E-ISSN: 2676-6159 | P-ISSN: 2538-5079

Journal of Decisions and Operations Research

www.journal-dmor.ir

J. Decis. Oper. Res. Vol. 6, No. 2 (2021) 242–255.

Paper Type: Original Article

Automatic Assessment of Short Answers Based on Computational and Data Mining Approaches

Hossein Sadr[1,*], Mojdeh Nazari Soleimandarabi[2], Zeinab Khodaverdian[2]

[1] *Department of Computer Engineering, Islamic Azad University, Rasht Branch, Guilan, Iran; sadr@qiau.ac.ir;*
[2] *Department of Computer Engineering, Ayandegan Institute of Higher Education, Tonekabon, Iran; zeinabkhodavardian@msc.aihe.ac.ir; zeinabkhodavardian@msc.aihe.ac.ir.*

Citation:

Sadr, H., Nazari Soleimandarabi, M., & Khodaverdian, Z. (2021). Automatic assessment of short answers based on computational and data mining approaches. *Journal of decisions and operations research, 6*(2), 242-255.

Received: 06/01/2021 Reviewed: 22/02/2021 Revised: 23/03/2021 Accept: 01/05/2021

Abstract

Purpose: Automatic short answer grading is known as the task of automatic assessment of answers based on natural language using computation methods and machine learning algorithms. The proliferation of large-scale intelligent education systems and the importance of assessment as a key factor in the education process have increased the need for highly flexible automated systems for scoring exams.

Methodology: While in the process of automatic short answer grading, student's answer is compared to an ideal response and scoring is done based on their similarity, semantic relatedness and similarity measures can also be employed for this aim. To this end, several semantic relatedness and similarity measures are firstly compared in application of short answer grading. In the following, a method for improving the performance of short answer grading systems based on semantic relatedness and similarity measures which leverages students' answers with the highest score as feedback is proposed.

Findings: In order to evaluate the performance of semantic and similarity relatedness methods in application of automatic short answer grading and the prposed model, various experiments were concucted on Mohler and Mihalcea dataset that contains 7 questions and 630 answers.

Originality/Value: Based on the empirical experiments not only semantic relatedness and similarity measures have great efficiency in automatic short answer grading but also using students' answers as feedback can considerably improve the accuracy and performance of semantic relatedness and similarity measures for this task.

Keywords: Data mining approaches, Short answer grading, Semantic relatedness, Semantic similarity.

Corresponding Author: sadr@qiau.ac.ir
10.22105/DMOR.2021.251713.1230

Journal of Decisions and Operations Research

تصمیم‌گیری و تحقیق در عملیات

دوره ۶، شماره ۲، (۱۴۰۰)، ۲۵۵-۲۴۲

www.journal-dmor.ir

نوع مقاله: پژوهشی

ارزیابی اتوماتیک آزمون‌های تشریحی مبتنی بر رویکردهای محاسباتی و داده کاوی

حسین صدر[۱*]، مژده نظری سلیمان دارابی[۲]، زینب خداوردیان[۲]

[۱] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد رشت، گیلان، ایران.

[۲] گروه مهندسی کامپیوتر، موسسه آموزش عالی آیندگان، تنکابن، ایران.

چکیده

هدف: نمره‌دهی خودکار آزمون‌های تشریحی فرآیند ارزیابی اتوماتیک پاسخ‌های سوالات مبتنی بر متن با استفاده از روش‌های محاسباتی و یادگیری ماشین است. گسترش استفاده از سیستم‌های آموزشی هوشمند و اهمیت ارزیابی نیاز به سیستم‌های خودکار برای نمره‌دهی آزمون‌ها را بیش از پیش افزایش داده است.

روش شناسی پژوهش: با توجه به اینکه در فرآیند نمره‌دهی خودکار، پاسخ‌های متنی ارائه شده توسط دانش‌آموزان با یک پاسخ ایده آل بر اساس میزان شباهت آن‌ها مورد مقایسه قرار می‌گیرد، می‌توان از تکنیک‌های محاسبه ارتباط و شباهت معنایی بین متون نیز برای اینکار بهره برد. در این راستا، در این مقاله ابتدا روش‌های مختلف محاسبه ارتباط معنایی در کاربرد ارزیابی خودکار آزمون‌های تشریحی با هم مقایسه و تاثیر دامنه و اندازه منبع دانش پیش‌زمینه‌ای روی دقت الگوریتم‌ها بررسی شد. در ادامه، یک رویکرد برای بهبود عملکرد سیستم نمره‌دهی خودکار آزمون‌های تشریحی معرفی شده که از پاسخ‌های ارائه شده توسط آزمون‌دهندگان که بالاترین نمره را دریافت کرده‌اند، به عنوان بازخورد استفاده می‌کند.

یافته‌ها: برای ارزیابی کارایی روش‌های محاسبه شباهت و ارتباط معنایی در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی و عملکرد مدل پیشنهادی، آزمایشاتی روی مجموعه داده ارائه شده توسط موهلرو میهالسیا که دارای ۷ سوال با ۶۳۰ پاسخ تشریحی است، صورت گرفت.

اصالت/ارزش افزوده علمی: بر اساس نتایج حاصل از آزمایش‌ها، نه تنها روش‌های محاسبه ارتباط معنایی از کارایی بالایی در حوزه ارزیابی خودکار آزمون‌های تشریحی برخوردارند، بلکه استفاده از از بازخورد اتوماتیک نیز می‌تواند دقت و کارایی روشهای محاسبه ارتباط معنایی برای این هدف به طور قابل توجهی افزایش دهد.

کلیدواژه‌ها: ارتباط معنایی، ارزیابی خودکار آزمون‌های تشریحی، رویکرد داده کاوی، شباهت معنایی.

۱- مقدمه

ارزیابی یکی از مهمترین جنبه‌های فرآیند یادگیری است. در سیستم‌های سنتی ارزیابی توسط یک فرد خبره صورت می‌گیرد که وی پاسخ‌های آزمون‌دهندگان را بررسی کرده و بر اساس میزان تطابق آن‌ها با پاسخ درست نمره‌ای را در بازه مشخص به آن‌ها اختصاص می‌دهد. استفاده از یک فرد برای ارزیابی پاسخ‌های ارائه شده با چالش‌های متعددی روبه‌رو است. یکی از مهمترین این چالش‌ها تعداد زیاد آزمون‌دهندگان

* نویسنده مسئول

sadr@qiau.ac.ir

10.22105/DMOR.2021.251713.1230

در مقابل تعداد محدود افراد تصحیح کننده و زمان‌بر بودن فرآیند تصحیح و ارزیابی است. از طرف دیگر ممکن است فرد تصحیح کننده هنگام تصحیح برگه با مشکلاتی از قبیل خستگی مواجه شود که این امر باعث کاهش دقت در تصحیح پاسخ سوالات می‌شود (سوزان و همکاران[۱]، ۲۰۲۰). ضمن این‌که بررسی‌ها نشان داده است که بین پاسخ‌های ارزیابی شده توسط افراد مختلف همبستگی بالایی وجود ندارد. درواقع نمرات دانش‌آموزان یک گروه با دانش‌آموزان گروه دیگر در یک آزمون مشابه کاملاً متفاوت است که این به نحوه تفکر و سلایق فرد تصحیح‌کننده بستگی دارد. در نتیجه می‌توان گفت که عمل تصحیح یک رفتار مبتنی‌بر فکر و عقیده شخصی فرد تصحیح‌کننده است. با توجه به چالش‌های موجود و بهره‌برداری روبه‌رشد از سیستم‌های جامع الکترونیکی، نیاز به یک سیستم نمره‌دهی خودکار بیش از پیش احساس می‌شود. سیستمی که بتواند پاسخ‌ها را در کوتاه‌ترین مدت و با دقت بالا ارزیابی کند، نیاز به معلم را از بین ببرد و تحت تاثیر عوامل محیطی نباشد. در چنین مواقعی استفاده از یک سیستم کامپیوتری هوشمند با سرعت و دقت بالا ضروری است (صدر و همکاران[۲]، ۲۰۲۱؛ فیلقرا و همکاران[۳]، ۲۰۲۰).

آزمون‌ها می‌توانند به صورت‌های مختلفی برگزار شوند. برخی از آزمون‌ها به صورت چندگزینه‌ای، درست/نادرست و یا پرکردن جاهای خالی می‌باشند. با توجه به اینکه این سیستم‌ها نیازمند تجزیه و تحلیل متون پیچیده ندارند، سیستم‌های هوشمند و خودکار مختلفی در طول زمان برای نمره‌دهی این نوع آزمون‌ها معرفی شده است. علیرغم قابلیت انعطاف و کاربرد فراوان این نوع آزمون‌ها، بسیاری از مدرسین آزمون‌ها با پاسخ تشریحی (مبتنی بر متن) را برای ارزیابی ترجیح می‌دهند. چالش سیستم‌های آموزشی هوشمند نیز زمانی آغاز می‌شود که برای ارزیابی آموخته‌های فرد آموزش‌دیده از آزمون‌ها با پاسخ‌های تشریحی استفاده شود. با توجه به اینکه افراد هنگام پاسخ دادن به سوالات ادبیات و نحوه نگارش متفاوتی دارند، تصحیح خودکار این نوع پاسخ‌ها بدون دخالت نیروی انسانی یکی از موانع پیش‌روی سیستم‌های آموزشی هوشمند است. در واقع می‌توان گفت هدف ارائه سیستم است که بتواند پاسخ‌های ارائه شده را با توجه به مفهوم پاسخ و بدون توجه به اشتباه‌های نگارشی و املایی ارزیابی کند (صدر و نظری سلیمان دارابی[۴]، ۲۰۱۹؛ ژانگ و همکاران[۵]، ۲۰۲۰). از آنجایی‌که در سیستم‌های نمره‌دهی خودکار پاسخ ارائه شده توسط پاسخ‌دهنده با یک یا چند پاسخ صحیح مورد مقایسه قرار می‌گیرد و میزان شباهت و ارتباط آن‌ها نمره مربوط به پاسخ را مشخص می‌کند، می‌توان از تکنیک‌های محاسبه شباهت و ارتباط معنایی بین متون نیز برای این‌کار بهره برد (نظری سلیمان دارابی و همکاران، ۲۰۱۵b[۶]؛ صدر و همکاران[۷]، ۲۰۱۹a).

