ایجاز: یک سامانهٔ عملیاتی برای خلاصه‌سازی تک‌سندی متون خبری فارسی

آصف پورمعصومی، محسن کاهانی، سید احمد طوسی، احمد استیری و هادی قائمی
آزمایشگاه فناوری وب، گروه کامپیوتر، دانشکدهٔ مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران

چکیده

امروزه با رشد چشم‌گیر اسناد منتشرشده در وب و نیاز اساسی به نگهداری، دسته‌بندی، بازیابی و پردازش آنها، توجه به پردازش زبان طبیعی و بهره‌گیری از ابزارهایی نظیر خلاصه‌سازهای خودکار و مترجم‌های ماشینی، بیش از پیش احساس می‌شود. خلاصه‌سازی خودکار به‌عنوان هستهٔ مرکزی طیف گسترده‌ای از ابزارهای پردازش‌گر متن مانند سامانه‌های تصمیم‌یار، سامانه‌های پرسش‌وپاسخ، موتورهای جستجو و غیره از سال‌ها پیش مطرح شده و همواره به‌عنوان یک موضوع مهم مورد بررسی و تحقیق قرار گرفته است. در این مقاله، سامانه‌ای به نام «ایجاز» به‌منظور خلاصه‌سازی‌خودکار تک‌سندی متون فارسی ارائه شده است. در این سامانه از تجربیات سامانه‌های مشابه داخلی و خارجی استفاده شده است؛ و با استفاده از پارامترهای جدید، دقت سامانهٔ خلاصه‌سازی ارائه شده به میزان قابل توجهی بهبود یافته است. همچنین برای اولین‌بار با بهره‌گیری از یک پیکرهٔ بزرگ خلاصه‌سازی و ابزار ارزیابی استاندارد، عملیات مقایسه و ارزیابی صورت گرفته است.

واژگان کلیدی: پردازش زبان فارسی، خلاصه‌سازی خودکار تک‌سندی، متون خبری.

۱- مقدّمه

در وب امروزی با توجه به گسترش روزافزون حجم اطلاعات و داده‌های انتشاریافته، دسترسی مناسب به اطلاعات مورد نیاز در کوتاه‌ترین زمان ممکن، همواره یکی از چالش‌های محققان و پژوهشگران حال حاضر است.

خلاصه‌سازی خودکار سند، تولید یک نسخهٔ مختصرتر از سند اصلی توسط یک برنامه رایانه‌ای است؛ به‌نحوی‌که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود (Mani, 1999).

سامانه‌های خلاصه‌سازی خودکار را می‌توان براساس اسناد ورودی به دو دستهٔ تک‌سندی[1] و چندسندی[2] تقسیم کرد. در سامانه‌های خلاصه‌سازی تک‌سندی، ورودی سامانهٔ خلاصه‌ساز تنها یک سند است؛ اما در سامانه‌های خلاصه‌سازی چندسندی، ورودی چندین سند با موضوعی مشترک است که مطالب اسناد در ارتباط با آن موضوع بوده ولی با دیدگاه‌های متفاوتی به آن پرداخته شده است (Mani, 1999). به‌عنوان مثال در ارتباط با موضوع "مشکل جهانی کمبود آب"، اسناد مختلفی می‌تواند وجود داشته باشد که از دیدگاه‌های متفاوت به این موضوع پرداخته‌اند. به‌عنوان مثال یکی در مورد "کمبود آب در ایران" و دیگری در خصوص "کمبود آب در پاکستان" باشد. هر دوی این سندها به موضوع کمبود آب از دیدگاه‌های مختلف پرداخته‌اند. یک سامانهٔ خلاصه‌ساز چندسندی مطلوب باید قادر باشد تا ضمن استخراج همهٔ مفاهیم مهم و اصلی پنهان در این اسناد، فاقد هرگونه افزونگی و یا مطالب متناقض باشد.

پیچیدگی خلاصه‌سازی تک‌سندی نسبت به خلاصه‌سازی چندسندی به مراتب کمتر است؛ چون در این مدل از خلاصه‌سازی، تنها با یک سند روبه‌رو هستیم که به‌احتمال در مورد یک موضوع و به‌صورت پیوسته صحبت

[1] Single Document
[2] Multi Document

می‌کند و فاقد زیرموضوعات ضد و نقیض خواهد بود (Mihalcea, 2005) (Svore, 2007).

انسان‌ها با توجه به هوش و شعور ذاتی خود قادر به درک و فهم مفاهیم موجود در متن و ارتباط بین آنها هستند و این در حالی است که انجام این عملیات توسط ماشین، کار بسیار دشوار و پیچیده‌ای است.

هدف نهایی سامانه‌های خلاصه‌سازی، تولید خلاصه‌هایی با کیفیت نزدیک به خلاصه‌های انسانی است؛ اما برای رسیدن به این مهم، چالش‌های زیادی وجود دارد. مهم‌ترین مشکل در گام نخست، انتخاب مناسب‌ترین جملات متن اصلی است به‌نحوی که مطالب مهم و اصلی متن را پوشش داده و در عین حال فاقد افزونگی و جملات تکراری یا شبیه به‌هم باشد. عمده تمرکز سامانه‌های خلاصه‌سازی به این موضوع اختصاص دارد.

پردازش‌هایی که بر روی متن‌های زبان طبیعی صورت می‌گیرد اغلب نیازمند عملیات پیش‌پردازش است؛ به‌طوری‌که دقت این پیش‌پردازش‌ها تأثیر به‌سزایی بر نتایج اعمال الگوریتم‌ها در مراحل بعدی دارد. هرقدر دقت این پیش‌پردازش‌ها بیشتر باشد، الگوریتم‌ها به نتایج واقعی خود نزدیک‌تر خواهند شد.

ساختار این مقاله به شرح زیر است. در بخش دوم مروری بر کارهای گذشته و در بخش سوم روش پیشنهادی برای خلاصه‌سازی تک‌سندی متون فارسی ارائه خواهد شد. ارزیابی کار نیز در بخش چهارم آمده است. در پایان نیز نتیجه‌گیری و کارهای آینده ذکر شده است.

۲- مروری بر کارهای گذشته

لازم به ذکر است که کارهای انجام‌شده در زمینهٔ پیاده‌سازی ابزارهای خلاصه‌سازی خودکار فارسی، بسیار اندک هستند (Dalianis, 2000) (Poormasoomi, 2011) (Kiyoumarsi, 2011) (Mazdak, 2004) (اخوانت، ۱۳۸۷) (کریمی، ۱۳۸۵) (مشکی, ۱۳۸۸) (مشکی، ۱۳۸۶) (ستوده، ۱۳۸۹). سنجش صحت و دقت روش‌های مطرح شده، به‌دلیل عدم استفاده از یک پیکرهٔ متون متحد و ابزار ارزیابی یکپارچه و استاندارد، قابل محاسبه و قیاس نیست.

از جمله خلاصه‌سازهای خودکار برای زبان فارسی می‌توان FarsiSum را نام برد (Mazdak, 2004) که نسخهٔ تغییریافتهٔ SweSum است (Dalianis, 2000) که برای متون سوئدی مورد استفاده قرار می‌گیرد. با توجه به اینکه ایدهٔ اصلی این سامانه برمبنای سامانهٔ SweSum است و در سامانهٔ SweSum همیشه اولین جمله به‌عنوان یکی از جمله‌های مهم متن در نظر گرفته می‌شود، در FarsiSum نیز به جمله‌های اول، وزن زیادی داده می‌شود که البته این موضوع، چندان مناسب نیست. تحقیقات آماری صورت‌گرفته بر روی پیکرهٔ پاسخ (Behmadi, 2013) (این پیکره شامل متون خبری و خلاصه‌های مربوط به آنها است) نشان می‌دهد که شانس انتخاب‌شدن جملات پایانی در متن خلاصه، حدوداً برابر با نصف شانس حضور جملات ابتدایی است. این مطلب تا حدودی با معیار تعریف‌شده در خلاصه‌سازهای ماشینی یادشده متفاوت است.

سامانهٔ ارائه‌شده در (کریمی، ۱۳۸۵)، یک خلاصه‌ساز تک‌سندی است که عملکردش بر مبنای گزینش جملات است. در این سامانه برای گزینش جملات خروجی از یک روش ترکیبی استفاده شده که ترکیبی از دو روش، زنجیرهٔ لغوی و نظریهٔ گراف است و خروجی نهایی می‌تواند کلی یا بر اساس پرس‌وجوی کاربر باشد. در (مشکی، ۱۳۸۶) پس از بررسی چالش‌های مربوط به پردازش متون فارسی، انواع روش‌های موجود در خلاصه‌سازی مورد بررسی قرار گرفته است. در (مشکی، ۱۳۸۸) یک نمونه ابزار برای خلاصه‌سازی متون فارسی پیاده‌سازی شده است. این ابزار برای خلاصه‌سازی چندسندی متون فارسی مورد استفاده قرار گرفته و مدل خروجی آن انتخابی بوده و از یک روش مبتنی بر خوشه‌بندی بهره می‌گیرد. در (اخوانت، ۱۳۸۷) سامانه PARSUMIST ارائه شده است که با استفاده از رویکرد آماری و در نظر گرفتن تعدادی از ویژگی‌های زبان‌شناختی زبان فارسی، خلاصه را تولید می‌کند.

