشد. برای مثال "کسی که شغل او مسافرت نباشد، باید قصد چند فرسخ بنماید تا نماز شکسته بخواند؟ هشت فرسخ." زیرگروه پول، دربرگیرنده سؤالاتی است که پاسخ آن‌ها، مقدار پولی باشد. برای مثال "زکات فطره به چه مقدار است؟". زیرگروه نسبت، حاوی سؤالاتی است که پاسخ آن‌ها، بیان‌کننده نسبت باشد. به عنوان مثال خمس و ثلث. زیرگروه غیره، شامل سایر مقادیر عددی است که در هیچ یک از طبقه‌های بالا جای نگرفته است.

گروه اسمی، حاوی سؤالاتی است که پاسخ آن‌ها، در مورد یک اسم خاص است. به عنوان مثال اسامی غسل‌ها و یا هر اسم دیگری که در هیچ یک از گروه‌های دیگر، قرار نگرفته است. برای مثال "هرگاه به قسمت جدا شده از بدن انسانی که دارای استخوان است، دست بزنیم، چه باید بکنیم؟ باید غسل مس میت انجام شود."

۳-۵- اجزای پیکره پیشنهادی

همان‌طور که پیشتر اشاره شد، اسناد موجود در پیکره در

رویداد است. برای مثال "راه‌های شناخت مجتهد چیست؟". زیرگروه شیوه‌ای دربرگیرنده سؤالاتی است که به‌طور معمول با کلمه کلیدی چگونه آغاز می‌شود و پاسخ آن، توضیحی در مورد نحوه انجام کار است. برای مثال "چگونه وضو می‌گیریم؟".

(جدول-۲): طبقه‌بندی نوع پاسخ موردانتظار در پیکره

پیشنهادی

(Table-2): classification of expected answer type in proposed corpus

درشت‌دانه	ریزدانه
احکام	حرام، حلال، مستحب، باطل، واجب، مکروه، مباح، احتیاط مستحب، احتیاط واجب، نجس، باطل، مستحب موكد
تاییدی	بله، خیر
توصیفی	تعریفی، توصیفی، شیوه‌ای،
موجودیت	حیوان، بدن، غذا، مرض-دارو، محصول، ورزش، رویداد، شخص، وسیله‌نقلیه، غیره
زمانی	اسمی، تاریخی، دوره،
مکانی	-
عددی	تعداد، فاصله، پول، نسبت، غیره
اسمی	-

گروه موجودیت، حاوی سؤالاتی است که پاسخ آن‌ها یک موجودیت است که شامل ده زیرگروه حیوان، بدن، غذا، مرض-دارو، محصول، ورزش، رویداد، شخص، و وسیله‌نقلیه و غیره است. زیرگروه حیوان، شامل سؤالی است که پاسخ آن، نام یک حیوان است. برای مثال "کدام حیوانات نجس هستند؟ سگ، خوک". زیرگروه بدن، حاوی سؤالی است که پاسخ آن، یک جزء از بدن باشد. مانند، دست، صورت و پا. زیرگروه غذا دربرگیرنده سؤالی است که پاسخ آن یک مواد غذایی مانند برنج، خرما و یا امثال آن باشد. برای مثال "به چه محصولاتی، زکات تعلق نمی‌گیرد؟ برنج". زیرگروه مرض-دارو حاوی سؤالاتی است که پاسخ آن، نام بیماری یا دارو باشد. برای مثال "وجود چه بیماری در زوجین، عقد را باطل می‌کند؟ دیوانگی". زیرگروه محصول، شامل سؤالی است که پاسخ آن، یک محصول غیرغذایی به عنوان مثال پنبه باشد. زیرگروه ورزش، حاوی سؤالی است که پاسخ آن، یک رشته ورزشی مثلاً شنا باشد. زیرگروه رویداد، حاوی سؤالی است که پاسخ آن، یک رویداد باشد. برای مثال "در چه مواقعی، خواندن نماز آیات واجب است؟ خورشیدگرفتگی، ماه‌گرفتگی، زلزله". زیرگروه شخص، شامل سؤالی است که پاسخ آن، یک شخص باشد. به عنوان مثال پیامبر اکرم. زیرگروه وسیله نقلیه، دربرگیرنده سؤالی است که پاسخ آن، یک وسیله نقلیه باشد. به عنوان مثال کشتی. زیرگروه غیره،

قالب فایل‌های XML ذخیره شده‌اند. در این بخش مهم‌ترین اجزای این فایل‌ها با جزییات بیشتری معرفی می‌شوند.

۱-۳-۵- فایل XML سؤال‌های غیرحقیقت استفتائات آیت‌الله العظمی خامنه‌ای

این فایل، حاوی ۲۱۱۸ سؤال غیرحقیقت است. ساختار پیشنهادی این فایل به همراه یک مثال، در شکل (۲) نشان داده شده است.