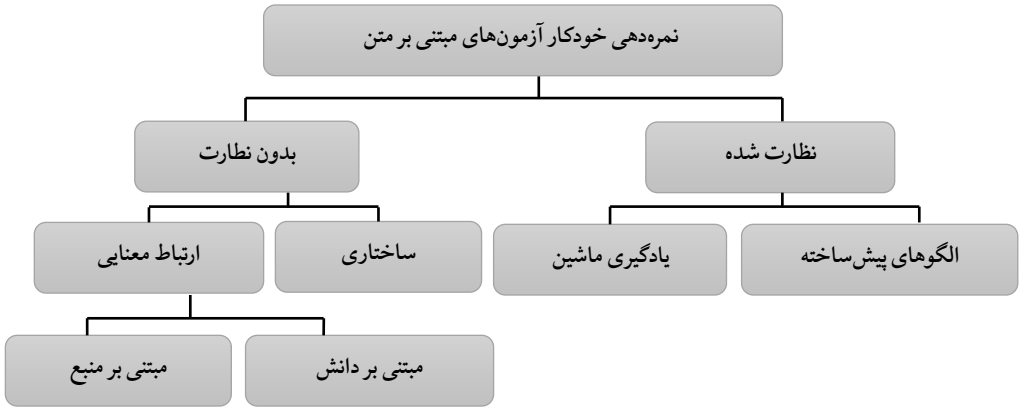
در این مقاله روش‌های مختلف محاسبه ارتباط و شباهت معنایی بین متون به دو دسته مبتنی بر منبع و مبتنی بر دانش تقسیم شده و در کاربرد نمره‌دهی خودکار سوالات تشریحی با هم مورد استفاده قرار گرفته و کارایی آن‌ها بررسی شده است. نتایج آزمایش‌ها نشان می‌دهد که روش‌های مبتنی بر منبع از دقت بالاتری نسبت به روش‌های مبتنی بر دانش برخوردارند. در ادامه نیز به منظور افزایش دقت الگوریتم‌های محاسبه ارتباط معنایی در کاربرد ارزیابی خودکار آزمون‌های تشریحی رویکردی معرفی شده است که در آن از پاسخ‌های ارائه شده توسط آزمون‌دهندگان که بالاترین نمره را دریافت کرده‌اند به عنوان بازخورد استفاده می‌کند. بر اساس نتایج بدست آمده، استفاده از بازخورد خودکار می‌تواند دقت روش‌های مختلف محاسبه ارتباط معنایی متون را در کاربرد ارزیابی خودکار آزمون‌های تشریحی به اندازه قابل توجهی افزایش دهد. سهم علمی این مقاله را می‌توان به طور خلاصه به صورت زیر بیان کرد:

- *با توجه به اینکه در نمره دهی خودکار میزان شباهت پاسخ ارائه شده با یک پاسخ صحیح مورد بررسی قرار می‌گیرد، استفاده از روش‌های محاسبه میزان شباهت و ارتباط معنایی بین متون می‌تواند راه‌حل مناسبی برای این هدف باشد. در این راستا، در این مقاله دقت انواع روش‌های مبتنی‌بر منبع و مبتنی‌بردانش برای نمره دهی خودکار متون مورد بررسی قرار گرفته است. که بر اساس نتایج آزمایش‌ها روش‌های روش‌های مبتنی‌بر دانش از دقت بالاتری در این حوزه برخوردار هستند.*
- *به‌منظور افزایش دقت نمره دهی خودکار در مدل پیشنهادی این مقاله از پاسخ‌های ارائه شده توسط آزمون‌دهندگان که بالاترین نمره را دریافت کرده‌اند به عنوان بازخورد استفاده خواهد شد. نتایج آزمایش‌ها نشان می‌دهد که استفاده از بازخورد خودکار می‌تواند دقت روش‌های محاسبه ارتباط معنایی را برای کاربرد تصحیح خودکار متون به طور قابل توجهی افزایش دهد. ادامه این مقاله به صورت زیر سازماندهی شده است. در بخش دوم کارهای انجام شده در زمینه نمره‌دهی خودکار آزمون‌های مبتنی بر متن مورد بررسی قرار گرفته است. با توجه به اینکه تاکید این مقاله روی روش‌های محاسبه ارتباط معنایی*

---

[۱] Süzen et al.
[۲] Sadr et al.
[۳] Filighera et al.
[۴] Sadr and Nazari Solimandarabi
[۵] Zhang et al.
[۶] Nazari Soleimandarabi et al.
[۷] Sadr et al.

– متون در کاربرد ارزیابی خودکار آزمون‌های تشریحی است، در ادامه این بخش رویکردهای مختلف محاسبه ارتباط و شباهت معنایی با توجه به نوع عملکردشان به دو دسته مختلف تقسیم شده و جزئیات مربوط به آن‌ها بیان شده است. ایده پیشنهادی این مقاله در بخش سوم آمده است. آزمایش‌های انجام شده و نتایج مربوط به ارزیابی‌ها در بخش چهارم بیان شده است. بخش پنجم شامل نتیجه‌گیری و کارهای آینده آمده است.



شکل ۱- طبقه‌بندی روش‌های موجود (صدر و نظری سلیمان دارابی، ۲۰۱۹).
Figure 1- Classification of existing methods.

۲- پیشینه تحقیق

هدف یک سیستم نمره‌دهی خودکار مقایسه پاسخ ارائه شده توسط پاسخ‌دهنده با یک پاسخ ایده آل و تخصیص نمره در یک بازه معین است. با توجه به اینکه در سازمان ها با مقیاس بزرگ فرآیند تصحیح بسیار زمان‌بر و هزینه بر است، استفاده از روش‌های نمره‌دهی خودکار می‌تواند منجر به افزایش کارایی و کاهش هزینه‌ها شود (یانگ[۱]، ۲۰۱۲؛ روی و همکاران[۲]، ۲۰۱۸). بدیهی است که در یک سیستم آموزشی با مقیاس بزرگ امکان ارزیابی تکالیف دانشجویان به‌صورت روش‌های سنتی وجود ندارد. لذا امروزه، سیستم‌های نرم‌افزاری نمره‌دهی خودکار متون یک رکن ضروری (و نه انتخابی) از سیستم آموزش آنلاین است و در عین حال به عنوان یکی از موضوعات تحقیقاتی داغ در حوزه پردازش زبان طبیعی، سیستم‌های اطلاعاتی و تکنولوژی آموزشی به شمار می‌رود. بنابراین، در آینده نزدیک با گسترش روش‌های نوین آموزش، ناگزیر به استفاده از سیستم‌های تصحیح خودکار متون خواهیم بود (ژو و همکاران[۳]، ۲۰۱۹؛ صدر و همکاران[۴]، ۲۰۱۹b).

روش‌های ارزیابی پاسخ آزمون‌های تشریحی به دو دسته کلی تقسیم می‌شوند. دسته اول گروهی از روش‌ها را شامل می‌شود که تاکید آن‌ها روی تصحیح متون از لحاظ نگارشی و دستوری است و به مفهوم پاسخ ارائه شده توجه نمی‌کنند (شرمیش و برستین[۵]، ۲۰۱۳؛ ژو و همکاران، ۲۰۱۹). دسته دوم شامل رویکردهایی است که هنگام تصحیح مفهوم پاسخ را در نظر می‌گیرند و اشتباه‌های نگارشی، دستوری و لغوی تاثیری در فرآیند تصحیح آن‌ها نخواهد داشت (فیلقرا و همکاران، ۲۰۲۰؛ باروز و همکاران[۶]، ۲۰۱۵). در این مقاله روش‌هایی مورد بررسی قرار می‌گیرند که تنها روی مفهوم پاسخ ارائه شده تاکید دارند. روش‌های نمره‌دهی خودکار سوالات تشریحی که روی مفهوم پاسخ ارائه شده تاکید دارند به دو دسته نظارت شده و بدون نظارت تقسیم می‌شوند. در **شکل ۱** طبقه‌بندی کلی از این روش‌ها مشخص شده است.

روش‌های با نظارت رایج برای مقایسه به الگوهایی که توسط یک انسان خبره و به صورت دستی ایجاد شده‌اند، نیاز دارند. به این معنی که اگر پاسخ ارائه شده توسط آزمون‌دهنده با الگوهای از پیش‌تعیین شده مطابقت داشته باشد، سوال به درستی پاسخ داده شده است و اگر با توجه به درست بودن مفهوم پاسخ، با الگوها تطابق نداشته باشد، نمره پایینی به آن اختصاص خواهد یافت. در این بین روش‌های نیمه

---

[۱] Young
[۲] Roy et al.
[۳] Zhu et al.
[۴] Sadr et al.
[۵] Shermis and Burstein
[۶] Burrows et al.

