مجله پژوهش های معاصر در علوم و تحقیقات
سال دوم، شماره ۱۰، اردیبهشت ۱۳۹۹
ISSN: 2676-5764
www.jocrisar.ir

طراحی و پیاده‌سازی سامانه خلاصه ساز خودکار و معنایی متون فارسی مبتنی بر رویکرد گراف وزن‌دار

سحر اسماعیلی شایان[۱]

[۱] دانش آموخته کارشناسی ارشد- دانشگاه الزهرا

چکیده

خلاصه سازی متون یکی از روش‌های استخراج اطلاعات مفید و مهم از حجم عظیم داده‌های متنی است که در اهدافی چون تحلیل داده‌های متنی به کار بسته می‌شود. طی سالیان متمادی، تکنیک‌های خلاصه سازی متن بسیاری توسعه داده شده اند که برخی تنها به انتخاب جملات کوتاه و آوردن آن‌ها در خلاصه بسنده کرده و برخی دیگر بدون توجه به همبستگی معنایی جملات، آن‌ها را گزینش و در خلاصه می‌آورند. تحلیل معنایی متون نیازمند روش‌های استخراج خلاصه با رویکرد معنایی است. در این پژوهش، سامانه‌ی خلاصه ساز زبان فارسی با استفاده از توسعه و به کارگیری الگوریتم TextRank گوگل و با بهره-گیری از مدل سازی سند متنی به صورت گرافی که در آن جملات به صورت گره و ارتباط جملات به صورت یال‌های گراف و میزان ارتباط معنایی میان جملات به صورت وزن هر یال مدل گردیده، توسعه داده شده است. نتایج پژوهش با بررسی ۱۱۴۶ مقاله فارسی خلاصه شده توسط این سامانه، نشان داد که سامانه توسعه داده شده با اختصاص رتبه‌ی بالاتر به جملات حامل معنای بیشتر و تهیه خلاصه نهایی از آن‌ها، عملکرد خوبی در استخراج خلاصه معنایی از متون الکترونیکی فارسی دارد.

واژه‌های کلیدی: متن کاوی، خلاصه سازی خودکار، اسناد الکترونیک فارسی، گراف وزن دار

۱. مقدمه

در دو دهه اخیر و با گسترش اینترنت، تولید محتوای الکترونیکی و بالاخص تولید اسناد متنی الکترونیکی بر بستر وب و اینترنت رشد فزاینده ای پیدا کرده است و این اسناد متنی را می‌توان به عنوان منابع عظیم اطلاعاتی مورد تحلیل قرار داد؛ با این حال برای دریافت معنای کلی اسناد متنی نیاز به در دست داشتن تمام متن نیست و می‌توان از خلاصه‌های معنایی استخراج شده از متن به جای کل سند استفاده کرد. خلاصه سازی خودکار متون، یکی از شاخه‌های علم پردازش زبان‌های طبیعی است که با هدف کاهش داده‌ی متنی به نفع مدیریت پایگاه‌های داده و نیز نگه‌داری بیش‌ترین اندازه از معنا به کار می‌رود و می‌توان آن را استخراج اطلاعات مهم و معنای مفید از حجم عظیم داده‌های متنی و از منابع رو به گسترشی چون سایت‌های اینترنتی دانست. ایده اصلی در بیشتر سامانه‌های خلاصه ساز نگه‌داری اطلاعات مفید و مهم و دور ریختن وارده‌های کم‌تر مرتبط است[2] ,[1].

از منظر ورودی به سامانه، دو دیدگاه در مورد سامانه‌های خلاصه ساز متن وجود دارد که یکی خلاصه ساز تک سندی و دیگر خلاصه ساز چند سندی است. خلاصه ساز تک سندی یک سند متنی را گرفته و یک سند خلاصه شده را به عنوان خروجی بازمی‌گرداند؛ در حالی که سامانه خلاصه ساز چند سندی، چند سند متنی ورودی را گرفته و از آن‌ها یک سند خروجی خلاصه سازی شده تهیه می‌کند. در منظر دوم تقسیم بندی سامانه‌های خلاصه ساز به خروجی سامانه توجه می‌شود و این روش نیز دو رویکرد را مد نظر قرار می‌دهد. رویکرد اول خلاصه سازی استخراجی است که در آن واحدهای اطلاعاتی از متن اصلی استخراج می شوند و رویکرد خلاصه سازی تجریدی است که در آن اطلاعات ساخته شده ای را می بینیم که ممکن است در متن اصلی نتوانیم آن ها را بیابیم و در واقع خلاصه‌های ایجاد شده حاصل ترکیب جملات با معنای نزدیک می‌باشند [۳].