عنصر TREATISE، تمامی سؤالات موجود در استفتائات آیت‌الله العظمی خامنه‌ای را در برمی‌گیرد. عنصر Question، کلیه اطلاعات مربوط به یک سؤال و پاسخ آن را شامل می‌شود. عنصر QuestionID، شماره سؤال را نگهداری می‌کند که این عنصر منحصربه‌فرد است و برای تمایز میان سؤالات پیکره در نظر گرفته شده است. در عنصر QuestionTitle، عنوان و دسته کلی سؤال و در عنصر QuestionSubTitle، زیرعنوان مربوط به سؤال ذخیره می‌شود. نگهداری عنوان و زیرعنوان سؤال، به منظور یافتن پاسخ مناسب در سامانه‌های پرسش و پاسخ، می‌تواند مورد استفاده قرار گیرد. برای مثال در صورتی که عنوان سؤال، تقلید تشخیص داده شود، بازه استخراج پاسخ را به فصل تقلید از کتاب رساله آیت‌الله العظمی مکارم محدود می‌توان کرد. تشخیص مناسب عنوان و زیرعنوان سؤال، منجر به بهبود عملکرد مؤلفه استخراج پاسخ می‌شود.

در عنصر QuestionContent، متن سؤال ذخیره و در عنصر QuestionType، نوع سؤال مدنظر نگهداری می‌شود. انواع سؤالات در بخش ۳-۲ به طور کامل معرفی شد که در تهیه پیکره، برای عنصر نوع سؤال، تنها مقدار حقیقت و غیرحقیقت نگهداری شده و از نگهداری دسته‌های ریزتر، اجتناب شده است. در آینده، دسته‌بندی ریزتر انواع سؤالات را به پیکره می‌توان اضافه کرد.

عنصر Answer، اطلاعات مربوط به پاسخ را به‌طور کامل نگهداری می‌کند. در عنصر AnswerContent، متن پاسخ مربوطه قرار خواهد گرفت. CourseGrainedCategory، نوع پاسخ مورد انتظار را در سطح درشت‌دانه و FineGrainedCategory، نوع پاسخ مورد انتظار را در سطح ریزدانه، طبق جدول (۲) نگهداری می‌کنند.

عنصر Answer، اطلاعات مربوط به پاسخ را به‌طور کامل نگهداری می‌کند. در عنصر AnswerContent، متن پاسخ مربوطه قرار خواهد گرفت. CourseGrainedCategory، نوع پاسخ مورد انتظار را در سطح درشت‌دانه و FineGrainedCategory، نوع پاسخ مورد انتظار را در سطح ریزدانه، طبق جدول (۲) نگهداری می‌کنند.

```
<TREATISE>
<Question>
<QuestionID>1</QuestionID>
<QuestionTitle>
احکام تقلید
</QuestionTitle>
<QuestionSubtitle>
راههای سه گانه: احتیاط، اجتهاد، تقلید
</QuestionSubtitle>
<QuestionContent>
آیا تقلید ادله شرعی دارد؟
</QuestionContent>
<QuestionType>
غیرحقیقت
</QuestionType>
<Answer>
<AnswerContent>
تقلید ادله شرعی دارد و عقل نیز حکم می‌کند
که شخص ناآگاه به احکام دین باید به مجتهد
جامع الشرائط مراجعه کند.
</AnswerContent>
<CourseGrainedCategory>
توصیفی
</CourseGrainedCategory>
<FineGrainedCategory>
توصیفی
</FineGrainedCategory>
</Answer>
<DifficultFromQuestioner>
متوسط
</DifficultFromQuestioner>
<DifficultFromResponder>
متوسط
</DifficultFromResponder>
</Question>
</TREATISE>
```

(شکل-۲): فایل XML سؤالات غیرحقیقت
(Figure-2): XML file of non-factoid questions

عنصر Answer، اطلاعات مربوط به پاسخ را به‌طور کامل نگهداری می‌کند. در عنصر AnswerContent، متن پاسخ مربوطه قرار خواهد گرفت. CourseGrainedCategory، نوع پاسخ مورد انتظار را در سطح درشت‌دانه و FineGrainedCategory، نوع پاسخ مورد انتظار را در سطح ریزدانه، طبق جدول (۲) نگهداری می‌کنند.

۲-۳-۵- فایل XML سؤالات حقیقت استفتائات آیت‌الله العظمی خامنه‌ای

علاوه بر سؤالات غیرحقیقت که در بخش ۱-۷ توضیح داده شد، پیکره پیشنهادی، ۲۰۵۱ سؤال حقیقت از استفتائات آیت‌الله العظمی خامنه‌ای را نیز شامل می‌شود. ساختار این فایل به همراه یک مثال، در شکل (۳) نشان داده شده است.

متن پاسخ حقیقت مربوطه را ذخیره می‌کند. سایر عناصر پیش‌تر معرفی شده‌اند.

