بسم الله الرحمن الرحيم

دانشکده آموزش های الکترونیکی

گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد مهندسی هوش مصنوعی

خلاصه سازی متون فارسی با استفاده از تکنیک‌های هوش مصنوعی

نگارنده: اعظم محمدزاده

استاد راهنما

دکتر مرتضی زاهدی

بهمن ۱۳۹۴

دانشگاه صنعتی شاهرود

دانشکده : آموزش های الکترونیکی

گروه : مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد خانم هاعظم محمدزاده

تحت عنوان: خلاصه سازی متون فارسی با استفاده از تکنیک‌های هموش مصنوعی

در تاریخ ..... 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 :دکتر مرتضی زاهدی		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

روح پدرم که آرزوهای او آرمانهای منند

مادر و همسر مهربانم که همیشه پشتیبان من بوده اند.

همه عزیزانی که دعای خیرشان بدرقه راهم بوده و هست.

تشکر و قدردانی

در ابتدا از استاد شایسته؛ جناب آقای دکتر زاهدی که در کمال سعه صدر، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛ کمال تشکر و قدردانی را دارم .

اینجانب اعظم محمدزاده دانشجوی دوره کارشناسی ارشد رشته هوش مصنوعی دانشکده آموزش های الکترونیکی دانشگاه صنعتی شاهرود نویسنده پایان نامه خلاصه سازی متون فارسی با استفاده از تکنیک‌های هوش مصنوعی تحت راهنمائی دکتر مرتضی زاهدی متعهد می شوم.

- 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 :

امروزه با توجه به حجم گسترده اطلاعات موجود، کمبود زمان بررسی و پردازش مطالب و همچنین عدم وجود یک سیستم خلاصه ساز کارآمد برای خلاصه سازی حجم انبوه اطلاعات، نیاز به یک سامانه قدرتمند جهت خلاصه سازی متون احساس می شود. خلاصه سازی عبارت است از فشرده سازی متن (متون) منبع و تولید یک نسخه کوتاه تر از آن به نحوی که محتوای اطلاعاتی آن حفظ شود. از کابردهای خلاصه سازی می توان به خلاصه سازی اتوماتیک اخبار و ارسال آن ها از طریق پست الکترونیکی یا پیامک اشاره نمود خلاصه‌سازی متون به دو صورت چکیده‌ای و استخراجی قابل انجام است. روش چکیده‌ای شبیه به روشی است که توسط انسان بکار برده می‌شود. این مورد به دلیل اینکه نیاز به درک معنی دارد، پیچیده می‌باشد از اینرو بیشتر روش‌های ارائه شده برای خلاصه‌سازی متون، از نوع استخراجی می‌باشد. در روش استخراجی، جملاتی از متن براساس متدهایی بعنوان چیکیده انتخاب می‌شوند.

یکی دسته از روش‌های استخراجی پیشنهاد شده برای خلاصه‌سازی متون، روش‌های آماری می‌باشند. در این روش‌ها، همه اطلاعات از خود متن بیرون کشیده شده و نیاز به هیچ دانش اولیه‌ای وجود ندارد. هدف از این پایان‌نامه ارائه یک روش آماری به منظور خلاصه‌سازی متون فارسی می‌باشد. این مدل که ترکیبی از یک الگوریتم آماری و یک مدل خوشه‌بندی است، ابتدا کلمات کلیدی را استخراج کرده سپس به هر یک از آنها وزنی اختصاص داده و سپس جملات را براساس آنها، وزن‌دهی می‌کند. سپس با استفاده از یک مدل خوشه‌بندی، جملات مشابه را در یک خوشه قرار داده و درنهایت براساس دو معیار وزن جملات و شباهت آنها، بهترین جملات را استخراج می‌کند. استفاده از شباهت جملات به همراه وزن آنها، باعث می‌شود جملاتی با اهمیت‌تری که نیز به یکدیگر وابسته هستند، استخراج گردد. به منظور ارائه نتایج، مجموعه داده که متشکل از مطالب روزنامه‌های همشهری و مجله نسیم آفتاب بود، تشکیل گردید. همچنین نتایج بدست آمده با استفاده از عامل انسانی ارزیابی شد.

کلمات کلیدی:خلاصه‌سازی، خوشه‌بندی، مجموعه دادهّ ,وزن دهی,

فهرست مطالب

۱-۱ مقدمه ---------- ۲

۱-۲ متن‌کاوی و خلاصه‌سازی متن ---------- ۳
۱-۲-۱ کاربردهای متن‌کاوی ---------- ٤
۱-۲-۲ ارتباط متن‌کاوی و خلاصه‌سازی متن ---------- ٦

فصل دوم تعاریف و مفاهیم ---------- ۹

۲-۱ مقدمه ---------- ۱۰

۲-۲ دسته‌بندی سیستم‌های خلاصه‌سازی متن ---------- ۱۲
۲-۲-۱ شکل متن خروجی ---------- ۱۳
۲-۲-۲ جزئیات ---------- ۱٤
۲-۲-۳ محتویات ---------- ۱٥
۲-۲-۴ محدودیت‌ها ---------- ۱٥
۲-۲-۵ تعداد متون ورودی ---------- ۱٦
۲-۲-٦ تعداد زبان‌های متن ---------- ۱٦

۲-۳ خلاصه‌سازی استخراجی ---------- ۱۷

فصل سوم روشهای خلاصه‌سازی ---------- ۲٥

۳-۱ مقدمه ---------- ۲٦

۳-۲ رویکردهای خلاصه‌سازی متون ---------- ۲٦

۳-۳ استفاده از روشهای آماری در خلاصه‌سازی متن ---------- ۲۷
۳-۱-۱ - خلاصه‌سازی متن با استفاده از زنجیره‌های لغوی ---------- ۳٥

۳-٤ خلاصه‌سازی متن و تئوری گراف ---------- ٤٥

۳-٥ خلاصه‌سازی متن با رویکرد خوشه بندی ---------- ٤۹

۳-٦ خلاصه‌سازی متن با رویکرد فازی ---------- ٥٤

۳-۷ چالشهای زبان فارسی در خلاصه‌سازی متن ---------- ٥۹

۳-۷-۱ چالشهای نگارش زبان فارسی ---------- ٦۰
۳-۷-۲ چالشهاي زبان فارسی در بخش ریشه یابی ---------- ٦۱

۳-۷-۳ چالش‌های ذاتی زبان فارسی ---------- ٦۲

۳-۸ مروی بر سابقه تحقیقات خلاصه‌سازی متن در زبان فارسی ---------- ٦٤
۳-۸-۱ سیستم خلاصه ساز فارسی FarsiSum ---------- ٦٥

۳-۸-۲ -سیستم Automatic Persian Text Summarization ---------- ٦٨
۳-۸-۳ -سیستم Hybrid Farsi Text Summarization ---------- ٦٨

فصل چهارم مدل پیشنهادی ---------- ۷۳

۴-۱ -مقدمه ---------- ۷٤

۴-۲ -مدل پیشنهادی ---------- ۷٤
۴-۲-۱ -بخشبندی ---------- ۷٥
۴-۲-۲ -یافتن ریشه کلمات ---------- ۷٦
۴-۲-۳ -استخراج کلمات کلیدی ---------- ۷۸
۴-۲-۴ -وزندهی به کلمات کلیدی ---------- ۷۸
۴-۲-۵ -وزندهی به جملات بر اساس کلمات کلیدی ---------- ۸۰
۴-۲-۶ - استخراج جملاتی که وزن بیشتری دارند ---------- ۸۰
۴-۲-۷ - استخراج جملات که پیوستگی بیشتری دارند ---------- ۸۳

فصل پنجم ---------- ۹۳

نتیجه‌گیری و پیشنهادات ---------- ۹۳

۵-۱ - نتیجه‌گیری ---------- ۹٤

۵-۲ -پیشنهاد آینده ---------- ۹٥

مراجع ---------- ۹٦

فهرست اشکال

شکل (۴-۱) نمونه ای از مجموعه داده همشهری.....۷۵

شکل (۴-۲) سه جمله کاندیدا برای انتخاب خلاصه.....۷۶

شکل (۴-۳) خوشه‌بندی جملات استخراج.....۸۱

شکل (۴-۴) نمونه‌ای از مجموعه داده جمع‌آوری شده.....۸۵

شکل (۴-۵) نمونه‌ای دیگری از مجموعه داده جمع‌آوری شده.....۸۹

فهرست جداول

جدول (۴-۱).....۷۷

جدول(۴ – ۲).....۷۹

جدول(۴ – ۳).....۸۲

جدول (۴ – ۴).....۸۳

جدول (۴ – ۵).....۸۶

جدول (۴ –۶).....۹۱

فصل اول:

مقدمه

۱-۱-مقدمه

امروزه با توجه به حجم گسترده اطلاعات موجود، کمبود زمان بررسی و پردازش مطالب و همچنین عدم وجود یک سیستم خلاصه ساز کارآمد برای خلاصه‌سازی حجم انبوه کتب، مقالات و اخبار در سطح وب نیاز به یک سامانه قدرتمند جهت خلاصه‌سازی متون احساس می‌شود. با توجه به اینکه بر روی موضوع خلاصه‌سازی تحقیقات زیادی انجام شده و مقالات زیادی در مورد آن منتشر گردیده، اما همچنان ضعف‌های زیادی در این زمینه وجود داشته و تا رسیدن به سیستم کارآمدی که بتواند این عمل را مشابه یک عامل هوشمند انسانی انجام دهد فاصله زیادی باقی مانده است. مشکل خلاصه‌سازی در زبان فارسی به مراتب بیشتر از زبانهای دیگر است. پیچیدگی‌های زبانی موجود و عدم وجود ابزارهای با دقت مناسب برای کار با زبان فارسی از جمله مشکلات فعلی پیش رو در پردازش متون فارسی است. از این رو بررسی عملیات و روشهای انجام شده بر روی سایر زبانها، توجه به حوزه معنایی و استفاده از این رابطه‌های معنایی، بکارگیری ابزارهایی مانند گراف‌ها و روشهای آماری می‌تواند کمک شایانی به این مهم بکند.

اجزای یک متن و موجودیت‌های یک زبان عبارتند از کلمات، عبارت و جملات کامل و این اجزا به صورت کاملا با مفهوم و با ارتباط معنی دار در کنار هم قرار بگیرند به آن یک سند متنی گفته می‌شود. یکی از مهمترین شاخصه‌ها در پردازش زبان طبیعی متن‌کاوی بوده که به معنای کشف اطلاعات جدید توسط استخراج خودکار آنها می‌باشد. در متن‌کاوی نوشته‌ها و متون تحلیل شده و الگوهای با ارزشی را که در آنان مخفی هستند یافته و استخراج می‌کنند. متن‌کاوی شامل کشف ارتباط بین کلمه‌ها و جملات، طبقه بندی متون و خلاصه‌سازی بوده که حوزه کاری ما بخش خلاصه‌سازی متون می‌باشد.

بیش از نیم قرن از اولین تحقیقات بر روی خلاصه‌سازی متن توسط کامپیوتر می‌گذرد[۱]. در ابتدای شروع کار پردازش متن‌ها، آنان را با یک شبکه معنایی شبیه‌سازی می‌کردند. از آن زمان به بعد

روشهای زیادی به همراه ابزارهای قدرتمندی به کار گرفته شدند تا بتوانند پروسه‌های پردازش متن مانند آنچه که در مغز انسان انجام می‌شود را شبیه سازی کنند. یکی از این ابزارها گراف‌ها هستند که با استفاده از ساختاری مبتنی بر ریاضیات ارتباطات بین اجزای یک متن را شبیه سازی کرده که در این ساختار واحدهای متن در جایگاه رئوس و ارتباط منطقی بین این واحدهای متنی نیز در جایگاه یال‌های گراف قرار می‌گیرند. گراف‌ها با توجه به توانایی بالایی که در تحلیل محتویات معنایی و ساختار نحوی اجزای متن دارند می‌توانند در پردازش طبیعی متون بکار بروند.

۱-۲-متن‌کاوی و خلاصه‌سازی متن

بخش قابل توجهی از اطلاعات قابل دسترس در پایگاه داده‌های متنی ذخیره شده‌اند که این پایگاه‌های داده شامل مجموعه بزرگی از اسناد منابع مختلف مانند مثلا مقالات خبری، مقالات علمی ، کتاب‌ها، ایمیل‌ها و صفحات وب می‌باشند. این پایگاه داده‌ها به علت افزایش مقدار اطلاعات موجود به فرم الکترونیکی سریعا در حال رشد بوده و امروزه بیشتر اطلاعات موجود در صنعت، کسب و کار و سایر سازمان‌ها به صورت الکترونیکی و به فرم پایگاه داده متنی ذخیره شده‌اند[۳۰].