در (Honarpisheh, 2007) یک سامانهٔ خلاصه‌ساز چندسندی چندزبانی ارائه شده است که براساس SVR[1] عمل می‌کند. خلاصهٔ حاصله، از نوع گزینشی بوده، روش مورد استفاده جزو دستهٔ آماری محسوب می‌شود. در (ستوده، ۱۳۸۹) با الهام از شیوهٔ انسان در تفکر، به جای گزینش مهم‌ترین جملات، زنجیره‌ای از جملات در متن انتخاب می‌شود که با یکدیگر، همبستگی و ارتباط بیشتری داشته باشند. اگرچه ممکن است هر یک از جملات زنجیره، به‌تنهایی مهم نباشند؛ اما ممکن است با در نظر گرفتن جملات در زنجیره، پازلی را تکمیل کنند و این پازل در کل بیان‌گر مفهوم متن است. در این خلاصه‌سازی، هدف آن است که از هر پاراگراف، جمله‌ای برجسته به شکلی انتخاب

[1] Support Vector Regression

شود که ارتباط و همبستگی خوبی با جملات برجسته پارگراف‌های دیگر داشته باشد.

در (Kiyoumarsi, 2011) با بهره‌گیری از خلاصه‌سازی فازی، سعی شده است با استفاده از ویژگی‌های خاص زبان و همچنین منطق فازی، متن ورودی به شکلی مناسب خلاصه شود. از آنجا که این خلاصه‌سازی، گزینشی بوده و سعی در انتخاب جملات مهم‌تر و مناسب‌تر برای خلاصه دارد، یک سری ویژگی برای جملات تعریف می‌شود و منطق فازی برای انتخاب جملات، استفاده شده است. در (Poormasoomi, 2011) با ارائهٔ یک روش جدید برای استخراج روابط معنایی موجود در متن (LSA[1]) و استفاده از تکنیک برچسب‌زنی معنایی نقش لغات (SRL[2])، روشی جدید برای خلاصه‌سازی چندسندی ارائه شده است.

با توجه به اینکه در سال‌های گذشته، ابزار و پیکرهٔ مناسب برای ارزیابی خلاصه‌سازها وجود نداشته است، به همین دلیل ارزیابی مناسبی از سامانه‌های معرفی شده، موجود نیست. در ادامه، مهم‌ترین خلاصه‌سازهای موجود برای سایر زبان‌های دنیا معرفی خواهند شد.

سامانهٔ CATS یک سامانهٔ خلاصه‌سازی چندسندی با توجه به‌عنوان مشخص‌شده توسط کاربر است (Farzindar, 2005). در این سامانه ابتدا اسناد به‌صورت موضوعی تجزیه شده و سپس این موضوعات با موضوعات مشخص‌شده در سؤال، تطبیق داده می‌شوند. قسمت‌های موضوعی که شامل موارد مرتبط با موضوع است، فهرست شده و با توجه به احتمالشان مرتب می‌شوند. بخش‌هایی که به سایر بخش‌ها شباهت دارند، حذف شده و در مرحلهٔ آخر جملات با بیشترین امتیاز تا زمانی‌که خلاصه‌ای شامل حداکثر ۲۵۰ کلمه ایجاد شود، انتخاب می‌شوند.

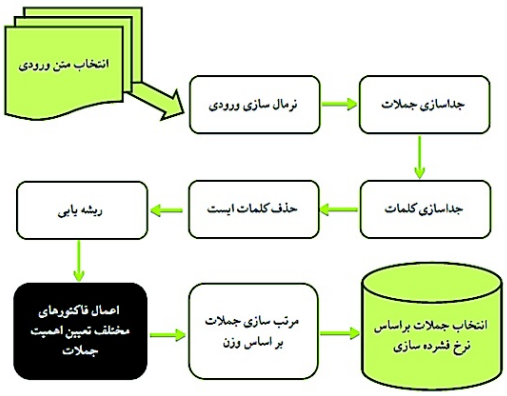
MEAD یکی از معروف‌ترین سامانه‌های خلاصه‌سازی چندسندی است (Radev, 2004). نسخهٔ ۱ و ۲ این خلاصه‌ساز در دانشگاه میشیگان در سال ۲۰۰۰ – ۲۰۰۱ پیاده‌سازی شده است. در حال حاضر نسخه‌های انگلیسی، چینی، ژاپنی و هلندی آن موجود است. سامانهٔ MEAD با استفاده از خوشه‌بندی، جملات مرکزی در هر خوشه را شناسایی کرده و خلاصه‌هایی را تولید می‌کند. این سامانه، الگوریتم‌های مختلفی از خلاصه‌سازی از جمله الگوریتم‌های مبتنی بر موقعیت، مبتنی بر مرکز جرم، بزرگ‌ترین زیررشتهٔ مشترک و کلیدواژه‌ها را پیاده‌سازی کرده است.

در (Schilder, 2009) سامانهٔ FastSum نیز ابتدا یک مرحلهٔ پیش‌پردازش انجام می‌گیرد که نشانه‌گذاری و جداسازی جملات را به عهده دارد؛ سپس روال ساده‌سازی جملات انجام می‌شود تا اجزای بی‌اهمیت جمله را حذف کند. تمام جملاتی که در کلماتشان حداقل دو تطابق کامل و یا سه تطابق فازی با توصیف موضوع[3] نداشته باشند، حذف می‌شوند. چهار ویژگی که در این روش استفاده می‌شوند عبارتند از: فرکانس عنوان موضوع، فرکانس توصیف موضوع، فرکانس کلمهٔ محتوا و فرکانس سند. درنهایت خلاصهٔ نهایی از فهرست جملات وزن داده شده و بعد از مرحلهٔ حذف اطلاعات اضافی، حاصل می‌شود.

در این مقاله با در نظر گرفتن تمامی عوامل دخیل در خلاصه‌سازی و همچنین ارائهٔ پارامترهای جدید، سامانهٔ ایجاز برای خلاصه‌سازی تک‌سندی متون فارسی ارائه شده است که خلاصهٔ تولید شده را نسبت به سایر سامانه‌های خلاصه‌ساز موجود به‌طور قابل ملاحظه‌ای بهبود بخشیده است.

۳- روش پیشنهادی

در پیاده‌سازی سامانهٔ ایجاز از روال شکل (۱) برای خلاصه‌سازی متون استفاده می‌شود.



(شکل- ۱): روال به‌کار رفته در سامانهٔ ایجاز جهت خلاصه‌سازی متون زبان فارسی

در سامانهٔ طراحی‌شده برای خلاصه‌سازی تک‌سندی متون، عملیات زیر به‌ترتیب انجام می‌شود:

[1] Latent Semantic Analysis
[2] Semantic Role Labeling
[3] Topic Description

انتخاب متن ورودی؛

نرمال‌سازی متن: در اولین گام باید متون برای استفاده در گام‌های بعدی به شکلی استاندارد درآیند. این اصلاحات شامل یکسان‌سازی انواع نویسه‌ها نظیر «ی» عربی و «ی» فارسی و موارد مشابه است. همچنین اصلاحات دیگری نیز به‌منظور پردازش دقیق‌تر متون در این مرحله صورت می‌گیرد که خارج از بحث اصلی این مقاله است (Shamsfard, 2011).

با توجه به اینکه به‌طور عمومی میزان فشرده‌سازی بسیار بالا است و خلاصه‌های تولیدی باید بسیار فشرده باشند، بنابراین باید متون کم‌اهمیت‌تر حذف شوند. یکی از متونی که می‌تواند به‌عنوان متن کم‌اهمیت قلمداد شود، متنی است که به‌عنوان توضیح اضافه‌تر مطرح می‌شود. یکی از راه‌های ایجاد توضیح اضافه، استفاده از کلمات خاص مانند "مثلاً" است؛ اما راه دیگری نیز وجود دارد که در آن، این اطلاعات اضافه، درون پرانتز یا علامتی شبیه به آن قرار می‌گیرند. این علائم، علائمی نظیر {} , [] , () و ... هستند. از آنجایی که هدف نهایی، خلاصه‌سازی متن و رسیدن به جملات اصلی است، می‌توان متون اضافه‌ای که در میان این علامت‌ها جهت توضیحات اضافه گنجانده می‌شوند، تشخیص داده و حذف کرد. برای انجام این کار نیز از عبارات منظم استفاده شده است. همچنین با استفاده از درخت تجزیه و به‌دست آوردن کلمات توصیفی و عبارات توصیفی در مورد افراد یا اماکن و یا نهادهای اسمی می‌توان آن عبارات را نیز حذف کرد.

جداسازی جملات: در این مرحله با استفاده از علائم جداکنندهٔ جملات و درنظرگرفتن یک سری قواعد خاص، جمله‌های تشکیل‌دهندهٔ متن استخراج می‌شود. به‌عنوان مثال علامت نقطه همیشه پایان‌دهندهٔ یک جمله نیست. این علامت می‌تواند به‌عنوان ممیز در یک عدد اعشاری هم لحاظ شود. از این‌رو می‌توان با توجه به اجزای قبل و بعد آن دقت نوع کاربرد آن را تا حدی قاعده‌مند ساخت.

