A Probabilistic Model to Determine the Coherence of Texts in Interactive Question Answering Systems

Mohammad Mehdi Hosseini
PhD Candidate in Computer Engineering; Department of Computer Engineering; Shahrood University of Technology; Corresponding Author hosseini_mm@shahroodut.ac.ir

Morteza Zahedi
PhD; Computer Engineering; Assistant Professor; Department of Computer Engineering; Shahrood University of Technology; zahedi@ganjineh.co.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 33 | No. 4 | pp. 1751-1776
Summer 2018

Received: 31, Jul. 2017 **Accepted: 02, Dec. 2017**

Abstract: Evaluation plays an important role in interactive question answering systems like many computational linguistics fields. The coherence between the questions and the answers exchanged between the user and the system is one of the important criteria in evaluating these systems. In this paper, a new approach to determine the degree of coherence of generated text by the IQA systems is presented. The proposed model is a probabilistic model in which for feature extraction, the similarity between different N-grams is derived based on four defined criteria. Then using a prediction of the best density function among the 18 functions considered for each feature, a model for determining the coherence is selected. The results of implementation on two databases provided by several interactive question answering systems indicate that the proposed probabilistic model is highly adapted and its accuracy in determining the degree of coherence in the conversation text has been made. The Kolmogorov-Smirnov, Anderson, Darling and Cramer van Meys trials were used to matching or non-matching probability density function. According to the presented results, the probability density factor with the least error was the best performance in determining the coherence of each conversation.

Keywords: Matematical Modeling, Coherence of Text, Intearctive Question Answering Systems (IQAs), N-gram, Statistical Similarity

ارائهٔ یک مدل احتمالاتی جهت تعیین انسجام متن در سیستم‌های پرسش و پاسخ تعاملی

محمدمهدی حسینی
دانشجوی دکتری کامپیوتر- هوش مصنوعی؛ دانشکدهٔ کامپیوتر؛ دانشگاه صنعتی شاهرود؛ شاهرود؛ ایران؛
پدیدآور رابط hosseini_mm@yahoo.com

مرتضی زاهدی
دکتری تخصصی کامپیوتر؛ استادیار؛ دانشکدهٔ کامپیوتر؛ دانشگاه صنعتی شاهرود؛ شاهرود؛ ایران؛
zahedi@ganjineh.co.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۶/۰۵/۰۹ پذیرش: ۱۳۹۶/۰۹/۱۱ مقاله برای اصلاح به مدت ۱۵ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۳ | شمارهٔ ٤ | ص‌ص ۱۷۳۹-۱۷٦٤
تابستان ۱۳۹۷

چکیده: همانند بسیاری از زمینه‌های زبان‌شناسی محاسباتی، ارزیابی نقش مهمی در سیستم‌های پرسش و پاسخ تعاملی ایفا می‌کند. یکی از معیارهایی که در زمینهٔ ارزیابی این سیستم‌ها دارای اهمیت است، میزان انسجام بین سؤال‌ها و پاسخ‌های ردوبدل شده بین کاربر و سیستم است. در این مقاله، یک راه حل اتوماتیک برای تعیین میزان انسجام متن تولیدشده ارائه شده است. مدل پیشنهادی، یک مدل احتمالاتی است که در آن برای استخراج ویژگی از میزان شباهت بین N-گرم‌های مختلف بر اساس چهار معیار تعریف‌شده بهره گرفته شده است. سپس، با استفاده از تخمین بهترین تابع چگالی از بین ۱۸ تابع در نظر گرفته‌شده برای هر ویژگی، یک مدل برای تعیین میزان انسجام انتخاب گردیده است. نتایج پیاده‌سازی بر روی دو پایگاه دادهٔ تهیه‌شده از چند سیستم پرسش و پاسخ تعاملی، حاکی از انطباق بسیار بالای مدل احتمالاتی پیشنهادی و دقت مناسب آن در تعیین میزان انسجام در متن مکالمهٔ صورت‌گرفته است. برای تطبیق یا عدم تطبیق تابع چگالی احتمال به‌دست‌آمده از آزمون‌های سه‌گانهٔ «کولموگروف-اسمیرنف»، «اندرسون دارلینگ» و «کرامر وان‌میس» استفاده شد. با توجه به نتایج ارائه‌شده، تابع چگالی احتمال «ناکاگامی» با داشتن کمترین اشتباه، بهترین عملکرد را در تعیین میزان انسجام هر مکالمه از خود نشان داد.

کلیدواژه‌ها: مدل‌سازی ریاضی، انسجام متن، سیستم پرسش و پاسخ تعاملی، N-گرم، شباهت آماری

۱. مقدمه

افزایش حجم اطلاعات، توسعهٔ سیستم‌های رایانه‌ای و گسترش استفاده از فناوری اطلاعات و کنترل و مدیریت آن را مشکل‌تر کرده است. بنابراین، تولید اطلاعات به تنهایی کافی نیست، بلکه باید ابزارهایی برای استفاده از آن فراهم شود. یکی از این ابزارها استفاده از سیستم‌های پرسش و پاسخ (QA)[1] متنی است. سیستم QA به‌عنوان سیستمی با پتانسیل بالا شناخته می‌شود که کاربران را قادر می‌سازد به منابع علمی به‌صورت زبان طبیعی (از طریق پرسش) دسترسی داشته باشند و یک پاسخ مرتبط، مناسب و مختصر دریافت کنند. بنابراین، انتظار می‌رود این سیستم‌ها در سال‌های آتی پیشرفت چشمگیری داشته باشند. با این حال، مشکلات چالش برانگیز فراوانی جهت مرتفع‌نمودن در این سیستم‌ها همچنان وجود دارد. یکی از وظایف مهم برای سیستم‌های QA موجود، درک صحیح سؤالات زبان طبیعی و استنتاج معنای دقیق، جهت ارائهٔ پاسخ‌های صحیح است. بهبود درک ماشین از سؤالاتی که با مشکلاتی نظیر طبقه‌بندی سؤال، فرمول‌بندی صحیح پرس‌وجوها، ابهام در تجزیه، تشخیص تقارن معنایی، تشخیص روابط ظاهری در سؤالات پیچیده، تشخیص یک جواب مناسب برای کاربر و مکانیزم اعتبارسنجی مناسب مواجه است، هنوز از چالش‌های موجود در این زمینه است (Dwivedi and Singh 2013).

از دیگر معضلات سیستم‌های QA، فقدان تعامل دو طرفه بین سیستم و کاربر است. این معضل از آنجا نشأت می‌گیرد که در سیستم‌های QA در صورتی که پرسشِ کاربر دارای ابهام باشد و یا اگر کاربر نیاز به اطلاعات بیشتری در مورد پاسخ دریافت‌شده داشته باشد، سیستم برای این مورد راهکاری ارائه ننموده است. بنابراین، در سیستم‌های پرسش و پاسخ تعاملی (IQA)[2] با افزودن سطح تعامل به رفع این معضل پرداخته شد تا در صورت وجود ابهام در سؤال یا درخواست اطلاعات بیشتر، سیستم جهت رفع ابهام و درج بهتر پرسش مکالمه‌ای با کاربر آغاز نماید (Amit and Jain 2016). بنابراین، می‌توان این طور تصور نمود که در IQA یک فرایند تکراری اتفاق می‌افتد. این نکته قابل ذکر است که سیستم‌های پرسش و پاسخ متقابل می‌توانند به‌عنوان یک سیستم دیالوگ در نظر گرفته شوند، اما برای در نظر گرفتن این دسته از سیستم‌ها به‌عنوان یک سیستم IQA،

1. Question Answering System (QA)
2. Interactive Question Answering System (IQA)

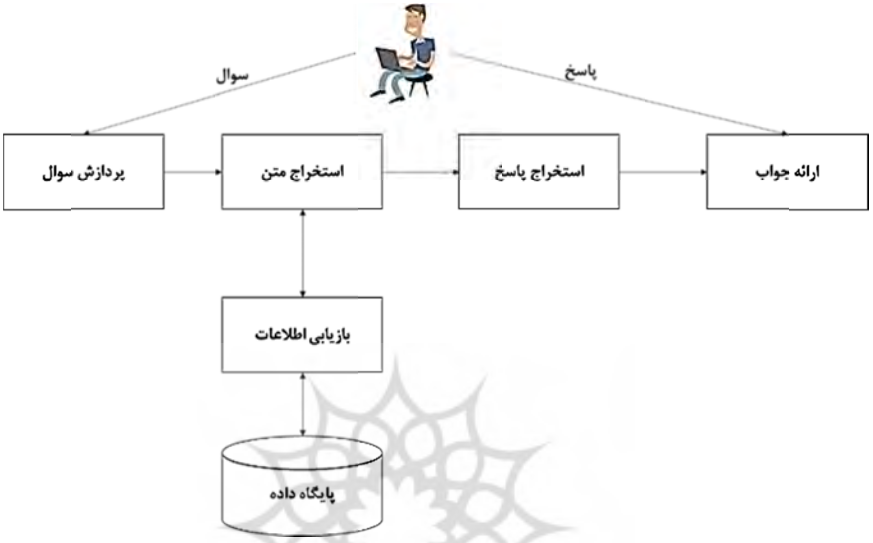
تنها آن دسته از سیستم‌های دیالوگی که در جست‌وجوی اطلاعات به‌صورت متنی هستند می‌توانند به‌عنوان سیستم پرسش و پاسخ متقابل در نظر گرفته شوند.