فارغ از اهمیت خلاصه سازی متن در مدیریت پایگاه‌های داده متنی، بسیاری از نرم افزارها و تکنیک‌های متن کاوی نیاز به خلاصه سازی اسناد متنی دارند تا با استفاده از تعداد کمتری از جملات بتوانند نگاهی سریع به متون طولانی و یا تعداد بالای اسناد متنی موجود در مورد یک موضوع خاص داشته باشند. اهمیت خلاصه سازی متون گاهی تا جایی بالا می‌رود که می‌توان آن را به عنوان یک مساله بهینه سازی برای استخراج تعداد بهینه ای از جملات برای رساندن معنای کلی سند متنی قلمداد کرد[۳], [۴]. یکی دیگر از موارد کاربرد خلاصه‌های متنی، استفاده از آن‌ها در برچسب زنی معنایی به اسناد متنی و بهینه‌سازی فرآیند جست و جوی اسناد متنی بر اساس معنا و محتوای آن‌ها است که بیشتر در سایت‌های آنلاین و خبری کاربرد دارد و لزوم توسعه روش‌های خلاصه سازی سریع و دقیق به جهت ارائه نتایج درست و به موقع هر جستجوی کاربران را بیش از پیش آشکار می‌سازد [۵], [۶].

تولید گسترده و انتشار متون فارسی در منابعی چون صفحات وب، ایمیل‌ها، سایت‌های خبری و منابعی از این دست، همانند دیگر زبان‌های زنده دنیا اجتناب ناپذیر بوده و با توجه به حجم عظیم داده‌های متنی موجود و نیز حجم فزاینده و شتاب نمایی تولید این جنس از داده‌ها نیاز به توسعه سامانه‌های خودکار خلاصه سازی متن به جهت کاهش اندازه داده‌ها و نگه‌داری داده‌های متنی با ارزش‌تر احساس می‌گردد.

۲. رویکردهای خلاصه سازی متنی

برخی از رویکردهای خلاصه سازی متون، مستقل از زبان نگارش سند متنی هستند و برخی دیگر وابسته به زبان هستند. روند پیشرفت رویکردها و تکنیک‌های خلاصه سازی اسناد متنی به همراه نیازمندی‌ها و مزایا و معایب آن‌ها در ادامه آمده است. خلاصه سازی آموزش دیده و تجزیه و تحلیل معناییِ پنهان متن به جهت خلاصه سازی متن که توسط [7] ارائه گردید. وی عنوان کرد MCBA + GA می توانند در کار با ژانر مشخص و همچنین برای اهداف آنلاین مورد استفاده قرار گیرند. از طرفی دیگر رویکرد LSA+TRM برای زمانی که کیفیت خلاصه اولویت بالایی دارد، مفیدتر واقع می‌شوند. با استفاده از LSA , TRM می‌توان از جملاتی که به یکدیگر وابستگی معنایی دارند خلاصه تهیه کرد. همانطور که پیداست محدودیت این روش الزام وجود وابستگی معنایی میان جملات سند متنی است.

استخراج اطلاعات با استفاده از روش انتزاع مبتنی بر جمله که توسط [8] معرفی شد. این رویکرد به تولید خلاصه از متن با استفاده از مفاهیم مرتبط با معنا و نیز با تمرکز بر فاکتورهای متنی مانند همبستگی متنی و همبستگی واژگانی می‌پردازد. ایراد موجود در این رویکرد این است که تنها یکپارچگی و همبستگی علت و معلولی در نظر گرفته می‌شود؛ در حالی که علیت، همبستگی زمانی و همبستگی مکانی است که همبستگی و یکپارچگی کلی متن را شکل می‌دهد.

درک متن و خلاصه سازی با استفاده از ساخت شبکه مفهوم از سند متنی که توسط [9] معرفی گردید و در آن تولید خلاصه ای از جملات همبسته با حداقل مفاهیم از دست‌رفته، برای تمرکز بیشتر بر معانی، پوشش همه مضامین متمایز و مرتبط با مفاهیم کلیدی و با حداقل تعداد کلمات مورد توجه بود. ایراد اساسی این روش بالا بودن هزینه‌ی محاسبه‌یDCL به دلیل در نظر گرفتن تمامی ترکیب‌های ممکن از مفاهیم به کاررفته در ساخت جملات متن اصلی است.

استخراج خلاصه بر اساس اطلاعات متنی و آماری سند متنی که توسط [10] ارائه گردید و روش به کار رفته در این پژوهش، مستقل از زبان نگارش اسناد متنی و مناسب خلاصه سازی اسناد متنی به زبان‌های مختلف به صورت یک سند یکتا بوده است. این روش چندین مزیت اساسی دارد و بزرگترین مزیت این رویکرد می تواند مستقل از زبان بودن آن باشد. مزیت دیگر مرتبه‌ی زمانی و مکانی کم الگوریتم به کاررفته در این روش است به طوری‌که پردازش‌ها به پردازنده و ظرفیت حافظه بالا نیاز ندارند. مزیت دیگر این است که جملات اسناد مورد نظر برای خلاصه سازی می‌توانند فاقد عنوان باشند.

روش دیگر خلاصه‌سازی خلاصه سازی متن با استفاده از روش شبکه پیچیده با معرفی در پژوهش [11] بوده است. روش به کار رفته در این پژوهش، مواقعی مفید واقع می‌شود که خلاصه می بایست از چندین زبان مختلف که بعضاً منبع زبانی درستی ندارند، ایجاد شود. روش معرفی شده، دارای رویکردی مستقل از زبان است و خلاصه‌ها تنها با استفاده از دانش ضمنی در مورد زبان‌ها ابجاد می‌شوند. ارائه‌ی مفاهیم موجود در یک شبکه‌ی پیچیده تنها در یک شبکه‌ی واحد از مهم‌ترین مزیت‌های روش مورد استفاده در پژوهش [11] است.

