کاربرد تحلیل گفتمان در خلاصه سازی خودکار متون علمی [1]

پگاه تاجر، عبدالرسول جوکار، سید مصطفی فخراحمد، علیرضا خرمایی، هاجر ستوده

چکیده

هدف این پژوهش آشکارسازی کاربرد ویژه تحلیل گفتمان در خلاصه‌سازی خودکار متون علمی است و به این منظور از رویکرد مرور و تحلیل نوشتارها استفاده شده است. متن یک واحد زبانشناسی پیچیده است و ساختار گفتمانی و تئوری های سازماندهی متن را می توان شاخصی برای تفسیر متن و انتخاب جمله مناسب به منظور تولید خودکار خلاصه های خواناتر و منسجم تر در نظر گرفت. نشانه گذاری بلاغی اطلاعات، برای فرایندهای جدید دسترسی به اطلاعات علمی مفید می باشد. استراتژی های ویژه خلاصه‌سازی خودکار مقالات علمی عموما تلفیقی از روش های سطحی و معنایی هستند. یکی از استراتژی هایی که از تجزیه و تحلیل کلام بهره می برد عبارت است از ناحیه بندی استدلالی. بر اساس این استراتژی طرح های حاشیه نویسی گفتمانی ویژه متون علمی به وجود آمده اند که بر اساس آنها می توان ابعاد چهارگانه حل مسئله، تخصیص فکری، استدلال علمی و نگرش به آثار دیگران را در متون علمی معلوم کرد و بر غنای خلاصه

کلیدواژه‌ها: تحلیل گفتمان، ژانر علمی، خلاصه سازی خودکار متن، خلاصه سازی گفتمان مدار، خلاصه سازی استنادی، ناحیه بندی استدلالی

۱. برگرفته از رساله دکتری

های خودکار افزود. مشهورترین طرح حاشیه نویسی نواحی بلاغی مقالات علمی، طرح تویفل می باشد. از کاربردهای مهم و ویژه تحلیل گفتمان در خلاصه سازی خودکار ژانر علمی، تولید خلاصه های استنادی گفتمان مدار می باشد که بر اساس آن می توان انواع خلاصه های خودکار مبتنی بر پرسش کاربر و خلاصه های چندسندی ای که تضادها و شباهت های مدارک در خلاصه نهایی مشهود است را تولید نمود.

۱- مقدمه

رشد فزاینده اطلاعات متنی در محیط وب و نیاز به دسترسی، ارزیابی و انتخاب سریع آنها، منجر به پدیدآمدن ابزارهایی برای خلاصه‌سازی خودکار متن شده‌است. این ابزارها، در ژانرهای مختلف از جمله خبری، حقوقی و علمی و نیز در خلاصه‌سازی صفحات وب به منظور نمایش در تلفن همراه و سیستم‌های بازیابی اطلاعات، کاربردهای فراوانی دارند.

خلاصه سازی خودکار متون از تکنیک های پردازش زبان طبیعی است که همواره مورد توجه زبان شناسان رایانشی و متخصصان بازیابی اطلاعات بوده است. خلاصه‌سازی خودکار متن عبارت است از نمایش فشرده، دقیق و منسجم متن ورودی به طوری که متن خروجی مفاهیم مهم متن ورودی را در بر داشته باشد (میبوری [۱]، ۱۹۹۵ نقل در رحمان و بورا[۲]، ۲۰۱۵). پژوهش در حوزه خلاصه‌سازی خودکار، در اواخر دهه پنجاه میلادی، زمانی که تمایل زیادی به خودکارسازی چکیده نویسی اسناد فنی وجود داشت، مورد توجه جامعه علمی قرار گرفت (لوهن[۳] ،۱۹۵۸) و با به کار گیری هوش مصنوعی به طور چشمگیری افزایش یافت (دیجانگ[۴]، ۱۹۸۲). از دهه نود، با ظهور کنفرانسهای علمی متعدد مرتبط با خلاصه‌سازی خودکار، تمایل به پژوهش در میان پژوهشگران این حوزه رنگ و بوی تازه ای به خود گرفت که نقطه اوج آن را می‌توان توسعه چارچوب‌های ارزیابی از سال ۲۰۰۰ در آمریکا دانست. دو برنامه ارزیابی مهم عبارتند از کنفرانس های درک سند (دی. یو. سی.)[۵] و کنفرانس های تحلیل متن (تی. ای. سی.)[۶] .

چگونگی انتخاب محتوای ضروری یک سند و نحوه بیان محتوای انتخاب شده به صورتی فشرده، دو عنصر اساسی فرآیند خلاصه‌سازی متن است (جونز[۷]، ۱۹۹۳؛ جونز و اندرس- نیگمیر[۸] ، ۱۹۹۵). به طور کلی، خلاصه‌های خودکار به دو روش استخراجی[۹] و چکیده‌سازی[۱۰] تولید می‌شود. خلاصه‌های استخراجی

۱ Maybury
۲ Borah
۳ Luhn
۴ DeJong
۵ Document Understanding Conference (DUC)
۶ Text Analysis Conference (TAC)
۷ Jones
۸ Endres-Niggemeyer
۹ Extractive
۱۰ Abstractive

شامل مجموعه‌ای از جملات سند ورودی هستند. این نوع خلاصه‌سازی یک روش ساده اما قوی برای خلاصه‌سازی متن است. لذا بسیاری از پژوهش‌های این حوزه، از نوع استخراجی هستند. استخراج و انتخاب جملات مناسب برای خلاصه نهایی، بر اساس ویژگی‌های متون صورت می‌پذیرند. این ویژگی‌ها، طیف وسیعی از ویژگی‌های سطحی مانند فراوانی کلمات و محل قرارگرفتن جمله در متن تا ویژگی‌های معنایی و گفتمانی را شامل می‌شوند (سیگون[1] و پوییبیو[2]، ۲۰۱۳). خلاصه‌های چکیده‌ای، تفسیری از متن اصلی را ارائه می‌کنند و در تولید آنها مفاهیم جملات اصلی به شکل کوتاه‌تر بازنویسی می‌شوند (مانی[3]، ۲۰۰۱). خلاصه‌سازی از منظر تعداد اسناد، می‌تواند تک‌سندی[4] یا چندسندی[5] باشد. در خلاصه‌سازهای تک‌سندی، ورودی تنها یک سند است. این در حالی است که خلاصه‌سازی چندسندی بازنمون مختصر و منسجمی از محتوای کلیدی مجموعه‌ای از مدارک مربوط به یکدیگر است که به صورت خودکار تولید شده باشد. (سیگون و پوییبیو، ۲۰۱۳).