نظارت شده نیز معرفی شدند که از انعطاف پذیری بالاتری نسبت به روش‌های نظارت شده برخوردارند. نکته قابل توجه درباره این روش ها این است که کلیه آن‌ها به یک منبع دانش از الگوهای از پیش‌تعریف شده نیاز دارند و عملکرد آن‌ها نیز به این الگوها وابسته است. از طرف دیگر، ساختن چنین منابع دانشی بسیار زمان‌بر و هزینه‌بر بوده و محدود به یک دامنه خاص است. مشکلات موجود در این روش ها منجر به ایجاد روش‌های بدون نظارت شد که به دخالت مستقیم نیروی انسانی وابسته نبوده و قابلیت اعمال بیشتری در مسایل دنیای واقعی دارند (ژانگ و همکاران[۱]، ۲۰۱۹؛ صدر و همکاران[۲]، ۲۰۲۱).

روش‌های بدون نظارت بر خلاف روش‌های نظارت شده به الگوهای از پیش‌تعیین شده نیاز ندارند، در نتیجه از انعطاف بیشتری برخوردار بوده و با محدودیت‌های کمتری در مسایل دنیای واقعی مواجه هستند. روش‌های بدون نظارت را می‌توان به دو دسته ساختاری و روش های محاسبه ارتباط معنایی تقسیم کرد. محاسبه ارتباط معنایی بین پاسخ ارائه شده توسط آزمون‌دهنده و پاسخ ایده‌آل به عنوان یک تکنیک برجسته در توسعه سیستم‌های نمره‌دهی خودکار به حساب می‌آید (ژانگ و همکاران، ۲۰۱۹؛ لی و همکاران[۳]، ۲۰۲۰). روش های محاسبه ارتباط معنایی را نیز می‌توان به دو دسته روش‌های مبتنی‌بر دانش[۴] و روش‌های مبتنی‌بر منبع[۵] تقسیم کرد (طیب و همکاران[۶]، ۲۰۲۰).

در روش‌های مبتنی‌بر منبع دانش‌پیش زمینه‌ای توسط اجرای آنالیز آماری از یک مجموعه اسناد بزرگ بدون علامت بدست می‌آید این روش‌ها یک فضای معنایی از کلمات را ایجاد می‌کنند که در آن کلمات در مجموعه‌ای از نوشته‌جات توزیع شده‌اند و از رخداد همزمان کلمات در منبع برای محاسبه ارتباط معنایی استفاده می‌شود. در این روش‌ها هر کلمه به یک بردار چندبعدی از مفاهیم نگاشت می‌شود که نشان‌دهنده مفاهیم پنهان است. این روش‌ها می‌توانند مفاهیم متون را بهتر از لغات نشان دهند و با توجه به استفاده از منابع غنی دانش پیش‌زمینه‌ای، دارای همبستگی بالایی با قضاوت‌های انسانی می‌باشند (موهلرو میهالسیا[۷]، ۲۰۰۹، موهلر و همکاران[۸]، ۲۰۱۱). در مقابل، روش‌های مبتنی‌بر دانش از روابط معنایی موجود در بین کلمات و مفاهیم در منابع دانش پیش‌زمینه‌ای مانند ووردنت (بودانیتسکی و هیرست[۹]، ۲۰۰۶) یا روگت (جارماسز و اسپاکویز[۱۰]، ۲۰۱۲، ۲۰۰۳) برای محاسبه ارتباط معنایی استفاده می‌کنند. به این معنی که از طول مسیر در گراف مفاهیم یا شبکه معنایی و میزان مفاهیم به اشتراک گذاشته شده در گراف مفاهیم برای محاسبه ارتباط معنایی هره می‌برند (ژو و همکاران، ۲۰۱۹).

۱-۲- رویکردهای محاسبه ارتباط معنایی متون

با توجه به اینکه یکی از اهداف این مقاله ارائه یک مقایسه بین روش‌های مختلف محاسبه ارتباط معنایی و بررسی کارایی آن‌ها در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی است، در این بخش مجموعه‌ای از روش مبتنی بر دانش و روش مبتنی بر منبع محاسبه ارتباط معنایی مورد بررسی قرار گرفته‌اند. با توجه به اینکه بیشتر روش‌های مبتنی بر دانش تنها توانایی محاسبه میزان ارتباط معنایی بین کلمات را دارند، از متدولوژی معرفی شده (میهالسیا و همکاران[۱۱]، ۲۰۰۶) برای محاسبه میزان ارتباط بین متون استفاده شده است. بر اساس این روش، به ازای هر کلمه در متن ورودی مقدار ماکزیمم ارتباط معنایی که می‌توان با هر کدام از کلمات در متن دیگر بدست آید، در نظر گرفته می‌شود. به عبارت دیگر، به ازای هر کلمه W با نقش C در پاسخ ایده آل می‌توان $maxsim(W,C)$ را به صورت زیر محاسبه کرد.

$$maxsim(W,C)=maxSIM_x(W, w_i). \quad (۱)$$

[۱] Zhang et al.
[۲] Sadr et al.
[۳] Lee et al.
[۴] Knowledge-based
[۵] Corpus-based
[۶] Taieb et al.
[۷] Mohler and Mihalcea
[۸] Mohler et al.
[۹] Budanitsky and Hirst
[۱۰] Jarmasz and Szpakowicz
[۱۱] Mihalcea et al.

در اینجا w_i کلمه‌ای در پاسخ دانش‌آموز با نقش Cاست و SIM_x یکی از معیارهای ارتباط معنایی است که در ادامه معرفی شده است. میزان ارتباط معنایی به ازای تمامی کلمات محاسبه شده، با هم جمع شده و بر اساس طول دو متن ورودی نرمال‌سازی می‌شوند. توضیح کوتاهی درباره هرکدام از روش‌های محاسبه ارتباط معنایی در ادامه آمده است.

۱-۱-۲- روش‌های مبتنی بر دانش

این روش‌ها از ساختار گراف، رده‌بندی و طبقه‌بندی منابع دانش‌پیش‌زمینه‌ای برای محاسبه ارتباط معنایی استفاده می‌کنند. یکی از مهمترین این منابع دانش منبع لغت و وردنت است. وردنت یک منبع لغت همه‌جانبه است که روابط بین کلمات آن به صورت گرافی از مفاهیم نشان داده شده است (صدر و همکاران[۱]، ۲۰۱۹c؛ زسچ و گورویچ[۲]، ۲۰۱۰). ساده‌ترین روش مبتی‌بر دانش که از وردنت بهره می‌برد روش *Path* یا همان کوتاهترین مسیر است که از ساختار گراف وردنت استفاده کرده و معکوس کوتاهترین مسیر بین دو مفهوم را برای محاسبه ارتباط معنایی در نظر می‌گرفت. هرچه طول این مسیر کوتاهتر باشد، میزان ارتباط معنایی افزایش خواهد یافت (پدرسن و همکاران[۳]، ۲۰۰۴).

$$rel_{Path}(c_1,c_2)= \frac{1}{max\ len\ (c_1,c_2)}. \qquad (۲)$$

در ادامه مدل *LCH* معرفی شد که با استفاده از عمق طولانی‌ترین گره از سمت ریشه به برگ عمق گراف را نرمال کرده و مشکل موجود در روش *Path* کلاسیک را برطرف می‌کرد. در فرمول زیر $len\ (c_1,c_2)$ طول مسیر بین دو مفهوم مورد مقایسه و $depth(c)$ طول طولانی‌ترین مسیر در گراف است (لی‌چاک و چودرو[۴]، ۱۹۹۸).

$$rel_{LCH}(c_1,c_2)=-log\frac{(len\ (c_1,c_2))}{2\times \underset{c\ \in wordnet}{depth(c)}}. \qquad (۳)$$

روش *WUP* پس از مدل *LCH* ارایه گردید که از عمق دو مفهوم مورد مقایسه در گراف و اولین مفهوم به اشتراک گذاشته شده (LCS[۵]) بین مفاهیم مورد مقایسه روی مسیر مفاهیم، به سمت ریشه سلسله مراتب برای محاسبه ارتباط معنایی استفاده کرد (وو و پالمر[۶]، ۱۹۹۴).

$$rel_{WUP}(c_1,c_2)=\frac{2\times depth\ (lcs)}{Depth\ (c_1)+depth(c_2\ )}. \qquad (۴)$$

روش *HSO*، برخلاف روش‌های فوق که تنها ساختار *is-a* موجود در بین اسم‌ها را در نظر می‌گرفتند، از کل روابط موجود در وردنت برای محاسبه ارتباط معنایی استفاده می‌کرد. بر اساس این معیار، دو کلمه در صورتی به هم مرتبطند که در یک مجموعه قرار داشته باشند، متضاد باشنند و یک کلمه بخشی از کلمه دیگر باشد (هیرست و اسینج[۷]، ۱۹۹۸).

$$rel_{res}(c_1,c_2)=IC\ (lcs(c_1,c_2)). \qquad (۵)$$

معیار دیگر موجود، روش *Resnik* است که مبتنی بر محتوای اطلاعاتی بوده و خود مفاهیم را در نظر نمی‌گیرد. بر اساس این معیار میزان ارتباط معنایی بین دو مفهوم را با استفاده از اطلاعات مشترک بین آن‌ها محاسبه می‌شود. به این معنی که اگر دو جفت کلمه مختلف *LCS* یکسانی داشته باشند، میزان ارتباط معنایی آن‌ها برابر خواهد بود (رسنیک[۸]، ۱۹۹۵).

$$IC(c)=-\log P(c). \qquad (۶)$$

---

[۱] Sadr et al.
[۲] Zesch and Gurevych
[۳] Pedersen et al.
[۴] Leacock and Chodorow
[۵] Lowest Common Subsumer
[۶] Wu and Palmer
[۷] Hirst and St-Onge
[۸] Resnik

در اینجا *IC* نشان دهنده محتوای اطلاعاتی است که به صورت زیر محاسبه می شود که در آن *P(c)* احتمال مواجه با یک نمونه از مفهوم *c* در یک مجموعه بزرگ می‌باشد (رسینک، ۱۹۹۵).