ارزیابی در سیستم‌های IQA همانند بسیاری دیگر از زمینه‌های مربوط به آن نقش مهمی را در ارتقای این سیستم‌ها ایفا می‌نماید. علی‌رغم این موضوع، در زمینهٔ ارزیابی سیستم‌های IQA تقریباً می‌توان گفت که روش خاصی وجود ندارد که به ارزیابی کلی این سیستم‌ها بپردازد و تنها روش‌هایی وجود دارند که با روش ارزیابی به‌کار برده‌شده در پرسش و پاسخ و سیستم‌های دیالوگ انطباق دارند (Bouziane, Bouchiha and Doumi 2015). مشکل اصلی طراحی یک روش اتوماتیک ارزیابی مربوط به سیستم‌های IQA ناشی از این است که به‌ندرت امکان پیشگویی بخش تعامل فراهم می‌شود. به همین منظور، باید انسان در فرایند ارزیابی آن مداخله داشته باشد (Forner et al. 2010). تحقیقات گذشته نشان داده که ارزیابی انسان بر اساس پارامترهای مختلفی صورت می‌پذیرد. بنابراین، برای ایجاد یک مدل اتوماتیک به اندازه‌گیری اتوماتیک این پارامترها نیاز است. خود این امر یکی از چالش‌های محققان این حوزه است. با توجه به مطالعات انجام‌شده در حوزهٔ ارزیابی سیستم‌های IQA، یکی از ویژگی‌هایی که در فرایند ارزیابی خروجی این سیستم‌ها مورد توجه قرار گرفته، انسجام بین سؤالات و پاسخ‌های ردوبدل‌شده بین کاربر و سیستم است که اندازه‌گیری اتوماتیک آن از اهمیت خاصی برخوردار است. بنابراین، در این مقاله قصد بر این است که یک روش آماری جهت تعیین اتوماتیک میزان انسجام در متن مکالمهٔ متنی صورت‌گرفته ارائه نماییم. ساختار مقالهٔ پیشنهادی بدین‌صورت است که در بخش دوم، مرور کوتاهی بر سیستم‌های QA، IQA و انسجام متن خواهیم داشت. تحقیقات صورت‌گرفته در این حوزه در بخش سوم آورده شده است. در بخش چهارم، سیستم مورد ارزیابی قرار گرفته، پایگاه داده، روش پیشنهادی، و نتایج به‌دست‌آمده از مدل ارائه‌شده آورده شده است و در بخش آخر، نتیجه‌گیری بیان شده است.

۲. سیستم QA

سیستم‌های QA شکل پیچیده‌تری از سیستم‌های بازیابی اطلاعات هستند که در این سیستم‌ها به جای ارائهٔ کل سند، تنها بخش‌های خاصی از اطلاعات سند به‌عنوان پاسخ بازگردانده می‌شود که ممکن است این پاسخ یک کلمه، یک جمله یا یک پاراگراف باشد. در این سیستم‌ها معمولاً از دو رویکرد بازیابی داده و بازیابی اطلاعات

جهت پاسخ‌دهی به پرسش‌ها استفاده می‌شود. بنابراین، در سیستم‌های QA ابتدا با کمک روش‌های بازیابی ساده، اسناد مرتبط را بازیابی نموده سپس، در میان حجم کاهش‌یافتهٔ اسناد، به‌دنبال پاسخ می‌گردند.



شکل ۱. ساختار کلی یک سیستم QA

ساختار کلی فرایند یک سیستم QA در شکل ۱، نشان داده شده است. مطابق با این شکل، یک سیستم QA شامل سه مرحلهٔ اساسی است (Amit and Jain 2016):

۱) پردازش پرسش: در این مرحله پرسش مطرح‌شده از سوی کاربر تجزیه و تحلیل شده و فرمت پاسخ مورد نظر از آن استنتاج می‌گردد.

۲) تحلیل متن یا بازیابی اطلاعات: در این مرحله استخراج متون مناسب و تشخیص پاسخ بر اساس بازیابی اطلاعات از پایگاه دادهٔ موجود برای سیستم صورت می‌پذیرد. در این مرحله، پس از ورود مخزنی از اسناد و اطلاعات متنی به همراه پرسش کاربر به این مرحله، تمامی اسناد از طریق یک تابع و بر اساس میزان ارتباط آن‌ها با پرسش کاربر، امتیازدهی و مرتب می‌شوند.

۳) تحلیل یا پردازش پاسخ: کارکرد کلی این مرحله بدین‌صورت است که قطعه‌های متنی بازگردانده شده از مرحلهٔ قبل پردازش شده و عباراتی که احتمال می‌رود دارای

پاسخ دقیق باشند، استخراج می‌شود. سپس، از بین عبارات به‌دست‌آمده عبارتی که بیشترین احتمال داشتن پاسخ را دارد، انتخاب شده و به‌عنوان پاسخ نهایی سیستم به کاربر ارائه می‌شود. معمولاً در این مرحله، سیستم‌ها جهت استخراج پاسخ مناسب، نیازمند به‌کارگیری دادگان آموزشی و همچنین، استفاده از تکنیک‌های یادگیری ماشین هستند.

۲-۱. سیستم IQA

گفت‌وگوی متقابل بین دو یا چند موجودیت را تعامل می‌گویند. در تعامل، گوینده مطلبی را بیان می‌کند و مطلب توسط شنونده تفسیر شده و تصمیم می‌گیرد که چه پاسخی ارائه نماید و این چرخهٔ گوش دادن، درک کردن، تصمیم‌گیری و پاسخ تکرار می‌شود. هرچند در مکالمهٔ بین انسان‌ها محدودیتی وجود ندارد، ولی در پیاده‌سازی سیستم QA مبتنی بر تعامل، معمولاً محدودیت‌هایی (مانند این‌که دقیقاً دو شرکت‌کننده با یکدیگر صحبت کنند، هر یک از شرکت‌کنندگان به نوبت صحبت کند و یا این‌که شرکت‌کنندگان در زمینهٔ مشترکی با یکدیگر صحبت کنند) در نظر گرفته می‌شود (Kolomiyets and Moens 2011). افزودن سطح تعامل به سیستم‌های QA با چند هدف صورت می‌گیرد. برای مثال، کاربر می‌تواند بجای طرح مجموعه‌ای از چندین سؤال مرتبط در قالب یک پرسش پیچیده، آن را به سؤالات ساده‌تر و کوتاه‌تر تجزیه نماید که این امر، مکالمهٔ طبیعی‌تر و پاسخ دقیق‌تری را در پی خواهد داشت. یا اگر پاسخ سیستم دلخواهِ کاربر نباشد و یا کاربر نیاز به اطلاعات بیشتری داشته باشد، مکالمه‌ای را با سیستم آغاز نماید تا پاسخ دلخواه خود را دریافت کند. همچنین، در این سیستم‌ها امکان طرح مجموعه‌ای از سؤالات مرتبط با هم از سوی کاربر وجود دارد. به‌عنوان مثال، کاربر می‌تواند پرسشی را مطرح کند که از طریق ضمایر به پرسش قبل ارجاع داده شده است و یا این‌که سیستم می‌تواند پس از ارائهٔ پاسخ به یک پرسش، مجموعه‌ای از سؤالات مرتبط با آن موضوع را به کاربر پیشنهاد نماید که انسجام گفت‌وگو و افزایش کارایی را در پی داشته باشد.

۲-۲. انسجام در متن

انسجام و پیوستگی متون تولید شده در سامانه‌های مختلف متنی همواره یکی از مهم‌ترین دغدغه در این سامانه‌ها بوده است. یک متن غیرمنسجم می‌تواند خروجی

سیستم‌های تولیدکننده متن مانند خلاصه‌سازی، ساده‌سازی، امتیازدهی خودکار مقالات، متن تولیدشده توسط یک سیستم پرسش و پاسخ و یا حتی متن تولیدشده توسط یک فرد اما با دانش نگارشی پایین باشد (Zhang and Feng 2015). انسجام و پیوستگی موضوعی متن خروجی در تمامی سیستم‌های پردازشی متنی یکی از مهم‌ترین تحقیقات در این حوزه است. بنابراین در این راستا رویکردهای متفاوتی شکل گرفته است. رویکردهای تعیین و ارزیابی انسجام توسط محققان این زمینه به سه دسته تقسیم شده است (Barzially and lapata 2005). دسته اول رویکردهایی هستند که ساختار هدفمند نامیده می‌شوند. در این رویکرد بیشتر بدنهٔ اصلی و هدف نهایی متن مورد توجه قرار دارد. دستهٔ دوم از این رویکردها به بخش‌بندی سخن‌ها می‌پردازند و بیشتر روش‌های ارائه‌شده در این دسته به انسجام ساختاری درون‌جمله‌ای می‌پردازند و گاهی این روش‌ها را مدل‌های ساختار زبانی نیز می‌نامند. دستهٔ سوم، ساختار تمرکزی نامیده می‌شوند که بیشتر روش‌های این دسته تمرکز موضوعی بر روی موضوع اصلی متن را ارزیابی می‌کنند. تا به حال، رویکردهای مبتنی بر ساختار تمرکزی بیشترین توجه را در مطالعات مربوط به انسجام متن به خود اختصاص داده و مدل‌های تئوری مرکزیت مهم‌ترین رویکردهای مورد مطالعه در این حوزه هستند. در این مدل‌ها فرض بر این است که توجه اصلی خواننده بر روی موضوع اصلی در متن بوده و این تمرکز تا انتهای متن بر روی نشانه‌های موضوع محوری حفظ می‌شود. مدل شبکه‌نهاد[1] مهم‌ترین و گسترده‌ترین رویکرد مبتنی بر تئوری مرکزیت است که تا به حال مورد استفاده قرار گرفته است. طبق ایدهٔ اصلی این نظریه می‌توان انسجام یک متن را توسط دنباله‌ای از نهاد گفتمان‌های تکراری مانند فاعل و مفعول اندازه‌گیری کرد. ارزیابی انسجام یک متن در این روش‌ها، معمولاً به دو صورت انسجام محلی و انسجام عمومی انجام می‌شود. تشخیص یا ارزیابی انسجام به‌صورت محلی به‌راحتی بر روی بخش‌های کوچکی از متن و با استفاده از رویکردهای محاسباتی قابل انجام است. تا به امروز، تحقیقات زیادی بر روی انسجام محلی انجام شده و مقالات زیادی بر روی آن تأکید داشته‌اند، اما تحقیقات در حوزهٔ انسجام عمومی کمتر مورد توجه قرار گرفته است.

