دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

خلاصه‌سازی خبر با استفاده از تکنیک‌های هوش مصنوعی

دانشجو : نداعظیمی

استاد راهنما:

دکتر مرتضی زاهدی

استاد مشاور:

مهندس مرضیه رحیمی

پایان‌نامه ارشد جهت اخذ درجه کارشناسی ارشد

بهمن۹۲

دانشگاه صنعتی شاهرود

دانشکده : کامپیوتر

گروه : هوش مصنوعی

پایان نامه کارشناسی ارشد خانم ندا عظیمی

تحت عنوان: خلاصه‌سازی خبر با استفاده از تکنیک‌های هوش مصنوعی

در تاریخ ۱۳۹۲/۱۱/۳۰ توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تقدیم به مادر و پدر عزیزم

خواهر و برادران بزرگوارم

و همسفر راه زندگی ام، همسر مهربانم

که همیشه به من لطف و محبت فراوان داشتند؛ آفتاب مهرشان در آستانه قلبم، همچنان پابرجاست.

تشکر و قدردانی

سپاس خدایی راکه نیکویی های آفرینش رابرای مابرگزیدو سپاس خدایی راکه سیاهی ندانستن راازمن زدود و هزار سپاس از برای او، به امید آنکه توفیق یابم جز خدمت به خلق او نکوشم.

بر خود لازم میدانم که از زحمات بی دریغ استاد راهنمای خود، جناب آقای دکتر زاهدی که اینجانب رادر تمامی مراحل پایان نامه، گام به گام راهنمایی وحمایت دلسوزانه نمودند کمال تشکر رانمایم.

از سرکارخانم مهندس رحیمی که استادمشاور بنده دراین پایان نامه بودندودر مقاطع مختلف پایان نامه، راهنمایی هایشان را از بنده دریغ ننمودند کمال تشکر رادارم.

همچنین از جناب آقای دکتر حسن پور و جناب آقای دکتر پویان که درکلیه ی مراحل تحصیل باحمایت های علمی و معنوی خویش،همواره ره گشای مسیر موفقیت اینجانب بوده اند، صمیمانه تشکر وقدردانی می نمایم.

تعهد نامه

اینجانب ..... دانشجوی دوره کارشناسی ارشد رشته .....

دانشکده ..... دانشگاه صنعتی شاهرود نویسنده پایان‌نامه .....

..... تحت راهنمایی متعهد می‌شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده :

با افزایش روز افزون منابع متنی الکترونیکی در شبکه جهانی وب، نیاز به دسترسی صحیح، سریع و آسان به اطلاعات بیشتر احساس می‌شود. برای یافتن سریع و مناسب سندهای مورد نظر کاربر، خواندن کامل متون بزرگ نامناسب است. در این میان سیستم‌های خلاصه‌سازی اتوماتیک نقش مهمی را ایفا می‌کنند و همچنین به عنوان یک راه حل برای مشکل سربار اطلاعات در نظر گرفته می‌شود. خلاصه‌سازی عبارت است از نمایش فشرده و دقیق متن ورودی به نحوی که متن خروجی مفاهیم مهم متن ورودی را داشته باشد. از کابردهای خلاصه‌سازی می‌توان به خلاصه‌سازی اتوماتیک اخبار و ارسال آن‌ها از طریق پست‌الکترونیکی یا پیامک اشاره نمود. از دیگر کاربردهای آن می‌توان خلاصه‌سازی تحقیقاتی، تجاری، خلاصه‌سازی صفحات وب برای آنکه در صفحه موبایل قابل نمایش باشد، در سیستم بازیابی اطلاعات، صنعت مخابرات، ویراستارها و سیستم‌های فیلترینگ را نام برد. دو چالش اساسی در طراحی سیستم‌های خلاصه‌سازی زبان فارسی وجود دارد. چالش اول پیوستگی میان برخی علائم با لغات و تنوع نگارشی در کلمات و دوم انتخاب جملات مهم و پیوسته برای حضور در خلاصه می‌باشد.

به علت چالش‌های خلاصه‌سازی در زبان فارسی ما سعی کرده‌ایم در این پایان‌نامه، مرحله پیش‌پردازش را به طور کامل انجام دهیم. استخراج جملات مرتبط و پیوسته برای خلاصه نیازمند کلمات‌کلیدی دقیق و مرتبط با هم است که در این تحقیق از روش استخراج کلمات‌کلیدی هم‌رخداد و همچنین یک روش مبتنی بر استخراج رویدادها از متن برای خلاصه‌سازی اخبار فارسی پیشنهاد شده است. سیستم پیشنهادی عنوان خبر، متن خبر و میزان فشرده‌سازی را از کاربر دریافت می‌کند و خلاصه‌ای مرتبط با عنوان خبر و متناسب با میزان فشرده‌سازی تولید می‌کند. روش پیشنهادی یک روش آماری و بدون ناظر است و خلاصه‌سازی انجام شده از نوع خلاصه‌ی گزینشی می‌باشد. برای ارزیابی سیستم پیشنهادی، از معیارهای دقت و بازخوانی و مقایسه با خلاصه مرجع استفاده شده است و همچنین سیستم پیشنهادی با سیستم FarsiSum مقایسه شده است. ارزیابی روی یک مجموعه‌ای از خبرهای فارسی منتخب از پیکره‌ی همشهری، از گروه‌های مختلف صورت گرفته‌است. نتایج نشان دهنده عملکرد ۸۱/۶۶ درصدی سیستم در مقایسه با خلاصه‌های انسانی بوده است و همچنین در مقایسه با سیستم FursiSum از عملکرد بهتری برخوردار است.

واژهای کلیدی:

خلاصه‌سازی اتوماتیک اخبار، شناسایی رویداد، خلاصه استخراجی،کلمات هم‌رخداد

مقالات استخراج شده از پایان‌نامه :

1. Neda Azimi، Morteza Zahedi، "Persian News Summarization by Event Detection" ،Journal of Intelligent Information System، JIIS(submitted on Jan27، 2014).

2. Neda Azimi، Morteza Zahedi، " Persian News Summarization by Event Detection" ،International Journal on Document Analysis and Recognition ،IJDA(submitted on Jan 27، 2014).

3. Neda Azimi، Morteza Zahed، "Improvement of Persion News Summarization Using Event Detection "،Internationl Journal of Computer Application(0975-8887) (revised، 2014).

۴. عظیمی ندا، مرتضی زاهدی، "خلاصه‌سازی خبر با استفاده از شناسایی رویدادها" ، نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران(ارسال شده در تاریخ ۱۳۹۲/۱۰/۱۸)

فهرست مطالب

عنوان مطالب **شماره صفحه**

فهرست شکل ها

عنوان **شماره صفحه**

فهرست جداول

عنوان **شماره صفحه**

فصل اول

مقدمه

۱-۱- مقدمه

با گسترش روز افزون حجم اطلاعات، نیاز به سیستم‌های کامپیوتری جهت پردازش و تحلیل اطلاعات بیشتر احساس می‌شود. از آنجا که درصد قابل توجهی از اطلاعات تولید شده به صورت متنی غیر ساختار یافته و نیمه‌ساختار یافته است، سیستمی که بتواند این اطلاعات را تحلیل و پردازش کند، به‌شدت مورد توجه قرار خواهد گرفت. یکی از انواع سیستم‌هایی که در تحلیل و پردازش متون وجود دارد، سیستم‌های خلاصه‌ساز متن است که حجم زیادی از متن را دریافت نموده و بر اساس الگوریتم‌ها و تکنیک‌های مختلف، آن را خلاصه می‌نماید. پیشرفت اینترنت در سال ۱۹۹۰ و پیشرفت روز افزون پایگاه داده‌های الکترونیکی نیاز به توسعه و تکمیل ابزار جستجوی اطلاعات را بیشتر نمود. در انتهای دهه ۱۹۹۰ پروژه‌های تحقیقاتی روی ایجاد سیستم‌های خلاصه‌سازی اتوماتیک در کشور آمریکا و کشورهای اروپایی تعریف شد[۱].

۱-۲- مزایا و کاربردهای خلاصه‌سازی اتوماتیک

در [۲]، سه مزیت عمده برای تولید خودکارخلاصه به وسیله ماشین برشمرده شده است، که عبارتند :

۱. اندازه‌ی خلاصه قابل کنترل است، یعنی ماشین می‌تواند خلاصه را با توجه به میزان فشردگی مورد نظر کاربر تهیه کند.
۲. محتوای آن قابل پیش‌بینی است.
۳. می‌توان مشخص کرد که هر بخش از خلاصه مربوط به کدام بخش یا بخش‌هایی از متن اصلی است.

از کابردهای خلاصه‌سازی می‌توان به خلاصه‌سازی اتوماتیک اخبار و ارسال آن‌ها از طریق پست-الکترونیکی یا پیامک اشاره نمود. از دیگر کاربردهای آن خلاصه‌سازی تحقیقاتی، تجاری و خلاصه

سازی صفحات وب برای آنکه در صفحه موبایل قابل نمایش باشد، است. در یک سیستم بازیابی اطلاعات، نمایش خلاصه‌ای که به صورت اتوماتیک ایجاد شده‌است مناسب و مفید می‌باشد، در این صورت کاربر می‌تواند سریعا تصمیم بگیرد که چه سندی مطلوب و ارزش باز کردن دارد. گوگل تا حدودی این کار را انجام می دهد و اطلاعات مختصری به همراه نتایج جستجو نشان می‌دهد. به خاطر علاقه روز افزون دولت، بخش بازرگانی، بخش دانشگاهی به این صنعت، تحقیقات و محصولات خلاصه‌سازی زیادی در سراسر دنیا موجود می‌باشد. سیستم‌های سنتی بین سال‌های ۱۹۵۰ تا۱۹۸۰ برای جستجوی اطلاعات علمی استفاده می‌شد. سیستم‌های خلاصه‌ساز امروزی در زمینه‌های جدیدی همانند صنعت مخابرات، ویراستارها، سیستم‌های فیلترینگ و یادگیری زبان خارجی مورداستفاده قرار می‌گیرد. کار اصلی سیستم خلاصه‌ساز کمک به کاربر برای پیدا کردن اطلاعات مورد نیازش است.

۱-۳- معرفی خلاصه‌سازی خودکار و انواع آن

منظور از خلاصه‌سازی، دریافت یک متن و تولید یا استخراج یک متن دیگر از آن متن است، به گونه‌ای که متن به دست‌آمده از متن اصلی کوتاه‌تر باشد، نکات اصلی و مهم آن را دربرداشته باشد و بین جملات آن پیوستگی وجود داشته باشد.

یکی از شاخه‌های مرتبط به خلاصه‌سازی متون، متن‌کاوی است که به معنای کشف اطلاعات جدید به وسیله استخراج اطلاعات به صورت اتوماتیک و بر اساس منابع مختلف می‌باشد. در متن کاوی تعداد زیادی متن تحلیل شده و الگوهای با ارزشی که در متن مخفی است یافته می‌شود. بخش متن کاوی در سال‌های اخیر شامل کشف ارتباط بین کلمات، طبقه بندی اسناد و خلاصه‌سازی شده-است. در سال‌های اخیر فعالیت گسترده‌ای روی ساخت و توسعه‌ی سامانه‌های خودکار خلاصه‌سازی متن برای زبان‌های مختلف انجام شده است.

به طور کلی، دو نوع اصلی برای خلاصه وجود دارد: خلاصه‌ی گزینشی و چکیده [۳].

- خلاصه‌ی گزینشی با توجه به معیارهای آرمای، شهودی[۱] و یا ترکیبی از این دو تهیه می-شود. از آنجا که در تولید این دسته از خلاصه‌ها، جملات متن تغییرات نحوی و معنایی ندارند، می‌توان آن را نوعی گزینش جملات قلمداد کرد. در واقع اگر متن خلاصه با انتخاب جملاتی از متن اصلی به دست آید، نوع خلاصه‌سازی، «استخراجی» یا «گزینشی» است. در حال حاضر خلاصه‌های استخراجی به دلیل آسان و ارزان بودن عمومیت بیش‌تری دارند.

- خلاصه‌ی چکیده، تفسیری از متن اولیه است. درتولید خلاصه‌ی چکیده، مفاهیم جملات متن اصلی به شکل کوتاه‌تر بازنویسی می‌شود. به عنوان مثال، جمله «اوسیب، انگور وگیلاس‌ها را خورد» را می‌توان به صورت «او میوه‌ها را خورد» نوشت. به عبارت دیگر اگر خلاصه متن پس از فهم مطالب موجود در متن اصلی تولید شود، خلاصه‌سازی از نوع «چکیده» است [۴].

هر دو نوع خلاصه‌سازی، با چالش‌های مختلفی مواجه هستند. در روش استخراجی، تشخیص جملات مهم متن، شناسایی و استخراج کلمات کلیدی، تجزیه متن اصلی و تولید متن خلاصه‌ای که دارای خوانایی و پیوستگی باشد، از چالش‌های اصلی به شمار می‌رود. در روش چکیده نیز باید ابتدا متن اصلی فهمیده شود و بر اساس معنای موجود در متن و به صورت معنایی، چکیده‌ای از متن اصلی تولید شود. از سوی دیگر، چکیده، خلاصه‌ای است که دست کم بخش‌هایی از آن در متن یا متون اولیه وجود ندارد. در تولید چکیده، پاره‌ای از جمله‌ها یا همه‌ی آن‌ها بازنویسی می‌شوند. در این روش، با چالش‌های موجود در زمینه پردازش زبان طبیعی و تجزیه و تحلیل معنایی متن، برای درک و تفسیر متن روبه‌رو هستیم.

[۱] Heuristic

تا کنون، روش‌های مختلفی برای خلاصه‌سازی متون فارسی انجام شده است که هر کدام یک یا چند پارامتر موثر در بهبود کیفیت خلاصه را مد نظر قرار دادند. روش‌های انجام شده بیش‌تر در حوزه خلاصه‌سازی استخراجی صورت گرفته است.

۱-۴- سیستم‌های خلاصه‌ساز معروف

پس از پی بردن به اهمیت خلاصه‌سازی خودکار در این قسمت در نظر داریم تعدادی از سیستم-های معروف خلاصه‌سازی را معرفی نماییم. ازآنجایی که تعداد این سیستم‌ها زیاد می‌باشد، فقط به بخشی از آنها خواهیم پرداخت.

❖ **SweSum**

یکی از معروف‌ترین سیستم‌های خلاصه‌سازی می‌باشد که به زبان‌های مختلف، عمل خلاصه‌سازی را انجام می‌دهد. اولین خلاصه‌ساز تولید شده برای زبان سوئدی می‌باشد[۵]. دقت آن برای متون خبری برای زبان سوئدی، برای خلاصه با میزان ۴۰ درصد، برابر ۸۴ درصد اندازه‌گیری شده است (با طول متوسط ۱۸۱ کلمه). نسخه نهایی این خلاصه‌ساز در حال حاضر برای زبان‌های انگلیسی، دانمارکی، نروژی و سوئدی تولید شده و نسخه آزمایشی آن هم برای زبان‌های فرانسوی، آلمانی، ایتالیایی، اسپانیایی و یونانی، تحت بررسی و آزمایش است. نسخه تحت وب این خلاصه‌ساز در حال حاضر دروب[۲] موجود است.

❖ **MEAD**

یکی از معروف‌ترین خلاصه‌سازهای موجود می‌باشد که در بسیاری از مقالات مورد استفاده قرار گرفته است. MEAD خلاصه‌سازی چندسندی می‌باشد. ورژن ۱ و ۲ این خلاصه‌ساز در دانشگاه میشیگان در سال ۲۰۰۰ با استفاده از زبان برنامه نویسی پرل پیاده‌سازی شده است. این خلاصه‌ساز

[۲] http://swesum.nada.kth.se/index-eng.htm

طوری پیاده‌سازی شده است که برای تمامی زبان‌ها، قابل اجرا باشد و در حال حاضر نسخه‌های انگلیسی، چینی، ژاپنی و هلندی آن موجود می‌باشد. نسخه‌ی هلندی این خلاصه‌ساز به همراه اطلاعات و مستندات کامل درباره این خلاصه‌ساز در سایت[۳] موجود می‌باشد.

❖ **DMSumm**

سیستم DMSumm یک رویکرد عمیق در مساله‌ی خلاصه‌سازی متن است که سه مرحله دارد: انتخاب محتوی، طرح ریزی متن و ادراک زبانی. پروسه انتخاب محتوا، اطلاعاتی که در خلاصه باید وارد شود را مشخص می‌کند. طرح ریزی متن، نگاشتی از روابط معنایی و مفهومی است به روابط rhetorical انجام می‌دهد که منجر به ساخت طرح‌های معنایی(rhetorical) از متن می‌شود. ادراکات زبانی، بیان کننده‌ی طرح‌ها در خلاصه‌های نوشته شده هستند. این عمل بر پایه مدل سخنی است که از سه منبع دانش مختلف (معنایی، مفهومی و rhetorical) ساخته شده است[۶]. این سیستم به چند محدودیت غلبه کرده است: تحقق، هدف، فصاحت و حفظ موضوع مرکزی. این خلاصه‌ساز بر روی زبان‌های انگلیسی و پرتغالی خلاصه تولید می‌کند. نسخه قابل دانلود این خلاصه‌ساز در وب[۴] موجود می‌باشد.

❖ **GLEANS**[۵]

این خلاصه‌ساز در سال ۲۰۰۰ معرفی شد و از نوع چندسندی می‌باشد[۷]. این سیستم، خلاصه-سازی را در چهار مرحله انجام می‌دهد:

---

[۳] http://www.summarization.com/mead

[4] http://www.icmc.usp.br/~taspardo/DMSumm.htm

[۵] A Generator of Logical Extracts and Abstracts for Nice Summaries

- اسناد تجزیه می‌شوند و بخش‌های اصلی هر جمله شناسایی می‌شود، بعضی عبارات تکراری رفع می‌شوند و در نهایت به یک نمایش استاندارد که مشخص‌کننده موجودیت‌های اصلی و روابط آنها است، نگاشت می‌شوند.
- هر مجموعه از اسناد توسط محتوایشان به دسته‌های افراد، تک‌رویداد، چندرویداد و فجایع طبیعی، دسته‌بندی می‌شوند.
- با در اختیار داشتن نوع دسته‌ی مجموعه و نمایش استاندارد اسناد یا انتخاب کلمات برجسته در مجموعه، موجودیت‌های اصلی و روابط آنها استخراج می‌شود.
- یک سرفصل بر اساس نوع مجموعه و موجودیت‌های اصلی و روابط آنها ایجاد می‌شود. برای مجموعه‌های چند رویدادی، یک چکیده کوتاه هم می‌تواند به همین روش تولید شود.

یک چکیده با اعمال یک کتابخانه از الگوهای استاندارد از تحلیل دستی چکیده‌ها در پیکره زبانی آموزشی تولید می‌شود. این الگوها تعیین می‌کنند چه جملاتی از متن منبع، نیازهای یک خلاصه‌ی استاندارد را برآورده می‌کنند و آنها را استخراج می‌کنند. بعد از پردازش، نشانه‌های معلق سخن حذف می‌شوند، در مورد عبارات تکراری که برای هر موجودیت، استفاده می‌شود تصمیم‌گیری می‌شود و عبارات زمانی در یک فرم استاندارد، ارائه می‌شوند.

❖ **PERSIVAL**[۶]

این سیستم، خلاصه‌ای از مقالات پزشکی مخصوص یک بیمار را از کتابخانه‌ی دیجیتالی توزیع شده‌ی چند رسانه‌ای مراقبت از بیماران، تهیه می‌کند[۸]. این سیستم شامل مجموعه بزرگی از داده-های ویدئویی است. اسناد ویدئویی قطعه‌بندی می‌شوند و یک خلاصه تولید می‌شود سپس ویدئوها در

[۶] Personalized Retrieval and Summarization of Image , Video and Language

سطوح نحوی و معنایی، شاخص‌بندی می‌شوند. یک مجموعه ابزارهای مبتنی بر محتوا برای جستجوی ویدئو، توسعه داده شده‌اند. این سیستم از معانی[۷] برای جستجوی تعاریف استفاده می‌کند.

