خلاصه‌سازی چکیده‌ای نظرات بر پایه ساختار داده گراف

پروانه خسروی‌زاده
استادیار گروه زبانشناسی رایانشی
دانشگاه صنعتی شریف
khosravizadeh@sharif.ir

سارا بازرگان
دانشجوی کارشناسی ارشد زبانشناسی رایانشی
دانشگاه صنعتی شریف
bazargan_sara@mehr.sharif.ir
نشانی و تلفن:

چکیده

با رشد سریع تکنولوژی‌های تبادل اطلاعات، حجم زیادی داده در دنیای وب تولید و جمع‌آوری شده و دائماً حجم این داده‌های الکترونیکی نیز درحال افزایش است. یکی از این منابع داده، متون حاوی نظرات کاربران راجع به اخبار، محصول یا موضوعی خاص است. در این پژوهش یک روش خلاصه‌سازی چکیده‌ای بر اساس ساختار داده گراف برای فشرده‌سازی نظرات ارائه می‌گردد. در این پژوهش سعی بر این است که موضوعات مهم داخل متن با استفاده از افزونگی موجود در این‌گونه متون و ویژگی‌های گراف کشف شوند، جملات با هم ترکیب شوند، اطلاعات تکراری حذف شوند و در نهایت نیز جملات خلاصه، متناسب با اندازه خروجی مورد نیاز کاربر، استخراج شوند. این روش، بدون نظارت است و کمتر از اطلاعات وابسته به زبان استفاده می‌کند.

کلیدواژه‌ها: خلاصه‌سازی نظرات، چکیده‌ای، افزونگی، گراف

مقدمه

با رشد سریع تکنولوژی‌های تبادل اطلاعات، حجم زیادی از داده‌های الکترونیکی در دنیای وب و کتابخانه‌های دیجیتال تولید و جمع‌آوری شده است و سالانه ۳۰ درصد نیز به حجم این داده‌های الکترونیکی افزوده می‌شود. این انفجار داده‌های الکترونیکی، استخراج اطلاعات را برای کاربران بسی مشکل ساخته است. در این شرایط حجم انبوه منابع اطلاعات از یک سو و محدودیت زمان از سوی دیگر خلاصه‌سازی را نیاز اساسی کاربرانی ساخته که برای پی بردن به موضوعی در حجم بالای داده موجود در وب جستجو می‌کنند.

تحقیق بر روی خلاصه‌سازی خودکار متن از اواسط دهه ۱۹۵۰ شروع شده و یکی از چالش‌های قدیمی در متن‌کاوی است که نیازمند توجه محققین در زمینه‌های هوش محاسباتی، فرایندهای یادگیری ماشین و زبان طبیعی بوده و روش-های مختلفی نظیر شبکه‌های عصبی، درخت تصمیم‌گیری، نمودار معنایی، مدل‌های رگرسیون، منطق فازی، هوش جمعی و ... را درگیر می‌کند (مهدی‌پور و همکاران، ۱۳۹۲). خلاصه‌سازی خودکار متن به عنوان هسته مرکزی طیف گسترده‌ای

از ابزارهای پردازشگر متن مانند خلاصه‌سازهای ماشینی، سیستم‌های تصمیم‌یار، سیستم‌های پاسخگو، موتورهای جستجو و غیره از سال‌ها پیش مطرح شده و همواره به عنوان یک موضوع مهم مورد بررسی و تحقیق قرار گرفته است (پورمعصومی، ۱۳۹۰).

یکی از دلایل خلاصه‌سازی نظرات، از یک سو استفاده گسترده از وسایل الکترونیکی کوچک همچون گوشی تلفن همراه است که می‌طلبد جملات فشرده شوند تا در چنین نمایشگرهایی نمایش یابند. از سوی دیگر اهمیت صرفه جویی در زمان، نیاز به راهکارهای تسریع کننده دستیابی به اطلاعات را بارزتر نموده است. امروزه با افزایش فعالیت‌های آنلاین از جمله خرید و گردشگری، یافتن محصول یا محلی مناسب مثلاً برای صرف غذا، و در دسترس بودن وسایل الکترونیکی کوچک، نیاز به چنین خلاصه‌هایی برای تصمیم‌گیری صحیح را به یک ضرورت تبدیل کرده‌است. به عنوان مثال به هنگام خرید محصول خاصی می‌توان با گردآوری اطلاعات خلاصه از مشخصات و ویژگی‌های آن توسط تولیدکننده‌های مختلف، مناسب‌ترین گزینه را انتخاب کرد.