$$rel_{jcn}(c_1,c_2)=\frac{1}{IC(c_1)+IC(c_2)-2\,IC(lcs)}. \qquad (۷)$$

JCN نیز روشی مبتنی بر تئوری *Resnik* است که به محدودیت‌های موجود در آن غلبه کرد (جیانگ و کونراث[۱]، ۱۹۹۷). *Lin* نیز معیار دیگری است که بر اساس تئوری *Resnik* معرفی شد. در این روش از فاکتور نرمال‌سازی که شامل محتوای اطلاعاتی دو مفهوم ورودی است استفاده می‌شود (لین[۲]، ۱۹۹۸).

$$rel_{lin}(c_1,c_2)=2.\frac{IC(lcs)}{IC(c_1)+IC(c_2)}. \qquad (۸)$$

در وردنت به ازای هر کلمه تعریف کوتاهی درباره آن ارائه شده است. *Lesk* روش محاسبه ارتباط معنایی کلمات مبتنی بر تعاریف کلمات است که از هم‌پوشانی کلمات موجود در تعاریف مربوط به هر مفهوم در وردنت استفاده می‌کند. در واقع هر چه میزان هم‌پوشانی تعاریف دو کلمه بیشتر باشد، آن‌ها به هم مرتبط‌تر خواهند بود (لسک[۳]، ۱۹۸۶). روش *Vector* نیز دیگر معیار مبتنی‌بر منبع دانش است که از تعاریف کلمات موجود در وردنت بهره می‌برد. در این روش به ازای هر مفهوم موجود در وردنت بر اساس تعاریف مربوط به آن یک بردار ساخته می‌شود و بردارهای مربوط به مفاهیم برای محاسبه میزان ارتباط بین آن‌ها با استفاده از معیار کسینوسی با هم مورد مقایسه قرار می‌گیرند (پتوازداهان و پترسن[۴]، ۲۰۰۶).

۲-۱-۲- روش‌های مبتنی بر منبع

مجموعه نوشته‌جات غیر ساخت‌یافته می‌توانند به عنوان یک منبع دانش پیش‌زمینه‌ای برای محاسبه ارتباط معنایی مورد استفاده قرار گیرند. برخلاف روش‌های مبتنی بر دانش، مجموعه نوشته‌جات غیر ساخت‌یافته شامل اطلاعات مربوط به کلمات، روابط معنایی روشن و آشکار بین مفاهیم و توضیحات مربوط به هر کلمه نمی‌باشند. در نتیجه اطلاعات پیش‌زمینه‌ای در سطح کلمات جمع آوری شده و کلمات به صورت ضمنی به هم مرتبط می‌باشند. این دسته از روش‌ها از رخداد لغات[۵] در یک منبع دانش غیرساخت یافته برای محاسبه ارتباط معنایی استفاده می کنند ودانش پیش زمینه خود را توسط اجرای آنالیز آماری از روی مجموعه اسناد بزرگ بدون برچسب بدست می‌آورند و در نهایت هدف آن ها کشف ساختار پنهان بین اسناد می‌باشد. این روش‌ها کاملا به صورت خودکار عمل کرده و نیاز به نیروی انسانی را از بین می‌برند. علاوه بر این، با توجه به اینکه این روش‌ها مفهوم متون را در نظر می‌گیرند و روی کلمات تاکید ندارند می‌توانند در کاربر تصحیح خودکار آزمون‌های تشریحی از دقت بالایی برخوردار باشند (نظری سلیمان دارابی و همکاران[۶]، ۲۰۱۵a؛ زسچ و گورویچ، ۲۰۱۰).

LSA[۷] یکی از مهمترین روش‌ها در این زمینه است که از نمایش برداری برای محاسبه ارتباط معنایی استفاده می‌کند. این تکنیک، توانایی کشف روابط پنهان بین کلمات را دارد. در اصل، *LSA* یک روش کاهش ابعاد فضا می‌باشد که با اعمال تجزیه ویژه مقدار[۸] روی ماتریس کلمه- سند قادر به نگاشت مجموعه اسناد به مجموعه کلمات می‌باشد (دومیس[۹]، ۲۰۰۴).

ESA[۱۰] یک روش مبتنی بر متون ویکی‌پدیا است که در آن هر کلمه به صورت برداری از مقالات ویکی‌پدیا نشان داده می‌شود. به بیان دیگر هر مقاله در ویکی‌پدیا نشان دهنده یک مفهوم در بردار مربوط به کلمات است. میزان ارتباط یک کلمه به یک مفهوم بر اساس نمره *tf*idf* آن کلمه در ویکی‌پدیا مشخص می‌شود و میزان ارتباط بین دو کلمه بر اساس معیار کسینوسی بین بردار مربوط به دو کلمه محاسبه

---

[۱] Jiang and Conrath
[۲] Lin
[۳] Lesk
[۴] Patwardhan and Pedersen
[۵] Term Co-Occurrence
[۶] Nazari Soleimandarabi et al.
[۷] LSA: Latent Semantic Analysis
[۸] SVD: Singular Value Decomposition
[۹] Dumais
[۱۰] ESA: Explicit Semantic Analysis

در فضای با ابعاد بالا مقالات ویکی‌پدیا محاسبه می‌شود (گابریلویچ و مارکویچ[۱]، ۲۰۰۹). لازم به ذکر است که لی وهمکارنش نیز به منظور غلبه بر مشکلات موجود در مدل *ESA*، این مدل را با ساختار گراف مقالات ویکی‌پدیا ترکیب کرده و به دقت بالاتری در این حوزه دست یافتند (لی و همکاران[۲]، ۲۰۱۷).

Wikirelate! مدل دیگری است که با استفاده از ویکی‌پدیا به محاسبه ارتباط معنایی بین کلمات می‌پردازد. این روش از ساختار رده‌بندی موجود در اسناد ویکی‌پدیا بهره می‌برد که شامل سه مرحله می‌باشد. در مرحله اول دو جفت کلمه وارد سیستم می‌شود و صفحات مرتبط با اسناد مربوط به این دو کلمه در ویکی‌پدیا بازیابی می‌گردد. در ادامه سیستم از طریق گراف طبقه‌بندی بین صفحات ویکی‌پدیا حرکت می‌کند تا صفحاتی که متعلق به آن دو کلمه است بازیابی شود. سرانجام، میزان ارتباط معنایی براساس صفحات استخراج شده و مسیرهای اتصال دسته ها در گراف طبقه‌بندی محاسبه می‌شود (استروب و پونزتو[۳]، ۲۰۰۶).

مدل *WOLM*[۴] نیز از ویکی‌پدیا به عنوان منبع دانش پیش‌زمینه‌ای برای محاسبه ارتباط معنایی استفاده می‌کند با این تفاوت که این روش تنها مبتنی‌بر گراف طبقه‌بندی نمی‌باشد و از لینک‌های ساختاری ویکی‌پدیا بجای طبقه‌بندی سلسه‌مراتبی آن بهره می‌برند. *WOLM* باعث ایجاد یک سازش مناسب بین روش *ESA* و *Wikirelate!* از لحاظ کارایی و پیچیدگی شده است. با توجه به اینکه پیوندهای بین مقاله به صورت دستی تعریف می شوند، این روش از نظر تئوری معیاری را ارائه می دهد که از *ESA* ارزان‌تر و دقیق‌تر است: ارزان‌تر، زیرا محتوای متنی گسترده ویکی پدیا را می‌توان تا حد زیادی نادیده گرفت و دقیق‌تر است زیرا وابستگی بیشتری با معناشناسی که به صورت دستی تعریف شده است، دارد (ویتن و میلن[۵]، ۲۰۰۸).

مدل *WCVM* از ریشه تمام مقالات موجود در سلسله مراتب ویکی‌پدیا برای ایجاد بردار مربوط به کلمات و محاسبه ارتباط معنایی بین کلمات با استفاده از فاصله کسینوسی بین کلمات بهره می‌برد. این روش با مرحله پیش‌پردازش آغاز می‌شود که در آن یک بردار تعریف معنایی برای هر مفهوم با استفاده از گراف سلسله مراتب ویکی‌پدیا ایجاد می‌شود که شامل وزن ریشه استخراج شده از مقالات مرتبط با هر مفهوم است. در ادامه، برای هر کلمه دسته‌های مرتبط با آن از گراف سلسله مراتب ویکی‌پدیا استخراج می‌شود. در انتها میزان شباهت معنایی با استفاده از بردارهای معنایی بدست آمده تعیین می‌گردد (طیب و همکاران[۶]، ۲۰۱۳).

در ادامه مسیر مطالعات انجام شده در این حوزه، (ژو و همکاران، ۲۰۱۹) از ترکیب لینک‌های ورودی و خروجی مفاهیم ویکی‌پدیا در یک بردار پیوند دو طرفه در مفسر معنایی به همراه وزن‌های دو طرفه مبتنی‌بر *TF-IDF* برای محاسبه ارتباط معنایی کلمات بهره بردند. در ادامه (لی و همکاران، ۲۰۲۰) از منبع لغت وردنت برای تقویت بردارهای بازنمایش کلمات استفاده کرده و محاسبه ارتباط معنایی بین کلمات را به کمک بردارهای تقویت شده حاصل از ترکیب بردارهای بازنمایش کلمات و مفاهیم وجود در وردنت انجام دادند. در پژوهشی مشابه، (پین‌لت و همکاران[۷]، ۲۰۲۰) نیز از روش بازنمایش ویژگی‌های کلمات *BERT* که یک روش مبتنی‌بر یادگیری عمیق و پردازش ترتیبی کلمات موجود در متن است برای محاسبه ارتباط معنایی بین کلمات استفاده کردند.