❖ **SUMMARIST**

این سیستم از نوع تک‌سندی می‌باشد [۹]. سیستم شامل سه مرحله پردازش است: شناسایی موضوع، تفسیر و تولید خلاصه. شناسایی موضوع، به کمک امضاهای موضوع که قبلا تهیه شده، انجام می‌شود. امضاهای سند، چندتایی‌هایی به فرم <topic, signature> است که امضا، لیستی از کلمات وزن‌دار به فرم <tn,wn> و ... و <t2,w2> و <t1,w1> است. امضاهای سند، می‌توانند به صورت خودکار یاد گرفته شوند. پس شناسایی موضوع، شامل قطعه‌بندی متن و مقایسه اندازه‌های[۸] متن با امضاهای موضوع موجود است. موضوع شناسایی شده در طول پروسه‌ی تفسیر، ترکیب می‌شود. سپس موضوعات ترکیب شده، مجددا فرمول‌بندی می‌شوند. مرحله‌ی آخر هم طبق معمول، استخراج است. این خلاصه‌ساز برای زبان‌های مختلفی عمل کرده و طوری نوشته شده که تمام زبان‌ها را پشتیبانی کند؛ مثل: انگلیسی، ژاپنی، اسپانیایی، عربی، اندونزیایی، کره‌ای.

۱-۵- مروری برکارهای انجام شده در زمینه‌ی خلاصه‌سازی زبان فارسی

کارهای انجام شده در زمینه‌ی پیاده‌سازی ابزارهای خلاصه‌سازی خودکار فارسی نسبت به کارهای انجام شده در زبان‌های دیگر بسیار اندک هستند. در ادامه، این کارها به طور مختصر، مورد اشاره قرار خواهند گرفت.

❖ **FarsiSum:**

[۷] Defind

[۸] Span

این سیستم توسط نیما مزدک در سال ۲۰۰۴ طراحی شده‌است، با استفاده از سیستم خلاصه‌ساز SweSum که یک سیستم خلاصه‌ساز عمومی است و با یک رویکرد آماری به خلاصه‌سازی متون فارسی پرداخته است، تهیه شده است. از خصوصیات زبان‌شناختی، تنها فهرست کلمات توقف را در پیش‌پردازش منظور کرده‌است. لیست کلمات‌توقف شامل ۲۰۰ کلمه است؛ که مجموعه‌ای از افعال، قیدها، حروف ربط و ضمیرهایی که بیشترین تکرار را در متن دارند تشکیل شده‌است. برای شناسایی مرز جملات از علائمی مانند " ؟"،"<"،">" و خط جدید استفاده شده و علامت" ." ،" :" ،">"،"<"،"؟" ،"!"، "،" برای شناسایی کلمات به کاربرده شده است. درFarsiSum برای محاسبه امتیاز هر کلمه از فرمول(۱-۱) و برای محاسبه امتیاز جملات از فرمول (۱-۲) استفاده شده است[۱۰].

$$Word\ Score = (word\ frequence) * (a\ keyword\ constant) \qquad (۱-۱)$$

Word frequence = تعداد تکرار کلمه در متن

A keyword constant = یک مقدار پیش‌فرض است که ۰,۳۳۳در نظر گرفته شده است

$$Sentence\ Score = \sum word\ score \qquad (۱-۲)$$

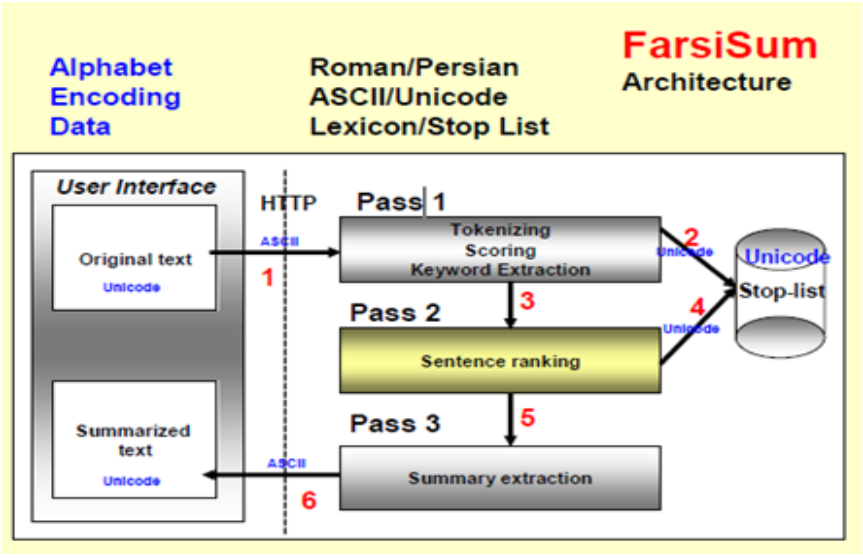
این سیستم برای خلاصه‌سازی اخبار به کار می‌رود و همواره به اولین جملات و اولین پاراگراف-ها، وزن بیشتری اختصاص می‌دهد.از فرمول (۱-۳) برای تعیین امتیاز موقعیت مکانی هر جمله استفاده شده است[۱۰].

$$Position\ Score = (\frac{1}{Line\ nr} * 10) \qquad (۱-۳)$$

در نهایت از مجموع دو امتیاز sentence score و position score امتیاز نهایی هر جمله بدست می‌آید، هر جمله با بیشترین امتیاز در خلاصه حضور پیدا می کند. در این سیستم همچنین برای کاهش افزونگی از متد LSA[۹] استفاده شده است.

این سیستم قادر به خلاصه کردن متون روزنامه‌های فارسی با قالب HTML و متن کد شده با فرمت یونیکد می‌باشد و دارای خروجی از نوع گزینشی می‌باشد. این سیستم تحت ویندوز و به زبان پرل ارائه شده و معروف‌ترین سیستم خلاصه‌ساز فارسی می‌باشد. پروسه‌ی تولید خلاصه در این خلاصه‌ساز را در شکل شماره‌ی یک مشاهده می‌کنید:

مرورگر، درخواست خلاصه را به همراه سند، از طریق سرور HTTP به سروری که FarsiSum در آن قرارگرفته است، می‌فرستد. سند در طی سه مرحله با استفاده از لیست توقف فارسی، خلاصه می‌شود. متن خلاصه‌ی تهیه شده از طریق HTTP به کاربر باز می‌گردد.



شکل۱-۱ ساختار کلی FarsiSum [10]

در این پایان‌نامه به دلیل اینکه در سیستم پیشنهادی و سیستمFarsiSum از خلاصه‌سازی استخراجی استفاده شده‌است و هر دو در زمینه خلاصه‌سازی خبر هستند و روش استفاده شده در FarsiSum

[۹] Latent Semantic Analysis

شباهت بیشتری با سیستم پیشنهادی دارد نتایج سیستم ما با FarsiSum مقایسه شده است. ایراداتی که در سیستم خلاصه‌ساز FarsiSum می‌باشد مراحل پیش‌پردازش کامل انجام نشده است در واقع مرحله‌ی پیش‌پردازش شامل یکسان‌سازی نگارشی و ریشه‌یابی نمی‌باشد و لیست کلمات توقف ۲۰۰ کلمه درنظر گرفته شده‌است که خیلی کم می‌باشد و باعث می‌شود کلمات‌کلیدی مناسبی استخراج نشود و ارتباط معنایی بین کلمات در نظر گرفته نشده است که ما سعی کرده‌ایم آنها را برطرف کنیم. در واقع می‌توان گفت سیستم پیشنهادی ما بهبود یافته سیستم خلاصه‌ساز FarsiSum می‌باشد.

❖ **PARSUMIST:**

این سیستم توسط مهرنوش شمس فرد، تارا اخوان و مونا عرفانی در سال ۲۰۰۹ ارائه شده‌است. نوع خلاصه‌سازی از نوع گزینشی است[۱۱]. با استفاده از رویکرد آماری و رعایت موارد بسیاری از ویژگی‌های زبان‌شناختی زبان فارسی در مرحله پیش‌پردازش، خلاصه‌سازی را انجام می‌دهد. سیستم خلاصه‌ساز PARSUMIST برای خلاصه‌سازی تک‌سندی و چندسندی مناسب است و برای زبان فارسی طراحی شده است و قابلیت دریافت درخواست از کاربر را دارد. در این سیستم ابتدا تک تک اسناد خلاصه می‌شوند و سپس خلاصه‌ها را به هم چسبانده و پردازشی روی آن‌ها انجام می‌دهند و خلاصه‌نهایی، خلاصه‌ای از خلاصه‌ها است. در مرحله پیش‌پردازش قطعه‌بندی و حذف کلمات غیرمهم و تشخیص اسامی‌خاص و انطباق مفهومی انجام گرفته است. در بخش انطباق مفهومی تمامی کلمات موجود در مجموعه‌های مترادف، توسط گروهشان جایگزین می‌شوند. از نتیجه این عمل در مراحل بعد جهت رفع افزونگی استفاده شده است.کلمات غیرمهم حدود ۴۰۰ کلمه از پرتکرارترین کلمات فارسی شامل افعال رایج، ضمایر، قیدها، حروف ربط، حروف اضافه و حروف تعریف ایجاد شده‌است[۱۱]. برای تعیین میزان شبات جملات، گرافی بدون جهت تشکیل شده است که نودهای آن، جملات متن هستند و وجود یال بین دونود نمایانگر شباهت میان آنهاست. شباهت میان جملات براساس تعداد کلمات مشابه و یا مرتبط براساس

منابع موجود مانند مجموعه‌های ترادف، محاسبه می‌شود. مقدار به دست آمده وزن یال میان دو جمله را می‌دهد. برای تولید خلاصه، جمله با بیشترین امتیاز و کمترین شباهت به جملات موجود در خلاصه، به خلاصه اضافه می‌شود. برای انجام این کار، در هر مرحله شباهت بین جمله کاندید برای حضور در خلاصه(دارای بیشترین امتیاز) و جملات موجود، بررسی می‌شود و در صورتی که این شباهت بیشتر از ۹۰درصد باشد، جمله کاندید با دریافت امتیاز منفی برای همیشه از لیست جملات کاندید حذف شده و دیگر جهت حضور در خلاصه انتخاب نخواهد شد. در غیر این صورت، اگر میزان شباهت بیشتر از یک حدآستانه مشخص باشد، جمله با امتیاز بالاتر برای حضور در خلاصه انتخاب شده و با کسر امتیازی از جمله دیگر،آن را مجددا در لیست جملات کاندید قرار می‌دهد.

- در[۱۲]، یک روش برای خلاصه‌سازی چند سندی پیشنهاد شده است. مبنای روش، استفاده از روش سلسله مراتبی اتصال میانگین برای خوشه‌بندی جمله‌ها و بکارگیری روش تجزیه به مقادیر منفرد یا SVD[۱۰] برای تعیین اهمیت جمله‌ها است. در این کار، از دو منبع زبانی ساده برای قطعه‌بندی متن به واژه‌ها و دیگری برای تعیین فراوانی واژه‌ها در سندها استفاده شده است.
- سیستم خلاصه‌ساز دیگری که توسط کریمی و شمس فرد ارائه گریده است، تک سندی و بر مبنای گزینش جمله‌ها کار می‌کند. این روش در گزینش جملات خروجی، ترکیبی از دو روش زنجیره‌ی لغوی و نظریه‌ی گراف را بکار می‌برد و خروجی نهایی می‌تواند کلی یا بر اساس پرس وجوی کاربر باشد[۱۳]. در روش مبتنی بر گراف، سند به شکل گرافی غیر جهتدار که گره‌های تشکیل دهنده‌ی آن، جمله‌ها هستند، ارائه می‌شود. در این گراف اگر بین دو جمله در دو گره، شباهت وجود داشته باشد، دو گره با یک لبه به یکدیگر متصل می‌شوند،

[۱۰] Singular value decomposition

این شباهت می‌تواند وجود کلمات مشترک یا رابطه‌ی بین کلمات آن با معیار کسینوسی باشد. زیرگراف‌ها، نشان دهنده‌ی موضوعات مجزای موجود در سند هستند. در خلاصه‌های مبتنی بر سوال، خلاصه ممکن است از زیرگراف خاصی انتخاب شود؛ درحالی که در خلاصه-های کلی، جملات می‌توانند از تمام زیرگراف‌ها انتخاب شوند. گره‌های با درجه‌ی بالا، جملات مهم در هر زیر گراف هستند و اولویت بیشتری برای بودن در خلاصه دارند.

- ❖ در [۱۴]، یک روش خلاصه‌سازی چندسندی معرفی شده است که نوع خروجی تولید شده توسط آن نیز گزینشی است.
- ❖ هنرپیشه و همکاران [۱۵] یک سیستم خلاصه‌ساز چندسندی چند زبانی ارئه کرده‌اند. خلاصه حاصله از نوع مستخرج بوده و از روش‌های آماری استفاده شده است.
- ❖ در [۱۶]، یک روشی برای خلاصه‌سازی چندسندی متون فارسی پیشنهاد شده‌است که در آن از الگوریتم خوشه‌بندی Kmeans برای خوشه‌بندی جمله‌ها استفاده شده است.
- ❖ در [۱۷]، نوعی از خلاصه‌ی خبر به نام خلاصه‌های پیشینه-خبر معرفی شده‌است. در این نوع از خلاصه، به همراه هر خبر خلاصه‌ای از خبرهای پیشین مرتبط با آن در اختیار کاربر قرار می‌گیرد.
- ❖ در [۱۸]، علاوه بر مروری بر روش‌های خلاصه‌سازی موجود، گزارشی از پیاده‌سازی یک نمونه‌ی عملی برای خلاصه‌سازی فارسی ارائه شده است.

همان‌گونه که مشخص است، در کارهای انجام شده بیشتر خلاصه‌سازی استخراجی انجام شده است و خلاصه‌سازی به صورت معنایی یا چکیده کار نشده است. علت آن هم پیچیدگی، پر هزینه بودن و آماده نبودن ابزارها و زیرساخت‌های لازم جهت خلاصه‌سازی به روش چکیده است.

۱-۶- بررسی روش‌های مربوط به پایان‌نامه

علاوه بر کارهایی که در بخش کارهای انجام شده در زبان فارسی به آن‌ها اشاره شد در این بخش به بررسی روش‌های دیگری که به موضوع پایان‌نامه مربوط هستند، می‌پردازیم. در حوزه زبان فارسی تا به حال در زمینه‌ی خلاصه‌سازی با استفاده از شناسایی رویدادها[۱۱] و استفاده از Topic Modeling کاری انجام نشده است؛ در این تحقیق ما با بررسی روش‌های موجود در زمینه شناسایی رویدادها و موضوعات متن در زبان‌های دیگر، خلاصه‌سازی با استفاده از شناسایی رویدادها را در زبان فارسی پیاده‌سازی کرده‌ایم.

در [۱۹] روشی پیشنهاد شده که در آن خلاصه‌سازی بر مبنای دسته‌بندی رویدادها صورت می-گیرد. در این روش از روابط معنایی سطحی و وردنت برای کاهش مشکل پراکندگی دادها و دسته-بندی کردن رویدادهای مشابه استفاده شده است. برای هر جمله فعل و فاعل و مفعول تعیین می‌شود و هر کدام به عنوا ن یک دسته در نظر گرفته می‌شوند، اگرپنج تا سه تایی مشابه وجود داشته باشد سه دسته می‌توانند اتصال پیدا کنند و یک ارتباط معنایی سطحی ایجاد می‌شود و آن جمله به عنوان یک رویداد در نظر گرفته می شود.

همچنین از میان کارهایی که به شناسایی موضوع[۱۲] در خلاصه‌سازی پرداخته‌اند، می‌توان به [۲۰] اشاره نمود. در این روش برای شناسایی جملات مرتبط با موضوع متن اصلی از VSM[۱۳] ، که مدلی برای نمایش موضوعات می‌باشد استفاده شده‌است و برای دسته‌بندی موضوعات از الگوریتم K-means استفاده شده‌است.

[۱۱] Event Detection

[۱۲] Topic Detection

[۱۳] Vector space model

۱-۷- نگاهی کوتاه بر پایان‌نامه

مشکلاتی که در سیستم‌های خلاصه‌سازی بالا وجود دارد عبارتند از : نحوه امتیاز دهی به جملات، نداشتن لیست کاملی از کلمات توقف، کامل نبودن مراحل پیش‌پردازش، نداشتن پیوستگی مطالب زیرا هر جمله از قسمتی از متن استخراج شده است. در این تحقیق ما سعی کرده‌ایم برای تولید خلاصه دقیق و مناسب تمامی این موارد را رعایت کنیم.

این پایان‌نامه، به معرفی یک روش جدید برای خلاصه‌سازی اخبار اختصاص دارد. روش پیشنهادی یک روش گزینشی است که خلاصه‌ی تولید شده در آن، گزیده‌ای از جمله‌های سندهای اولیه است. ما برای تولید خلاصه ابتدا متن و عنوان خبر و میزان فشرده‌سازی را از کاربر دریافت می‌کنیم. سپس کلمات و جملات با استفاده از جداکننده‌های کلمه و جمله در آرایه‌های جداگانه‌ای ذخیره می‌شوند؛ سپس عملیات یکسان‌سازی نگارشی روی متن انجام می‌شود و در مرحله بعد کلمات توقف (لیستی شامل ۸۰۰ کلمه و قیدهاو افعال و... تهیه کرده‌ایم) از متن حذف می‌شوند. برای استخراج کلمات کلیدی ما از روش استخراج سریع خودکار کلمات کلیدی[۱۴] استفاده کرده‌ایم و توانستیم کلمات کلیدی هم‌رخداد[۱۵] را از متن استخراج کنیم(منظور از هم رخدادی کلمات، رخدادکلمات درمجاور یکدیگر است که حداقل دو بار درمتن تکرار شده باشند) سپس با استفاده از کلمات‌کلیدی، رویدادهای مهم و مرتبط با عنوان خبر را از متن اولیه استخراج کرده‌ایم(که رویدادها مجموعه‌ای از کلمات کلیدی هستند) و در مرحله‌ی بعد با امتیازدهی به جملات، جملات مهم برای حضور در خلاصه را تعیین کرده‌ایم.

در مروری بر کارهای گذشته، روش FarsiSum شباهت بیش‌تری با روش پیشنهادی دارد. مهمترین شباهت این دو روش، گزینشی بودن آنها و اینکه هر دو سیستم در زمینه خلاصه‌سازی خبر

[۱۴] Rapid automatic keyword extraction

[۱۵] Co-occurrence

هستند. با وجود شباهت‌های بین این دو روش، تفاوت‌هایی نیز بین آنها وجود دارد. تفاوت‌ها، مربوط به نحوه گزینش جمله‌ها، استخراج کلمات کلیدی و لیست کلمات توقف در آنها می‌باشد.

در فصل دوم از پایان‌نامه، به معرفی خلاصه‌سازی متون فارسی و چالش‌های آن در زبان فارسی می‌پردازیم. فصل سوم، به بررسی فرآیند پیش‌پردازش و همچنین نحوه‌ی تولید منابع زبانی مورد نیاز برای روش پیشنهادی اختصاص دارد. در فصل چهارم، به معرفی فرآیند کلی روش پیشنهادی می‌پردازیم. فصل پنجم پایان‌نامه، حاوی نتایج پیاده‌سازی و ارزیابی روش پیشنهادی و همچنین مقایسه‌ی آن با سیستم خلاصه‌ساز FarsiSum است. در فصل پایانی نیز به جمع‌بندی وکارهای آتی می‌پردازیم.

فصل دوم

خلاصه سازی متون فارسی

۲-۱- مقدمه

بسیاری از متخصصین در زمینه هوش مصنوعی بر این باورند که مهمترین وظیفه‌ای که هوش مصنوعی باید به آن بپردازد پردازش زبان طبیعی است. دلیلی که ایشان برای این اعتقاد خود ارائه می کنند آن است که پردازش زبان طبیعی راه ارتباط مستقیم انسان و کامپیوتر را از طریق مکالمه باز می کند. به این ترتیب دیگر برنامه‌نویسی معمولی و قراردادهای مربوط به سیستم‌عامل کنار گذاشته خواهد شد، اگر یک کامپیوتر بتواند یک زبان انسانی را درک کرده و به آن صحبت کند دیگر به بسیاری از وظایفی که باید توسط مهندسین نرم افزار طراحی شوند نیازی نخواهد بود.