۴-۵- فایل XML رساله آیت‌الله العظمی مکارم شیرازی

این فایل، شامل ۲،۴۵۶ مسأله از کتاب رساله آیت‌الله العظمی مکارم شیرازی است. ساختار این فایل به همراه یک مثال، در شکل (۴) نشان داده شده است.

```
<?xml version="1.0" encoding="utf-8"?>
<MakaremCorpus>
  <Problem>
    <ProblemID>
      1
    </ProblemID>
    <ProblemTitle>
      تقلید
    </ProblemTitle>
    <ProblemSubTitle>
      احکام تقلید
    </ProblemSubTitle>
    <ProblemContent>
      هیچ مسلمانی نمی‌تواند در اصول دین تقلید نماید
      بلکه باید آنها را به فراخور حال خویش بداند
      ولی در فروع دین یعنی احکام و دستورات عملی
      اگر مجتهد باشد به عقیده خود عمل می‌کند
      و اگر مجتهد نباشد باید از مجتهدی تقلید کند
    </ProblemContent>
    <PageNumber>
      19
    </PageNumber>
  </Problem>
</MakaremCorpus>
```

(شکل-۴): فایل XML آیت‌الله مکارم شیرازی
(Figure-4): XML file of Ayatollah Makarem shirazi

عنصر MakaremCorpus، تمامی مسأله‌های رساله آیت‌الله العظمی مکارم را دربرمی‌گیرد. عنصر Problem، اطلاعات مربوط به یک مسأله را به‌طور کامل نگهداری می‌کند. ProblemID، شماره مسأله مربوط به رساله است که به‌عنوان شماره پاراگراف نیز می‌تواند به کار رود که این شماره منحصر به فرد است و برای تمایز میان مسأله‌های رساله آیت‌الله العظمی مکارم، در نظر گرفته شده است.

ProblemTitle، عنوان و دسته کلی مسأله استخراج شده و ProblemSubTitle، زیرعنوان یا زیرعنوان‌های مربوط به یک مسأله را نگهداری می‌کنند. یکی از دلایل نگهداری عنوان و زیرعنوان برای استفتانات آیت‌الله العظمی خامنه‌ای و رساله آیت‌الله العظمی مکارم، این است که در صورت تشخیص مناسب عنوان و زیرعنوان یک سؤال، بتوان بازه تشخیص پاسخ را در رساله، محدودتر کرد. ProblemContent،

```
<?xml version="1.0" encoding="utf-8"?>
<FactoidQuestionsCorpus>
  <FactoidQuestion>
    <FactoidQuestionId ParentId="1">
      1
    </FactoidQuestionId>
    <FactoidQuestionContent>
      آیا تقلید، اهل‌ء شرعی دارد؟
    </FactoidQuestionContent>
    <Answer>
      <FactoidAnswerContent>
        بله
      </FactoidAnswerContent>
      <CourseGrainedCategory>
        تائیدی
      </CourseGrainedCategory>
      <FineGrainedCategory>
        بله
      </FineGrainedCategory>
    </Answer>
    <DifficultFromQuestioner>
      آسان
    </DifficultFromQuestioner>
    <DifficultFromResponder>
      آسان
    </DifficultFromResponder>
  </FactoidQuestion>
</FactoidQuestionsCorpus>
```

(شکل-۳): فایل XML سؤالات حقیقت
(Figure-3): XML file of factoid questions

این سؤالات حقیقت، که برگرفته از سؤالات غیر حقیقت بخش ۷-۱ هستند، توسط اعضای تیم، با استفاده از نرم‌افزاری که در بخش ۸ معرفی خواهد شد، ایجاد شده است. هدف ایجاد مجموعه سؤالات حقیقت، این است که پیکره تهیه شده، پیکره جامعی باشد و بتواند در طراحی سامانه‌های پرسش و پاسخ حقیقت و غیرحقیقت به کار گرفته شود.

عنصر FactoidQuestionsCorpus، تمامی سؤالات حقیقت موجود را شامل می‌شود. عنصر FactoidQuestion، اطلاعات مربوط به یک سؤال حقیقت را نگهداری می‌کند. عنصر FactoidQuestionId، یک شناسه منحصر به فرد را برای هر سؤال حقیقت، ذخیره می‌کند. این عنصر دارای ویژگی parentId است که این ویژگی برابر با شماره سؤال (QuestionID) مربوطه از پیکره سؤالات غیرحقیقت است که این سؤال حقیقت از آن ناشی شده است. لازم به ذکر است که برای هر سؤال غیرحقیقت، بیش از یک سؤال حقیقت می‌تواند ذخیره شود.

در عنصر FactoidQuestionContent، متن سؤال حقیقت مدنظر نگهداری می‌شود. FactoidAnswerContent،

Keys در بردارنده اطلاعات مربوط به سؤالات مرتبط با این مسأله است. هر سؤال مرتبط با این مسأله در عنصر Question، ذخیره می‌شود. عنصر Key، میزان ارتباط سؤال با مسأله را نشان می‌دهد. این عنصر، دو مقدار ۱ یا ۲ را می‌تواند شامل شود.

مقدار ۲، نشان‌دهنده ارتباط کامل و مقدار ۱ نشان‌دهنده ارتباط جزئی مسأله با سؤال است. نگهداری ارتباط مسأله با سؤال در دو سطح، در ارزیابی سامانه‌های پرسش و پاسخ، موثر خواهد بود تا بتوان ارزیابی مناسب‌تری از عملکرد سامانه داشت.

مثلاً اگر سیستم ۱ نتواند پاسخی با درجه‌ی ارتباط ۲ را بازیابی کند، اما قادر باشد پاسخ‌هایی با درجه‌ی ارتباط ۱ را بازیابی کند، در مقابل سیستم ۲ که نتواند پاسخی با درجه ارتباط ۱ یا ۲ را بازیابی کند، عملکرد بهتری خواهد داشت. ویژگی QuestionNumber، نشان‌دهنده شماره سؤال مرتبط با این مسأله است. سایر عناصر پیش‌تر معرفی شده‌اند.