از آنجایی که متن یک واحد زبان شناسی پیچیده است، ساختار گفتمانی و یا تئوری های سازماندهی متن را می‌توان شاخصی برای تفسیر متن و انتخاب جمله مناسب به منظور تولید خودکار خلاصه های خواناتر و منسجم تر در نظر گرفت.

۲- بیان مسأله

بهترین منابع باید در مناسب‌ترین قالب و در کمترین زمان در دسترس محققان قرار گیرد. با توجه به رشد روزافزون انتشارت علمی، خلاصه‌سازی خودکار متون، راهکار مناسبی برای تسهیل و تسریع فرآیند درک محتوای اسناد علمی است.

امروزه، خلاصه‌سازی خودکار چندسندی متون علمی اجتناب‌ناپذیر است. پایگاه‌های اطلاعات علمی علاوه بر متن اصلی، چکیده‌های نویسنده را در اختیار کاربر قرار می‌دهند اما در مواجه با انبوهی از مقالات، یک سیستم خلاصه‌ساز چندسندی می‌تواند فشرده‌ای از اهداف، روش‌ها و نتایج مجموعه‌ای از مدارک پژوهشی را در اختیار محقق قراردهد و او را از صرف وقت زیاد برای بررسی انبوهی از چکیده مقالات و یا متن کامل آنها به ویژه در مراحل آغازین پژوهش مانند انتخاب موضوع و مرور پیشینه‌ها تا حد زیادی بی‌نیاز کند.

همانطور که در مقدمه گفته شد، فرآیند خلاصه‌سازی چندسندی دشوارتر از نوع تک‌سندی است و با وجود ارائه راهکارهای مختلف در سال‌های گذشته، همچنان چالش‌هایی وجود دارد. بیشترین تلاش‌ها برای رفع معضلات خلاصه‌سازی چندسندی در ژانر خبری انجام شده اند که بعضا به نتایج

۱ Saggion
۲ Poibeau
۳ Mani
٤ Single document
٥ Multi-document

قابل قبولی رسیده‌اند. با این حال، با توجه به اینکه متون خبری با متون علمی ماهیتا تفاوت دارند، نتایج به دست آمده در ژانر خبری را نمی‌توان به ژانر علمی تعمیم داد. با توجه به این مهم، مسأله این پژوهش آشکارسازی کاربرد ویژه تحلیل گفتمان در خلاصه‌سازی خودکار متون علمی است و به این منظور از رویکرد مرور و تحلیل نوشتارها استفاده شده است.

۳- اهمیت پژوهش

تولید خلاصه‌های خودکار باکیفیت مطلوب همواره از چالش‌های خلاصه‌سازی خودکار متن بوده‌است. به همین دلیل، خلاصه‌سازی خودکار متن را از جمله مباحث پیچیده پردازش زبان طبیعی می‌دانند که علی رغم پیشرفتها، همواره مسائل زیادی را پیش روی جامعه علمی قرار می‌دهد (سیگون و پویبیو، ۲۰۱۳). این در حالی است که اکثر پژوهش‌های خلاصه‌سازی خودکار بر مقالات خبری تمرکز دارند و به دیگر ژانرها از جمله متون علمی توجه کمتری شده‌است (تویفل[۱] و موئنز[۲] ، ۲۰۰۲).

از طرف دیگر، اطمینان از پیوستگی متن خلاصه های تولید شده به صورت خودکار کار دشواری است زیرا نه تنها به درک محتوای هر قطعه متنی بلکه به دانش ساختار گفتمانی متن هم نیازمند است (رادو[۳]، هاوی[۴] و مک کئون[۵] ، ۲۰۰۲). بنابراین آشکارسازی کاربرد های ویژه تحلیل گفتمان در خلاصه‌سازی ژانر علمی می تواند منجر به اکتشاف و ارائه الگوریتم های نوین خلاصه‌سازی خودکار متون علمی گردد.

۴- پیشینه پژوهش

خلاصه‌سازی خودکار متن از حوزه هایی است که همواره مورد توجه پژوهشگران بوده و در ۲۵ سال گذشته نیز انواع رویکردهای خلاصه‌سازی ارائه شده‌است. تعدادی از این رویکرد ها بر تخمین توزیع موضوع/محتوای خلاصه نهایی از طریق تکنیک هایی چون تحلیل معنایی پنهان[۶]، مدل سازی موضوعی[۷] و مدل های بیزی[۸] تمرکز کرده‌اند. برخی دیگر، خلاصه‌سازی را یک مسئله بهینه سازی دانسته و عده ای دیگر با اتخاذ رویکرد طبقه‌بندی نظارتی سعی کرده‌اند راه حل هایی

۱ Teufel
۲ Moens
۳ Radev
٤ Hovy
٥ McKeown
٦ Latent Semantic Analysis (LSA)
۷ Topic Modelling
۸ Bayesian models

برای خلاصه‌سازی خودکار ارائه دهند. مدل های گراف‌مدار، رویکرد دیگری است که در ادبیات خلاصه‌سازی متن، بسیار موردتوجه قرارگرفته است. در این‌گونه مدل‌ها هدف اصلی، یافتن مرکزی ترین جمله در گراف شباهت جملات یک سند می‌باشد. خلاصه‌ساز هایی چون «لکس رنک»[۱] و «تکس رنک»[۲] از این رویکرد بهره می‌برند. علاوه بر این‌ها، حداکثرسازی تازگی و پرهیز از افزونگی در یک خلاصه سیستمی با بهره گیری از الگوریتم انتخاب حریصانه[۳] راهکار دیگری است که بررسی شده است. از جمله ایده هایی که اخیرا موردتوجه بوده‌است، بررسی ساختار بلاغی مدارک بر اساس تئوری «آر. اس. تی.» برای تجزیه گفتمانی مدارک می‌باشد (کهن و گوهریان، ۲۰۱۵).

