مجله علمی- پژوهشی «پدافند الکترونیکی و سایبری»
سال دوم، شماره ۳، پاییز ۱۳۹۳؛ ص ۶۳-۵۱

روش مبتنی بر شباهت معنایی در خلاصه‌سازی متون فارسی براساس عبارت پرس‌وجوی کاربر

زهرا سپهریان[1*]، سعیده‌سادات سدیدپور[2]، حسین شیرازی[3]

۱- کارشناسی ارشد، مجتمع فناوری اطلاعات، ارتباطات وامنیت، دانشگاه صنعتی مالک‌اشتر
۲- دانشجوی دکتری، مجتمع فناوری اطلاعات، ارتباطات و امنیت، دانشگاه صنعتی مالک‌اشتر
۳- دانشیار، مجتمع فناوری اطلاعات، ارتباطات و امنیت، دانشگاه صنعتی مالک‌اشتر
(دریافت: ۹۳/۰۲/۰۸، پذیرش: ۹۳/۰۶/۰۱)

چکیده

سیستم‌های خلاصه‌سازی خودکار متون، یکی از انواع سیستم‌های مدیریت اطلاعات حجیم هستند. این مقاله به یکی از انواع خلاصه‌سازی استخراجی به نام خلاصه‌سازی مبتنی بر پرس‌وجوی کاربر بر روی زبان فارسی می‌پردازد که بسیار برای مرور اطلاعات بر روی موضوعات مشخص توسط فرماندهان مفید می‌باشد.

مهم‌ترین فاز در این نوع خلاصه‌سازی، محاسبه شباهت بین عبارت پرس‌وجو و اجزای متن اصلی است. برای رسیدن به این مهم، پس از فاز پیش‌پردازش، تبدیل عبارت پرس‌وجو به جمله و بهره بردن از ابهام‌زدایی معنایی کلمات، شباهت معنایی بین عبارت پرس‌وجو و جملات متن با استفاده از فارس‌نت محاسبه می‌شود. سپس، جملاتی که بیشترین شباهت معنایی را با عبارت پرس‌وجو داشته باشند برای حضور در خلاصه انتخاب می‌شوند. ارزیابی‌های حاصل از رویکرد پیشنهادی مقاله، نشان از مطلوب بودن نسبی الگوریتم موردنظر دارد. با توجه به نوپا بودن زبان فارسی در زمینه پردازش زبان طبیعی، توسعه آنچه در این مقاله بررسی شده‌است و نظایر آن می‌تواند کمک شایانی به بهبود نتایج کند.

واژه‌های کلیدی: خلاصه‌سازی مبتنی برکاربر، عبارت پرس‌وجو، شباهت معنایی، ابهام‌زدایی معنایی کلمات، فارس‌نت

۱. مقدمه

در دوران دیجیتالی امروز، با رشد نمایی فناوری‌هایی از جمله اینترنت، دنیا شاهد افزایش روزافزون اطلاعات متنی و منابع اینترنتی می‌باشد. فراوانی تولید اطلاعات آن‌قدر تأثیرگذار بوده که باعث ورود اصطلاحات و لغات جدید به زبان شده‌است. سیل و طوفان اطلاعات، بیماری شناخته‌شده‌ای است که جهانیان از آن رنج می‌برند [۱].

این حجم زیاد اطلاعات موجب شده تا دسترسی به اطلاعات مورد نیاز، کاری زمان‌گیر قلمداد شود. این مسئله برای عموم مردم و به‌ویژه فرماندهان و رهبران، بسیار مشکل‌ساز می‌باشد. در اصل، فضای سایبری و این حجم عظیم اطلاعات، نوعی صحنه نبرد می‌باشد که باید به‌نوعی مدیریت شود. استخراج اطلاعات مفید و خلاصه برای درک شرایط موجود و یافتن راه‌حل مناسب، بهترین راهکار برای مدیریت این صحنه نبرد به‌شمار می‌رود. استفاده از روش‌های کارا در بازیابی اطلاعات راه‌حلی برای مشکل افزونگی اطلاعات می‌باشند که در این میان، خلاصه‌سازی متون نقش مهمی در استخراج مفاهیم اصلی و مهم یک متن به شیوه فشرده بازی می‌کند. به بیان دیگر، با در اختیار قرار دادن خلاصه متون مرتبط با کلیدواژه‌های مورد نظر کاربری که می‌تواند در هر سطحی از سلسله مراتب قرار داشته باشد، وی می‌تواند به مرور سریع اطلاعات مورد نیاز خود پرداخته و تصمیم‌های لازم را اتخاذ نماید.

خلاصه‌سازی خودکار متن عبارت است از فرآیند تولید خودکار گونه‌ای فشرده از متن ورودی که اطلاعات مفید را به کاربر ارائه می‌کند. در واقع، مهم‌ترین مزیت خلاصه نسبت به متن اصلی، کاهش زمان خواندن آن است؛ به‌شرطی که از خوانایی و پیوستگی بین جمله‌ای نیز برخوردار باشد.

ساده‌ترین نوع خلاصه خودکار موجود، خلاصه استخراجی است که در آن با استفاده از تکنیک‌ها و الگوریتم‌های متعدد، جملات مهم متن انتخاب و به تعدادی که نسبت فشردگی اجازه دهد، در خلاصه حاضر می‌شوند [۲].

یکی از انواع خلاصه‌های استخراجی، خلاصه‌های مبتنی‌بر پرس‌وجوی[1] ارائه‌شده توسط کاربر یا به اختصار خلاصه مبتنی‌بر کاربر است که در نقطه مقابل خلاصه‌های عمومی قرار دارد. خلاصه‌سازی عمومی، سعی در بیان نظر نویسنده داشته و در آن، عناوین اصلی به‌منظور تولید خلاصه استفاده خواهندشد؛ در حالی که ممکن است خروجی سیستم خلاصه‌سازی به‌عنوان ورودی سایر سیستم‌ها به خصوص سیستم‌های بلادرنگ به‌کار رود؛ بنابراین در خلاصه‌سازی

* رایانامه نویسنده مسئول: sadidpur@aut.ac.ir

1. Query

مبتنی‌بر کاربر، به جنبه یا موضوع خاصی توجه و نقطه نظر مورد نظر در ایجاد خلاصه اعمال شده و تنها عناوینی که با عبارت پرس‌وجو مرتبط باشند، برای تولید خلاصه استفاده خواهند شد. در واقع گاهی متن اصلی همزمان به چند موضوع مرتبط می‌پردازد ولی کاربر و یا سیستم تنها نیاز به خلاصه در یک راستا دارد در نتیجه مطلوب این است که اطلاعات اضافه با توجه به عبارت پرس‌وجو از خلاصه حذف شود. بنابراین، می‌توان گفت که در این سیستم‌ها، تشخیص میزان ارتباط و شباهت بین عبارت پرس‌وجو و اجزای متن اصلی، مهم‌ترین فاز می‌باشد.

این مقاله به بررسی روش‌های موجود در این نوع خلاصه‌سازی پرداخته و روشی جهت تولید خلاصه مبتنی بر کاربر در زبان فارسی ارائه می‌دهد. در ادامه، در بخش ۲، تاریخچه مختصری از خلاصه‌سازی و روش‌های موجود درخلاصه‌سازی مبتنی‌بر کاربر ارائه می‌شود. سپس، بخش ۳ به روش پیشنهادی مقاله می‌پردازد و در بخش ۴ ارزیابی از روش پیشنهادی صورت می‌گیرد. در نهایت، در بخش ۵، نتیجه‌گیری و پیشنهاداتی برای کارهای آینده بیان خواهد شد.

۲. کارهای انجام شده

اولین گام‌ها در زمینه خلاصه‌سازی خودکار متن در دهه ۵۰ برداشته شد و Luhn، روشی مبتنی‌بر فراوانی واژه‌ها پیشنهاد کرد که در آن، جملات براساس میزان فراوانی کلماتشان رتبه‌گذاری و در خلاصه حاضر شدند [۳]. سپس، در سال ۱۹۶۹، Baxendale، ویژگی موقعیت جمله را به فراوانی واژه اضافه کرد. در ادامه، Edmondson سیستم خلاصه‌سازی را معرفی کرد که چهار ویژگی فراوانی واژه، موقعیت جمله، عنوان متن و حضور کلمات نشانه را مدنظر قرار می‌داد [۴]. Kupiec نیز در دهه‌ی ۹۰، از تکنیک‌های یادگیری ماشین برای تولید خلاصه استفاده کرد و تکنیک تحلیلی‌ای را به‌کار برد که فرآیند یادگیری را با استفاده از آمارهای بیزی انجام می‌داد [۵]. Chauang در ادامه کار Kupiec، روش‌های دیگری همانند درخت تصمیم و شبکه‌ی عصبی را معرفی کرد [۶]. در سال‌های اخیر، برخی روش‌های محاسبات نرم همانند منطق فازی و الگوریتم ژنتیک برای مسئله‌ی استخراج جملات مهم و تولید خلاصه به‌کار گرفته شدند.

از سال ۲۰۰۰، ایده خلاصه‌سازی براساس پرس‌وجوی ارائه‌شده توسط کاربر مطرح شد تا به‌وسیله‌ی آن، کاربران از منابع اطلاعاتی از قبیل کتاب‌های الکترونیکی و سایت‌های اینترنتی، استفاده‌ی بهینه‌تری داشته باشند و خلاصه‌ها متناسب با نیاز آنها تولید شوند.

رویکردهای متفاوتی در زمینه خلاصه‌سازی مبتنی‌بر کاربر وجود دارد که می‌توانند در ۳ دسته مبتنی‌بر گراف، مبتنی‌بر یادگیری ماشین و مبتنی‌بر زبان‌شناسی تقسیم‌بندی شوند. این روش‌ها، مستقل از زبان بوده و در صورت وجود ابزارها و امکانات مورد نیاز هر روش، در زبان موردنظر قابل پیاده‌سازی می‌باشند.