مروری بر کارهای گذشته:

آغاز فعالیت سیستم‌های خلاصه‌سازی متن مربوط به سال ۱۹۵۰ می‌شود. به دلیل کمبود کامپیوترهای قدرتمند و مشکلات موجود برای پردازش زبان‌های طبیعی، کارهای اولیه بر روی مطالعه ظواهر متن مانند موقعیت جمله و عبارات اشاره متمرکز شده بود. سال‌های ۱۹۷۰ تا ۱۹۸۰ هوش مصنوعی به کار آمد که ایده آن استخراج نمایش‌های دانش مانند فریم‌ها یا الگوها برای شناسایی موجودیت‌های مفهومی از متن و استخراج روابط بین موجودیت‌ها با مکانیزم‌های استنتاج بود. از اوایل ۱۹۹۰ تا به حال نیز روش‌های بازیابی اطلاعات به کار گرفته شده است (پورمعصومی، ۱۳۹۰). به طور ویژه از اوایل قرن بیستم به توسعه سیستم‌های خلاصه‌سازی خودکار توجه شد که رقابت‌های سالانه DUC[1] و TAC[2] نمونه‌ای از این توجه و علاقه می‌باشند (پورغلامعلی، ۱۳۹۰).

بررسی‌ها نشان داد که در زبان فارسی تا کنون سامانه‌ای برای خلاصه‌سازی نظرات طراحی و پیاده‌سازی نشده است؛ ولیکن در اینجا تعدادی از پژوهش‌های داخلی که از ساختار داده گراف و یا روش چکیده‌ای جهت خلاصه‌سازی متون استفاده کرده‌اند معرفی می‌گردند. در سال ۱۳۸۱ از روشی چکیده‌ای برای خلاصه‌سازی متون چندسندی استفاده گردید (به نقل از ریاحی و همکاران، ۱۳۹۱). سپس در سال ۱۳۸۵ (کریمی و شمس‌فرد) یک روش خلاصه‌سازی تک سندی پیشنهاد شد که بر مبنای گزینش جمله‌ها کار می‌کند. ایده بکار رفته در گزینش جمله‌ها در این خلاصه‌ساز، ترکیبی از دو روش زنجیره لغوی و نظریه گراف است. بهره‌پور و همکاران در سال ۱۳۸۷ از یک روش ترکیبی مبتنی بر گراف، TF-IDF و الگوریتم ژنتیک استفاده کردند که در این روش پس از امتیازدهی، جملات خلاصه با استفاده از الگوریتم ژنتیک انتخاب می‌شدند. پورغلامعلی و همکاران (۱۳۹۱) یک خلاصه ساز چندسنده بر روی زبان انگلیسی

[1] - Document Understanding Conference
[2] - Text Analysis Conference

طراحی کرده‌اند که ترکیبی از روش‌های گزینشی و چکیده‌ای می‌باشد. این روش یک فاز فشرده‌سازی بر مبنای شباهت با سایر جملات بر اساس نقش‌های معنایی اعمال می‌نماید سپس جملات به دسته‌هایی تقسیم شده تا جملات مشابه ادغام و یا حذف شوند. مهدی‌پور و همکاران (۱۳۹۲) با استفاده از ترکیب روش‌های مبتنی بر گراف و TF-IDF جملات را وزن‌دهی کرده و با استفاده از الگوریتم ترکیبی SA-GA که ترکیبی از الگوریتم ژنتیک و الگوریتم شبیه‌سازی حرارت است، خلاصه را تولید کرده‌اند.

در پژوهش‌های خارجی نیز، روش‌های زیادی برای خلاصه‌سازی متون ارائه شده است که به دلیل کثرت کارهای انجام شده تنها به تعدادی از آن‌ها که مرتبط با این پژوهش می‌باشند اشاره می‌کنیم. Ganesan و همکاران (۲۰۱۰) یک خلاصه‌ساز چکیده‌ای برمبنای گراف برای خلاصه‌سازی نظرات معرفی کردند. آن‌ها سپس در سال ۲۰۱۲ روشی بدون نظارت برای خلاصه‌سازی نظرات به کار بردند. نویسندگان این مقاله از اطلاعات متقابل برای پوشش مطالب عمده متن و از مدل زبانی n-gram مایکروسافت برای حفظ خوش‌ساختی خلاصه استفاده کرده‌اند. Fillippova (۲۰۱۰) نیز از گراف کلمات برای خلاصه‌سازی استفاده نموده است.

در پژوهش حاضر نیز از گراف کلمات بر پایه روش Ganesan و همکارانش (۲۰۱۰) و Filippova (۲۰۱۰) برای جملات تفکیک شده، استفاده کردیم. علاوه بر این، به هنگام افزودن کلمات جدید به گراف، از یک بانک اطلاعاتی کلمات مترادف نیز استفاده شد. از گراف عموماً برای خلاصه‌های استخراجی استفاده می‌شود ولی در این روش‌ها، گراف‌ها اکثراً بدون جهت هستند و معمولاً جمله‌ها به عنوان نودهای گراف و تشابه آن‌ها به عنوان یال تعریف می‌شوند. ولی ساختار داده گراف مورد استفاده در این پژوهش متفاوت است چرا که هر نود آن نماینده یک کلمه است و یال‌های جهتدار آن نشان‌دهنده ساختار جملات هستند. به علاوه نودها حاوی اطلاعات موقعیتی نیز می‌باشند. اطلاعات موقعیتی به هنگام انتخاب مسیرهای کاندید حائز اهمیت می‌شوند.