۶- ابزار نشانه‌گذاری و وضعیت در دسترس بودن پیکره

جهت انجام فرآیند نشانه‌گذاری اسناد XML، یک نرم‌افزار به نام ابزار نشانه‌گذاری پرسش و پاسخ رساله، توسعه داده شده است که بخشی از ویژگی‌های این نرم‌افزار در شکل (۶) نشان داده شده است. با استفاده از این نرم‌افزار، به راحتی می‌توان برچسب‌های مربوط به پیکره را مشاهده، به روزرسانی و حذف کرد. همچنین امکان ایجاد سؤالات حقیقت، ارتباط سؤال‌های استفتائات آیت‌الله العظمی خامنه‌ای و مسائل آیت‌الله العظمی مکارم شیرازی نیز در این نرم‌افزار فراهم شده است تا کلیه برچسب‌های معرفی شده در بخش قبل، به راحتی توسط افراد متخصص اعمال شود.

نشانه‌گذاری یک پیکره بدون استفاده از نرم‌افزار، بسیار دشوار است. استفاده از نرم‌افزار به منظور نشانه‌گذاری اسناد موجب افزایش سرعت و دقت در تهیه پیکره می‌شود. جهت انجام فرآیند نشانه‌گذاری اسناد XML، یک نرم‌افزار به نام ابزار نشانه‌گذاری پرسش و پاسخ مذهبی^۱ به زبان C# توسعه داده شده است که رابط گرافیکی این نرم‌افزار در شکل (۶) نشان داده شده است. مهم‌ترین ویژگی‌های این نرم‌افزار عبارتند از:

¹ Treatise QA Annotation Tool (TQAAT)

² Religious QA Annotation Tool (RQAAT)

محتوای متن یک مسأله را شامل می‌شود. در عنصر PageNumber، شماره صفحه مسأله از پیکره آیت‌الله العظمی مکارم ذخیره می‌شود. یکی از دلایل نگهداری شماره صفحه و پاراگراف، این است که مؤلفه بازیابی اطلاعات بتواند عملیات بازیابی را بر اساس شماره صفحه و یا شماره پاراگراف انجام دهد. به عنوان مثال مؤلفه بازیابی اطلاعات، پاراگراف‌های مرتبط با یک سؤال را بازیابی کند. بدین صورت مؤلفه بازیابی اطلاعات در چند سطح بازیابی صفحه، پاراگراف و پاسخ دقیق، می‌تواند اعمال شود.

۱- ۴-۵- فایل XML ارتباط سؤالات استفتائات آیت‌الله العظمی خامنه‌ای و مسائل رساله آیت‌الله العظمی مکارم

در این فایل، ارتباط میان سؤالات موجود در استفتائات آیت‌الله العظمی خامنه‌ای و مسأله‌های موجود در پیکره آیت‌الله العظمی مکارم، نگهداری می‌شود. ساختار این فایل به همراه یک مثال در شکل (۵) نشان داده شده است.

```
<?xml version="1.0" encoding="utf-8"?>
<MakaremCorpus>
  <Problem>
    <ProblemNumber>
      2
    </ProblemNumber>
    <Keys>
      <Question>
        <Key QuestionNumber="1">
          2
        </Key>
        <DifficultFromQuestioner>
          آسان
        </DifficultFromQuestioner>
        <DifficultFromResponder>
          آسان
        </DifficultFromResponder>
      </Question>
    </Keys>
  </Problem>
</MakaremCorpus>
```

(شکل-۵): فایل xml روابط بین سؤالات و پاسخ‌ها
(Figure-5): XML file of relation between questions and answers

عنصر MakaremCorpus، شامل تمامی مسأله‌هایی از رساله آیت‌الله العظمی مکارم است که با سؤالات موجود در استفتائات آیت‌الله العظمی خامنه‌ای، ارتباط دارند. عنصر Problem، حاوی اطلاعات مسأله‌ای است که دست‌کم با یکی از سؤالات استفتائات آیت‌الله العظمی خامنه‌ای ارتباط دارد. ProblemNumber، شماره مربوط به مسأله است. عنصر



۶-۱- افزایش خوانایی

این نرم‌افزار فایل خام اولیه سؤال و جواب را در قالب متن دریافت می‌کند. از آن جایی که پردازش فایل متنی کار دشواری است، این نرم‌افزار فایل خام اولیه را به شکل xml تبدیل می‌کند. ساختار فایل xml ذخیره شده در این سامانه در بخش قبل به‌طور کامل شرح داده شده است. تبدیل فایل متنی به XML امکان خوانایی و توسعه پذیری پیکره را افزایش می‌دهد.

۶-۲- فرم مربوط به سؤالات غیر حقیقت

به‌منظور نشانه‌گذاری اسناد غیر حقیقت، فرمی طراحی شده است که بتوان به‌سرعت برچسب‌های مناسب را به پیکره اضافه کرد. کاربر برای هر سؤال غیر حقیقت، نوع سؤال، دسته‌بندی معنایی پاسخ در سطح درشت‌دانه، دسته‌بندی معنایی پاسخ در سطح ریزدانه، سختی سؤال از نظر پرسش‌گر و سختی سؤال از نظر پاسخ‌دهنده را می‌تواند نشانه‌گذاری کند. به‌منظور سهولت فرآیند نشانه‌گذاری، این عناصر به صورت فهرست باز شو در نظر گرفته شده‌اند تا کاربر بتواند مقدار مورد نظر خود را از فهرست انتخاب کند.