جداسازی کلمات: جداسازی کلمات با استفاده از علائم جداکننده و با درنظرگرفتن اصلاحات اعمال‌شده در مرحله نرمال‌سازی، انجام می‌شود.

حذف کلمات ایست: پس از جداسازی کلمات، در این مرحله، کلمات پرتکرار و بی‌اهمیت (ایست‌واژه‌ها) نظیر "که"، "است"، "در" و ... با استفاده از فهرست گردآوری شده، حذف می‌شوند (Tashakori, 2002).

ریشه یابی[1]: به‌منظور افزایش دقت پردازش‌های آماری، در این مرحله ریشه‌یابی کلمات انجام می‌شود.

اعمال عوامل مختلف تعیین اهمیت جملات: در این گام از پارامترهای مختلفی برای تعیین اهمیت جملات استفاده شده است. پس از محاسبهٔ مقادیر این پارامترها برای هر یک از جملات، ترکیب خطی از این معیارها به‌عنوان وزن جمله که بیان‌گر میزان اهمیت آن است، محاسبه می‌شود. در ادامه جزئیات این معیارها و نحوهٔ محاسبهٔ آنها توضیح داده شده است.

مرتب‌سازی جملات بر اساس وزن: پس از محاسبهٔ وزن جملات، جمله‌ها براساس وزنشان مرتب‌سازی می‌شوند.

انتخاب جملات براساس نرخ فشرده‌سازی: در گام نهایی جملات براساس نرخ فشرده‌سازی و به‌ترتیب اهمیتشان انتخاب شده و براساس موقعیتشان در متن اصلی، در خلاصه قرار داده می‌شوند. در این گام با استفاده از محاسبهٔ شباهت معنایی جملات، جمله‌های تکراری حذف می‌شوند.

۳-۱- اعمال فاکتورهای مختلف جهت تعیین اهمیت جملات

برای تعیین جمله‌های مهم، باید پارامترهای مختلفی را درنظرگرفت. در این مرحله، با کمک دانشجویان زبان‌شناسی و افرادی که در تولید پیکرهٔ خلاصه‌سازی حضور داشته‌اند و همچنین با بهره‌گیری از روش‌های آماری، پارامترهای مختلفی که می‌تواند در اهمیت جملات، نقش داشته باشد، استخراج شده است. ویژگی‌های مهم جملات در جدول (۱) آورده شده است.

پس از تعیین هر یک از این فاکتورها برای جملات، ترکیب خطی این وزن‌ها، به‌عنوان وزن نهایی هر جمله در نظر گرفته می‌شود و درنهایت جملات با بیشترین وزن، متناسب با نرخ فشرده‌سازی، انتخاب خواهند شد. برای محاسبهٔ وزن پارامترهای مختلف از روش‌های یادگیری ماشین می‌توان استفاده کرد. در این مقاله از روش کم‌ترین مربعات برای محاسبه وزن فاکتورها استفاده شده است.

۳-۱-۱- میزان شباهت با زمینه

هر متنی که با شیوهٔ نگارش مناسب، نوشته شده باشد، به‌حتم دارای زمینهٔ (Context) است. به‌طور عمومی در متون خبری تک‌سندی، تنها یک زمینه وجود دارد. برای انتخاب

[1] Stemming

جملات مهم، باید در ابتدا زمینهٔ اصلی متن را استخراج کنیم. زمینهٔ متن درحقیقت کلماتی هستند که با تکرارهای فراوان، اهمیت یک مفهوم خاص را در آن متن می‌رسانند. برای استخراج بردار زمینهٔ متن، کارهای زیر انجام می‌شود:

۱. استخراج کلمات موجود درجمله؛
۲. حذف کلمات ایست؛
۳. ریشه‌یابی کلمات؛
۴. محاسبهٔ فراوانی کلمات.

(جدول- ۱): ویژگی‌های استخراج‌شده در مورد جملات متن

نماد	توضیح ویژگی
x_1	میزان شباهت با زمینه
x_2	تأثیر کلمات ایست
x_3	تأثیر طول جمله
x_4	موقعیت جمله در متن
x_5	عبارات پر اهمیت
x_6	تأثیر ضمایر
x_7	عبارات اشاره و عبارات تعیین‌کننده درون جملات
x_8	عبارات خاص
x_9	میزان شباهت با عنوان متن

فرکانس کلمه w_j در جملهٔ i ام که آن را با $F(w_j^i)$ نشان می‌دهیم، برابر است با تقسیم تعداد تکرار ریشهٔ کلمه j ام از جملهٔ i ام متن. پس از محاسبهٔ بردار فرکانس کلمات که به‌عنوان بردار زمینهٔ متن درنظرگرفته می‌شود، شباهت بردار جملات با بردار زمینه محاسبه می‌شود. برای تعیین شباهت بین جملات با بردار زمینه از فرمول زیر استفاده می‌شود:

رابطه (۱)

$$x_1^i = \frac{\sum_j F(w_j^i)}{\max_k \{\sum_j F(w_j^k)\}}$$

با توجه به رابطهٔ (۱)، هر چه جمله‌ای شامل کلمات پرتکرار متن باشد، آنگاه شباهت آن با بردار زمینه بیشتر بوده و لذا اهمیت آن جمله بیشتر می‌شود.

۳-۱-۲- تأثیر کلمات ایست

گاهی اوقات تعداد کلمات یک جمله زیاد امّا کلمات کلیدی و مهم آن کم است. به‌طور عمومی چنین جملاتی دارای تعداد زیادی کلمات ایست هستند و به همین دلیل بار معنایی آنها کم است. هر چقدر جمله‌ای با استفاده از کلمات کمتری بتواند معنای بیشتری را منتقل نماید، اهمیت آن بیشتر خواهد بود. معیار زیر برای تعیین میزان اطلاع‌رسانی جملات در نظر گرفته شده است:

رابطه(۲)

$$x_2^i = \frac{|S_{NS}^i|}{|S^i|}$$

در این رابطه $|S_{NS}^i|$ تعداد کلمات غیر ایست جملهٔ i ام و $|S^i|$ هم طول جمله (برحسب تعداد کل کلمات) آن است.

۳-۱-۳- تاثیر طول جمله

به‌طور عمومی جملات با طول متوسط برای خلاصه‌سازی مناسب و جملاتی که دارای طول زیاد و یا طول خیلی کوتاه هستند، چندان مناسب خلاصه‌سازی نیستند. برای این‌که جملات خیلی طولانی و یا خیلی کوتاه در خلاصه انتخاب نشوند، می‌توان حدّ آستانه‌ای تعریف کرد و سپس جملاتی را که خارج از آن حدّ آستانه هستند، حذف کرد. البته در روش پیشنهادی، هیچ جمله‌ای براساس طول به‌صورت کامل حذف نمی‌شود. به‌همین دلیل معیار زیر برای مشخص‌کردن تأثیر طول جملات در نظر گرفته شده است:

رابطه (۳)

$$x_3^i = -RS^i.Log(RS^i) - (1-RS^i).Log(1-RS^i)$$

به‌طوری‌که

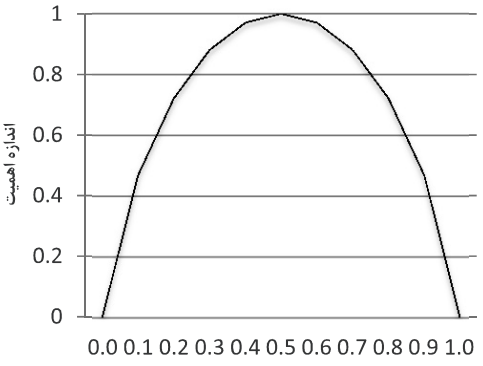
$$RS^i = \frac{|S^i|}{Max_{j=1:n}(|S^j|)}$$

در این معیار، RS^i اندازهٔ نسبی طول جمله بر حسب کلمه است. نمودار پارامتر x_3^i برای جملات مختلف در شکل (۲) نشان داده شده است. جملات دارای طول نسبی خیلی زیاد و یا خیلی کم ارزش کمتری در این پارامتر دارند.