مفاهیم نظری:

۱-خلاصه‌سازی خودکار:

خلاصه‌سازی خودکار سند، تولید یک نسخه مختصرتر از سند اصلی توسط یک برنامه رایانه‌ای است؛ به نحوی که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود (پورمعصومی و همکاران، ۱۳۹۳). ورودی فرایند خلاصه‌سازی می‌تواند اطلاعات چندرسانه‌ای، صوت، ویدئو، اطلاعات آنلاین، متن و یا ابرمتن باشد. فرایند خلاصه‌سازی خودکار متن که یکی از شاخه‌های پردازش زبان طبیعی محسوب می‌شود، روی ورودی‌های متنی متمرکز است (رمضانی، ۱۳۹۱).

خلاصه‌سازی از دیدگاه‌های مختلف انواع گوناگونی خواهد داشت. روش‌های خلاصه‌سازی را می‌توان به دو دسته با نظارت[3] و بدون نظارت[4] تقسیم نمود. در روش‌های با نظارت به منظور رسیدن به دقت کافی، نیاز به استفاده از دانش

[3] -Supervised
[4] - Unsupervised

پیشین و مجموعه بزرگی از خلاصه‌های تولید شده توسط انسان می‌باشد. این دسته از خلاصه‌ها، با تغییر نوع داده‌ها و ویژگی‌های آن‌ها، نیاز به تولید مجدد داده‌های آموزشی دارند یا نیاز به میزان قابل‌ملاحظه‌ای کار دستی تا الگوهایی تعریف شوند که بتوان با استفاده از تکنیک‌های استخراج اطلاعات آن‌ها پرکرد و یا اینکه به حوزه خاصی وابسته‌اند. ولی روش‌های بدون نظارت نیازی به خلاصه‌های انسانی برای آموزش ندارند (). تقسیم‌بندی مهم دیگر به روند و نحوه خلاصه‌سازی مربوط می‌شود. در این تقسیم‌بندی خلاصه‌ها می‌توانند گزینشی و یا چکیده‌ای باشند. در خلاصه‌سازی گزینشی گزیده‌ای از مجموع قطعات متن اولیه بدون تغییر به عنوان خلاصه برگردانده می‌شود. اغلب جمله به عنوان واحد گزینش انتخاب می‌گردد (پورغلامعلی، ۱۳۹۰). انتخاب جملات خلاصه با توجه به ویژگی‌هایی از جمله صورت می‌گیرد که در گذشته میزان تأثیر این ویژگی‌ها یکسان در نظر گرفته می‌شده است و یا با استفاده از تکنیک‌های سعی و خطا ضرایب متغیری به هر یک از این پارامترها اختصاص داده می‌شد (ریاحی و همکاران، ۱۳۹۱). بنابراین خروجی اغلب لیستی از جملات متن یا متون اولیه است که بر مبنای اهمیت‌شان مرتب شده‌اند. در خلاصه‌سازی چکیده‌ای، جملات متن اولیه تغییر کرده و در واقع برگرفته‌ای از متن اولیه را خواهیم داشت (پورغلامعلی، ۱۳۹۰). مشکل روش-های خلاصه‌سازی گزینشی این است که در بسیاری موارد ما شاهد یک میزان اشتراک اطلاعاتی بین جملات هستیم. در این موارد در صورتی که برخی جملات حذف شوند، میزانی از اطلاعات را از دست داده‌ایم، و در صورتی که همه جملات آورده شوند، دچار افزونگی خواهیم شد و از هدف خلاصه‌سازی دور خواهیم شد (پورغلامعلی و همکاران، ۱۳۹۱) و این مشکل بیشتر زمانی بروز می‌کند که بخواهیم خلاصه‌های تولید شده را روی صفحه نمایش‌های کوچک مثل گوشی‌های تلفن همراه یا تبلت و PDA مشاهده کنیم (Ganesan و همکاران، ۲۰۱۰). از سوی دیگر فرایند خلاصه‌سازی چکیده‌ای بسیار پیچیده و دشوار است، چرا که نیازمند نمایشی مفهومی از متن می‌باشد و رسیدن به این نمایش بسیار مشکل خواهد بود. علاوه بر این برای ساخت جمله‌ای جدید نیاز به اطلاعات زبانشناسی بسیار قوی می-باشد. به همین سبب در این زمینه کارهای کمی صورت گرفته است.

همانطور که گفته شد خلاصه‌سازی بخصوص از نوع چکیده‌ای، یکی از پیچیده‌ترین کاربردهای پردازش متن است که معمولاً با چند سطح از پردازش متن درگیر است. پردازش لغوی یکی از مؤلفه‌های جدانشدنی خلاصه‌سازی است و در عملیات پیش‌پردازشی به طور گسترده مورد استفاده قرار می‌گیرد که از آن جمله، می‌توان به تعیین مرز لغات و واژه-ها اشاره نمود. بکارگیری سایر انواع پردازش، بستگی مستقیم به نوع روش خلاصه‌سازی دارد (مشکی و آنالویی، ۱۳۸۸).