خلاصه‌سازی یکی از زیر مجموعه مشکلات پیچیده پردازش زبان طبیعی می‌باشد که در حال حاضر همچنان حل نشده‌است و هنوز زمان باقی است تا به حدی برسیم که ماشین همانند انسان به طور کامل متن را متوجه شود.

هدف از خلاصه‌سازی اتوماتیک، آن است که مراحلی که توسط فرد انجام می‌شود، در خلاصه سازی، اتوماتیک انجام شود. بدین صورت که تمام متن خوانده و فهمیده شود و سپس خلاصه تولید شود. فهمیدن متن شامل تشخیص قسمت‌های مهم و غیرمهم متن است. مرحله تبدیل متن ورودی به متن خلاصه شامل مشخص کردن کلمات کلیدی، مفهوم اصلی، کلمه‌های مهم و جمله‌های مهم می‌باشد. مسائلی که در خلاصه‌سازی باید در نظر داشت آن است که:

۱. به چه صورت یک توصیف نرمال ازمتن ورودی می‌توانیم داشته باشیم.

۲. چگونه متن را تجزیه کنیم؟

۳. چگونه مهمترین بخش متن را تشخیص دهیم و متن خلاصه را ایجاد نماییم؟

در این فصل ابتدا به فرآیند خلاصه‌سازی، تفاوت‌های خلاصه‌سازی متون فارسی با سایر زبان‌ها پرداخته شده‌است و پس از آن، چالش‌های خلاصه‌سازی متون فارسی تشریح شده است و در نهایت، نتیجه‌گیری ارائه شده است.

۲-۲- فرآیند خلاصه‌سازی

فرآیند خلاصه‌سازی متون به طور کلی شامل سه فاز: پیش‌پردازش، تحلیل و انتخاب است. در فاز پیش‌پردازش، پردازش‌های اولیه مورد نیاز بر روی متن صورت می‌گیرد. وجود این فاز، در فرآیند خلاصه‌سازی بسیار ضروری است؛ به طوری که اگر وظایف این فاز به خوبی انجام نشود، کیفیت خلاصه تولید شده به میزان چشمگیری کاهش می‌یابد.

فاز دوم، فاز تحلیل است. در این فاز، وظایف اصلی خلاصه‌سازی انجام می‌شود و بخش‌های مهم و اصلی متن شناسایی و امتیاز دهی می‌شوند. در فاز سوم، با توجه به بخش‌های مهم متن که در فاز دوم شناسایی شده‌اند، خلاصه نهایی تولید می‌شود. باید توجه داشت که اطلاعات تکراری، اضافی و یا ناخوانا در خلاصه وجود نداشته باشد.

۲-۲-۱- فاز پیش‌پردازش

پردازش متن شامل چهار سطح است[۲۲]: پردازش لغوی، پردازش ساخت‌واژه‌ای، پردازش نحوی و پردازش معنایی. هریک از کاربردهای فراوان پردازش متن، ازجمله بازیابی اطلاعات، خلاصه‌سازی، درک متن، تولید متن، ترجمه‌ی ماشینی، پرسش و پاسخ به زبان طبیعی، استخراج دانش از متون و موارد دیگر، با توجه به گستردگی و پیچیدگی، در یک یا چند سطح فوق به انجام می‌رسد. مهمترین قسمت در هر سیستم کاربردی پردازش زبان طبیعی تجزیه متن ورودی خواهد بود. از جمله مهم‌ترین وظایفی که در این فاز انجام می‌شود، می‌توان به موارد زیر اشاره نمود :

۱. یکسان‌سازی نگارشی :

در زبان فارسی برخی کلمات و عبارات را به چند شیوه مختلف می‌توان نگارش کرد. به عنوان مثال: «کتاب‌ها» و«کتابها» باید به یک شکل تبدیل شوند.

۲. مشخص کردن کلمات مترادف :

کلماتی مانند: «کامپیوتر» و «رایانه» مترادف هستند. در مرحله پیش‌پردازش لازم است کلمات مترادف یافته و به یک شکل تبدیل شوند.

۳. اعمال گروه‌های نوعی:

به عنوان مثال، جمله: «در ایران برنج، پرتقال، پسته و... کشت می‌شود» را می‌توان با جمله «در ایران محصولات کشاورزی کشت می‌شود» جایگزین نمود.

۴. حذف لغات و بخش‌های اضافی:

حروف اضافه و کلماتی که تقریبا در تمامی متون یافت می‌شوند، به کلمات توقف موسوم هستند. کلمات توقف و همچنین توضیحات اضافی که معمولا داخل پرانتز، خط‌تیره و ... ظاهرمی‌شوند، باید از متن اصلی حذف شوند.

۵. تعیین مرز جملات:

از آنجایی که جمله، یک واحد متنی پایه است و بلافاصله بعد از کلمه و عبارت قرار می-گیرد تعیین و محدوده جمله یک مسله اساسی برای بسیاری از کاربردهای پردازش زبان طبیعی مانند تجزیه، استخراج اطلاعات، ماشین ترجمه، خلاصه‌سازی، برچسب‌گذاری نحوی و معنایی و... است [۲۳]. در این زمینه تا به‌حال درمورد زبان فارسی کار چندانی صورت نگرفته

است، اما در زبان‌های دیگر، مثل انگلیسی، پرتغالی وژاپنی و... ، کارهای زیادی با استفاده از روش‌های مختلف صورت گرفته‌است.

Tokenizer ابزاری برای شکستن یک متن بر اساس واحدهای با معنی مانند کلمه، پاراگراف، نمادهای معنادار مانند space و tab و ... می‌باشد. لازمه‌ی ایجاد این ابزار، جمع‌آوری واحدهایی است که در زبان به عنوان واحدهای مستقل معنایی شناخته می‌شوند. سپس بر اساس انتخاب هر کدام از این واحدها، متن بر اساس آن، شکسته خواهد شد. این ابزار نیز با توجه به بررسی میزان تشابه واحدهای معنایی در دو متن خلاصه در ارزیابی خلاصه‌ها دارای اهمیت بسزایی می‌باشد. از نمونه‌های انگلیسی این ابزار می‌توان به Flex، JLex، JFLex، ANTLR، Ragel و Quex اشاره کرد.

روش‌هایی که در زمینه متون فارسی برای تعیین حدود جمله پیشنهاد شده است عبارتند از[۲۳] :

- تعیین حدود جمله با استفاده از تشخیص فعل
- تعیین حدود جمله با استفاده از مدل چندتایی
- استخراج بردار ویژگی و استفاده از رده‌بندها

بدلیل اشکالاتی مانند ابهام هم‌نویسه‌ها(مَردم ، مُردم)، افعال سبک وتغییرات ساختاری از روش‌های بالا نتایجی با درصد بالا بدست نیامده است(میزان دقت۴۹ درصد)، ولی با تحقیقاتی که انجام شده است می‌توان فهمید که با مسئله تعیین حدود جمله در پیکره‌های متنی زبان فارسی می توان به صورت یک مسئله رده‌بندی برخورد کرد و همچنین استخراج خصیصه‌های مناسب تاثیر به سزایی در نتیجه خواهد داشت[۲۳].

۶. استخراج کلمات کلیدی

کلمات‌کلیدی مجموعه‌ای از لغات مهم در یک مستند هستند که توصیفی از محتوای مستند را فراهم می‌آورند و برای اهداف مختلفی قابل استفاده هستند، کلمات‌کلیدی نکات اصلی متن را توصیف می‌کنند، به همین دلیل می‌توانند به عنوان یک ابزار برای اندازه‌گیری شباهت متون مختلف مورد استفاده قرار گیرند. در مجموع کلمات‌کلیدی ابزار مفیدی برای جستجوی حجم زیادی از مستندات در زمان کوتاه هستند. استخراج کلمات‌کلیدی به طور دستی فرآیندی بسیار دشوار و زمان‌بر است. بنابراین نیاز به یک فرآیند خودکار است که کلمات کلیدی را از مستندات استخراج نماید.

فرآیند کلی برای استخراج کلمات‌کلیدی بدین ترتیب است که ابتدا کلمات کلیدی کاندید تشخیص داده می‌شود، به هرکلمه‌ی کاندید وزنی اختصاص داده شده و در نتیجه کلمات کلیدی با بیشترین وزن انتخاب می‌شوند. در حالت کلی سه روش متداول برای استخراج کلمات کلیدی وجود دارد:

- **روش TF-IDF[۱۶]:** این روش میزان تکرار یک کلمه در یک مستند را درمقابل تعداد تکرار آن در کل مستندات در نظر می‌گیرد.
- **روش یادگیری ماشین:** در این روش بامجموعه‌ای از سندهای آموزشی و کلمات‌کلیدی مشخص برای آنها، فرآیند استخراج کلمات‌کلیدی به عنوان یک مساله طبقه‌بندی

[۱۶] Term Frequency Inversr Document Frequency

مدلسازی می‌شود.کلمات بر اساس مشخصه‌هایی که دارا هستند، به کلمات‌کلیدی وکلمات غیرکلیدی طبقه‌بندی می‌شوند.

- **ترکیب روش‌های تحلیل آماری و زبان‌شناختی:** علت ترکیب این دو روش این است که اطلاعات آماری بدون پالایش زبان‌شناختی نتایج مفیدی را در اختیار قرار نمی‌دهد. براساس بررسی‌های انجام شده و مقایسه‌ی بین روش‌ها ، نشان داده‌شده که روش آخر یعنی تحلیل آماری و اطلاعات زبان‌شناختی نتایج بهتری را در استخراج کلمات‌کلیدی بدست می‌آورد[۲۵]. راه حلی که ما در این پایان‌نامه استفاده کرده‌ایم.

۷. شناسایی کلمات غیر فارسی موجود در متن:

کلمات غیر فارسی موجود در متن را می‌توان به عنوان یک کلمه کلیدی در نظر گرفت و یا آنها را با فرم فارسی معادل جایگزین نمود؛ به عنوان مثال، در جمله: «Unix یک سیستم عامل است» ممکن است بخواهیم Unix را یک کلمه کلیدی جدا در نظر بگیریم و یا آن را با «یونیکس» جایگزین کنیم.

۸. ریشه‌یابی:

کلمات استخراج شده از جملات باید ریشه‌یابی شده و با ریشه خود جایگزین شوند. یافتن ریشه کلمات و به خصوص ریشه افعال، یکی از چالش‌های زبان فارسی در این وظیفه است. ریشه‌یابی به فرآیند کاهش دادن لغات به ریشه‌های آنها اطلاق می‌گردد. بنابراین "computer"، "compute" و "computing" به "compute" که ریشه‌ی اصلی است، کاهش می‌یابند. معمولا ریشه‌یابی لغات بر اساس قواعد ساخت‌واژه‌ای و سپس حذف پسوندها می‌باشد. تاکنون روش مؤثری برای حذف پیشوندها ارائه نشده است. تمامی سیستمای بازیابی اطلاعات نوع یکسانی از "ریشه‌یاب" را مورد استفاده قرار نمی‌دهند. در انگلیسی، معروف‌ترین

ریشه‌یاب، الگوریتم ریشه‌یاب "مارتین پورتر" است. ریشه‌یاب پورتر، ریشه‌یاب کاهش دهنده‌ی ادغامی برای زبان انگلیسی است که توسط مارتین پورتر در دانشگاه کمبریج در سال ۱۹۸۰ ارائه شد. در طول سالیان گذشته، بسیاری مزایا و معایب استفاده از ریشه‌یابی را متذکر شده-اند. اما مشکل ریشه‌یابی مخصوصا در زبان فارسی، دقت آن می‌باشد. معروف‌ترین الگوریتم ریشه‌یاب فارسی هم الگوریتم نوشته شده توسط کاظم تقوی می‌باشد[۲۶] که البته دقت بالایی ندارد. تقریبا در اکثر مقالات خلاصه‌سازی، ریشه‌یابی، عملیاتی است که همواره در فاز پیش‌پردازش صورت می‌پذیرد. در زبان فارسی هم مقالاتی برای ریشه‌یابی موجود می‌باشد. همانطور که ذکر گردید، یکی از این مقالات، نوشته آقای کاظم تقوی بوده که از دیاگرام حالت برای تشخیص ریشه استفاده می‌کند.

۹. یافتن ارتباط متون:

در مرحله پیش‌پردازش می‌توان ارتباط بین جملات یا معانی آنها را یافت و تشابه یا تفاوت بین برخی از بخش‌های متن را دریافت. برچسب زنی معنایی یکی از مهم‌ترین نقش‌ها را در پردازش‌های زبانی برعهده دارد. دقت در این ابزار بسیار حائز اهمیت است. این ابزار باید نقش‌های گرامری کلمات در جمله‌ها مانند فعل، فاعل، مفعول مستقیم، مفعول غیر مستقیم و ... را تشخیص دهد. برچسب‌زنی معنایی کلمات هم عملی اساسی برای بسیاری از حوزه‌های دیگر پرازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و شباهت معنایی می‌باشد. در حال حاضر، این ابزار در ارزیابی خلاصه‌سازهای ماشینی به کار گرفته نمی‌شود، در حالی‌که به نظر می‌رسد به کارگیری صحیح این ابزار می‌تواند نتایج حاصل از ارزیابی را بهبود ببخشد. از جمله نمونه‌های انگلیسی این ابزار می‌توان به OpenNlP، Illinois SRL، Swirl و LTHSRL اشاره کرد. این ابزارها از الگوریتم پارسینگ استفاده می‌کنند.

Chunker ابزاری برای تشخیص گروه‌های اسمی، فعلی و ... در یک جمله می‌باشد. جهت تقویت الگوریتم‌های وابسته به SRL لازم است نه تنها نقش‌های کلمات مشخص گردند، بلکه باید وابستگی‌های کلمات به لحاظ نقشی در جمله مشخص گردند. از جمله نمونه‌های انگلیسی این ابزار می‌توان به Illinois Chunker اشاره کرد.

۱۰. تعیین مرز کلمات:

در مرحله پیش‌پردازش، ابتدا باید بتوان جملات را از یکدیگر تفکیک نمود. در زبان فارسی جملات می‌توانند با نقطه، ویرگول، علامت سؤال، علامت تعجب و ... تمام شوند؛ در حالی که کلمات نیز می‌توانند با نقطه یا ویرگول و فاصله و غیره از همدیگر جدا شوند و این نیز چالش دیگری است که باید برای آن راه حلی در نظر گرفت.

آنچه که در مرحله پیش‌پردازش متون فارسی قابل توجه است، چالش‌ها و استثنائات مختلف موجود در زبان فارسی، در انجام هر یک از این وظایف است. همچنین ممکن است با توجه به وظایف در نظر گرفته‌شده برای فاز تحلیل، تنها بخشی از این وظایف در مرحله پیش‌پردازش انجام شوند.

۲-۲-۲- فاز تحلیل

در این فاز، جملات موجود در متن با توجه به اطلاعات به دست‌آمده از فاز پیش‌پردازش، تحلیل و امتیازدهی می‌شوند. این امتیازدهی ممکن است با استفاده از تکنیک‌های آماری (در خلاصه‌سازی استخراجی یا گزینشی) و یا تکنیک‌های پردازش زبان طبیعی و مبتنی بر درک متن (در خلاصه‌سازی معنایی یا چکیده) صورت بگیرد.

چنانچه از تکنیک‌های آماری استفاده شود، خصوصیات کمّی متن، فرکانس کلمات و عبارات، مکان جمله در متن، مکان محلی پاراگراف، کلمات کلیدی موجود در جمله و سایر ویژگی‌های متن، جهت

تحلیل و امتیازدهی، مورد استفاده قرار می‌گیرد. همچنین می‌توان از روش‌های معناگرا همانند: روش‌های موجودیتی ـ معنایی، هم‌مکانی، مبتنی بر گراف، زنجیره لغوی، ساختار کلامی و یا روش‌های ترکیبی استفاده نمود.

۲-۲-۳- فاز انتخاب

در این فاز، با توجه به تحلیل صورت گرفته در فاز دوم، جملات و یا مفاهیم موجود در متن، بر اساس امتیاز و وزن مرتب می‌شوند. سپس جملات و مفاهیم با امتیاز و وزن بیشتر، با توجه به اندازه خلاصه مورد نظر که معمولاً به صورت درصد بیان می‌شود، جهت استفاده در خلاصه‌ی نهایی انتخاب می‌شوند. در این فاز، در حین و یا پس از تولید خلاصه، راهکارهایی اتخاذ می‌شود تا خلاصه تولید شده، یکپارچه و خوانا بوده و حاوی مطالب تکراری نباشد.

۲-۳- عوامل تاثیرگذار در خلاصه‌سازی

محتوای خلاصه بستگی زیادی به ورودی، طبیعت متن، هدف، قصد خواننده و دیگر موارد دارد. عواملی که در خلاصه‌سازی مهم می‌باشد به شرح زیر می‌باشد[۲۷]:

۱. **عوامل ورودی**: عوامل ورودی به عوامل زیر تقسیم می‌شود:

- ✓ **فرم ورودی:** شامل طول متن ورودی و ساختار متن(پاراگرافی، نهاد-گزاره‌ای)است. ساختار متن به طور مستقیم روی پردازش متن تاثیر می‌گذارد. زبان متن(انگلیسی، فرانسه و ...)، نوع متن(متن ادبی،داستانی و ..) در خلاصه‌سازی تاثیر مستقیم دارد.
- ✓ **موضوع متن:** شامل سه مورد عادی، خاص و محدود شده می‌باشد. متن عادی شامل موضوعاتی است که دامنه دانش وسیعی دارند همانند ورزش و باغبانی. متن خاص، متن هایی هستند که بستگی به دانش فرد خواننده دارند. متن محدود شده، متنی است که موضوع خاص مربوط به یک سازمان یا انجمن می باشد.

۲. **عوامل خروجی:** شامل سه عامل محتوا، فرمت و شیوه است.

- ✓ **عامل محتوا:** این عامل مربوط به آن است که خلاصه استخراجی و یا چکیده است.
- ✓ **عامل فرمت:** روان، منسجم و یا غیرمنسجم بودن خلاصه ی متن در عملکرد خلاصه‌سازی تاثیر مستقیم دارد.
- ✓ **عامل شیوه** : انواع مختلف خلاصه‌سازی را بیان می‌دارد. شیوه خلاصه‌سازی خبری و یا آگاهی‌دهنده باشد.

۳. **فاکتورهای هدف:** علت خلاصه‌سازی و مورد استفاده متن چه می‌باشد؛ که به سه زیر عامل شنوندگان، شرایط و نحوه استفاده بستگی دارد.

- ✓ **عامل شنوندگان:** دانش شنوندگان در زمینه‌ی متن مورد خلاصه تا چه حدی می‌باشد. این مساله به طور مستقیم روی نتیجه خلاصه تاثیر می‌گذارد.
- ✓ **عامل شرایط:** خلاصه قرار است در کجا مورد استفاده قرار بگیرد. اگر زمینه‌ی خلاصه در سطح وسیعی شناخته شده باشد جزئیات را حذف می‌نمایند .
- ✓ **عامل کاربرد**: موارد استفاده خلاصه چه می‌باشد. این مساله به طور مستقیم روی نحوه تولید متن خلاصه تاثیر خواهد گذاشت. به عنوان مثال وسیله‌ای برای بازیابی متن‌است، جایگزینی برای متن ورودی می‌باشد و یا اینکه مروری بر متنی است که قبلا خوانده شده است.

۲-۴- چالش‌های خلاصه‌سازی

۲-۴-۱- مشکلات خلاصه‌سازی‌از دید ماشین

برای درک زبان طبیعی به اشکال مختلفی از دانش نیاز است[۲۷] که در زیر به آنها اشاره می‌شود:

۱. **دانش مورفولوژی**: روش ساخت عبارات متنوع و مختلف از روی ریشه کلمات را بیان می-دارد .