۲. ۱. رویکردهای مبتنی‌بر گراف

در رویکرد مبتنی‌بر گراف، بعد از توکن‌بندی و تجزیه جملات به گروه‌های اسمی، روابطی براساس قوانین اکتشافی تولید می‌شود. سپس، با در نظر گرفتن همه منابع، یک گراف مرکزی ایجاد می‌شود که به انتخاب جملات خلاصه کمک می‌کند [۷]. در بهترین روش این رویکرد گرافی از متن و گرافی از عبارت پرس‌وجو ساخته شده و شباهت بین هر جمله و عبارت پرس‌وجو محاسبه می‌شود. سپس، بهترین جملات برای حضور در خلاصه انتخاب می‌شوند.

۲. ۲. رویکردهای مبتنی‌بر یادگیری ماشین

در رویکرد مبتنی بر یادگیری ماشین، عموماً برای تولید خلاصه استخراجی، تکنیک‌های بازیابی اطلاعات و خلاصه‌سازی با هم ترکیب می‌شوند [۸]. درواقع، برداشتی جدید از اهمیت جمله، مستقل از عبارت پرس‌وجو، در امتیازدهی نهایی دخیل می‌شود. در این رویکرد، جملات، به‌وسیله مجموعه‌ای از ویژگی‌ها امتیازدهی و کل امتیاز هر جمله که در این بخش به‌دست آمده، با ترکیب خطی و وزن‌دار ویژگی‌ها حاصل می‌شود.

۲. ۳. رویکردهای مبتنی‌بر زبان‌شناسی

در رویکرد مبتنی‌بر زبان‌شناسی، از الگوریتم‌ها و ابزارهای موجود در زبان‌شناسی مانند مدل مفخی مارکوف[1] (HMM)، ابزار تشخیص موجودیت‌های نامدار[2] (NER)، پارسر و غیره استفاده می‌شود. نتایج نشان می‌دهد که این رویکرد به دلیل شبیه‌سازی تقریبی معنا، نسبت به دو رویکرد دیگر به نتایج انسانی شباهت بیشتری داشته و موفقیت‌های بزرگتری کسب نموده‌است.

یکی از راهکارهای مهم و مؤثر در رویکرد مبتنی‌بر زبان‌شناسی، تعیین میزان شباهت معنایی عبارت پرس‌وجو و جملات متن می‌باشد. براساس این راهکار، جملاتی که شباهت معنایی بیش‌تری به عبارت پرس‌وجو داشته باشند، شانس بیشتری برای حضور در خلاصه مبتنی‌بر کاربر خواهند داشت [۲, ۴]. برای این منظور، به محاسبه میزان شباهت اجزای تشکیل‌دهنده عبارت پرس‌وجو و جملات متن یعنی کلمه نیاز می‌باشد. در سال‌های گذشته، روش‌های متعدد و متنوعی در این زمینه گسترش یافته که هر کدام در کاربرد خاصی مؤثر واقع شده‌است.

به‌طور کل می‌توان روش‌های موجود در زمینه محاسبه شباهت کلمات را در دو دسته مبتنی بر پایگاه‌داده‌های لغوی و مبتنی بر وب تقسیم‌بندی نمود.

رویکرد مبتنی بر پایگاه‌داده‌های لغوی، از پایگاه داده‌هایی که به‌صورت دستی و توسط انسان ایجاد شده‌است، بهره می‌برد. در این راستا، پایگاه‌داده‌های متنوعی در زبان‌های مختلف دنیا توسعه یافته است که از بین آنها می‌توان به Hownet و Wordnet [۹] در زبان

1. Hidden Markov Model
2. Named Entity Identifier

انگلیسی اشاره کرد که به‌دلیل جامعیت، کامل بودن نسبی و ساختار سلسله‌مراتبی، دارای محبوبیت بالایی در این عرصه در این زبان می‌باشند. به‌طور معمول در این مجموعه‌ها، شباهت بین کلمات با محاسبه تعداد یال‌های موجود در مسیر سلسله‌مراتبی آنها به‌دست می‌آید. در کارهای جدید مانند [۱۰] به یال‌های مسیر با توجه به عمق آنها وزنی اختصاص می‌یابد که این امر به دقیق‌تر بودن میزان شباهت محاسبه شده کمک قابل توجهی می‌کند. مجموعه‌ی فارس‌نت در زبان فارسی نیز تلاشی جهت پیاده‌سازی چنین پایگاه‌داده‌ای می‌باشد. رویکرد دیگر در این حوزه از پیکره وب و اطلاعات مختص آن مانند بازدیدها[1] [۱۱]، قطعات وب[2] [۱۲-۱۳] و تعداد صفحه [۱۴-۱۵] جهت محاسبه میزان شباهت کلمات استفاده می‌کند.

پس از محاسبه میزان شباهت کلمات ، از روش‌های موجود جهت محاسبه میزان شباهت جملات استفاده و شباهت میان عبارت پرس‌وجو و جملات متن اصلی تعیین می‌شود. کارهای انجام شده در این راستا در سه دسته محاسبات مبتنی‌بر هم‌پوشانی کلمات، مبتنی‌بر اطلاعات آماری و مبتنی‌بر زبان‌شناسی قرار دارند [۱۶].

از کارهای تأثیرگذار در این بخش می‌توان به تحلیل معنایی پنهان (LSA)[3] [۱۷]، TF-IDF[4] [۱۸]، رویکردی براساس اطلاعات معنایی و چیدمان کلمات [۱۹] و رویکردی مبتنی بر استخراج رویداد جمله [۲۰] اشاره کرد.

در LSA یک مجموعه از کلمات، نماینده تعداد زیادی از مفاهیم بوده و به‌ازای هر جمله، یک بردار در فضای کاهش‌یافته پر شده و سپس، شباهت بین دو جمله، با اندازه‌گیری شباهت بین این دو بردار به‌دست می‌آید [۱۷]. به‌دلیل محدودیت ابزارهای به‌کارگرفته شده در این روش، ابعاد کاهش زیادی می‌یابند و ممکن است بعضی از کلمات مهم جمله ورودی با طول نامعلوم در فضای ابعاد LSA قرار نگیرند. به‌علاوه، به‌علت ثابت بودن نماینده‌های کلمات، واحدهای کوتاهی مانند جمله، با بازنمایی تنکی مواجه می‌شوند.

TF-IDF به دو پارامتر تعداد تکرار عبارت در دو جمله و عکس تعداد تکرار عبارت در مجموعه متون توجه دارد؛ به‌طوری‌که با افزایش تعداد تکرار عبارت موردنظر در مجموعه متون، وزن اختصاص داده‌شده به آن کاهش می‌یابد [۱۸]. ایراد این روش آن است که جملاتی با معنی مشابه، لزوما کلمات مشترک زیادی ندارند.

به این ترتیب جهت محاسبه شباهت معنایی جملات، رویکردی براساس اطلاعات معنایی و چیدمان کلمات پیشنهاد شد [۱۹]. در این رویکرد، ابتدا شباهت معنایی ازطریق پایگاه دانش لغوی و یک پیکره به‌دست می‌آید. سپس، شباهت چیدمان کلمات، براساس تعداد کلمات متفاوت و تعداد جفت کلمات در یک چیدمان متفاوت محاسبه می‌شود. در نهایت، شباهت بین دو جمله، با ترکیب این دو مقدار شباهت به‌دست‌آمده تعریف می‌شود [۱۹].

در سال‌های اخیر ابزارهای کارا در زمینه پردازش زبان طبیعی، سهم زیادی را در محاسبات شباهت معنایی به خود اختصاص داده‌اند. به‌عنوان مثال رویکردی، برای ترکیب شباهت نحوی و معنایی از ابزار توسعه‌یافته در زبان انگلیسی یعنی پارسر بهره برده و به استخراج رویداد جمله پرداخته است. در واقع جهت محاسبه شباهت معنایی، شباهت بین عناصر رویداد اندازه‌گیری شده است [۲۰]. در رویکرد دیگری، جملات براساس چند قاعده گرامری و با استفاده از پارسر به چند قطعه (بدنه و سایر قطعات) تقسیم شده‌اند و در محاسبه شباهت معنایی دو جمله، شباهت معنایی هر کدام از قطعات با توجه به اهمیت‌شان و با وزن‌های مختلف در نظر گرفته شده‌است [۲۱]. پژوهش دیگری از ابهام‌زدای معنایی کلمات بهره برده و از هم‌پوشانی senseها برای محاسبه این شباهت استفاده نموده است. همانطور که در ادامه و در بخش ۰ به آن پرداخته می‌شود کلمات یکسان در کاربردهای متفاوت با معانی متفاوت ظاهر می‌شوند که به این معانی متفاوت اصطلاحا senseهای آن کلمه گفته می‌شود. در این رویکرد جهت محاسبه میزان شباهت دو جمله، پس از انتساب بهترین sense به هر کلمه و در واقع ابهام‌زدایی معنایی کلمات (با توجه به جمله‌ای که در آن قرار گرفته است)، از میزان هم‌پوشانی این senseها در دو جمله استفاده شده است [۲۲].

در زبان فارسی که در عرصه پردازش زبان طبیعی نوپا محسوب می‌شود کارهای مختلفی درعرصه خلاصه‌سازی متون انجام شده است که از جمله بارزترین آنها می‌توان به FarsiSum [۲۳]، سیستم خلاصه‌سازی خودکار متون فارسی [۲۴] و سیستم خلاصه‌سازی با استفاده از رویکرد انسان شناختی [۲۵] اشاره کرد.