ایجاد خودکار خلاصه‌ی generic از طریق فاکتورسازی ماتریس غیر منفی نیز یکی دیگر از تلاش‌های صورت گرفته در حوزه خلاصه‌سازی اسناد متنی است که توسط [12] صورت گرفته است. این تکنیک مناسب برای موقعیت‌هایی است که در آن‌ها خلاصه‌های آموزش دیده برای آموزش سیستم در اختیار نیست و خصیصه‌های معنایی متن باید در نظر گرفته‌شوند. روش آموزش به کار رفته در این روش، بدون ناظر و بدون نیاز به مجموعه آموزشی داده‌های متنی است و هدف استخراج جملات معنادارتر و بررسی عمیق‌تر معنا در زیرمجموعه‌های متنی سند است.

خلاصه سازی خودکار متن با استفاده از رویکردهای مبتنی بر بر آمار توسط [13] صورت گرفت. این روش بیش‌تر در خلاصه سازی متون برای ایجاد خلاصه متنی از زبان مبدایی خاص به زبان مقصدی متفاوت، مناسب است. مزیت این روش،

آموزش زبان با استفاده از مدل‌های نام‌برده شده در یک زبان و آزمایش آن در زبانی دیگر است و تمامی خصیصه‌ها برای ساخت دیتاست مستقل از زبان به جز کلمات کلیدی و کلمات کلیدی منفی در نظر گرفته می‌شوند.

خلاصه سازی اسناد متنی متعدد بر اساس مدل‌های رگرسیونی مبتنی بر پرس و جو[1] که توسط [14] ارائه شد و در آن خلاصه سازی چند سنده بر اساس روش‌های یادگیری ماشین و بر مبنای خصیصه‌های مبتنی بر پرس و جو صورت گرفت. این روش برای ایجاد داده‌های شبه آموزشی برای اندازه‌گیری میزان کارایی روش‌های دیگر مورد توجه است. مزیت روش استفاده شده در پژوهش ذکر شده، مناسب‌تر بودن روش‌های رگرسیونی به نسبت روش‌های طبقه بندی و آموزشی برای رتبه‌دهی است. خلاصه سازی متن با استفاده از حداکثر پوشش و حداقل افزونگی با رویکرد بهینه سازی برای حل مساله و عدم وجود خلاصه‌های آموزش دیده و با هدف پوشش محتوای مهم با حداقل افزونگی توسط [15] ارائه شد. ویژگی اساسی روش به کار گرفته شده، ایجاد خلاصه با رویکرد بدون ناظر و بدون نیاز به خلاصه‌های آموزشی با بیشترین پوشش محتوای مهم و کمترین جملات و محتوای تکراری است.

خلاصه سازی متن با استفاده از حداکثر پوشش و حداقل افزونگی با رویکرد بهینه سازی برای حل مساله و عدم وجود خلاصه‌های آموزش دیده و با هدف پوشش محتوای مهم با حداقل افزونگی توسط [15] ارائه شد. ویژگی اساسی روش به کار گرفته شده، ایجاد خلاصه با رویکرد بدون ناظر و بدون نیاز به خلاصه‌های آموزشی با بیشترین پوشش محتوای مهم و کمترین جملات و محتوای تکراری است.

استخراج خلاصه‌ی تک سند متنی از طریق عملگرهای ژنتیک و جستجوهای محلی هدایت شده با هدف خلاصه سازی سندمتنی واحد و مستقل از زبان و با نیاز به هدایت الگوریتم به سمت فضای جست و جوی امیدبخش توسط [16] صورت گرفت. مزیت روش، مستقل از زبان و دامنه بودن تمامی خصیصه‌های به کار رفته در ساخت دیتاست است. استفاده از الگوریتم Memetic برای هدایت در فضای جست و جو و انتخاب بهترین راه حل و نیز استفاده از جهش چند بیتی برای ایجاد بیشترین چگالی اطلاعات از دیگر ویژگی‌های روش به کار رفته در پژوهش نام برده شده است.

خلاصه سازی جملات مرتبط بر اساس روش مبتنی بر یادگیری توسط [17] معرفی شد و هدف استفاده از فشرده-سازی چند جمله و تبدیل آن‌ها به یک جمله‌ی گرامری واحد و صحیح با نگه‌داشتن بیشترین حد از معنا بود. خلاصه سازی با استفاده از حداکثر ۵ خصیصه و کاهش مرتبه‌ی الگوریتم ساخت گراف نسبت به روش‌های دیگر مبتنی بر گراف مزیت استفاده از روش به کار رفته در این پژوهش بود.

خلاصه سازی سند متنی واحد با استفاده از ساختار درخت تو در تو (درخت نهفته) با هدف ساخت خلاصه‌ی معنی محور با استفاده از تشخیص جملات و کلمات همبسته در متن و بهینه‌سازی وابستگی بین آن‌ها و توسط [18] صورت گرفت. مزیت روش مورد استفاده در این پژوهش توجه به وابستگی میان کلمات و وابستگی میان جملات برای ساخت درخت تو در تو (درخت وابستگی) است.