۲. **دانش نحوی**: نحوه ساخت جملات را با استفاده از ترکیب کلمات مختلف نشان می‌دهد. دانش ساختاری هر کلمه‌ی موجود در جمله را تعیین می‌نماید.

۳. **دانش معنایی:** معنی هر لغت چه می‌باشد و چطور این معانی ترکیب می‌شوند تا معنی هر جمله تشکیل شود. این دانش، مطالعه معنی جملات را بدون توجه به متن شرح می‌دهد.

۴. **عدم موفقیت در تعریف**: عدم وجود استاندارد مشخص برای نمایش معنی جمله.

۵. **عدم موفقیت در پیاده‌سازی توسط قوانین:** معمولا برای درک زبان طبیعی از سیستمی استفاده می‌شود که نحوه تجزیه جملات، روش تبدیل به فرم انتزاعی جملات و سایر کارها برای رسیدن به معنا را با قوانین نمایش می‌دهد. نقض قوانین به دلیل مسائل مختلف همانند استثنا و یا برخورد با متون جدید و پیش‌بینی نشده، باعث ناتوانی در پیاده‌سازی این قبیل کارها در زمینه درک متن شده‌است.

۶. **نفوذ و تاثیر زمینه متن:** از دیگر دلایل عدم موفقیت، وابستگی معنای واقعی جمله به زمینه سخن است که مثلا از نتایج آن کاربرد یک لغت در متون مختلف است که معنای مختلفی را نسبت به زمینه متن تولید می‌کند.

۷. **ابهام:** وجود انواع ابهام در متون هم مساله‌ی مهم و قابل توجهی می‌باشد.

۸. **تجزیه درست متن به جملات:** تجزیه‌ی نحوی نه تنها باید نشانه‌های فرمت گذاری متن و قواعد نقطه‌گذاری را در نظر بگیرد بلکه کلمه‌های توقف و حروف اختصار را نیز در نظر بگیرد. از کارهایی که می‌توان در تولید متن انجام داد، ترکیب مناسب جملات و استفاده از شیوه‌های آیین نگارش مانند حذف به قرینه لفظی و معنوی است. تغییر معنای جمله نیز از طریق تعویض قسمت‌هایی ازمتن باعث بهبود تاثیر بر شنونده خواهد شد.

۲-۴-۲- چالش‌های زبان فارسی با رویکرد خلاصه‌سازی

برخلاف زبان انگلیسی که در آن هم حروف و هم لغات کاملا متمایز از یکدیگرند، در زبان فارسی پیوستگی میان برخی علائم با لغات وجود دارد و علاوه بر آن تنوع نگارش در کلمات نیز موجود می باشد. ریشه‌یابی فعل که یکی از مراحل مهم پیش‌پردازش متن برای خلاصه‌سازی می‌باشد، در زبان فارسی چالش‌های خاص خود را دارد. به عنوان مثال در یک لغت به هم پیوسته هم بن فعل، شناسه، علامت زمان فعل و حتی شناسه‌های مفعولی می‌توان داشت که کار پردازش لغات را پیچیده‌تر می نماید به طوری که نمی‌توان از دانش، تجربه و نرم‌افزارهای موجود در این زمینه استفاده نمود و تولید نرم‌افزاری که قادر به حل تمامی این پیچیدگی‌ها باشد، فرآیندی زمان‌بر و مستلزم تلاش فراوان می‌باشد. تفاوت‌های ذاتی زبان‌های گسسته‌ای مانند انگلیسی با زبان‌هایی مانند فارسی که با یکدیگر تفاوت‌های بنیادین در قواعد دستوری دارند، منجر به آن شده است که ادعای اعمال تغییرات در ساختار یک نرم‌افزار انگلیسی و به دست آوردن نتایج خوب برای زبان فارسی لزوما امکان‌پذیر نباشد و مستلزم آزمایش‌های فراوان برای اثبات صحت آن خواهد بود.

۲-۴-۲-۱- چالش های ذاتی زبان فارسی

زبان فارسی در ساختار و قاعده با زبان انگلیسی متفاوت می باشد. برخی از مشکلات ذاتی مربوط به متون فارسی در زیر طبقه بندی شده‌است .

۱. نبود نکات گرامری تعریف شده همانند آن چه که در زبان انگلیسی وجود دارد . این مساله درریشه‌یابی و پیش‌پردازش متن تاثیر گذار خواهد بود.
۲. وجود لغات ترکیبی چند جزئی همانند " آب سرد کن".

۳. در زبان فارسی، ضمائر مفعولی و نشانه های جمع به لغات متصل می‌شوند. این مسئله نیز منجر به پیچیدگی تفکیک لغات برای رایانه می‌شود زیرا حروف ضمایر مفعولی و نشانه‌های جمع باتعدادی از وندها و نیز واژگان فارسی مشابهت دارند در نتیجه رایانه نمی‌تواند به سادگی در مورد بخش اصلی لغت قضاوت کند. به عنوان نمونه "ان" در"درختان" نشانه‌ی جمع است در حالی که "ان" در "خوران" نشانه‌ی فاعلی می‌باشد .

۴. در زبان فارسی قواعد تبدیل بن مضارع به بن ماضی و بالعکس به صورت قانون‌مند وجود ندارد، در نتیجه رایانه برای آنکه بتواند لغات حاصل از بن مضارع یا بن ماضی را تشخیص دهد (مثلا برای آنکه بداند" ان "در لغت" خوران" نشانه‌ی جمع است یا نشانه‌ی صفت فاعلی)، هیچ راه مشخصی ندارد.

۵. ابهام ساختاری به نحوی که یک کلمه می‌تواند معانی مختلف داشته‌باشد. به عنوان مثال، شیر۳ معنی متفاوت دارد. شیر حیوان، شیر آب، شیر خوراکی.

۶. عدم وجود قاعده خاص برای تشخیص اسامی و مکان‌های خاص همانند آنچه در زبان انگلیسی موجود می‌باشد.

۷. ابهام در معنی کلمه به علت نبود اطلاعات آوایی همانند"مرد " و" ٬مرد".

۸. عدم وجود دستورالعمل قطعی برای استفاده از نیم فاصله .

۹. استفاده از کلمات انگلیسی به صورت‌های مختلف همانند سورس و source.

۱۰. استفاده از"ا" و "آ" به جای یکدیگر همانند"فرایند" و "فرآیند".

۱۱. تنوع استفاده از "می"چسبان و غیرچسبان همانند کلمات "می‌تواند " و " میتواند".

۱۲. تنوع نحوه به کاربردن"ها " چسبان و غیرچسبان همانند" آنها"و" آن‌ها".

۱۳. تنوع نگارش"ی" اضافه در کلمات مختوم به"ه" همانند"خانه سبز" و "خانه‌ی سبز".

۱۴. ناآگاهی رایانه از نقش لغت در جمله همانند " آیین‌نامه نوشتن" و " آیین نامه رانندگی". از آنجا که در این ترکیبات، تمام کلمات به تنهایی معنی دارند، رایانه قادر به تعیین مرز لغات نمی‌باشد .

۱۵. وجود اشتباهات نگارشی در مستندات فارسی همانند نگارش با کاراکترهای متفاوت همانند آئین نامه و آیین نامه، به کار بردن همزه به صورت های مختلف، و نیاز به یکسان‌سازی این موارد برای پیش‌پردازش متن.

۱۶. ابهام در دستور خط زبان فارسی: پیوسته‌نویسی کلمات مرکب ترکیب شده با پسوند.

۱۷. فقدان هر نوع منابع داده‌ای مانند یک گرامر کامل، یک مجموعه از جملات تجزیه شده و یا آمارهای ارزشمند از کاربرد لغات در جملات مختلف و جایگاه‌های مختلف در جمله.

عدم در نظر گرفتن مسائل نگارشی بیان شده باعث می‌شود تا تفکیک لغات غیرضروی و بی-اهمیت از لغات ضروری و مفید به آسانی انجام پذیر نباشد، ریشه لغات به درستی تشخیص داده نشود و یا لغات همسان، غیرهمسان تشخیص داده‌شود که در مرحله امتیازدهی به لغات، تاثیر خواهد گذاشت. به همین علت، مرحله پیش‌پردازش برای امکان‌پذیر شدن پردازش دقیق متن از اهمیت بالایی برخوردار می‌باشد. در مرحله پیش‌پردازش ابتدا الگوریتم ریشه‌یابی اعمال می‌شود. اگر همانند انگلیسی ابتدا لیست کلمه‌های توقف اعمال شود، این احتمال وجود خواهدداشت که حروف پیشوندی فعل‌ها به عنوان کلمه توقف در نظر گرفته شوند. پس از آن کلمه‌های ترکیبی تشخیص داده خواهند شد. پس از آن، کلمه‌های توقف در مرحله آخر حذف می‌شوند.

در زبان فارسی به علت نبود یک پایگاه داده لغوی همانند وردنت[۱۷] [۱۲]، (در ادامه به توضیح وردنت می‌پردازیم) اعمال کامل خصوصیات زبان‌شناختی برای خلاصه‌سازی غیرممکن می‌باشد. البته فارس‌نت که یک وردنت نیمه اتوماتیک می‌باشد توسط دانشگاه شهید بهشتی ارائه شده است که

[۱۷] WordNet

دارای دوبخش واژه‌نامه‌ی معنایی و آنتولوژی واژه‌ای می‌باشد. آرمایشگاه فناوری وب دانشگاه فردوسی مشهد نیز یک نمونه از این مجموعه با نام فردوس‌نت را تولید کرده‌است که توسط تعدادی از دانشجویان کارشناسی ارشد دانشگاه فردوسی مشهد در حال گسترش است.

از جمله مسائل دیگر در زبان فارسی نبود حروف کوچک و بزرگ همانند زبان انگلیسی برای تشخیص اسامی خاص (همانند نام، روزهفته، ماه و مناطق جغرافیایی) می‌باشد. تشخیص اختصارات موجود در زبان فارسی از دیگر چالش‌های موجود می‌باشد.

۲-۵- نتیجه‌گیری

در این فصل به بررسی خلاصه‌سازی متون فارسی پرداخته‌شد. برای خلاصه‌سازی سه فاز کلی: پیش‌پردازش، تحلیل و انتخاب وجود دارد. در مرحله پیش‌پردازش، عملیات اولیه روی متن ورودی انجام می‌شود و متن جهت انجام پردازش‌های فاز تحلیل آماده می‌شود. در فاز تحلیل، جمله‌ها بر اساس معیارهای مختلف معنایی و یا آماری امتیازدهی می‌شوند و نهایتاً در فاز انتخاب، تعدادی از جملات با امتیاز بیشتر کاندیدا می‌شوند و سپس، به عنوان خلاصه نمایش داده می‌شوند. در این فصل همچنین چالش‌ها و مشکلات خلاصه‌سازی متون فارسی بیان شده‌است. کارهایی که تا کنون در زمینه‌ی خلاصه‌سازی متون فارسی انجام شده‌است بیشتر بر پایه‌ی پردازش سطحی متن انجام پذیرفته است و مناسب متون خبری می‌باشد و در زمینه‌ی پیش‌پردازش متون به ایجاد پایگاه داده غیر جامع از کلمه‌های توقف اکتفا شده است.

فصل سوم

پیش پردازش و تولید منابع زبانی روش پیشنهادی

۳-۱- مقدمه

هسته مرکزی هر سیستم پردازش زبان طبیعی تجزیه‌کننده[۱۸] آن است . تجزیه‌کننده آن قسمت از برنامه است که هر جمله را کلمه به کلمه خوانده و تصمیم می‌گیرد که هر کلمه چه معنی می‌دهد. همان‌طور که گفته شد، در مرحله‌ی پیش‌پردازش متن ورودی به ساختاری قابل پردازش برای مراحل بعد تبدیل می‌شود. مهم‌ترین فرآیندهای پیش‌پردازشی مورد استفاده در خلاصه‌سازی عبارتند از: یکسان‌سازی نگارشی، مشخص کردن مرز جمله‌ها و واژه‌ها، ریشه‌یابی و حذف واژهای غیرمهم، استخراج کلمات‌کلیدی. در ادامه‌ی این بخش، هریک از آنها به طور مجزا مورد بررسی قرار خواهد گرفت.

۳-۲- یکسان‌سازی نگارشی

گاهی نویسه‌های به کار رفته در دو واژه‌ی یکسان، با هم متفاوت هستند؛ این باعث می‌شود که هنگام شمردن واژه‌ها، دو واژه‌ی یکسان با املاهای متفاوت به عنوان دو واژه‌ی مختلف در نظر گرفته شوند.

برای جلوگیری از بروز این مشکل، نیازمند یک‌دست‌سازی پیکره‌ی متنی هستیم. به عنوان مثال، پیشوند«می» و پسوند«ها» در ابتدا و انتهای واژه‌ها، ممکن است به سه صورت مختلف زیر دیده شوند:

چسبان	جدا با فاصله	جدا بدون فاصله
کتابها	کتاب ها	کتاب‌ها
میرود	می رود	می‌رود

[۱۸] Parser

و یا واژه‌های «مسئول»، «مجموعه« و« پاییز» به صورت‌های زیر در پیکره‌ی متنی دیده می‌شوند:

مسئول	مسؤول	مسوول
مجموعه‌ی	مجموعه	مجموعه
پاییز	پائیز	

به عنوان یک مثال از یکسان‌سازی، در مورد پسوند«ها» می‌توان شکل‌های «جدا با فاصله» و«جدا بدون فاصله» را نادیده‌گرفت و با مشاهده‌ی هر یک از این دو، آن را به شکل چسبان در آورد. در [۳۰]، یک مجموعه تبدیل‌های نسبتا جامع و موثر برای یکسان‌سازی متون فارسی ارائه شده است. تجربه نشان می‌دهد که با استفاده از تبدیل‌های زیر، می‌توان پیکره‌ی متنی را به خوبی یکسان‌سازی کرد:

- تبدیل نویسه‌های «ی» و «ک» عربی به نوع فارسی آن.
- تبدیل نویسه‌های «ؤ» به «و»،«ئ» به «ی» و«أ» به «ا».
- تبدیل نویسه‌های «ۀ» به «ه» در آخر واژه‌ها.
- حذف «ی» از آخر واژه‌هایی مانند «خانه‌ی».
- حذف پسوندهای«تر» و «ترین» به آخر واژه‌ها.
- حذف فاصله بعد از پیشوند«بر» در واژه‌هایی مانند «برمی‌گردد».
- حذف شناسه‌ی «ء» در آخر بعضی واژه‌ها مانند «شهداء».
- حذف پیشوندهای «می»، «نمی»، «بی»، «درمی»، «برمی» از ابتدای واژه‌ها.
- حذف پسوندهای «ها»، «های»، «هایی»،«هایم»،«هایت»، «هایش»، «هایمان»، «هایشان»، «هایتان» از انتهای واژه‌ها.
- چسباندن پیشوند«به» به واژه‌هایی مانند«به‌ندرت» به ابتدای واژه‌ها.

- چسباندن پیشوند«هم» به واژه‌هایی مانند«هم‌چنین» به ابتدای واژه‌ها.

۳-۳- تعیین مرز جمله‌ها و واژه‌ها

تمام روش‌های خلاصه‌سازی نیازمند شناسایی مرز واژه‌ها و جمله‌ها برای جداسازی آنها از هم هستند. [۳۱] علاوه بر بررسی روش‌های مختلف تعیین مرز واژه‌ها وجملات، مشکلات هر یک از آنها را مورد بررسی قرارداده است. [۳۲] روشی برای شناسایی انتهای واژه‌ها معرفی کرده است که امکان جداسازی واژه‌ها را فراهم می‌آورد. در[۳۳] نیز یک روش آماری برای تعیین کسره اضافه معرفی شده است که با استفاده از آن می‌توان مرز گروه‌های اسمی را به طور دقیق‌تری شناسایی نمود.

در بیشتر مواقع، تعیین مرز جمله‌ها از طریق بررسی علائم جداکننده انجام می‌شود. در این پایان-نامه، شناسایی واژه‌ها و جمله‌ها با بررسی علائم نامبرده انجام گرفته است. علائمی که برای تعیین مرز جمله از آن‌ها استفاده می‌شود، عبارتند از: ".","؟"،"? "،":"،"!". باید توجه داشت که جستجو برای یافتن این علائم به تنهایی کافی نیست و در بسیاری از موارد مشکلاتی را به همراه دارد. به عنوان مثال، در زبان فارسی بعضی از واژه‌های اختصاری به صورت چند حرف که با نقطه از هم جدا شده‌اند، ظاهر می‌شوند(مانند«ه.ق.» که مخفف «هجری قمری» است.) در صورتی که تعداد این حالات مشکل-ساز در پیکره زیاد باشد، باید از روش‌های پیچیده‌تری استفاده شود، که قادر به شناسایی این موارد باشند. برای حل این مشکل از شمارش حروف وکلمه استفاده شده است؛ از آنجایی که کوچکترین جمله شامل دوکلمه است وکوچکترین کلمه از سه حرف تشکیل می‌شود (به غیر حروف اضافه مانند از، به) پس می‌توان واژه‌های اختصاری را مشخص کرد و به عنوان جمله در نظر نگرفت؛ از این رو فقط علائم در متن بررسی نشده‌اند بلکه اگر کلمات بین این علائم بیشتراز یک حد آستانه بود(که ما در اینجا دو درنظر گرفته‌ایم) به عنوان جمله در آرایه قرار می‌گیرند.

شناسایی واژه‌ها نیز از طریق بررسی علائم قابل انجام است. این علائم عبارتند از: فضای خالی، علامت خط جدید، ", "، "،"، ">"، "<"، "]"، "["، "_"، "/"، "–"، "؛".

۳-۴- ریشه‌یابی

ریشه‌یابی، کاهش یک کلمه به شکل عمومی است. این شکل عمومی باید برای همه‌ی کلمه‌های هم‌ریشه یکسان باشد. روش‌های ریشه‌یابی را می‌توان به دو نوع اصلی دسته‌بندی نمود[۳۴]: مبتنی بر شبکه‌های واژه و مبتنی بر ریخت‌شناسی.

در روش‌های مبتنی برشبکه‌های واژه، از ارتباط واژه‌ها در یک شبکه معنایی استفاده می‌شود. واژه-هایی که دارای ریشه یکسان هستند در یک خوشه قرار می‌گیرند. این شبکه‌ها ساختاری گراف مانند دارند. این دسته از روش‌ها نیازمند بهره‌گیری گسترده از دانش انسانی هستند، به همین دلیل ایجاد و نگه‌داری آن‌ها بسیار پرهزینه است. هنوز الگوریتم‌هایی که بتواند به صورت خودکار این شبکه‌ها را به گونه‌ای قابل اطمینان ایجاد کنند، توسعه نیافته است. به دلیل کارایی بالای این دسته از روش‌ها، به طور گسترده در سامانه‌های داده‌کاوی و متن‌کاوی استفاده می‌شود.

رویکرد دوم استفاده از ریخت‌شناسی برای ریشه‌یابی است. این دسته از روش‌ها با بررسی ساخت واژه‌ها، ریشه‌ی کلمه را پیدا می‌کنند[۳۵]. این روش نیازی به معلوم بودن ریشه‌ی واژه‌ها ندارد و تنها نیازمند آگاهی از قواعد ساخت‌واژی است. از آنجا که در خلاصه‌سازی بیشتر ریشه‌یابی تصریفی مورد نظر است و قواعد تصریفی اندک هستند، این دسته از روش‌ها به سادگی قابل پیاده‌سازی هستند. نمونه‌ای از افعال و واژه‌ها در جدول(۳-۱) آمده است.