در ادبیات پردازش زبان طبیعی، مطالعات تویفل و موئنز در سال ۲۰۰۲ را می‌توان اولین پژوهش خلاصه‌سازی خودکار ژانر علمی دانست. اما ریشه های خلاصه‌سازی خودکار متون علمی را می‌توان در ادبیات علم اطلاعات و دانش شناسی و مباحث چکیده نویسی خودکار سال های دهه پنجاه قرن بیستم جست.

تمایل به خلاصه‌سازی خودکار متون علمی از زمانی آغازشد که کتاب ها و مقالات علمی به صورت دیجیتالی ذخیره شدند. نمونه های اولیه چنین تمایلاتی را در دهه پنجاه میلادی و در پژوهش‌هایی که به چکیده نویسی و نمایه سازی خودکار مقالات و کتاب های علمی می پرداختند، می‌توان دید (لوهن، ۱۹۵۸ ؛ بکسندال[۴]، ۱۹۵۸). با این وجود، جامعه پژوهشی خلاصه‌سازی خودکار، از خلاصه‌سازی مقالات علمی به خلاصه‌سازی مقالات خبری روی آورد و این روند تا دوباره فعال شدن خلاصه‌سازی خودکار ژانر علمی در اواخر دهه نود میلادی، همچنان ادامه داشت (یلوگلو[۵] و همکاران، ۲۰۱۱). رویکرد خلاصه‌سازی پیکره‌بنیاد، در سال ۱۹۹۵ توسط کوپیک[۶] و همکارانش در هشتمین کنفرانس بین المللی سالانه «ای. سی. ام. سیگیر»[۷] پیشنهاد شد (کوپیک و همکاران، ۱۹۹۵). لازم به ذکر است که این کنفرانس به موضوع پژوهش و توسعه در بازیابی اطلاعات می پرداخت.

مرور ادبیات خلاصه‌سازی نشان می‌دهد که با وجود فعال شدن این حوزه از دهه نود، تعداد پژوهش‌های آن در مقایسه با پژوهش‌های خلاصه‌سازی خودکار متون خبری اندک است. اغلب پژوهش‌های خلاصه‌سازی خودکار مقالات علمی، به خلاصه‌سازی تک‌سندی مقالات پرداخته اند

۱ LexRank
۲ TextRank
۳ Greedy selection

این الگوریتم، بهترین انتخاب را با توجه به شرایط مسئله انجام می‌دهد به امید آنکه با ادامه همین روش بهینه سازی انجام شود.

٤ Baxendale
٥ Yeloglu
٦ Kupiec
٧ ACM SIGIR

و از روش‌های یادگیری ماشینی بهره برده اند. پژوهشگرانی چون کوپیک و همکاران (۱۹۹۵)، تویفل و موئنز (۲۰۰۲) یادگیری ماشینی نظارتی و برخی دیگر مانند قزوینیان و رادو (۲۰۰۸) و (۲۰۱۰)، الکیس[۱] و همکاران (۲۰۰۸) و ابوجبارا[۲] و رادو (۲۰۱۱) رویکرد غیرنظارتی را به کار گرفته اند. در این میان پژوهشگرانی هم هستند که خلاصه‌ساز های چندمنظوره برای ژانرهای مختلف و با اهداف مختلف، طراحی کرده‌اند که بعضا ژانر علمی را هم شامل می‌شود. برای مثال، لورت[۳] و پالومار[۴](۲۰۱۲) در پژوهشی با عنوان «کامپندیوم[۵] : یک ابزار خلاصه‌سازی متن برای تولید خلاصه‌های چندمنظوره از ژانرها و حوزه های گوناگون»، یک سامانه خلاصه‌ساز طراحی کردند که قادراست متداول‌ترین انواع خلاصه‌ها از جمله تک چندی، چندسندی، استخراجی، چکیده ای، پرسش‌مدار و مبتنی بر احساس[۶] را به صورت خودکار تولیدکند.

۵- چارچوب نظری پژوهش

در این بخش ضمن تعریف اجمالی گفتمان و تحلیل آن، مباحث نظری خلاصه‌سازی گفتمان مدار که اساس نظری این مطالعه را تشکیل می دهد، تشریح می گردد.

۵-۱- تعریف گفتمان

چادرون[۷] و ریچاردز[۸] (۱۹۸۵)، دو تعریف برای گفتمان ارائه کرده اند که عبارتند از: الف. گفتمان اصطلاحی است عام برای نمونه های کاربرد زبان؛ یعنی زبان که برای برقراری ارتباط تولید شده است، ب. بر خلاف دستور زبان که با عبارت ها و جمله ها سر و کار دارد، گفتمان به واحدهای زبانی بزرگ تر چون بند، مصاحبه، مکالمه و متن نظر دارد (بشیری، ۱۳۹۴).

براون[۹] و یول[۱۰] (۱۹۸۳) معتقدند، گفتمان عبارت است از تجزیه و تحلیل زبان در کاربرد. بنابراین گفتمان نمی تواند به توصیف صورت های زبانی، جدا از اهداف و نقش هایی که این صورت ها برای پرداختن به آنها در امور انسانی به وجود آمده اند، بپردازد.