[1] Query

خلاصه سازی مبتنی بر موضوع سند متنی و از طریق انتخاب گروه‌ها و برای ایجاد خلاصه بر مبنای موضوعات مختلفی که باید در خلاصه‌ی ایجاد شده در نظر گرفته شده باشند، توسط [19] صورت گرفت. بهبود کلی کیفیت خلاصه‌ی ایجاد شده از طریق اضافه کردن گروه‌های متمایز برای شرکت در مفهوم خلاصه‌ی نهایی مهم‌ترین مزیت روش نام برده شده است و از طرفی دیگر هر دو پارامتر پوشش معنایی و تنوع معنایی در این روش در نظرگرفته‌شده است.

خلاصه سازی مبتنی بر گراف با توجه به اهمیت معنایی جملات، فقدان افزونگی و انسجام معنایی متون توسط [20] و با هدف ساخت خلاصه‌ی منسجم محلی و فاقد افزونگی از ژانرها و حوزه‌های متفاوت صورت گرفت. روش خلاصه‌سازی به کار رفته در این پژوهش، مستقل از پارامتر و بدون نیاز به آموزش داده‌ها (روش خلاصه سازی بدون ناظر) بوده است و منتج به ایجاد خروجی با کیفیت شد. محدود شدن بازده رویکرد تنها به خلاصه سازی سند متنی واحد از جمله محدودیت‌های این روش است.

ازجمله اخیرترین روش‌های خلاصه‌سازی متن مبتنی بر گراف نیز در پژوهش خلاصه سازی تجریدی متن با استفاده از گراف بهبودیافته و با هدف استفاده از گراف بهبودیافته معنایی برای غلبه بر مشکل عدم شناسایی جملات افزونه و بیان و از دست رفتن معنا در نمایش متن در قالب کیف کلمات و توسط [21] صورت گرفت. مدل‌سازی سند متنی با نگاشت هر جمله‌ی سازنده‌ی متن به یک گره از گراف انجام می‌گیرد و یال‌های گراف جملاتی هستند که از نظر معنایی با هم در ارتباط هستند و وزن هر یال مبین میزان ارتباط معنایی جملات با یکدیگر است. مزیت اساسی این روش خلاصه‌سازی در استقلال روش از حوزه‌ی معنایی متن و عدم نیاز به ساخت آنتولوژی از متن است که این مورد را می‌توان نیازمندی و محدودیت اکثر روش‌های خلاصه‌سازی مبتنی بر گراف دانست. نتیجه پژوهش صورت گرفته، ایجاد گراف معنایی برای خلاصه سازی تجریدی متن با استفاده از الگوریتم رتبه دهی PageRank گوگل است.

3. رویکردهای ارزیابی خلاصه‌های متنی

پس از ایجاد خلاصه‌های معنایی از متون الکترونیکی، نوبت به ارزیابی آنها می‌رسد. رویکردهای ارزیابی خلاصه‌های متنی را می‌توان با استفاده از ۴ تکنیک MCS, ROUGE-N, Co-Selection و قضاوت متخصصان انسانی پیاده سازی کرد. در ادامه توضیح مختصری از هر یک از تکنیک‌ها آمده است.

تکنیک ارزیابی Mean Coverage Score (MCS) و یا Pyramid Score: پس از ایجاد خلاصه با استفاده از تشکیل گراف معنایی، نتایج با خلاصه‌های ایجاد شده توسط انسان که در پایگاه داده‌ی (DUC, 2002) جمع آوری شده اند، مقایسه شده و امتیاز MCS مطابق رابطه (۱) محاسبه می‌شود. پس از محاسبه‌ی MCS هم می‌بایست دو معیار Precision و F-Measure به ترتیب بر اساس رابطه‌های (۲) و (۳) اندازه گیری شوند [22].

۱- تکنیک ارزیابی ROUGE – N: بر اساس معیار N-gram Recall میان خلاصه‌های ایجاد شده توسط انسان و نیز خلاصه‌های ایجاد شده توسط سیستم و بر اساس رابطه (۴) که در ادامه آمده است [23].

۲- تکنیک ارزیابی ارزیابی محتوایی با استفاده از Co-Selection: Precision, Recall و F-score معیارهای دقت، فراخوانی و امتیاز F- اساسی‌ترین معیارهای ارزیابی خلاصه‌های تولید شده توسط سامانه‌های خودکار خلاصه‌ساز هستند [۲۴]. روش محاسبه‌ی Recall، Precision و F-score به ترتیب در رابطه‌های (۵)، (۶) و (۷) آمده است.