۳-۱-۴- موقعیت جمله در متن

یکی از مهم‌ترین اصول حرفه‌ای نگارش متون خبری و علمی این است که جملات مهم در ابتدای متن قرار گیرند. برای متون انگلیسی این موضوع بارها و بارها آزموده و در سامانه‌های مختلف، نتایج آن بررسی شده است. با بررسی‌های انجام شده بر روی متون خبری فارسی این قاعده تا حدود زیادی تأیید می‌شود. البته به‌دلیل ضعف شیوهٔ انتقال خبر و عدم حرفه‌ای‌بودن اکثر خبرگزاری‌ها، متأسفانه

در بخش زیادی از متون خبری، شاهد آن هستیم که جملات انتهایی، نیز می‌توانند جزء مهم‌ترین جملات باشند. به همین دلیل برای متون خبری فارسی، فرمول زیر برای تعیین اهمیت موقعیت جمله در متن پیشنهاد شده است:



(شکل- ۲): نمودار ارزش‌گذاری جملات براساس معیار طول

رابطه (۴)

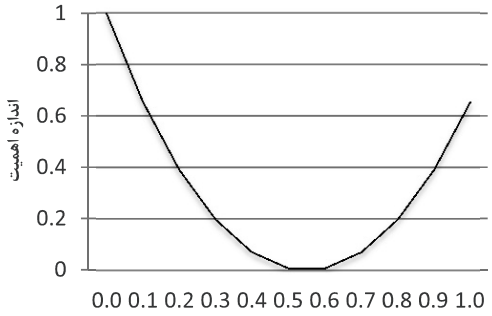
$$x_4^i = \frac{\hat{PS}^i}{Max_{j=1:n}(\hat{PS}^j)}$$

به‌طوری‌که

$$\hat{PS}^i = 1 + (0.5 * (PS^i + C) * \log(0.5 * (PS^i + C))) + ((1 - 0.5 * (PS^i + C)) * \log(1 - 0.5 * (PS^i + C)))$$

در این رابطه x_4^i موقعیت نسبی جمله در متن است. $\hat{PS}^i$ ارزش مؤثّر موقعیت نسبی جمله i اُم است که براساس ($0 \le PS^i \le 1$) یعنی موقعیت نسبی این جمله در متن و مبتنی بر مدل آنتروپی تعیین می‌شود. ضریب C در این رابطه به‌منظور ایجاد فاصلهٔ تمایز میان اهمیت جملات ابتدایی و انتهایی متن به آن اضافه شده است. در این مقاله مقدار C به‌طور شهودی و با توجه به نمونه‌های متن خبری که توسط تیم تولید پیکره بررسی شده، به‌دست آمده که برابر با 0.47 در نظر گرفته شده است. نمودار تأثیر این پارامتر و میزان تفاوت جملات ابتدایی و انتهایی در شکل (۳) نشان داده شده است.

بدین ترتیب در رابطهٔ به‌کار برده شده که مبتنی بر بررسی‌های آماری صورت گرفته بر روی پیکرهٔ پاسخ (Behmadi, 2013) است، جملات ابتدایی دارای اهمیت بیشتری حتی در مقایسه با جملات انتهایی هستند و جملات میانی در درجه اهمیت پایین‌تری قرار می‌گیرند.



(شکل- ۳): نمودار تعیین اهمیت جملات براساس موقعیت نسبی آنها

۳-۱-۵- عبارات پراهمیت

برخی عبارات خاص درون جمله هستند که اهمیت جمله را افزایش می‌دهند. این عبارات درواقع اطلاعات خاص و مهمی را در اختیار قرار می‌دهند و همین امر موجب می‌شود جملات حاوی آنها اهمیت ویژه‌ای داشته باشند. البته تعیین میزان اهمیت هر یک از عبارات خاص می‌تواند بسته به نوع متن صورت گیرد. در این مقاله دسته‌های زیر برای عبارات خاص توصیه می‌شود:

تاریخ

زمان

عدد

قیمت

وزن

کلمات انگلیسی

نقل قول

درصد

اسامی خاص (افراد، مکان و زمان)

برای جمع آوری اسامی افراد از پیکرهٔ استاندارد همشهری استفاده شده (AleAhmad, 2009) است. پس از بررسی اطلاعات به‌دست آمده، مشخص شد برخی از این اسامی در سایر نقش‌ها نیز به کرات به‌کار می‌روند، از این‌رو از فهرست نهایی حذف شدند. نمونه‌هایی از این اسامی عبارتند از «پیام»، «نیاز»، «عظیم»، «بهمن» و غیره.

یکی از مسائلی که به‌شدت در تعیین اهمیت جملات و درنهایت خلاصه‌سازی کمک می‌کند، مشخّص کردن اسامی مکان است. درواقع جملاتی که به مکان خاصی در قالب اسم مکان اشاره می‌کنند، می‌توانند اهمیت ویژه‌ای داشته باشند. می‌توان این اسامی را در گروه‌های:

اسامی کشورها
اسامی استان‌ها
اسامی شهرها
اسامی بخش‌ها
دسته بندی و استخراج کرد.

همان‌طور که ذکر شد اگر جمله‌ای دارای هر کدام از موارد فوق باشد براهمیت آن افزوده می‌شود.

به‌منظور تشخیص عبارات خاص در متن، ابتدا قاعدهٔ گرامری هر یک از دسته‌ها تعیین شده و برای هر کدام، جهت تطبیق یک عبارت باقاعده نوشته می‌شود.

پس از این‌که تعداد تکرار هریک از موارد بالا نسبت به طول جمله به‌دست آمد، ترکیب خطی آنها به‌عنوان وزن x_5^i در نظر گرفته شد.

رابطه (۵)

$$x_5^i = \frac{\left|S_{IMP}^i\right|}{\left|S^i\right|}$$

۳-۱-۶- تأثیر ضمایر

وجود ضمایر می‌تواند نشان‌دهندهٔ عدم اهمیت جمله باشد چون ضمایر به چیزهای دیگری اشاره دارند که در جملات قبل آمده‌اند. به عبارت دیگر، جمله‌ای که حاوی ضمیر، به‌ویژه در کلمات ابتدایی جمله باشد، حاوی توضیحات بیشتر برای جمله‌ای است که ضمیر به آن ارجاع می‌دهد. بنابراین نسبت به جمله‌ای که به آن ارجاع داده شده، اهمیت کمتری دارد. همچنین وجود ضمیر در یک جمله درصورتی‌که عملیات ادغام ضمایر صورت نگیرد و جملهٔ مرجع آن ضمیر هم در خلاصه انتخاب نشود، باعث می‌شود تا خوانایی خلاصه پایین بیاید. البته موقعیت ضمیر در جمله نیز اهمیت دارد. اگر ضمیری در کلمات ابتدایی جمله ظاهر شود، به‌طور تقریبی می‌توان گفت که آن جمله برای توصیف بیشتر جملات قبل است و به همین دلیل اهمیت کمتری دارد.

اگر ضمیر در کلمات میانی و انتهایی یک جمله ظاهر شود مشکل حالت قبل را ندارد؛ اما مشکلی دیگری که برای آن جمله پیش می‌آید، این است که خوانایی آن پایین می‌آید. به‌همین دلیل برای وجود ضمیر در جمله، ضریب منفی در نظر گرفته شده است. فرمول محاسبهٔ تأثیر ضمیر در جمله به‌صورت زیر در نظر گرفته شده است:

رابطه (۶)

$$x_6^i = -\frac{BP_3^i + \left|S_{PR}^i\right|}{1+\left|S^i\right|}$$

در رابطهٔ بالا $\left|S_{PR}^i\right|$ تعداد ضمایر موجود در جمله و BP_3^i در مواردی که ضمیر در سه کلمه اول ظاهر شود، برابر با یک و در بقیهٔ موارد صفر است. در رابطهٔ پیشنهادی، هر چه تعداد ضمایر بیشتر باشد وزن منفی جمله، بیشتر و اهمیت آن کمتر خواهد بود. همچنین درصورتی‌که ضمیر در سه کلمهٔ اول جمله رخ دهد، وزن منفی آن بیشتر خواهد بود. برای یافتن ضمایر در جمله بایستی براساس فهرستی از ضمایر، جستجویی درون جملات صورت گیرد. فهرست ضمایر فارسی در جدول (۲) نشان داده شده است.

(جدول- ۲): لیست ضمایر فارسی

ضمیر پیوسته	مثال	ضمایر گسسته	مثال
م	کتابم	من	کتاب من
ت	کتابت	تو	کتاب تو
ش	کتابش	او	کتاب او
مان	کتابمان	ما	کتاب ما
تان	کتابتان	شما	کتاب شما
شان	کتابشان	آن‌ها	کتاب آنها

برای تعیین ضمایر از عبارات باقاعده[1] جهت سرعت بیشتر در اجرای کار استفاده شده است. به این صورت که درون جملات به دنبال ضمایر در قالب عبارات باقاعده هستیم. تعداد و محل تطابق نیز مهم خواهد بود.

۳-۱-۷- عبارات اشاره و عبارات تعیین‌کنندهٔ درون جملات

تعیین عبارات اشاره نیز می‌تواند تأثیر به‌سزایی در افزایش دقت خلاصه‌سازی داشته باشد. عبارات اشاره درواقع عباراتی هستند که برای اشاره به شخصی یا موضوعی خاص بیان می‌شوند. برای استخراج این عبارات نیز می‌توان از عبارات باقاعده استفاده کرد. به‌عنوان مثال به‌طور معمول قبل از نام افراد، کلماتی مانند «آقای»، «خانم»، «دکتر»، «حجت‌الاسلام»، «سرهنگ»، «سروان»، «ستوان»، «سرلشگر»، «سردار»، «تیمسار»، «بانو» و از این قبیل موارد می‌آید. برخی عبارات اشاره نیز وجود دارند که وجود آنها درون جمله بر اهمیت جمله می‌افزاید. این قبیل عبارات می‌توانند مانند «در نتیجه»، «بنابراین»، «در مجموع» و «در پایان» باشند.

[1] Regular Expression

همچنین جملات حاوی کلمات موجود در عنوان اصلی مقاله یا خبر و کلمات کلیدی مشخص‌شده توسط کاربر، امتیاز بیشتری کسب می‌کنند. بدیهی است درصورتی‌که درون جمله، کلمهٔ کلیدی وجود داشته باشد، آن جمله، مطلب مهمی را بیان می‌کند و بایستی نمرهٔ بالایی از لحاظ اهمیت به آن جمله، نسبت داده شود.