۲- پیش پردازش:

پردازش‌هایی که بر روی متن‌های زبان طبیعی صورت می‌گیرد اغلب نیازمند عملیات پیش‌پردازش است؛ به طوری که دقت این پیش‌پردازش‌ها تأثیر بسزایی بر نتایج اعمال الگوریتم‌ها در مراحل بعدی دارد. هر قدر دقت این پردازش‌ها بیشتر باشد، الگوریتم‌ها به نتایج واقعی خود نزدیک‌تر خواهندشد (پورمعصومی و همکاران، ۱۳۹۳). اعمال پیش

پردازشی شامل تعیین مرز واژه‌ها و جمله‌ها، یکسان‌سازی متن، حذف واژه‌های عمومی (کلمات پرتکرار بی‌ارزش) و ریشه‌یابی است. (مشکی و آنالویی، ۱۳۸۸).

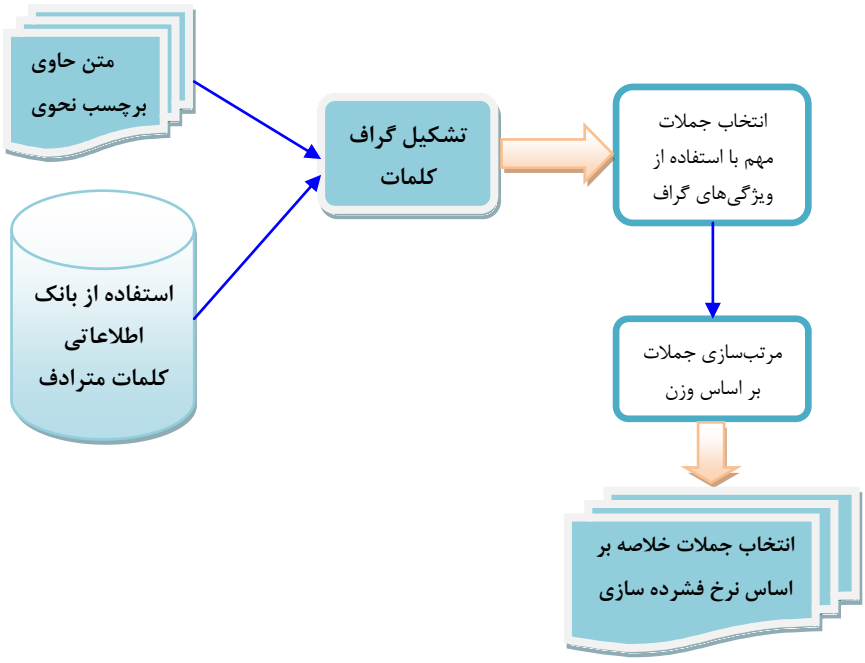
۳- برچسب زنی نحوی:

برچسب زنی نحوی عمل ابهام زدایی از کلمات با توجه به زمینه را انجام می‌دهد. از مهمترین برچسب‌های نحوی می‌توان به اسم، فعل، صفت، قید، حرف اضافه و غیره اشاره کرد.

ما در این پژوهش عملیات پیش‌پردازشی و برچسب‌گذاری نحوی را به صورت دستی انجام دادیم تا آماده عملیات پردازش شوند. درواقع، هیچ یک از واژه‌ها حتی ایست واژه‌ها نیز حذف نشدند و کلمات به همان صورتی که در متن آمده بودند در گراف وارد شدند.

روش پیشنهادی:

در این پژوهش به بیان روشی برای خلاصه‌سازی نظرات خواهیم پرداخت. روش مذکور بر اساس ساختار داده گراف به شرح زیر می‌باشد که پس از انجام پیش‌پردازش‌های لازم اجرا می‌شود:



معماری سیستم خلاصه‌سازی پیشنهادی (شکل۱)

ما یک روش ساده و دقیق با استفاده از گراف برای تولید خلاصه‌های کوتاه استفاده می‌کنیم که با حداقل اطلاعات مربوط به برچسب مقوله نحوی (POS) قادر به انجام خلاصه‌سازی باشد. بنابراین جملات ورودی باید حاوی برچسب مقوله نحوی کلمات باشند.

گراف کلمات گرافی است جهت‌دار که اگر رشته «abcd» یک جمله از زبان باشد، a و b و c و d به ترتیب نودهای این گراف را تشکیل می‌دهند و یک یال از a به b نشان‌دهنده رابطه مجاورت آن‌ها در جمله است. سپس کلمات جمله دوم بایستی به این گراف اضافه شوند. اگر کلمه جدید یا مترادف آن از قبل در گراف موجود نباشد، باید یک نود جدید درج گردد، ولی در صورتی که کلمه یا مترادف آن در جملات قبلی وجود داشته است، به نود موجود اختصاص خواهد یافت. در صورتی کلمه‌ای از یک جمله جدید به یک نود موجود اختصاص می‌یابد که POS آن‌ها یکسان باشد و همچنین کلمه‌ای از همان جمله قبلاً به این نود اختصاص داده نشده باشد. نیز با استفاده از اطلاعات POS، احتمال اختصاص فعل و اسمی که صورتی مشترک دارند به جای یکدیگر منتفی خواهد شد. بنابراین مسیرهای غیردستوری و بدساخت تولید نخواهند شد (برای مشاهده جزئیات بیشتر در خصوص گراف حاصل به شکل ۲ مراجعه کنید).