1. Entity Based Model

۳. پیشینهٔ پژوهش

امروزه، محققان برای پاسخ به سؤال در سیستم‌های QA و IQA، از تکنیک‌هایی مانند هوش مصنوعی، پردازش زبان طبیعی، تحلیل استاتیک، مطابقت الگو، بازیابی اطلاعات و در نهایت، استخراج اطلاعات و ترکیب آن‌ها با یکدیگر بهره می‌جویند. بر این اساس، سه دستهٔ کلی از این روش‌ها شکل گرفته است که این سه دسته را رویکرد زبان‌شناختی، رویکرد آماری و رویکرد تطابق الگو تشکیل می‌دهند. سیستم‌های IQA از سیستم‌های QA دقیق‌تر هستند. این موضوع از این حقیقت نشأت می‌گیرد که سیستم‌های IQA به ساختمان زبان‌های مبهم مرتبط هستند. به‌علاوه، زمانی که سیستم‌های IQA با ساختار مبهم مواجه می‌شوند، برای آشکار و مشخص‌شدن درخواست، دیالوگ آغاز می‌شود. سیستم‌های موجود در زمینه IQA، می‌توانند با توجه به شرایط و کاربردهایشان در یکی از سه گروه، IQA به‌عنوان یک مدیریت محدودیت، QA ارتقاء یافته و سؤالات متوالی قرار گیرند.

کارهای صورت‌گرفتهٔ قابل توجهی در ارزیابی سیستم‌های QA در زمینهٔ استفاده از کاربران واقعی وجود دارد. به‌طور کلی، پاسخ صحیح در یک سیستم IQA و ارزیابی جواب‌های ممکن با روش‌های متفاوتی بیان می‌شود. اکثر سیستم‌های موجود در ارزیابی، از ارزیابی انسان بهره می‌گیرند. مسائل طراحی کلی مربوط به این سیستم‌ها توسط (Tague-Sutcliffe 2001) ارائه شده است. بنابراین، ارزیابی سیستم‌های QA بسته به ارزیابی سؤالات پیچیده یا ساده (مثل تعریف، روابط و سناریوهای مربوط به سؤالات) متفاوت است. یکی از روش‌های ارزیابی مورد استفاده در سیستم‌های QA استفاده از مجموعه‌ای از سؤالات و پاسخ به نام «مجموعهٔ استاندارد طلایی» است. در این روش با استفاده از میزان منطبق‌بودن سیستم با این مجموعهٔ استاندارد طلایی توانایی یک سیستم سنجیده می‌شود. البته، این روش برای سؤالات پیچیده و مبهم هنوز تقویت نشده است. بیشتر ارزیابی سیستم‌های QA توسط TREC[1] انجام شده است که این ارزیابی‌ها بیشتر به جای این‌که مبتنی بر یک سیستم باشد، بر اساس کاربر صورت پذیرفته است. در حقیقت بیشتر کارهای صورت‌گرفتهٔ ارزیابی در زمینهٔ استخراج پاسخ و نحوهٔ تعامل و استفاده از آن انجام شده است. با برگزاری کنفرانس TREC، سالانهٔ دورهٔ جدیدی در طراحی و

1. Text Retrieval Evaluation Conference

ارزیابی سیستم‌های پرسش و پاسخ آغاز گردید. کنفرانس TREC با اعلام رویکرد جدید خود، طراحان سیستم‌های پرسش و پاسخ را به طراحی سیستم‌هایی تشویق نمود که با بهره‌گیری از مجموعهٔ بزرگی از اسناد متنی در موضوعات مختلف بتوانند پاسخی کوتاه برای سؤال کاربر که دربارهٔ موضوعات مختلف با زبان طبیعی مطرح شده است، ارائه نمایند که بعدها این‌گونه سیستم‌ها را سیستم‌های با دامنهٔ نامحدود[1] و گاهی سیستم‌های پویا نامیدند. به‌طور کلی، در سیستم‌های QA ارائه‌شده از یک ارزیاب انسانی بهره گرفته می‌شد که این ارزیاب‌ها باید قبل از ارزیابی سیستم، آموزش می‌دیدند.

ارزیابی سیستم‌های IQA به‌منظور تعیین و ارتقای کارایی آن‌ها از اهمیت زیادی برخوردار است. با وجود این، هنوز روش استاندارد و مخصوصی برای ارزیابی این سیستم‌ها ارائه نشده است و بیشتر روش‌های انجام‌شده برای کنترل صحت و درستی جواب‌های برگشت داده‌شده به کاربران در حیطهٔ پاسخ‌دهی این سیستم‌ها صورت پذیرفته است. با وجود این، روش‌های مطرح‌شده نمی‌توانند اطلاعات کافی دربارهٔ کیفیت سیستم فراهم نمایند. به همین دلیل، بیشتر روش‌های ارزیابی به‌کاررفته در این سیستم‌ها، در سیستم‌های مکالمه‌محور مورد استفاده قرار گرفته است. ارزیابی‌های انجام‌شده برای سیستم‌های IQA معمولاً به دو صورت ارزیابی کیفی و کمی صورت می‌پذیرد. در ارزیابی کمی، ارزیابی کارایی سیستم‌های پرسش و پاسخ از نظر صحت و دقت پاسخ‌گویی آن‌ها با توجه به نوع سیستم اندکی متفاوت است و روش‌های متفاوتی در این نوع ارزیابی مطرح گردیده است. هدف از ارزیابی کمی سیستم‌های پرسش و پاسخ تعاملی، تعیین میزان صحت پاسخ بازگردانده‌شده توسط این سیستم‌هاست. این ارزیابی، اطلاعات کافی دربارهٔ کیفیت تعامل سیستم با کاربر و این‌که آیا تعامل موفقیت‌آمیز خاتمه یافته است یا خیر، ارائه نمی‌کند. به همین دلیل، علاوه بر ارزیابی کمی، این سیستم‌ها از نظر کیفی نیز مورد سنجش و ارزیابی قرار می‌گیرند تا کیفیت تعامل سیستم با کاربر و میزان رضایت‌مندی کاربر تعیین شود. جهت ارزیابی کیفیت تعامل، معمولاً پرسشنامه‌ای تهیه شده و در اختیار کاربران قرار داده می‌شود تا با تکمیل آن میزان رضایت آن‌ها از سیستم سنجیده شود. اما همچنان روش جامعی که به ارزیابی کلی سیستم‌های پرسش و پاسخ تعاملی بپردازد، مطرح نگردیده است. تحقیقات متعددی در زمینهٔ ارزیابی سیستم‌های IQA

1. Open Domain Question Answering systems (ODQA)

صورت پذیرفته است که در این باره می‌توان به موارد زیر اشاره نمود.