البته برخی عبارات نیز درون جمله وجود دارند که باعث کاهش اهمیت جمله می‌شوند. از جملهٔ این موارد می‌توان به «مثلاً»، «مانند»، «همانند» و«همچون» اشاره کرد. از نوع کلمات استفاده‌شده، معلوم است که جملات حاوی این عبارات، درصدد بیان توضیح بیشتری در مورد مطالب قبلی بیان شده هستند. درنتیجه در هنگام خلاصه‌سازی، گزینهٔ مناسبی برای حذف از خلاصهٔ نهایی هستند.

وزن مربوط به عبارات اسامی اشاره مثبت و منفی به‌صورت زیر در نظر گرفته شد:

رابطه (۷)
$$x_7^i = \frac{|S^i| + |S_{PP}^i| - |S_{NP}^i|}{2.|S^i|}$$

که $|S_{NP}^i|$ برابر است با تعداد کلمات اشارهٔ مثبت در جمله iأم و $|S_{pp}^i|$ برابر است با تعداد کلمات اشاره منفی در جمله iأم است.

۳-۱-۸- عبارات خاص

عبارات خاص عباراتی هستند که با علائمی از قبیل گیومه و یا مانند آن از بقیه متن مجزا شده‌اند. بررسی متون خلاصه انسانی نشان می‌دهد متن‌های موجود در این بخش‌ها مورد توجه خلاصه‌گرهای انسانی قرار گرفته و لذا در سامانهٔ پیشنهادی مؤلّفه‌ای با همین عنوان برای تعیین میزان اهمیت جملات حاوی آنها نیز در نظر گرفته شده است.

رابطه (۸)
$$x_8^i = \frac{|S_{SPC}^i|}{|S^i|}$$

که $|S_{SPC}^i|$ برابر با تعداد عبارات خاص به‌کار رفته در داخل علائم ذیل است:

«»،<< >>،“”

$|S^i|$ هم طول جمله بر اساس تعداد کلمه است.

۳-۱-۹- میزان شباهت با عنوان متن

یکی از روش‌های قدیمی و ابتدایی در خلاصه‌سازی متون، روش مبتنی بر عنوان است (Edmundson, 1969). در این روش، شباهت جمله با عنوان محاسبه و سپس جملاتی که بیشترین شباهت با عنوان را داشته باشند، به‌عنوان جملات خلاصه برگردانده می‌شوند. در روش پیشنهادی در سامانهٔ طراحی‌شده هم این معیار تعبیه شده است و درصورتی‌که کاربر عنوان خبر را وارد کند، خلاصهٔ متن با توجه به‌عنوان تولید خواهد شد.

شباهت هر جمله با عنوان خبر با استفاده از رابطهٔ زیر محاسبه می‌شود:

رابطه (۹)
$$x_9^i = \frac{|S^i \cap T|}{\sqrt{|S^i| . |T|}}$$

در رابطهٔ بالا $|S^i \cap T|$ تعداد کلمات مشترک بین جمله S^i و عنوان بوده و مخرج کسر هم جذر حاصل‌ضرب طول جمله و عنوان (برحسب کلمه) است.

۳-۲- ترکیب فاکتورهای مختلف برای تعیین جملات مهم و تولید خلاصه

پس از محاسبهٔ ویژگی‌های مختلف که در بخش‌های پیشین به آنها پرداخته شد، ترکیب خطی این پارامترها به‌عنوان ارزش نهایی جملات در نظر گرفته می‌شود.

رابطه (۱۰)
$$Y = W_0 + \sum_{i=1}^{9} W_i X_i$$

برای محاسبهٔ وزن پارامترهای مختلف از روش‌های یادگیری ماشین می‌توان استفاده کرد. یکی از این روش‌های پرکاربرد، بهره‌گیری از روش کمترین مربعات (Strutz, 2010) می‌باشد.

برای این منظور ابتدا پانزده سند خبری برای عمل خلاصه‌سازی انتخاب و به کمک پنج نفر از افراد خبره خلاصه‌های مرتبط تولید شد. پس از آنکه برای هر سند پنج خلاصهٔ مجزا فراهم شد، میانگین نظرات داوران در مورد حضور (۱) یا عدم حضور (۰) هر جمله از متن اصلی در متن خلاصه به عنوان *برچسب عددی جمله* مربوطه تعیین شد.

مثال ۱: اگر بنابر جدول (۳)، متن اصلی یک سند شامل جملات { S_1, S_2, S_3 } فرض شود و { E_1, E_2, E_3, E_4, E_5} نشان‌دهندهٔ مجموعه افراد خبره باشد؛ آنگاه تلاقی سطر i ام و ستون j ام، بیان‌گر انتخاب (۱) یا عدم انتخاب (۰) جمله i ام توسط فرد خبره j ام است. میانگین داده‌های سطر i اُم تعیین‌کنندهٔ برچسب جمله i ام است. این برچسب می‌تواند به‌عنوان میزان مقبولیت جملهٔ مربوطه جهت انتخاب‌شدن برای قرارگرفتن در متن خلاصه تلقی شود. اگر مقدار برچسب برای یک جمله برابر با یک باشد به این معناست که این جمله توسط تمامی افراد خبره به‌عنوان جملهٔ مهم، تشخیص داده شده و حائز بالاترین درجهٔ اهمیت است. همچنین اگر مقدار برچسب جمله‌ای برابر با صفر باشد، نشان‌دهندهٔ آن است که جملهٔ مذکور توسط هیچ یک از افراد خبره به‌عنوان جملهٔ مهم متن، تشخیص داده نشده و لذا دارای پایین‌ترین درجه اهمیت است.

پس از محاسبهٔ بردار مؤلّفه‌های هر یک از جملات متن اصلی (به‌صورتی که در بخش ۳-۱ بیان شد) و اختصاص برچسب هر جمله (به‌نحوی‌که در قسمت قبل بیان شد)، حال می‌توان با استفاده از روش LMS مقادیر مربوط به وزن هر مؤلّفه (W_i) را بر اساس رابطه ۱۱ و ۱۲ محاسبه کرد.

$$Y_{n\times 1} = X_{n\times p}.W_{p\times 1} \qquad \text{رابطه (۱۱)}$$

$$W = [X^T X]^{-1}[X^T Y] \qquad \text{رابطه (۱۲)}$$

در روابط ۱۱ و ۱۲، ماتریس X شامل بردارهای ویژگی کلیه جملات (با p مؤلّفه) موجود در کل اسناد پیکرهٔ آموزشی است. همچنین برچسب هر یک از جملات مربوطه در بردار Y قرار داده شده است. با توجه به این نکته که میزان اهمیت هر جمله (y_i) را می‌توان به‌صورت ترکیب خطی مؤلّفه‌های $x_j \quad 1 \le j \le p$ در نظر گرفت، مقدار ضرایب وزنی $w_j \quad 0 \le j \le p$ توسط رابطهٔ ۱۲ قابل محاسبه است.

در جدول (۲) مقادیر W_i های به‌دست آمده مربوط به پیکرهٔ آموزشی درج شده است.

برای خلاصه‌سازی سند ورودی جدید Dk با m جمله، ابتدا بردار ویژگی (X) هر یک از جملات، تعیین و براساس رابطه ۱۱ میزان اهمیت هر جمله (Y) مربوطه مشخص می‌شود. در ادامه بردار Y به‌صورت نزولی مرتب می‌شود. بنابراین y1 بیان‌گر مهم‌ترین جمله و ym بیان‌گر کم اهمیت‌ترین جمله است؛ سپس با توجه به نرخ فشرده‌سازی، جملات با بیشترین وزن انتخاب می‌شوند.

(جدول- ۲): مقدار تخصیص داده شده به پارامترها

وزن	توضیح ویژگی	مقدار
w_0	مقدار فاصله از مبدأ	۹۱
w_1	میزان شباهت با زمینه	۱۰
w_2	تأثیر کلمات ایست	۸
w_3	تأثیر طول جمله	-۱
w_4	موقعیت جمله در متن	۶
x_5	عبارات پر اهمیت	-۱۰
x_6	تأثیر ضمایر	۲۹
x_7	عبارات اشاره و عبارات تعیین‌کننده درون جملات	-۱۴۲
x_8	عبارات خاص	۲۰۸
x_9	میزان شباهت با عنوان متن	۲۴

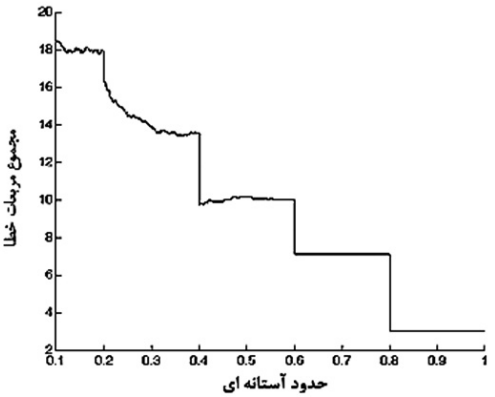
(جدول- ۳): نمونه نظرات داوران برای سه جمله S_i از متن

	E1	E2	E3	E4	E5	میانگین آراء
S1	1	1	0	1	1	**0.8**
S2	0	1	1	0	0	**0.4**
S3	1	0	1	1	1	**0.8**