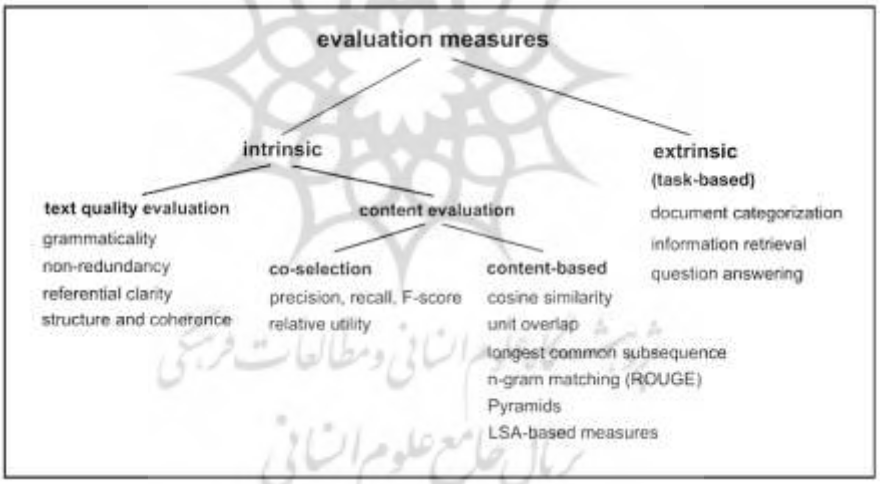
۳- تکنیک ارزیابی قضاوت متخصصان انسانی Human Coherence Judgements: به دلیل مشکلات موجود در ارزیابی‌های فوق اعم از نادیده گرفته شدن ترتیب حضور جملات در خلاصه‌ی ایجاد شده از روش قضاوت انسانی برای ارزیابی خلاصه‌های تولید شده استفاده می‌شود. روش کار به این صورت است که ۳ نفر از متخصصان حوزه‌ی آی‌تی و پردازش زبان‌های طبیعی متون خلاصه شده را قرائت کرده و بسته به خلاصه‌ی ایجاد شده امتیازهای زیر را به خلاصه‌ها می‌دهند و در نهایت از امتیازهای اخذ شده برای خلاصه‌ها میانگین گرفته می‌شود.

امتیاز ۱: خلاصه تولید شده منسجم و شامل است.

امتیاز ۲: خلاصه تولید شده نسبتاً منسجم و شامل است.

امتیاز ۳: خلاصه تولید شده فاقد انسجام و شمول مورد قبول است.

به طور کلی، تمامی معیارهای لازم برای ارزیابی خلاصه‌های تولید شده توسط سیستم‌های خلاصه ساز خودکار متون را می‌توان در شکل ۱. خلاصه کرد.



شکل ۱ – معیارهای ارزیابی خلاصه‌های تولیدی توسط سامانه‌ی خلاصه ساز [۲۴]

بر اساس میانگین به دست آمده از نظرات متخصصان، هرچه میانگین تولیدی به عدد ۱ نزدیک‌تر باشد، عملکرد خلاصه‌ساز مورد استفاده بهتر خواهد بود [۲۰].

$$\mathbf{MCS} = \frac{\sum \text{وزن } SCUs \text{ موجود در خلاصه‌ی کاندیدا}}{\text{میانگین } SCU \text{ در خلاصه‌ی تولیدی توسط متخصصان}}$$

(۱)

که در آن SCU[۲] واحدهای محتوایی موجود در خلاصه‌ها و وزن SCUs برابر باتعداد خلاصه‌های ایجاد شده توسط متخصصان است.

$$P = \frac{\text{تعداد } SCU \text{ های موجود در خلاصه‌ی مرجع و کاندیدا}}{\text{میانگین } SCU \text{ در خلاصه‌ی کاندیدا}} \quad (۲)$$

که در آن SCU واحدهای محتوایی موجود در خلاصه‌ها هستند و P بیانگر معیار Precision می‌باشد. در انتها زیر معیار F-Measure برای خلاصه‌های کاندیدا، بر اساس دو رابطه (۱) و (۲) و مطابق با رابطه (۳) محاسبه می‌شود.

$$\mathbf{F-Measure} = \frac{۲ \times MCS \times P}{\mathbf{MCS + P}} \quad (۳)$$

که در آن MCS امتیاز میانگین شمول میان جملات تولید شده در خلاصه‌ی کاندیدا و خلاصه‌ی مرجع است و از رابطه (۱) استخراج می‌شود.

$$\mathbf{ROUGE-N} = \frac{\sum_{S \in (\text{ReferemceSummaries})} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in (\text{ReferenceSummaries})} \sum_{gram_n \in S} Count(gram_n)} \quad (۴)$$

که در آن متغیر n برابر با طول n-gram و $(gram_n)$ و $Count_{match}(gram_n)$ برابر با بیشترین تعداد n-gram های ظاهر شده در هر دو سند خلاصه‌ی کاندیدا و خلاصه‌ی مرجع (خلاصه‌ی انسانی تولید شده توسط متخصصان و حاضر در پایگاه داده) است. متغیر $Count_{match}(gram_n)$ نیز مبین تعداد n-gram های استخراج شده از متن خلاصه‌شده است. همانطور که معلوم است، ROUGE-N از معیار Recall برای ارزیابی خلاصه‌های ایجاد شده استفاده می‌کند؛ چرا که درصد رخداد n-gram های مطابق در دو سند خلاصه‌ی کاندیدا و مرجع را اندازه‌گیری و محاسبه می‌کند.

(۵)

[2] **Summarization Content Units**

$$Recall = \frac{\sum \text{تعداد جملات منطبق در هر دو خلاصه کاندیدا و مرجع}}{\sum \text{تعداد جملات در خلاصه مرجع}}$$

(۶)

$$\text{Precision} = \frac{\sum \text{تعداد جملات منطبق در هر دو خلاصه کاندیدا و مرجع}}{\sum \text{تعداد جملات در خلاصه کاندیدا}}$$

(۷)

$$\text{F} - \text{score} = \frac{(\beta^{2}+1) \times \text{Recall} \times \textit{Precision}}{\beta^{2} \times \textit{Precision} + \text{Recall}}$$

که در آن β یک پارامتر وزن‌دهی است و اگر $\beta > 1$ باشد تمایل امتیاز F به سمت دقت و در صورت $\beta < 1$ تمایل امتیاز F به سمت فراخوانی خواهد بود [۲۴].