متن‌کاوی یک حوزه تحقیقاتی بوده که از ترکیب چند حوزه تحقیقاتی دیگر بوجود آمده است. این حوزه‌های تحقیقاتی عبارتند از داده‌کاوی، پردازش زبان طبیعی و بازیابی اطلاعات. اما بین داده‌کاوی و متن‌کاوی تفاوت‌هایی وجود دارند که مهمترین آنها عبارتند از:

داده‌کاوی معمولا با داده‌های ساخت یافته سر و کار دارد اما برعکس متن‌کاوی با داده‌های ساخت نیافته یا شبه ساخت یافته درگیر بوده، خصوصا زمانی که این متن‌کاوی درون یک مقاله یا سند متنی انجام شود. از طرف دیگر حتی اگر کمی ساخت‌یافتگی هم در متن وجود داشته باشد، باز هم دلیلی دیگر وجود دارد که متن‌کاوی را بسیار مشکل تر از داده‌کاوی می‌کند و آن هم مفاهیم درون متن بوده که بسیار انتزاعی تر است و به سختی می‌توان آنرا با ساختارهای ارائه دانش مرسوم مدل کرد. از

طرف دیگر در متن‌کاوی با مفاهیم مشکلی مانند کلمات هم‌خانواده (کلماتی که دارای معنی مشابه اما دیکته و تلفظ متفاوت) و کلمات متشابه (کلماتی که دیکته یا تلفظ مشابه داشته اما معنی کاملا متفاوت دارند) در گیر بوده که کار با متن‌کاوی را به مراتب مشکلتر نسبت به داده‌کاوی می‌کنند.

۱-۲-۱-کاربردهای متن‌کاوی

تعاریف گسترده ای از متن‌کاوی وجود دارد، در نتیجه در باره کاربردهای آن نیز نظریات متفاوتی وجود دارد. با اینکه متن‌کاوی یک فیلد جدید است اما نرم‌افزارهای آنالیز داده متنی از اواخر سال ۱۹۹۰ در دسترس بوده‌اند. از جمله متداولترین کاربردهای متن‌کاوی می‌توان به موتورهای جستجو اشاره کرد که در آنان کاربر یک عبارت یا کلمه را که حتی ممکن است دارای غلط املایی هم بوده را تایپ کرده و موتور جستجو که حتی با حجم بسیار بزرگی که از متن‌های موجود در وب مرتبط ترین متون را یافته و لیست می‌کند. از جمله کاربردهای دیگر متن‌کاوی می‌توان به موارد زیر اشاره نمود.

- شناسایی spam: آنالیز کردن عنوان و محتوای ایمیل برای تشخیص اینکه آیا spam هست یا نه.
- نظارت: نظارت کردن رفتار شخص یا گروهی از انسانها به صورت پنهان. نرم افزار هایی وجود دارند که می‌توانند تلفن، اینترنت و دیگر وسایل ارتباطی را برای شناسایی رفتار افراد کنترل کنند.
- شناسایی نام‌های مستعار
- روابط میان مفاهیم: از جمله واقعیتهایی که می توان از یک مجموعه متون دریافت، ارتباط و وابستگی برخی مفاهیم با مفاهیم دیگراست. به این صورت که پدیدار شدن بعضی کلمات ممکن است که به ظاهر شدن بعضی از کلمات دیگر وابسته باشد. منظور این است که هرگاه مجموعه ای از یک سری لغات دیده شوند، انتظار می‌رود که مجموعه ای دیگر از لغات نیز پس از آنان دیده شوند.
- جستجو و بازیابی: بازیابی اطلاعات دارای این هدف است که از میان مجموعه ای از مدارک و متون، آنهایی را که به نیاز اطلاعاتی کاربر ارتباط بیشتری دارند، جدا کرده و به کاربر نشان دهند. بنابر این

همیشه شرایطی وجود دارد که کاربر نتواند به نیاز اطلاعاتی خود، حتی از میان نتایج دریافت شده، برسد. به عنوان مثال ممکن است لازم باشد تا ۱۰۰ متن مطالعه شود باشد تا کاربر آنچه را که می خواهد بدست آورد. اگرچه این نوع از سیستمهای بازیابی اطلاعات از تکنولوژی های NLP و یادگیری ماشین استفاده می‌کنند اما آنچه در نهایت در این سیستمها، نقش اصلی را داراست، یک پایگاه دانش است که از طریق روشهای مبتنی بر NLP و یا روشهای آماری بر روی مدارک موجود در مجموعه ساخته شده است. در استخراج این روابط، می‌توان از متن‌کاوی استفاده کرد.

- طبقه بندی و خوشه بندی داده ها: یکی از مواردی که می تواند به کاربر در یافتن سریعتر اطلاعات مورد نظرش کمک کند، دسته بندی اطلاعات موجود است. این دسته بندی به کاربر یک نگاه کلی از آنچه در مجموعه متون موجود است می‌دهد. در ساختن این دسته بندی دو روش کلی وجود دارد. در روش اول کلاسهای از پیش تعریف شده ای از مفاهیم وجود دارند که این روشها به طبقه بندی[۱] معروفند و سیستم تلاش می‌کند که داده‌ها و مستندات را به یکی از این کلاسها، نگاشت کند و در روش دوم ایجاد این کلاسها به صورت اتوماتیک است که به خوشه بندی معروف است. در واقع در خوشه بندی قصد بر این است که مشخص شود تمرکز مفاهیم در مجموعه متون حول چه چیزهایی است و کلاس از پیش تعریف شده ای وجود ندارد.
- خلاصه‌سازی متن: منظور از خلاصه‌سازی، روند تلخیص مجموعه ای مفاهیم پایه ای از متن است تنها در چند خط می‌باشد. در این نوع از متن‌کاوی اطلاعات جدیدی از متن استخراج نشده، به این دلیل که خود نویسنده می‌دانسته است که چه چیزی را خواهد گفت و خلاصه نوشته‌های او، اطلاع جدیدی را اضافه نمی کند. موضوع اصلی تحقیق پیش رو خلاصه‌سازی متن بوده که آنرا از دیدگاه‌های متفاوت مورد بررسی قرار خواهد داد. استراتژیهای مهم در خلاصه‌سازی متن عبارت از روشهای آماری و روشهای مکاشفه بوده که به تفصیل به آنان پرداخته خواهد شد.

---

[۱] - Classification

۱-۲-۲-ارتباط متن‌کاوی و خلاصه‌سازی متن

آنچه مسلم است پروسه خلاصه‌سازی متن را نمی توان از موضوع متن‌کاوی جدا کرد. از آنجا که تحقیقات بسیاری در حوزه متن‌کاوی انجام شده است و خلاصه‌سازی هم نوعی کاوش در متن به جهت یافتن بخشهای مهم آن می‌باشد می‌توان از یافته‌های موجود در این مبحث برای خلاصه‌سازی متن کمک گرفت. مسلما مشکلات موجود در بخش متن‌کاوی در حوزه خلاصه‌سازی متن نیز وجود داشته و بررسی این مشکلات می‌تواند کمک شایانی در بخش خلاصه‌سازی متن بکند.

هدف از نگارش این بخش این است تا دست کم بخشی از این مشکلات مشخص و ریشه یابی گردند. بررسی‌ها حاکی از آن هستند که عمده مشکلات پیش روی عملیات پردازش متن و خلاصه‌سازی در حال حاضر ریشه در سه مورد زیر دارند:

- نبود زبانشناسان مجرب در گروه برنامه نویسان
- نحوـ محوربودن تئوری‌های زبانی
- نبود قدرت تعقل و تفکر در ماشین‌های پردازشگر متن

مشکل اول را می‌توان با به کارگیری نیروی متخصص رفع کرد. مشکل نوع دوم را نیز می‌توان با تجدید نظر در تئوری زبانی و توجه به معنا رفع کرد، اما دست کم در شرایط کنونی نمی‌توان قدرت تعقل و تفکر را به ماشین داد و بنابراین مشکلات مربوط به مورد سه، لااقل در وضعیت کنونی قابل حل نیستند. اما می‌توان با قاطعیت عنوان کرد که در برخورد با متون اینترنتی و همچنین اطلاعات با حجم زیاد، بویژه وقتی هدف درک کلیات متن باشد، استفاده از ماشین به جای انسان راحت تر، سریعتر و معقولانه تر است.

در آغاز به دیدگاه‌های متفاوتی که در مورد خلاصه‌سازی متن ماشینی و خلاصه‌سازی توسط انسان وجود دارد می‌پردازیم::

- عدم تفاوت بین خلاصه انسانی و خلاصه ماشینی

- برتری خلاصه انسانی بر خلاصه ماشینی
- برتری خلاصه ماشینی بر خلاصه انسانی

در برخی از باورها اعتقاد بر این است که در ترجمه و خلاصه‌سازی متن، بین انسان و ماشین تفاوت چندانی وجود ندارد. این باورها ضمن اقرار به مشکلات موجود در کار ماشین و قبول عملکرد ضعیف آن نسبت به عملکرد انسان، اظهار می‌دارند که در آینده ماشین در این زمینه‌ها از عملکردی بهتر نسبت به انسان برخوردار خواهد شد و دلیل این ادعا را سرعت بیشتر عملیات ریاضی و محاسباتی ماشین نسبت به انسان می‌داند. به بیان دیگر انسان ابداً در بسیاری از زمینه‌ها قادر به رقابت با ماشین رقابت نیست. مثلاً تصور کنید که یک فرد بخواهد دو عدد ۱۲۳۹۴۷۶۵ و ۸۹۴۵۶۳۰ را در هم ضرب، حاصل را از ۴۴۳۹۵۴ کسر و نتیجه را بر ۱۲۴۵۹ تقسیم کند. مسلماً حتی با استفاده از کاغذ، انجام چنین کاری برای انسان چندین دقیقه به طول خواهد انجامید، درحالی که تمام این محاسبات را یک ماشین حساب ساده می‌تواند با دقت لازم و تنها در چند ثانیه به انجام برساند. محققانی که بر این باورند یادگیری ریاضیات را به مراتب دشوارتر از یادگیری زبان می‌دانند و بر این نکته تأکید می‌کنند که در آینده‌ای نزدیک، کامپیوترها و برنامه‌های کامپیوتری جدید در امر پردازش اطلاعات زبانی از سرعت و دقتی بالاتر از انسان برخوردار خواهند بود. سه مزیت عمده برای تولید خودکار خلاصه به وسیله ماشین برشمرده شده است، که عبارتند از:

- اندازه خلاصه قابل کنترل است، یعنی ماشین می تواند خلاصه را با توجه به میزان فشردگی مورد نظر کاربر تهیه کند.
- محتوای آن قابل پیش بینی است.
- می‌توان مشخص کرد که هر بخش از خلاصه مربوط به کدام بخش یا بخش‌ها از متن اصلی است.

اما نظریه برتری پردازش و خلاصه‌سازی توسط انسان بر این باور است که چون عملکرد ذهن انسان با ماشین متفاوت بوده و ماشین‌های فعلی از قدرت تصمیم گیری و انتخابی بسیار کمتری از انسان برخوردار هستند هرگز قادر به پردازش اطلاعات زبانی بهتر و یا حتی همانند انسان نخواهند بود. آنچه قابل تاکید است اینکه با توجه به پیشرفت‌های به دست آمده کنونی در علوم، ماشین قادر نیست مانند انسان به ترجمه بپردازد اما این گفته به این معنی نیست که ماشین قادر به پردازش و خلاصه‌سازی نیست، بلکه بر کیفیت پایین تر این اعمال در مقایسه با انسان تأکید دارد.

در ادامه ساختار این پایان‌نامه بدین شکل است که در فصل دوم به بیان مسئله و تعریف مفاهیم پرداخته می‌شود. در فصل سوم، روش‌های خلاصه‌سازی متون که تاکنون پیشنهاد شده است، مرور خواهند شد. در فصل چهارم به روش پیشنهادی این پایان‌نامه پرداخته شده و در انتها نتایج و پیشنهادهای آینده مطرح خواهند شد.