در زمینه خلاصه‌سازی مبتنی بر کاربر متون فارسی اکثر تلاش‌های انجام شده به روش‌های مبتنی‌بر گراف متمرکز بوده است. به‌عنوان مثال سیستم خلاصه‌سازی خودکار متون فارسی [۲۴] قابلیت تولید هر دو شکل خلاصه (عمومی و مبتنی بر کاربر) را دارد و در آن متن به شکل گرافی غیرجهت‌دار که گره‌های تشکیل‌دهنده‌ی آن جمله‌ها هستند ارائه و شباهت بین دو جمله با لبه اتصال بین آنها نمایش داده می‌شود. معیار محاسبه این شباهت وجود کلمات مشترک یا رابطه بین کلمات آن با معیار کسینوسی است. زیرگراف‌ها نشان‌دهنده موضوعات موجود در سند هستند و در تولید خلاصه مبتنی‌بر کاربر، جملات خلاصه از زیرگراف خاصی مرتبط با عبارت پرس‌وجو انتخاب می‌شوند. [۲۶] نیز رویه‌ای مشابه سیستم خلاصه‌سازی خودکار متون فارسی در پیش می‌گیرد با این تفاوت که با جایگزینی مفاهیم متناظر کلمات، از رابطه‌های ترادف (synonym)، شمول (hypernymy) و زیرشمول (hyponymy) که بین این مفاهیم وجود دارد، جهت محاسبه شباهت با عبارت پرس‌وجو بهره می‌گیرد. [۲۷] نیز مبتنی‌برگراف بوده و در راستای محاسبه شباهت میان جملات از ویژگی‌های تعداد کلمات مشترک، تعداد کلمات کلیدی مشترک، تعداد کلمات انگلیسی مشترک و تعداد کلمات مشترک که توضیح آنها در پانویس آمده است با وزن‌های متفاوت استفاده

1. Web Hits
2. Snippets
3. Latent Semantic Analysis
4. Term Frequency-Inverse Document Frequency

می‌کند. این مقاله به معرفی روشی جهت محاسبه شباهت معنایی جملات که کمتر در کارهای بالا مورد توجه قرار گرفته است و کاربرد آن در ایجاد خلاصه مبتنی‌بر کاربر می‌پردازد.

۳. روش پیشنهادی

در این بخش به شرح روش پیشنهادی جهت تولید خلاصه‌ی مبتنی‌بر کاربر پرداخته می‌شود. در این روش، بعد از طی مراحل پیش‌پردازش موردنیاز و جداسازی جملات و کلمات مربوط به آنها، ابهام‌زدایی معنایی کلمات انجام می‌گیرد و سپس به محاسبه شباهت بین عبارت پرس‌وجو و جملات موجود در متن پرداخته می‌شود. در این راستا پس از تبدیل عبارت پرس‌وجو به جمله، ابتدا شباهت میان کلمات محاسبه شده و سپس براساس [۱۹] بردار شباهت چیدمان و شباهت معنایی تشکیل و میزان شباهت بین عبارت پرس‌وجو و جملات متن محاسبه می‌شود. در نهایت براساس این مقادیر، خلاصه‌سازی صورت می‌گیرد. در طی مراحل فوق، سیستم پیشنهادی از مجموعه‌ها و منابع موجود در زبان فارسی بهره برده است که پیش از بررسی فازهای سیستم پیشنهادی به معرفی این مجموعه‌ها پرداخته می‌شود.

۳. ۱. منابع مورد استفاده

این مقاله جهت محاسبه شباهت میان کلمات، تبدیل عبارت پرس‌وجو به جمله و محاسبه شباهت بین عبارت پرس‌وجو و جملات متن به ترتیب از ۳ مجموعه "فارس‌نت"، "فرهنگ ظرفیت افعال فارسی" و " بیجن‌خان" استفاده کرده‌است. در ادامه و به‌طور جداگانه هریک از این مجموعه‌ها مورد بررسی قرار می‌گیرد.

۳. ۱. ۱. فارس‌نت

در زبان فارسی، در زمینه شناسایی روابط بین کلمات تلاش‌هایی صورت گرفته است که از کارهای مهم انجام‌شده می‌توان به مجموعه فارس‌نت اشاره کرد. فارس‌نت، تلاشی جهت پیاده‌سازی مجموعه‌های مهمی همانند Wordnet در زبان فارسی است که در دانشگاه شهید بهشتی تدوین شده است [۲۸].

فارس‌نت شامل مجموعه‌ای از کلمات فارسی به همراه روابط بین آنها است. هر کلمه در این مجموعه دارای مشخصات زیر می‌باشد.

۳. ۱. ۱. ۱. Sense

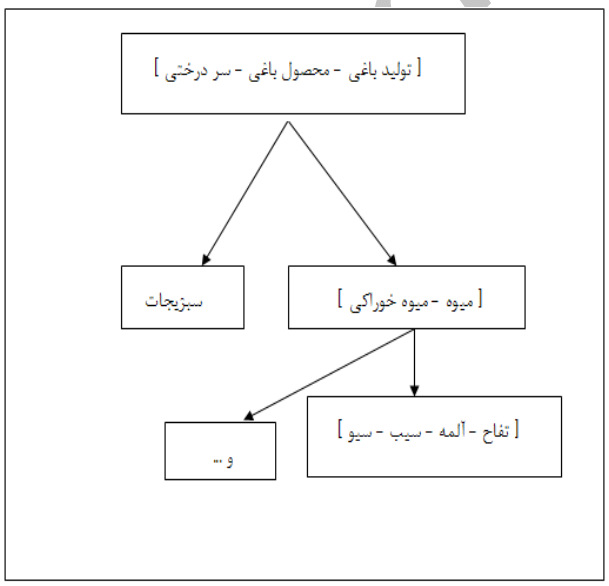
در مجموعه شبکه واژگانی، کلمات واحدی که در کاربردهای مختلف و با معانی مختلف ظاهر می‌شود، اصطلاحا دارای sensهای مختلف می‌باشند. به هر کدام از این sesnseها، شناسه‌ای اختصاص داده شده‌است که مستقل از زبان بوده و در تمام زبان‌های دنیا یکسان می‌باشد. روابط زیادی بین sense کلمات مختلف تعریف شده است؛ به‌طوری که این امکان وجود دارد که هر sense متعلق به کلمه‌ای با چندین sense، روابط متمایز و جداگانه‌ای با مجموعه‌ای از sesnseهای متعلق به سایر کلمات داشته باشد. تعداد روابط تعریف شده در فارس‌نت بیش از Wordnet می‌باشد و این مزیتی برای این مجموعه به حساب می‌آید.

۳. ۱. ۱. ۲. Synset

یکی از روابط مهم بین Sesnseها، رابطه Synonym یا هم‌معنی بودن است. دو Sense متعلق به دو کلمه، زمانی هم‌معنی هستند که حامل مفهوم یکسانی بوده و قابلیت جایگزینی با یکدیگر را داشته باشند. Synset به مجموعه‌ای از Senseها گفته می‌شود که دارای رابطه هم‌معنی بودن باشند؛ به‌عنوان مثال، کلمات "والد- بابا- پدر" یک Synset محسوب می‌شوند. در این مجموعه، هر Sense به یک Synset تعلق دارد. هر Synset در مجموعه فارس‌نت دارای مشخصات زیر می‌باشد:

- شناسه: به هر Synset، شماره منحصر به‌فردی اختصاص داده شده‌است.
- مقوله نحوی: نقش نحوی Sesnseهای موجود در Synset را نشان می‌دهد.
- اطلاعات نحوی: اطلاعاتی در مورد نحو هر کدام از کلمات موجود را نمایش می‌دهد.
- آوا: چگونگی تلفظ هر Sense را بیان می‌کند.
- تعریف[1]: مفهوم انتقالی Sesnseهای موجود را در قالب جمله کوتاهی به نمایش می‌گذارد.
- مثال: نمونه‌ای از کاربرد آنها را در جمله نشان می‌دهد.

روابط: نشان‌دهنده روابط Senseهای موجود در آن Synset با سایر مجموعه Synsetها می‌باشد. شکل ۱ نمایی از ساختار سلسله‌مراتبی این مجموعه را نشان می‌دهد [۲۹].



شکل ۱. نمایی از ساختار سلسله مراتبی فارس‌نت

1. Gloss

۳. ۱. ۲. فرهنگ ظرفیت افعال فارسی

این مجموعه حاوی اطلاعات مربوط به ظرفیت نحوی بیش از ۴۵۰۰ فعل در زبان فارسی است. در این منبع، شکل‌های مختلف افعال ساده، مرکب، پیشوندی و عبارات فعلی به‌همراه متمم‌های اجباری و اختیاری آنها مشخص شده است.

با توجه به اینکه شناخت افعال مرکب، چه از لحاظ انسانی و چه از لحاظ پردازشی، کاری دشوارتر از شناخت افعال ساده است، گردآورندگان این منبع سعی در فراهم کردن فهرستی از افعال مرکب به همراه ساخت‌های ظرفیتی آنها کرده‌اند. بنا به نظر گردآوران، ظرفیت از آن جهت اهمیت دارد که تعیین‌کننده وجود عناصری از جمله است که حضورشان به واسطه ویژگی‌های منحصر به فرد فعل مرکزی جمله ضروری بوده (ظرفیت‌های اجباری) یا توجیه می‌شود (ظرفیت‌های اختیاری). این منبع در کارهای پردازشی زبان و کمک به غیرفارسی‌زبانان، نقش مهمی ایفا می‌کند [۳۰].

۳. ۱. ۳. پیکره بیجن‌خان

این پیکره شامل تعداد زیادی کلمه به همراه برچسب و فرکانس آنها است که در دانشگاه تهران و از برخی اخبار روزنامه‌ها و متون معمولی جمع‌آوری شده‌است. برخی از ویژگی‌های این پیکره به قرار زیر است [۳۱]:

- این پیکره شامل ۲۵۹۸۲۱۵ واژه و ۵۵۰ برچسب می‌باشد که به‌طور دستی برچسب زده شده است.
- هر سند در این مجموعه دارای یک عنوان می‌باشد. به‌عنوان مثال، اسناد تحت عناوین (سیاسی، فرهنگی، اقتصادی) دسته‌بندی شده‌اند.
- در این پیکره ۴۳۰۰ عنوان مختلف وجود دارد. این دسته بندی بزرگ نشان‌دهنده کیفیت بالای این پیکره می‌باشد.

آمار بالای پیکره بیجن‌خان این امکان را می‌دهد که از آن به‌عنوان منبعی برای محاسبه فرکانس کلمات در دنیای واقعی استفاده شود. در این مقاله از متونی با عنوان "سیاسی" استفاده شده است.