4. روش پژوهش

برای خلاصه سازی مقالات اخباری، با استفاده از روش درج شده در مقاله‌ی [۲۱] ابتدا گراف معنایی را بر پایه متن ورودی که همان نسخه‌ی نرمال‌شده‌ی اسناد هستند، ساخته و سپس با استفاده از این گراف نودهای با اولویت بالا را انتخاب و به عنوان جملات اصلی متن معرفی می کنیم و در حقیقت متن را خلاصه سازی می کنیم. روند کار بدین صورت است که هر جمله‌ی مقالات اخباری به عنوان یک گره از گراف انتخاب شده و روابط جملات بر اساس ارتباط معنایی اجزای جملات بر روی گراف به شکل یال ها نمود پیدا می کند.

اولین بخش بعد از پارس شدن متن اصلی، محاسبه SRL هست که این مهم بر اساس بسته‌ی معروف SENNA انجام می شود. این بسته مستقل از زبان و حوزه‌ی معنایی متن نیست و تنها برای پردازش زبان‌های طبیعی زبان انگلیسی بوده و بر روی جملات انگلیسی قابل اجرا می باشد. از آنجا که ورودی متون در پژوهش حاضر فارسی است؛ لذا پردازش معنایی با استفاده از بسته‌ی SENNA با مشکل در استفاده از ابزار اولیه برای ادامه کار مواجه می شود. به همین جهت می بایست روشی مشابه که رویکردی مستقل از زبان و حوزه‌ی معنایی متن داشته باشد و بتواند نیازمندی‌های خلاصه سازی را با استفاده از گراف معنایی تامین کند برگزیده شود.

در این پژوهش، با استفاده از الگوریتم TextRank که به نحوی الهام گرفته از الگوریتم PageRank گوگل در ایندکس کردن صفحات اینترنتی می باشد، متن را خلاصه سازی کردیم. TextRank الگوریتم بدون ناظر و مبتنی بر گرافی است که اولین بار توسط Rada Mihalcea و Paul Tarau معرفی شد و برای خلاصه سازی خودکار اسناد متنی به کار رفت. با استفاده از TextRank می‌توان کلیدواژه‌های مهم در متن را استخراج کرد. روش خلاصه‌سازی با استفاده از TextRank روشی است مستقل از حوزه و زبان و نیاز به دانش عمیق در مورد زبان ندارد. همین مساله هم باعث شده است که این الگوریتم به صورت گسترده‌ای در پژوهش‌های پردازش زبان طبیعی مورد استقبال قرار بگیرد. از این روش بیشتر در خلاصه -

سازی مقالات اخباری، اعتبارسنجی محتوای وب و نیز خلاصه‌سازی گزارش‌های جلسات استفاده می‌شود. در این روش با استفاده از اعمال مکرر الگوریتم رتبه‌دهی به صفحات وب[3] که توسط گوگل ارائه و استفاده می‌شود، گراف ساخته شده از متن را برای ایجاد خلاصه‌ی نهایی آماده می‌کنند [۲۵].

مدل سازی سند متنی به صورت گراف به این صورت است که هر جمله‌ی سند متنی به عنوان یک نود در گراف درنظر گرفته شده و وزن یالها میزان شباهت معنایی جملاتی است که با آن یال‌ها به هم مرتبط شده‌اند؛ به این ترتیب، گره ای دارای وزن بیشتری است که بیشترین ارتباط معنایی را با دیگر گره ها یا جملات دارد و گره‌های با وزن بیشتر برای شرکت در متن خلاصه انتخاب می‌شوند.

خلاصه‌سازی مقالات اخباری در این پژوهش با استفاده از الگوریتم بهبود یافته‌ی TextRank صورت گرفته‌است. پاره‌های متنی که برای ساخت گراف معنایی استفاده می‌شوند، می‌توانند عبارات، جملات و یا پاراگراف‌های سازنده‌ی متن باشند. شاید جملات را بتوان بهترین کاندیدا برای گره بودن در گراف دانست؛ چرا که در این‌صورت هم معنای کلمات در کنار یکدیگر و هم قواعد گرامری لحاظ خواهند شد. در روش اتخاذ شده برای ساخت گراف، جملات با بیشترین ارتباط معنایی با سایر جملات موجود در متن (گره‌های با بالاترین درجه) برای ساخت خلاصه‌ی نهایی استفاده می‌شوند. این جملات می‌توانند نماینده‌ی بهتری برای ارائه‌ی مفهوم کلی متن باشند و معنای کلی متن را منتقل کنند. با استفاده از این روش دیگر نیازی به آموزش و برچسب‌زنی به ادات متن نیست. برای مشخص شدن ارتباط میان جملات (ارتباط میان جملات را با یال‌هایی وزن‌داری که جملات مرتبط را به یکدیگر متصل می‌کنند، نمایش می‌دهیم)، می‌توان از چندین معیار استفاده نمود که از آن جمله می‌توان به کلمات هم‌پوشا، فاصله‌ی کسینوسی و شباهت حساس به جست‌وجو اشاره کرد [۲۶].