---

۱ Elkiss
۲ Abu-Jbara
۳ Lloret
۴ Palomar
۵ COMPENDIUM
۶ Sentiment-based
۷ Chaudron
۸ Richards
۹ Brown
۱۰ Yule

تحلیل گفتمان برچسبی است برای آن دسته از تحلیل های متون که منبعث از نظریه ها و روال های زبان شناسی هستند. تحلیل گفتمان در وهله نخست به شرایط خوش ساختگی (پیوستگی و انسجام) و قواعد قیاسی می پردازد (البرزی ورکی، ۱۳۸۶).

تحلیل گفتمان، رویکردی میان رشته ای است که ریشه در زبان شناسی دارد. هدف عمده تحلیل گفتمان این است که تکنیک و روش جدیدی را در مطالعه متون، رسانه ها، فرهنگ ها، علوم، سیاست، اجتماع و مواردی مانند آن به دست دهد. تحلیل گفتمان جزء روش های تحقیق کیفی بوده که جهت کشف معنای به کار رفته در متن یا سخن به کار می رود (کلانتری و دیگران، ۱۳۸۸).

۵-۲-نشانگرهای گفتمانی

اولین تلاش ها برای معرفی نشانگرهای گفتمانی به عنوان یک موجودیت زبانی توسط لابو[۱] و فانشل[۲] در سال ۱۹۷۷ م. صورت گرفت. این دو، نشانگر گفتمانی را موجودیتی می دانستند که به موضوع[۳] دانش توزیع شده بین افراد شرکت کننده در بحث ارجاع می دهد (فراسر[۴]، ۱۹۹۹).

نشانگرهای گفتمانی وابسته به بافت هستند و بر اساس نگرش ها، شرکت کنندگان و متن طبقه بندی می شوند. بنابراین دارای کارکردهای بلاغی هم در سطح متن و هم در سطح بین فردی می باشند. این نشانگرها، نقش مهمی در درک گفتمان و پیشرفت اطلاعات دارند (شیفرین[۵]، ۱۹۸۷ نقل در یانگ[۶]، ۲۰۱۱). کلمات و عباراتی چون But، like، right، yeah، so، well، all in، all، after all، you know و okay نمونه هایی از نشانگرهای گفتمانی در زبان انگلیسی هستند (یانگ، ۲۰۱۱). شیفرین (۱۹۸۷) اهمیت نشانگرهای گفتمانی را آشکار ساخت و مدل انسجامی خود را ارائه داد. مدل او شامل سه سطح معنا، نحو[۷] و سازماندهی گفتمان بود. وی با ارائه این مدل درصدد بود تا بررسی کند که نشانگرهای گفتمانی چگونه می توانند به انسجام شفاهی کمک کنند (آرچاکیس[۸]، ۲۰۰۱ نقل در یانگ، ۲۰۱۱). او ۱۱ نوع از نشانگرهای گفتمانی را در مدل خود توصیف نمود (شیفیرین، ۱۹۸۷). در حالی که بسیاری از زبان شناسان به تحلیل توصیفی نشانگرهای گفتمانی در زبانی خاص و کلام بومی پرداختند، مولر[۹] (۲۰۰۴) بر نشانگرهای گفتمانی زبان مقصد که غیربومیان در کلام به کار می گیرند، تمرکز نمود (یانگ، ۲۰۱۱).

---

۱ Labov
۲ Fanshel
۳ Topic
۴ Fraser
۵ Schiffrin
۶ Yang
۷ Syntactic
۸ Archakis
۹ Muller

در ادبیات تجزیه و تحلیل کلام، عبارت های متنوعی وجود دارند که در برابر عبارت «نشانگرهای گفتمانی»به کار رفته اند. برخی از آنها عبارتند از: عبارت های نشانه، ذرات گفتمانی، عبارات پراگماتیک، نشانه های پراگماتیک، عملگرهای پراگماتیک، اتصال دهنده های جمله و عملگرهای گفتمانی (فراسر، ۱۹۹۹).

۵-۳- خلاصه‌سازی خودکار گفتمان مدار

انسجام و خوانایی خلاصه‌های سیستمی از چالشهای خلاصه‌سازی خودکار به ویژه نوع چندسندی آن است که برای مقابله با آن راهکار تعیین ناحیه های بلاغی متن[۱] ارائه شده‌است. این روش، با تعیین ساختار گفتمانی متن تلاش می‌کند جملات را به نحوی طبقه‌بندی کند که خلاصه‌هایی منسجم وخوانا تولید شود. بر این اساس، نوعی از خلاصه‌سازی به نام خلاصه‌سازی گفتمان‌مدار پیشنهاد شده‌است. ساختار گفتمانی یک متن شامل تمام روابط بین قطعه[۲]های یک متن است که ساختار آن متن را می‌سازد. ایده به کار گیری ساختار گفتمانی متن برای خلاصه‌سازی متون در سال ۱۹۹۷ توسط مارکو[۳] برای تولید خلاصه‌های کلی پیشنهاد شد (باسما[۴]، ۲۰۰۸) و حوزه ای از خلاصه‌سازی خودکار به نام خلاصه‌سازی گفتمان‌مدار را پدید آورد. این نوع خلاصه سازی از نظریه‌های سازماندهی گفتمان بهره می برد. یکی از مشهورترین این نظریه‌ها نظریه ساختار بلاغی[۵] است. بر اساس این نظریه می‌توان متن را به صورت ساختاری درختی یا درخت بلاغی بازنمون نمود که در آن محدوده‌های متنی[۶] با به کارگیری مجموعه‌ای از روابط گفتمانی از پیش تعیین شده به هم پیوند داده می‌شوند (مان[۷]و تامپسون[۸]، ۱۹۸۸).