در ادامه به بررسی فازهای پیشنهادی مقاله، جهت تولید خلاصه مبتنی بر عبارت پرس‌وجوی کاربر پرداخته می‌شود.

۳. ۲. پیش‌پردازش

پیش‌پردازش متن، مرحله آماده‌سازی متن برای ورود به سامانه‌های پردازشی متون می‌باشد. سامانه‌های پردازشی متون، در صورتی عملکرد مناسبی ارائه می‌دهند که ورودی مناسب برای آنها تامین شود. هرچه پیش‌پردازش متن، با کیفیت بالاتری انجام شود، کارایی سامانه بیشتر خواهد شد. الگوریتم پیشنهادی، با کمک ابزارهای زیر مراحل پیش‌پردازش را انجام می‌دهد.

۳. ۲. ۱. نرمال ساز[1]

نرمال‌ساز استفاده شده، آشفتگی موجود در متون فارسی مانند انواع مختلف نگارش برای یک کلمه، تنوع استفاده از "می" و "ها" چسبان و غیرچسبان، تنوع نگارش "ی" اضافه در کلمات مختوم به "ه" و فواصل متفاوت میان کلمات را از بین برده و اصلاح می‌نماید. در واقع، یکسان‌سازی‌هایی در این بخش اعمال می‌شود که در نتیجه آن، نظم موردنظر در متن حکمفرما می‌شود.

۳. ۲. ۲. واحد ساز[2]

واحدساز با تشخیص کران کلمات در متن، دنباله کلمات آن را استخراج می‌کند. با توجه به نظام پیچیده صرف فارسی، تعداد وندهای زیاد و تصریف‌های متعدد کلمات، این بخش با چالش‌های زیادی روبه‌رو خواهد بود که به‌عنوان نمونه می‌توان موارد زیر را نام برد:

- تشخیص کلمات مرکب
- تشخیص اسامی خاص

به این منظور، از لیستی از کلمات مرکب (متشکل از ۱۶۰۰۰ کلمه)، لیستی از اسامی افراد مطرح در حوزه" سیاسی" (به همراه سمت آنها) و لیستی از اسامی شهرها و کشورها استفاده و به این ترتیب، کلمات از متن استخراج می‌شوند.

۳. ۲. ۳. برچسب‌گذار ادات سخن[3]

این ابزار، عمل انتساب مقولات واژگانی به کلمات متن را انجام می‌دهد. از مهم‌ترین مقولات واژگانی اسم، فعل، صفت و قید می‌باشد. درصد بالایی از کلمات، از نقطه‌نظر برچسب‌واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های متفاوتی دارند. بنابراین، برچسب‌گذاری واژگانی باید با توجه به بافت کلمات انجام شود.

۳. ۲. ۴. جمله‌یاب

در این بخش علاوه بر علائم نگارشی "."، "؛" و ":" کلمات ربط "و"، "بلکه"، "که"، "حال‌آنکه"، "ضمن‌آنکه" و "تا" نیز در صورتی که برچسب کلمه قبل آنها فعل باشد، به‌عنوان انتهای جمله در نظر گرفته شدند. شایان ذکر است که برای حفظ پیوستگی و خوانایی متن، در صورت انتخاب هر یک از جملات تفکیک‌شده با "بلکه"، "که"، "حال‌آنکه"، "ضمن‌آنکه" و "تا" می‌توان جمله دیگر را نیز در خلاصه شرکت داد.

۳. ۲. ۵. ریشه‌یاب[4]

در ریشه‌یابی کلمات، انتخاب ریشه مناسب کلمات با توجه به

1. Normalizer
2. Tokenizer
3. Part-Of-Speech Tagger
4. Stemmer

بافتی که کلمه در آن ظاهرشده و برچسب کلمات همسایه آن انجام می‌شود. در همین راستا از مجموعه‌ای از قوانین اکتشافی[1] استفاده شده‌است.

۳. ۳. ابهام‌زدایی معنایی کلمات

ابهام معنایی یکی از انواع ابهام در سطوح مختلف زبان می‌باشد و ابهام‌زدایی معنایی کلمات (WSD[2]) در واقع پردازشی است که براساس آن معنی یا sense درست کلمه مبهم براساس متنی که در آن به‌کار برده شده‌است شناسایی می‌شود [۳۲]. برای محاسبه میزان شباهت کلمات نیز در وهله‌ی اول نیاز به ابهام‌زدایی معنایی آنها داریم. بنابراین پس از طی کردن فاز پیش‌پردازش، نیاز به رفع ابهام کلمات احساس می‌شود.

یکی از روش‌های رفع ابهام sense کلمات و یا همان WSD استفاده از روش‌های آماری است. مدل‌های مبتنی بر آمار که از روش‌های با نظارت برای نسبت دادن sense مناسب به کلمات مبهم استفاده می‌نمایند، نیاز به یک پیکره زبانی غنی و برچسب‌گذاری‌شده دارند. متاسفانه در زبان فارسی چنین پیکره‌ای موجود نمی‌باشد؛ به همین منظور پیکره کوچکی در این بخش ایجاد شده‌است.

از آنجایی که معمولاً کلمات در حوزه‌های موضوعی مختلف دارای کاربردهای متفاوتی هستند تفکیک این حوزه‌ها در ایجاد پیکره زبانی و به تبع آن استفاده از روش‌های آماری تا حدودی راه‌گشا خواهد بود. به‌عنوان مثال جملات زیر در نظر گرفته می‌شود:

- "کاهش تورم از مسائل اساسی کشور است."
- "از جمله علائم این بیماری تورم مفاصل است."

کلمه "تورم" در حوزه‌ی موضوعی "سیاسی" اغلب با sense "افزایش قیمت" به‌کار برده می‌شود در حالی که در حوزه‌ی "پزشکی"، "حجیم‌شدن بخشی از بدن" مد نظر است.

با توجه به تعدد حوزه‌های موضوعی و وقت‌گیر بودن بررسی همه آنها، در اینجا از حوزه‌ی موضوعی "سیاسی" استفاده شده است. جهت ساخت پیکره مورد نیاز، ۵۰ خبر فارسی در حوزه‌ی موردنظر و از پایگاه اطلاع‌رسانی "فارس‌نیوز" انتخاب و توسط ۲ فرد خبره برچسب‌گذاری شده است. به این ترتیب در مجموعه ایجاد شده به ازای هر کلمه، sense مرتبط با جمله انتخاب شده‌است. از خروجی حاصل می‌توان جهت آموزش ابهام‌زدایی معنایی کلمات استفاده کرد.

در این مقاله برای تعیین معنای صحیح یک کلمه مبهم با استفاده از پیکره‌های زبانی، از تلفیق دو روش استفاده از اطلاعات توزیعی و استفاده از واژه‌های بافتی به صورت زیر استفاده شده‌است:

- با توجه به پیکره آموزشی ایجادشده، senseی که بیش‌ترین احتمال رخداد را داشته باشد، به‌عنوان sense برگزیده خواهد بود.
- در صورت عدم وجود کلمه موردنظر در پیکره و یا مساوی بودن احتمال sense‌ها، سیستم به ازای هر sense ممکن برای کلمه موردنظر، sense‌های دیگری را که براساس مجموعه فارس‌نت، با آن کلمه رابطه‌ی "مرتبط" و یا "هم‌وقوع" دارند، می‌یابد. senseی که مجموعه sense‌های "مرتبط" و "هم‌وقوع" با آن، بیش‌ترین هم‌پوشانی را با جمله‌ای که کلمه موردنظر در آن قرار دارد داشته باشد، به‌عنوان sense برگزیده خواهدشد.

۳. ۴. تبدیل عبارت پرس‌وجو به جمله

در اغلب موارد، عبارت پرس‌وجوی واردشده توسط کاربر در قالب یک عبارت است و شکل یک جمله کامل را ندارد. از آنجایی که الگوریتم پیشنهادی این مقاله، از روشی مناسب محاسبه شباهت دو جمله برای خلاصه‌سازی استفاده می‌کند، در این مرحله روشی برای ساخت جمله از روی عبارت پرس‌وجو ارائه می‌شود.

در اینجا به دلیل متنوع بودن عبارت‌های پرس‌وجو، این عبارت‌ها به عبارت‌های تیتروار محدود می‌شود که در آن اولین کلمه عبارت، بار معنایی فعل را دارد. مثال زیر نمونه‌ای از این عبارت‌ها را نمایش می‌دهد:

- "سفر رئیس‌جمهور به نیویورک"

همان‌طور که گفته شد، اولین کلمه از عبارت‌های پرس‌وجویی که نظایر آن در مثال‌های فوق دیده می‌شود، بار معنایی فعل را دارا می‌باشند. به‌همین دلیل، جهت ساخت جمله از روی این عبارت‌ها، کلمه ابتدایی به انتهای عبارت منتقل می‌شود.

با توجه به اینکه در بخش شناسایی روابط میان کلمات، از ریشه‌ی آنها استفاده می‌شود؛ برای ساخت جمله از روی عبارت‌های پرس‌وجوی مورد بررسی، فقط کافی است که کلمات منتقل‌شده، به مصدر فعلی مناسب خود تبدیل شوند. در این میان، نکته مهم و قابل‌توجه این است که شکل‌های مصدری متفاوتی برای بعضی از این کلمات متصور می‌باشد؛ بنابراین، باید به نحوی به صحیح‌ترین شکل دست یافت.

در این راستا و برای رسیدن به صحیح‌ترین شکل، نیاز به حضور لیستی از حالت‌های مصدری ممکن برای این کلمات می‌باشد. به همین منظور، تمام حالت‌های ممکن در فارس‌نت که با کلمه موردنظر آغاز شده و دارای مقوله نحوی فعل باشند، بررسی و synsetی که مصدر موردنظر به آن تعلق دارد، تعیین می‌شود. در صورت متفاوت بودن این synset‌ها، مصدر صحیح از بین حالت‌های به‌دست آمده و با کمک فرهنگ ظرفیت افعال فارسی انتخاب می‌شود. در واقع با استفاده از این فرهنگ، تعداد اجزای موردنیاز (ظرفیت‌های اجباری) برای هر یک از افعال به‌دست می‌آید و براساس آن، تصمیم‌گیری صورت می‌گیرد.