فصل دوم

تعاریف و مفاهیم

۲-۱-مقدمه

در این فصل به بررسی تعاریف و مفاهیم خلاصه‌سازی متون و همچنین سیستم‌های خلاصه‌سازی خواهیم پرداخت. بزرگترین مسئله و مشکل در متن‌کاوی این است که یک متن یک مجموعه بسیار بزرگ از کلمات می‌باشد. اگر هر کلمه را یک المنت از یک بردار فرض کرده و حالات مختلف حضور آن را در متن مورد نظر قرار دهیم آنگاه متوجه خواهیم شد که با چه مسئله بزرگ با ابعاد زیاد روبرو هستیم. از این رو کاهش لغات که منجر به کاهش ابعاد مسئله می‌شود یکی از حیاتی ترین کارهایی است که در بخش پیش پردازش متن باید انجام داد. علاوه برآن این پیش پردازش و کاهش لغات باید آنقدر کارا و قوی باشد که بتواند در مورد متن‌های دارای نویز (دارای اشکالات گرامری و تایپی) نیز بکار رود. پیش پردازش متن به غیر از کاهش ابعاد و لغات شامل مراحل دیگری می‌باشد که برخی از آنان به شرح زیر باشند:

Case Folding: در این مرحله تمامی لغات از نظر بزرگی و کوچکی حروف به یک شکل در می‌آیند. البته این مورد بیشتر در متون انگلیسی کاربرد داشته و درمورد متون فارسی کاربردی ندارد. ضمنا همیشه سعی بر این شده است که تمامی حروف بزرگ به حروف کوچک تبدیل شوند.

Stemming: در این مرحله سعی می‌شود که تمامی لغات به شکل ریشه اصلی خود در بیایند. این تغییر حالت بدون در نظر گرفتن حالات مختلف آن مانند جمع یا مفرد بودن، زمان آن و تمامی پیشوندها و پسوندها انجام شود. به عنوان مثال دو کلمه compression و compressed به شکل واحد compress در بیایند. این عمل نیازمند داشتن دانش زبانی خاص هر زبان بوده و الگوریتم هایی نیز برای آن تا به حال پیشنهاد شده است.

Stop Words: این مرحله عبارت از حذف کلمات پر تکرار و بدون مفهوم می‌باشد. در هر متن کلماتی وجود دارند که دارای تکرار زیادی در طی متن می‌باشند. از آنجا که این کلمات دارای مفهوم کمی در

مورد اصل موضوع متن بوده می‌توان آنان را حذف کرد. از نمونه‌های این نـوع کلمـات در زبـان انگلیسـی می‌توان به can, do, does, will و غیره اشاره کرد.

N-gram: N-gramها عبارتند از شکل کوتاه شده اسامی و مفاهیم چند کلمه ای به شکل یک کلمـه بـا استفاده از حرف اول هر کلمه آنها. N-gramها کمک زیـادی بـه درک مفـاهیم یـک مـتن کـرده و بایـد حفظ شوند ولی بهتر است که به شکل یکنواخت در آیند. چون ممکن است در مـورد یـک مفهـوم چنـد N-gram وجود داشته باشد.

Tokenization: جداسازی متن یا Tokenization عبارت است تقسیم بندی متن به واحدهای کوچکتر که اغلب این واحدهای کوچک کلمات هستند. اما در برخی از موارد تعریـف کلمـه و محـدوده آن عملـی مشکل است. توکن واحدی ازمتن بوده که دارا خصوصیات زیر می‌باشد:

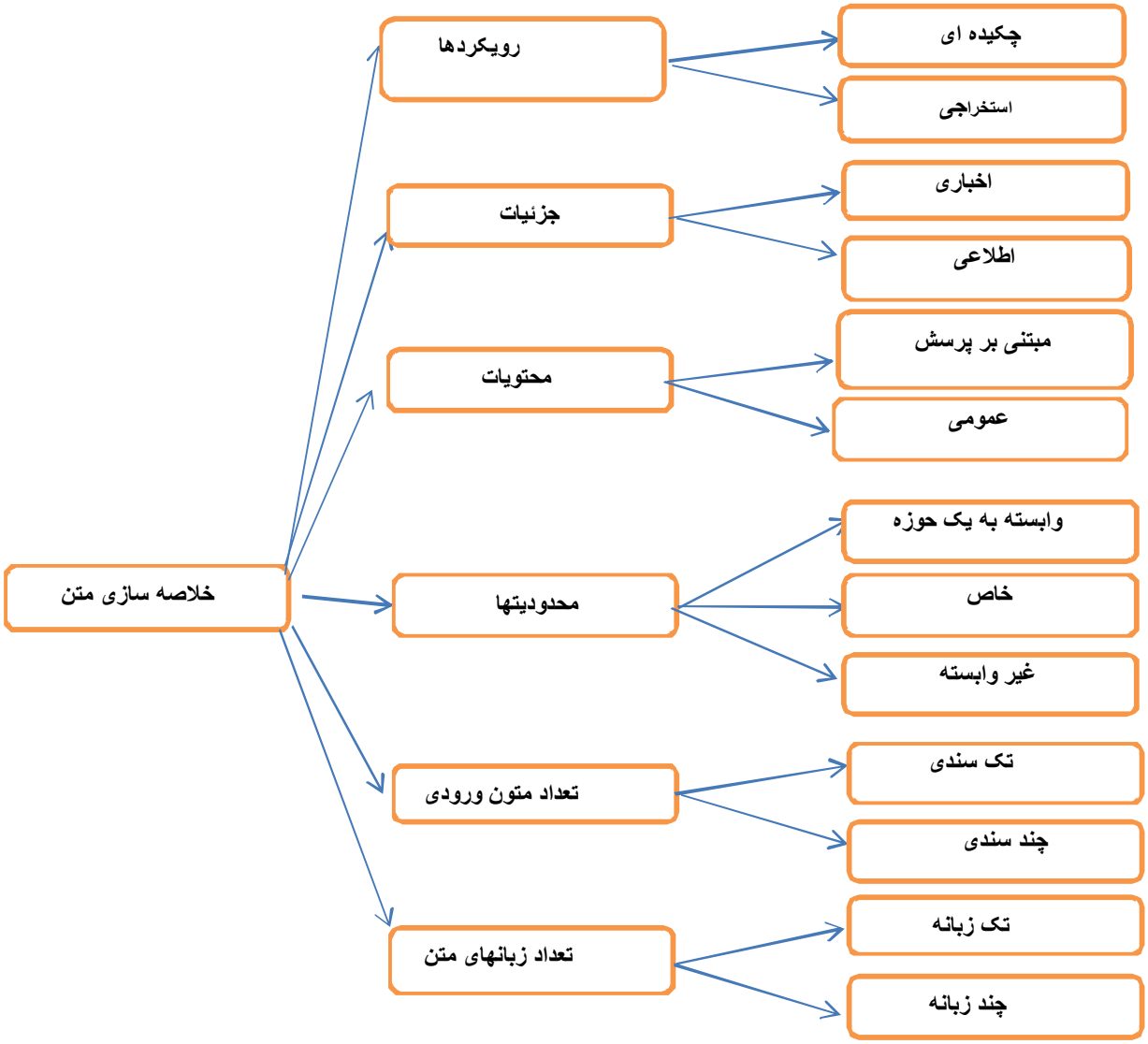
- هر مجموعه از کاراکترهای پیوسته یک توکن می‌باشد. این کاراکترها حتی می‌توانند عدد نیز باشند.
- محدوده یک توکن می‌تواند یک کاراکتر فضای خـالی، کـاراکتر انتهـای خـط و یـا علامـت گـذاری‌هـای انتهای جمله باشد.
- گاهی اوقات کاراکتر فضای خالی و علامت گذاری‌ها محدوده کلمه یا توکن نیست.

در زبانهایی مانند که محدوده کلمات با فضای خالی مشخص می‌شوند یافتن مرز کلمات راحـت بـوده امـا در همین زبانها گاهی اوقات مواردی مشاهده می‌شوند که این تشخیص مشـکل مـی‌شـود. از ایـن جملـه می‌توان به کلمات اختصاری که درون آنان نقطه بکار رفته است، کلمات موجـود در ابـرمتن هـا، کلمـاتی که با کاراکترهای سمبولیک از هم جدا شده‌اند و یا ساختارهای بزرگتـری ماننـد آدرسـهای URL اشـاره کرد. همچنین کلماتی مانند "سیب زمینی" در فارسی و "New York" در انگلیسی کـه در آنهـا فضـای خالی وجود داشته اما واقعا یک کلمه هستد[۱۴].

یکی از عملیات مهم در متن‌کاوی خوشه بندی متن می‌باشد. منظور از خوشه بندی یک متن قرار دادن جملات شبیه و مرتبط به هم درون خوشه‌ها می‌باشد. تعداد خوشه‌های تشکیل دهنده متن می‌تواند توسط کاربر و یا توسط برنامه به صورت خودکار تعیین شود. تعیین تعداد خوشه‌ها بستگی به الگوریتم خوشه بندی بکار گرفته شده در برنامه دارد. برای خوشه بندی جملات ابتدا متن به جملات تشکیل دهنده خود تقسیم می‌شود. تفکیک جملات بسادگی انجام شده و هر جمله به عنوان یک داده در نظر گرفته می‌شود. علامات جدا کننده یک جمله می‌توانند نقطه، علامت سوال و سایر علامت‌های پایانی جمله باشد. همچنین می‌توان با اعمال روشهایی که در تحقیقات بعدی بررسی خواهند شد حتی جملات بزرگ را که خود شامل دو یا چند جمله هستند را نیز شناسایی و تجزیه کرد.

۲-۲-دسته‌بندی سیستم‌های خلاصه‌سازی متن

سیستم‌های خلاصه‌سازی متن که تا به امروز ارائه شده‌اند به یک شکل نبوده اند. روشهای ارائه شده در مقالات مختلف معیارهای متفاوت و خاصی را برای خلاصه‌سازی در نظر گرفته‌اند که در این بخش سعی شده است که به صورت کلی به اکثر این روشها مختلف اشاره شود. نمودار زیر یک طبقه بندی کلی از معیارهای مختلفی که برای خلاصه‌سازی معرفی شده‌اند را نشان داده که در ادامه به توضیحی کوتاه در مورد هر کدام پرداخته می‌شود[۸].



۲-۲-۱-شکل متن خروجی

خلاصه‌سازی متن توسط ماشین به دو صورت عمده زیر تقسیم می‌شود:

- خلاصه استخراجی (Extraction)
- خلاصه چکیده ای (Abstraction)

در خلاصه‌سازی از نوع استخراجی از کلمات، جملات و پاراگرافهای موجود در متن اصلی استفاده می‌شود. به این صورت که جملات با اهمیت بیشتر پیدا شده و یا عین جملات و یا با کم کردن

بخشهایی از جمله و تبدیل آنها به جملات کوتاه تر به متن خلاصه شده منتقل می‌شوند. در این نوع خلاصه‌سازی جملات و پاراگرافهای مهم و کلیدی در متن بر اساس معیارهایی استخراج و در متن خلاصه قرار داده می‌شوند..

در خلاصه‌سازی از نوع چکیده هدف اصلی استخراج مفهوم اصلی متن ورودی می‌باشد بی آنکه لزوما ازجملات متن اصلی استفاده شود. هر خلاصه ساز از نوع چکیده از دو بخش اصلی درک مطلب و تولید کننده متن جدید تشکیل شده است. به این معنی که مفهوم متن اصلی باید به طور کلی درک شده و سپس بر اساس مفهوم درک شده متنی جدید و خلاصه تولید شود. در این نوع خلاصه‌سازی برای درک مطلب باید در متن اصلی پردازش دانش انجام شده و به هیچ عنوان پردازش متن بر اساس معنای لغوی کلمات انجام نمی شود.

۲-۲-۲-جزئیات

این نوع طبقه بندی نیز به دو دسته تقسیم می‌شود:

- خلاصه اخباری (Indicative)
- خلاصه اطلاعی (Informative)

در خلاصه‌سازی اخباری فقط ایده اصلی متن آمده که معمولا حدود ۵ تا ۱۰ درصد متن اصلی است. این نوع خلاصه‌سازی برای این بکار می‌رود که خواننده را به خواندن متن اصلی ترغیب کند. به عنوان مثال می‌توان به چکیده کوتاهی که در مورد یک فیلم یا داستان در بورشور تبلیغاتی آن اشاره کرد که فقط موجب ترغیب خواننده به دیدن فیلم و خواندن داستان می‌شود.

در خلاصه‌سازی اطلاعی خلاصه و چکیده متن اصلی آمده به صورتی که کل مباحث مهم متن در آن قرار گرفته و خواننده را از خوندن متن اصل بی نیاز کند. اندازه این نوع خلاصه بین ۲۰ تا ۳۰ درصد متن اصلی بوده و حاوی تمام نکات اصلی و مهم متن می‌باشد. داستان‌های خلاصه مثال خوبی از این نوع می‌باشند.

۲-۲-۳- محتویات

این نوع طبقه بندی نیز به دو دسته تقسیم می‌شود:

- خلاصه عمومی (Generic)
- خلاصه مبتنی بر پرسش (Query-Based)

در خلاصه عمومی سیستم به هیچ عنوان وابسته به موضوع متن نیست و فرض بر این است که خواننده هیچ دانش پایه ای در مورد متن نداشته و متن طوری خلاصه می‌شود که تمام جوانب و موضوعات مهم را در بر داشته که خواننده بدون دانش اولیه در مورد متن نیز بتواند به درک کاملی از موضوع آن برسد.