«کلی» و همکارانش در مقالهٔ خود به ارزیابی عملکرد چهار سیستم IQA با کاربر واقعی پرداخته‌اند. هدف آن‌ها از این کار، شناسایی پتانسیل معیارهای ارزیابی برای سیستم‌های IQA با تجزیه و تحلیل نظرات ارزیابی ایجادشده توسط کاربران برای چنین سیستم‌هایی بود. آن‌ها در کار خود از داده‌های کیفی که ارزیاب‌ها در طول مصاحبه‌ها با کاربران برای شناسایی موضوعات مشترک مربوط به عملکرد، استفاده، و قابلیت استفادهٔ مجدد سیستم جمع‌آوری نموده بودند، بهره بردند (Kelly et al. 2009). «سان، کانتور و مورس» روشی را برای ارزیابی سیستم‌های IQA معرفی نمودند که X-EVAL نامیده می‌شد. مدلی که آن‌ها معرفی کردند شامل دو مرحله بود. در مرحلهٔ اول، به‌کارگیری سیستم‌های آزمایشگاهی و در مرحلهٔ دوم، تست مدل ارزیابی بر روی این سیستم‌ها صورت پذیرفت. آن‌ها برای انجام آزمایشات خود سه سیستم توسعه‌یافته در ARDA را که برای برنامهٔ AQUAINT بود، مورد مطالعه قرار دادند. ارزیابی‌ها بر روی همهٔ سیستم‌ها توسط تحلیگران انسانی صورت پذیرفت و از روش X-EVAl به‌عنوان روش سنجش پایه استفاده گردید. نتایج به‌دست‌آمده از آزمایشات نشان داد که ارزیابی از یک سیستم IQA به‌دلیل این‌که شامل دو موجودیت سیستم و فرد است، و عوامل بسیاری وجود دارد که کاربر باید تحت تأثیر آن‌ها با سیستم کار کند، کار بسیار سخت و پیچیده‌ای است (Sun, Kantor and Morse 2011). «فورنر» و همکارانش یک مرور کلی از مسائل مهم مطرح‌شده طی هفت سال بر روی یک سیستم دو زبانهٔ QA انجام دادند (Forner et al. 2010). «شاه و پومرانتز» روشی را برای ارزیابی و پیش‌بینی کیفیت پاسخ در سیستم‌های پرسش و پاسخ اجتماعی[1] ارائه نمودند که مورد توجه بسیاری از محققان این حوزه قرار گرفت. در این مقاله، تفسیر در مورد کیفیت، بر اساس مجموعهٔ دادهٔ موجود ارائه شد. همچنین، سعی بر این بود که یک پیش‌بینی دربارهٔ این‌که آیا پاسخ انتخاب‌شده توسط کاربر می‌تواند به‌عنوان بهترین پاسخ باشد یا خیر، صورت پذیرد (Shah and Pomerantz 2010).

«واچلدر» و همکارانش گزارشی از توسعهٔ عناصر طرح ارزیابی برای طراحی کلی ارزیابی و قابلیت استفاده از سیستم HITIQA (که یک سیستم تعاملی پرسش و پاسخ برای تهیهٔ گزارش گسترده در مسائل پیچیده است) ارائه نمودند. در این گزارش دو هدف اساسی

1. Community Question Answering

پیگیری شد. هدف اول، یک ارزیابی واقع‌بینانه از سودمندی و قابلیت استفاده از HITIQA به‌عنوان یک سیستم پایان‌به‌پایان[1] با استفاده از سؤالات اولیهٔ جست‌وجوگر اطلاعات برای تکمیل یک پیش‌نویس گزارش بود. و دومین هدف، توسعهٔ معیارهای مقایسهٔ پاسخ به‌دست‌آمده توسط تحلیلگران مختلف و ارزیابی کیفیت پشتیبانی فراهم‌شده توسط HITIQA بود. آن‌ها از ابزار کمی و کیفی برای به‌دست آوردن اطلاعات در مورد راحتی تحلیلگر با سیستم HITIQA استفاده کردند؛ به‌خصوص از ویژگی‌های جدید در سنجش توانایی یافتن پاسخ به سؤالات پیچیده و گفت‌وگوی تعاملی بهره گرفتند. به‌دلیل این‌که کیفیت خروجی اندازه‌گیری‌شدهٔ سیستم HITIQA با معیارهای استاندارد دقت و بازیابی سنجیده می‌شد، آن‌ها یک کار جدید (ارزیابی تقاطعی)[2] برای اندازه‌گیری غیرمستقیم کیفیت پاسخ به‌دست‌آمده با استفاده از سیستم HITIQA را توسعه دادند و و توانستند سیستم رأی دادن تحلیلگران به کیفیت گزارش خود و همکارانشان در ارزیابی سیستم را طراحی نمایند(Wacholder et al. 2004). اگرچه روش‌های استانداردی وجود دارد که می‌توان اطلاعات مربوط به عملکرد سیستم از قبیل زمان، دقت و یا بازیابی را با استفاده از آن‌ها به‌دست آورد، اما هنوز به شناسایی سهم سیستم و کاربران در عملکرد مطلوب یک سیستم نیاز هست و تنها نتیجه‌گیری که می‌توان متصور شد، این است که عملکرد یک سیستم از کاربری به کاربر دیگر با ارزش‌تر خواهد بود. به‌عبارت دیگر، روش ارزیابی باید قادر به بیان میزان سهم کاربر یا سیستم در موفقیت و یا شکست تولید خروجی مطلوب باشد.

تاکنون کارهای متعددی در زمینهٔ انسجام متن برای سیستم‌های پردازش متن صورت گرفته است. بررسی تحقیقات ارائه‌شده در این زمینه نشان می‌دهد که محققان اندکی به‌دنبال ارائهٔ یک روش اتوماتیک جهت ارزیابی انسجام متن تولیدشده در سیستم‌های IQA بوده‌اند و بیشتر کارهای صورت‌پذیرفته در زمینهٔ خلاصه‌سازی یا متن‌های تولیدشده به‌صورت اتوماتیک است. یکی از مهم‌ترین رویکردهای پیشنهادشده برای ارزیابی میزان انسجام محلی مدل مبتنی بر نهاد است. این مدل با توجه به تحلیل تغییر و تحولات برخی از عوامل انسجامی در جملات مجاور، الگوهایی را برای سنجش میزان انسجام و پیوستگی بین آنان استخراج می‌کند. ایدهٔ رویکردهای مبتنی بر نهاد این است که

1. End-to-End System
2. cross evaluation

تغییرات ایجادشده بین عوامل انسجامی جملات مجاور در یک متن منسجم دارای الگوهای منظم و باقاعده‌ای هستند. در این روش هر متن با یک ماتریس دو بعدی که به آن شبکه‌نهاد نیز گفته می‌شود، نمایش داده می‌شود. در ماتریس ایجادشده سطرها نشان‌دهندهٔ جملات و ستون‌ها نشان‌دهندهٔ عوامل انسجامی موجود در جمله بوده و سطرهای متوالی مشخص‌کنندهٔ جملات متوالی خواهند بود. به ازای هر عامل انسجامی در جمله، درایه‌ای در ماتریس مربوط وجود دارد که حاوی اطلاعاتی مانند وجود یا عدم وجود و نقش آن در جمله خواهد بود. در این روش، در صورت تکرار یک عامل در یک جمله با بیش از یک نقش گرامری، نقش گرامری دارای رتبهٔ بالاتر انتخاب می‌شود. فرضیهٔ اساسی در این رویکرد مبتنی بر شکل توزیع و توپولوژی عوامل انسجامی در ماتریس تولیدی ایجاد شده است. بنابراین، متن‌های منسجم دارای ستون‌هایی با چگالی بالا بوده و یا این‌که ستون‌هایی که عوامل موجود در آن‌ها بیشتر فاعل یا مفعول هستند، بیشتر موجب ایجاد انسجام در متن می‌شوند. لازم به ذکر است که این رویکرد فقط قادر به تشخیص انسجام محلی است (Barzilay and Lapata 2008).

«گینودو و استروب» روشی مبتنی بر ترکیب تئوری گراف و مدل مبتنی بر نهاد ارائه دادند. در روش پیشنهادشده تعاملات بین جملات و عوامل انسجامی موجود در آن‌ها به یک گراف دو قسمتی مدل می‌شوند. بخش اول، گراف دوقسمتی ایجادشده حاوی جملات و بخش دوم، حاوی عوامل انسجامی استخراج‌شده از جملات است. هر نود[1] جمله با یک نود عامل ارتباط برقرار می‌کند اگر و فقط اگر در جملهٔ عامل مورد نظر وجود داشته باشد. گراف مورد نظر جهت‌دار بوده و همیشه جهت آن از بخش جملات به بخش عوامل انسجامی است. وزن هر ارتباط با توجه به نقش گرامری عامل مورد نظر در جمله تعیین می‌شود (Guinaudeau and Strube 2013).

یکی دیگر کارهای انجام‌گرفته در زمینهٔ انسجام متن، استفاده از رویکردهای مبتنی بر خوشه‌بندی است. در این رویکردها عقیده بر این بود که با انتخاب جملات مرتبط به‌هم و قرار دادن آنان درون خوشه‌هایی جداگانه، بزرگ‌ترین خوشه به‌عنوان موضوع غالب متن تشخیص داده خواهد شد. بنابراین، با قرار دادن جملات موجود در آن خوشه در متن خلاصه مرتبط‌ترین جملات انتخاب و منسجم‌ترین خلاصه تولید خواهد شد. بعدها از

1. Node

خوشه‌بندی در ترکیب با سایر الگوریتم‌ها برای ایجاد خلاصه‌های منسجم استفاده شد. یکی از مهم‌ترین این ترکیب‌ها استفادهٔ همزمان خوشه‌بندی و ماشین‌های بردار پشتیبان بود (Shivakumar and Soumya 2015).