بر اساس نظریه ساختار بلاغی، استراتژی ناحیه بندی استدلالی متون علمی به منظور خلاصه‌سازی خودکار متون علمی شکل گرفت. در این استراتژی، فرض بر این است که گفتمان علمی دربردارنده توصیف هایی از نظرات مثبت و منفی راجع به سهم پژوهشگران از یک حوزه علمی است. این نوع گفتمان، حاصل یک بازی بلاغیاست که سعی دارد جایگاه و سهم پژوهشگران از یک حوزه علمی را ارتقاء دهد. بر این اساس، متون علمی باید ادعاهای دانش را در تقابل با آثار پژوهشی قبلی به

---

۱ Rhetorical Zones
۲ Passage
۳ Marcu
٤ Bosma
٥ Rhetorical Structure Theory (RST)
٦ Text spans
۷ Mann
۸ Thompson

وضوح نشان دهند (فیساس[1] ، رونزانو[2] و سیگون، ۲۰۱۵). از این منظر، مدل گفتمانی ادعاهای دانش[3] معروف به «مدل سیمون تویفل[4]»یا همان ناحیه بندی استدلالی به وجود آمده است (تویفل، ۱۹۹۹؛ تویفل و موئنز، ۲۰۰۲؛ تویفل، سیدهارتان[5] و بچلور[6] ، ۲۰۰۹) در ناحیه بندی استدلالی، وضعیت بلاغی هر جمله به منظور بازنمون بافت گفتمانی جملات استخراج شده، تعیین می شود. به عبارت دیگر در این استراتژی، فرایند استخراج جمله با تحلیل گفتمان ترکیب شده است (تویفل و موئنز، ۲۰۰۲).

۶- کاربرد تحلیل گفتمان در خلاصه‌سازی ژانر علمی

در راستای مباحثی که در بخش چارچوب نظری مطرح شد، استراتژی های ویژه خلاصه‌سازی خودکار مقالات علمی توسعه یافته اند که عموما تلفیقی از روش های سطحی و معنایی هستند. یکی از استراتژی هایی که از تجزیه و تحلیل کلام بهره می برد عبارت است از ناحیه بندی استدلالی[7] بر اساس این استراتژی طرح های حاشیه نویسی گفتمانی ویژه متون علمی به وجود آمده اند که در ادامه مورد بحث قرار می گیرند.

۶-۱- طرح های مبتنی بر ناحیه بندی استدلالی

نشانه گذاری بلاغی اطلاعات، برای فرایندهای جدید دسترسی به اطلاعات علمی مفید می باشد. مثلا در حوزه بازیابی اطلاعات علمی، می توان از اطلاعات بلاغی به عنوان گزاره های بیان گر تغییر پارادایم استفاده کرد (لیساکیک[8] و همکاران، ۲۰۰۵ نقل در تویفل، سیدهارتان و بچلور ، ۲۰۰۹). زیرا مقالاتی که شامل این نوع گزاره ها هستند، مقالات تأثیرگذار یک حوزه محسوب می شوند. برای مثال، ۷۵ درصد از مقالات «اف. ۱۰۰۰»[9]که یک سرویس توصیه گر مقالات حوزه های پزشکی و زیست شناسی است، حاوی جملات تغییر پارادایم است. این مقالات از منظر خبرگان،

۱Fisas
۲Ronzano
۳Knowledge Claim Discourse Model(KCDM)
٤Simone Teufel's model
٥Siddharthan
٦Batchelor
۷Argumentative Zoning(AZ)
۸Lisacek
۹Biology ۱۰۰۰ Faculty of

از اهمیت خاصی در حوزه برخوردارند (تویفل، سیدهارتان و بچلور، ۲۰۰۹). بنابراین ساختار بلاغی مقالات علمی در خلاصه‌سازی مقالات علمی نیز حائز اهمیت می باشد. زیرا بر اساس آن، می توان خلاصه های تک سندی وچندسندی تولید کرد که به تفاوت ها و شباهت های آثار مورد استناد قرار گرفته هم توجه می کنند.برای تعیین وضعیت بلاغی جملات منتخب، ابتدا لازم است ساختار یک مقاله علمی تشریح شود. ساختار یک مقاله علمی ابعاد مختلفی دارد که عبارتند از:

حل مسئله: پژوهش عموما به صورت فعالیتی برای حل مسئله توصیف می شود (جردن[۱]، ۱۹۸۴؛ زاپن[۲]، ۱۹۸۳ نقل در تویفل و موئنز، ۲۰۰۲). انتظار می رود در هر مقاله پژوهشی سه نوع اطلاعات وجود داشته باشد که عبارتند از: مسئله و اهداف پژوهش، راه حل ها (روش ها) و نتایج. این ساختار که ساختار مسئله راه حل نام دارد در بسیاری از رشته ها به ویژه در علوم تجربی ساختاری غالب است که به صورت ساختار ثابت مقدمه/روش/نتایج/بحث مشهود است (ون دیجک[۳]، ۱۹۸۰ نقل در تویفل و موئنز، ۲۰۰۲). این در حالی است که متونی علمی هم وجود دارند که عینا از این ساختار پیروی نمی کنند.

تخصیص فکری[۴]: متون علمی باید سهم خود از دنیای علم و اصالت اثر[۵] را به صورت واضح نشان دهند و جنبه های جدید کار را در مقابله با رویکردهای پژوهشگران دیگر و بیانیه های علمی عموما پذیرفته شده، آشکار سازند. نویسندگان عموما دلیل محکمی برای نوشتن مقاله خود دارند و آن را به گونه ای می نگارند که تفکر خاص شان آشکارا بیان شود.