اما در خلاصه‌سازی مبتنی بر پرسش فرض بر این است که خواننده یک دانش عمومی در مورد موضوع متن داشته و فقط به دنبال اطلاعات خاصی درون متن است. در این حالت سیستم خلاصه ساز بر پایه سوالی که از سوی کاربر مطرح می‌شود خلاصه ای از متن متناسب با نوع سوال و مفاهیم مربوط به آن ایجاد می‌کند در این نوع سیستم‌ها که بیشتر از نوع استخراجی نیز می‌باشند جملاتی از متن درون بخش خلاصه شده قرار می‌گیرند که در ارتباط با پاسخ سوال مطرح شده باشند.

۲-۲-۴- محدودیت‌ها

طبقه بندی دیگری از سیستم‌های خلاصه‌سازی متن وجود دارد که بستگی به محدودیت‌های متن ورودی دارد. در این گروه نیز سه دسته وجود داشته که از قرار زیر هستند:

- خلاصه غیر وابسته (Independent)
- خلاصه خاص (Genre specific)
- خلاصه وابسته به یک حوزه (Domain dependant)

خلاصه‌سازی غیر وابسته چیزی شبیه به خلاصه‌سازی عمومی بوده و هر گونه متنی مربوط به هر حوزه را پذیرفته و یک خلاصه عمومی بدون توجه به حوزه و یا نوع و الگوی متن ایجاد می‌کنند.

در خلاصه‌سازی خاص سیستم فقط نوع خاصی از یک حوزه از متون را برای خلاصه‌سازی می‌پذیرد و این متن ورودی محدود به یک الگوی خاص می‌باشد. الگوهای متنی خاصی که از این نوع می‌توان نام برد که اهم آنها عبارت از متنهای خبری، علمی، داستانی و غیره می‌باشند. در این سیستم‌ها خلاصه‌سازی بر اساس الگویی متناسب با نوع متن ورودی انجام می‌شود.

خلاصه‌سازی وابسته به یک حوزه نیز مانند خلاصه‌سازی خاص بوده اما به صورت کلی تر به مسئله نگاه کرده که متون یک حوزه خاص را پذیرفته و خلاصه می‌کند.

۲-۲-۵- تعداد متون ورودی

سامانه‌های خلاصه‌سازی خودکار را می‌توان بر اساس متن‌های ورودی به آنها به دو دسته کلی زیر تقسیم کرد:

- خلاصه سازهای تک سندی (Single document)
- خلاصه سازهای چند سندی (Multi document)

در سامانه‌های خلاصه‌سازی تک سندی ورودی تنها یک متن بوده اما در سامانه‌های چند سندی ورودی چندین متن اما با موضوع مشترک بوده که ممکن است از دیدگاههای متفاوت بررسی شده باشند. در کل پیچیدگی در مدل‌های تک سندی به مراتب کمتر از مدل‌های چند سندی است.

۲-۲-۶- تعداد زبان‌های متن

یک متن ورودی به سیستم خلاصه ساز بر اساس تعداد زبانهای بکار گرفته شده در متن به دو دسته تقسیم می‌شود:

- متن تک زبانه (Mono lingual)
- متن چند زبانه (Multi lingual)

خلاصه ساز متن تک زبانه فقط متونی را برای خلاصه‌سازی می‌پذیرد که فقط به یک زبان خاص مانند فارسی، انگلیسی و غیره نوشته شده اند. اما یک خلاصه ساز چند زبانه متون با زبانهای متفاوت را پذیرفته و قادر به خلاسه سازی آنان می‌باشد.

۲-۳-خلاصه‌سازی استخراجی

همانطور که در قبل اشاره شد در خلاصه‌سازی از نوع استخراجی از جملات موجود در متن اصلی استفاده می‌شود. به این صورت که جملات با اهمیت بیشتر پیدا شده و یا عین جملات و یا با کم کردن بخشهایی از جمله و تبدیل آنها به جملات کوتاه تر به متن خلاصه شده منتقل می‌شوند. برای تعیین جمله‌های مهم باید پارامترهای مختلفی را در نظر گرفت. پس از تعیین مقدار عددی هر یک از این پارامترها، جمع جبری آنان به عنوان وزن نهایی هر جمله در نظر گرفته شده و در نهایت جملات با بیشترین وزن و با به نرخ فشرده سازی مورد نظر انتخاب و به ترتیب حضور در متن اصلی درمتن خلاصه شده ظاهر می‌شوند. با توجه به اینکه برخی از پارامترهای ذکر شده از اهمیت یک جمله می‌کاهند مقدار عددی آنان به صورت منفی محاسبه می‌شود[۲]. این پارامترها و ویژگی‌ها عبارتند از:

۱. کلمات کلیدی (Content word (Keyword) feature)

کلمات کلیدی اغلب اسم هستند و توسط معیار tf * idf تعیین می‌شوند. جملاتی که دارای این کلمات هستند شانس بسیار بیشتری برای قرار گرفتن در بخش خلاصه شده دارند. روش‌های دیگری نیز برای استخراج کلمات کلیدی پیشنهاد شده‌اند که از آن جمله می‌توان به روشهایی که کلمات کلیدی را با سه معیار زیر استخراج می‌کنند[۳]:

- تجزیه و تحلیل مورفولوژیکی

تجزیه و تحلیل مورفولوژیکی نقش مهمی در پردازش زبان طبیعی داشته و کمک شایان به رفع ابهام در کلمات می‌کند. این بخش به بررسی و مطالعه ریشه کلمات، پیشوندها و پسوندهای متصل به یک کلمه و بطور کلی شکل گیری لغات می‌پردازد.

- استخراج عبارات اسمی و امتیازدهی به آنان
- استخراج عبارات اسمی، خوشه بندی و امتیازدهی به آنان

۲. میزان شباهت با عنوان متن

یکی دیگر از معیارهامیزان شباهت بین جمله و عنوان مقاله است. شباهت جمله با عنوان موضوع عبارت است از اشتراک لغات بین جمله و عنوان موضوع بوده و جملاتی که شامل کلمات موجود در عنوان متن باشند از درجه اهمیت بالایی برخوردارند. در اکثر روشهای پیشنهادیمیزان شباهت هر جمله با عنوان متن با رابطه زیر محاسبه می‌شود:

$$X_i = |Si \cap T| / \sqrt{|Si| * |T|}$$

در این معادله صورت کسر تعداد کلمات مشترک بین جمله مربوطه و عنوان متن و مخرج کسر نیز جذر حاصلضرب طول جمله در طول عنوان متن بر حسب کلمه می‌باشد.

۳. موقعیت مکانی جملات

موقعیت مکانی جملات از اهمیت زیادی برخوردار است. معمولا در هر متن جمله اول و آخر هر پاراگراف از اهمیت بالایی برخوردار بوده و با خواندن جمله اول و آخر هر پاراگراف با احتمال بسیار بالا می‌توان منظور پاراگراف را استخراج کرد. این موضوع برای متنهای انگلیسی بارها و بارها آزموده شده و نتایج آن بررسی و صحت آن تایید شده است. بررسی بر روی متون فارسی نیز این قاعده را تا حدود زیادی تایید می‌کند. از این رو این جملات شانس بسیار بالایی برای قرار گرفتن در متن خلاصه دارند. رابطه زیر برای محاسبه میزان اهمیت موقعیت مکانی یک جمله در متن به کار می‌رود:

$$PS_i = 1 + (0.5 * (PS_i + C) * \log (0.5 * (PS_i + C))) + ((1 - 0.5 * (PS_i + C)) * \log(1 - 0.5 * (PS_i + C)))$$

$$X_i = PS_i / Max_{j=1:n} (PS_i)$$

در این رابطه PS_i ارزش موثر موقعیت نسبی جمله i ام بوده که بر اساس مدل آنتروپی محاسبه می‌شود. مقدار C یک مقدار ثابت بین صفر تا یک بوده و مقدار X_i به عنوان معیار ارزش موقعیت مکانی جمله بر اساس تقسیممیزان ارزش مکانی هر جمله برمیزان ارزش بالاترین موقعیت مکانی جملات در کل متن بدست می‌آید.

۴. کلمات و عبارات پراهمیت

برخی از کلمات و عبارات خاصی درون جمله وجود داشته که اهمیت یک جمله را افزایش می‌دهند. این عبارات حاوی اطلاعات خاص و مهمی بوده که تعییننمیزان اهمیت هر کدام آنها تا حدودی به نوع متن نیز بستگی دارد. این کلمات و عبارات را می‌توان به دسته‌های زیر تقسیم بندی کرد:

- اعداد (شامل تاریخ، زمان، قیمت، درصد، تعداد و وزن)
- اسامی خاص (شامل افراد، مکان، زمان)

تعیین میزان اهمیت یک جمله با توجه به کلمات و عبارات پر اهمیت آن با توجه به رابطه زیر و با محاسبه نسبت وجود تعداد این کلمات به تعداد کل کلمات جمله بدست می‌آید:

$$X_i = |S_{imp}| / |S_i|$$

۵. کلمات با حروف بزرگ و کلمات اختصاری

کلمات با حروف بزرگ و یا کلمات اختصاری دارای مفاهیم ویژه ای می‌باشند. جملات حاوی آنها نیز حتما دارای بار موضوعی بیشتری نسبت به سایر جملات متن می‌باشند. کلمات با حروف بزرگ در زبانهای با حروف لاتین وجود داشته و در زبان فارسی چنین کلماتی وجود ندارد، اما کلمات اختصاری در همه زبانها وجود داشته و در زبان فارسی هم می‌توان از آنان می‌توان برای تعیین اهمیت یک جمله استفاده کرد.

۶. **عبارات نشانه ای (Cue-Phrase)**

جملات شامل عباراتی مانند "در نهایت"، "در نتیجه"، "در این مقاله"، ..... حتما شامل موضوع مهمی بوده و حتما باید حفظ شوند. این جملات از درجه اهمیت بالایی برخوردار هستد.

۷. **کلمات کلیدی و خاص**

کلمات کلیدی و خاص عبارتند از کلمات خاصی که دارای معنایی خاص بوده و ممکن است با کلمات کلیدی متن فرق می‌کند. این کلمات می‌توانند موضوعات و مباحث علمی مانند "ریاضیات"، روانشناسی" و غیره باشند و وجود آنان در یک جمله اهمیت آن را بالا می‌برد.

۸. **جملات حاوی کلمات با فونت و جلوه‌های فونت متفاوت**

برخی کلمات بنا به اهمیتی که دارند و یا دارای بار موضوعی خاصی که هستند با حروف بزرگ آغاز شده و یا به صورت بولد، ایتالیک و یا زیر خط دار در متن قرار می‌گیرند. جملات حاوی آنها نیز شانس بالاتری برای قرارگرفتن در متن خروجی خواهند داشت.

۹. **میزان پیوستگی و شباهت یک جمله با سایر جملات**

برای هر جمله موجود در متنمیزان شباهتش با سایر جملات بررسی شده و به تعداد جملات مشابه به آن امتیاز داده می‌شود. این عمل برای تمام جملات متن تکرار شده و جملات با امتیاز بالاتر شانس بیشتری برای حضور در متن خلاصه شده خواهند داشت. معیار شباهت بین دو جملهمیزان تعداد کلمات مشابه، طول جمله و موارد و معیارهای ذکر شده قبل می‌باشد.

۱۰. **میزان شباهت یک جمله با جمله موضوعی پاراگراف**

در این بخشمیزان شباهت هر جمله از پاراگراف با جمله موضوعی آن پاراگراف بررسی می‌شود. معمولا جمله موضوعی هر متن جمله ای در هر پاراگراف می‌باشد که بیشترین بار موضوعی را داشته و اغلب جمله اول هر پاراگراف می‌باشد. معیارهای شباهت در این بخش نیز همان معیارهای شباهت در بخش قبل می‌باشند.

۱۱. عبارات خاص

عبارات خاص عباراتی هستند که در متن توسط علائمی مانند " "، << >>، ..... مشخص شده اند. بررسی متون خلاصه شده توسط عامل انسانی نشان می‌دهد که جملات حاوی این عبارات از توجه بیشتری برخوردار هستند. در اکثر روشهای پیشنهاد شده از رابطه زیر برای تعیینمیزان اهمیت جملات حاوی این عبارات استفاده می‌شود:

$$X_i = |S_{ip}| / |S_i|$$

در معادله فوق S_{ip} تعداد عبارات خاص قرار گرفته در جمله، S_i طول جمله بر اساس تعداد کلمات و X_i درجه اهمیت جمله بر اساس این معیار می‌باشد.