جدول (۳-۱) واژه‌های غیر مهم دسته‌ی دوم(شامل فعلا ها در حالت‌های تصریفی)[۳۵]

آمد	آیی	بخواهم	بگویند	بیایی	توانسته	داریم	کنیم	گویم	یابد
آمدم	آیید	بخواهی	بگوید	بیایند	خواستم	دارند	کنند	گویند	یابیم
آمدن	آییم	بخواهد	بگیر	بیاند	خواستی	دارید	کنید	گویید	یابند
آمدند	باش	بخواهیم	بگیر	بیایم	خواست	داشتم	گرفتم	گیرم	یابید
آمده	باشم	بخواهند	بگیرد	بیاید	خواستیم	داشتی	گرفتی	گیرد	یافتم
آمدی	باشی	بخواهید	بگیری	بیاورم	خواستند	داشت	گرفتیم	گیری	یافتی
آمدید	باشد	بکن	بگیریم	بیاوری	خواستید	داشتیم	گرفتند	گیریم	یافت
آمدیم	باشیم	بکنم	بگیرند	بیاورد	خواسته	داشتند	گرفتید	گیرند	یافتیم
آوردم	باشند	بکنی	بگیرد	بیاوردیم	خواهم	داشتید	گفتم	گیرید	یافتند
آوردی	باشید	بکند	بودم	بیاوردند	خواهی	کردم	گفتی	هستم	یافتید
آورد	بتوان	بکنیم	بودی	بیاورید	خواهد	کردی	گفت	هستی	یافته
آوردیم	بتوانم	بکنند	بود	تواند	خواهیم	کرد	گفتیم	هست	
آوردند	بتوانی	بکنید	بودند	توانستم	خواهند	کردیم	گفتند	هستیم	
آوردید	بتواند	بگو	بودیم	توانست	خواهید	کردند	گفتید	هستند	
آیم	بتوانیم	بگویم	بودید	توانستی	دارم	کردید	گویم	هستید	
آید	بتوانند	بگویی	بوده	توانستیم	دارد	کنم	گوی	است	
آیم	بتوانید	بگویند	بیا	توانستند	داری	کنی	گویم	یابم	
آیند	بخواه	بگوییم	بیایم	توانستید	دارند	کند	گویند	یابی	

روشی که در این پایان نامه استفاده شده است یک روش مبتنی بر حذف پسوندها و پشوندها می-باشد. این روش مشابه به روش پورتر در زبان انگلیسی می‌باشد، با این تفاوت که روش پورتر فقط حذف پسوندها را انجام می‌دهد.

۳-۵- حذف کلمات و واژه‌های غیرمهم

در این مرحله می‌خواهیم کلماتی را که در محتوای اصلی متن تاثیری ندارند، حذف نماییم. حذف کلمات اضافی نظیر حروف اضافه، بسیاری از قیدها و صفات، برخی از افعال، حروف ربط و غیره عموما در مضمون کلی متن تاثیر منفی نمی‌گذارند، بلکه باعث خلاصه‌سازی متن می‌شوند. از آنجایی که لیست کاملی از این کلمات در دسترس نبود خودمان لیستی از کلمات توقف جمع‌آوری کردیم که این لیست شامل ۶۰۰ کلمه است که در آن حروف اضافه، قیدهایی که زیاد به کار می‌روند و افعال اسنادی در زمان‌های مختلف و با شناسه‌های گوناگون آمده‌اند. (پرتکرارترین کلمات در متون خبری پایگاه‌داده همشهری استخراج شده‌اند)؛ در روش پیشنهادی سیستمی ارائه شده است که به کاربر اجازه می‌دهد کلماتی که در متن جز کلمات مهم نیستند را حذف نماید.

چند نکته در مورد کلمات غیر مهم حائز اهمیت است:

۱. شناسه‌ی افعال ماضی بعید در آن وجود دارد، دلیل آن این است که نمی‌توان آن‌ها را با فعل سرهم نوشت و بنابراین سیستم آن‌ها را به صورت دو کلمه در نظر می‌گیرد. با وارد کردن آنها در کلمات غیرمهم این موارد که نقشی در معنا ندارند در متن در نظر گرفته نمی‌شوند. علامت جمع "ها" و دو مورد "ها" و "هایی" را اضافه کرده‌ایم تا لزومی به نوشتن کلمات جمع با علامت "ها" به صورت سرهم نباشد.
۲. "می" که پیشوند افعال ماضی استمراری و مضارع اخباری است به دلیل این‌که بتوان آن را جدا از فعل نوشت ولی دو کلمه‌ی جدا در نظر گرفته نشوند در جدول کلمات غیر مهم وارد شده‌است.
۳. "تر" و دو حالت "ترین" و"تری" آن که نشان دهنده‌ی عالی وتفضیلی بودن صفت هستند، چون در برخی صفات به صورت جدا آورده می شوند در جدول کلمات غیر مهم قرار داده‌ایم

جدول لیستی از این واژه‌ها را نشان می‌دهد.

جدول (۳-۲) واژه‌های غیر مهم(شامل اسم، صفت وحروف و وندها)

در	نیز	برای	یا	را	بلکه
به	تا	ها	دو	های	شاید
از	ما	آن	آنها	و	وقتی
که	باید	وی	اما	ممکن	اندک
این	اند	یک	دیگر	هر	کمی
با	هم	خود	اگر	ای	بسیار
شما	همچنین	بر	چه	اندکی	زیاد

۳-۶- استخراج کلمات‌کلیدی[۱۹]

از آنجایی که تعداد مستندات الکترونیکی فارسی رو به رشد است، به کارگیری روش‌های کارآمد جهت بازیابی اطلاعات فارسی بسیار مورد اهمیت است. کلمات کلیدی مجموعه‌ای از لغات مهم در یک سند هستند که توصیفی از محتوای مستند را فراهم می‌آورند و برای اهداف مختلفی قابل استفاده هستند. به عنوان مثال در مقالات علمی برای اینکه خواننده بتواند درک مناسبی از مقاله داشته باشد، کلمات‌کلیدی مقاله عنوان می‌شوند. استخراج کلمات‌کلیدی از مستندات، یک عملیات مهم در فرآیندهایی مانند خوشه‌بندی، طبقه بندی، خلاصه‌سازی، استخراج اطلاعات معنایی می‌باشد و همچنین در موتورهای جستجو به منظور بازگرداندن نتایج دقیق‌تر در زمان کوتاه‌تر مورد استفاده قرار می‌گیرد. در مجموع کلمات‌کلیدی ابزار مفیدی برای جستجوی حجم زیادی از مستندات در زمان

[۱۹] keyword

کوتاه هستند، به همین دلیل به عنوان یکی از مراحل مهم پیش‌پردازش در خلاصه‌سازی محسوب می‌شود. استخراج کلمات‌کلیدی به طور دستی فرآیندی بسیار دشوار و زمان‌بر است. بنابراین نیاز به یک فرآیند خودکار است که کلمات‌کلیدی را از مستندات استخراج نماید[۳۶]. روش به کار رفته در این پایان‌نامه استفاده از استخراج سریع خودکار کلمات کلیدی(RAKE[۲۰]) می‌باشد. هدف از استفاده-ی این روش استخراج واژه‌های کلیدی با حداکثر دقت برای اسناد خبری است.

۳-۶-۱- استخراج کلمات کلیدی کاندیدا

RAKE یک روش بدون مربی [۲۱]، مستقل از دامنه و مستقل از زبان، برای استخراج کلمات‌کلیدی می‌باشد. این روش برای استخراج کلمات‌کلیدی در اسناد، تک‌سندی و چندسندی قابل استفاده است. در روش حاضر با معانی واژه‌ها کاری نداریم و پیوند و ترکیب کلمات با یکدیگر مورد بحث قرار می‌گیرد. البته، از این خصوصیت، که فاصله‌ی کمتر واژه‌ها از یکدیگر بر جنبه خاصی از یک موضوع دلالت می‌کند، استفاده شده است. بنابراین، رخداد بالای کلمات مختلف در مجاورت یکدیگر، نشان دهنده‌ی احتمال مرتبط بودن این واژه‌ها با محتوای مقاله است. رتبه اهمیت مجاورت می‌تواند بر اساس خصوصیات زبانی نوشته‌ها متفاوت باشد. به صورت فیزیکی، واژه‌هایی که برای بیان تصورهای متجانس ذهنی به کار می‌روند از لحاظ مکانی در موقعیت نزدیکتری نسبت به هم قرار دارند.

پارامترهای ورودی برای استخراج خودکار کلمات‌کلیدی شامل لیست کلمات غیرمهم، مجموعه‌ای از جداکننده‌های کلمه و عبارت می‌باشد. RAKE برای استخراج کلمات‌کلیدی از متن، ابتدا متن ورودی را به کلمات کلیدی کاندیدا تجزیه[۲۲] می‌کند. برای این کار متن سند توسط جداکننده‌های جمله مشخص شده ودر آرایه‌ای قرار می‌گیرد، سپس توسط جداکننده‌های کلمه به آرایه‌ای از کلمات

[۲۰] Rapid automatic keyword extraction

[۲۱] Unsupervised

[۲۲] Parsing

تبدیل می‌شود بعد از آن کلمات توقف[۲۳] را از متن حذف می‌کنیم. کلماتی که باقی مانده‌اند به عنوان کلمات کلیدی کاندیدا شناخته می‌شوند که دنباله‌ای از کلمات محتوای متن هستند که در متن اصلی آمده‌اند. نمونه‌ای از کلمات کاندیدا برای خبر زیر مشخص شده‌است:

جدول۳-۳ خبری از پیکره‌ی همشهری

عنوان خبرشماره یک : دریافتی مستمری بگیران تامین اجتماعی ۱۷ درصد افزایش می یابد
مدیر عامل سازمان تامین اجتماعی بودجه پیشنهادی ۲۰۰۰ میلیارد تومانی سال ۱۳۷۹ این سازمان را جهت بررسی و تصویب به شورای عالی تامین اجتماعی ارائه کرد. رقم اختصاص یافته برای مصارف سازمان نیز برابر منابع در نظر گرفته شده است که به این ترتیب کل منابع و مصارف این سازمان در سال ۱۳۷۹ رقمی معادل ۳۹۹۶ میلیارد تومان خواهد بود. وی همچنین با اشاره به این نکته که در سال مالی ۱۳۷۹ شرایط جدیدی برای برقراری مستمری بازماندگان در نظر گرفته شده است. خاطر نشان کرد: تا به حال سازمان تنها برای بازماندگان بیمه شدگانی که بین ۱۰ تا ۲۰ سال سابقه پرداخت حق بیمه داشتند، پرداخت غرامت مقطوع در نظرمی گرفت و بازماندگان افرادی که زیر این میزان سابقه داشتند از این پرداخت محروم بودند، اما با توجه به این که خانواده های بازماندگان از نظر مالی در شرایط نامساعد اقتصادی قرار دارند، تصمیم گرفته شد که برای بازماندگان آن دسته از بیمه شدگان متوفا که بین ۱ تا ۱۰ سال سابقه پرداخت حق بیمه دارند نسبت به پرداخت غرامت مقطوع به آنان اقدام کند که متعاقبا " آئین نامه" آن تدوین و اعلام خواهد شد.

همان‌طور که در شکل(۳-۱) آورده شده کلمات کاندید به ترتیب حروف الفبا نمایش داده می-شوند(البته در اینجا ما تعدادی از آنها را آورده‌ایم). در سیستم ارائه شده این امکان وجود دارد که کاربر هر کلمه را که لازم است به لیست کلمات توقف اضافه شود با کلیک روی دکمه StopWord به لیست اضافه کند.

[۲۳] Stop word

شکل ۳-۱ نمایش کلمات کاندیدا از متن خبر شماره یک

۳-۶-۲- محاسبه امتیاز کلمات کلیدی

بعد از شناسایی کلمات کلیدی کاندید، گراف هم‌رخداد[۲۴] کلمات رسم می‌شود. حال ما می‌توانیم کلمات هم‌رخداد در متن را بدون استفاده از سایز پنجره مشخص پیداکنیم (روش‌های موجود با استفاده از سایز پنجره کلمات هم‌رخداد را شناسایی می‌کنند). برای اندازه گیری رتبه اهمیت یک کلمه، تعداد رخداد کلمه[۲۵] و مجاورت مکانی واژگان با یکدیگر [۲۶] در نظر گرفته شده است. دلیل استفاده از تعداد برای اندازه‌گیری رتبه اهمیت، بر این باور استوار است که نویسنده معمولا از واژگان معینی برای پیشبرد بحث یا تشریح دقیق جنبه‌های مختلف موضوع مورد نظر استفاده وآنها را تکرار

[۲۴] Co-occurrence
[۲۵] Frequence
[۲۶] Degree

می‌کند. تعداد رخداد هر واژه می‌تواند به عنوان عامل تعیین درجه اهمیت واژگان مورد استفاده قرار گیرد. در بیشتر اوقات، واژه‌های معینی وجود دارند که با یکدیگر یک گروه را تشکیل می‌دهند، باید به این واژه‌ها رتبه اهمیت بالاتری اختصاص داد. در این میان، بعضی کلمات برای نشان دادن میزان ارتباط واژه‌ها با یکدیگر و گروه‌بندی آن‌ها به کار می‌روند؛ به این واژه‌های رابط، نمره‌ای اختصاص داده نمی‌شود و این کلمات به عنوان کلمات توقف از متن حذف می‌شوند. شکل(۳-۲) نمونه‌ای از گراف هم‌رخداد بین کلمات و امتیازی که به هر کلمه تعلق گرفته است را نشان می‌دهد نحوه محاسبه امتیاز کلمات را در ادامه شرح می‌دهیم.

	اجتماعي	اختصاص	اقتصادي	بازماندگان	بودجه	بيمه	تدوين	تصويب	خانواده	سازمان	شوراي	عالي	عامل	مالي	متوفا	محروم	مدير	مستمري	مقطوع	ميليارد	نامساعد	deg	freq	score	
اجتماعي	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	1	اجتماعي
اختصاص	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	اختصاص
اقتصادي	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	2	اقتصادي
بازماندگان	0	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	7	5	1	بازماندگان
بودجه	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	2	بودجه
بيمه	0	0	0	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	1	بيمه
تدوين	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	تدوين
تصويب	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	تصويب
خانواده	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	خانواده
سازمان	0	0	0	0	0	0	0	0	0	5	0	0	1	0	0	0	0	0	0	0	0	6	5	1	سازمان
شوراي	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2	1	2	شوراي
عالي	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2	1	2	عالي
عامل	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	3	1	3	عامل
مالي	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	2	1	مالي
متوفا	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	متوفا
محروم	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	محروم
مدير	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	2	1	2	مدير
مستمري	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	1	2	مستمري
مقطوع	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	2	1	مقطوع
ميليارد	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	2	1	ميليارد
نامساعد	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	2	نامساعد

شکل ۳-۲ نمایش گراف هم‌رخداد کلمات متن خبر شماره یک

این گراف نشان می‌دهد هر کلمه چند بار در متن تکرار شده و هر کلمه با چه کلمات دیگری آمده است. مثلا سطر اول کلمه‌ی اجتماعی دو بار در متن خبر تکرار شده و همراه کلمه‌ی بودجه هم در متن خبر آمده است.

استانداردهایی که برای محاسبه امتیاز هر کلمه در نظر می‌گیریم عبارتند از:

۱. فرکانس کلمه (freq(w)): کلمه w به تنهایی چند بار در متن تکرارشده‌است.

۲. درجه کلمه(deg(w)): درجه کلمه در گراف.

۳. نسبت درجه به فرکانس: (deg(w)/freq(w))

نسبت درجه به فرکانس امتیاز نهایی هر کلمه را مشخص می‌کند.

۳-۶-۳- کلمات کلیدی مجاور

واژه‌های کلیدی در اکثر مقالات و اسناد به صورت یک واژه‌ای، دو واژه‌ای و یا بیش‌تر مشخص می‌شوند. مشکلی که در بعضی روش‌های استخراج کلمه‌های‌کلیدی وجود دارد این است که، واژه‌های کلیدی یک واژه‌ای استخراج می‌کنند و واژه‌های کلیدی دو واژه‌ای یا بیشتر را استخراج نمی‌کنند. ما در این پایان‌نامه با استفاده از محاسبه‌ی رخداد همزمان پسین وپیشین، در جملات مختلف سند، واژه‌های کلیدی بیش از یک واژه‌ای را نیز استخراج می‌کنیم.

RAKE در استخراج کلمات‌کلیدی کلمات و حروف اضافه بین کلمات را حذف می‌کند، درنتیجه کلمات کلیدی شامل کلمات اضافه نیستند. برای پیدا کردن کلمات هم‌رخداد، RAKE جفت‌هایی از کلمات‌کلیدی که مجاور هم هستند و حداقل دو بار در متن تکرار شده‌اند را جستجو می‌کند. برای محاسبه امتیاز کلمه کلیدی جدید، امتیاز تک تک کلمات را با هم جمع می‌کنیم .

۳-۶-۴- استخراج کلمات کلیدی

بعد از امتیاز دادن به کلمات کلیدی کاندیدا، T تا ازکلمات کاندیدی که بالاترین امتیاز را دارند انتخاب می‌شوند.T یک سوم تعداد کلمات گراف در نظر گرفته شده است [۳۷]. در نمونه داده شده، متن شامل ۲۱کلمه می‌باشد بنابراین ۷ تا کلمه‌کلیدی داریم، (T=۷). جدول زیر کلمات کلیدی که توسط RAKE استخراج شده و کلمات‌کلیدی که توسط انسان از متن استخراج شده است را نشان داده شده‌است.

جدول ۳-۴ کلمات کلیدی استخراج شده توسط سیستم و انسان

استخراج شده توسط RAKE	استخراج شده توسط انسان
مدیر عامل	مدیر عامل
سازمان تامین اجتماعی	سازمان تامین اجتماعی
خانواده‌های بازماندگان	خانواده‌های بازماندگان
شرایط نامساعد اقتصادی	شرایط نامساعد اقتصادی
بازماندگان بیمه شدگان	-
مصارف سازمان	-
شورای عالی تامین اجتماعی	-
-	بیمه‌شدگان

همان طور که در جدول می‌بینیم سیستم استخراج خودکار کلمات کلیدی فقط یکی از کلمات کلیدی را تشخیص نداده است و سه کلمه کلیدی هم‌جوار در متن خبر شماره یک را انتخاب کرده‌است که توسط انسان انتخاب نشده است. که می‌تواند در انتخاب جملات خلاصه مهم باشد.