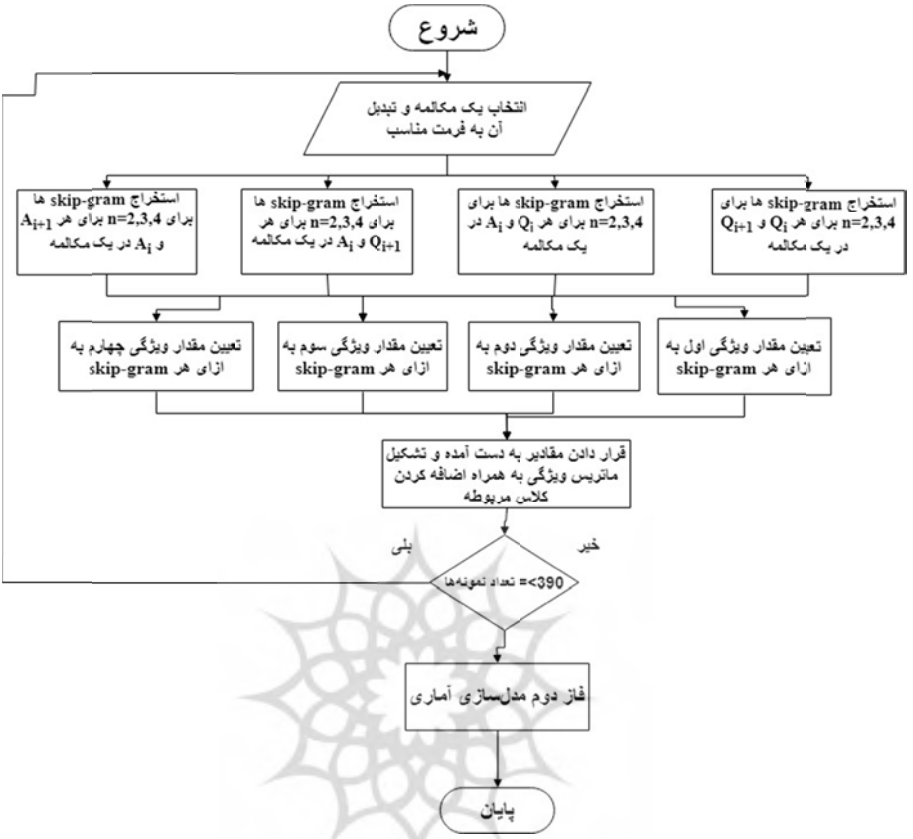
«جِین و ساخین» از رویکردهای مبتنی بر زنجیره‌های لغوی در مقالهٔ خود استفاده کردند. در این روش ایده بر آن است که ابتدا کلمات مرتبط به‌هم استخراج و در یک زنجیرهٔ لغوی قرار گیرند. سپس، با استفاده از این زنجیره، مرتبط‌ترین جملات انتخاب و در متن خلاصه قرار گیرند. در نتیجه، خلاصهٔ ایجادشده شامل مرتبط‌ترین و منسجم‌ترین جملات خواهد بود (Jain and Sachin 2016).

رویکردهای دیگری نیز از گراف‌های G-Flow برای تولید خلاصه‌های منسجم استفاده کرده‌اند. «جانرا» و همکارانش روشی برای تولید خلاصه‌هایی منسجم در خلاصه‌سازی چندسندی معرفی کردند. این روش موجب شد که انسجام خلاصهٔ تولیدشده در هر دو حوزهٔ انتخاب جملات مبهم و مرتبط و همچنین، چینش صحیح جملات به‌دنبال هم رعایت شود. یکی از بزرگ‌ترین چالش‌ها در خلاصه‌های چندسندی، قرار دادن جملات انتخاب‌شده در مکانی صحیح در متن خروجی است. گراف G-Flow انسجام جملات انتخابی را با وزنی که به یال‌های هر گراف داده است، تخمین زده و در نهایت، جملات با ارتباطی قوی‌تر را انتخاب و در متن خلاصه قرار می‌دهد (Pooja and Sachin 2016). یکی از حوزه‌های تولید و ارزیابی خلاصه‌های منسجم دیدگاه‌های شناختی انسان است. در روش پیشنهادشده توسط «ژان و فنگ» با به‌کارگیری مدلی شناختی جملات مهم‌تر متن تشخیص داده شده و خلاصه‌های منسجم‌تری تولید گردیده است. در این روش با استفاده از مجموعهٔ متن‌های روایی و تئوری شناختی، ارتباطات بین جملات کشف و مهم‌ترین آنان استخراج گردیده است (Zhang and Feng 2015). بیشتر کارهای انجام‌شده در حوزهٔ انسجام متن، در زمینهٔ انسجام متون خلاصه‌سازی شده و تولید متون اتوماتیک صورت گرفته، ولی در زمینهٔ سیستم‌های IQA چنین کاری صورت نپذیرفته است و بیشتر ارزیابی‌ها به‌صورت کیفی توسط کاربران انجام گرفته است. بنابراین، در این مقاله یک روش اتوماتیک بر اساس n-گرم‌ها پیشنهاد شده است تا بتواند جایگزین روش‌های کیفی در ارزیابی‌ها گردد. نوآوری روش پیشنهادی در این است که با استفاده از یک مدل احتمالاتی بر اساس شباهت بین سؤالات و پاسخ‌ها، میزان انسجام را محاسبه می‌نماید. با توجه به معیارهای تعریف‌شده، هم انسجام محلی و هم انسجام کلی متن در نظر گرفته

شده است که نتایج حاصل، حاکی از کارایی بالای روش پیشنهادی در تعیین انسجام متن خروجی سیستم‌های IQA است.

٤. روش پژوهش

همان‌طور که قبلاً اشاره شد، ویژگی‌های متعددی در ارزیابی یک سیستم IQA دخالت دارند و اندازه‌گیری اتوماتیک آن‌ها برای ایجاد یک مدل دارای اهمیت است. به‌دلیل پیچیدگی بسیار بالای این حوزه و با توجه به کار اندک انجام‌شده بر روی تعیین اتوماتیک انسجام در متن خروجی حاصل از یک سیستم IQA، در این مقاله مدلی جدید برای تعیین اتوماتیک انسجام پیشنهاد شده است. در این روش، با حذف پیچیدگی‌های ساختاری روش‌های تعیین انسجام متن، از رابطهٔ آماری شباهت و ارتباط بین سؤال‌ها و جواب‌ها استفاده شده است که در بخش ویژگی‌ها به تعریف و نحوهٔ استفاده از هر کدام آن‌ها پرداخته شده است. استفاده از ویژگی‌های آماری و تنوع معیارهای استفاده‌شده در اعتبارسنجی مدل، از دیگر امتیازات روش پیشنهادی است. شکل ۲ و ۳، ساختار روش پیشنهادی را در دو فاز نمایش می‌دهد.



شکل ۲. فلوچارت فاز اول روش پیشنهادی

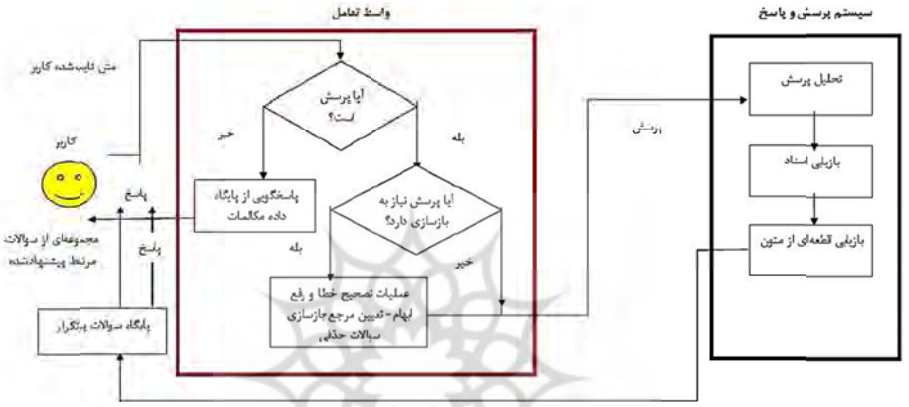
٤-۱. سیستم تعاملی پایه

سیستم تعاملی پایهٔ در نظر گرفته‌شده برای انجام آزمایش‌ها، در آزمایشگاه وب‌کاوی «دانشگاه صنعتی شاهرود» طراحی گردید. در سیستم پایهٔ طراحی‌شده، هدف طراحی یک سیستم پرسش و پاسخ مبتنی بر تعامل با استفاده از تکنیک‌های آماری جهت استخراج دانش نهفته در متون ساختارنیافته بود. سیستم ارائه‌شده بر خلاف سیستم‌های پیشین که عمدتاً از تجزیه و تحلیلی معنایی و گرامری استفاده می‌کردند، مستقل از زبان طراحی شده بود و در آن از روش‌های مبتنی بر معنا و گرامر استفاده نشده بود.

شکل ۳. فلوچارت فاز دوم روش پیشنهادی

بنابراین، این سیستم با در اختیار داشتن پایگاه دادگان مناسب هر زبان می‌توانست به سؤالات مطرح‌شده به آن زبان پاسخ دهد (شهرآئینی و زاهدی ۱۳۹۵). در سیستم

طراحی‌شده از پایگاه دادگان فارسی که به این منظور طراحی شده بود، جهت آموزش و ارزیابی استفاده گردید. معماری سیستم پایهٔ تعاملی طراحی‌شده در شکل ۴، نمایش داده شده است. بررسی نظرات ارائه‌شده توسط کاربران نشان‌دهندهٔ رضایت آن‌ها از کیفیت تعامل برقرار شده با سیستم است. این نکته قابل ذکر است که با ایجاد تغییراتی در این سیستم توانستیم به نتایج بهتری در جهت افزایش عملکرد و کارایی سیستم دست پیدا کنیم. نتایج حاصل از این بهینه‌سازی در (حسینی و زاهدی، ۱۳۹۵) آورده شده است.