با توجه به اینکه روش‌های متعددی برای محاسبه ارتباط معنایی بین متون وجود دارند که هرکدام دارای نقاط قوت و ضعف خود می‌باشند، می‌توان گفت تاکنون مقایسه‌ای روی نحوه عملکرد آن‌ها در کاربرد تصحیح خودکار متون صورت نگرفته است. در نتیجه می‌توان گفت که چالش اول این مقاله این است که کدامیک از روش‌های محاسبه ارتباط معنایی دارای کارایی بهتری در زمینه نمره‌دهی خودکار متون هستند و چالش دوم نیز ارائه یک راهکار جدید است که بتوان با استفاده از آن دقت روش‌های محاسبه ارتباط معنایی در کاربرد نمره‌دهی خودکار آزمون های تشریحی را افزایش داد.

۳- رویکرد مبتنی بر بازخورد مرتبط خودکار

نمره‌دهی خودکار آزمون‌های تشریحی را می‌توان با استفاده از روش‌های محاسبه ارتباط معنایی و از طریق مقایسه پاسخ ارائه شده توسط دانش‌آموز و پاسخ ایده‌آل ارائه شده توسط معلم انجام داد. با توجه به اینکه برای محاسبه ارتباط معنایی بین پاسخ ارائه شده توسط دانش‌آموز با پاسخ ایده‌آل با داده‌های متنی روبه‌روهستیم، ابتدا باید داده‌های متنی مورد پیش‌پردازش قرار بگیرند تا بتوان از آن‌ها به عنوان ورودی مدل

---

[۱] Gabrilovich and Markovitch
[۲] Li et al.
[۳] Strube and Ponzetto
[۴] Wikipedia out-link vector-based measures
[۵] Witten and Milne
[۶] Taieb et al.
[۷] Peinelt et al.

آموزشی بهره برد. پیش پردازش داده‌ها برای آماده‌سازی داده‌ها نیاز است تا آن‌ها را از شکل و حالت اولیه، خارج کرده و به شکلی که برای الگوریتم مناسب باشد.

پیش‌پردازش متون شامل چند مرحله است. مرحله اول قطعه‌بندی است. در قطعه‌بندی، متن به رشته‌های کوچك شناخته‌شده‌ای به نام توکن تقسیم می‌شود. از آنجاییکه هدف استخراج کلمات معنادار برای کاربرد خاص محاسبه ارتباط معنایی و پیرو آن نمره‌دهی خودکار است، باید از بین توکن‌ها، کلمات با ارزش معنایی بالاتر انتخاب شده و کلمات بی‌اهمیت حذف شوند. حذف کلمات بی‌اهمیت مرحله دوم پیش‌پردازش است که در آن کلماتی مانند ضمایر، قیود، حروف اضافه و ربط هستند که تاثیری بر ارزش محتوایی متن ندارند، حذف می‌شوند. مرحله سوم ریشه‌یابی است، هدف از ریشه‌یابی زدودن الحاقات و یافتن جوهره اصلی کلمه است که پیرو آن از ریشه کلمات به‌جای حالت‌های دستوری متفاوت آن‌ها استفاده می‌شود. ریشه‌یابی کلمات، حجم عملیات و حجم ماتریسی که در مراحل بعدی ساخته می‌شود را به مقدار قابل توجهی کاهش می‌دهد.

پس از اینکه عملیات پیش‌پردازش به اتمام رسید، باید ارتباط معنایی بین کلمات محاسبه شود. تاکید این مقاله روی روش *ESA* و *Graph-Based ESA* برای محاسبه ارتباط معنایی بین کلمات است. روش *ESA* از مفاهیم دانش که به صورت واضح تعریف شده و توسط انسان‌ها قابل دستکاری می‌باشند، استفاده می‌شود. در این روش مفاهیم با استفاده از متدهای تجزیه‌وتحلیل متون، از یک هستانشناسی مفاهیم صریح که کاملاً منطبق با شناخت و درک انسانی می باشد، استخراج شده‌اند. در *ESA* از ویکی‌پدیا به عنوان هستانشناسی که نشان‌دهنده مفاهیم صریح مبتنی بر درک و شناخت انسان می‌باشد، استفاده می‌شود. ویژگی قابل توجه این است که اطلاعات موجود در آن مبتنی بر شناخت انسان بوده و قسمت اعظم مقالات آن توسط افراد مختلف، قابل ویرایش است.

روش *ESA* برای تولید بردارهای نهایی کلمات از یک مفسر معنایی استفاده می‌کند. وقتی کلمات یک قطعه متن به مفسر معنایی داده می شود، مفسر معنایی تمام مفاهیم موجود در ویکی‌پدیا را بر اساس ارتباطشان با قطعه متن ورودی، امتیازبندی می‌کند. در ابتدای کار هر قطعه متن ورودی با استفاده از بردار *TF-IDF* نشان داده می‌شود. مفسر معنایی کلمه‌های متن را پیمایش می‌کند، سپس درایه‌های متناسب را از نمایه معکوس بازیابی کرده (برخی از درایه‌ها صفر بوده و یا مقدار بسیار کمی دارند به همین دلیل حذف می‌شوند) و آن‌ها را با یک بردار وزن‌دار از مفاهیمی که توسط متن ارایه می‌شود، ترکیب و ادغام می‌کند. (به ازای هر کلمه از متن یک بردار ساخته می‌شود پس از آن این بردارها با هم جمع می‌شوند تا بردار مربوط یک متن تولید شود). اگر $t=\{w_i\}$ متن ورودی باشد و $<v_i>$ بردار *TF-IDF* آن متن باشد که v_i وزن کلمه w_i می‌باشد. همچنین $<k_j>$ درایه از فهرست معکوس برای کلمه w_i می‌باشد که k_j شدت رابطه معنایی بین کلمه w_i و مفهوم C_j از ویکی‌پدیا را بیان می‌کند که $\{C_j \in C_1,\ldots,C_N\}$ می‌باشد و N تعداد همه مفاهیم موجود در ویکی‌پدیا است. بنابراین بردار تفسیر معنایی V برای متن T برداری به طول N خواهد بود که وزن هر مفهوم C_j از رابطه $\sum_{w_i \in T} V_i.K_j$ به‌دست می‌آید.

درایه های این بردار میزان ارتباط مفاهیم با متن را نشان می‌دهند. برای محاسبه رابطه معنایی دو متن، بردارهای آن‌ها با استفاده از معیار کوسینوسی مقایسه می‌شوند. روش *Graph-Based ESA* نیز از مشابه روش *ESA* می‌باشد اما برای افزایش دقت محاسبه ارتباط معنایی بین کلمات از گراف رده‌های ویکی‌پدیا[1] *(WCG)* نیز بهره می‌برد. در این مدل برای هر کلمه در جفت کلماتی که قرار است میزان ارتباط آن‌ها مشخص شود، k مرتبط‌ترین مفاهیم ویکی‌پدیا به کمک روش *ESA* بازگردانده می‌شود. در ادامه برای هر مفهوم موجود در هر دو مجموعه مفاهیم مرتبط، دسته‌های مرتبط با آن‌ها از گراف رده‌های ویکی‌پدیا استخراج می‌شود. در ادامه میزان ارتباط معنایی به کمک ضریب ارتباط بین مفاهیم در گراف رده‌های ویکی‌پدیا محاسبه خواهد شد.

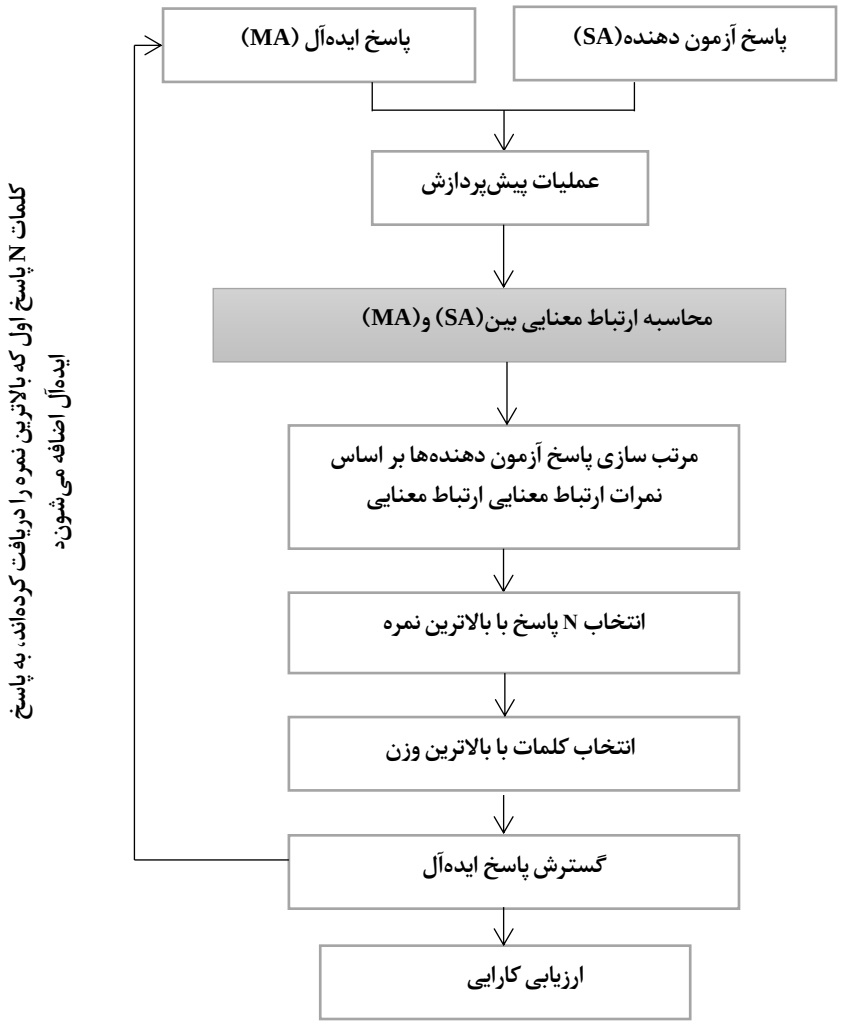
با اینکه مدل‌های *ESA* و *Graph-Based ESA* از دقت بالایی در محاسبه ارتباط معنایی بین متون برخوردارند، اما در پاسخ‌های تشریحی تاکید روی مفهوم پاسخ مطرح شده است و تنها یک پاسخ ایده‌آل برای مقایسه وجود دارد، ممکن است پاسخ یک دانش‌آموز با وجود درست بودن به دلیل این شباهت کمی که با پاسخ ایده‌آل دارد، به درستی نمره‌دهی نشده و نمره پایینی به آن اختصاص (سوزان و همکاران، ۲۰۲۰).