فصل چهارم

معرفی روش پیشنهادی

(تولید خلاصه با استفاده از استخراج رویدادها از متن خبر)

۴-۱- مقدمه

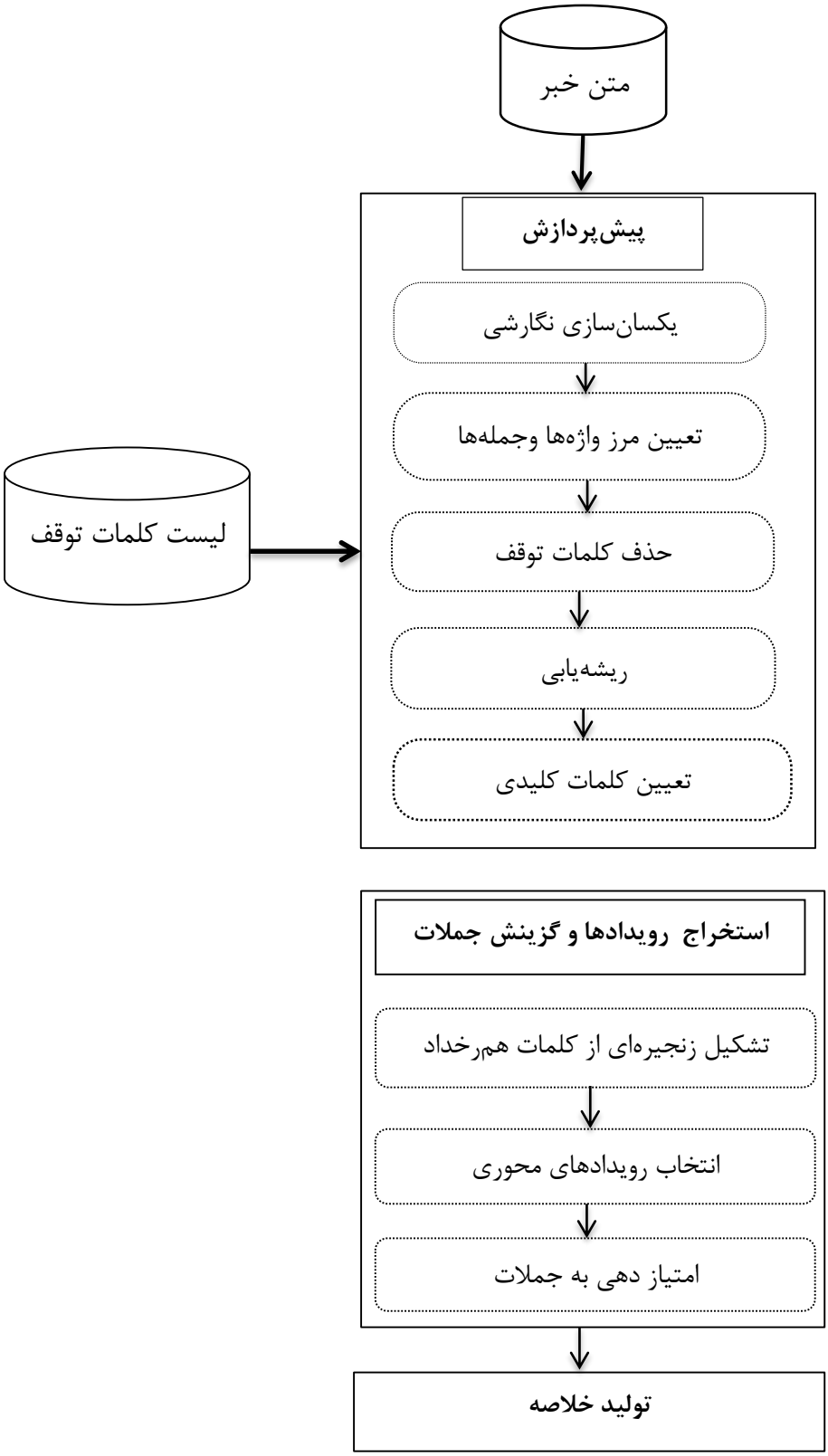
روش پیشنهادی، یک روش مبتنی بر استخراج رویداد از متن خبری است. استخراج رویداد راه‌حلی برای دو چالش گزینش اطلاعات و حذف افزونگی است. در این روش، فرض ما بر این است که هر جمله نشان دهنده‌ی یک رویداد است و کلمات‌کلیدی استخراج شده از هر جمله نماینده رویدادهای آن جمله هستند. بعداز پیدا کردن رویدادهای جمله‌ها، رویدادهایی که ارتباط بیش‌تری با عنوان خبرداشته باشند به عنوان رویدادهای محوری انتخاب می‌شوند و جملات شامل این رویدادها، در جملات خلاصه قرار می‌گیرند. گزینش یک جمله از هر رویداد منجر به جلوگیری از آمدن چند جمله شبیه به‌هم درخلاصه می‌شود.

۴-۲- روش پیشنهادی

روش پیشنهادی شامل سه مرحله است:

- شناسایی پیش پردازش
- استخراج رویدادها از متن خبر وگزینش
- تولید خلاصه

شکل ۴-۱، معماری کلی روش پیشنهادی را نشان می‌دهد. این شکل، فرآیند و مراحل روش پیشنهادی را همراه با منابعی که در هر مرحله استفاده می‌شوند، نشان می‌دهد. سه مرحله اصلی روش پیشنهادی در شکل (۴-۱)مشاهده می‌شود. در مورد مرحله‌ی پیش‌پردازش در فصل ۳ توضیحات کاملی ارائه شد. دو مرحله‌ی اصلی دیگر در ادامه‌ی این فصل مورد بررسی قرار خواهند گرفت.



شکل ۴-۱ معماری کلی روش پیشنهادی

۴-۳- استخراج رویدادها[۲۷] از متن خبر با استفاده از زنجیره‌ی کلمات هم-رخداد(روش پیشنهادی اول)

برای استخراج رویدادها از متن فرض ما بر این است که هر جمله شامل یک رویداد است. برای استخراج رویداد از هر جمله، برای هر جمله زنجیره‌هایی از کلمات هم‌رخداد تشکیل می‌دهیم، هر زنجیره یک رویداد از متن شناخته می‌شود؛ ممکن است که یکسری از کلمات نماینده‌ی خوبی برای شناسایی رویداد نباشند و از آنجایی که عنوان اصلی خبر منظور کلی متن را نشان می‌دهد و گاه آن را فشرده‌ترین خلاصه از متن می‌دانند به همین دلیل برای تشخیص رویدادهای اصلی متن خبر، کلمات هر زنجیره با کلمات‌کلیدی عنوان خبر مقایسه می‌شوند، زنجیره‌هایی که به عنوان خبر شبیه‌تر هستند به عنوان رویدادهای اصلی شناخته می‌شوند و امتیاز بیشتری نسبت به سایر زنجیره‌ها دریافت می-کنند امتیاز هر زنجیره از مجموع امتیازات کلماتی است که در آن زنجیره قرار گرفته‌اند، می‌باشد ولی زنجیره‌ایی که دارای حداقل نصف تعداد کلمات کلیدی عنوان هستند علاوه بر مجموع امتیازات کلمات زنجیره امتیاز W_B هم دریافت می‌کند.

برای تعیین جملاتی که می‌خواهیم در خلاصه قرار بگیرند علاوه بر شناسایی رویدادهای اصلی متن، چند ویژگی از جملات هم بررسی شده است؛ که برای هر ویژگی یک امتیاز در نظر گرفته شده است. امتیاز نهایی جمله، مجموع امتیازاتی است که آن جمله دارا می‌باشد، سپس با تعیین یک حدآستانه (در این تحقیق حدآستانه ۰٫۶در نظر گرفته شده است؛ این مقدار با توجه به مجموعه ارزیابی که شامل ۲۰متن خبری انتخابی ازپیکره‌ی همشهری می‌باشد تعیین شده‌است و امتیاز هر جمله در بازه

[۲۷] Event

صفر تا یک نرمال شده است) رویدادهای اصلی متن انتخاب شده و سپس جمله‌های شامل رویدادهای اصلی تا جایی به خلاصه اضافه می‌شوند تا به طول مورد نظر برسد.

منظور از طول خلاصه، درصد فشرده‌سازی است که کاربر در ابتدا در سیستم وارد کرده است. امتیاز نهایی جملات شامل رویدادهای اصلی با W_E نشان داده‌می‌شود. برای پیدا کردن نزدیک‌ترین رویدادها با عنوان خبر از فاصله همینگ استفاده کرده‌ایم، که در ادامه به توضیح آن می‌پردازیم.

❖ فاصله همینگ

فاصله همینگ دو کلمه، تعداد حروف متناظر نامشابه است[۳۸]. فرمول(۴-۱) روش محاسبه فاصله همینگ را نشان می‌دهد. به عنوان مثال فاصله همینگ دو کلمه «امید» و «نماد»، ۲ است که برای تبدیل آن به فرم نرمال شده در بازه‌ی ۰تا ۱، این فاصله به طول کلمات تقسیم می‌گردد. این معیار میزان تفاوت را نشان می‌دهد جهت محاسبه معیار شباهت، همانند فرمول (۴-۲) ، این مقدار را از ۱کم می‌کنیم بنابراین میزان تشابه همینگ دو کلمه فوق $1 - \frac{2}{4} = 0.5$ است. زمانی که فاصله‌ای همینگ دو کلمه صفر شود دو کلمه مشابه هستند.

$$f_n(i) = 0 \quad if\left(q_t = l_f\right),\ else \quad f_n(i) = 1 \qquad (۱-۴)$$

$$Similarity = 1 - \frac{Hamming\ Distance(A,B)}{\max(|A|,|B|)} \qquad (۲-۴)$$

❖ **محاسبه امتیاز جملات**

• **موقعیت مکانی:** جمله‌های متن خبر از اولین جمله تا جمله‌های میانی بر حسب نزدیکی به ابتدای خبر امتیازدهی می‌شوند به طوری که جمله اول امتیاز ۱را دارد و به همین ترتیب امتیاز جملات کم می‌شود تا جمله‌های میانی کمترین امتیاز را دریافت کنند و سپس امتیاز جمله‌ها از آخرین جمله متن خبر محاسبه می‌شود(جمله آخر امتیازش از جمله اول کمتر ولی از جمله‌های میانی بیشتر است) تا جمله‌های میانی که کمترین امتیاز را دریافت کنند.امتیاز جملات را با W_P نشان داده‌ایم. نمونه‌ای از امتیازدهی به جملات در قسمت پیوست (ب) آورده شده است.

➢ فرمول محاسبه امتیازجمله اول تا جمله $n/2$:

$$P_i = \frac{n-i+1}{n} \qquad (۳-۴)$$

➢ فرمول محاسبه امتیازجمله‌آخر تا جمله $(n/2) + 1$:

$$P_i = \frac{i-3}{\mathrm{n}} \qquad (۴-۴)$$

• **وجود نقل و قول:** چنانچه جمله‌ای در بردارنده‌ی نقل و قول باشد، امتیاز W_T به آن داده می‌شود.

- **وجود اطلاعات عددی:** چنانچه جمله‌ای حاوی اطلاعات عددی باشد، امتیاز W_N به آن داده می‌شود.
- **حذف مثال‌ها**
- **امتیاز مثبت به جملات حاوی علامت"٪":** وجود علامت "٪"در جمله در متون خبری معمولا نشان از آماردهی دارد و بهتر است جملات حاوی این علائم هم امتیاز مثبت گرفته و در خلاصه ظاهر شوند. بنابراین چنانچه جمله‌ای شامل این علامت باشد، امتیاز W_M به آن داده می‌شود.

امتیاز W_B،W_T، W_M برابر یک در نظر گرفته شده‌است.امتیاز پایانی هر جمله، با جمع امتیازات به صورت زیرمحاسبه می‌شود:

$$Score(s) = W_T + W_N + W_P + W_M + W_E + W_B \qquad (۴-۴)$$

۴-۴- تولید خلاصه

تولید خلاصه، آخرین مرحله از فرآیند خلاصه‌سازی است. بعد از محاسبه $score_s$ برای همه جملات متن اصلی، برای جلوگیری از ورود جمله‌های کم‌اهمیت به خلاصه، حدآستانه‌ای در نظر گرفته می‌شود و تنها جمله‌هایی که امتیازشان بیشتر از آن حدآستانه باشد، برای درج در خلاصه گزینش می‌شوند. بعد از نرمال کردن امتیازات هر جمله، وبا توجه به مجموعه‌ی ارزیابی که ۲۰متن از پیکره همشهری می‌باشد، حدآستانه ۰/۶ در نظر گرفته شده است. نحوه قرار گرفتن جملات در خلاصه به همان ترتیبی که درمتن خبر آمده‌اند در خلاصه هم قرار می‌گیرند.

۴-۵- نتیجه‌گیری

یکی از واقعیت‌های پذیرفته شده در حوزه‌ی خلاصه‌سازی خودکار آن است که برای تولید خلاصه-های با کیفیت نیازمند روش‌هایی هستیم که مبتنی بر درک متن باشند. اگرچه پیشرفت‌های کنونی در این حوزه، هنوز فاصله‌ی زیادی با فرآیندهای شناختی سطح بالا مانند برخی فرآیندهای شناختی انسان دارند؛ با این حال ، می‌توان از برخی ویژگی‌های متنی ساده‌تر مانند همبستگی لغوی سود برد که نیاز چندانی به درک کامل متن ندارند. یک نمونه از اطلاعاتی که در بسیاری از کاربردهای متن-کاوی اهمیت فراوانی دارد، وابستگی بین اجزای متنی است. دو دسته‌ی مهم از وابستگی‌های متنی عبارتنداز : همبستگی[۲۸] و ارتباط معنایی[۲۹]. پدیده‌ی همبستگی، معادل با این واقعیت است که بعضی عناصر متنی(مانند واژه‌ها) در کنار هم ظاهر شوند، در حالی که پدیده‌ی ارتباط معنایی بر این حقیقت اشاره داردکه یک ارتباط هوشمندانه بین جملات متن وجود دارد[۳۹].

در این فصل به توضیح روش پیشنهادی پرداختیم که با توجه به همبستگی و ارتباط مکانی کلمات رویدادهای مرتبط با عنوان خبر شناسایی شده‌اند. چند نمونه خلاصه‌ی تولید شده توسط سیستم پیشنهادی، از متن خبری که از پیکره همشهری انتخاب شده‌است توسط سیستم خلاصه شده‌اند که در در پیوست یک آورده‌شده است. در ادامه در فصل پنجم به ارزیابی نتایج می‌پردازیم.

---

[۲۸] Cohesion

[۲۹] Coherence

فصل پنجم

نتایج تجربی

۵-۱- مقدمه

یکی از مشکل‌ترین کارها در حوزه‌ی خلاصه‌سازی، ارزیابی سامانه‌های خلاصه‌سازی است، زیرا معیار دقیقی برای سنجش آن‌ها وجود ندارد. در زبان انگلیسی مجموعه‌ی داده‌ای[۳۰] شامل خلاصه‌های انسانی وجود دارد که می‌توانند به عنوان محک[۳۱] استفاده شوند که از معروف‌ترین آن‌ها می‌توان به DUC اشاره کرد که بسیاری ازمحققان نتایج کار خود را با آن مقایسه می‌کنند اما برای زبان فارسی چنین محک‌هایی وجود ندارد.

۵-۲- ویژگی‌های پایه‌ی ارزیابی خلاصه‌ها

ارزیابی خلاصه‌ها و سیستم‌های خودکار خلاصه‌سازی متن، فرایندهای مشخص و واضحی نیستند. بطور کلی حداقل دو خصوصیت از خلاصه وجود دارد که در هنگام ارزیابی و همچنین در سیستم‌های خلاصه‌سازی، باید مورد سنجش قرار گیرند[۹].

نرخ فشرده‌سازی میزان کوتاه بودن خلاصه نسبت به متن اصلی:

$$CR = \frac{Length\ of\ Summary}{Length\ of\ Full\ Text} \qquad (۱-۵)$$

۳۰ DtaSet

۳۱ Benchmark

و نرخ حفظ(میزان اطلاعاتی که حفظ می‌شوند):

$$RR = \frac{Information\ in\ Summary}{Information\ in\ \ Full\ Text} \qquad (۲-۵)$$

در ارزیابی یک سیستم خلاصه‌سازی باید هر دوی این‌ها مشخص شوند. در بسیاری از موارد، CR برای یک خلاصه که طول آن برابر یک اندازه تعریف شده است، نادیده گرفته می‌شود. بنابراین نرخ حفظ، بیشترین توجه را به خود اختصاص می‌دهد.

به طور کلی، دو رهیافت برای ارزیابی خلاصه‌های ایجاد شده توسط سامانه‌های خلاصه‌سازی خودکار وجود دارد:

- **ارزیابی درونی**[۳۲]: در این روش، کیفیت خلاصه‌ها به صورت مستقیم مشخص می‌شود. این کار می‌تواند به صورت مقایسه خلاصه سیستم با خلاصه‌های مرجع و یا با نظر مستقیم چند داور انسانی انجام گیرد. اکثر روش‌های ارزیابی خلاصه، ارزیابی مستقیم هستند. در ارزیابی درونی، توجه اصلی بر روی پیوستگی و اطلاع‌رسانی خلاصه‌ها است و در نتیجه، تنها کیفیت خروجی را مورد سنجش قرار می‌دهد. روش‌هایی که در این گروه قرار می‌گیرند عبارتند از: دقت و بازخوانی جمله [۳۳]، رتبه‌دهی به جملات، روش سودمند، پیوستگی خلاصه، اطلاع-رسانی خلاصه و... .

- **ارزیابی بیرونی**[۳۴]: برخلاف ارزیابی درونی، در ارزیابی بیرونی، توجه بر روی کاربر نهایی معطوف می‌شود. در این روش کیفیت خلاصه با توجه به اینکه چگونه در انجام یک وظیفه

[۳۲] Intrinsic

[۳۳] Precision and recall

[۳۴] Extrinsic

خاص (مانند دسته‌بندی) مفید است، مشخص می‌گردد. این روش، گاهی مبتنی بر وظیفه[۳۵] نیز نامیده می‌شود. در این روش چندین سناریوی بازی مانند،به عنوان روش‌های سطحی برای ارزیابی خلاصه، پیشنهاد داده شده که ترتیب‌های مختلفی دارند. در میان آن‌ها بازی Shannon(تئوری اطلاعات) را می‌توان نام برد[۹].

روش‌های دقت و بازخوانی برای ارزیابی سیستم پیشنهادی استفاده شده است ودر ادامه به معرفی آنها می‌پردازیم در این قسمت به معرفی معیار ارزیابی مبتنی بر سودمندی اکتفا می‌کنیم.

- **معیار ارزیابی مبتنی بر سودمندی**

در روش‌های مبتنی بر سودمندی، به جای ایجاد خلاصه‌های دستی، جمله‌ها بر حسب اهمیت آن‌ها توسط افراد خبره امتیازدهی می‌شوند. به این ترتیب، می‌توان ارزیابی خلاصه‌ها را بر مبنای میزان سودمندی جمله‌های آن‌ها انجام داد. برای نمونه، جدول ۵-۱ را درنظر بگیرید. این جدول، امتیاز مربوط به چند جمله وهمچنین خروجی دو سامانه‌ی خلاصه‌سازی را نشان می‌دهد. اگر فرض کنیم خلاصه‌ی ایده‌آل شامل با ارزش‌ترین جمله‌ها یعنی جمله‌ی اول، جمله‌ی دوم وجمله-ی چهارم باشد، دقت سامانه‌ی ۱ برابر ۶۶درصد ودقت سامانه‌ی دوم برابر با ۳۳ درصد خواهد بود. به این ترتیب معیار دقت، خلاصه‌ی نخست را خلاصه‌ی بهتری از خلاصه‌ی دیگر می‌داند.

[۳۵] Task-based

جدول(۵-۱) مقایسه سامانه‌های خلاصه‌سازی بر مبنای سودمندی جمله‌ها

جمله	ایده‌آل	سامانه‌ی ۱	سامانه‌ی ۲
جمله‌ی اول	۱۰(+)	+	+
جمله‌ی دوم	۹(+)		
جمله‌ی سوم	۶		+
جمله‌ی چهارم	۷(+)	+	
جمله‌ی پنجم	۳	+	
جمله‌ی ششم	۶		+

در روش سودمندی، مجموع سودمندی خلاصه‌ی ایده‌آل ۲۶، خلاصه‌ی نخست ۲۰، خلاصه دوم۲۲ است. به این ترتیب با در نظر گرفتن سودمندی جمله‌ها، خلاصه دوم بهتر از خلاصه نخست است.

۵-۲- روش‌های ارزیابی

در این پایان‌نامه برای ارزیابی سیستم خلاصه‌ساز پیشنهاد شده، از روش‌های دقت و بازخوانی و مقایسه با خلاصه مرجع که در گروه ارزیابی‌درونی قرار می‌گیرند استفاده شده است. و همچنین ارزیابی میزان افزونگی در خلاصه‌ی تولید شده توسط سیستم نیز به طور جداگانه انجام شده‌است. روش‌های استفاده شده به شرح زیر هستند:

۵-۲-۱- معیار دقت و بازخوانی:

بازخوانی، تعداد جملات خلاصه‌ی مرجع که در خلاصه‌ی تولید شده، حضور دارند را مشخص می‌کند. به همین ترتیب می‌توان دقت را به صورت تعداد جملات خلاصه تولید شده که در خلاصه مرجع وجود دارند، تعریف کرد. بازخوانی و دقت، معیارهای استاندارد در بازیابی اطلاعات هستند و اغلب از

ترکیب آنها، تحت عنوان F_measure یاد می‌شود. نحوه محاسبه هریک از معیارهای دقت و بازخوانی[۴۰] در صفحه‌ی بعد آورده شده است:

$$\text{بازخوانی} = \frac{\text{اشتراک جملات خلاصه ایجاد شده توسط سیستم و انسان}}{\text{تعداد جملاتی که توسط انسان تشخیص داده شده است}} \quad (۵-۱)$$

$$\text{دقت} = \frac{\text{اشتراک جملات خلاصه ایجاد شده توسط سیستم و انسان}}{\text{تعداد جملاتی که توسط سیستم برای خلاصه ایجاد شده است}} \quad (۵-۲)$$

ما برای ارزیابی روش خلاصه‌سازی پیشنهادی، مجموعه‌ای شامل ۱۰سند متنی خبری را به عنوان پیکره[۳۶] در نظرگرفتیم. این متن‌ها را از نسخه دو مجموعه همشهری انتخاب کرده‌ایم. میزان فشرده-سازی در حالت۴۰ درصد و۶۰ درصد مورد آزمایش قرارگرفت. ۳ شخص(دو دانشجویی کارشناسی ارشد و یک دانشجوی دکترا) به صورت جداگانه هر یک از ۱۰متن اولیه را با میزان فشرده‌سازی مشخص شده خلاصه کردند و این خلاصه‌ها به عنوان خلاصه‌های مرجع در نظر گرفته شد. خلاصه-سازی با میزان فشرده‌سازی فوق توسط سیستم خلاصه‌سازی پیشنهادی ما نیز انجام شدکه در ادامه به شرح نتایج می‌پردازیم. از کارهای دیگران تنها موردی که شبیه سیستم ما می‌باشد، وخلاصه‌ساز گزینشی است و از روش‌های آماری استفاده کرده و همچنین برای خلاصه‌سازی اخبار نیز می‌باشد FarsiSum بود که برای مقایسه نتایج از آن استفاده شده است.