درصورتی که بیش از یک نود برای اختصاص دادن یک کلمه جدید وجود داشته باشد، بایستی به نودهای قبلی و بعدی در گراف و کلمات قبلی و بعدی در جمله توجه شود و نودی از گراف انتخاب شود که همپوشانی بیشتری با جمله دارد و یا دارای فرکانس بیشتری در گراف تا به حال بوده‌است (یعنی کلمات بیشتری به آن نود اختصاص داده شده‌اند). کلمات عمومی نیز زمانی به نودهای موجود اختصاص می‌یابند که بافت جمله حاوی آن دارای اشتراکی در همسایگی آن نود باشد، در غیر اینصورت بایستی نود جدیدی به ازای آن اضافه شود (Filippova، ۲۰۱۰).

نکته قابل توجهی که وجود دارد این است که افزونگی باید به اندازه کافی باشد تا نه تنها کلمات مهم مشخص گردد، بلکه ارتباطی نیز بین کلمات ظاهر گردد. در ادامه ویژگی‌های مهم این گراف را برای یافتن بهترین مسیرهایی که منجر به تولید خلاصه‌ای از متن ورودی می‌شوند را معرفی می‌کنیم:

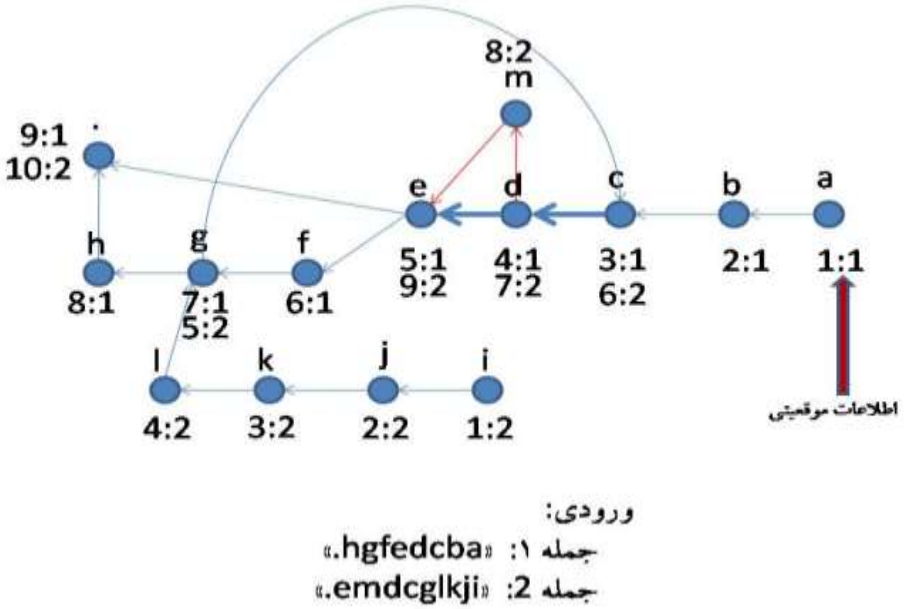
۱) عبارت‌هایی که دارای تکرار بالایی هستند به صورت زیر گراف‌هایی قابل تفکیک هستند.

۲) ساختار ایجاد شده برای جملات مسیرهایی را تولید می‌کند که جملات جدیدی را ساخته‌اند و عبارت‌های موجودی را که دارای اندک تفاوتی با هم بوده‌اند را پررنگ‌تر نشان می‌دهند.

۳) نودهایی را که دارای انشعاب هستند به سادگی نشان می‌دهد و با استفاده از این نودها می‌توان جمله جدیدی را به خلاصه اضافه کرد. بدین صورت که اگر یک یا چند نود به دنبال هم در چندین جمله تکرار شده باشند و سپس نودی حاوی انشعاب بعد از آن‌ها آمده باشد، چنین ساختاری می‌تواند به سادگی استخراج شده و کاندیدای مناسبی برای تولید خلاصه باشد. می‌توان برای نودهای حاوی انشعاب شرطی قائل شد مبنی بر اینکه نودهایی که حاوی POS خاصی باشند می‌توانند به عنوان کاندید انتخاب شوند.

با توجه به تعریف خلاصه‌سازی چکیده‌ای به نظر می‌رسد این روش بسیار شبیه به روش‌های استخراجی باشد زیرا خلاصه تولید شده حاوی عباراتی است که در متن ورودی وجود دارند با این تفاوت که در روش‌های پیشین با استخراج جملات دست به خلاصه‌سازی می‌زدند ولی در این روش ما با استخراج کلمات این کار را انجام می‌دهیم و بنابراین جنبه چکیده‌ای به خلاصه ما می‌دهد چرا که ما از ترکیب واحدهای استخراج شده و حذف اطلاعات اضافی از جملات، فشرده‌سازی انجام داده‌ایم. بنابراین جملات موجود در خلاصه عین جملات موجود در متن اصلی نیستند. از این جهت می‌توان این روش را یک روش نسبتاً چکیده‌ای محسوب کرد.