۳. ۵. محاسبه شباهت میان کلمات

جهت محاسبه شباهت میان کلمات در روش به‌کار گرفته‌شده، از مجموعه فارس‌نت استفاده شده‌است. به ازای هر دو کلمه، توابع

1. Heuristic
2. Word Sense Disambiguation

فارس‌نت فراخوانده شده، رابطه بین آن دو کلمه به‌دست آمده و با توجه به این رابطه، میزان شباهت میان دو کلمه مقداردهی شده‌است.

الگوریتم محاسبه میزان شباهت، این گونه طراحی شده‌است که اگر رابطه به‌دست‌آمده میان دو کلمه، یکی از روابط موجود در فارس‌نت باشد، با توجه به آن رابطه، مقداری به میزان شباهت اختصاص داده می‌شود؛ در غیر این‌صورت، الگوریتم به‌دنبال اولین پدر مشترک دو کلمه در ساختار سلسله‌مراتبی می‌گردد و از محاسبه تعداد یال‌های موجود در مسیر، برای محاسبه شباهت دو کلمه استفاده می‌کند [۲۰]. به عنوان مثال در شکل ۱ اولین پدر مشترک بین دو synset [تفاح - آلمه - سیب - سیو] و [سبزیجات]، synset [تولید باغی- محصول باغی- سر درختی] و تعداد یال‌های سپری شده در مسیر بین دو synset، ۳ می‌باشد. مسلماً هر چقدر این تعداد یال بیش‌تر باشد، فاصله معنایی بین دو synset نیز بیش‌تر خواهد بود.

۳. ۵. ۱. محاسبه میزان شباهت بین دو کلمه با توجه به رابطه بین آنها

پس از به‌دست آوردن نوع رابطه بین کلمات، به انتساب میزان دقیق شباهت نیاز می‌باشد. به‌همین منظور، لیستی از جفت کلمات به همراه نوع رابطه آنها در اختیار ۲ فرد خبره قرار گرفت تا در رابطه با میزان شباهت میان جفت کلمات اظهار نظر کنند.

خروجی این لیست که از هر رابطه به تعداد کافی نمونه در آن آورده شده است، میانگین میزان شباهت‌های به‌دست آمده به ازای هر رابطه می‌باشد. جدول این مقادیر را نمایش می‌دهد.

جدول ۱. میانگین میزان شباهت به‌دست‌آمده براساس رابطه بین دو کلمه

نام رابطه	میزان شباهت
اشتقاقی (derivational)	۰/۸
شمول (hypernymy)	۰/۸
زیرشمول (hyponymy)	۰/۹
جزءواژگی (meronym)	۰/۸
کل واژگی (holonymy)	۰/۸
تضاد (Antonym)	-۱
ترادف (synonym)	۱
سببیت (cause)	۰/۸
استلزام (entailment)	۰/۸
مرتبط (related-to)	۰/۵
هم‌وقوع (co-occurrent)	۰/۸
صفت برجسته	۰/۷
ویژگی (attribute)	۰/۷
ابزار (instrument)	۰/۸
عامل (agent)	۰/۸
پذیرا (patient)	۰/۸۵

اگر نوع رابطه بین دو کلمه هیچ کدام از روابط موجود در مجموعه فارس‌نت نباشند، سیستم به‌دنبال اولین پدر مشترک دو کلمه در ساختار سلسله مراتبی می‌گردد و با محاسبه تعداد یال‌های این مسیر، میزان شباهت را محاسبه می‌کند. رابطه (۱) این رابطه را نمایش می‌دهد [۱۹].

$$sim(w_1,w_2)=e^{-\alpha len} \quad (۱)$$

که در آن len طول مسیر بین دو کلمه در ساختار سلسله‌مراتبی و α ضریب ثابتی است که از افزایش بیش از حد میزان محاسبه‌شده شباهت جلوگیری کرده و با سعی و خطا به‌دست می‌آید (در اینجا مقدار ۰.۲ برای α در نظر گرفته شده است).

به این ترتیب و با توجه به رابطه (۱) دو شرط زیر ارضا می‌شود [۱۹]:

- اگر دو کلمه عین هم باشند و در واقع به ازای len=0، میزان شباهت آنها برابر ۱ خواهد بود.
- میزان شباهت بین هر دو کلمه، مقداری بین ۰ و ۱ را اختیار خواهد کرد.

۳. ۶. محاسبه شباهت بین عبارت پرس‌وجو و جملات متن ورودی

پس از اعمال فاز پیش‌پردازش و تبدیل عبارت پرس‌وجو به جمله متناظر، میزان شباهت جملات متن و جمله ایجاد شده از عبارت پرس‌وجو محاسبه می‌شود. روش به‌کار گرفته‌شده در این مقاله جهت محاسبه میزان شباهت بین دو جمله به این ترتیب می‌باشد که به ازای هر جمله، یک بردار شباهت معنایی و یک بردار شباهت چیدمان کلمات ساخته شده و میزان شباهت نهایی بین دو جمله با ترکیب وزن‌دار این دو بردار محاسبه می‌گردد. شکل ۲ روند این الگوریتم را نمایش می‌دهد.

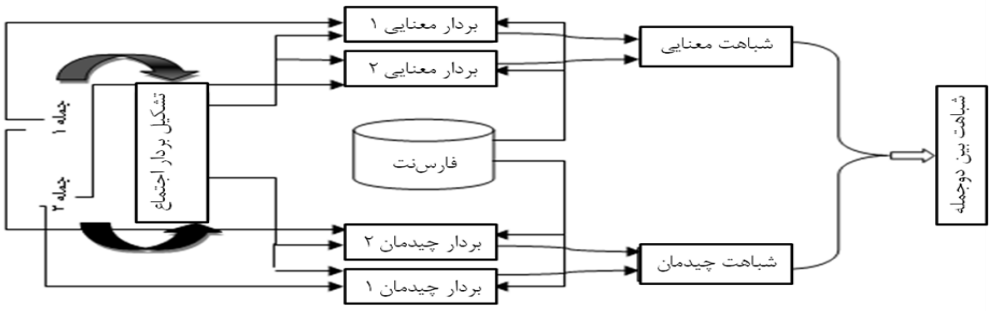
۳. ۶. ۱. بردار شباهت معنایی

برای ایجاد بردار شباهت معنایی، اجتماع کلمات دو جمله در بردار T تشکیل می‌شود که شامل تمام کلمات متمایز دو جمله است. به بیان دیگر بردار T به‌صورت رابطه (۲) به‌دست می‌آید.

$$T = T_1 \cup T_2 = \{w_1 w_2 \dots w_m\} \quad (۲)$$

به‌دلیل نقش غیرقابل انکار کلمات توقف در شکل‌دهی نحوی و معنایی جملات، این کلمات حذف نمی‌شوند و در ایجاد بردار T این کلمات نیز لحاظ می‌شوند. سپس به‌ازای هر جمله، برداری به‌طول m به‌صورت $S_{i\,(i=1,2,..,m)}$ ساخته می‌شود.

برای پر کردن این بردار، ابتدا مقدار $\widehat{S}$ محاسبه می‌شود. برای هر کلمه از بردار T، اگر عین آن کلمه در جمله موجود باشد، $\widehat{S}$ برابر ۱ و در غیر این‌صورت $\widehat{S}$ برابر میزان شباهت شبیه‌ترین کلمه از جمله به w_i خواهد بود (این میزان شباهت با یک حد آستانه (β) مقایسه و در صورتی‌که از آن کمتر باشد، مقدار صفر برای $\widehat{S}$ در نظر گرفته می‌شود).



شکل ۲. روند الگوریتم محاسبه شباهت معنایی بین دو جمله

در این الگوریتم، تمامی کلمات از جمله کلمات توقف حفظ خواهند شد. البته تمامی این کلمات ارزش یکسانی در محاسبه شباهت بین دو جمله نخواهند داشت و این مسئله باید مدنظر قرار گیرد [۱۱-۱]. بنابراین، S_i به‌صورت رابطه (۳) به‌دست می‌آید:

$$s_i = \hat{s} * I(w_i) * I(\hat{w}_i) \quad (۳)$$

که $\hat{w}_i$ برابر شبیه‌ترین کلمه از جمله به w_i و $I(w_i)$ و $I(\hat{w}_i)$ به ترتیب متناسب با فرکانس w_i و $\hat{w}_i$ در "پیکره بیجن‌خان" است.

به این ترتیب و با پیمودن مراحل فوق، شباهت معنایی بین دو جمله به‌صورت رابطه (۴) به‌دست می‌آید:

$$S_s = \frac{s_1 . s_2}{\|s_1\| . \|s_2\|} \quad (۴)$$

۳. ۶. ۲. بردار شباهت چیدمان

برای محاسبه میزان شباهت چیدمان کلمات دو جمله نیز، از بردار T که حاوی کلمات متمایز دو جمله است و در رابطه (۲) معرفی شد، استفاده می‌شود. برای ایجاد بردار شباهت چیدمان کلمات دو جمله، ابتدا کلمات موجود در هر دو جمله به ترتیب اندیس‌گذاری می‌شوند. به ازای هر کلمه w_i در T، اندیس شبیه‌ترین کلمه از جمله ۱ و ۲ در بردار r متناظر آن جمله یعنی r_1 و r_2 به این صورت قرار داده می‌شود:

- اگر عین w_i در جمله موجود باشد، اندیس آن در بردار r قرار داده می‌شود.
- اگر w_i در جمله موجود نباشد و میزان شباهت آن به شبیه‌ترین کلمه‌ی جمله از یک حد آستانه بیشتر باشد، اندیس شبیه‌ترین کلمه در r قرار می‌گیرد.
- اگر w_i در جمله موجود نباشد و میزان شباهت شبیه‌ترین کلمه به آن از حد آستانه کمتر باشد، مقدار ۰ در r قرار داده می‌شود.