[۳۶] Corpus

۵-۲-۳- ارزیابی میزان افزونگی جمله‌ها

میزان اهمیت جمله‌های یک خلاصه به تنهایی نمی‌تواند معیار مناسبی برای ارزیابی خلاصه‌ها باشد. باید معیار دیگری معرفی کنیم که میزان افزونگی اطلاعات را بسنجد. معیار دومی که برای ارزیابی خلاصه‌ها استفاده خواهد شد، افزونگی جمله‌ها است. برای این کار، از چند داور انسانی خواسته می‌شود تا اطلاعاتی که تکرای هستند را مشخص کنند. نحوه کار به این صورت است که هر داور، با مشاهده‌ی تکرار جزیی یا کلی اطلاعات آن را برحسب تعداد جمله‌ی افزونه مشخص کند. برای نمونه، اگر در یک خلاصه دو جمله‌ی یکسان وجود داشته باشد ارزیابی داور باید میزان افزونگی آن را ۱جمله تشخیص دهد. بر حسب میزان افزونگی بین دو جمله، داور می‌تواند یکی از مقادیر۰,۲۵، ۰,۵، ۰,۷۵ یا ۱ را به عنوان میزان افزونگی آن دو انتخاب کند. در نهایت، میانگین افزونگی هر خلاصه به ازای همه-ی داوران مشخص می‌شود و ثبت می‌گردد. بعد از بررسی خلاصه‌ی تولید شده از متون خبری مشخص شد، درسیستم پیشنهادی هیچ جمله‌ای تکراری در خلاصه وجود ندارد در نتیجه سیستم افزونگی ندارد.

۵-۳- نتایج و مقایسه

جدول‌های ۵-۲ و ۵-۳ روش پیشنهادی ما را در مقایسه با FarsiSum براساس معیارهای مختلف نشان می‌دهد.

جدول ۵-۲ مقایسه روش پیشنهادی با FarsiSum را استفاده از معیار دقت و بازخوانی ما را در نشان می‌دهد.

جدول ۵-۲ مقایسه‌ی روش پیشنهادی در مقایسه با FarsiSum

سیستم خلاصه-سازی	نرخ فشرده‌سازی ۴۰درصد		نرخ فشرده سازی ۶۰درصد	
	دقت	بازخوانی	دقت	بازخوانی
روش پیشنهادی	۰,۸۱	۰,۸۱	۰,۷۳	۰,۷۳
FarsiSum	۰,۵۵	۰,۵۵	۰,۶۵	۰,۶۵

روش دوم ارزیابی نتایج، مقایسه سیستم پیشنهادی با خلاصه مرجع می‌باشد. نتایج مقایسه خلاصه روش پیشنهادی با خلاصه مرجع برای هر یک از ده سند در ده جدول جداگانه آمده‌اند دو نمونه در پیوست (ج) آورده شده‌است که بعد از بررسی ده جدول نتایج کلی در جدول ۵-۳ نشان داده شده است.

جدول ۵-۳ مقایسه نتایج خلاصه مرجع وخلاصه سیستم پیشنهادی

حضور یا عدم حضور جملات در سیستم پیشنهادی	درصد حضور جملات در خلاصه‌های انسانی	شماره جمله برگزیده (به ترتیب بیشترین حضور)
۱۰۰	۱۰۰	**۱**
۹۵	۹۵	**۲**
۴۰	۴۰	**۳**
<۴۵	<۴۰	۴جمله تا جمله $\frac{n}{2}$
<۳۰	<۴۰	جمله$\frac{n}{2}$ تا جمله $n-2$
۶۰	۶۰	$n-1$
۸۰	۸۰	n

ستون اول جدول نمایانگر تعداد جملات مستندات، ستون دوم درصد حضور جمله در خلاصه‌های انسانی (به عنوان مثال ۱۰۰ در سطر اول مربوط به جمله اول، به معنای حضور این جمله در همه‌ی خلاصه‌های انسانی است) و ستون سوم نمایانگر حضور یا عدم حضور جمله در سیستم پیشنهادی است.

فصل ۶

جمع‌بندی و پیشنهادات برای ادامه کار

۶-۱- جمع‌بندی و نتیجه‌گیری

با گسترش روز افزون حجم اطلاعات، نیاز به سیستم‌های کامپیوتری جهت پردازش وتحلیل اطلاعات بیشتر احساس می‌شود. یکی از انواع سیستم‌هایی که در تحلیل و پردازش متون وجود دارد، سیستم‌های خلاصه‌ساز متن است که حجم زیادی از متن را دریافت نموده و بر اساس الگوریتم‌ها و تکنیک‌های مختلف، آن را خلاصه می‌نماید. از کابردهای خلاصه‌سازی می‌توان به خلاصه‌سازی اتوماتیک اخبار و ارسال آن‌ها از طریق پست‌الکترونیکی یا پیامک اشاره نمود. از دیگر کاربردهای آن خلاصه‌سازی تحقیقاتی، تجاری و خلاصه‌سازی صفحات وب برای آنکه در صفحه موبایل قابل نمایش باشد، است.

خلاصه‌سازی یکی از زیر مجموعه مشکلات پیچیده پردازش زبان طبیعی می باشد که در حال حاضر همچنان حل نشده است و هنوز زمان باقی است تا به حدی برسیم که ماشین همانندانسان به طور کامل متن را متوجه شود.

در این تحقیق، سعی شده است با بهره‌گیری از روش‌های موجود و همچنین ارائه‌ی یک روش متفاوت در مقایسه با کارهای انجام شده سیستم خلاصه‌سازی ارائه شود از دقت بالاتری برخوردار باشد. سیستم پیشنهادی، سیستم خلاصه‌سازی تک سندی اخبار فارسی می‌باشد ولی در بقیه‌ی متون فارسی قابل استفاده می‌باشد، ولی ارزیابی نتایج آن روی متون خبری انجام شده است. ما با استفاده از استخراج کلمات کلیدی هم‌رخداد و استخراج رویدادهای مهم متن خبر سعی کرده‌ایم تا جملاتی را از متن خبر استخراج کنیم تا هم دارای مطالب مهم خبری و هم پیوستگی مطالب باشند.

روش ارزیابی استفاده شده در این پایان‌نامه، یک روش ارزیابی مستقیم است. این روش شامل دو بخش است. در بخش اول، خلاصه‌ی خودکار با تعدادی خلاصه‌ی مرجع که توسط افراد خبره تهیه

شده است مقایسه می‌شود و با اهمیت بودن جمله‌های موجود در خلاصه مورد ارزیابی قرار می‌گیرد. در بخش دوم، میزان اطلاعات تکراری در جمله‌های گزینش شده ارزیابی می‌شود.

نتایج حاصل از ارزیابی روش پیشنهادی نشان می‌دهند که سیستم ما دقت بهتری نسبت به سیستم خلاصه‌ساز اخبارFasiSum داشته و نزدیک به خلاصه‌های انسانی باشد.

۶-۲- کارهای آینده

راه‌های مختلفی برای بهبود روش پیشنهادی وجود دارد. در روشی که در این پایان‌نامه استفاده شده است فرض بر این است که هر جمله شامل یک رویداد است، به دلیل اینکه تعداد کلمات موجود در یک جمله به تنهایی کمتر از آن است که ملاک مناسبی برای بررسی رویداد به ما بدهد؛ به همین دلیل بهتر است در یک پنجره n تایی از جملات بررسی رویدادهای متن انجام شود.

استفاده از الگوریتم‌های دسته‌بندی برای دسته‌بندی رویدادها می‌تواند به بهبود کیفیت خلاصه کمک شایانی کند وهمچنین از منطق فازی می‌توان در شناسایی رویدادها استفاده کرد.

تمامی کارهای انجام شده در زمینه‌ی ارزیابی خلاصه‌سازهای ماشینی تا کنون، صرفا به بررسی میزان تشابه نحوی دو متن پرداخته و همچنان در سطح معنا کار نمی‌کند. می‌توانیم درشناسایی رویدادها معنای کلمات وروابط معنایی آنها را هم در نظر بگیریم، به همین منظور می‌توان برای مقایسه در سطح واژگان، از ابزار WordNet استفاده نمود. بدین صورت که واژگانی که قرار است با هم مقایسه شوند، اگر کاملا شبیه هم باشند بالاترین امتیاز را می‌گیرند. در غیر این صورت به WordNet داده می‌شوند تا روابط معنایی بین آن‌ها از قبیل مترادف، متضاد و ... مشخص گردد و با توجه به هر یک از این روابط، امتیازی به آنها داده شود که در واقع این گام، در حرکت به سوی معنایی‌تر شدن این مقایسه‌ها نقش بسیار موثری خواهد داشت.

پیوست‌ها

پیوست الف: نمونه‌ای از خروجی سیستم پیشنهادی میزان فشرده‌سازی ۶۰درصد

خبر (۱) : گزارش خبرگزاری فرانسه از دستور رییس جمهوربرای حضور بانوان در ورزشگاه

دستور رییس جمهور به رییس سازمان تربیت بدنی مبنی بر ایجاد راهکاری برای حضور بانوان در ورزشگاه‌ها در خبرگزاری فرانسه بازتاب داشت.

به گزارش خبرگزاری دانشجویان ایران(ایسنا) :محمود احمدی نژاد رییس جمهور ایران روز دوشنبه اعلام کرد که زنان می‌توانند سرانجام برای تماشای وقایع ورزشی به ورزشگاه‌ها بیایند. احمدی نژاد به تلوزیون محلی ایران گفت: راهکاری باید مشخص شود که به بانوان احترام گذاشته شود و بهترین مکان‌ها برای تماشای دیدارهای ملی و مهم به آنها اختصاص یابد. پس از چندین سال تنها عده‌ای از بانوان می‌توانند با دعوت از سوی مقامات امنیتی رسمی با حضور درقسمت VIP ورزشگاه دیدارهای ورزشی را به صورت زنده تماشاکنند. حتی ورزشی‌نویسان زن هم برای حضور در ورزشگاه‌ها محدودیت دارند. مهدی فراهانی مدیر کل حراست سازمان تربیت بدنی به ایسنا اعلام کرد ما برای ورزشگاه‌ها نیاز به زمان داریم تا بتوانیم آنها را مجهز کنیم و باید برنامه‌ای ریخته شود تا ورودی‌ها وخروجی‌های جداگانه‌ای ترتیب داده شود همچنین او افزود: اجازه به بانوان و خانواده‌ها برای حضوردر ورزشگاه‌ها می‌تواند فضای آنجا را تلطیف دهد.

خلاصه خبر (۱) توسط سیستم پیشنهادی:

دستور رییس جمهور به رییس سازمان تربیت بدنی مبنی بر ایجاد راهکاری برای حضور بانوان در ورزشگاه‌ها در خبرگزاری فرانسه بازتاب داشت. محمود احمدی نژاد رییس جمهور ایران روز دوشنبه اعلام کرد که زنان می‌توانند سرانجام برای تماشای وقایع ورزشی به ورزشگاه‌ها بیایند. مهدی فراهانی مدیر کل حراست سازمان تربیت بدنی به ایسنا اعلام کرد ما برای ورزشگاه‌ها نیاز به زمان داریم تا بتوانیم آنها را مجهز کنیم و باید برنامه‌ای ریخته شود تا ورودی‌ها وخروجی‌های جداگانه‌ای ترتیب داده شود همچنین او افزود: اجازه به بانوان و خانواده‌ها برای حضوردر ورزشگاه‌ها می‌تواند فضای آنجا را تلطیف دهد.

خبر ۲: روزنامه" مشارکت "منتشرشد.

نخستین شماره روزنامه مشارکت ارگان جبهه مشارکت ایران اسلامی منتشر شد. در سرمقاله نخستین شماره مشارکت که به قلم سید محمدرضا خاتمی صاحب امتیاز و مدیرمسئول این نشریه نوشته شده آمده است: "تجربه انتشار روزنامه مشارکت، تجربه ای جدید و در نوع خود کم نظیر به حساب می آید و آن انتشار روزنامه توسط دست اندرکاران یک حزب است با این تاکید که ارگان رسمی یا سخنگوی جبهه مشارکت نخواهد بود ولی قرابت فکری و عملی زیادی با جبهه مشارکت خواهد داشت و خوانندگان را با دیدگاهها و مواضع حزب بیشتر آشنا خواهد کرد. "در بخش دیگری از این سرمقاله آمده است: رعایت اخلاق وپرهیز از حرمت شکنی سرلوحه کار ما خواهد بود و در عین اعتقاد و عمل به شفافیت و جدیت در انعکاس حقایق وواقعیت های سیاسی و اجتماعی، حرمت شکنی را روا نخواهیم داشت . جفا بر مخالفان را نیز نخواهیم پسندید، نقد وانتقاد، مرام ما خواهد بود و در این راه می کوشیم هیچکسی را محروم از این نعمت نگذاریم. کادر تحریریه روزنامه مشارکت اعضای تحریریه روزنامه سلام می باشند ستون "الو سلام "از ستونهای ثابت مشارکت است. روزنامه مشارکت در ۱۲ صفحه در قطع مترویی با مطالب فرهنگی، هنری، اقتصادی، جوان، دانشگاه، سیاسی، اجتماعی، ورزشی و بین الملل منتشر می شود. روزنامه همشهری پیوستن مشارکت به خانواده بزرگ مطبوعات کشور را تبریک می گوید وبرای همکاران مطبوعاتی در آن نشری آرزوی توفیق دارد.

خلاصه خبر (۲) توسط سیستم پیشنهادی:

نخستین شماره روزنامه مشارکت ارگان جبهه مشارکت ایران اسلامی منتشر شد . در سرمقاله نخستین شماره مشارکت که به قلم سید محمدرضا خاتمی صاحب امتیاز و مدیرمسئول این نشریه نوشته شده آمده است:تجربه انتشار روزنامه مشارکت، تجربه ای جدید و در نوع خود کم نظیر به حساب می آید و آن انتشار روزنامه توسط دست اندرکاران یک حزب است با این تاکید که ارگان رسمی یا سخنگوی جبهه مشارکت نخواهد بود ولی قرابت فکری و عملی زیادی با جبهه مشارکت خواهد داشت و خوانندگان را با دیدگاهها و مواضع حزب بیشتر آشنا خواهد کرد. روزنامه مشارکت در ۱۲ صفحه در قطع مترویی با مطالب فرهنگی، هنری، اقتصادی، جوان، دانشگاه، سیاسی، اجتماعی، ورزشی و بین الملل منتشر می شود.

خبر۳: دریافتی مستمری بگیران تامین اجتماعی ۱۷ درصد افزایش می یابد.

مدیر عامل سازمان تامین اجتماعی بودجه پیشنهادی ۲۰۰۰ میلیارد تومانی سال ۱۳۷۹ این سازمان را جهت بررسی و تصویب به شورای عالی تامین اجتماعی ارائه کرد. در جلسه اخیر شورای عالی تامین اجتماعی، دکتر ستاری فرمدیرعامل سازمان تامین اجتماعی گفت: بودجه اصلی سازمان تامین اجتماعی در بخش بیمه ای شامل ۱۲۸۱ میلیارد تومان درآمد متعلق به سازمان(ناشی از حق بیمه، سرمایه گذاری، خسارات و جرایم و سایر)و ۷۱۷ میلیارد تومان منابع شرکت سرمایه گذاری است که در مجموع به رقمی معادل ۱۹۹۸میلیارد تومان می رسد. رقم اختصاص یافته برای مصارف سازمان نیز برابر منابع در نظر گرفته شده است که به این ترتیب کل منابع و مصارف این سازمان در سال ۱۳۷۹ رقمی معادل ۳۹۹۶ میلیارد تومان خواهد بود. مدیر عامل سازمان تامین اجتماعی میزان بودجه تخصیص یافته برای تعهدات قانونی بلندمدت و کوتاه مدت این سازمان رادر بودجه سال آینده رقمی معادل ۶۵۱ میلیارد و ۱۳۸ میلیون تومان ذکر کرد و گفت: تعهدات قانونی شامل مستمری بازنشستگی بالغ بر ۳۱۵ میلیارد و ۱۴۹ میلیون تومان، مستمری های از کارافتادگی و بازماندگان بیمه شده متوفا معادل ۲۲۶ میلیارد و ۸۹ میلیون تومان و تعهدات کوتاه مدت (نظیر کمک هزینه بارداری، غرامت دستمزد ایام بیماری و...) رقمی معادل ۱۶ میلیارد و ۷۳۰ میلیون تومان و قالب بیمه بیکاری رقمی معادل ۵۷ میلیارد و ۱۷۰ میلیون تومان در نظرگرفته شده است. وی همچنین با اشاره به این نکته که در سال مالی ۱۳۷۹ شرایط جدیدی برای برقراری مستمری بازماندگان در نظر گرفته شده است، خاطر نشان کرد: تا به حال سازمان تنها برای بازماندگان بیمه شدگانی که بین ۱۰ تا ۲۰ سال سابقه پرداخت حق بیمه داشتند، پرداخت غرامت مقطوع در نظرمی گرفت و بازماندگان افرادی که زیر این میزان سابقه داشتند از این پرداخت محروم بودند، اما با توجه به این که خانواده های بازماندگان از نظر مالی در شرایط نامساعداقتصادی قرار دارند، تصمیم گرفته شد که برای بازماندگان آن دسته از بیمه شدگان متوفا که بین ۱ تا ۱۰ سال سابقه پرداخت حق بیمه دارند نسبت به پرداخت غرامت مقطوع به آنان اقدام کند که متعاقبا" آئین نامه آن تدوین و اعلام خواهد شد. دکتر ستاری فر افزود: رشد مستمری ها براساس ماده ۹۶ قانون تامین اجتماعی که سازمان را مکلف می کند، حداقل سالی یک بار براساس تورم، حقوق بازنشستگان ومستمری بگیران را افزایش دهد، که در بودجه سال آینده میزان مستمری ها با ۱۷ درصد رشد پیش بینی شده است.

خلاصه خبر (۳) توسط سیستم پیشنهادی:

مدیر عامل سازمان تامین اجتماعی بودجه پیشنهادی ۲۰۰۰ میلیارد تومانی سال ۱۳۷۹ این سازمان را جهت بررسی و تصویب به شورای عالی تامین اجتماعی ارائه کرد. در جلسه اخیر شورای عالی تامین اجتماعی، دکتر ستاری فرمدیرعامل سازمان تامین اجتماعی گفت: بودجه اصلی سازمان تامین اجتماعی در بخش بیمه ای شامل ۱۲۸۱ میلیارد تومان درآمد متعلق به سازمان(ناشی از حق بیمه، سرمایه گذاری، خسارات و جرایم و سایر)و ۷۱۷ میلیارد تومان منابع شرکت سرمایه گذاری است که در مجموع به رقمی معادل ۱۹۹۸میلیارد

تومان می رسد. رقم اختصاص یافته برای مصارف سازمان نیز برابر منابع در نظر گرفته شده است که به این ترتیب کل منابع و مصارف این سازمان در سال ۱۳۷۹ رقمی معادل ۳۹۹۶ میلیارد تومان خواهد بود. دکتر ستاری فر افزود: رشد مستمری ها براساس ماده ۹۶ قانون تامین اجتماعی که سازمان را مکلف می کند، حداقل سالی یک بار براساس تورم، حقوق بازنشستگان ومستمری بگیران را افزایش دهد، که در بودجه سال آینده میزان مستمری ها با ۱۷ درصد رشد پیش بینی شده است.