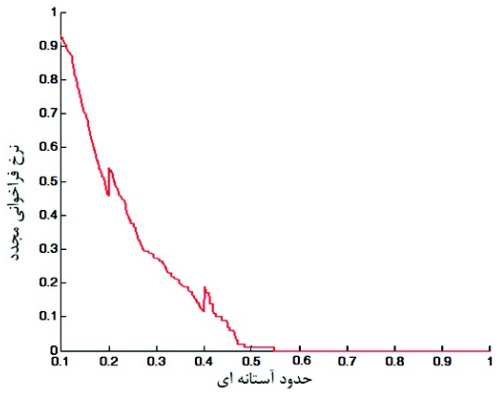
برای اینکه در خلاصهٔ تولیدشده، جملات تکراری وجود نداشته باشد، هر جملهٔ نامزد، با جملات انتخاب‌شدهٔ قبلی مقایسه می‌شود و فقط درصورتی‌که شباهت آن با جملات قبلی از حد آستانه α کمتر باشد (در این سامانه α برابر با ۰/۵ در نظر گرفته شده است)، به‌عنوان جملهٔ خلاصه، در نظر گرفته شده و به فهرست جملات انتخاب‌شده، افزوده می‌شود. برای تعیین شباهت جملات با یکدیگر هم معیار زیر در نظر گرفته شده است:

$$Sim(i,j) = \frac{|S^i \cap S^j|}{\sqrt{|S^i| . |S^j|}} \qquad \text{رابطه (۱۳)}$$

که $|S^i|$ و $|S^j|$ به ترتیب طول جملات i ام و j اُم است. در این سامانه α برابر با در نظر گرفته شده است.



(شکل- ۴): رابطهٔ خطا در گزینش جملات با آستانه‌های حدی متفاوت

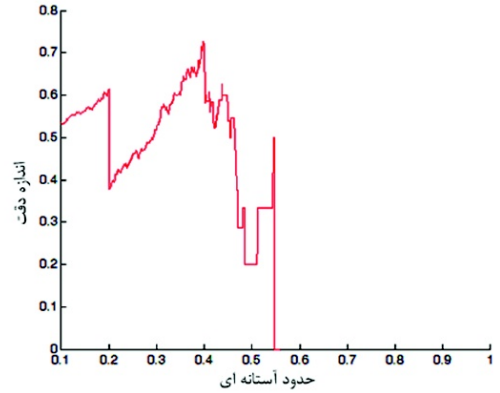


(شکل- ۵): رابطهٔ نرخ فراخوانی مجدد با آستانه‌های حدّی مختلف

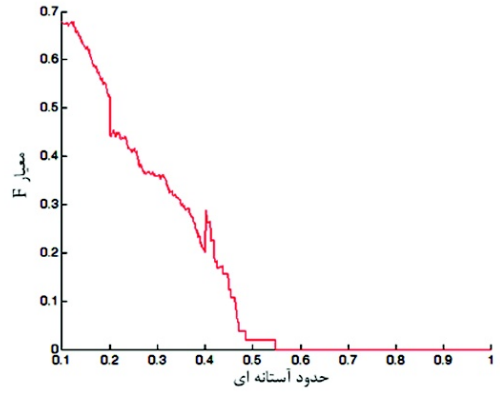
جهت ارزیابی دقت وزن‌های محاسبه‌شده با استفاده از روش کمترین مجذور مربعات، مجدداً برچسب تمامی جملات پیکره به کمک رابطهٔ خطی ارائه‌شده محاسبه شد. آنگاه برای محاسبهٔ خطای اندازه‌گیری این روش از یک حد آستانه β برای تعیین انتخاب یا عدم انتخاب یک جمله در خلاصه نهایی استفاده شد. به‌عنوان نمونه اگر $\beta = 0.5$ باشد همه جملاتی در خلاصهٔ نهایی منتخب افراد خبره قرار می‌گیرند که هم توانسته باشند در حداقل نیمی از خلاصه‌های تولیدشده توسط این افراد حضور یابند و هم برچسب به‌دست آمده از رابطهٔ خطی مذکور از این حد آستانه‌ای بزرگتر باشد. در غیر این‌صورت در تشخیص جمله مناسب خطا رخ داده است. با این توصیف کیفیت روش محاسبهٔ وزن‌های به‌دست آمده برای حدود آستانه‌ای مختلف مورد بررسی قرار گرفت.

در شکل (۴) نمودار خطا به‌ازای حدود آستانه‌ای متمایز نشان داده شده است. برای به‌دست آوردن حد آستانه‌ای مناسب به غیر از توجه به مقدار خطا، نرخ دقت و نرخ فراخوانی مجدد نیز مورد ملاحظه قرار گرفته است.

با فرض بازیابی ۳۰٪ از جملات یک متن در فرآیند خلاصه‌سازی، همان‌گونه که در نمودارهای ۵، ۶ و ۷ نیز دیده می‌شود با انتخاب حد آستانه‌ای برابر با ۰/۴ می‌توان به کم‌ترین خطای ممکن دست پیدا کرد. در این مجموع، مجذور مربعات خطا برای 782 جمله، حدود ۹/۵ به‌دست آمده است.



(شکل- ۶): رابطهٔ اندازهٔ دقت با آستانه‌های حدّی مختلف



(شکل- ۷): رابطهٔ اندازهٔ معیار F با آستانه‌های حدی مختلف

بدیهی است هر چه میزان حد آستانه‌ای بیشتر باشد، تعداد جملات قابل استفاده در مجموعهٔ خلاصه کمتر می‌شود.

۴- سامانهٔ ایجاز

سامانهٔ ایجاز در حال حاضر در مورد متون خبری با دقت بالا قابل استفاده است و می‌توان با بررسی انواع متون علمی، ادبی و غیره و در نظر گرفتن عوامل دخیل در خلاصه‌سازی هر کدام از انواع متون، سامانهٔ تولیدشده را گسترش داده و دقت آن را به‌ازای انواع متون بهبود بخشید. در تولید این سامانه، با بهره‌گیری از یک تیم خبرهٔ زبان‌شناسی، ویژگی‌های مهمی که در تولید خلاصه می‌توانند مؤثّر باشند جمع‌آوری شد و در ادامه با استفاده از روش‌های یادگیری ماشین، ترکیب خطی‌ای از این پارامترها به‌عنوان وزن مؤثّر جملات به‌کار گرفته شد.

این سامانه در دو نسخهٔ تحت وب و تحت ویندوز تولید شد. در شکل (۸) تصویری از نسخهٔ تحت وب را مشاهده می‌کنید که از طریق نشانی ijaz.um.ac.ir قابل استفاده است.

(شکل- ۸): تصویری از نسخهٔ تحت وب ایجاز

سامانهٔ تحت وب ایجاز قابلیت دریافت متن به سه صورت: الف) فایل متنی ب) متن ج) نشانی اینترنتی صفحه وب خبر را دارد و می‌توان نرخ فشرده‌سازی را نیز به دلخواه تولید کرد.

۵- ارزیابی

برای یک ارزیابی مناسب و دقیق، احتیاج به یک مجموعه دادهٔ مناسب و استاندارد است. به همین منظور، از پیکرهٔ استاندارد پاسخ[1] با تعداد یکصد سند که با رعایت استانداردهای لازم تولید شده‌اند، استفاده شد. به‌علاوه با توجه به اینکه تنظیم وزن‌ها براساس متون خبری صورت گرفته، لذا این پیکره بهترین گزینه برای ارزیابی خلاصه‌ساز پیشنهادی است. مشخصات این پیکره در جدول (۴) ذکر شده است.

(جدول- ۴): مشخصات پیکرهٔ پاسخ

تعداد اسناد پیکره	تعداد موضوعات	تعداد خبرگزاری‌های انتخاب شده	تعداد خلاصه‌ها به ازای هر سند
۱۰۰	۶	۷	۵

جهت ارزیابی خلاصه‌های تولیدشده نیاز به ابزاری جهت تعیین میزان شباهت خلاصه‌های تولید شده توسط انسان (خلاصه‌های هدف) و خلاصه‌های تولیدشده با ماشین است. به همین منظور برای ارزیابی کیفیت خلاصه‌های تولیدشده، از ابزار ارزیابی خودکار خلاصه‌سازهای ماشینی[2](استیری، ۱۳۹۱) استفاده شده است. این ابزار با استفاده از معیارهای شباهت مبتنی‌بر هم‌رخدادی، میزان شباهت بین خلاصه‌های انسانی و خلاصه‌های تولیدشده به‌وسیلهٔ ماشین را مشخص می‌کند.

تنها سامانهٔ خلاصه‌ساز فارسی در دسترس (قابل دانلود یا روی خط) سامانهٔ خلاصه‌ساز FarsiSum است. به‌همین دلیل از این سامانه برای مقایسه با سامانهٔ تولیدشده، استفاده شده است. برای تولید خلاصه توسط سامانهٔ FarsiSum، متون موجود در پیکرهٔ ایجادشده به FarsiSum داده شده و خلاصه‌های برگردانده‌شده در فایل ذخیره شدند. یکصد سند برای ارزیابی خلاصه‌سازی انتخاب شد. نرخ فشرده‌سازی هم سی‌درصد در نظر گرفته شد. از یکصد سند ورودی، سامانهٔ FarsiSum تنها برای ۶۸ عدد از آنها، خلاصه تولید کرد. برای ۲۷ مورد هیچ گونه خلاصه‌ای برگردانده نشد. یک مورد، کل متن برگردانده شد و درچهار مورد هم سامانه با خطای نرم‌افزاری مواجه شده و خروجی مشخصی نداشت.