برای انتخاب جملات کاندید پیش‌شرط‌هایی نیز لازم است از جمله اینکه مسیر کاندید باید دارای شروع و پایان طبیعی در زبان باشد. یعنی نودهای شروع و پایان زیرگراف انتخابی باید شرایط آغاز جمله و پایان طبیعی آن را داشته باشند تا بتوانند به عنوان جمله کاندید انتخاب شوند. این شروط نیز با استفاده از اطلاعات موقعیتی نودها و همچنین علائم نقطه‌گذاری و حروف ربط تأمین می‌گردند. همچنین مسیر کاندید باید دارای توالی مجازی از POSها در زبان طبیعی باشد. به هنگام انتخاب مسیرهایی که در اثر افزونگی کاندید شده‌اند نیز عبارت‌هایی را انتخاب می‌کنیم که با اندکی تفاوت در جملات متعددی تکرار شده باشند. آستانه‌ای برای این تفاوت بایستی تعریف گردد که معرف شکاف بین جملات است. این مقدار آستانه کنترل می‌کند که عبارت‌های بدساخت تولید نشوند و یا اطلاعات مفید از دست نروند (Ganesan و همکاران، ۲۰۱۰).



شکل (۲)

امتیاز دهی

چه ویژگی‌هایی نشان‌دهنده یک خلاصه خوب هستند؟ مسیرهایی که نه خیلی طولانی و نه خیلی کوتاه باشند. مسیری باید انتخاب گردد که نودهای آن مفاهیم اصلی را شامل می‌شوند ولی از یک نود نباید چندین بار عبور کند و ضمناً مسیر باید حاوی POSهای مجاز در زبان باشند. بدین منظور نودهای متصل به همی را که دارای فرکانس بالایی هستند و کوتاهترین مسیر بین دو نود شروع و پایان مجاز را شامل می‌شوند، انتخاب می‌کنیم. در نهایت مسیر بدست آمده شامل رشته‌ای از کلماتی است که در بسیاری از جملات ورودی وجود داشته است. هر چه طول مسیری که در نتیجه تکرار متوالی کاندید شده است بیشتر باشد، بهتر است زیرا اطلاعات موجود را بیشتر پوشش می‌دهد. بنابراین یکی از معیارهای امتیازدهی به مسیرهای کاندید می‌تواند اندازه جمله باشد. بدین منظور یک حد آستانه‌ای برای حداقل طول جمله انتخابی می‌توان در نظر گرفت. همچنین به منظور آنکه خلاصه نهایی حداکثر پوشش را از موضوعات اصلی متن ورودی نمایش دهد، مسیرهایی که حاوی فعل نباشند را نیز انتخاب نمی‌کنیم.

امتیازدهی در جملاتی که از خاصیت انشعابی برخی نودها جهت ترکیب عبارت‌ها استفاده می‌کنند، پس از ترکیب و بدست آوردن جمله کاندید محاسبه می‌گردد. بدین صورت که امتیاز مسیر به ازای هر انشعاب یکبار محاسبه می‌گردد و سپس امتیاز مسیر ترکیبی از میانگین امتیازات محاسبه شده بدست می‌آید.

تولید خلاصه

پس از مرحله امتیازدهی نوبت به تولید خلاصه می‌رسد که شامل دو گام است: ابتدا تمامی مسیرهای کاندید به ترتیب نزولی مرتب می‌شوند. جملات کاندید ممکن است همچنان شامل جملاتی باشند که شباهت زیادی با هم دارند. در گام دوم باید جملات دارای محتوای تکراری حذف شوند که با استفاده از معیار شباهت (مثل کسینوسی یا Jaccard و ...) شباهت جملات اندازه‌گیری می‌شود. سپس از جملاتی که کمترین فاصله را با هم دارند، یکی که دارای بیشترین امتیاز بوده، انتخاب خواهد شد. نهایتاً تعداد اندکی از مسیرهای دارای بیشترین امتیاز جهت خلاصه نهایی باقی می‌مانند که در پایان نسبت به اندازه خلاصه موردنظر بایستی انتخاب گردند. برای کاهش تعداد جملات کاندید هم یک مقدار آستانه‌ای تعریف می‌کنیم که به هنگام انتخاب جملات حاصل از افزونگی، صرفاً جملاتی انتخاب شوند که بیش از آستانه تعریف شده، در جملات ورودی مورد اشتراک بوده‌اند. مثلاً اگر عبارتی در بیش از ۵ جمله تکرار شده باشد به عنوان کاندید انتخاب شود.