به این ترتیب شباهت چیدمان دو بردار r_1 و r_2 به‌صورت رابطه (۵) به‌دست می‌آید:

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (۵)$$

۳. ۶. ۳. شباهت نهایی بین دو جمله

در بخش‌های قبل با محاسبه دو بردار S_s و S_r، شباهت معنایی و شباهت چیدمان کلمات دو جمله به‌دست آمد. برای محاسبه شباهت نهایی بین دو جمله از ترکیب وزن‌دار این دو استفاده می‌شود [۷]. رابطه (۶) به این منظور به کار برده می‌شود:

$$S(T_1, T_2) = \delta S_s + (1-\delta) S_r \quad (۶)$$

۳. ۷. تولید خلاصه

پس از گذراندن تمام مراحل فوق، در این مرحله می‌توان به تولید خلاصه مورد نیاز کاربر پرداخت (خلاصه‌ای که با توجه به‌عبارت پرس‌وجوی کاربر و نرخ فشردگی مورد نیاز باشد). در این راستا، به هر جمله امتیازی با توجه به رابطه (۷) اختصاص داده می‌شود:

$$Score_{s_i} = S(s_i, q) \quad (۷)$$

که در آن $S(s_i, q)$ میزان شباهت جمله iام و عبارت پرس‌وجو را نشان می‌دهد. به این ترتیب، جملاتی که بیشترین شباهت را به عبارت پرس‌وجو داشته باشند، امتیاز و شانس بیشتری برای حضور در خلاصه خواهد داشت.

به این ترتیب، جملات با توجه به نرخ فشرده‌سازی، امتیازات اختصاص داده‌شده به آنها و ترتیب حضور در متن اصلی، در خلاصه حضور می‌یابند. در ضمن، همانطور که در بخش ۰ به آن اشاره شد؛ جهت حفظ پیوستگی و خوانایی متن، اگر یکی از جملاتی که با "بلکه"، "که"، "حال‌آنکه"، "ضمن‌آنکه"، "تا"و علامت نقل‌قول تفکیک شده‌اند، امتیاز لازم برای حضور در خلاصه را کسب کنند، جمله دیگر نیز در خلاصه حضور پیدا می‌کند.

۴. ارزیابی نتایج

جهت ارزیابی، روش پیشنهادی بر روی ۵۰ خبر برگزیده از منبع خبری "فارس‌نیوز" و در حوزه‌ی موضوعی "سیاسی" پیاده شد. در ادامه به بررسی نتایج کسب‌شده پرداخته می‌شود.

۴. ۱. ارزیابی نحوه‌ی تفکیک جملات

به‌منظور ارزیابی روش ارائه شده جهت تفکیک جملات، الگوریتم موردنظر بر روی ۱۷ خبر متشکل از ۵۰۰ جمله اعمال شد. جدول ۲ دقت به‌دست آمده از روش موردنظر در مقایسه با تفکیک انسانی را نمایش می‌دهد.

جدول ۲. میزان دقت روش پیشنهادی جهت تفکیک جملات

تعداد اخبار	تعداد جملات	دقت به‌دست آمده
۱۷	۵۰۰	۰/۹۱

اکثر ناسازگاری‌های روش موردنظر مربوط به جملاتی است که از فرم محاوره‌ای برخوردار هستند مانند جمله زیر که در متن یکی از اخبار برگزیده به چشم می‌خورد:

- "نفوذ در وجود ملتی بزرگ با این همه بسیجی آگاه و فداکار آرزویی است دست نیافتنی که می‌بایست همراه خود به گور ببرند."

روش پیشنهادی، مثال فوق را یک جمله در نظر می‌گیرد؛ در حالی که مثال از دو جمله تشکیل شده است این امر از فرم محاوره‌ای مثال بالا ناشی می‌شود و در غیر اینصورت جملاتی مانند مثال زیر به‌درستی تفکیک می‌شوند:

- "۱. نفوذ در وجود ملتی بزرگ با این همه بسیجی آگاه و فداکار آرزویی دست نیافتنی است که ۲. می‌بایست همراه خود به‌گور ببرند."

۴. ۲. ارزیابی الگوریتم محاسبه شباهت میان جملات

جهت ارزیابی روش محاسبه شباهت معنایی بین جملات که در بخش ۰ تشریح شد، الگوریتم موردنظر بر روی ۵۰ جمله از اخبار منتخب اعمال شد. با توجه به اینکه مجموعه داده محک مناسبی در زمینه محاسبه شباهت جملات به‌خصوص در زبان فارسی وجود ندارد؛ برای ارزیابی نتایج حاصل از روش ذهنی[1] استفاده و جملات به ۲ فرد خبره ارائه شد تا نظر خود را درمورد شباهت این جملات، در چهار سطح فازی مورد بررسی اعلام نمایند. این سطح‌بندی فازی در جدول نمایش داده شده است.

سپس نظرات این افراد توسط معیاری به‌نام کاپا ارزیابی‌گردید. این معیار برای ارزیابی سازگاری نظرات خبرگان به‌عنوان یک ابزار اندازه‌گیری کیفیت و دقت اطلاعات استفاده و در آن میزان توافق نظرات افراد خبره محاسبه می‌شود. اگر این میزان یا به‌عبارتی کیفیت حاشیه‌نویسی انجام شده از ۰/۷ بیشتر باشد، به‌عنوان ارزیابی روش پیشنهادی استفاده می‌شود. جدول نمونه‌هایی از مقایسه نتایج روش‌پیشنهادی با حاشیه‌نویسی‌های انسانی را نشان می‌دهد. همان‌طور که این جدول نشان می‌دهد، روش پیشنهادی به نتایجی هم‌سو با حس انسانی رسیده است و این نشان از موفقیت نسبی روش مورد نظر دارد.

جدول ۳. محدوده‌های سطوح فازی میزان شباهت جملات

محدوده‌ها	سطوح فازی
[۰-۰/۱]	بدون شباهت
(۰/۱-۰/۴)	کم
[۰/۴-۰/۷)	متوسط
[۰/۷-۱]	زیاد

۴. ۳. ارزیابی روش پیشنهادی خلاصه‌سازی مبتنی‌بر کاربر

در این بخش به‌ازای هر خبر، دو عبارت پرس‌وجو در نظر و خلاصه‌سازی براساس آن صورت گرفته است. از بین این اخبار، ۳۰ خبر در مرحله ابهام‌زدایی معنایی کلمات، برچسب‌گذاری شده و سایر اخبار براساس یادگیری صورت گرفته، رفع ابهام شد.

جهت ارزیابی روش خلاصه‌سازی پیشنهادی، از یکی از روش‌های ارزیابی درونی یعنی ارزیابی مبتنی‌بر انتخاب همزمان[2] استفاده شد که در آن خلاصه‌های تولیدی با خلاصه‌های ایده‌آل مقایسه می‌شوند. در این روش، مقایسه‌ای بین جملات انتخاب شده و جملات خلاصه صورت گرفته و براساس آن در مورد روش پیشنهادی قضاوت می‌شود. در این راستا، سه معیار دقت[3]، بازخوانی[4] و F[5] مورد استفاده قرار گرفتند.

معیار دقت به‌صورت نسبت تعداد جملاتی که همزمان در خلاصه تولیدی و در خلاصه مرجع رخ داده‌اند، به تعداد جملات موجود در خلاصه تولیدی و معیار بازخوانی، به صورت نسبت تعداد جملات رخ داده‌شده در خلاصه تولیدی و مرجع به تعداد جملات موجود در خلاصه مرجع، تعریف می‌شود. همچنین با ترکیب هارمونیک این دو معیار، معیار F حاصل می‌شود. به همین منظور، از ۲ فرد خبره خواسته شد تا براساس اخبار موجود و عبارت‌های پرس‌وجوی مدنظر خلاصه‌سازی استخراجی را انجام دهند. به این ترتیب و براساس خلاصه‌های مرجع، نتایج جدول ۵ حاصل شد.

1. Subjective

2. Co-Selection Evaluation
3. Precision
4. Recall
5. F_Measure

جدول ۴. مقایسه میزان شباهت جملات با روش ذهنی

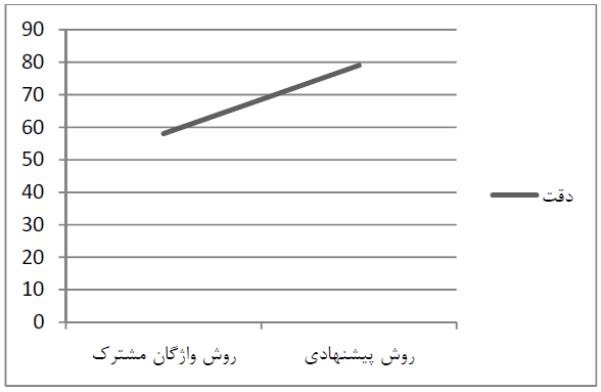
جملات	سطح فازی شباهت با استفاده از روش پیشنهادی	سطح فازی شباهت با روش ذهنی
- دولت ایران روز چهارشنبه حمله اسرائیل به سودان را محکوم کرد. - اسرائیل در حمله به سوریه یک ساختمان نظامی و موشکی را هدف قرار داده است.	متوسط	متوسط
- ارتش سوریه حمله جنگنده‌های اسرائیلی به یک مرکز پژوهشی نظامی در حومه دمشق را تأیید کرد. - دبیر شورای عالی امنیت ملی جمهوری اسلامی ایران لحظاتی پیش به فرودگاه دمشق وارد شد.	بدون شباهت	بدون شباهت
- موضع ولایت برای ما حجت شرعی است. - مردم با حضور پرشور و گسترده خود در انتخابات ریاست جمهوری توطئه‌های دشمنان را خنثی می‌کنند.	کم	کم
- علی لاریجانی با ۱۷۷ رأی موافق نمایندگان ملت برای اولین سال فعالیت مجلس نهم رئیس مجلس شد. - تصویب بودجه سه‌شنبه در دستور کار مجلس قرار خواهد گرفت.	متوسط	متوسط
-مسئولین باید آرامش سیاسی کشور را حفظ کنند. -جلسه علنی مجلس شورای اسلامی با ریاست علی لاریجانی آغاز شد.	بدون شباهت	بدون شباهت
- تجربه نظام در برگزاری انتخابات رقابتی ذی قیمت است. - میزان مشارکت در انتخابات آمریکا اندکی کمتر از ۵۰ درصد بوده است.	کم	کم
- مردم با حضور پرشور و گسترده خود در انتخابات ریاست جمهوری توطئه‌های دشمنان را خنثی می‌کنند. - فرآیند اخذ رأی الکترونیکی در انتخابات سال آینده برای تعداد معدودی از شعب در نظر گرفته شده است.	متوسط	متوسط
-۹دی مهر پایان جریان‌های برانداز انقلاب بود. - جزیره خارک نماد مقاومت و پایداری است.	کم	کم
- سید علی اکبر طاهایی در پیامی درگذشت حسن حبیبی را تسلیت گفت. - رهبر معظم انقلاب اسلامی در پیامی رحلت آیت‌الله مجتبی تهرانی را تسلیت گفتند.	زیاد	زیاد