جدول۱: طرح حاشیه نویسی نواحی بلاغی تویفل و موئنز (۲۰۰۲)

ناحیه بلاغی	توصیف
AIM	ناحیه هدف خاص پژوهشی مقاله
TEXTUAL	ناحیه گزاره هایی درباره ساختار بخش
OWN	ناحیه توصیف اثر خود: روش شناسی، نتایج و بحث (ناحیه خنثی)
BACKGROUND	ناحیه پیشینه علمی عموما پذیرفته شده
CONTRAST	ناحیه گزاره هایی در مقایسه یا تضاد با اثر دیگر؛ ضعف اثر دیگر
BASIS	ناحیه گزاره هایی در موافقت با اثر دیگریا ادامه اثر دیگر
OTHER	ناحیه توصیف اثر دیگران (ناحیه خنثی)

۱ Jordan
۲ Zappen
۳ van Dijk
٤ Intellectual attribution
٥ Contribution

خوانندگان نیز معمولا آن را به آسانی درک می کنند (تویفل و موئنز، ۲۰۰۲).
استدلال علمی[۱]: علم تنها کارخانه ای از واقعیت ها نیست بلکه جنبه های اجتماعی قوی ای دارد. زیرا موفقیت یک پژوهشگر، به توانایی او برای متقاعدکردن دیگران در مورد اعتبار استدلال هایش وابسته است (اسویلز[۲]، ۱۹۹۰ نقل در تویفل و موئنز، ۲۰۰۲). به عبارت دیگر، نویسنده تلاش می کند اعتبار «ادعای دانش»[۳] خود را نشان دهد. دانشی که در واقع قطعه دانش جدیدی است که از طریق داوری و انتشار با مخازن دانش آن حوزه یکپارچه خواهدشد (تویفل، سیدهارتان و بچلور، ۲۰۰۹). او این کار را از طریق یک قطعه متنی که از نظر بلاغی انسجام دارد، انجام می دهد (میرز [۴]، ۱۹۹۲ نقل در تویفل و موئنز، ۲۰۰۲).
نگرش به آثار دیگران: نویسندگان برای مستدل کردن افکار خود به آثار دیگران ارجاع می دهند. هر استناد دلیلی دارد. به عبارت دیگر، نویسنده هر ارجاع را بر اساس هدفی ذکر می کند. در یک متن خوب نگارش شده، این گونه اهداف قابل درک است (تویفل و موئنز، ۲۰۰۲). حوزه تحلیل محتوای استناد[۵]، به چگونگی و چرایی استنادهی نویسندگان می پردازد. در تحلیل محتوای استناد با بررسی بافتار استناد[۶] تلاش می شود تا چگونگی ارتباط بین مقالات آشکار شود. هدف از استناد انواع گوناگونی چون علمی (نقد مثبت و منفی و...) و اجتماعی (رعایت ادب، سنت و ...) دارد (زیمان[۷]، ۱۹۶۹ نقل در تویفل و موئنز، ۲۰۰۲).
ابعاد مختلف ساختار یک مقاله علمی، راهنمایی جهت تعیین وضعیت بلاغی جملاتی است که قرار است برای درج در خلاصه نهایی انتخاب شوند. جملات دارای وضعیت بلاغی یکسان ناحیه های بلاغی متن را تشکیل می دهند. کوچک ترین ناحیه فقط یک جمله را شامل می شود و ناحیه های بزرگ تر جملات بیشتری را شامل می شوند (تویفل و موئنز، ۲۰۰۲). یکی از مشهورترین طرح های حاشیه نویسی نواحی بلاغی مقالات علمی، طرحی است که تویفل و موئنز (۲۰۰۲) ارائه داده اند. این طرح، ۷ ناحیه بلاغی[۸] را برای مقالات علمی حوزه زبان شناسی رایانشی شناسایی کرده است (جدول ۱). سه ناحیه OWN، OTHER و BACKGROUND بر این اساس تعریف می شوند که در بخش مربوطه، چه کسی صاحب آن «ادعای دانش» است. دو ناحیه BASIS و CONTRAST با توجه به ارتباط شان با آثار موجود تعریف می شوند. بنابراین بخشی از این طرح به طرح های

۱Scientific argumentation
۲Swales
۳Knowledge claim
٤Myers

۵. متن پیرامونی یک استناد را گویند. به عبارت دیگر جمله ای است که شامل استناد به اثر است.
٦Content citation analysis
۷Ziman
۸Rhetorical Zones

جدول۲: طرح حاشیه نویسی نواحی بلاغی تویفل، سیدهارتان و باتچلور (۲۰۰۹)

ناحیه بلاغی	توصیف	ناحیه بلاغی	توصیف
AIM	اهداف یا فرضیه های پژوهش در مقاله حاضر	OWN-CONC	نتیجه گیری (غیر قابل اندازه گیری) نویسنده در مقاله حاضر
NOV-ADV	تازگی و مزیت رویکرد خاص پژوهشگر در مقاله حاضر	CODI	مقایسه، تضادها و تفاوت های مقاله با دیگر راه حل ها (توصیف و مقایسه خنثی)
CO-GRO	زمینه های مشترک، پیشینه علمی	GAP-WEAK	کمبود راه حل در حوزه، مشکل با راه حل های دیگر (ترکیب شکاف های دانش با رویکرد انتقادی)
OTHER	ادعای دانش قابل توجهی که توسط افراد دیگر در مقاله هاشان بیان شده است (توصیف خنثی)	ANTISUPP	مقابله و برخورد با نتایج یا نظریه دیگران، برتری اثر خود
PREV-OWN	ادعای دانش قابل توجهی که توسط نویسنده در مقالات گذشته اش آورده است (توصیف خنثی)	SUPPORT	آثار دیگر، مقاله حاضر را پشتیبانی می کنند یا اثر دیگری، این مقاله را حمایت می کند
OWN-MTHD	ادعای دانش جدید نویسنده مقاله: روش شناسی	USE	کارهای دیگران که در اثر نویسنده مقاله حاضر به کارگرفته شده است
OWN-FAIL	راه حل، روش و آزمایش نویسنده مقاله که کارساز نبوده است (خطاهای قابل بازیافتنی نویسنده)	FUT	محدودیت ها و پیشنهادات پژوهش، پژوهش های آینده
OWN-RES	بروندادهای عینی و قابل اندازه گیری نویسنده در مقاله حاضر (یافته ها)		

رده بندی عملکرد استناد[۱] مربوط است که بر گرفته از حوزه تحلیل محتوای استناد است. AIM به اصلی ترین ادعای دانشی مقاله که معمولا در مقدمه و نتیجه گیری تکرار می شود اشاره می نماید و بالاخره TEXTUAL، محل فیزیکی اطلاعات را از طریق مرور کلی یک بخش و یا ارائه خلاصه ای از بخش های فرعی توضیح می دهد (تویفل، سیدهارتان و بچلور ، ۲۰۰۹).