شکل ٤. معماری سیستم پرسش و پاسخ تعاملی پایه (شهرآئینی و همکارانش ۱۳۹٤)

٤-۲. پایگاه داده

به‌دلیل نبودِ پایگاه دادهٔ استاندارد ابتدا باید برای انسجام مکالمه یک پایگاه داده ساخته می‌شد. بنابراین، در مرحلهٔ نخست می‌بایست یک سیستم پرسش و پاسخ تعاملی آماده می‌گردید که در این راستا سیستم طراحی و پیاده‌سازی‌شدهٔ آزمایشگاه وب‌کاوی «دانشگاه صنعتی شاهرود» مورد استفاده قرار گرفت. برای این کار، جهت آموزش سیستم و ساخت سیستم تعاملی اولیه، از سه پایگاه دادگان فارسی با نام WMPR-QA1-2015، WMPR-QA2-2015 و WMPR-QA3-2015 استفاده شده است که در زمان تست و ارزیابی سیستم طراحی‌شده از ۲۲ پرسش و پاسخ از مجموع سه پایگاه دادگان فوق به‌کار گرفته شد. پایگاه دادگان اول با نام WMPR-QA1-2015 دارای چهار فایل متنی با محتوای آئین‌نامهٔ آموزشی «دانشگاه شاهرود» است که در قالب ۲۹۲ جمله و با فرمت UTF-8 گردآوری شده است و به‌عنوان دادهٔ آموزشی شناخته می‌شود. ۸۱ پرسش و پاسخ مطرح‌شده از

این آئین‌نامه نیز به‌عنوان مجموعهٔ تست پایگاه دادگان فوق در نظر گرفته شده است. پایگاه دادگان دوم با نام WMPR-QA2-2015 دارای یک فایل متنی با محتوای آئین‌نامهٔ مالی شهرداری‌هاست که در قالب ۷۵ جمله و با فرمت UTF-8 گردآوری شده است و از آن به‌عنوان مجموعهٔ آموزش استفاده می‌شود. ۳۳ پرسش و پاسخ مطرح‌شده از این آئین‌نامه نیز به‌عنوان مجموعهٔ تست پایگاه دادگان WMPR-QA2-2015 در نظر گرفته شده است. پایگاه دادگان سوم با نام WMPR-QA32015 شامل دو مجموعهٔ آموزش و تست است. مجموعهٔ آموزش آن دارای یک فایل متنی با محتوای آئین‌نامهٔ استخدام هیئت علمی دانشگاه‌هاست که در قالب ۲۵۶ جمله و با فرمت UTF-8 گردآوری شده است و مجموعهٔ تست آن دربردارندهٔ ۳۱ پرسش و پاسخ مطرح‌شده از این آئین‌نامه است. سه پایگاه دادگان فوق هم‌اکنون از وب‌سایت آزمایشگاه وب‌کاوی و شناسایی الگوی «دانشگاه صنعتی شاهرود»[1] قابل دریافت هستند. حال، لازم بود با سیستم کار شده و متن خروجی از سؤالات ردوبدل‌شده بین سیستم و کاربر تولید شود تا با برچسب‌گذاری مناسب این سؤالات، پایگاه داده ساخته شود. در این مرحله سامانه‌ای تحت وب طراحی گردید که متن تبادل‌شده و امتیاز داده‌شدهٔ کاربران به انسجام به‌صورت اتوماتیک در پایگاه داده‌ای ذخیره گردد. در این راستا تعداد ۱۲۰ کاربر برای ۴ موضوع مختلف با سیستم کار کردند و مکالمات هر یک با توجه به موضوع به‌صورت جداگانه ذخیره گردید. از این مجموعهٔ ۴۸۰ تایی، ۳۹۰ نمونه توسط فرد خبره به‌عنوان نمونهٔ مناسب‌تر انتخاب شد. جدول ۱، یک نمونه از خروجی متن ذخیره‌شده در پایگاه داده (بدون فرمت) را نشان می‌دهد. همچنین، برای سنجش مدل پیشنهادی از یک پایگاه دادهٔ دیگر که شامل ۳۰ مکالمهٔ ذخیره‌شده از حاصل کار با یک سیستم مکالمه‌محور به نام Cleverbot بود، استفاده گردید.

1. http://wmpr.ir/fa/index/category/53

جدول ۱. نمونهٔ 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؟
دانشجوی کارشناسی ارشد	منظور شما معدل کل است؟
خیر، معدل یک ترم	در صورت کمتر شدن معدل از ۱۴ در هر نیمسال، دانشجو مشروط تلقی می‌گردد.
دانشجوی کارشناسی چطور؟	در صورت کمترشدن معدل از ۱۲ در هر نیمسال، دانشجو مشروط تلقی می‌گردد.
چند ترم متوالی مشروط‌شدن دانشجو اخراج می‌شود؟	دانشجوی کارشناسی یا ارشد؟
دانشجوی کارشناسی	در صورتی که دانشجوی کارشناسی سه ترم متوالی یا چهار ترم غیرمتوالی مشروط شود، در آستانهٔ اخراج قرار می‌گیرد.

٤-۳. استخراج ویژگی

استخراج ویژگی یکی از مهم‌ترین قسمت‌های هر سیستم تشخیص یا مدل‌سازی محسوب می‌شود. از آنجا که در بیشتر روش‌های آماری در پردازش متن از N-گرم‌ها بهره گرفته شده، ما نیز از این ویژگی در انسجام متن استفاده کردیم. با توجه به این‌که برای N-گرم‌ها، برای n=2 به بالا انسجام محلی وجود دارد، ما از این نکته بهره گرفته و از آن برای مدل‌سازی تعیین انسجام در متن خروجی یک سیستم IQA استفاده کردیم. برای اندازه‌گیری انسجام محلی و رسیدن به انسجام کلی، ما چهار رابطه برای ارتباط بین N-گرم‌ها تعریف کردیم و از این چهار رابطه توانستیم ۴۸ ویژگی ایجاد نماییم. این چهار رابطهٔ تعریف‌شده به شرح زیر است.

$$Likeness(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1\cup S_2|} \quad (۱)$$

$$Cosine\_dis=\frac{|S_1\cap S_2|}{\sqrt{|S_1|*|S_2|}} \quad (۲)$$

$$Containment(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_2|} \quad (۳)$$

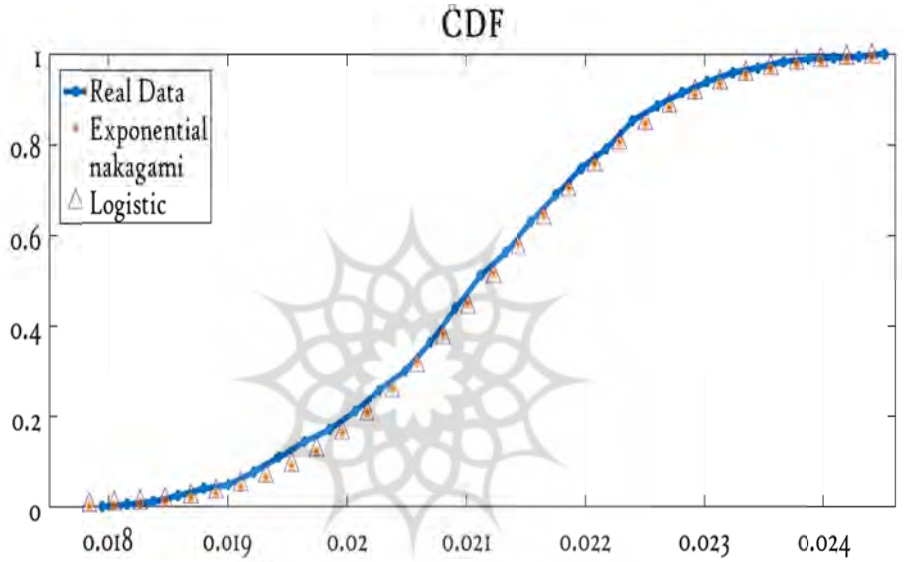
$$Overlap(S_1,S_2)=\frac{|S_2|}{|S_1|} \quad (۴)$$

همان‌طور که در روابط بالا نمایش داده شده، دو مجموعهٔ S_1 و S_2 تعریف شده که به‌عنوان مثال، برای یک جفت پرسش و پاسخ، مجموعهٔ S_1 مجموعهٔ N-گرم‌های مربوط به سؤال و مجموعهٔ S_2 مجموعهٔ N-گرم‌های مربوط به جواب است. همان‌طور که می‌دانیم یک مکالمه از تعدادی پرسش و پاسخ تشکیل شده است. بنابراین، مقدار هر یک از این روابط به‌صورت جداگانه برای هر جفت پرسش و پاسخ محاسبه شده و در نهایت، میانگین برای هر مکالمه محاسبه و در ماتریس ویژگی مربوط به هر مکالمه قرار گرفت. قابل ذکر است که برای تشکیل ماتریس ویژگی از این روابط یک‌بار برای جفت Q_i و A_i سپس، برای جفت‌های Q_i و Q_{i+1}، A_i و Q_i و در نهایت، برای Q_i و A_{i+1} در محاسبات استفاده شد که در آن A_iها جواب‌های یک مکالمه و Q_iها سؤالات یک مکالمه هستند. مقدار N را ۲ تا ۴ فرض کردیم. بنابراین، با توجه به چهار گروه در نظر گرفته شده، مقدار N و چهار رابطهٔ تعریف‌شده، ۴۸ ویژگی حاصل شد. با توجه به تعداد مکالمات، اندازهٔ ماتریس ویژگی ۴۹×۳۹۰ حاصل گردید که ۴۸ ستون آن مربوط به ویژگی‌ها و ستون آخر مربوط به کلاس مکالمه است. ما در پایگاه داده مکالمه‌ها را به سه کلاس تقسیم کردیم: در کلاس اول مکالمات منسجم، در کلاس دوم مکالمات نیمه‌منسجم و در کلاس سوم مکالمات غیرمنسجم قرار گرفته بودند. دسته‌بندی هر مکالمه بر اساس امتیازات داده‌شدهٔ کاربران به انسجام مکالمه و با نظارت فرد خبرهٔ انسانی صورت پذیرفته و تعداد مکالمه‌های هر کلاس ۱۳۰ نمونه بود که با توجه به سه دستهٔ موجود مجموعاً ۳۹۰ نمونه خواهد شد.