برای حل این مشکل، در مدل پیشنهادی این مقاله رویکرد نوین مبتنی بر بازخوردهای پاسخ‌های آزمون دهندگان معرفی است که از تکنیکی مشابه بازخورد شبه مرتبط[2] معرفی شده در سیستم‌های بازیابی اطلاعات استفاده می‌کند (رویتمن و کورلند[3]، ۲۰۱۹). بازخورد شبه مرتبط یکی از روش‌های بهبود نتایج موتور جستجو است که در آن با استخراج خودکار اطلاعات از نتیجه جستجوی قبلی، یک پرس وجو گسترش می‌یابد و دوباره جستجو انجام می‌شود. به طور کلی، در این روش در مرحله اول بازیابی به صورت نرمال انجام می‌شود و

[1] Wikipedia Category Graph
[2] Pseudo-Relevance Feedback
[3] Roitman and Kurland

N سند که بالاترین ارتباط را با پرسجوی مطرح شده دارند، به عنوان بازخورد مرتبط در نظر گرفته می‌شوند. در ادامه در بین این N سند تعداد مشخصی از کلمات که بالاترین وزن *TF-IDF* را دارند، به پرس‌وجوی اصلی اضافه شده (گسترش پرس‌وجو) و بازیابی بر اساس پرس وجوی جدید صورت می‌گیرد.

در مدل پیشنهادی از ایده بازخورد شبه مرتبط با این تفاوت که در اینجا هدف توسعه پاسخ ایده‌آل و کاربرد نمره‌دهی خودکار آزمون‌های تشریحی استریال استفاده شده است. با استفاده از این تکنیک و تفسیر پاسخ‌های دانش‌آموزان می‌توان کلمات موجود در پاسخ ایده‌آل را افزایش داد. به بیان دیگر می‌توان از پاسخ‌های درست برای گسترش پاسخ ایده‌آل استفاده کرد و پیرو آن دقت روش‌های محاسبه ارتباط معنایی در کاربر نمره‌دهی خودکار آزمون‌های تشریحی را افزایش داد. دیاگرام کلی رویکرد معرفی شده در **شکل ۲** نشان داده شده است. به طور خلاصه، پس از اینکه میزان ارتباط معنایی بین پاسخ آزمون‌دهنده و پاسخ ایده‌آل با استفاده از روش *ESA* محاسبه شد، نمره‌های بدست آمده به صورت نزولی مرتبط می‌شوند. سپس از کلمات N پاسخ اول که بالاترین وزن *TF-IDF* را دارند به پاسخ ایده‌آل اضافه خواهند شد و آن پاسخ‌های باقی‌مانده بر اساس پاسخ ایده‌آل جدید دوباره نمره‌دهی می‌شوند. در واقع نمرات بدست آمده از اجرای اول برای N پاسخی که بالاترین نمره را داشته‌اند تغییر نخواهد کرد (بدون بازخورد) اما سایر پاسخ‌ها بر اساس پاسخ ایده‌آل جدید دوباره نمره



شکل۲ - دیاگرام کلی رویکرد پیشنهادی.

Figure 2- Diagram of the proposed model.

دهی می‌شوند. با این‌کار نمره‌اصلی پاسخ‌هایی که در اجرای اول بالاترین نمره را داشته‌اند حفظ شده و تضمین می‌شود که هیچ کدام از پاسخ‌های باقی‌مانده که دوباره نمره‌دهی شده‌اند نمره‌ای بالاتر از آن‌ها بدست نیاورند. نتایج آزمایش‌ها نشان می‌دهد که استفاده از بازخورد مرتبط می‌تواند دقت روش‌ها تا به مقدار قابل توجهی افزایش دهد.

۴- پیاده سازی و نتایج

آزمایش‌های انجام شده را می‌توان به دو دسته کلی تقسیم کرد. در بخش اول آزمایش‌ها، روش‌های محاسبه ارتباط معنایی مبتنی‌بر منبع و مبتنی‌بر دانش که در بخش ۲ معرفی شدند در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی باهم مقایسه شده و کارایی آن‌ها مشخص می‌شود. در بخش دوم نیز تاثیر رویکرد معرفی شده در بخش ۳ روی دقت روش‌ها معرفی شده مورد آزمایش قرار گرفته است.

۴-۱- نحوه اجرا

به منظور ارزیابی جامع انواع مدل‌های معنایی در مساله نمره‌دهی خودکار آزمون‌های تشریحی، پیاده‌سازی جامعی از روش‌های معرفی شده در بخش ۲ صورت گرفته است. در این بخش، جزئیات پیاده‌سازی و پیکربندی انواع مدل‌های معنایی به تفکیک هر مدل آمده است. در شرایطی که پیکربندی‌های متعددی برای هر مدل وجود دارد، سعی شده است تا موثرترین پیکربندی با توجه به تحقیقات پیشین مورد ارزیابی قرار گیرد.

برای پیاده‌سازی روش‌های مبتنی بر دانش از پیاده‌سازی های مبتنی‌بر وردنت موجود در بسته *WordNet::Similarity* استفاده شده است (پدرسن و همکاران، ۲۰۰٤). برای پیاده سازی *LSA* نیز از بسته *Gensim* استفاده است (ریهورک و سوجکا[۱]، ۲۰۱۰). لازم به ذکر است که پیاده‌سازی *ESA* (گابریلویچ و مارکویچ، ۲۰۰۹)، *WikiRelate!* (استروب و پونزتو، ۲۰۰٦)، *WOLM* (ویتن و میلن، ۲۰۰۸)، *Graph-based ESA* (لی و همکاران، ۲۰۱۷) و *WCVM* (طیب و همکاران،۲۰۱۳) نیز بر اساس روش مطرح شده درمطالعات اصلی صورت گرفته است. لازم به ذکر است که نتایج بدست آمده توسط تمامی روش‌ها نرمال‌سازی شده‌اند تا در محدوده صفر تا یک قرار بگیرند. علاوه بر این از ویکی‌پدیا ۲۰۱٦ به عنوان منبع دانش‌پیش‌زمینه‌ای برای پیاده‌سازی روش‌های مبتنی‌بر منبع و از وردنت ۱/۲ برای پیاده‌سازی روش‌های مبتنی بر دانش استفاده شده است.

۴-۲- مجموعه داده

مجموعه داده مورد استفاده شامل سه تمرین به زبان انگلیسی است[۲] که هرکدام از این تمرین‌ها شامل هفت سوال با پاسخ تشریحی مبتنی بر متن است که توسط ۳۰ دانش‌آموز پاسخ دهی شده‌اند. در نتیجه مجموعه داده شامل ٦۳۰ پاسخ است. پاسخ‌ها به صورت مستقل توسط دو نفر در بازه صفر (کاملا نادرست) تا پنج (کاملا درست) نمره‌دهی شده‌اند که نمرات ارائه شده توسط آن‌ها دارای ضریب همبستگی $r=0.6443$ است (موهلرو میهالسیا ، ۲۰۰۹).

ارزیابی نتایج بدست آمده از آزمایش‌ها به وسیله تجزیه و تحلیل ریاضی و تعیین میزان همبستگی نتایج حاصل از آزمایش‌ها با قضاوت های انسانی صورت می‌گیرد با توجه به اینکه در حوزه نمره‌دهی خودکار آزمون‌های تشریحی مجموعه داده‌هایی وجود دارد که شامل یکسری سوال و پاسخ‌های ارائه شده توسط دانش‌آموزان مختلف است که توسط معلمان نمره دهی شده‌اند، می‌توان رویکردهای معرفی شده را روی این مجموعه داده اعمال کرد و با محاسبه ضریب همبستگی پیرسن مابین نتایج بدست آمده از آزمایش‌ها و قضاوت‌های انسانی دقت روش‌ها را در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی بدست آورد (ژانگ و همکاران[۳]، ۲۰۱۳).

۴-۳- روش ارزیابی

هدف روش‌های محاسبه ارتباط معنایی ساختن ماشین‌هایی است که بتوانند همانند انسان عمل کنند. به همین منظور رویکردهای محاسبه ارتباط معنایی در مقابل قضاوت‌های انسانی سنجیده می‌شوند. هرچه نتایج حاصل به قضاوت‌های انسانی نزدیک‌تر باشند، دقت روش بالاتر است. برای مقایسه رویکرد پیشنهادی با قضاوت‌های انسانی از ضریب همبستگی پیرسن[٤] استفاده شده است. به‌طورکلی می توان گفت ضریب همبستگی[٥] یک ابزار آماری برای تعیین نوع و درجه رابطه یک متغیر کمی با متغیر کمی دیگر است. ضریب همبستگی،

---

[۱] Rehurek and Sojka
[۲] https://web.eecs.umich.edu/~mihalcea/downloads.html#saga
[۳] Zhang et al.
[٤] Pearson
[٥] Correlation Coefficient

یکی از معیارهای مورد استفاده در تعیین همبستگی دو متغیر است. ضریب همبستگی شدت رابطه و همچنین نوع رابطه (مستقیم یا معکوس) را نشان می‌دهد.