اغلب روش‌های موجود از اطلاعات نحوی برای تولید خلاصه‌های خوش ساخت استفاده می‌کنند. اتفاقاً نحو سرنخ-هایی راجع به آنچه دارای اهمیت است نیز فراهم می‌کند، مثلاً در کاربردهایی همچون خلاصه‌سازی متن، ممکن است فاعل و فعل جمله پایه مهمتر از عبارت حرف‌اضافه‌ای یا فعل جمله پیرو باشند؛ شاید بهتر باشد بگوییم تفکیک اطلاعات نحوی در یافتن آنچه که بار معنایی را منتقل می‌کند دارای اهمیت است. البته نحو تنها راه سنجش اهمیت کلمات یا عبارات نیست. مثلاً تکرار یک کلمه یا وقوع کلماتی که به لحاظ معنایی با هم شباهت دارند، می‌توانند معیار سنجش اهمیت باشند که در خلاصه‌سازهای اولیه از آن‌ها استفاده می‌شده‌است. هنوز هم پارسرهای نحوی، یکی از

ابزارهای ضروری در فرایند خلاصه‌سازی هستند چرا که محدودیت‌های نحوی، تنها راه ممکن جهت کنترل خوش-ساختی خروجی هستند. در این مقاله ما قصد داشتیم ضرورت استفاده از این ابزار برای کنترل خوش‌ساختی خروجی را تا حدی انکار کنیم زیرا در متون نظرات، که تعداد تکرار زیاد است، استفاده از افزونگی می‌تواند یک روش قابل اطمینان جهت تولید جملات خوش‌ساخت و مطابق با دستور زبان باشد (Filippova، ۲۰۱۰).

نتیجه‌گیری

در این پژوهش روشی بر پایه ساختار داده گراف برای خلاصه‌سازی چکیده‌ای نظرات فارسی معرفی شد. روش مورد استفاده بدون نظارت بوده و کمتر از اطلاعات وابسته به زبان استفاده می‌کند. در این روش، برای خلاصه‌سازی متن ورودی و برای سازگاری با وسایل الکترونیکی کوچکتر، از گراف و بانک اطلاعاتی کلمات مترادف برای تولید خلاصه-های چکیده‌ای استفاده کردیم (با استفاده از روش Ganesan و همکارانش، ۲۰۱۰). این روش از دانش خاصی استفاده نمی‌کند و صرفاً روش‌های ساده NLP را بکار می‌گیرد. بدین صورت که از ترتیب کلمات موجود در متن و افزونگی موجود در آن برای تولید خلاصه‌های چکیده‌ای و جامع استفاده می‌کند. در نهایت نیز جملات خلاصه، با استفاده از ویژگی‌های بدست آمده از گراف و اندازه خروجی، استخراج شدند. چون این روش از دانش زمینه‌ای خاصی استفاده نمی‌کند، بنابراین برای خلاصه‌سازی هر گونه محتوایی که حاوی افزونگی باشد قابل استفاده است.

منابع:

1- Erkan, G. & Radev, D. R. (2004). "LexRank: Graph-based lexical centrality as salience in text summarization". J Artif Intell ResJAIR, 22(1).

2- Filippova, K. (2010). **"**Multi-sentence compression: Finding shortest paths in word Graphs". Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, 322–330.

3- Ganesan, K., Zhai, Ch., & Han J. (2010). "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions". Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, 340–348.

4- Ganesan, K., Zhai, Ch., & Viegas, E. (2012). "Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions". Proceedings of the 21st international conference on World Wide Web, April 16–20, Lyon, France, 869-878.

5- Kolla, M. (2002). "Automatic text summarization using lexical chains: Algorithms and experiments" (Master thesis). University of Lethbridge, Canada.

6- Mani, I. & Bloedorn, E. (1997). "Multi-document summarization by graph search and matching". ArXiv Prepr, Cmp-Lg9712004.

7- Nenkova, A. & McKeown, K. (2012). "A survey of text summarization techniques". Chapter in Mining Text Data, 43-76.

۸- استری، ا.، کاهانی، م.، پورمعصومی، آ.، و عباسی، م. (۱۵-۱۶ شهریور ۱۳۹۱). «ارائه یک ابزار ارزیابی خودکار خلاصه‌سازهای چکیده‌ای فارسی با بهره‌گیری از شبکه واژگان». نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، دانشگاه سمنان.

۹- بازقندی، م.، تدین تبریزی، ق.، وفایی جهان، م.، و بازقندی، ع. (۱۵-۱۶ شهریور ۱۳۹۱). «خلاصه‌سازی گزینشی متون فارسی مبتنی بر خوشه‌بندی PSO». نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، دانشگاه سمنان.

۱۰- بهره‌پور، م.، مهدی‌پور، ا.، کامل، آ.، امیری، م.، طهماسبی، آ.، و اکبرزاده توتونچی، م. (۲۰-۲۱ اسفند ۱۳۸۷). «سیستم خلاصه‌ساز خودکار متن‌های فارسی». چهاردهمین کنفرانس سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر، تهران.

۱۱- پورغلامعلی، ف. (۱۳۹۰). «خلاصه‌سازی چکیده‌ای مبتنی بر مشابهت جملات» (پایان نامه کارشناسی ارشد). دانشگاه فردوسی مشهد. خراسان.