۱۲. وجود کلمات اشاره و تعیین کننده

کلمات اشاره کلماتی هستند که برای اشاره به شخص یا موضوعی خاص به کار رفته و موجب افزایش اهمیت آن شخص یا موضوع و به طبع افزایش اهمیت خود جمله می‌شوند. به عنوان مثال برای افراد می‌توان به کلماتی مانند "استاد"، "آقای"، "خانم"، "دکتر"، "سردار"، ..... اشاره کرد. برخی از کلمات و عبارات نیز وجود دارند که بیشتر در ابتدای جمله آمده و وجود آنان نیز براهمیت جمله می‌افزاید. از این دسته می‌توان به عباراتی مانند "در نتیجه"، "در پایان"، "بنابراین"، ..... اشاره کرد.

برخی از کلمات که در ابتدای یک جمله قرار می‌گیرند نشان دهنده این موضوع هستند که این جمله در تکمیل جمله قبل آمده و می‌توان آنرا ندیده گرفت. این کلمات عبارتند از کلماتی مانند "چون"، "بعلاوه"، "همچنین"، "مانند"، ..... که وجود آنها در ابتدای هر جمله موجب کاهش امتیاز آن جمله و شانس حضور آن در متن خلاصه شده خواهد شد مگر اینکه موارد ذکر شده در بخشهای قبلی امتیاز جمله مربوطه را افزایش دهد. وزن مربوط به جملات حاوی این کلمات و عبارات اشاره در هر دو حالت مثبت و منفی از رابطه زیر محاسبه می‌شود:

$$X_i = (|S_i| + |S_{pi}| - |S_{ni}|) / 2*|S_i|$$

در این معادله S_i تعداد کلمات جمله i ام، S_{pi} تعـداد کلمـات اشـاره مثبـت و S_{ni} تعـداد کلمـات اشـاره منفی موجود در آن جمله می‌باشد.

۱۳. ضمایر

جملاتی که در آنان به جای استفاده از اسامی از ضمایر استفاده شده اسـت از درجـه اهمیـت کمتـری برخوردار خواهند بود. چون ضمایر به مواردی اشاره می‌کنند که در جملات قبلی آمده و توصـیف شـده اند. به عبارت دیگر جمله حاوی ضمیر حاوی توضیحات بیشتر در مورد جمله ای اسـت کـه ضـمیر بـه آن ارجاع می‌دهد. البته موقعیت ضمیر در جمله نیز دارای اهمیت است. مثلا اگر ضمیر در ابتدای یـک جمله قرارگرفته باشد به طور قطع می‌توان گفت آن جمله برای توصیف بیشـتر جمـلات قبـل آمـده و اهمیت بسیار کمتری دارد. اما اگر ضمیر در کلمات می‌انـی و انتهـایی یـک جملـه قـرار گرفـت درجـه اهمیت جمله کمی بیشتر می‌شود. جملات حاوی ضمیر در صورتی که موارد ذکر شـده در مراحـل بـالا را دارا باشند موجب افزایش درجـه اهمیتشـان مـی‌شـوند.میزان اهمیـت جملـه ای کـه حـاوی ضـمیر می‌باشد از رابطه زیر بدست می‌آید:

$$X_i = - (BP_i + |S_{Pi}|) / (1 + |S_i|)$$

در این رابطه مقدار $|S_{pi}|$ تعداد ضمایر موجود در جملـه بـوده و مقـدار BP_i مقـدار ثـابتی اسـت کـه در صورتی که ضمیر در سه کلمه اول جمله بوده برابر یک و در غیر این صورت برابر صفر خواهد بـود. $|S_i|$ نیز طول جمله مورد نظر از نظر تعداد کلمه خواهد بود. همانطور که در معادله مشـخص مـی‌باشـد ایـن مقدار در یک منفی ضرب شده که هر چه مقدار آن بزرگتر باشد تاثیر منفی آن در امتیاز جملـه بیشـتر است.

۱۴. طول جملات

اندازه و طول یک جمله تاثیر زیادی در قرار گیری آن در متن خلاصه دارد. بهترین جملات بـرای قـرار گیری درخروجی خلاصه شده جملات با طول متوسط بوده و جملات بسیار کوتاه یا بلنـد بهتـر اسـت

که در خروجی قرار نگیرند. برای اینکه جملات خیلی کوتاه و یا خیلی بلند انتخاب نشوند می‌توان حد آستانه ای دو سطحی تعیین و جملات خارج از آن حد را حذف کرد. اما گاهی در اثر اعمال معیارهای قبلی یک جمله بلند نیز دارای اهمیت شده که نمی توان آنرا صد در صد حذف کرد. برای این کار می‌توان در صورت امکان آن را به دو یا چند جمله کوچکتر تقسیم کرده و با اهمیت ترین آنان را در متن خلاصه شده قرار داد. در غیر این صورت می‌توان برای طول خارج از حد آستانه امتیاز منفی در نظر گرفت که با جمع جبری با سایر معیارها از ارزش جمله کم کند. معادله زیر برای مشخص کردن تاثیر طول جمله درمیزان امتیاز در نظر گرفته می‌شود:

$$X_i = - RS_i * Log (RS_i) - (1 - RS_i) * Log (1 - RS_i)$$

در این معادله RS_i از رابطه زیر بدست آمده که در آن نیز S_i طول جمله مورد نظر و $Max_{j=1:n}(|S_i|)$ اندازه بلند ترین طول جمله متن از نظر تعداد کلمه می‌باشد. مقدار منفی ضرب شده در معادله بالا مشخص کننده تاثیر منفی طول جمله‌های بیش از کوتاه یا بلند است:

$$RS_i = |S_i| / Max_{j=1:n}(|S_i|)$$

فصل سوم

روشهای خلاصه‌سازی

۳-۱ مقدمه

اگر چه متن‌های خلاصه شده توسط انسان معمولا از نوع چکیده ای هستند ولی اکثر روشهای خلاصه‌سازی ماشینی موفق ایجاد شده از نوع استخراجی هستند. در خلاصه‌های ایجاد شده توسط ماشین اغلب یک خلاصه استخراجی خوب دارای نتیجه ای بسیار بهتر از نوع چکیده ای بوده و این به دلیل وجود مشکلات خاصی که سر راه خلاصه چکیده ای مانند ابهامات معنایی، استنباطی موجود در تولید زبان طبیعی در هر زبانی وجود دشته و عملا کار را بسیار سخت تر در مقایسه با استخراج جملات مهم متن می‌کند. در واقع سیستم‌های تولید کننده خلاصه چکیده ای هنوز تا رسیدن به حد کمال مطلوب فاصله دارند. اولین رویکردهای خلاصه‌سازی از نوع استخراجی بوده و جملات مهم موجود در یک متن را با توجه به ویژگی‌های مکاشفه ای ساده آن مانند موقعیت جمله در متن، فرکانس کلمات موجود در آن و برخی از کلمات کلیدی مهم مرتبط با موضوع متن انجام می‌دادند[۱][۳۳]. مراحل خلاصه‌سازی متن عبارت از تشخیص موضوع متن، تفسیر متن و تولید متن خلاصه شده می‌باشند.

۳-۲-رویکردهای خلاصه‌سازی متون

رویکردهای متفاوت خلاصه‌سازی متن دسته‌های زیر تقسیم می‌شوند:

۱- رویکردهای مبتنی بر روشهای آماری (Statistical_based)

۲- رویکردهای مبتنی بر زنجیره لغوی (Lexical Chain_based)

۳- رویکردهای مبتنی بر تئوری گراف (Graph_based)

۴- رویکردهای مبتنی بر خوشه بندی (Cluster_based)

۵- رویکردهای مبتنی بر منطق فازی (Fuzzy logic_based)

در ادامه برخی از رویکردهای ذکر شده بررسی شده و مقالاتی از هر کدام مورد تجزیه و تحلیل قرار می‌گیرد.

۳-۳-استفاده از روشهای آماری در خلاصه‌سازی متن

استفاده از روشهای آماری رویکردی موثر در خلاصه‌سازی متن بوده که در برخی از مقالات و تحقیقات اخیر بکار گرفته شده است. رایجترین روشهایی که در این تحقیقات استفاده شده‌اند دو رویکرد بیزین کلاسیفایر (Bayesian Classifier) و ارتباط مفهومی (Concept Relevance) می‌باشند. یکی از روشهای آماری که در برخی از مقالات از آن برای خلاصه‌سازی متن استفاده شده است روش LSA می‌باشد[۳۲].

آنالیز معنایی پنهان LSA (Latent Semantic Analysis): LSA یک روش جبری-آماری برای استخراج ساختارهای پنهان معنایی کلمات و جملات می‌باشد [۱۱]. این روش یک رویکرد بدون ناظر بوده که عمل استخراج ساختارهای متنی را بدون یادگیری و نیاز به دانش خارجی انجام می‌دهد. روش مورد نظر از محتویات موجود در متن ورودی استفاده کرده و اطلاعاتی نظیر کلماتی که همراه با هم در یک جمله بکار رفته و یا کلمات مشترک استفاده شده در جملات متفاوت را بدست می‌آورد. ایده این روش به این صورت است که هر چه کلمات مشترک بین جملات متفاوت بیشتر باشند دلیلی بر وابستگی معنایی این جملات به هم می‌باشد. LSA همچنین توانایی نمایش معنی و مفهوم جملات و لغات را به طور همزمان داشته که این دو لازم و ملزوم یکدیگرند. این روش از یک متد جبری به نام تجزیه مقدار منفرد یا SVD (Singular Value Decomposition) برای تشخیص ارتباط بین جملات و کلمات استفاده می‌کند. SVD علاوه بر اینکه قادر به مدل کردن ارتباط بین کلمات و جمله بوده توانایی کاهش نویز را نیز داشته که موجب بهبود دقت نیز می‌شود.

الگوریتم خلاصه‌سازی که بر پایه روش LSA باشد شامل سه مرحله زیر است:

۱- ایجاد ماتریس ورودی: ابتدا با متن ورودی یک ماتریس ایجاد می‌شود.

۲- SVD: اعمال روش SVD بر روی ماتریس ایجاد شده جهت مشخص کردن رابطه بین کلمات، عبارات و جملات.

۳- استخراج جملات: استفاده از نتیجه خروجی SVD برای استخراج جملات مهم.

LSA دارای محدودیت هایی نیز می‌باشد. مهمترین این محدودیت‌ها از قرار زیر هستند:

۱- این الگوریتم از اطلاعات مربوط به ترتیب و آرایش کلمات در جمله، ارتباطات گرامری و ریخت شناسی کلمات استفاده نمی کند. هر چند این اطلاعات می‌توانند برای درک بهتر لغات و متن مفید باشند.

۲- الگوریتم از هیچ دانش لغوی به جز اطلاعات موجود در متن ورودی استفاده نکرده و به هیچ عنوان یک پایگاه داده لغوی دیگر را بکار نمی گیرد.

۳- با افزایش تعداد کلمات متفاوت و داده‌های ناهمگن کارایی الگوریتم بشدت کاهش می‌یابد. کاهش کارایی به دلیل پیچیدگی بالای زمانی و حافظه ای روش SVD می‌باشد.

ایجاد ماتریس ورودی: برای اینکه کامپیوتر بتواند متن ورودی را درک و بر روی آن اعمال محاسباتی انجام دهد متن مورد نظرباید درون یک ساختمان داده قرار گیرد[۱۱]. در این الگوریتم از ماتریس برای قرار دادن متن استفاده می‌شود. در این ماتریس ستونها مشخص کننده جملات، سطرها مشخص کننده کلمات و درایه‌های ماتریس نیز جهت درج اهمیت کلمات در جمله بکار می‌روند. برای پر کردن مقادیر درایه‌های ماتریس می‌توان رویکردهای متفاوتی را بکار برد. از آنجا که همه کلمات موجود در متن در همه جملات موجود نمی باشند ماتریس تشکیل شده یک ماتریس اسپارس خواهد بود.

اولین قدم در تولید ماتریس ورودی $m * n$ بوده که در آن m تعداد ترمها و n تعداد جملات می‌باشند. ماتریس ایجاد شده را می‌توان به صورت $A=[A_1, A_2, \ldots, A_n]$ نشان داد که در آن A_i نشان دهنده بردار وزن جمله i ام متن است. هر ترم می‌تواند یک کلمه یا عبارت بوده که در جمله i ام دیده شده و یا قبل از تولید ماتریس توسط عملیات پیش پردازش ایجاد شده باشد.