جدول ۵. نتایج حاصل از روش پیشنهادی

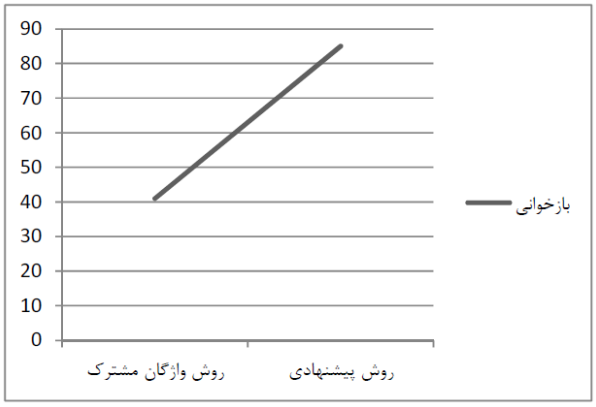
میانگین دقت	میانگین بازخوانی	میانگین F
٪۷۷	٪۸۵	٪۸۱

۴. ۴. مقایسه روش پیشنهادی با روش واژگان مشترک

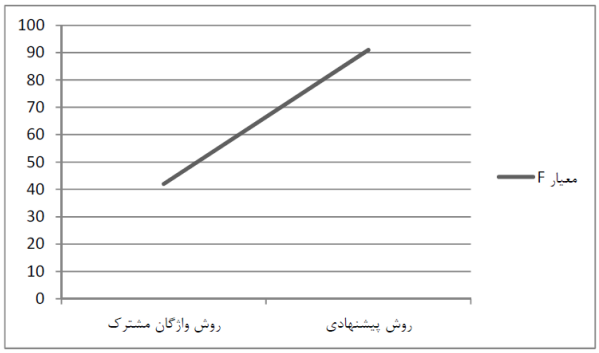
جهت مقایسه روش پیشنهادی، از روش پایه واژگان مشترک استفاده شد. بر مبنای این روش، به هر جمله، امتیازی متناسب با تعداد کلمات مشترک با عبارت پرس‌وجو اختصاص داده می‌شود و جملات با امتیاز بالاتر در خلاصه حضور پیدا می‌کنند[۳۳]. شکل‌های ۳، ۴ و ۵ به‌ترتیب مقایسه دقت، بازخوانی و معیار F روش پیشنهادی با روش پایه واژگان مشترک را نمایش می‌دهند.



شکل ۳. مقایسه دقت روش پیشنهادی با روش واژگان مشترک



شکل ۴. مقایسه بازخوانی روش پیشنهادی با روش واژگان مشترک



شکل ۵. مقایسه معیار F روش پیشنهادی با روش واژگان مشترک

نمودارها بیانگر این مطلب است که روش پیشنهادی به موفقیت قابل‌قبولی نسبت به روش پایه دست یافته‌است.

۵. نتیجه‌گیری

در این مقاله به یکی از انواع خلاصه‌سازی‌های خودکار یعنی خلاصه‌سازی مبتنی بر عبارت پرس‌وجوی ارائه شده توسط کاربر بر روی زبان فارسی پرداخته شده است. این نوع سیستم‌ها بر خلاف خلاصه‌سازهای عمومی، خلاصه‌ای با توجه به نیاز کاربر تولید می‌کنند. در واقع با توجه به عبارت واردشده، جملاتی از متن اصلی برای حضور در خلاصه برگزیده خواهند شد.

رویکرد پیشنهادی این مقاله ابتدا و در فاز پیش پردازش نرمال‌سازی، واحدسازی، ریشه‌یابی و برچسب‌گذاری متن را انجام می‌دهد و در ادامه به ابهام‌زدایی معنایی کلمات و تعیین روابط میان آنها با استفاده از مجموعه فارس‌نت می‌پردازد.

بعد از محاسبه شباهت میان کلمات، الگوریتم پیشنهادی پس از تبدیل عبارت پرس‌وجو به جمله، به سراغ محاسبه شباهت میان عبارت پرس‌وجو و جملات متن خواهد رفت.

پس از طی مراحل فوق، به هر جمله از متن اصلی براساس میزان شباهت با عبارت پرس‌وجو امتیازی اختصاص می‌یابد. در نهایت جملات، براساس امتیاز کسب شده و تا زمانی که از نرخ فشردگی تجاوز نکنند در خلاصه حاضر خواهند شد.

روش پیشنهادی بر روی ۵۰ خبر فارسی و در حیطه موضوعی "سیاسی" اعمال شده است. ارزیابی‌های صورت گرفته براساس معیار دقت، بازخوانی و در نهایت معیار F نشان از مطلوبیت نسبی روش فوق دارد.

برای بهبود نتایج روش پیشنهاد شده در این مقاله پیشنهاداتی مطرح می‌گردد:

- برای ابهام‌زدایی معنایی کلمات، پیکره‌ای برچسب‌گذاری شده در یک حوزه معنایی خاص تهیه و طریق آن فرآیند رفع ابهام به‌صورت آماری انجام شد. برای بهبود نتایج این بخش می‌توان به توسعه پیکره تولیدی پرداخت. همچنین می‌توان از یکسری قواعد اکتشافی دیگر در این زمینه بهره برد.
- در این مقاله جهت تعیین میزان شباهت میان کلمات از مجموعه فارس‌نت استفاده شد و سایر روش‌ها و راهکارهای ممکن در زبان فارسی به‌دلیل محدودیت در کاربرد نادیده گرفته شد. می‌توان از ترکیب روش‌های موجود بهره برد.
- همان‌طور که گفته شد در هنگام محاسبه میزان شباهت دو کلمه، اگر دو کلمه هیچ ارتباطی با هم نداشتند سیستم به‌دنبال اولین پدر مشترک دو کلمه در سیستم سلسله‌مراتبی فارس‌نت خواهد رفت و بر اساس آن میزان شباهت را تعیین خواهد کرد. در این راستا از رابطه (۱) استفاده شد. در این رابطه ضریب $\propto$ با سعی و خطا به‌دست آمد که می‌توان به جای آن از روش‌های هوشمند استفاده نمود.
- در مرحله محاسبه میزان شباهت بین دو جمله نیز، ضرایب ترکیب شباهت نحوی و معنایی براساس سعی و خطا و به صورت دستی تنظیم شده است. می‌توان این ضرایب را هم، با روش‌هایی مانند الگوریتم ژنتیک به‌صورت هوشمند به‌دست آورد.
- برای امتیازدهی به جملات، می‌توان از ترکیب میزان شباهت با عبارت پرس‌وجو و سایر پارامترها مانند پارامترهای آماری استفاده کرد.

۷. مراجع

[1] Esmaeilpour, R., "The Review of automatic summarization tools Documents in Various Language for using in Persian Texts Summarization," Proposal in Supreme Council of Information and Communication Technology, Computer Engineering College, Iran University of Science and Technolog. (in Persian)

[2] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," Expert Systems with Applications, vol. 36, pp. 7764-7772, 2009.

[3] H. Luhn, "The Automatic Creation of Literature Abstracts. Advances in Automatic Text Summarization," ed: MIT Press, Cambridge, Massachusetts, USA, 1956.

[4] D. Das and A. F. Martins, "A survey on automatic text summarization," Literature Survey for the Language and Statistics II course at CMU, vol. 4, pp. 192-195, 2007.

[5] I. Mani and E. Bloedorn" ,Machine learning of generic and user-focused summarization," in AAAI/IAAI, 1998, pp. 821-826.

[6] W. T. Chuang and J. Yang, "Extracting sentence segments for text summarization: a machine learning approach," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 152-159.

[7] A. Mohamed and S. Rajasekaran, "Query-Based Summarization Based on Document Graphs," in IEEE International Symposium on Signal Processing and Information Technology, Vancouver, Canada, 2006, pp. 408-410.

[8] J. Jagadeesh, et al., "Capturing Sentence Prior for Query-Based Multi-Document Summarization," 2007.

[9] G. A. Miller, "WordNet: a lexical database for english " presented at the Comm. ACM, 1995.

[10] M. G. Ahsaee, et al., "Semantic similarity assessment of words using weighted WordNet," International Journal of Machine Learning and Cybernetics, vol. 5, pp. 479-490, 2014.

[11] Y. Matsuo, et al., "Graph-based word clustering using a web search engine ",in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 542-550.

[12] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in Proceedings of the 15th international conference on World Wide Web, 2006, pp. 377-386.

[13] H.-H. Chen, et al., "Novel association measures using web search with double checking," in Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006, pp. 1009-1016.

[14] R. L. C. a. P. M. B. Vit´anyi, "The Google Similarity Distance," presented at the TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2007.

[15] D. Bollegala, et al., "Measuring semantic similarity between words using web search engines," www, vol. 7, pp. 757-766, 2007.