بنابراین با توجه به مشکلاتی که سامانهٔ FarsiSum برای تولید خلاصه داشت، درنهایت ارزیابی برروی ۶۸ سند خبری انجام شد.

پس از اجرای دو نرم‌افزار به‌ازای ۶۸ سند موجود، خلاصه‌های تولیدشده، با استفاده از ابزار ارزیابی خلاصه‌سازهای فارسی با خلاصه‌های انسانی مقایسه شدند که نتایج آنها به تفکیک معیار به‌کار رفته در ذیل آورده شده

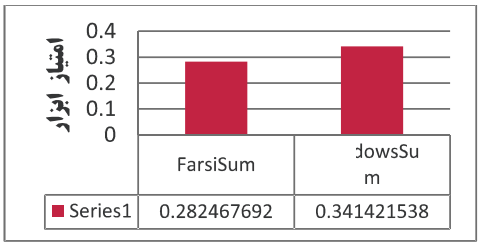
[1] این پیکره در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد تولید شده است.

[2] این ابزار در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد طراحی و پیاده‌سازی شده است.

است. نتایج سامانهٔ طراحی‌شده در نمودار با برچسب Ijaz نمایش داده شده است.

۵-۱- ارزیابی با بررسی n گرم‌های مشابه در کل متن

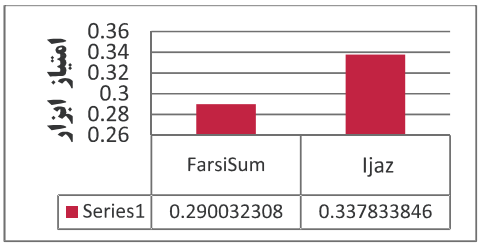
این معیار، روشی است که مبتنی بر فراخوانی n تایی‌ها بین یک خلاصهٔ سیستمی و مجموعه‌ای از خلاصه‌های انسانی است. در این معیار تعداد n تایی‌های مشترک بین خلاصه‌های انسانی و خلاصه ماشینی بر کل تعداد n تایی‌های موجود در خلاصه انسانی تقسیم می‌شود.



(شکل- ۹): مقایسهٔ میانگین n گرم‌های مشابه در کل متن

۵-۲- ارزیابی با بررسی n گرم‌های مشابه در جملات

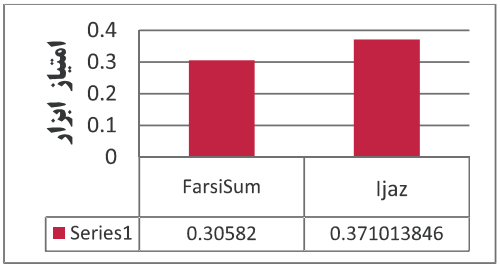
در این معیارn تایی‌هایی که مورد ارزیابی قرار می‌گیرند، فقط از لغات موجود در یک جمله و نه تمام متن ساخته می‌شوند.



(شکل- ۱۰): مقایسهٔ میانگین n گرم‌های مشابه در جملات

۵-۳- ارزیابی با بررسی ویژه چندکلمه‌ای‌های مشابه در کل متن

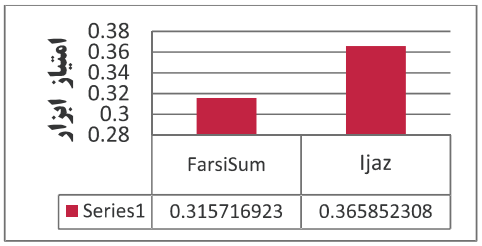
در این معیار به جای در نظر گرفتن تنها nتایی‌های مشترک بین دو متن، تمام iتایی‌های مشترک بین دو متن $(i < n)$ تشکیل و میزان انطباق آن‌ها مورد ارزیابی قرار می‌گیرد.



(شکل- ۱۱): مقایسهٔ میانگین بررسی ویژه n گرم‌های مشابه درکل متن

۵-۴- ارزیابی با بررسی ویژه چندکلمه‌ای‌های مشابه در جملات

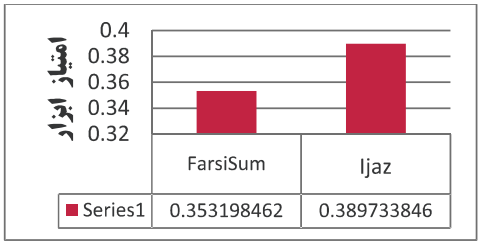
در این معیار، تمام iتایی‌های مشترک بین دو متن $(i < n)$ در جملات، تشکیل و میزان انطباق آنها مورد ارزیابی قرار می‌گیرد. این معیار بیشترین شباهت به عملکرد امتیازدهی انسانی به خلاصه‌ها را دارد.



(شکل- ۱۲): ارزیابی میانگین با بررسی ویژه n گرم‌های مشابه در جملات

۵-۵- ارزیابی با بررسی طولانی‌ترین زیر رشته مشترک

در این معیار ارزیابی، از الگوریتم‌های محاسبه طولانی‌ترین زیررشتهٔ مشترک بین دو رشته[1] استفاده می‌شود.

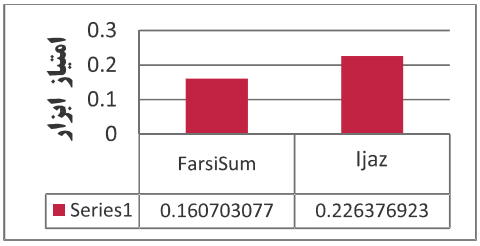


(شکل- ۱۳): ارزیابی میانگین با بررسی طولانی‌ترین زیررشتهٔ مشترک

[1] LCS (Longest Common Subsequence)

۵-۶- ارزیابی با بررسی ۲کلمه‌ای‌های مشابه با فاصله آزاد

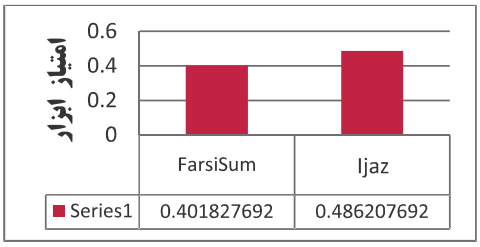
به هر جفت کلمه (با حفظ ترتیب) در جمله، Skip-bigram گفته می‌شود. این معیار با اندازه‌گیری تعداد Skip-bigramهای مشترک بین خلاصهٔ ماشینی و خلاصه‌های مرجع محاسبه می‌شود.



(شکل- ۱۴): ارزیابی با بررسی ۲گرم‌های مشابه با فاصله آزاد

۵-۷- ارزیابی با بررسی تعداد لغات مشابه دو متن

به‌عنوان ساده‌ترین معیار شباهت دو متن می‌توان تعداد واژگان مشترک بین آنها را در نظر گرفت. در این معیار تعداد واژگان مشترک بین خلاصه ماشینی با هریک از خلاصه‌های انسانی محاسبه و بر تعداد کل واژگان در خلاصه‌های انسانی تقسیم می‌شود.



(شکل- ۱۵): ارزیابی با بررسی تعداد لغات مشابه دو متن

همان‌طور که مشخص است در تمامی معیارهای به‌کار رفته در ابزار ارزیابی خودکار خلاصه‌سازها، سامانهٔ خلاصه‌ساز طراحی‌شدهٔ ایجاز برتری قابل ملاحظه‌ای نسبت به ابزار FarsiSum دارد. با توجه به اینکه پرداختن به جزییات و ویژگی این معیارها خارج از حوزه این مقاله است، از ذکر توضیحات بیشتر در ارتباط با جنبه‌های کیفی این معیارها خودداری شده است؛ اما تفاوت قابل توجه سامانهٔ ایجاز و FarsiSum در تمامی این معیارها، نشان‌گر عملکرد مناسب این سامانه است.

۶- نتیجه‌گیری و کارهای آینده

امروزه با توجه به حجم گستردهٔ مطالب موجود و کمبود وقت و از طرفی ناکارآمدی سیستم‌های خلاصه‌سازی موجود، وجود یک سامانهٔ قدرتمند برای خلاصه‌سازی حجم انبوه کتب، مقالات و اخبار در سطح وب به‌شدت احساس می‌شود. چنان‌که شاهد هستیم علی‌رغم اینکه بحث خلاصه‌سازی از سال ۱۹۶۰ مطرح شده است، امّا همچنان ضعف‌های زیادی در این زمینه وجود دارد و کارهای بسیاری برای رسیدن به وضعیتی مطلوب در این زمینه بایستی انجام پذیرد.

مشکل خلاصه‌سازی در زبان فارسی به‌مراتب بیشتر از زبان‌های دیگر است. پیچیدگی‌های زبانی موجود و همچنین عدم وجود ابزارهای با دقت مناسب برای کار با زبان فارسی از جمله مشکلات فعلی پردازش متون فارسی است. به عبارت دیگر با استفاده از ابزارهای پردازش متن دقیق‌تر، برای زبان فارسی می‌توان دقت نهایی مقادیر تولیدشده برای هر یک از ویژگی‌های استخراج‌شده دراین مقاله را افزایش داد. از طرف دیگر توجه به حوزهٔ معنایی و استفاده از رابطه‌های معنایی میان جملات و مفاهیم موجود در سند از دیگر پیشنهادهایی است که برای کارهای آتی می‌تواند مطرح شود.