با توجه به اینکه LSA از الگوریتم SVD برای مقدار دهی بهمیزان اهمیت کلمات در جمله استفاده کرده و با افزایش سایز ماتریس پیچیدگی زمانی و حافظه آن شدیدا افزایش خواهد یافت می‌توان با کاهش اندازه ماتریس ورودی کارایی الگوریتم را افزایش داد. به دلیل اینکه ستونهای ماتریس یاد شده

برابر با تعداد جملات متن بوده و جملات متن را در مرحله اول نمی توان کاهش داد این کـاهش سـایز در سطرها انجام می‌شود. سطرها حاوی لغات بوده که می‌توان با اعمـال عملیـات پـیش پـردازش مـتن مانند حذف Stop World ها، بکارگیری ریشه کلمات، استفاده از عبارات به جای کلمات تا حـد ممکـن و غیره اندازه آن را کاهش داد. ناگفته نماند که عملیات پیش پردازشی ذکر شده شدیدا بـه زبـان مـتن ورودی وابسته بوده و تاثیر عملیات پیش پردازشی مختلف در زبانهای گوناگون متفـاوت خواهـد بـود. در این مرحله برخی از عملیات پیش پردازشی در این الگوریتم بررسی می‌شود[۱۱]:

قطعه بندی (Segmentation): قبل از شروع به عمل خلاصه‌سازی، ابتدا باید کوچکترین واحد قابـل استخراج متن را مشخص کرد. استخراج کوچکترین واحد متن tokenization نام داشته که مـی‌تـوانـد پاراگراف، جمله و یا عبارت بوده و معمولا در اغلب روشهای خلاصه‌سازی اسـتخراجی جملـه مـی‌باشـد. جداسازی جملات و قرار دادن آنان درون ماتریس نیازمند یافتن مرزهای جمله بوده کـه در ایـن مرزهـا در اغلب زبانها با فضای خالی و نقطه گذاری مشخص می‌شود. اما در زبانهایی مانند چینـی و ژاپنـی کـه مرز کلمات در آنها مخص نبوده این عمل نیازمند عملیات آماری پیچیده تری می‌باشد.

یافتن ریشه لغات (Stemming): در یک متن یک لغت می‌تواند به فرمت‌های متفاوتی ماننـد جمـع، مفرد، حال، گذشته و غیره ظاهر شود که در بیشتر اوقات مفهوم همگی یکی بوده و نیازی نیسـت آنـان را به صورت متفاوت در نظر گرفته شوند. از این رو برای کاهش تعداد لغات موجود در متن ریشـه آنهـا را استخراج کرده و در درایه مورد نظر ماتریس قرار مـی‌دهنـد. ریشـه یـابی همچنـین موجـب کـاهش حجم ماتریس و همچنین کاهش نویز می‌شود.

فیلتر کردن Stop word ها: هر متن شامل کلماتی بوده که بار مفهومی خاصی نداشته و صرفا از نظـر گرامری برای تشکیل جمله لازم می‌باشند. از جمله این کلمات در انگلیسـی مـی‌تـوان بـه the, is, a, ... اشاره کرد. در روشهای خلاصه‌سازی استخراجی نیازی به عملیـات پردازشـی بـر روی ایـن کلمـات

نبوده و می‌توان آنان را در ماتریس مورد نظر قرار نداد. قرار ندادن Stop wordها در ماتریس ورودی موجب کاهش سطرهای ماتریس شده ومیزان نویز را شدیدا کاهش می‌دهد.

حذف Stop wordها از یک متن به دو شکل امکان پذیر است. در روش اول یک لیست از این کلمات توسط فرد خبره ایجاد شده و در هنگام درج کلمات درون ماتریس با کلمات لیست مقایسه و در صورت عدم تطبیق در ماتریس جای می‌گیرند. این روش نیازمند ایجاد اولیه لیست بوده و وابسته به زبان ورودی متن می‌باشد. اما در روش دوم می‌توان این حذف را به صورت آماری انجام داد. به این صورت که فرکانس حضور لغات مختلف در متن را محاسبه و با تعیین یک حد آستانه، Stop word را تشخیص داد. تعیین این حد آستانه نیز نیازمند محاسبات آماری خاصی است که می‌توان مشخص کرد. روش دوم وابسته به زبان نبوده و می‌توان آنرا برای متون با زبانهای مختلف بکار برد.

تعیین وزن درایه‌های ماتریس:میزان ارزش هر درایه از ماتریس نشان دهندهمیزان اهمیت کلمه مورد نظر درون جمله بوده و مقدار آن در تعیین نتیجه خروجی الگوریتم SVD تاثیر زیادی دارد. رویکردهای متفاوتی برای تعیین وزن هر درایه وجود داشته که در ادامه به آنان پرداخته می‌شود[۱۱].

۱- فرکانس تکرار کلمات: در هر درایه عدد تکرار کلمه مربوطه در جمله قرار داده می‌شود.

۲- نمایش یک عدد باینری: در صورت وجود لغت مورد نظر در جمله عدد یک و درصورت عدم وجود مقدار صفر در درایه مورد نظر قرار می‌گیرد.

۳- Tf_Idf (Term Frequency–Inverse Document Frequency): درایه مورد نظر با مقدار Tf_Idf پر می‌شود. با توجه به تعریف این معیار هر چه تعداد تکرار کلمه مورد نظر در جمله زیاد اما در کل متن کم باشد این مقدار بزرگتر است. مقدار معیار Tf_Idf با معادله زیر بدست می‌آید:

$$Tf_Idf = tf - idf$$

مقدار tf (term frequency) با معادله زیر محاسبه می‌شود. در این معادله مقدار n(I, j) برابر با تکرار کلمه i ام در جمله jام بوده و $\sum_k n\ (k,j)$ برابر با مجموع کلمات موجود در جمله jام می‌باشد:

$$tf\ (I, j) = n\ (I, j)\ /\ \sum_k n\ (k,j)$$

مقدار inverse document frequency (idf) نیز با معادله زیر محاسبه خواهد شد. در این معادله مقدار |D| تعداد جملات متن ورودی بوده و d_i نیز تعداد جملاتی است که کلمه i در آن دیده شده است:

$$idf = \log (|D| / d_i)$$

۴- لگاریتم آنتروپی (Lig Entropy) : در این حالت درایه مورد نظر با مقدار لگاریتم آنتروپی کلمه مورد نظر پر می‌شود. این مقدار نشان دهنده این است که کلمه مورد نظر به چهمیزان دارای اطلاعات خاصی در جمله است. روند محاسبه این مقدار به صورت زیر است:

$$P (i, j) = f_{ij} / gf_i$$
$$sum = \sum_j p\ (i,j) * log2\ p(i,j)$$
$$global\ (i) = 1 + (sum / \log2\ n)$$
$$local\ (I, j) = \log2\ (1 + f\ (I, j))$$
$$Log_Entropy = global * local$$

در این معادله‌ها مقدار p (I, j) احتمال حضور کلمه I ام در جمله j ام، gf_i تعداد دفعات حضور کلمه i ام در تمام جملات، f (i, j) تعداد دفعات حضور کلمه i ام در جمله j ام و n نیز تعداد کل جملات متن است.

۵- نوع ریشه (Root Type): اگر کلمه مورد نظر اسم بود درایه مورد نظر با تعداد تکرار آن در متن پر شده و در غیر این صورت مقدار صفر خواهد گرفت.

۶- مقدار Tf_Idf تعدیل شده: در این حالت مقدار درایه مورد نظر با مقدار Tf_Idf پر شده وپس از اینکه تمام درایه‌های سطر مورد نظر مقدار گرفتند میانگین مقادیر محاسبه شده وتمام کلماتی که درایه مربوط به آنها دارای آنها دارای مقداری کمتر از میانگین محاسبه شده بود مقدار صفر خواهند گرفت. مقدار دهی درایه‌ها با این روش موجب حذف نویز از ماتریس ایجاد شده خواهد شد.

الگوریتم‌های متفاوتی در زمینه خلاصه‌سازی متن مبتنی بر روش LSA ارائه شده اند. در این مرحله برخی از آنان بررسی می‌شوند:

الگوریتم ارائه شده توسط Gong & Liu, 2001: در تحقیق ارائه شده توسط این دو فرد دو الگوریتم برای خلاصه‌سازی عمومی متن پیشنهاد کرده که الگوریتم اول از روشهای IR استاندارد جملات با اهمیت و ارتباط بین آنها را تشخیص و الگوریتم دوم LSA را برای این امر انتخاب کرده

است. با توجه به اینکه تحقیق حاضر تکیه بر الگوریتم‌های مبتنی بر LSA داشته بخش دوم از مقاله این دو فرد مورد بررسی قرار می‌گیرد[۱۲].

الگوریتم ارائه شده توسط Gong و Liu یکی از روشهای پایه و اصلی خلاصه‌سازی استخراجی متن بر پایه LSA می‌باشد. در این روش پس از قرار دادن متن درون ماتریس ورودی و اعمال الگوریتم SVD، از ماتریس مفاهیم استخراج شده x جمله (V^T) برای انتخاب جملات مهم استفاده می‌شود. در این ماتریس ترتیب سطر نشان دهنده اهمیت مفاهیم موجود است. به عنوان مثال اولین سطر ماتریس نشان دهنده مهمترین مفهوم استخراج شده است. درایه‌های هر سطر نشان دهنده ارتباط بین جمله و مفهوم مورد نظر بوده و مقدار بالای هر درایه مشخص کننده بیشتر مرتبط بودن جمله مربوطه با مفهوم مورد نظر می‌باشد. در این رویکرد جملات به ترتیب اهمیت و با توجه به حد آستانه تعریف شده بر اساس تعداد جملات متن خلاصه شده در متن خروجی ظاهر می‌شوند. تعداد جملات خروجی به عنوان یک پارامتر ورودی در ابتدای الگوریتم دریافت می‌شود.

روش ارائه شده دارای معایبی نیز بوده که از قرار زیر است[۱۳]:

۱- ممکن است در هر مفهوم (concept) چندین جمله با اهمیت وجود داشته باشد اما این الگوریتم از هر مفهوم فقط یک جمله با بالاترین درجه اهمیت را بر می‌گزیند. ممکن است جمله انتخابی از مفهوم بعدی از درجه اهمیت کمتری نسبت به جمله با درجه اهمیت دوم در مفهوم اول باشد.

۲- از آنجا که تعداد جملات خروجی در این روش توسط یک پارامتر ورودی دریافت می‌شود، اگر تعداد جملات قرار گرفته در خلاصه خروجی زیاد باشد با توجه به مشکل شماره یک قطعا جملات بی اهمیت تر نیز در متن قرار خواهند گرفت.

۳- ممکن است که conceptها با درجه اهمیت یکسان ایجاد شده اما درجه اهمیت جملات درون هر کدام تفاوت زیادی داشته باشند که در انتخاب جمله خروجی تاثیر خواهند گذاشت.

الگوریتم ارائه شده توسط Steinberger & Jezek, 2004: با توجه به مشکلات ذکر شده (Josef Steinberger & Karel Jezek, 2004) تغییراتی را در SVD اعمال شده ایجاد کرده و معایب ذکر

شده را بهبود بخشیده اند[۱۵]. در روش پیشنهادی ایشان مانند روش‌های قبل ابتدا توسط SVD ماتریس جملات ایجاد شده و سپس برای هر بردار جمله در ماتریس اولیه V طول آن بدست آمده است. معادله زیر نحوه محاسبه طول بردار جمله را مشخص می‌کند:

$$S_k = \sqrt{\sum_{i=1}^{n} V_{k,i}^2 * \alpha_i^2}$$

در این معادله S_k طول بردار k امین جمله در فضای بردار بوده که یک معیار بسیار خوب برای خلاصه‌سازی است، n ابعاد فضای بردار جدید بوده که پارامتر برای روش است و مستقل از تعداد جملات درخواست شده برای اندازه متن خلاصه است. در نهایت جملات با بالاترین مقدار S در متن خلاصه قرار خواهند گرفت.