٤-٤. مدل پیشنهادی

یک الگوریتم مناسب برای محاسبهٔ انسجام باید قادر باشد مجموعهٔ تفاوت‌های هر مکالمه را با مکالمه‌های دیگر به‌صورت کاملاً پررنگ ببیند و در همان حال، مجموعهٔ تفاوت‌های هر مکالمه با خودش را در حد امکان از بین ببرد. با این دیدگاه و نظر به این‌که در الگوریتم محاسبهٔ انسجام بیش از یک نمونه از مکالمه‌های منسجم و غیرمنسجم در اختیار داریم. این مجموعهٔ ویژگی‌های به‌دست‌آمده را به دو گروه آموزش و تست تقسیم کردیم. سپس، بهترین تابع چگالی احتمال که بر این داده‌ها منطبق بود، تعیین گردید. توابعی که ما برای این قسمت در نظر گرفتیم، ۱۸ تابع چگالی احتمال بودند. پس از این مرحله با توجه به فرم هیستوگرام به‌دست‌آمده، سه تابع چگالی احتمال نمایی،

«ناکاگامی»[1] و لجستیک به‌عنوان کاندیداهای احتمالی انتخاب شدند. یک نمونه تابع توزیع تجمعی برای هیستوگرام به‌دست‌آمده و فرم تقریب زده‌شدهٔ این سه تابع برای یک ویژگی از یک مکالمه در شکل ۵، آورده شده است. همان‌طور که در شکل ۵، نشان داده شده، هر سه تابع انطباق بسیار عالی با فرم اطلاعات واقعی دارند. برای این‌که بررسی شود کدام‌یک از این سه تابع چگالی احتمال انطباق بهتری با اطلاعات مجود دارند، چند معیار سختگیرانهٔ آماری به‌عنوان سنجش میزان انطباق در نظر گرفته شد.



شکل ۵. تابع توزیع تجمعی ۳۲ نقطه‌ای ویژگی اول برای یک مکالمه و سه تابع توزیع تجمعی تقریب زده‌شده

٤-۵. تجزیه و تحلیل یافته‌ها

در مرحلهٔ اول، میزان نیکویی برازش[2] بین منحنی اطلاعات واقعی و منحنی توزیع تجمعی سه تابع تقریب زده شد. این معیار به‌ازای تمام ویژگی‌ها و تمامی مکالمات و برای دو دسته اطلاعات آموزش و تست به‌صورت مجزا محاسبه شد. علاوه بر آن، در مواردی که اطلاعات هنگام تطبیق با تابع چگالی احتمال همخوانی نداشته باشد و نتواند تابع مناسبی برای پوشش اطلاعات واقعی با یک تابع چگالی احتمال به‌خصوص یافت

1. Nakagami
2. goodness of fit

نمایـد، یـک پیغـام خطـا مبنـی بـر عـدم تطابـق اطلاعـات بـا تابـع مفـروض به‌وجـود می‌آیـد. بـرای سـه تابـع در نظـر گرفته‌شـده نتایـج جـدول ۲، نشـان می‌دهـد کـه تقریبـاً هیـچ خطـای تطبیقـی در سـه حالـت رخ نـداده (در بدتریـن حالـت ۱۱۸ خطا در مجمـوع ۱۸۷۲۰ حالـت داریـم کـه زیـر ۱ درصـد محسـوب می‌شـود و به‌طـور قطـع، می‌تـوان هـر سـه تابـع چگالـی احتمـال را در ایـن وضعیـت بسـیار دقیـق ارزیابـی کـرد) و معیـار نیکویـی بـرازش نیـز به‌صـورت مطلوبـی بـرای هـر سـه تابـع بالاسـت. بـا توجـه بـه اعـداد بسـیار مناسـب نیکویـی بـرازش نمی‌تـوان تفـاوت معنـاداری بیـن سـه تابـع یافـت و بـه همیـن دلیـل سـراغ آزمون‌هـای آمـاری دقیق‌تـر رفتیـم.

٤-٦. آزمون‌های سه‌گانهٔ تطبیق یا عدم تطبیق تابع چگالی احتمال

در اولیـن گام، بـا تابـع چگالـی احتمـال به‌دسـت‌آمده اطلاعـات تصادفـی جدیـد تولیـد می‌کنیـم. لازم بـه ذکـر اسـت کـه مـا طـول ایـن بـردار تصادفـی را به‌صـورت پیش‌فـرض ۲۰۰ در نظـر گرفتیـم. سـپس، بیـن اطلاعـات تولیدشـده بـا تابـع چگالـی احتمـال به‌دسـت‌آمده از نمونه‌هـای آمـوزش یـک کلاس از مکالمـه و نمونه‌هـای آزمایـش واقعـی همـان کلاس از مکالمـه به‌وسـیلهٔ آزمون‌هـای آمـاری یکسـان‌بودن تابـع احتمـال را کنتـرل می‌کنیـم.

جدول ۲. نتایج تطبیق سه تابع چگالی احتمال با کل ویژگی‌ها در سه گروه مکالمهٔ انتخاب‌شده

نیکویی برازش	خطای تطبیق	تعداد ویژگی	مجموع مکالمه‌ها	
۰/۹۰۷	۱۱۸	۴۸	۳۹۰	نمایی
۰/۹۴۷	۲۱	۴۸	۳۹۰	ناکاگامی
۰/۹۱۲	۸۹	۴۸	۳۹۰	لجستیک

بدیهـی اسـت جـواب بایـد مثبـت باشـد و آزمـون متناظـر یکسـان‌بودن دو نمونه را اعـلام کند. در غیـر ایـن صـورت معنـا آن اسـت کـه تابـع چگالـی احتمـال تقریـب زده‌شـده بـرای نمونه‌هـای آمـوزش فـرم مناسـبی نداشـته و پایـدار نیسـت. نتایـج حاصـل از ایـن آزمایـش در جـدول ۳، آورده شـده اسـت. سـه آزمـون آمـاری استفاده‌شـده در ایـن مرحلـه شـامل «کولموگروف-اسـمیرنف»[1]، «اندرسـون دارلینـگ»[2] و «کرامـر وان میـس»[3] بوده‌انـد کـه هـر سـه از معیارهـای متـداول بررسـی

1. Kolmogorov-Smirnov
2. Anderson Darling
3. Cramér–von Mises

یکسان‌بودن توزیع بین دو دسته اطلاعات با طول یکسان محسوب می‌شوند. باز هم بر اساس نتایج به‌دست‌آمده می‌توان مشاهده نمود که هر سه تابع با دقت فوق‌العاده‌ای توانسته‌اند تابع چگالی احتمال یک مکالمه را پیش‌بینی و مشخص کنند. در آخرین تست به سراغ ایجاد تمایز بین یک کلاس از مکالمه و سایر کلاس‌ها می‌رویم. تابع چگالی احتمال مناسب باید بین مکالمهٔ یک کلاس و کلیهٔ کلاس‌های دیگر در مکالمه تفاوت قائل شده و هر سه آزمون یادشده شباهت اطلاعات دو مکالمهٔ مختلف را رد کنند. طبعاً تعداد مقایسات در این حالت بسیار بالاتر است؛ چرا که یک کلاس از مکالمه با تمامی کلاس‌های دیگر در تمام ویژگی‌ها باید مقایسه و امتیازدهی شود. از آنجا که درصد انطباق هر سه تابع بسیار بالا بوده، در این مرحله به سراغ درصد درستی کل نرفتیم و تک‌تک اشتباهات در هر سه تابع را جداگانه شمارش کردیم. در این مرحله، در مجموع ۱۳۰ مکالمه برای هر کلاس داشتیم که هر کلاس از یک مکالمه با مکالمهٔ هم‌گروه خود در سه تابع چگالی احتمال و به ازای هر تابع دو بار در ۴۸ ویژگی مقایسه شد. در مجموع، در هر حالت ۱۸۷۲۰ مقایسه صورت می‌گیرد. نتیجهٔ به‌دست‌آمده باز هم فوق‌العاده عالی است. اما در این حالت بین سه تابع می‌توان تفاوت‌های مشخصی دید و تابع مناسب‌تر را انتخاب نمود. نتیجهٔ حاصل از این کار در جدول ۴، آورده شده است. بر اساس نتایج این جدول، تابع چگالی احتمال «ناکاگامی» با مجموع ۱۱۸۵ اشتباه نسبت به دو تابع دیگر که هر دو بیش از ۱۶۰۰ اشتباه داشته‌اند، بهتر عمل کرده است.