تویفل، سیدهارتان و بچلور (۲۰۰۹) این طرح را به ۱۵ ناحیه برای مقالات دو حوزه شیمی و زبان شناسی رایانشی گسترش دادند (جدول ۲). در این طرح گسترش یافته، ناحیه AIM مانند طرح قبل حفظ شده است. نام ناحیه BACKGROUND به CO-GRO COMMON GROUND تغییر یافته است. ناحیه OTHER به دو ناحیه OTHER و PREV-OWN، ناحیه BASIS به دو ناحیه USE و SUPPORT، ناحیه CONTRAST به سه ناحیه ANTISUPP ،CODI و

[۱] Citation function classification schemes

GAP-WEAK و بالاخره ناحیه OWN به چهار ناحیه OWN-، OWN-RES، OWN-MTHD CONC و OWN-FAIL تقسیم شده است. با توجه به این که ناحیه TEXTUAL نسبت به دیگر نواحی اطلاعات کمتری در اختیار قرار می دهد، در این طرح حذف شده است.

با توجه به اهمیت بیش از پیش تمایزات در حوزه های مختلف علم و با هدف نزدیک تر شدن به طرح های عملکرد استناد موجود، طرح تویفل، سیدهارتان و بچلور (۲۰۰۹) جزئی تر و مفصل تر از طرح تویفل و موئنز (۲۰۰۲) می باشد. لازم به ذکر است که ادعاهای دانشی که شامل استناد هستند و استنادهایی که در مثال های متعدد به وقوع می پیوندند، یکی از عوامل تصمیم گیری برای تعیین ناحیه بلاغی یک جمله است. اما، استنادها لزوما عامل کلیدی برای چنین تصمیم گیری هایی نیستند. این در حالی است که تأثیرگذارترین عامل در این فرآیند، تمایز خاص نویسنده در آن حوزه از علم می باشد که از طریق نواحی چهارگانه OWN به خوبی قابل درک خواهدبود (تویفل، سیدهارتان و بچلور ، ۲۰۰۹).

۷- بحث

با وجود اینکه بیش از ۵۰ سال است که پژوهش در حوزه خلاصه‌سازی خودکار متن آغاز شده‌است اما هنوز چالش اصلی، تولید خلاصه‌های با کیفیت است. بنابراین اکثر پژوهش‌ها بر بهبود نتایج تمرکز دارند. یکی از رویکردهایی که به منظور بهبود کیفیت خلاصه ها مورد توجه جامعه علمی قرار گرفته است، خلاصه‌سازی گفتمان مدار می باشد که همان گونه که در بخش پیشین تشریح شد، تلاش می کند اطلاعات ساختارمند مفیدی از نوشتارهای علمی استخراج کند تا در مرحله بعد بتوان آنها را در ابزارهای خودکار خلاصه‌سازی متن مورد استفاده قرارداد. با اختصاص برچسب های گفتمانی به جملات و با بهره گیری از روش های یادگیری ماشینی نظارتی می توان خلاصه هایی از این دست را تولیدنمود.

در ژانر علمی، می توان رویکرد خلاصه‌سازی گفتمان مدار را با رویکرد استنادی تلفیق کرد آن را به عنوان راهکاری برای تولید الگوریتم هایی که قادر هستند خلاصه هایی با خوانایی بالاتر که محتوای اطلاعاتی مفیدتر و بیشتری دارند، ارائه داد.

در خلاصه‌سازی استنادی از مجموعه‌ای از استنادها به یک مقاله مرجع، برای تولید خلاصه بهره می برند (قزوینیان و رادو، ۲۰۰۸؛ قزوینیان و همکاران، ۲۰۱۳). مجموعه استنادها به یک مقاله، را در واقع می‌توان خلاصه ای از آن مقاله دانست که توسط جامعه علمی تولید شده‌است (الکیس و همکاران، ۲۰۰۸؛ قزوینیان و همکاران، ۲۰۱۳؛ کهن، سولداینی[1] و گوهریان، ۲۰۱۴) که در بردارنده

[1] Soldaini

مهمترین نکات مقاله اصلی و در واقع سهم آن اثر در جامعه علمی است. به این ترتیب، یک خلاصه استنادی با فراهم آوردن اطلاعات اضافی نسبت به چکیده نویسنده، می‌تواند بینش عمیق تری راجع به تأثیر آن مقاله در جامعه علمی را فراهم آورد.
اختصاص چهریزه های گفتمانی به جملات استنادی مقالات استنادکننده می تواند در طبقه بندی و رتبه بندی آنها به منظور درج در خلاصه نهایی مفید واقع شود. به عبارت دیگر، از طریق تحلیل گفتمان جملات استنادی مقالات استنادکننده می توان عملکرد استنادی را نیز معلوم کرد و بر اساس آنها جهت تولید خلاصه های منسجم تر اقدام نمود.