خبر۴: سال ۲۰۰۰ بدون بحران رایانه ها آغازشد.

ساکنان کره زمین از سال دوهزار، قرن بیست ویکم و هزاره سوم استقبال کردند. سال جدید میلادی ظرف یک روز که از حوالی ظهر پنجشنبه به وقت تهران و از جزایر تونگا در منتهی الیه شرق اقیانوس آرام آغاز و ظرف ۲۴ ساعت تا ظهرجمعه در جزیره ساموآ تکمیل شد، تمام کره زمین رادربرگرفت. برخلاف تمام سناریوهای بدبینانه ای که در مورد اختلال درسامانه های رایانه ای در سراسر جهان اعلام شده بود و ازاختلال در ارائه خدمات آب، برق، مخابرات تا نقص بالقوه فاجعه بار در زرادخانه های هسته ای برخی کشورها را شامل می شد، هیچ واقعه ناگواری در کشورهای مختلف رخ نداد و.مردم و کارشناسان رایانه که از حدود دوسال پیش با دلهره آغاز این روز را انتظار می کشیدند، نفس راحتی کشیدند. انتظار دلهره آور اختلال در فعالیت رایانه ها به هیچ وجه متناسب با آنچه رخ داد، نبود۸۹۰ کشور جهان با گذشت چند ساعت از تحویل سال جدید میلادی به مرکز بین المللی همکاری گزارش کردند دچار هیچگونه اختلالی در ۱۱ بخش عمده صنعتی از قبیل برق و مخابرات نشده اند. در این حال، فرانسه اعلام کرد اختلالات کوچکی در نظام بانکی بروز کرد که به سرعت مرتفع شد.

برزیل بزرگترین اقتصاد آمریکای لاتین و دیگر کشورهای آمریکای جنوبی ونیز کانادا نیز اعلام کردند هیچگونه مشکلی در بخشهای خدمات آب، برق، هواپیمایی و مخابرات آنها بروز نکرده است. در ژاپن و در ساعات پس از تحویل سال، بروز مشکلاتی درچهار نیروگاه برق گزارش شد ولی عملیات این نیروگاههامتوقف نشد و به علاوه مشخص نشد آیا این مشکلات ناشی از بوده است یا منشاء دیگری داشته است. در آمریکا وضعیتی مشابه در ۸ نیروگاه برق پیش آمد که باعث توقف ارائه خدمات نشد و این مشکلات به سرعت برطرف شد. شرکت آمریکایی آی بی ام، بزرگترین شرکت رایانه سازی جهان اعلام کرد تمام سامانه های آن به طور عادی کارمی کند و به علاوه دارندگان تولیدات این شرکت در سراسرجهان هیچگونه اختلالی را گزارش نکرده اند. مقامات نظامی آمریکا نیز گزارش کردند از میان شصت و اندی پایگاه هوایی این کشور در سراسر جهان، چهار پایگاه از جمله یک پایگاه در آمریکا و سه پایگاه در خارج این کشور به طور موقت دچار مشکلاتی شدند که به سرعت برطرف شد. گروهی از فرماندهان نظامی آمریکایی و روسیه در خلال هفته های گذشته در پایگاه هوایی پترسون در ایالت کلرادومشغول هماهنگی فعالیتهای رایانه ای نظامی دو کشور بودندتا فاجعه ای در آغاز هزاره جدید رخ ندهد. روسیه به خوبی توانست تهدید آفت هزاره را نه تنها در بخش موشکهای هسته ای که در بخشهای غیرنظامی همچون شبکه های تلفن وبرق در سر تا سر این کشور پهناور پشت سربگذارد. یکی از بخشهایی که تا پیش از این در موردآمادگی آن برای مقابله با آفت هزاره نگرانی زیادی وجود داشت، صنایع

هواپیمایی جهان بود در این بخش نیز همچون دیگر بخشهای اقتصادی حتی یک مشکل نیز بروز نکرد. در صنایع نفت خاورمیانه نیز که شریان حیاتی اقتصاد این کشورها محسوب می شود، هیچ نقصانی گزارش نشد روی هم رفته کارشناسان معتقدند از میان مشکلات گزارش شده، بخش کوچکی از آنها ناشی Y2K از بوده است. با این حال، آنان هشدار می دهندهنوز نباید این خطر را دست کم گرفت. از روز دوشنبه فعالیتهای اداری و اقتصادی در اکثر کشورهای جهان آغازمی شود و ممکن است برخی مشکلات از اواسط هفته خود رانشان دهد، هرچند که بالقوه احتمال بقا و بروز این مشکل تا روزها و حتی هفته های بعد ادامه دارد. افزون بر Y2K تهدید، احتمال پخش گسترده ویروسهای رایانه ای و حمله به نشانی های اینترنتی در آغاز هزاره سوم نیز وجود داشت که عملا" قسمت اعظم آن تحقق نیافت. کارشناسان خاطرنشان می کنند آمادگی کشورها برای مقابله با این نقص فنی، عمده علت پیشگیری از حوادث ناخواسته دراولین روز سال میلادی جدید بود. بر اساس آمار رسمی دولت آمریکا، در این کشور بیش از یک صد میلیارد دلار ظرف دوسال گذشته صرف برطرف شدن این مشکل شد و به همین اندازه نیز دیگر کشورها در همین زمینه هزینه کردند. برخی دیگراز برآوردها رقمی بین ۲۰۰ تا ۶۰۰ میلیارد دلار را درسراسر جهان نشان می دهد.

خلاصه خبر (۴) توسط سیستم پیشنهادی:

ساکنان کره زمین از سال دوهزار، قرن بیست ویکم و هزاره سوم استقبال کردند. سال جدید میلادی ظرف یک روز که از حوالی ظهر پنجشنبه به وقت تهران و از جزایر تونگا در منتهی الیه شرق اقیانوس آرام آغاز و ظرف ۲۴ ساعت تا ظهرجمعه در جزیره ساموآ تکمیل شد، تمام کره زمین رادربرگرفت. برخلاف تمام سناریوهای بدبینانه ای که در مورد اختلال درسامانه های رایانه ای در سراسر جهان اعلام شده بود و ازاختلال در ارائه خدمات آب، برق، مخابرات تا نقص بالقوه فاجعه بار در زرادخانه های هسته ای برخی کشورها را شامل می شد، هیچ واقعه ناگواری در کشورهای مختلف رخ نداد. گروهی از فرماندهان نظامی آمریکایی و روسیه در خلال هفته های گذشته در پایگاه هوایی پترسون در ایالت کلرادومشغول هماهنگی فعالیتهای رایانه ای نظامی دو کشور بودندتا فاجعه ای در آغاز هزاره جدید رخ ندهد.مقامات نظامی آمریکا گزارش کردند از میان شصت و اندی پایگاه هوایی این کشور در سراسر جهان، چهار پایگاه از جمله یک پایگاه در آمریکا و سه پایگاه در خارج این کشور به طور موقت دچار مشکلاتی شدند که به سرعت برطرف شد. کارشناسان خاطرنشان می کنند آمادگی کشورها برای مقابله با این نقص فنی، عمده علت پیشگیری از حوادث ناخواسته دراولین روز سال میلادی جدید بود. بر اساس آمار رسمی دولت آمریکا، در این کشور بیش از یک صد میلیارد دلار ظرف دوسال گذشته صرف برطرف شدن این مشکل شد و به همین اندازه نیز دیگر کشورها در همین زمینه هزینه کردند. برخی دیگراز برآوردها رقمی بین ۲۰۰ تا ۶۰۰ میلیارد دلار را درسراسر جهان نشان می دهد.

پیوست (ب): نمونه‌ای از امتیاز دهی به جملات

گروه خبر : علمی و فرهنگی

عنوان خبر : کتابخانه بین المللی مونیخ اعلام کرد: مرادی کرمانی کتاب" مربای شیرین "برگزیده سال شد.

شماره‌ی جمله	متن جمله	امتیاز جمله
۱	کتاب داستان مربای شیرین نوشته هوشنگ مرادی کرمانی که سال گذشته از سوی انتشارات معین به چاپ رسیده بود، به عنوان کتاب برگزیده سال ۱۹۹۹ کتابخانه بین المللی مونیخ معرفی شد.	۱
۲	کتابخانه بین المللی مونیخ آلمان ویژه نوجوانان که درسال گذشته میلادی پنجاهمین سالگرد تاسیس خود را جشن گرفت با مجموعه ای بیش از پانصد هزار عنوان کتاب به بیش از صد زبان مختلف در سراسر دنیا شهرت خاصی دارد.	۰,۸
۳	کتابخانه مونیخ علاوه بر فعالیت های گوناگون مانند برگزاری نمایشگاه ها، انتشار کتاب و... دارای مجموعه ای تحت Whith Raven عنوان است .	۰,۶
۴	همه ساله متخصصین ادبیات کودک از میان کتابهای منتشرشده در سراسر جهان تعدادی را انتخاب و به صورت مجموعه ای در دسترس علاقه مندان قرارمی دهند .	۰,۱
۵	مجموعه ۱۹۹۹ از کتاب مربای شیرین هوشنگ مرادی کرمانی که توسط انتشارات معین منتشرشده است با عنوان یکی از کتابهای برگزیده این کتابخانه در سال ۱۹۹۹ نام برده شده است .	۰,۳
۶	نام دو کتاب پاپر نوشته محمد میرکیانی و پولک ماه نوشته اسدالله شعبانی در بخش معرفی ویژه این کتابخانه به چشم می خورد.	۰,۵

مراجع

مراجع:

[1] Yatsko V and Vishnyakov T (2007)، "Some Problems in the development of Systems of Automatic Text Summarization"، *Automatic Documentation and Mathematical Linguistics*،vol. 41، no. 5، pp. 185-193.

[2] Visser W.T.،Wieling M.B (2008)، "Sentence-Based Summarization of Scientific Documents"،M.S. Project، University of Groningen.

[3] E. Hovy (1990)، "Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations" in Workshop on NLG.

[۴] مشکی م.، آنالویی م.(۱۳۸۷)، "خلاصه سازی چند سندی متون فارسی با استفاده از یک روش مبتنی بر خوشه بندی"، *اولین کنفرانس ملی مهندسی نرم‌افزار ایران*، تهران.

[5] Dalianis، H.(2000). SweSum –A Text Summarizer for Swedish، Technical report، TRITANA-P0015، IPLab-174.

[6]http://wtlab.um.ac.ir/index.php?option=com_content&view=category&id=36&Itemid=148&lang=fa&limitstart=5.

[7] H.Daume، A.Echihabi، D.Marcu، D.Stefan، R.Soricut.(2000) GLEANS: A Generator of Logical Extracts and Abstracts for Nice Summaries..

[8] K.Mckeown and S.F.Chang and J.Cimino,S and K.Feiner، (2001)، PERSIVAL :a System for Personalized Search and Summarization over Multimedia Healthcare Information. JCDL'01، June 24–28، Roanoke،Virginia, USA. ACM.

[9] E.Hovy، C.Lin(1998). Automated Text Summarization in SUMMARIST.

[10] Mazdak، N، (2004)، Master thesis، "FarsiSum-a Persian text summarizer"، Department of linguistics، Stockholm University.

[11] Shamsfard M and Akhavan، T and Erfani M (2009). PARSUMIST:" A Persian Text Summarization"، *IEEE conference on natural language processing and knowledge Engineering* (NLPKE'09)، Dalian، China.

[12] Honarpisheh M.(2008)، Ghassem-Sani G.R.، Mirroshandel G.. "A Multi-Document Multi-Lingual Automatic Summarization System"، *Proceedings of the Third International Joint Conference on Natural Language Processing*.

[۱۳] کریمی ز، شمس‌فرد م.(۱۳۸۵)، "سیستم خلاصه‌سازی خودکار متون فارسی"، *دوازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران*، تهران ۱۳۸۵.

[۱۴] شهابی ا.ش. (۱۳۸۱)،"چکیده‌سازی متون زبان فارسی"، *دومین کنفرانس علوم شناختی*، صفحه ۵۶، تهران.

[15] Mohamad Ali Honarpisheh and Gholamreza Ghassem-Sani and Ghassem Mirroshandel، (2007)،"A Multi-Document Multi-Lingual Automatic Summarization System"، *Proceedings of the 3rd International Joint Conference on natural language processing (IJCNLP)* ، pp. 733-738.

[۱۶] مشکی ، آنالویی م. (۱۳۸۸)،"خلاصه‌سازی چند سندی متون فارسی با استفاده از یک روش مبتنی بر خوشه-بندی"، *اولین کنفرانس ملی مهندسی نرم‌افزار*، دانشگاه آزاد رودهن.

[۱۷] مشکی،، آنالویی م. (۱۳۸۸)،"بازیابی خبرهای مرتبط پیشین برای تولید خلاصه‌های پیشینه خبر" ، *اولین کنفرانس ملی مهندسی نرم‌افزار*، دانشگاه آزاد رودهن.

[۱۸] ورفریاری و. (۱۳۷۶)، "بررسی جامع سیستم‌های خلاصه‌ساز فارسی و تولید یک نمونه‌ی عملی"، پایان‌نامه کارشناسی ارشد، دانشکده‌ی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران.

[19] Shiping. L and Jiabin. L،(2008)، "Summarization Based on Event-cluster"،*J.of.IEEE*،978-1-4244-3531-9/08.

[20] Dan. Z and Shengdong. L،(2011) " Topic Detection Based on K-means" ،*J.of.IEEE*،978-1-4577-0321-8/11.

[21] Shengdong. L and Xueqiang. L and Tao. W،(2010) "The Key Technology of Topic Detection Based on K-means" ،*International Conference on Future Information Technology and Management Engineering.* ، *IEEE*،978-1-4244-9088-2/10.

[۲۲] شمس‌فرد م.(۱۳۸۵)، "پردازش متون فارسی: دستاوردهای گذشته، چالش‌های پیش‌رو"، *دومین کارگاه پژوهشی زبان فارسی و رایانه*، ص. ۱۷۲ تا ۱۸۹، تهران.

[۲۳] دبیرخانه شورای عالی اطلاع‌رسانی، (۱۳۸۸)، استخراج نیازمندی‌های تعیین حدود جمله برای پیکره متنی ربان فارسی، نسخه ۱۰، دانشگاه علم و صنعت.

[24] Haoyi W، Hung y،(2001)،PHD Thesis ،"Bondec- A Sentence Boundary detector" ، Berkeley Engineering Information.

[25] Zhang Yand zheng.M and Zincir H Nur،(2005)، "Comparing Key Phrase Extraction Methods in Automatic Web Site Summarization",*4th Dalhousie Computer Science In-House Conference.*

[26] Taghva k، Sadeh M (2005)، "A Stemming Algorithm for the Farsi Language"،ITCC(1):158-162.

[27] Spark-Jones K،(1999)، "Automatic Summarizing: Factors and Direction".

[28] Calero C.، Ruiz F.، Piattini M. (2006) "Ontologies for Software Engineering and Software Technology". Springer-Verlag Berlin Heidelberg.

[29] Lin F., Sandkuhl k.,(2008), "A Survey of Exploiting WordNet in Ontology Matching", In IFIP International Federation for Information Processing, Volume 276; Artificial Intelligence in Theory and Practice II,Max Bramer; (Boston: Springer), pp. 341-350,

[۳۰] حافظی م.م ، ثامتی ح، منصوری ن، بحرانی،موثق. (۱۳۸۵)." ارائه یک مدل دستوری برای بهبود دقت سیستم-های بازشناسی گفتار پیوسته‌ی فارسی"، دومین گارکاه پژوهشی زبان فارسی و رایانه، ص. ۸۰تا ۹۱،تهران.

[۳۱] ایران‌پور مبارکه م. (۱۳۸۶)، "بررسی مشکلات تعیین حدود جمله وکلمه"، سمینار کارشناسی ارشد، دانشگاه علم وصنعت ایران.

[۳۲] دادش میری پ. (۱۳۸۰)،"تشخیص انتهای کلمات وایجاد فاصله میان کلمات"، پایان‌نامه کارشناسی، دانشگاه علم وصنعت ایران.

[۳۳] بیجن‌خان م. (۱۳۸۴)،"تشخیص کسره‌ی اضافه"، طرح تحقیقاتی، پژوهشگاه فرهنگ، هنر وارتباطات، تهران.

[۳۴] حسامی‌فرد ر. ، قاسم ثانی غ.ر. (۱۳۸۴) ، "طراحی یک الگوریتم ریشه‌یابی برای زبان‌فارسی"، *یازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران*، ص. ۵۱۵ تا۵۱۹، تهران.

[۳۵] مشکی م،(۱۳۸۸)، پایان‌نامه ارشد، "خلاصه‌سازی گزینشی چند سندی متون فارسی"، دانشکده کامپیوتر،دانشگاه علم و صنعت.

[36]Turney, P.D. ، (1999) " Learning Algorithms for Key phrase Extraction", Information Retrieval.

[37**]**Mihalcea R and Tarau P (2004)، Textrank: Bringing order into texts. In Proceedings of EMNLP (ed. Lin D and Wu D)، Association for Computational Linguistics، Barcelona، Spain.

[38] Hamming R. W، (1950)، "Error detecting and error correcting codes"، *Bell System Tech*. J., vol. 29, no. 2, pp. 147-160.

[۳۹] Jing.H and Barzilay R and Mckeown K and Elhadad M، "Summarization evaluation methods: Experiments and analysis."، *In AAAI Symposium on Intelligent Summarization*، pages 60–68، 1998.

[۴۰] شورای عالی ازطلاع رسانی ، (۱۳۸۸)، "بررسی مستندات ابزارهای خودکار خلاصه‌سازی زبان‌های دنیا برای به کارگیری در متون فارسی"، دانشگاه علم وصنعت.

Abstract

By increasing electronic text resources on the World Wide Web, the need to accuratly, fast and easy access to more information is vital. In order to find fast and suitable the user's documents, complete reading great literature is inappropriate. The automatic summarization system play an important role and is also a solution to the problem of information overload. Summarization is actually a compact and accurate text input representation so that the output text represent important concepts of the input text. Some of the applications of summarization are automatic news summarization while sending them via e-mail or SMS. other applications can be named research-oriented summarization, business-oriented summarization, information retrival system,industry communication ,editing and filtering systems and web page summarization which are to be displayed on mobile screens. There are two major challenges in designing a system for Persian language summarization. The first challenge is continuity between some symptoms and various forms of words and the second is selecting the most important continuous sentence that are to be present in the summary.

In this thesis, duo to the challenges in the summarization of Persian language , Preprocessing step is completely done. Extacting related and continuous sentence in proposed summarization method, requires related and accurate keywords and therefore we have proposed a method for extracting co-occurrences of words and event detection of text for persian news summarization. In Proposed method, news title, news text and compression rate are received from users and produce summary according to compression rate and related news title. The Proposed method is an unsupervised and statistical method and is a kind of the extraction summary. In order to evaluate proposed method, measures of precision, recall and comparison to human summary and to comparison to FarsiSum are used. Evaluations have be done on set of persion news of Hamshahry database in different scope and domain. According to our experiment ,our system has the best performance among other Persian summarization system(%81/66),the closest to human summaries and it also has a better performance to comared FursiSum system.

Keyword:

Automatic news summarization, event detection, extraction summary, co-occurences word.

Shahrood University of Technology

Faculty of Computer

News Summarization using Artificial Intelligence Techniques

Neda Azimi

Supervisor(s):

Morteza Zahedi

Marzieh Rahimi

Feb 2014