۶-۳- فرم مربوط به سؤالات حقیقت

در نرم‌افزار پیشنهادی این امکان فراهم شده است که کاربر بتواند برای هر سؤال غیر حقیقت، سؤالات حقیقت مناسب را استخراج کند. برای این منظور، در فرم مربوط به سؤالات غیر حقیقت، برای هر سؤال غیر حقیقت وارد فرم مربوط به اطلاعات حقیقت می‌توان شد. در این فرم فهرست سؤالات حقیقت مربوط به سؤال مدنظر نشان داده می‌شود و کاربر سؤالی را به مجموعه سؤال حقیقت اضافه و یا از مجموعه می‌تواند حذف کند. همچنین امکان ویرایش سؤالات حقیقت ذخیره‌شده نیز فراهم شده است. برای هر سؤال حقیقت، کاربر متن سؤال حقیقت و متن پاسخ حقیقت را می‌تواند وارد کند. همچنین امکان نشانه‌گذاری دسته‌بندی معنایی پاسخ در سطح درشت‌دانه، دسته‌بندی معنایی پاسخ در سطح ریزدانه، سختی سؤال از نظر پرسش‌گر و سختی سؤال از نظر پاسخ‌دهنده فراهم شده است. به‌طور مشابه به‌منظور سهولت فرآیند نشانه‌گذاری، این عناصر به صورت فهرست باز شو در نظر گرفته شده‌اند.

۶-۴- مستقل بودن از زبان

به‌منظور ساخت اسناد XML، از xml schema استفاده شده است. استفاده از xml schema امکان تغییر اسناد xml را به‌صورت پویا فراهم می‌کند. با اعمال تغییرات بر روی Xml schema، تغییرت مدنظر بر روی فایل xml ایجاد می‌شود و توسعه‌پذیری برنامه افزایش می‌یابد. همچنین این امر موجب می‌شود که برنامه مستقل از زبان باشد و با اعمال تغییرات کوچک در ساختار xml schema، ابزار طراحی شده بتواند در سایر زبان‌ها نیز به کار گرفته شود. همچنین این امکان فراهم شده است که طراحی فرم به‌صورت پویا باشد، به‌طوری که با تغییر ویژگی‌های xml schema، تغییرات مدنظر به‌صورت پویا بر روی فرم ایجاد شود. برای مثال با افزودن یک ویژگی به ساختار xml schema مربوط به سؤالات غیر حقیقت، این ویژگی به‌صورت پویا به فرم سؤالات غیر حقیقت اضافه می‌شود.

۶-۵- فرم جستجو

از دیگر ویژگی‌های این نرم‌افزار به امکان جستجو و گزارش‌گیری نیز می‌توان اشاره کرد. فرمی به‌منظور جستجو طراحی شده است که کاربر بتواند سؤالات را بر اساس نوع سؤال، سختی سؤال، عنوان سؤال و دسته معنایی پاسخ جستجو کند. برای مثال کاربر فهرست سؤالات حقیقت مربوط به فصل نماز را که در پیکره ذخیره شده است، می‌تواند مشاهده کند.

این نرم‌افزار خاص دامنه مذهبی نیست و به‌راحتی قابل توسعه است و می‌تواند به‌منظور نشانه‌گذاری کلیه سامانه‌های پرسش و پاسخ مورد استفاده قرار گیرد. اگر چه ابزار نشانه‌گذاری پرسش و پاسخ مذهبی به‌منظور نشانه‌گذاری اسناد فارسی ارائه شده ولی مستقل از زبان است و به‌راحتی در سایر زبان‌ها می‌تواند به کار گرفته شود.

نخستین نسخه پیکره شامل ۲۱۱۸ سؤال غیر حقیقت و ۲۰۵۰ سؤال حقیقت، از طریق نشانی گروه پردازش زبان طبیعی دانشگاه گیلان <http://nlp.guilan.ac.ir> به‌صورت رایگان در دسترس عموم قرار دارد.

۷- ارائه سامانه پرسش و پاسخ

به‌منظور ارزیابی صحت عملکرد پیکره پیشنهادی، لازم است یک سامانه پرسش و پاسخ معرفی شود. سامانه پیشنهادی از دو مؤلفه پیش‌پردازش سؤال و بازیابی سند تشکیل شده است.

۱-۷- پیش پردازش سؤال

گرفته است. یکی از معروف ترین موتورهای جستجوی بولین، موتور جستجوی لوسین^۶ است که به رایگان در دسترس عموم قرار دارد.^۷ این موتور جستجو بر اساس پرس و جوی بولین عمل می کند و بر پایه مدل فضای بردار $TF \cdot IDF$ ^۸ است.

نحوه عملکرد این موتورهای جستجو به این صورت است که ابتدا مجموعه پاسخ ها را نمایه گذاری می کنند، سپس یک سؤال را به عنوان ورودی دریافت کرده و با استفاده از مدل فرمول بازیابی خود، به هر یک از اسناد با توجه به میزان شباهتشان با سؤال، یک نمره نسبت می دهند؛ سپس اسناد را بر اساس نمره به ترتیب نزولی مرتب می کنند تا بهترین سند در رتبه یک قرار گیرد.