[16] Y. Liu and Q. Liu, " Sentence Similarity Computation Based on Feature Set," in 13th International Conference on Computer Supported Cooperative Work in Design, 2009.

[17] T. K. Landauer, et al., "Introduction to Latent Semantic Analysis," in Discourse, 1998, pp. 259-284.

[18] J. Allan, et al., "Retrieval and novelty detection at the sentence level," in SIGIR'03, 2003, pp. 314-321.

[19] Y. Li, et al., "Sentence Similarity Based on Semantic Nets and Corpus Statistics," presented at the TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2006.

[20] S. Jian-fang, et al., "Sentence Similarity Measure Based on Events and ContentWords," 2010.

[21] Y. Liu and Y. Liang, "A Sentence Semantic Similarity Calculating Method Based on Segmented Semantic-Comparision," Journal of Theoretical and Applied Information Technology, vol. 48, pp. 231-235, 2013.

[22] J. Xu and Q. Lu, "PolyUCOMP-CORE TYPED: Computing Semantic Textual Similarity using Overlapped Senses," Atlanta, Georgia, USA, p. 90, 2013.

[23] M.Hassel and N.Mazdak, "FarsiSum : A Persian Text Summarizer," presented at the 20th International Conference on Computational Linguistic, 2004.

[24] Karimi, Z. and Shamsfard, M., "the automatic summarization system of Persian texts," 12th international conference of computer society of Iran , Tehran, 1385. (in Persian)

[25] Akbarzadeh, S. and Teshnehlab, "Text Summarization based on Extraction using human cognitive Approach," 18th Iranian Conference on Electrical Engineering, Isfahan University of Technology, 1389. (in Persian)

[26] M. Shamsfard, et al., "Persian Document Summarization by Parsumist," World Applied Sciences Journal, vol. 7, pp. 199-205, 2009.

[27] H. Shakeri, et al., "A New Graph-Based Algorithm for Persian Text Summarization," in Computer Science and Convergence, ed: Springer, 2012, pp. 21-30.

[28] M. Shamsfard, et al., "Semi automatic development of farsnet; the persian wordnet," in Proceedings of 5th Global WordNet Conference, Mumbai, India, 2010.

[29] Z.Sepehrian, "Persian text Summarization based on Query", Malek Ashtar University of Tehran , Computer and Information Technology Department, Tehran, 1392. (in Persian)

[30] M. S. Rasooli, et al., "A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank," in 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics, 2011, pp. 227-231.

[31] S. Tasharofi, et al., "Evaluation of statistical part of speech tagging of Persian text," in Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on, 2007, pp. 1-4.

[32] R. NAVIGLI, "Word Sense Disambiguation: A Survey," ACM Computing Surveys, vol. 41, February 2009.

[33] Z. Karimi and M. Shamsfard" ,Summarization of Persian texts," in Proceedings of 11th International CSI computer Conference, Tehran, Iran, 2006.

پیوست

مثالی از تبدیل عبارت پرس‌وجو به جمله

همان‌طور که در بخش ۰ گفته شد در این مقاله عبارت‌های پرس‌وجو بررسی شده به عبارت‌های تیتروار محدود می‌شود که در آن اولین کلمه عبارت، بار معنایی فعل را دارد. مثال‌های زیر چند نمونه از این عبارت‌ها را نمایش می‌دهد:

- "سفر رئیس‌جمهور به نیویورک"
- "دعوت مجلس از وزیر علوم"
- "برگزاری انتخابات"

همان‌طور که گفته شد، اولین کلمه از عبارت‌های پرس‌وجویی که نظایر آن در مثال‌های فوق دیده می‌شود، بار معنایی فعل را دارا می‌باشند. به‌همین دلیل، جهت ساخت جمله از روی این عبارت‌ها،

کلمه ابتدایی به انتهای عبارت منتقل می‌شود. بنابراین مثال‌های فوق به عبارات زیر تبدیل می‌شوند:

- "رئیس‌جمهور به نیویورک سفر"
- "مجلس از وزیر علوم دعوت"
- "انتخابات برگزار"

با توجه به اینکه در بخش شناسایی روابط میان کلمات، از ریشه آنها استفاده می‌شود؛ برای ساخت جمله از روی عبارت‌های پرس‌وجوی مورد بررسی، فقط کافی است که کلمات منتقل‌شده، به مصدر فعلی مناسب خود تبدیل شوند. در این میان، نکته مهم و قابل‌توجه این است که شکل‌های مصدری متفاوتی برای بعضی از این کلمات متصور می‌باشد؛ بنابراین، باید به نحوی به صحیح‌ترین شکل دست یافت.

در این راستا و برای رسیدن به صحیح‌ترین شکل، نیاز به حضور لیستی از حالت‌های مصدری ممکن برای این کلمات می‌باشد. به همین منظور، تمام حالت‌های ممکن در فارس‌نت که با کلمه موردنظر آغاز شده و دارای مقوله نحوی فعل باشند، بررسی می‌شوند. جدول ۶ حالت‌های ممکن افعال را برای مثال‌های فوق نشان می‌دهد.

سیستم پیشنهادی حالت‌های ممکن به‌دست‌آمده را در فارس‌نت بررسی کرده و synsetی که مصدر مورد نظر به آن تعلق دارد، تعیین می‌کند و در صورت متفاوت بودن این synsetها، به انتخاب مصدر صحیح می‌پردازد. به‌عنوان مثال، همان‌طور که جدول ۶ نشان می‌دهد، با توجه به فرهنگ ظرفیت افعال، برای جمله اول دو حالت مصدری "سفر کردن" و "سفر رفتن" متصور است که هر دوی این افعال به یک synset تعلق دارند. بنابراین، نیازی به انتخاب بین این دو حالت نیست و با توجه به الگوریتم پیشنهادی، در نظر گرفتن هر یک از حالت‌ها مطلوب می‌باشد. جدول ۶ نتایج به‌دست آمده از فارس‌نت را برای این افعال نمایش می‌دهد.

جدول ۶. مثالی از حالت‌های مصدری متفاوت کلمات

جمله	حال‌های ممکن مصدری
رئیس‌جمهور به نیویورک	سفر کردن- سفر رفتن
مجلس از وزیر علوم	دعوت کردن
انتخابات	برگزار شدن– برگزار نمودن- برگزار کردن

جدول ۷. synset متناظر حالت‌های مصدری

جمله	حالت‌های ممکن مصدری	شناسه synset متناظر
رئیس‌جمهور به نیویورک	سفر کردن	۹۰۸۹
	سفر رفتن	۹۰۸۹
مجلس از وزیر علوم	دعوت کردن	۸۴۹۳
انتخابات	برگزار شدن	۸۲۵۶
	برگزار نمودن	۱۴۵۰۲
	برگزار کردن	۱۴۵۰۲-۷۹۳۱

برای تشخیص مصدر مناسب در حالتی که synsetهای به‌دست‌آمده متفاوت می‌باشند (مانند جمله سوم جدول ۶)، از "فرهنگ ظرفیت افعال فارسی" استفاده می‌شود. بنابراین، با استفاده از این فرهنگ، تعداد اجزای مورد نیاز (ظرفیت‌های اجباری) برای هر یک از افعال به‌دست می‌آید و براساس آن، تصمیم‌گیری صورت می‌گیرد. جدول ۸ حاکی از ظرفیت‌های اجباری صورت‌های مختلف فعل جمله سوم می‌باشد.

جدول ۸. نمونه‌ای از ظرفیت‌های اجباری افعال

فعل	ظرفیت‌های اجباری
برگزار شدن	<فا>
برگزار نمودن	<فا،مف[را+/-]>
برگزار کردن	<فا،مف[را+/-]>

بنابر جدول فوق، فعل "برگزار شدن" نیاز به فاعل اما دو فعل "برگزار نمودن" و "برگزار کردن" علاوه بر فاعل نیاز به مفعول نیز دارند. از طرفی فاعل و مفعول هر دو دارای مقوله‌ی نحوی اسم می‌باشند؛ به این ترتیب فعل "برگزار شدن" نیاز به یک اسم و دو فعل دیگر نیاز به دو اسم در جمله‌ای که در آن قرار می‌گیرند دارند.

همان‌طور که در بخش ۰ گفته شد، در فاز پیش‌پردازش، جملات متن برچسب‌گذاری می‌شوند؛ بنابراین، براساس این برچسب‌ها و اطلاعات به‌دست‌آمده، نظیر آنچه در جدول ۸ مشاهده می‌شود، می‌توان به بهترین انتخاب رسید. در واقع، از آنجایی که عبارت پرس‌وجوی موردنظر دارای یک کلمه با مقوله‌ی نحوی اسم می‌باشد (انتخابات)، فعل "برگزار شدن" برای این عبارت انتخاب می‌شود و عبارت مورد نظر به جمله زیر تبدیل می‌شود:

"انتخابات برگزار شدن"

An Approach Based on Semantic Similarity in Persian Query-Based Summarization

Z. Sepehrian[1*], S. S. Sadidpour[2], H. Shirazi[2]

1- Master Student, Malek Ashtar University of Technology
2- PhD Student, Malek Ashtar University of Technology
3- Associated Professor, Malek Ashtar University of Technology

(Reccive: 2014/04/28, Accept: 2014/08/23)

Abstract

Automatic text summarization systems are one type of management systems of huge information. This paper discusses one type of Persian text summarization based on a query named "an extractive text summarization" which is very useful for leaders to review information about special topics.

The most important phase in this type of summarization is calculation of the similarity between the query phrase and components of the original text. For this purpose, after preprocessing the phase, converting the query to a sentence, and clarifying the word sense, it is possible to calculate the similarity between the query phrase and sentences using Farsnet. Then, those sentences that are the most similar to those in the query are selected to be used in the summary. The results of the proposed method show that this method results in quite acceptable success. Since Persian is very young in processing the original language, this paper and all alike can be a great help to its result improvement.

Keywords:

Query-Based Suummarization, Query, Semantic Similarity, Word Sense Disambiguation, Farsnet

*** Corresponding Author Email:** sadidpur@aut.ac.ir