جدول ۳. بررسی تطبیق یک کلاس از مکالمه با خودش در سه تابع چگالی احتمال مورد بررسی

	مجموع مکالمات	تعداد ویژگی	تعداد تطبیق	خطای تطبیق	درصد خطای تطبیق
نمایی	۳۹۰	۴۸	۱۸۷۲۰	۴۵۱	٪۲/۵
ناکاگامی	۳۹۰	۴۸	۱۸۷۲۰	۲۱۲	٪۱/۱
لجستیک	۳۹۰	۴۸	۱۸۷۲۰	۳۷۴	٪۲

در انتهای کار برای مقایسهٔ بسیار دقیق روی یک پایگاه دادهٔ دیگر، روش پیشنهادی بر روی پایگاه داده حاصل از کار با سیستم cleverbot پیاده‌سازی و تست شد. از مجموع ۳۰ مکالمهٔ صورت‌گرفته، در برآورد تابع چگالی احتمال یک کلاس با خودش تابع چگالی احتمال «ناکاگامی» توانست در تمامی موارد درست عمل کند و هیچ مشکلی نداشت. در مقایسه بین دو کلاس در این حالت، از مجموع ۱۴۴۰ وضعیت ممکن در ۱۳۱ مورد دچار

اشتباه شد که از نظر درصد خطا برابر عدد ۹ محسوب می شود و باز هم نشان از قابل قبول بودن مدل پیشنهادی است.

جدول ٤. مقایسه بین تابع چگالی احتمال کلاس‌های مکالمهٔ مختلف و آزمون‌های آماری

	اندرسون-دارلینگ		کرامر وان میس		کولموگروف-اسمیرنف	
تعداد تطبیق	اشتباه	درصد درستی	اشتباه	درصد درستی	اشتباه	درصد درستی
۱۸۷۲۰	۴۷۳	۹۷/۴۸	۵۸۴	۹۶/۸۸	۷۴۲	۹۶/۰۳	نمایی
۱۸۷۲۰	۲۱۹	۹۸/۸۳	۳۱۲	۹۸/۳	۶۵۴	۹۶/۵	ناکاگامی
۱۸۷۲۰	۴۸۶	۹۷/۴۰	۶۳۱	۹۶/۶	۹۱۷	۹۵/۰۱	لجستیک

۵. نتیجه‌گیری

در این مقاله روشی برای مدل‌کردن تعیین میزان انسجام متن خروجی یک سیستم پرسش و پاسخ تعاملی ارائه شد. با توجه به کارهای کم انجام‌شده در زمینهٔ تعیین انسجام متن خروجی و در دسترس نبودن سیستم‌های پرسش و پاسخ تعاملی، ابتدا ارتقایی بر روی سیستم پایه صورت پذیرفت. سپس، ۴۸ ویژگی از هر مکالمهٔ صورت‌پذیرفته استخراج گردید. در نهایت، با استفاده از دسته‌بندی تصادفی و تولید یک هیستوگرام پایدار برای هر ویژگی به ازای هر مکالمه، با استفاده از فرم تابع توزیع تجمعی به‌دست‌آمده از هیستوگرام‌ها، سه تابع چگالی احتمال که مشابه با هیستوگرام به‌دست‌آمده بودند، انتخاب شدند. توابع انتخابی بررسی و با آزمون‌های مختلف ارزیابی شدند. نتایج نشان داد که تابع چگالی احتمال «ناکاگامی» بهترین و کم‌خطاترین مدل متناظر را بر اساس پروسهٔ طی‌شده نشان می‌دهد. از امتیازات مدل پیشنهادی علاوه بر دقت بسیار بالای آن، قدرت مدل‌سازی هر ویژگی با دو عدد است که در هیچ‌یک از کارهای مشابه مشاهده نشد. معمولاً در استفاده از n-گرم‌ها کاراکترهای مشترک بین کلمات در یک n-گرم در نظر گرفته نمی‌شود که خود این امر می‌تواند بر روی نتایج در تعیین انسجام تأثیرگذار باشد. بنابراین، پیشنهاد می‌شود که علاوه بر این معیار از زیررشتهٔ مشترک نیز در بین n-گرم‌ها استفاده گردد و همچنین، با توجه به نتایج به‌دست‌آمده می‌توان به‌دنبال تعریف ویژگی‌های دیگری در جهت بهترشدن مدل پیشنهادی در آینده گام برداشت.

فهرست منابع

محمدمهدی حسینی و مرتضی زاهدی. ۱۳۹۵. بهبود پاسخ ارائه‌شده در سیستم‌های پرسش و پاسخ تعاملی به کمک شبکه عصبی. *هشتمین کنفرانس بین‌المللی فناوری اطلاعات و دانش، همدان، انجمن فناوری اطلاعات و ارتباطات ایران، دانشگاه بوعلی‌سینا، شهریور ماه*، صفحات ۷۶-۸۲.

سلیمه سادات شهرآئینی و مرتضی زاهدی، ۱۳۹۵، *ارائه مدل تعاملی جهت استفاده در سیستم‌های پرسش و پاسخ اتوماتیک*، کنفرانس بین‌المللی دانشگاه فردوسی، مشهد، صفحات ۱۱۴-۱۲۰.

M. Amit, and S. K. Jain, 2016, *A survey on question answering systems with classification*. Journal of King Saud University-Computer and Information Sciences vol. 28, no. 3, pp. 345-361.

Barzilay, R., and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34: 1-34.

R. Barzially, M. Lapata, 2005, *Automatic Evaluation of Text Coherence: Models and Representations*, Proceedings of the 19th international joint conference on Artificial intelligence, vol. 5, pp. 1085-1090.A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, 2015, *Question Answering Systems: Survey and Trends*, Procedia Computer Science, vol. 73, pp. 366-375.

R. Barzilay, M. Lapata, 2008, *Modeling local coherence: An entity-based approach*, Computational Linguistics, vol 34, pp. 1-34.

Christensen, Janara, Stephen Soderland Mausam, Stephen Soderland, and Oren Etzioni. 2013. *Towards Coherent Multi-Document Summarization,* In HLT-NAACL, pp. 1163-1173.

S.K. Dwivedi, and V. Singh, 2013, *Research and Reviews in Question Answering System.* Procedia Technology, vol. 10, pp. 417-424.

F. Pamela, D. Giampiccolo, B. Magnini, A. Peñas, Á. Rodrigo, and R. Sutcliffe, *2010, evaluating multilingual question answering systems at CLEF*, Proceedings of Language Resource and Evaluation Conference 2010, Malta, pp. 2774-2781.

C. Guinaudeau, M. Strube, 2013, *Graph-based Local Coherence Modeling*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pp. 93–103.

J. Pooja, and J. Sachin, 2016, *Summarizing Text Using Lexical Chains,* International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no.4, pp.524-530.

D. Kelly, P. Kantor, E. Morse, J. Scholtz, and Y. Sun, 2009, *Questionnaires for eliciting evaluation data from users of interactive question answering systems*. Natural Language Engineering, vol. 15, no. 1, PP119-141 ..

O. Kolomiyets, and M. F. Moens, 2011, *A survey on question answering technology from an information* C. Shah, and J. Pomerantz, 2010, *Evaluating and predicting answer quality in community QA,* In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval ACM, USA, pp. 411-418.

Shahraini, S,and M. Zahedi. 2015. A New Statistical Model for Interactive Question Answering Systems. 5th International Conference on Computer and knowledge Engineering, Mashhad.

ShivaKumar, K. M., and R. Soumya. 2015. Text Summarization using Clustering Technique and SVM Technique. *International Journal of Applied Engineering Research* 10 (12): 28873-28881.

Y. Sun, P. Kantor, and E. Morse, 2011, *Using cross-evaluation to evaluate interactive QA systems*. Journal of the American Society for Information Science and Technology, vol.62, no.9, pp. 1653-1665.

J. Tague-Sutcliffe, 2001, *The pragmatics of information retrieval experimentation, revisited*. Information Processing & Management, vol.24, no.4, pp.467-490.

N. Wacholder, S. Small, B. Bai, D. Kelly, R. Rittman, and P. Kantor, 2004, *Designing a Realistic Evaluation of an End-to-end Interactive Question Answering System, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04), Lisbon, Portugal, pp.989-992.*

M. Zhang, V. W. Feng, 2015, *Encoding World Knowledge in the Evaluation of Local Coherence*, Proceedings of the conference, Chapter of the Association for Computational Linguistics: Human Language Technologies, North American, pp. 1087-1096.

محمدمهدی حسینی

متولد سال ۱۳۶۲، دانشجوی مقطع دکتری تخصصی رشتهٔ مهندسی کامپیوتر، هوش مصنوعی در دانشگاه صنعتی شاهرود است. ایشان هم‌اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد شاهرود است. پردازش متن، سیستم‌های پرسش و پاسخ تعاملی، داده‌کاوی از جمله علایق پژوهشی وی است.

مرتضی زاهدی

متولد سال ۱۳۵۴، دارای مدرک تحصیلی دکتری تخصصی کامپیوتر از دانشگاه RWTH-Aachen آلمان است. ایشان هم‌اکنون استادیار گروه کامپیوتر دانشگاه صنعتی شاهرود است. تعامل انسان و کامپیوتر، شناسایی الگو، پردازش تصویر از جمله علایق پژوهشی ایشان است.