۸- نتیجه گیری

در سال های اخیر، رویکردهای گفتمانی خلاصه‌سازی ژانر علمی گسترش یافته‌اند. در این رویکردها فرض بر این است که با تکیه بر ویژگی‌های گفتمانی متن، نه تنها می‌توان خلاصه‌هایی با خوانایی بالاتر تولید کرد بلکه امکان بازنمون خودکار اطلاعات ساختارمند مفید برای پژوهشگران نیز از متون علمی فراهم می‌شود. پژوهش‌های انجام شده با این رویکرد، از نشانه‌های گفتمانی برای ناحیه بندی بلاغی متون علمی بهره می‌برند.
از کاربردهای مهم و ویژه تحلیل گفتمان در خلاصه‌سازی خودکار ژانر علمی، تولید خلاصه های استنادی گفتمان مدار می باشد. از آنجایی که تحلیل گفتمان جملات پیکره های علمی و اختصاص برچسب های گفتمانی و ناحیه بندی استدلالی آنها توسط متخصصان، به صرف وقت و هزینه زیاد نیازمند است. لازم است مؤسسات تحقیقاتی پشتیبان این مهم باشند. زیرا در صورت تحقق آن برای پیکره های عظیم ژانر علمی می توان انواع خلاصه های خودکار مبتنی بر پرسش کاربر، خلاصه های چندسندی که تضادها و شباهت ها شان در خلاصه نهایی مشهود است و نیز خلاصه های استنادمدار را تولید نمود و بدین وسیله خدمت بزرگی به بازیابی اطلاعات علمی مورد نیاز پژوهشگران ارائه داد.

۹- پیشنهادات

با توجه به مباحث مطرح شده در این مقاله پیشنهاد می شود گروه های تحقیقاتی متشکل از متخصصین علم اطلاعات و دانش شناسی، زبان شناسان و مهندسین نرم افزار در دانشگاه ها و مؤسسات تحقیقاتی کشور برای تولید پیکره های حاشیه نویسی شده ژانر علمی به زبان فارسی ایجاد شود. همچنین، در پایان نامه های رشته زبان شناسی که به تجزیه و تحلیل کلام در متون

علمی می پردازند، رویکرد کیفی بیشتر از پیش تقویت گردد تا بتوان با اجتناب از گزارش های صرفا آماری نشانه های گفتمانی موجود در متون علمی به سمت کابردی کردن تحلیل گفتمان علمی در حوزه هایی چون بازیابی اطلاعات خودکار حرکت نمود و از این طریق مطالعات بین رشته ای پرثمرتری را انجام داد.

مراجع

البرزی ورکی ، پرویز (۱۳۸۶). «مبانی زبان شناسی متن». تهران: انتشارات امیر کبیر.

بشیری، حسن (۱۳۹۴). زبان شناسی و تحلیل گفتمان اخبار: رویکردی عملیاتی، فصلنامه نقد کتاب: اطلاع‌رسانی و ارتباطات، شماره ۷و۸. ۲.

کلانتری، صمد و دیگران (۱۳۸۸). تحلیل گفتمان با تاکید بر گفتمان انتقادی به عنوان روش تحقیق کیفی. جامعه شناسی. شماره ۱،۴. ۲۸-۷.

Abu-Jbara, A., and Radev, D. (2011, June). Coherent citation-based summarization of scientific papers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. pp. 500-509.

Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. IBM Journal of Research and Development, 2(4). pp. 354-361.

Bosma, W. E. (2008). Discourse oriented summarization. University of Twente.

Brown, G and Yule, G. (1983) Discourse Analysis, Cambridge University Press.

Cohan, A., and Goharian, N. (2015). Scientific article summarization using citation-context and article›s discourse structure. Proceedings of the 2015 Conference on Empirical Methods in Nathral Language Processing. Lisbon, Portugal, 17-21 September. pp. 390-400.

Cohan, A., Soldaini, L., and Goharian, N. (2014). Towards citation-based summarization of biomedical literature. In Proceedings of the Text Analysis Conference (TAC'14).

DeJong, G. (1982). An overview of the FRUMP system. Strategies for natural language processing, 113. pp. 149-176.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article?. Journal of the American Society for Information Science and Technology, 59(1). pp. 51-62.

Fisas, B., Ronzano, F., and Saggion, H. (2015, June). On the discoursive structure of computer graphics research papers. In The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015. pp. 42.

Fraser, B. (1999). What are discourse markers?, Journal of Pragmatics, 31. pp. 931-952.

Jones, K. S. (1993). What might be in a summary?. Information retrieval, 93. pp. 9-26.

Jones, K. S., and Endres-Niggemeyer, B. (1995). Automatic summarizing. Information Processing & Management, 31(5). pp. 625-630.

Kupiec, J., Pedersen, J., and Chen, F. (1995, July). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. pp. 68-73.

Lloret, E., and Palomar, M. (2013). COMPENDIUM: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. Natural Language Engineering, 19(2). pp.147-186.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2). pp.159-165.

Mani, I. (2001). Automatic summarization (Vol. 3). John Benjamins Publishing.
Mann, W. C., and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3), 243-281.

Qazvinian, V., and Radev, D. R. (2008, August). Scientific paper summarization using citation summary networks. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 Association for Computational Linguistics. pp. 689-696.

Qazvinian, V., and Radev, D. R. (2010, July). Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics. pp. 555-564.

Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. Computational linguistics, 28(4). pp. 399-408.

Rahman, N., and Borah, B. (2015, September). A survey on existing extractive techniques for query-based text summarization. In Advanced Computing and Communication (ISACC), 2015 International Symposium on IEEE. pp. 98-102.

Saggion, H., and Poibeau, T. (2013). Automatic text summarization: Past, present and future. In Multi-source, multilingual information extraction and summarization (pp. 3-21). Springer Berlin Heidelberg.

Schiffrin, D. (1987). Discourse Markers. Cambridge: Cambridge University Press.

Teufel, S. (1999). Argumentative Zoning: Information Extraction from Scientific Text. School of Cognitive Science, University of Edinburg, UK..

Teufel, S., and Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. Computational linguistics, 28(4). pp.409-445.

Teufel, S., Siddharthan, A., and Batchelor, C. (2009, August). Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3 . Association for Computational Linguistics. pp. 1493-1502.

Yang, S. H. A. N. R. U. (2011). Investigating discourse markers in pedagogical settings: a literature review. ARECLS, 8. pp. 95-108.

Yeloglu, O., Milios, E., and Zincir-Heywood, N. (2011, March). Multi-document summarization of scientific corpora. In Proceedings of the 2011 ACM Symposium on Applied Computing .ACM. pp. 252-258.