برای خلاصه‌سازی اسناد متنی که از مدل TextRank استفاده می‌کنند، سند متنی به مثابه‌ی گراف مدل می‌شود. اجزای سازنده‌ی گراف به شرح زیر هستند:

- گره‌های گراف: جملات سازنده‌ی سند متنی
- یال‌های گراف: ارتباط میان جملات دارای شباهت معنایی به یکدیگر
- وزن گره‌های گراف: بر اساس رابطه (۸) شباهت معنایی میان جملات صورت گرفته و وزن دهی یال‌های گراف معنایی بر اساس محتوای مشترک میان جملات انجام می‌شود. محاسبه‌ی هم‌پوشانی معنایی میان دو جمله بر اساس نسبت توکن[4]های مشابه میان آن دو جمله به طول جملات مورد مقایسه است.

(۸)

$$\mathbf{Sim}(s_i, s_j) = \frac{|\{w_k | w_k \in s_i \ \& \ w_k \in s_j\}|}{\log(|s_i|) + \log(|s_j|)}$$

که در آن S_i , S_j دو جمله‌ی اختیاری از متن هستند که هر کدام از آن‌ها توسط n کلمه ساخته‌شده اند. کلمات سازنده‌ی جمله‌ی S_i را میتوان توسط معادله‌ی $S_i = w^i_1 , w^i_2 , ..., w^i_n$ مدل کرد.

[3] PageRank

[4] Token

$Sim(S_i, S_j)$ تابع شباهت میان دو جمله است و وزن نهایی یالی است که دو گره را در گراف به یکدیگر متصل می-کند و W_k نیز بیانگر kامین کلمه سازنده‌ی جمله است.

خروجی این فرآیند گراف چگال معنایی است که نماینده‌ی معنایی متن است. در مرحله‌ی بعدی با اجرای الگوریتم PageRank اهمیت (وزن) هر گره (جمله) مشخص می‌شود و جملات مهم‌تر (گره‌های با وزن بیشتر) برای شرکت در خلاصه‌ی نهایی انتخاب شده و با همان ترتیبی که در متن اصلی آمده‌اند در خلاصه ظاهر می‌شوند.

از آنجا که پایه اصلی الگوریتم خلاصه سازی gensim می باشد، لذا تغیرات جهت فارسی سازی این الگوریتم در فایل اصلی این کتابخانه اعمال شد و در نهایت با اعمال فارسی سازی در فایل textcleaner از بسته gensim، این خلاصه‌ساز به شکل کتابخانه در پایتون استفاده شد. اساس خلاصه‌سازی gensim بر پایه الگوریتم TextRank گوگل است.

برای فارسی سازی کتابخانه‌ی genism، فایل textcleaner.py به منظور ایمپورت تابع POS کتابخانه هزم و اعمال برچسب POS بر روی متن مقالات خبری تغییر پیدا کردند. به این ترتیب، ابتدا فایلهای خزش شده از سایت خبری فارسی کوک موبایل و نرمال‌سازی شده را بر روی پروژه بارگذاری کرده و سپس یک به یک به تابع خلاصه سازی پاس داده شدند. خروجی نهایی که شامل متون خلاصه‌شده است، در فایل های شاخه ای با نام Sum ذخیره شدند.

5. نتیجه گیری

پس از ایجاد خلاصه از متون اخبار، می‌بایست آن‌ها را ارزیابی کرد تا میزان انتقال مفاهیم اساسی و معنا توسط خلاصه‌های ایجاد شده سنجیده شود. برای این منظور با بررسی برخی از مقاله‌های مرتبط با ارزیابی سامانه‌های خلاصه ساز اسناد متنی که در بخش رویکردهای ارزیابی خلاصه‌های متنی آمد، در مورد انتخاب روش مناسب ارزیابی خلاصه‌های ایجاد شده که ورودی به فاز استخراج برچسب‌های معنایی هستند، تصمیم خواهیم گرفت.

روبکرد مورد استفاده در ارزیابی سامانه توسعه داده‌شده، محاسبه F-Score است و محاسبه دقت، فراخوانی و F-Score سامانه در مقایسه با عملکرد سامانه‌های خلاصه ساز ایجاز و Noortm و بر اساس رابطه‌های (۵)، (۶)، (۷) عنوان شده در بندهای بالاتر، در جدول ۱ آورده شده است. در محاسبه F-Score اندازه پارامتر β برابر با عدد ۱ در نظر گرفته شده است.

جدول ۱- ارزیابی عملکرد سامانه خلاصه ساز توسعه داده شده در پژوهش

	دقت (Precision)	فراخوانی (Recall)	F-Score
سامانه پیشنهادی در مقایسه با سامانه ایجاز	٪۴۷,۹	٪۳۳,۵	٪۳۹,۴
سامانه پیشنهادی در مقایسه با سامانه متن کاوی نور	٪۷۴,۸	٪۷۴,۶	٪۷۸,۷

پس از ارائه اسناد خلاصه شده به ۳ نفر از متخصصان حوزه آی‌تی، امتیاز داده شده به انسجام و شمول خلاصه‌های ایجاد شده توسط سامانه توسعه داده شده در پژوهش عدد ۲ و عملکرد سامانه ایجاز ۳ ارزیابی گردید که صحت ادعای شمول و جامعیت معنایی جملات موجود در خلاصه‌های تولید سامانه توسعه‌داده شده را تایید می‌کند. مقدار F-Score سامانه توسعه داده شده در مقایسه با عملکرد سامانه ایجاز را می‌توان به دلیل ارجحیت سامانه ایجاز به انتخاب جملات کوتاه‌تر و آوردن آن‌ها در خلاصه نهایی توجیه نمود. از نتایج این پژوهش می‌توان در جهت توسعه سامانه‌های خلاصه ساز برخط در سایت‌های خبری به زبان فارسی و ایجاد خلاصه‌های معنایی از اسناد الکترونیکی فارسی استفاده نمود.