۱۲- پورغلامعلی، ف.، کاهانی، م.، و پورمعصومی، آ. (۱۵-۱۶ شهریور ۱۳۹۱). «خلاصه‌سازی چکیده‌ای مبتنی بر مشابهت جملات». نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، دانشگاه سمنان.

۱۳- پورمعصومی، آ.، کاهانی، م.، کامیار، م.، و کامیار، ح. (۱۳۸۹). «خلاصه‌سازی خودکار چندسندی مبتنی بر مفاهیم». شانزدهمین کنفرانس ملی انجمن کامپیوتر ایران. دانشگاه صنعتی شریف، تهران، ۳۳۲-۳۳۷.

۱۴- پورمعصومی، آ. (۱۳۹۰). «خلاصه‌سازی خودکار چندسندی مبتنی بر استخراج مفاهیم» (پایان نامه کارشناسی ارشد). دانشگاه فردوسی مشهد. خراسان.

۱۵- پورمعصومی، آ.، کاهانی، م.، طوسی، س.ا.، استیری، ا.، و قائمی، ه. (۱۳۹۳). «ایجاز: یک سامانه عملیاتی برای خلاصه‌سازی تک سندی متون خبری فارسی». دوفصلنامه پردازش علائم و داده‌ها، ۱(۲۱)، ۳۳-۴۸.

۱۶- حافظی، م.م .، ثامتی، ح.، منصوری، ن.، منتظری، ن.، بحرانی، م.،و موثق، ح. (۱۳۸۵). «ارائه یک مدل دستوری برای بهبود دقت سیستم‌های بازشناسی گفتار پیوسته فارسی». دومین کارگاه پژوهشی زبان فارسی و رایانه،80 - ۹۱.

۱۷- رمضانی، م. (۱۳۹۱). «خلاصه‌سازی خودکار متون فارسی مبتنی بر هستی شناسی» (پایان نامه کارشناسی ارشد). مؤسسه آموزش عالی نبی اکرم. تبریز.

۱۸- ریاحی، ن.، غزالی، ف.، و غزالی، م.ع. (۱۵-۱۶ شهریور ۱۳۹۱). «بهبود کارایی سیستم خلاصه‌سازی متون فارسی با استفاده از الگوریتم هرس در شبکه‌های عصبی». نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، دانشگاه سمنان.

۱۹- ستوده، آ.، پویان، ع.، و ستوده، ح. (۳-۴ اسفند ۱۳۹۰). «استخراج شباهت معنایی بوسیله سیستم استنتاج فازی در خلاصه-سازی متن». کنفرانس ملی فناوری اطلاعات و جهاد اقتصادی، دانشگاه سلمان فارسی کازرون، فارس، ۱۱۰۷- ۱۱۱۸.

۲۰- شورای عالی اطلاع‌رسانی. (۱۳۸۸). «بررسی مستندات ابزارهای خودکار خلاصه‌سازی زبان‌های دنیا برای به کارگیری در خلاصه‌سازی متون زبان فارسی». فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی.

۲۱- صادقی، س.س. (۱۳۹۲). «خلاصه‌ساز استخراجی تک سندی متون روایی مبتنی برعملکرد ذهن انسان» (پایان نامه کارشناسی ارشد). دانشگاه صنعتی شریف، تهران.

۲۲- کامیار، ح.، کاهانی، م.، کامیار، م.، پورمعصومی، آ.، و کیاده، ح. (۱۷-۱۹ اسفند ۱۳۸۹). «روش جدید خلاصه‌سازی استخراجی تک سندی با استفاده از نظریه مرکزیت». شانزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی شریف، تهران، ۴۷۹-۴۸۴.

۲۳- کریمی، ز.، و شمس فرد، م. (۱-۳ اسفند ۱۳۸۵). «سیستم خلاصه‌سازی خودکار متون فارسی». دوازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، دانشگاه شهید بهشتی، تهران۱۲۸۶-۱۲۹۴.

۲۴- مشکی، م.، و آنالویی، م. (۱۳۸۸). «خلاصه‌سازی چند سندی متون فارسی با استفاده از یک روش مبتنی بر خوشه‌بندی». اولین کنفرانس ملی مهندسی نرم افزار ایران، ۱۵۰-۱۵۶.

۲۵- مهدی‌پور، ا.، باقری قرقوروک، م.، و رضایی، ا. (۱۳۹۲). «سیستم خلاصه‌ساز خودکار متن فارسی با استفاده از الگوریتم ترکیبی SA-GA». هشتمین سمپوزیوم پیشرفت‌های علوم و تکنولوژی، مشهد.

۲۶- نوری‌زاده، ر.، و خسروی، ح. (۱۳۹۲). «خلاصه‌سازی متن مبتنی بر ویژگی‌های جملات با استفاده از الگوریتم رقابت استعماری گسسته». کنگره ملی مهندسی برق، کامپیوتر و فناوری اطلاعات (8thSASTech)، مشهد، ایران.

۲۷- هنرپیشه، م. (۱۳۸۶). «طراحی و پیاده‌سازی یک خلاصه‌ساز متون فارسی» (پایان نامه کارشناسی ارشد). دانشگاه صنعتی شریف، تهران.