از موارد دیگری که در این مقاله مورد بررسی قرارگرفته و بهبود یافته است نحوه ارزیابی روش خلاصه‌سازی پیشنهاد شده و مقایسه آن با سایر موارد مشابه در تحقیقات قبلی است[۱۵]. از آنجا که ارزیابی روشها یکی از مهمترین موارد هر روش پیشنهاد شده در خلاصه‌سازی متن بوده و متدهای متفاوتی در مقالات گوناگون برای این امر ایجاد گردیده است موضوع روشهای متفاوت ارزیابی در بخشی جداگانه به پرداخته، مقایسه و گزارش خواهد شد. اما ساده ترین روش ارزیابی که بیشتر مقالات به آن استناد کرده‌اند مقایسه خلاصه خروجی با متن خلاصه شده توسط یک عامل انسانی است. روشی که مقاله مورد بحث برای ارزیابی متن خلاصه شده بکار برده است به شرح زیر می‌باشد:

روش ارزیابی ارائه شده در این مقاله نوع بهبود یافته شده ارزیابی با روش ارزیابی محتوی محور (Content-based method) بوده که از معیار تشابه کسینوسی استفاده می‌کند. در مقایسه و ارزیابی کسینوسی متن خلاصه شده با متن اصلی مقایسه شده ومیزان تشابه ارزیابی می‌شود. معادله معیار تشابه کسینوسی به صورت زیر است:

$$\cos(X, Y) = \frac{\sum X_i * Y_i}{\sqrt{\sum X_i^2 * \sqrt{Y_i \wedge 2}}}$$

روش ارزیابی پیشنهاد شده در این تحقیق به جای مقایسه متن اصلی با متن خلاصه شده با توجه به معیار تشابه کسینوسی از ماتریس‌های خروجی SVD ایجاد شده برای مقایسه استفاده می‌کند. به این صورت که پس از ایجاد متن خلاصه شده دوباره الگوریتم SVD را بر روی متن خلاصه شده اعمال و ماتریس‌های مورد نظر را ایجاد کرده، حال برخی از این ماتریس‌ها را با ماتریس‌های SVD اولیه متن اصلی ورودی ایجاد شده و با استفاده از معیار تشابه کسینوسی مقایسه می‌کند. معیار مقایسه شده با توجه به معادلیه زیر مبنای ارزیابی متن خلاصه شده می‌باشد:

$$\cos\theta = \sum_{i=1}^{n} Ue_i * Uf_i$$

در این معادله Ue اولین ماتریس بردار یکه چپ ایجاد شده الگوریتم SVD متن خلاصه شده بوده و Uf اولین ماتریس بردار یکه چپ ایجاد شده متن اصلی ایجاد شده توسط SVD می‌باشد. n نیز تعداد unique termهای متن اصلی می‌باشد.

Makbule Gulcin Ozsoy& Ilyas Cicekli, 2010: در این تحقیق از روشهای ارائه شده قبلی برای اعمال LSA و SVD استفاده شده است. اما دارای نوآوری هایی در مورد استخراج جملات می‌باشد که به بررسی آن پرداخته می‌شود[۱۴].

برخی از جملات موجود در متن‌های ورودی بخصوص جملاتی که در بخشهای مقدمه و نتیجه گیری متون قرار دارند به بیشتر از یک موضوع وخصوصا به موضوعاتی غیر از موضوع اصلی متن تعلق دارند. تشخیص اینکه آیا این جملات جزو جملات مهم متن بوده و یا اینکه به دلیل نویز ایجاد شده در ماتریس‌های LSA انتخاب شده‌اند یکی از مسائل مشکل در روشهای خلاصه‌سازی متن بر پایه الگوریتم LSA می‌باشد. مقاله یاد شده روشی برای حل این مسئله با نام cross معرفی کرده است.

از موارد دارای ابهام دیگری که در روشهای خلاصه‌سازی مبتنی بر الگوریتم LSA ارائه شده قبلی وجود دارد این است که concept هایی که پس از اجرای SVD ایجاد می‌شوند ممکن است به موضوع اصلی مربوط نبوده و به موضوعی با اهمیت کمتر موجود در متن و یا یک زیر موضوع مربوط باشند.

برای تشخیص و تعیین ارتباط جمله استخراجی با موضوع اصلی و یا سایر موضوعات، تحقیق مورد بررسی روشی جدید به نام Topic ارائه کرده است تا توسط آن بتوان تا حد امکان فقط جملات مربوط به موضوع اصلی را انتخاب و در متن خلاصه قرار داد.

روش cross ارائه شده در مقاله: در این رویکرد همانند سایر روشهای ارائه شده قبل ابتدا دو مرحله اول الگوریتم LSA اجرا و ماتریس VT ایجاد می‌شود. اما قبل از انتخاب جملات مهم ابتدا میانگین امتیاز جمله در هر concept موجود در سطر محاسبه شده و هر درایه ای که دارای مقداری کمتر از این میانگین داشت برابر صفر می‌شود. ایده اصلی این عمل این است که این جملات می‌توانند به موضوع اصلی متن کمی ارتباط داشته باشند ولی حتما جمله موضوعی اصلی نخواهند بود. در این مرحله مقادیر درایه‌های این ماتریس (ماتریس VT) در مقادیر ماتریس $\sum$ ضرب شده و طول بردارهای ماتریس حاصل محاسبه می‌شوند. حال بردارهای با طول بزرگتر انتخاب و جملات مربوط به آنان در متن خلاصه شده قرار می‌گیرند.

روش Topic ارائه شده در مقاله : در این رویکرد نیز ابتدا دو مرحله اول الگوریتم LSA اجرا، میانگین درایه‌های ماتریس VT محاسبه و درایه‌های کمتر از این مقدار میانگین برابر صفر قرار می‌گیرند. با انجام این عمل جملات با اهمیت کمتر حذف شده، سپس ماتریس جدیدی به نام ماتریس concept * concept با جمع مقادیر درایه‌های نظیر ایجاد و سطربا بالاترین مجموع مقدار موضوع اصلی متن می‌باشد.

۳-۱-۱- خلاصه‌سازی متن با استفاده از زنجیره‌های لغوی

پیوستگی لغوی در ساختار یک متن می‌تواند اهمیت بخشهای مهم آن را مشخص کند. لغات هم معنی را می‌توان به صورتی با هم مرتبط کرد و زنجیره ای از لغات را بوجود آورد. زنجیره‌های لغوی موجود در یک متن می‌توانند ابزار مفیدی برای تحلیل پیوستگی بخشهای متفاوت آن متن بوده و از این رو در بسیاری از روشهای خلاصه‌سازی ارائه شده از تکنیک استخراج زنجیره‌های لغوی استفاده

شده است. یکی از مهمترین مزایای استفاده از هبستگی‌های لغوی سادگی شناسایی ارتباطات و محاسبات بین آنان می‌باشد. اولین مدل محاسباتی زنجیره‌های لغوی توسط Moris و Hirst در سال ۱۹۹۱ ارائه شد.

استفاده از زنجیره‌های لغوی در خلاصه‌سازی متن ابتدا توسط Barzilay و Elhadad معرفی شد[۳۲][۱۹]. این روش از دیتابیس word net برای تعیین ارتباط و پیوستگی بین ترم‌های متن استفاده کرده و سپس یک زنجیره بین این ترمها ایجاد می‌کند. امتیازهای در نظر گرفته شده برای این ترم‌ها با توجه به تعداد و نوع ارتباط موجود در این زنجیره‌ها تعیین شده و خلاصه نهایی شامل جملاتی است که قویترین ارتباطات زنجیره ای بین آنان برقرار است. به عبارتی ساده تر زنجیره‌های لغوی به عنوان خوشه هایی از کلمات هستند که از نظر معنایی با هم مرتبط هستند. به عنوان مثال کلمات خانه، منزل، سرا، اتاق یک زنجیره لغوی را تشکیل داده که در آنها خانه، سرا، منزل هم معنی بوده و اتاق بخشی از یک خانه می‌باشد که با آن مرتبط است. در حالت کلی اغلب سیستم‌های خلاصه‌سازی متن با رویکرد زنجیره‌های لغوی از روش یکسانی در مرحله پردازش متن بکار می‌برند که شامل مراحل زیر است:

۱- تولید زنجیره‌های لغوی
۲- امتیاز دهی به زنجیره ها
۳- یافتن قویترین زنجیره برای امتیازدهی و استخراج

در استفاده از زنجیره‌های لغوی دو مسئله را باید در نظر گرفته و راهکارهایی جهت مدیریت آن ایجاد کرد. این دو مسئله عبارتند از:

۱. ابهام در زنجیره لغوی: ارتباط معنایی بین کلمات بستگی به معنای هر کلمه داشته و اگر برخی از این کلمات دارای ابهام معنایی باشند زنجیره ایجاد شده نیز دارای ابهام خواهد بود.

۲. مربوط بودن یا نبودن زنجیره ایجاد شده با موضوع متن: تمام زنجیره‌های ایجاد شده به موضوع اصلی متن مرتبط نمی باشند. از این روشی برای تشخیص و تفکیک زنجیره‌های مرتبط و غیر مرتبط باید ابداع کرد.

روش ارائه شده توسط R. Barzilay & M. Elhadad: همانطور که در قبل اشاره شد استفاده از زنجیره‌های لغوی در خلاصه‌سازی متن ابتدا در مقاله مورد بررسی مورد معرفی شد[۱۹]. این روش از دیتابیس word net برای تعیین ارتباط و پیوستگی بین ترم‌های متن استفاده کرده و سپس یک زنجیره بین این ترمها ایجاد می‌کند. امتیازهای در نظر گرفته شده برای این ترم‌ها با توجه به تعداد و نوع ارتباط موجود در این زنجیره‌ها تعیین شده و خلاصه نهایی شامل جملاتی است که قویترین ارتباطات زنجیره ای بین آنان برقرار است.

در روش ارائه شده برای ایجاد زنجیره لغوی از همان مدل محاسباتی زنجیره‌های لغوی Moris و Hirst استفاده شده و برای تعیین امتیاز هر زنجیره از دو پارامتر Length (طول) و Homogeneity index استفاده می‌کند. پارامتر طول تعداد اعضای زنجیره بوده و Homogeneity index تعداد اعضای متمایز از هم موجود در زنجیره تقسیم بر مقدار طول می‌باشد.میزان امتیاز هر زنجیره با توجه به پاراکترهای فوق به صورت زیر محاسبه می‌شود:

Score (Chain) = Length * Homogeneity index

حال با توجه به محاسبهمیزان امتیاز زنجیره ها، زنجیره‌های قوی نیز به صورت زیر محاسبه می‌شوند:

Strength Chain = Score (Chain) > Average (Scores) + 2 * Standard Deviation (Scores)

انتخاب جملات با اهمیت: پس از اینکه زنجیره‌های لغوی متن مشخص شدند مرحله استخراج جملات مهم آغاز می‌شود. انتخاب جملات مهم در این تحقیق به سه طریق انجام می‌شود.

۱- به ازای هر زنجیره لغوی قوی انتخاب شده، اولین جمله موجود در متن اصلی که شامل کلماتی از زنجیره مربوطه باشد انتخاب می‌شود.

۲- به ازای هر زنجیره لغوی انتخاب شده، اولین جمله موجود در متن اصلی که شامل اولین کلمه موجود در زنجیره باشد انتخاب می‌شود. خلاصه خروجی ایجاد شده در این روش تقریبا شبیه به روش قبلی می‌باشد.

۳- از آنجایی که موضوع اصلی یک متن ممکن است شامل جملاتی باشد که در سرتاسر متن باشند و همچنین زنجیره لغوی مربوط به موضوع اصلی متن چگال ترین زنجیره بوده و جملات شامل کلمات آن در سرتاسر متن پراکنده هستند. در این روش جملاتی که بیشترین ارتباط با این زنجیره را دارند در خروجی قرار خواهند گرفت. در صورت نیاز به جملات بیشتر در خروجی می‌توان از زنجیره‌های با درجه امتیاز پایین تر نیز استفاده کرد.

در [۲۰]روشِی که از نوع استخراجی بوده و ابتدا مفهوم صحیح هر کلمه را یافته و سپس بر اساس آن یک زنجیره لغوی را ایجاد کرده و بر اساس معیارهایی به آنها امتیاز می‌دهد. در این هنگام زنجیره هایی که دارای امتیاز کمتری نسبت به سایرین بوده حذف کرده و موضوع اصلی متن را با توجه به زنجیره‌های باقیمانده کشف می‌کند. سپس با استفاده از روشهایی مشابه سایر رویکردهای پیشنهاد شده تقسیم بندی و استخراج جکلات مهم را انجام می‌دهد. در این مرحله به برخی از موارد بهبود یافته شده در این مقاله اشاره می‌شود.

از موارد مهمی که در ایجاد زنجیره لغات ممکن است موجب ایجاد ابهام شود کلماتی که دارای یک سبک نوشتاری بوده ولی دارای معانی مختلف هستند. به عنوان مثال می‌توان کلمه شیر در فارسی و یا bank در انگلیسی را نام برد. اینگونه موارد زنجیره ایجاد شده دارای ابهام بوده و یا ممکن است یک کلمه در دو زنجیره متفاوت قرار گرفته و یا موجب اتصال دو زنجیره متفاوت به هم بشود. تحقیق مورد بررسی در این گونه موارد سعی در بررسی لغات به صورت محلی کرده و معتقد است که ایجاد زنجیره با کلمات به صورت محلی با احتمال بسیار بالا بدون ابهام خواهد بود.

مراحل الگوریتم بکار گرفته شده در این تحقیق به صورت زیر است:

۱- ورود متن

۲- استخراج جملات

۳- پیش پردازش متن و حذف stop word ها

۴- ریشه یابی و POS

۵- ایجاد گراف WordNet

۶- تولید زنجیره‌های لغوی

۷- حذف زنجیره‌های لغوی ضعیف با ترکیب آنها با هم

۸- خوشه بندی زنجیره‌های لغوی

۹- قطعه بندی متن

۱۰-استخراج جملات مهم

مراحل دریافت متن، استخراج جملات، پیش پردازش متن، حذف stop word ها، ریشه یابی لغات و عملیات شناسایی جایگاه کلمه در جمله مانند روشهای انجام شده در سایر رویکردها می‌باشد. اما مراحل بعدی از ایده‌های این تحقیق می‌باشد.