این پژوهش تنها بر اساس رویکرد خلاصه سازی تک سندی و با اجرا بر روی هر سند خبری به صورت مجزا اجرا و ارزیابی گردیده است. سایر پژوهشگران علاقه‌مند به حوزه پردازش زبان‌های طبیعی می‌توانند روش مورد استفاده در پژوهش را برای خلاصه سازی چند سندی نیز به کار برده و نتایج آن را تحلیل و بررسی نمایند.

6. قدردانی

در انتها بر خود واجب می‌دانم که از دانشجویان مقطع کارشناسی ارشد رشته مدیریت فناوری اطلاعات دانشگاه الزهرا (س) که اینجانب را در انجام فاز ارزیابی سامانه یاری رساندند، تشکر و قدردانی کنم.

7. مراجع

[1] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, “A text summarization method based on fuzzy rules and applicable to automated assessment,” *Expert Syst. Appl.*, vol. 115, pp. 264–275, 2019.

[2] F. Kiyani and O. Tas, “A survey automatic text summarization,” *Pressacademia*, vol. 5, no. 1, pp. 205–213, 2017.

[3] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” *arXiv Prepr. arXiv1707.02919*, 2017.

[4] R. Alguliyev, R. Aliguliyev, N. Isazade, A. Abdi, and N. Idris, “A Model for Text Summarization,” *Int. J. Intell. Inf. Technol.*, vol. 13, pp. 67–85, Jan. 2017.

[5] R. Pieters, *Deep Learning for Natural Language Processing: Word Embeddings*, no. December. 2015.

[6] S. Kumar and K. K. Bhatia, “Semantic similarity and text summarization based novelty detection,” *SN Appl. Sci.*, vol. 2, no. 3, p. 332, 2020.

[7] J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I.-H. Meng, “Text summarization using a trainable summarizer and latent semantic analysis,” *Inf. Process. Manag.*, vol. 41, no. 1, pp. 75–95, 2005.

[8] S. W. K. Chan, “Beyond keyword and cue-phrase matching: A sentence-based

abstraction technique for information extraction," *Decis. Support Syst.*, vol. 42, no. 2, pp. 759–777, 2006.

[9] S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu, "Document concept lattice for text understanding and summarization," *Inf. Process. Manag.*, vol. 43, no. 6, pp. 1643–1662, 2007.

[10] Y. Ko and J. Seo, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1366–1371, 2008.

[11] L. Antiqueira, O. N. Oliveira, L. da Fontoura Costa, and M. das Graças Volpe Nunes, "A complex network approach to text summarization," *Inf. Sci. (Ny).*, vol. 179, no. 5, pp. 584–599, 2009.

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, p. 788, Oct. 1999.

[13] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 126–144, 2009.

[14] Y. Ouyang, W. Li, S. Li, and Q. Lu, "Applying regression models to query-focused multi-document summarization," *Inf. Process. Manag.*, vol. 47, no. 2, pp. 227–237, 2011.

[15] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, "MCMR: Maximum coverage and minimum redundant text summarization model," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14514–14522, 2011.

[16] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos, and E. León, "Extractive single-document summarization based on genetic operators and guided local search," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4158–4169, 2014.

[17] E. Tzouridis, J. A. Nasir, and U. Brefeld, "Learning to Summarise Related Sentences," *Proc. COLING 2014, 25th Int. Conf. Comput. Linguist.*, pp. 1636–1647, 2014.

[18] Y. Kikuchi, T. Hirao, H. Takamura, M. Okumura, and M. Nagata, "Single Document Summarization based on Nested Tree Structure EDU sentence subtree sentence selection selection," *52nd Annu. Meet. Assoc. Comput. Linguist.*, pp. 315–320, 2014.

[19] H. Fang *et al.*, "Topic aspect-oriented summarization via group selection," *Neurocomputing*, vol. 149, pp. 1613–1619, 2015.

[20] D. Parveen and M. Strube, "Integrating importance, non-redundancy and coherence in graph-based extractive summarization," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2015-Janua, no. Ijcai, pp. 1298–1304, 2015.

[21] A. Khan *et al.*, *Abstractive Text Summarization based on Improved Semantic Graph Approach*. 2018.

[22] a Nenkova and R. Passonneau, "Evaluating content selection in summarization: The

pyramid method," *Proc. HLT-NAACL*, vol. 2004, pp. 145–152, 2004.

[23] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.

[24] Stratton, "Chapter 3," pp. 39–52.

[25] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," *Proc. EMNLP*, vol. 85, pp. 404–411, 2004.

[26] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization," 2016.