۳-۷- ارزیابی سامانه پیشنهادی

به منظور ارزیابی سامانه های پرسش و پاسخ، استفاده از دو معیار دقت^۹ و میانگین معکوس رتبه^{۱۰} (MRR)، متداول است. معیار دقت، درصد سؤالاتی را نشان می دهد که در میان n پاسخ برگردانده شده، دست کم دارای یک پاسخ درست هستند.

معیار MRR برای رتبه بندی سامانه هایی مناسب است که چندین پاسخ را برای یک سؤال برمی گردانند. فرض کنید Q یک مجموعه سؤال و T_i رتبه نخستین پاسخ درست برای سؤال i است که اگر هیچ پاسخ درستی برگردانده نشود، مقدارش برابر با صفر خواهد بود. این معیار طبق رابطه (۱) محاسبه می شود [8].

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

آزمایش های انجام شده بر روی سامانه معرفی شده نشان می دهد که سامانه پیشنهادی توانسته است به دقت ۸۲/۲۹ و MRR برابر با ۵۶/۷۳ درصد دست یابد. از آنجاکه تاکنون در زبان فارسی سامانه پرسش و پاسخی ارائه نشده و این سامانه نخستین سامانه پرسش و پاسخ در زبان فارسی است، در نتیجه امکان مقایسه سامانه پیشنهادی با سایر سامانه ها میسر نیست.

در این بخش ابتدا متن ورودی نرمال و تمیز می شود. یکی از مهم ترین اقداماتی که در نرمال سازی انجام می شود، اصلاح نیم فاصله های متن است. برای مثال کلماتی مانند "می شود" و "آن ها"، به ترتیب به شکل "می شود" و "آن ها" تبدیل می شوند. بعد از اعمال نرمال سازی، عمل جداسازی جملات بر روی متن نرمال شده، انجام می گیرد تا متن نرمال شده به تعدادی جمله تبدیل شود؛ سپس بر روی هر یک از جملات، جداسازی کلمات^۱ انجام می شود. یکی دیگر از پیش پردازش هایی که در این مرحله انجام شده، ریشه یابی است تا هر کلمه به شکل اصلی و ریشه آن تبدیل شود به عنوان مثال کلمه "کتاب ها" به "کتاب" تبدیل می شود.

به منظور اعمال این پیش پردازش ها در زبان فارسی، از ابزار هضم استفاده شده است.^۲ از ویژگی های این برنامه به قابلیت تمیز و مرتب کردن متن، تقطیع جمله ها و واژه ها، ریشه یابی واژه ها، تحلیل صرفی و تجزیه نحوی جمله می توان اشاره کرد.

یکی دیگر از اقداماتی که در این بخش انجام شده، حذف کلمات توقف است. در زبان فارسی فهرستی حاوی ۳۳۲ کلمه توقف ارائه شده^۳ که در این پژوهش از این مجموعه کلمات توقف استفاده شده است.

سؤال ورودی پس از اعمال پیش پردازش های بیان شده، به عنوان پرس و جوی بازیابی اطلاعات در اختیار مؤلفه بازیابی سند قرار می گیرد. لازم به ذکر است که کلیه پیش پردازش های انجام شده بر روی سؤال، به طور مشابه بر روی مجموعه پاسخ ها نیز اعمال می شود و مجموعه پاسخ ها پس از اعمال پیش پردازش های بیان شده در اختیار مؤلفه بازیابی سند قرار می گیرد.

۲-۷- بازیابی سند

این مؤلفه، پرس و جوی تولید شده از مؤلفه پردازش سؤال را دریافت کرده، سپس مشابه ترین اسناد موجود با این پرس و جو را از میان مجموعه پاسخ ها بازیابی می کند. به منظور بازیابی اسناد، به طور عمومی از موتورهای جستجوی منبع باز استفاده می شود که به تازگی استفاده از موتورهای جستجوی بولین در سامانه های پرسش و پاسخ مورد توجه پژوهشگران قرار

⁶ lucence

⁷ <https://lucene.apache.org/>.