ضریب همبستگی پیرسن برای محاسبه همبستگی دو متغیر فاصله‌ای یا نسبی به کار برده می شود. مقدار ضریب همبستگی همواره بین ۱+ و ۱- است. اگر مقدار بدست آمده مثبت باشد به معنی این است که تغییرات دو متغیر به طور هم جهت اتفاق می‌افتد یعنی با افزایش در هر متغیر، متغیر دیگر نیز افزایش می‌یابد و برعکس اگر مقدار r منفی شد یعنی اینکه دو متغیر در جهت عکس هم عمل می‌کنند یعنی با افزایش مقدار یک متغیر مقادیر متغیر دیگر کاهش می یابد و برعکس. اگر مقدار بدست آمده صفر شد نشان می‌دهد که هیچ رابطه‌ای بین دو متغیر وجود ندارد و اگر ۱+ شد همبستگی مثبت کامل و اگر ۱- شد همبستگی کامل و منفی است.

۴-۴- نتایج آزمایش‌ها

در این بخش، آزمایش‌های جامعی به منظور ارزیابی روش‌های محاسبه ارتباط معنایی در مساله نمره‌دهی خودکار آزمون‌های تشریحی صورت گرفته است. در **جدول ۱** میزان همبستگی مقادیر ارتباط معنایی بدست آمده توسط هر مدل با مقادیر تعیین شده توسط عوامل انسانی برمبنای معیار همبستگی پیرسن *(r)* نشان داده شده است. نتایج آزمایش‌ها حاکی از آن است کارایی روش‌های مبتنی‌بر منبع به طور محسوسی بهتر از روش‌های مبتی‌بر دانش در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی است. به طوریکه میانگین همبستگی روش‌های مبتنی‌بر منبع با قضاوت‌های انسانی برابر $r=0.55$ است که در مقایسه با میانگین همبستگی روش‌های مبتنی‌بر دانش با قضاوت‌های انسانی $(r=0.33)$ بسیار بهتر است. در واقع بهترین روش مبتنی بر دانش دارای همبستگی $r=0.451$ و بهترین روش مبتنی بر منبع دارای همبستگی $r=0.651$ با قضاوت‌های انسانی است. کمترین مقدار مربوط به روش *HSO* $(r=0.196)$ و بیشترین مقدار مربوط به روش *Graph based ESA* $(r=0.651)$ است. البته لازم به ذکر است که روش‌های مبتنی بر دانش به منابع دانش از پیش‌ساخته‌ای برای اجرای مدل نیاز دارند که ساخت این منبع بسیار زمان‌بر هزینه‌بر و نیازمند نیروی انسانی است در حالیکه روش‌های مبتنی بر منبع از مجموعه‌ای از نوشته‌جات استفاده می‌کنند که هیچ محدودیتی در دامنه ندارند و به عنوان یک منبع دانش همه‌جانبه به حساب می‌آیند.

جدول ۱- مقایسه روش‌های مبتنی بر دانش و مبتنی بر منبع در کاربرد تصحیح خودکار آزمون‌های تشریحی.
Table 1- Comparision of knowledge based and corpus based methods for automatic short answer grading.

ضریب همبستگی پیرسن	الگوریتم	نوع
0.451*	Path (پدرسن و همکاران، ۲۰۰۴)	روش‌های مبتنی بر دانش
0.223	LCH (لی‌چاک و چودرو، ۱۹۹۸)	
0.363	Lesk (لسک، ۱۹۸۶)	
0.336	WuP(وو و پالمر، ۱۹۹۴)	
0.252	Resnik (رسینک، ۱۹۹۵)	
0.391	Lin (لین، ۱۹۹۸)	
0.449	JCN (جیانگ و کونراث، ۱۹۹۷)	
0.196	HSO (هیرست و اسیانج، ۱۹۹۸)	
0.382	Vector (پتوازداهان و پترسن، ۲۰۰۶)	روش‌های مبتنی بر منبع
0.502	ESA (گابریلویچ و مارکویچ، ۲۰۰۹)	
0.438	LSA (دومیس،۲۰۰۴)	
0.511	WikiRelate! (استروب و پونزتو، ۲۰۰۶)	
0.562	WOLM (ویتن و میلن، ۲۰۰۸)	
0.601	WCVM (tf) (طیب و همکاران، ۲۰۱۳)	
0.613	WCVM (tf-icf) (طیب و همکاران، ۲۰۱۳)	
0.651*	Graph-based ESA (لی و همکاران، ۲۰۱۷)	

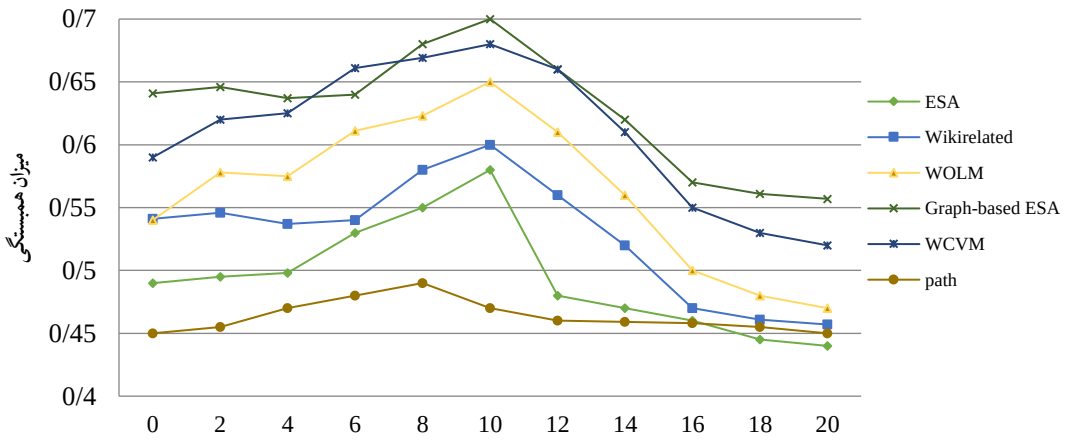
بخش دوم شامل آزمایش‌های صورت گرفته به منظور بررسی میزان تاثیر رویکرد معرفی شده در بخش ۳ روی دقت الگوریتم‌های محاسبه ارتباط معنایی متون است. **جدول ۲** نتایج حاصل از استفاده از تکنیک بازخورد مرتبط به منظور افزایش دقت الگوریتم محاسبه ارتباط معنایی در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی را نشان می‌دهد.

همان‌طور که مشخص است استفاده از این رویکرد منجر به افزایش دقت تمامی الگوریتم‌ها شده است (در این مرحله از رویکرد *Path* به عنوان نماینده روش‌های مبتنی بر دانش به علت دقت بالا استفاده شده است). لازم به ذکر است که این نتایج بعد از ٦ بار تکرار الگوریتم و استفاده از بازخوردهای اتوماتیک بدست آمده‌اند. همان‌طور که مشخص است استفاده از بازخورد دقت روش محاسبه ارتباط معنایی را در حوزه تصحیح خودکار متون به طور قابل توجهی افزایش داده است.

شکل ۳ تاثیر بازخورد خودکار را بر اساس اندازه *N* نشان می‌دهد. این نمودار بر اساس تعداد پاسخ‌های با نمره بالا استفاده شده در بازخورد *(N)* و میزان ضریب همبستگی ترسیم شده است. همان‌طور که مشخص است، افزایش بیش از حد مقدار *N* (بالای ۱۰) باعث کاهش دقت الگوریتم‌ها می‌گردد. نمودار نشان می‌دهد که تاثیر بازخورد خودکار مرتبط روی الگوریتم *Graph based ESA* نسبت به سایر روش‌ها بیشتر است و استفاده از بازخورد خودکار منجر به افزایش قابل توجه دقت الگوریتم *Graph based ESA* در کاربرد ارزیابی خودکار آزمون‌های تشریحی می‌شود.

جدول ۲- نتایج بدست آمده با استفاده از رویکرد بازخورد مرتبط.
Table 2- Results obtained using a related feedback approach.

الگوریتم	ضریب همبستگی پیرسن
Path (پتوازداهان و پترسن، ۲۰۰۶)	0.488
ESA (گابریلویچ و مارکویچ، ۲۰۰۹)	0.581
LSA (دومیس،۲۰۰۴)	0.539
WikiRelate! (استروب و پونزتو، ۲۰۰۶)	0.581
WOLM (ویتن و میلن، ۲۰۰۸)	0.593
WCVM (tf) (طیب و همکاران، ۲۰۱۳)	0.634
WCVM (tf-icf) (طیب و همکاران، ۲۰۱۳)	0.683
Graph-based ESA (لی و همکاران، ۲۰۱۷)	0.731



شکل ۳- تاثیر بازخورد خودکار روی عملکرد الگوریتم‌های محاسبه ارتباط معنایی.
Figure 3- The effect of automatic feedback on the performance of semantic relatedness algorithms.