سپاس‌گزاری

در این قسمت، لازم می‌دانیم از زحمات جناب آقای دکتر جهانگیری، دانشجویان گروه زبان‌شناسی و اعضای آزمایشگاه فناوری وب دانشگاه فردوسی مشهد به‌ویژه جناب مهندس رضا سعیدی، کمال سپاس‌گزاری را داشته باشیم.

مراجع

اخوانت.، شمس فرد م.، عرفانی جورابچی م.، "PARSUMIST: خلاصه‌ساز تک‌سندی و چندسندی متون فارسی"، چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر، ایران، تهران، ۱۳۸۷.

استیری، احمد. کاهانی، محسن.، کیوانلو شهرستانکی، زهره.، پورمعصومی، آصف.، "ارائه یک ابزار ارزیابی خودکار خلاصه‌سازهای ماشینی فارسی"، چهارمین کنفرانس فناوری اطلاعات و دانش، دانشگاه صنعتی نوشیروانی بابل، ۱۳۹۱.

ستوده حمیدرضا.، اکبرزاده توتونچی، محمدرضا.، تشنه لب، محمد .، "خلاصه‌سازی متن بر اساس گزینش با استفاده از رویکرد انسان‌شناختی"، مجموعه مقالات هجدهمین کنفرانس مهندسی برق ایران، دانشگاه صنعتی اصفهان، اصفهان، ایران، ۱۳۸۹.

کریمی، زهره.، شمس فرد، مهرنوش.، "سیستم خلاصه‌سازی خودکار متون فارسی"، دوازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران- تهران، ۱۳۸۵، صفحه ۱۲۷۶ .

مشکی، محسن.، "بررسی روش‌های خلاصه‌سازی متون غیرساخت یافته‌ی فارسی"، سمینار کارشناسی ارشد، دانشکده ی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، ۱۳۸۶.

مشکی، محسن.، "خلاصه‌سازی چندسندی اخبار فارسی برای تولید خلاصه‌های پیشینه-خبر"، هفدهمین کنفرانس مهندسی برق ایران، صفحه ۳۰۴، تهران ۱۳۸۸.

AleAhmad ,A., and et al, "Hamshahri: A standard Persian text collection", Journal of Knowledge-Based Systems, July 2009, Vol. 22 No.5, p.382-387.

Behmadi ,B. , Kahani ,M. , Toosi ,SA., Poormasooumi ,A., Estiri ,A., "Pasokh: A standard corpus for the evaluation of Persian text summarizers," in 3th International eConference of Computer and Knowledge Engineering (ICCKE), 2013, pp. 471-475.

Dalianis, H., "SweSum - A TextSummarizer for Swedish, Technical report", TRITANA-P0015, IPLab-174, NADA, KTH, 2000.

Edmundson ,H.P., "New Methods in Automatic Extracting.", Journal of the Association for Computing Machinery, April 1969, 16(2):p264-285 .

Farzindar, A., Rozon, F., and Lapalme, G.,"CATS a topic-oriented multi-document summarization system", at DUC 2005.

Honarpisheh, M.,Ghassem-Sani, Gh.,Mirroshandel, Gh., "A Multi-Document Multi-Lingual Automatic Summarization System", Proceedings of the 3rd International Joint Conference on natural language processing (IJCNLP), pp. 733-738, 2007.

Kiyoumarsi, F., Rahimi Esfahani F., "Optimizing Persian Text Summarization Based on Fuzzy Logic Approach" International Conference on Intelligent Building and Management Proc of CSIT, 2011, vol.5 ,IACSIT Press, Singapore.

Mani, I., Maybury, M., "Advances in Automatic Text Summarization", The MIT Press, 1999.

Mazdak, N., Hassel, M., "FarsiSum-a persiantext summarizer", Master thesis, Departmentof linguistics, Stockholm University, 2004.

McKeown, K., Barzilay, R., Vasileios Hatzivassiloglou, D,. "Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster", 2001.

Mihalcea, R. and Tarau, P. "An Algorithm for Language Independent Single and Multiple Document Summarization". In Proceedings of the International Joint Conference on Natural Language Processing, Korea, 2005.

Poormasoomi, A., Kahani, M., Varasteh, S., Kamyar, H., "Context-based Persian multi-document summarization (global view)", IALP, Malaysia, 2011.

Radev, D., Allison, T. et all, "MEAD - A platform for multidocument multilingual text summarization.", In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 2004.

Shamsfard, M., "Challenges and open problems in Persian text processing," in 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics, 2011, pp. 65-69.

Strutz ,T., "Data Fitting and Uncertainty," A practical introduction to weighted least squares and beyond. Vieweg+ Teubner,2010.

Svore, K., Vanderwende, L. and Burges, C., "Enhancing single-document summarization by combining Rank Net and third-party sources". In Proceedings of the EMNLP-CoNLL, 2007.

Tashakori, M., and et al., "Bon: The persian stemmer," in EurAsia-ICT 2002: Information and Communication Technology, ed: Springer, 2002, pp. 487-494.

آصف پورمعصومی مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (گرایش نرم‌افزار) به‌ترتیب در سال‌های ۱۳۸۷ و ۱۳۹۰ از دانشگاه فردوسی مشهد دریافت کرده است. زمینهٔ تحقیقاتی مورد علاقهٔ ایشان شامل ابزارهای پردازش زبان طبیعی، کاوش فرآیند، سیستم‌های تصمیم‌یار است.

نشانی رایانامهٔ ایشان عبارت است از:

Asef.pms@gmail.com

محسن کاهانی استاد گروه مهندسی کامپیوتر دانشگاه فردوسی مشهد و مدیر آزمایشگاه فناوری وب است. ایشان دکترای خود را در رشتهٔ مهندسی کامپیوتر از دانشگاه ولونگونگ استرالیا در سال ۱۳۷۷ اخذ کرده است. زمینهٔ تحقیقاتی مورد علاقه ایشان شامل وب‌معنایی، پردازش زبان طبیعی، سیستم‌های تصمیم‌یار و مهندسی نرم‌افزار است.
نشانی رایانامهٔ ایشان عبارت است از:
kahani@um.ac.ir

سیّداحمد طوسی مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (گرایش هوش مصنوعی) به‌ترتیب در سال‌های ۱۳۸۱ و ۱۳۹۱ از دانشگاه فردوسی مشهد دریافت کرده است.
زمینهٔ تحقیقاتی مورد علاقه ایشان شامل ابزارهای پردازش زبان طبیعی، سیستم‌های چندعامله، سیستم‌های تصمیم‌یار است.
نشانی رایانامهٔ ایشان عبارت است از:
ahmad.toosi@alumni.um.ac.ir

احمد استیری مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (گرایش نرم‌افزار) به‌ترتیب در سال‌های ۱۳۸۷ و ۱۳۹۱ از دانشگاه فردوسی مشهد دریافت کرده است. زمینهٔ تحقیقاتی مورد علاقهٔ ایشان شامل ابزارهای پردازش زبان طبیعی، سیستم‌های چندعامله، سیستم‌های تصمیم‌یار است.
نشانی رایانامهٔ ایشان عبارت است از:
ahmad.estiry66@gmail.com

هادی قائمی در حال تحصیل در مقطع کارشناسی ارشد در دانشگاه فردوسی مشهد است. زمینهٔ تحقیقاتی مورد علاقهٔ ایشان شامل ابزارهای پردازش زبان طبیعی، پردازش الگو، است.
نشانی رایانامهٔ ایشان عبارت است از:
Hadi.qaemi@stu.um.ac.ir

پیوست

x_k^i	ویژگی kام برای جمله iام متن
$\mid S^i \mid$	طول جمله بر اساس تعداد کلمات (ایست واژه + غیر ایست واژه)
$\mid S_{PP}^i \mid$	طول جمله براساس تعداد اشارات مثبت
$\mid S_{NP}^i \mid$	طول جمله براساس تعداد اشارات منفی
$\mid S_{IMP}^i \mid$	طول جمله براساس تعداد واژه‌های بااهمیت
$\mid S_{LOC}^i \mid$	طول جمله براساس تعداد اسامی خاص مکان
$\mid S_{Time}^i \mid$	طول جمله براساس تعداد اسامی خاص زمان
$\mid S_{SPC}^i \mid$	طول جمله براساس تعداد زیر متن خاص قرار گرفته در میان جفت علائم ویژه
$F(W_j^i)$	بسامد کلمه jام جمله iام در کل متن
RS^i	طول نسبی جمله iام در متن
PS^i	شماره ترتیب جمله i ام در متن
$\hat{PS}^i$	ارزش مطلق موقعیت ترتیبی جمله iام در متن
$\mid S_{PR}{}^i \mid$	طول جمله براساس تعداد ضمایر
BP_3^i	اگر در سه کلمه اول جمله i ام ضمیر وجود داشته باشد برابر با یک و در غیر اینصورت برابر با صفر است