⁸ term frequency-inverse document frequency

⁹ accuracy

¹⁰ mean reciprocal rank

¹ Normalize

² Sentence Tokenizer

³ Word Tokenizer

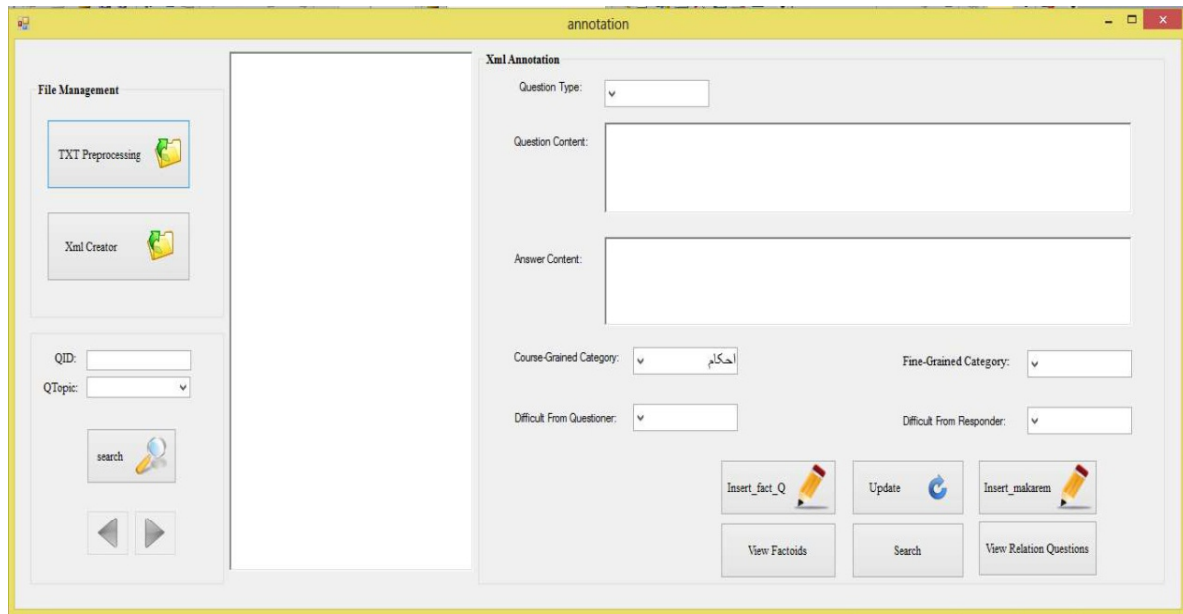
⁴ <https://github.com/mojtaba-khllash/JHazzm>

⁵ <http://members.unine.ch/jacques.savoy/clef/index.html>

۸- نتیجه گیری

در این مقاله، مراحل کامل توسعه یک پیکره متنی پرسش و پاسخ به زبان فارسی به طور کامل مطرح شد. منابع داده‌ای اولیه، نحوه توسعه آن و همچنین ساختار فایل‌های پیکره به طور کامل شرح داده شد. با توجه به ویژگی‌های ذکر شده، پیکره *رسائل* و *مسائل* حاوی ۲۰۱۱۸ سؤال غیرحقیقت و ۲۰۵۱ سؤال حقیقت، نخستین پیکره در زبان فارسی است که برای کلیه مراحل سامانه‌های پرسش و پاسخ می‌تواند

مورد استفاده قرار گیرد. لازم به ذکر است که هیچ یک از پیکره‌های انگلیسی مطرح شده، به تنهایی برای کلیه مؤلفه‌های سامانه‌های پرسش و پاسخ نمی‌توانند مورد استفاده قرار گیرند؛ همچنین پیکره‌ای در زبان انگلیسی وجود ندارد که بتواند همزمان برای داده‌های حقیقت و غیرحقیقت مورد استفاده قرار گیرد؛ در حالی که پیکره پیشنهادی به گونه‌ای طراحی شده است که بتواند برای انواع سؤالات حقیقت و غیرحقیقت مورد استفاده قرار گیرد؛ از این رو، پیکره پیشنهادی قابل قیاس با پیکره‌های انگلیسی مطرح شده است.



(شکل-۶): ابزار نشانه‌گذاری پرسش و پاسخ رساله
(Figure-6): Treatise Question Answering Annotation Tool

خامنه‌ای را در راستای این پژوهش در اختیارمان قرار دادند، سپاس گزاریم.

در ادامه یک سامانه پرسش و پاسخ بر روی پیکره پیشنهادی معرفی شد که این سامانه توانست به دقت ۸۲/۲۹ و MRR برابر با ۵۶/۷۳ درصد دست یابد.

۹- مراجع

[۱] حسینی، پدرام و همکاران، "پیکره فارسی تحلیل احساس سنتی پرس: توسعه یک پیکره تحلیل احساس متنی برای زبان فارسی". سومین کنفرانس زبانشناسی رایانشی، ۱۳۹۳.

به عنوان فعالیت‌های آینده، با استفاده از روش‌های پرکاربرد یادگیری ماشین و بازیابی اطلاعات، سامانه پرسش و پاسخ معرفی شده را می‌توان بهبود و دامنه اطلاعات موجود در این پیکره را افزایش داد، به گونه‌ای که این پیکره، حاوی اسناد رساله تعداد مراجع تقلید بیشتری باشد. انواع سؤالات پیکره را طبق تعریف ۲-۳ در سطح ریزتر می‌توان ذخیره کرد. همچنین با استفاده از سؤال و جواب‌های موجود در سایت‌های مذهبی، تعداد سؤالات این پیکره را می‌توان گسترش داد.

[1] (Hosseini, P. et al., "Persian Sentiment Analysis Corpus: Developing a textual sentiment corpus for Persia". *Third conference on Computational Linguistics*, Sharif University of Technology, 2014.)