۵- بحث و نتیجه‌گیری

در این مقاله رویکردهای مختلف محاسبه ارتباط معنایی در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی مورد بررسی قرار گرفتند. به همین منظور در ابتدا روش‌های مختلف محاسبه ارتباط معنایی به دو دسته مبتنی بر دانش و مبتنی بر منبع تقسیم شده و آزمایش‌های جامعی به منظور مقایسه کارایی این روش‌ها در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی انجام شد. نتایج نشان می‌دهد که روش‌های

مبتنی بر منبع در مقایسه با روش‌های مبتنی بر دانش از دقت بالاتری در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی برخوردار بوده و با محدودیت‌هایی کمتری در مسایل دنیای واقعی مواجه هستند. به بیان دیگر، روش‌های مبتنی بر دانش به منبع دانش پیش‌زمینه‌ای از پیش‌ساخته‌ای نیاز دارند که ایجاد آن وابسته به نیروی انسانی بوده و دامنه و زبان آن محدود است، در مقابل روش‌های مبتنی بر منبع مستقل به زبان نبوده قابلیت اعمال بیشتری در مسائل دنیای واقعی دارند. در ادامه نیز یک تکنیک جدید برای ترکیب جواب پاسخ‌دهنده با بالاترین نمره با پاسخ ایده‌آل به منظور افزایش دقت روش‌های محاسبه ارتباط معنایی در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی معرفی شد. رویکرد پیشنهادی همانند رویکرد بازخورد شبه مرتبط در بازیابی اطلاعات عمل می‌کند و با گسترش پاسخ ایده‌آل به وسیله کلمات پاسخ‌های که بالاترین نمره را دریافت کرده‌اند منجر به افزایش قابل توجه دقت روش‌های محاسبه ارتباط معنایی در کاربرد تصحیح خودکار آزمون‌های تشریحی می‌شود. نتایج آزمایش‌ها نشان می‌دهد که استفاده از بازخورد خودکار روی روش *Graph based ESA* تاثیرگذارتر بوده و می‌تواند دقت این الگوریتم را تا اندازه ۰/۷۳۱ افزایش دهد.

در کارهای آینده نیز می‌توان به منظور افزایش دقت روش‌های محاسبه ارتباط معنایی در کاربرد نمره‌دهی خودکار آزمون‌های تشریحی آن‌ها را با تکنیک‌های یادگیری ماشین ترکیب کرد و از مزایایی‌آن‌ها به صورت توامان بهره برد. علاوه بر این می‌توان با استفاده از منبع دانش پیش‌زمینه‌ای مناسب و با محدوده خاص، سیستمی طراحی را کرد که بتواند عملیات نمره‌دهی خودکار آزمون‌های تشریحی را در زبان فارسی و یا زبان‌های دیگر انجام دهد.

منابع

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, *32*(1), 13-47. https://doi.org/10.1162/coli.2006.32.1.13

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, *25*(1), 60-117.https://doi.org/10.1007/s40593-014-0026-8

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, *38*(1), 188-230. https://doi.org/10.1002/aris.1440380105

Filighera, A., Steuer, T., & Rensing, C. (2020, July). Fooling automatic short answer grading systems. *International conference on artificial intelligence in education* (pp. 177-190). Cham: Springer. https://doi.org/10.1007/978-3-030-52237-7_15

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of artificial intelligence research*, *34*, 443-498. DOI: https://doi.org/10.1613/jair.2669

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, *305*, 305-332.

Jarmasz, M., & Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In N. Nicolov., K. Bontcheva., G. Angelova., & R. Mitkov (Eds.), *Recent advances in natural language processing iii*.John Benjamins Publishing Co.

Jarmasz, M., & Szpakowicz, S. (2012). Roget's Thesaurus and semantic similarity. *Proceedings of conference on recent advances in natural language processing*. arXiv:1204.0245

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, *49*(2), 265-283.

Lee, Y. Y., Ke, H., Yen, T. Y., Huang, H. H., & Chen, H. H. (2020). Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. *Journal of the association for information science and technology*, *71*(6), 657-670. https://doi.org/10.1002/asi.24289

Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on systems documentation*, 24-26. https://doi.org/10.1145/318723.318728

Li, P., Xiao, B., Ma, W., Jiang, Y., & Zhang, Z. (2017). A graph-based semantic relatedness assessment method combining wikipedia features. *Engineering applications of artificial intelligence*, *65*, 268-281. https://doi.org/10.1016/j.engappai.2017.07.027

Lin, D. (1998). An information-theoretic definition of similarity. *The fifteenth international conference on machine learning* (pp. 296-304). https://dl.acm.org/doi/10.5555/645527.657297

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st national conference on artificial intelligence* (pp. 775-780). https://dl.acm.org/doi/10.5555/1597538.1597662

Mohler, M., & Mihalcea, R. (2009, March). Text-to-text semantic similarity for automatic short answer grading. *The 12th conference of the european chapter of the ACL* (pp. 567-575). Athens, Greece. DOI: 10.3115/1609067.1609130

Mohler, M., Bunescu, R., & Mihalcea, R. (2011, June). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *The 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 752-762). https://dl.acm.org/doi/10.5555/2002472.2002568

Nazari Soleimandarabi, M., Mirroshandel, S. A., & Sadr, H. (2015a). The significance of semantic relatedness and similarity measures in geographic information science. *International journal of computer science and network solutions*, *3*(2), 12-23.

Nazari Soleimandarabi, M., Mirroshandel, S. A., & Sadr, H. (2015b). A Survey of semantic relatedness measures. *International journal of computer science and network solutions*, 3(2), 1-11 .

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *The workshop on making sense of sense: bringing psycholinguistics and computational linguistics together*.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, July). WordNet:: similarity-measuring the relatedness of concepts. *HLT-NAACL--Demonstrations '04: Demonstration Papers at HLT-NAACL* 2004 (pp. 38–41). Association for Computational Linguistics.

Peinelt, N., Nguyen, D., & Liakata, M. (2020, July). tBERT: Topic models and BERT joining forces for semantic similarity detection. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7047-7055).

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of LREC 2010 workshop new challenges for NLP frameworks*. Valletta, Malta: University of Malta. https://is.muni.cz/publication/884893/en/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007

Roitman, H., & Kurland, O. (2019, July). Query performance prediction for pseudo-feedback-based retrieval. *The 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 1261-1264). https://doi.org/10.1145/3331184.3331369

Roy, S., Rajkumar, A., & Narahari, Y. (2018). Selection of automatic short answer grading techniques using contextual bandits for different evaluation measures. *International journal of advances in engineering sciences and applied mathematics*, *10*(1), 105-113. https://doi.org/10.1007/s12572-017-0202-9

Sadr, H., & Nazari Solimandarabi, M. (2019). Presentation of an efficient automatic short answer grading model based on combination of pseudo relevance feedback and semantic relatedness measures. *Journal of advances in computer research*, *10*(2), 17-30.

Sadr, H., Nazari, M., Pedram, M. M., & Teshnehlab, M. (2019a). Exploring the efficiency of topic-based models in computing semantic relatedness of geographic terms. *International journal of web research*, *2*(2), 23-35.

Sadr, H., Pedram, M. M., & Teshnehlab, M. (2019c). A robust sentiment analysis method based on sequential combination of convolutional and recursive neural networks. *Neural processing letters*, *50*(3), 2745-2761. https://doi.org/10.1007/s11063-019-10049-1

Sadr, H., Pedram, M. M., & Teshnehlab, M. (2020). Multi-view deep network: A deep model based on learning features from heterogeneous neural networks for sentiment analysis. *IEEE access*, *8*, 86984-86997. DOI: 10.1109/ACCESS.2020.2992063

Sadr, H., Pedram, M. M., & Teshnehlab, M. (2021). Convolutional neural network equipped with attention mechanism and transfer learning for enhancing performance of sentiment analysis. *Journal of AI and data mining*. *9*(2), 141-151. DOI:10.22044/jadm.2021.9618.2100

Sadr, H., Pedram, M. M., & Teshnelab, M. (2019b). Improving the performance of text sentiment analysis using deep convolutional neural network integrated with hierarchical attention layer. *International journal of information and communication technology research*, *11*(3), 57-67. http://ijict.itrc.ac.ir/article-1-416-en.html

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Strube, M., & Ponzetto, S. P. (2006, July). WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI'06 Proceedings of the 21st national conference on Artificial intelligence* (pp. 1419-1424). https://dl.acm.org/doi/10.5555/1597348.1597414

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia computer science*, *169*, 726-743. https://doi.org/10.1016/j.procs.2020.02.171

Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2013). Computing semantic relatedness using Wikipedia features. *Knowledge-based systems*, *50*, 260-278. https://doi.org/10.1016/j.knosys.2013.06.015

Taieb, M. A. H., Zesch, T., & Aouicha, M. B. (2020). A survey of semantic relatedness evaluation datasets and procedures. *Artificial intelligence review*, *53*(6), 4407-4448. https://doi.org/10.1007/s10462-019-09796-3

Witten, I. H., & Milne, D. N. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *The 32nd annual meeting on association for computational linguistics*. arXiv preprint cmp-lg/9406033.

Young, J. R. (2012). Inside the Coursera contract: How an upstart company might profit from free courses. *The chronicle of higher education*, *19*(07), 2012.

Zesch, T., & Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words. *Natural language engineering*, *16*(1), 25-59.

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments*, 1-14. https://doi.org/10.1080/10494820.2019.1648300

Zhang, Y., Lin, C., & Chi, M. (2020). Going deeper: Automatic short-answer grading by combining student and question models. *User modeling and user-adapted interaction*, *30*(1), 51-80. https://doi.org/10.1007/s11257-019-09251-6

Zhang, Z., Gentile, A. L., & Ciravegna, F. (2013). Recent advances in methods of lexical semantic relatedness–a survey. *Natural language engineering*, *19*(4), 411-479.

Zhu, X., Guo, Q., Zhang, B., & Li, F. (2019). An efficient approach for measuring semantic relatedness using Wikipedia bidirectional links. *Applied intelligence*, *49*(10), 3708-3730. https://doi.org/10.1007/s10489-019-01452-1