[۲] قائمی، هادی، کاهانی، محسن. "دسته‌بندی پرسش‌ها با استفاده از ترکیب دسته‌بندها". *پروازش علائم و داده‌ها*. ۱۳۹۵؛ ۱۳ (۳): ۹۹-۱۱۲

سپاس‌گزاری

از مرجع تحقیقات علوم اسلامی (نور)، که منبع داده خام رساله آیت‌الله العظمی مکارم و استفتائات آیت‌الله العظمی

- [16] Mollaei, A., Rahati-Quchani, S. and Estaji, A., "Question classification in Persian language based on conditional random fields", *2nd International eConference on Computer and Knowledge Engineering (ICCKE)*, 2012, pp.295–300.
- [17] Rasooli, M.S et al., "Development of a Persian syntactic dependency treebank". In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 306–314.
- [18] Smith, N.A., Heilman, M. and Hwa, R., "Question generation as a competitive undergraduate course project", In *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008, pp. 4–6.
- [19] Tellex, S. et al., "Quantitative evaluation of passage retrieval algorithms for question answering", In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 41–47.
- [20] Tom, D. et al., "TrainQA: a Training Corpus for Corpus-Based Question Answering Systems", In *Proc. 8th Int. Conf. on Computational Linguistics and Intelligent Text Processing*, 2007, pp. 1–7.
- [21] Voorhees, E.M., "Building a question answering test collection", *ACM SIGIR*, 2000.
- [22] Voorhees, E.M., "The TREC-8 Question Answering Track Report", In *TREC*, 1999, pp. 77–82.
- [23] Yaghoobzadeh, Y. et al., "ISO-TimeML Event Extraction in Persian Text", *COLING*, 2012, pp.2931-2944.
- [24] Zhang, D. and Lee, W.S., "Question classification using support vector machines", In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 26–32.
- [2] (Ghaemi, .H, kahani, .M. "Question Classification using nsemble Classifiers ", *JSDP*, 13 (3), 2016, 99-112)
- [3] AleAhmad, A. et al., "Hamshahri: A standard Persian text collection", *Knowledge-Based Systems*, 22(5), 2009, pp.382–387.
- [4] Bijankhan, M. et al., "Lessons from building a Persian written corpus: Peykare". *Language Resources and Evaluation*, 45(2), 2010, pp.143–164.
- [5] Greenwood, M.A., "Open-domain question answering", *Foundations and Trends in Information Retrieval*, 2005.
- [6] Gupta, P. and Gupta, V., "A survey of text question answering techniques", *International Journal of Computer Applications*, 53(4), 2012, pp.1–8.
- [7] Hirschman, L. and Gaizauskas, R., "Natural language question answering: the view from here", *Natural Language Engineering*, 7(04), 2000, pp.275–300.
- [8] Kolomiyets, O. and Moens, M.-F., "A survey on question answering technology from an information retrieval perspective", *Information Sciences*, 181(24), 2011, pp.5412–5434.
- [9] Lee, G. et al., "SiteQ: Engineering High Performance QA System Using", *Lexico-Semantic Pattern Matching and Shallow NLP*. In *TREC*, 2001.
- [10] Li, X. and Roth, D., "Learning question classifiers", In *Proceedings of the 19th international conference on Computational linguistics.*, 2002.
- [11] Li, X. and Roth, D., "Learning question classifiers: the role of semantic information", *Natural Language Engineering*, 12(03), 2006, pp.229–249.
- [12] Magnini, B. et al., "Creating the DISEQuA corpus: a test set for multilingual question answering", In *Comparative Evaluation of Multilingual Information Access Systems*, 2004, pp. 487–500.
- [13] Manning, C.D., Raghavan, P. and Schütze, H., "Introduction to information retrieval", *Cambridge university press Cambridge*, 2008.
- [14] Moghaddas, B.B. et al, "Pasokh: A standard corpus for the evaluation of Persian text summarizers". In *Computer and Knowledge Engineering (ICCKE)*, 2013, pp. 471–475.
- [15] Moll, D. and Vicedo, L., "Question Answering in Restricted Domains: An Overview", *Computational Linguistics*, 2007.



یاسمن برشبان، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرمافزار در سال ۱۳۹۲ از دانشگاه تربیت دبیر شهید رجایی تهران اخذ کرد. وی تحصیلات خود را در مقطع کارشناسی ارشد همان رشته در سال ۱۳۹۴ در دانشگاه گیلان به پایان رساند. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: پردازش

زبان طبیعی، بازیابی اطلاعات و یادگیری ماشین.
نشانی رایانامه ایشان عبارت است از:

boreshban@msc.guilan.ac.ir



حامد یوسفی نسب، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۹۴ از دانشگاه گیلان اخذ کرد. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: بازیابی اطلاعات و پردازش زبان طبیعی.

نشانی رایانامه ایشان عبارت است از:

hdyousefi@gmail.com



سید ابوالقاسم میرروشندل، مدرک کارشناسی خود را در رشته مهندسی کامپیوتر در سال ۱۳۸۴ از دانشگاه تهران و مدرک کارشناسی ارشد و دکترای خود را در همان رشته به ترتیب در سال‌های ۱۳۸۶

و ۱۳۹۱ از دانشگاه صنعتی شریف تهران دریافت کرد. ایشان از سال ۱۳۹۱ استادیار گروه مهندسی کامپیوتر دانشگاه گیلان هستند. از ایشان بیش از سی مقاله فنی در نشریات و همایش‌های معتبر به چاپ رسیده است. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: داده‌کاوی، یادگیری ماشین و پردازش زبان طبیعی.

نشانی رایانامه ایشان عبارت است از:

mirroshandel@guilan.ac.ir

Archive