ایجاد گراف WordNet: فرض شود که Wi یک لغت درون متن بوده و Wi دارای n مفهوم به صورت {Wi1, Wi2, …, Wik, …, Wkn} می‌باشد. در این روش هر لغت تمام مفاهیم متفاوت آن در نظر گرفته می‌شود. سپس کلمات مرتبط با هر معنی آن به صورت درختی به آن متصل شده و در سطح دوم مفاهیم متفاوت هر لغت قرار گرفته و به همیت ترتیب سطح‌های بعدی نیز ایجاد می‌شود. در برخی از حالات این سطوح با هم مرتبط شده و تشکیل گراف می‌دهند.

خوشه بندی زنجیره‌های لغوی: پس از ایجاد زنجیره لغوی برای متن مورد نظر برخی از زنجیره‌های ایجاد شده ضعیف بوده که باید حذف شوند. عمل حذف در این مرحله حذف مجازی بوده و زنجیره‌های ضعیف با هم ادغام می‌شوند. برای هر زنجیره لغوی LCi یک بردار وقوع جمله به صورت Vi = {S1i, S2i, …, Ski, …, Sni} ایجاد می‌شود. در این بردار n تعداد جملات موجود در متن بوده و هر Ski تعداد اعضای LCi در جمله k ام است. به عنوان مثال اگر جمله k دارای سه عضو از

LCi باشد آنگاه Ski برابر ۳ خواهد بود. دو زنجیره لغوی LCi و LCj درون یک خوشه قرار خواهند گرفت اگر بردار Vi و Vj شبیه هم باشند. الگوریتم خوشه بندی با n خوشه شروع به کار کرده که در اینجا n برابر با تعداد زنجیره‌های لغوی ایجاد شده می‌باشد. الگوریتم به صورت تکراری اجرا شده و در هر بار خوشه‌های شبیه به هم پیدا شده و با هم ادغام می‌شوند. این ادغام خوشه‌ها تا زماتی ادامه می‌یابد تامیزان شباهت بین دو خوشه از یک حد آستانه کمتر شود.میزان شباهت دو خوشه با استفاده از معیار کسینوسی و به صورت زیر محاسبه می‌شود. در این فرمول |Vi| طول اقلیدسی بردار خواهد بود.

$$\text{Cos}(\theta) = \frac{V\square}{|Vi| * |Vj|}$$

استخراج دنباله جمله ای: در این تحقیق قطعه بندی متن با توجه به هر خوشه زنجیره لغوی و با یافتن نقاط اصلی هر موضوع انجام می‌شود. برای هر خوشه یک دنباله از جملات متصل به هم به عنوان یک قطعه استخراج می‌شود. چون جملاتی که به صورت منسجم به هم متصل شده‌اند معمولا مربوط به یک موضوع می‌باشند برای هر خوشه زنجیره لغوی CLj یک دنباله جمله ای جداگانه ایجاد می‌شود. برای هر جمله Sk، اگر جمله Sk دارای یک عضو زنجیره لغوی در CLi باشد یک دنباله جدید ایجاد شده و یا جمله به دنباله موجود افزوده می‌شود. با این رویه متن با توجه به خوشه‌ها قطعه بندی می‌شود. برای هر دنباله جمله ای امتیازی در نظر گرفته می‌شود که این امتیاز با معادله زیر محاسبه می‌شود:

$$\text{Score (Sequence i)} = \text{Score (CLi)} * \text{Li} * \frac{(1+SLC\square)*PLCi}{f^2}$$

در این معادله Li تعداد جملات موجود در دنباله جمله ای i ، SLCi تعداد زنجیره‌های لغوی که در دنباله جمله ای i استارت زده شده اند، PLCi نیز تعداد زنجیره‌های لغوی هستند که دارای عضوی در دنباله جمله ای i هستند و f نیز تعداد زنجیره‌های لغوی موجود در خوشه هستند. امتیاز هر خوشه

عبارت است از امتیاز میانگین هر زنجیره لغوی درون آن خوشه و معادله امتیاز دهی تعریف شده سعی در مدل کردن ارتباط هر قطعه با استفاده از امتیاز خوشه ای است.

ارزیابی سیستم پیشنهادی: ارزیابی سیستم یاد شده با معیارهای Precision (p) , Recall (R) , F1-measure سنجیده می‌شود. معادله‌های زیر این ارزیاب‌ها را مشخص می‌کنند:

$$P = (|sum_ref \cap sum_cand|) / |sum_cand|$$
$$R = (|sum_ref \cap sum_cand|) / |sum_ref|$$
$$F1 = 2PR / (P + R)$$

در این معادلات sum_ref مشخص کننده خلاصه استخراجی توسط فرد خبره و sum_cand متن خلاصه شده کاندید توسط سیستم می‌باشند. معیار ارزیابی دیگر استفاده شده معیار ROUGE-N بوده که با معادله زیر محاسبه می‌شود. این معیار تعداد N- gramهای متن خلاصه شده توسط مشین و خبره را با هم مقایسه می‌کند.

$$\text{ROUGE-N} = (\sum S \in sum_{ref} * \sum N-gram \in s * \text{count_match (N-gram)}) / (\sum S \in sum_{ref} * \sum N-gram \in s * \text{count (N-gram)})$$

در این معادله N تعداد N-gram ها، count_match (N-gram) ماکزیمم تعداد N-gram که به صورت همزمان در متن خلاصه شده کاندید و مجموعه خلاصه‌های انجام شده توسط خبره وجود دارند و count (N-gram) نیز تعداد N-gramهای موجود در خلاصه مرجع انجام شده توسط فرد خبره می‌باشد.

در [21] روشی جدید برای ایجاد زنجیره‌های لغوی با استفاده از پایگاه دانش Lexico-Semantic (LS) پیشنهاد گردیده است. در این روش کلمات موجود در متن درون خوشه هایی معنایی قرار گرفته تا بتوان توسط این خوشه‌ها الگوهایی معنایی را از متن استخراج کرد. پایگاه دانش LS یک منبع ایجاد شده بوده که در آن کلمات با معنی مشابه به صورت گروه در آمده اند. خوشه‌های ایجاد

شده به صورت سلسله مراتبی در یک درخت باینری قرار گرفته و گروه بندی شده‌اند که در آن یک لغت می‌تواند به خوشه‌های متفاوتی تعلق داشده باشد.

در روش ارائه شده برای تعیین ارتباط یک لغت به یک زنجیره ازیک معیار شباهت معنایی استفاده شده است که ترکیبی از روشهای پیشنهادی Hist & St-onge, 1997 و Barzilay & Elhadad, 1997 می‌باشد.

در این مرحله نوآوری‌های پایان نامه مورد بررسی قرار می‌گیرد. مراحل ایجاد زنجیره‌های لغوی در این تحقیق به صورت زیرمی باشند:

اولین قدم در ساخت زنجیره‌های لغوی با اولین کلمه موجود در متن شروع می‌شود. این کلمه درون یک زنجیره لغوی اولیه قرار داده می‌شود. برای درج کلمه بعدی و کلمه‌های بعدی ارتباط آن با کلمات موجود در زنجیره‌های لغوی موجود بررسی شده و در صورت وجود ارتباط با تمامی لغات موجود در آن زنجیره لغت مورد نظر در آن زنجیره درج شده و در غیر این صورت زنجیره لغوی جدیدی ایجاد می‌شود.

برای تعیین ارتباط و سازگاری معنایی یک لغت به یک زنجیره از روشهایی مانند اصلاح نامه، آنتولوژی و دیکشنری‌ها استفاده می‌شود. اغلب روشها از WordNet برای تعیین ارتباط معنایی استفاده می‌کنند.

در الگوریتم پیشنهادی در این تحقیق به جای استفاده از منبع استاندارد زبانی Wordnet از ارتباط و پیوستگی لغوی بین کلمات موجود در متن و پایگاه دانش LS ایجاد شده استفاده می‌شود. به جهت انتخاب لغات کاندید برای پروسه زنجیره سازی لغات برچسب گذاری می‌شوند. روش برچسب گذاری لغات توسط الگوریتم TNT_tagger ارائه شده توسط Brants, 2000 انجام می‌شود [۲۲]. برای استخراج اسامی ترکیبی از الگوریتم SENTA ارائه شده توسط Dias, 2002 استفاده شده و در نهایت

اسامی، اسامی ترکیبی و اسامی خاص فیلتر شده و بـه عنـوان لغـات کاندیـد بـرای پردازشـهای بعـدی زنجیره‌های لغوی به کار می‌روند.

هر لغت کاندید ممکن است دارای چندید معنی بوده که این معانی با توجه بـه متنـی کـه لغـت در آن قرار دارد تعیین می‌شود. ابهام زدایی، تعیین معنی مرتبط با متن و سایر لغات شـبیه توسـط الگـوریتم خوشه بندی متن PoBOC ارائه شده توسط Cleuziou, 2004 انجام شده است [۲۳]. مقالات مربـوط به روشهای ارائه شده دریافت شده ولی بررسی کامل نگردیده و صرفا فقط بـه عنـوان مرجـع الگـوریتم ذکر می‌گردند. لازم به ذکر است که روش پیشنهادی در این پایان نامه در مورد زبانهـایی ماننـد فارسـی که دارای یک پایگاه داده لغوی قوی مانند WordBet نیستند بسیار مفید و کارا می‌باشد. بـا توجـه بـه روش پیشنهادی لغاتی که در یک گروه قرار می‌گیرند دارای مفهوم و کاربرد یکسانی می‌باشند.

قبل از شروع به اجرای روشهای ذکر شده، متد پیشنهادی همانند سایر روشها متن مـورد نظـر را پـیش پردازش می‌کند. مراحل پیش پردازش بکار گرفته شده به شرح زیر می‌باشند:

برچسب زنی (Tagging) : در این مرحله به هر لغت برچسب سمت کلمه در متن داده می‌شود. ایـن برچسب زنی توسط الگوریتم TNT (Trigrams N Tags) که یـک روش آمـاری مـوثر POS مـی‌باشـد انجام می‌شود.

استخراج عبارات متنی چند قسمتی (N-ary Textual words): در این مرحله کلماتی کـه دارای چند قسمت بوده تشخیص و استخراج می‌شوند. N-gramها نمونه هایی از این نوع می‌باشند. بـا توجـه به اشاره قبل الگوریتم SENTA برای تشخیص و استخراج این کلمات بکار می‌رونـد. چنـد روش دیگـر مانند نیز Mutual Expectation و الگوریتم GenLocalMaxs برای انجام این کار در پایان نامـه ذکـر شده است که در این مرور کوتاه مجال پرداختن به آن نمی باشد.

بازیابی و استخراج اسم ها: استخراج اسم‌ها توسط الگوریتم‌های برچسب زنی کلمات انجام می‌شود. اما در این مطالعه چون اسامی چند بخشی نیز مورد توجه قرار گرفته و جهت قرار گرفتن در زنجیره‌های لغوی باید ابهام زدایی شوند از الگوریتم SENTA برای تشخیص آنان استفاده می‌شود. از جمله این لغات در فارسی سیب زمینی و در انگلیسی New York را می‌توان نام برد.

بردار word-context: این بردار یک بردار چند بعدی بوده که شامل مقادیر عددی اجزای متن می‌باشد. مقادیر موجود در این بردار مشخص کننده نزدیکی و ارتباط کلمه یا عبارت به موضوع اصلی زنجیره می‌باشد.

ماتریس شباهت: در این مرحله شباهت لغت محاسبه می‌شود. بنابراین برای هر لغت موجود در متنمیزان شباهت آن با تمام کلمات متن محاسبه می‌شود.

معیار شباهت در پایگاه دانش: منظور از شباهت معنایی ارتباط مفهومی بین دو کلمه است. به عنوان مثال بین دو کلمه ماشین و اتومبیل ارتباط مفهومی بسیار نزدیکی وجود دارد. بین ماشین و بنزین نیز ارتباط مفهومی وجود دارد. می‌توان بین اتومبیل و دوچرخه نیز یک ارتباط مفهومی ایجاد کرد. اما در این مثالها درجه شباهت از اولی به سومی کم می‌شود. در این تحقیق از یک روش آماری برای تعیینمیزان شباهت بین دو مفهوم استفاده کرده است که در معادله زیر به صورت خلاصه نشان داده می‌شود:

$$\mathrm{Sim}\,(x1, x2) = \frac{2*\log P(C0)}{\log p(C1)+\log P(C2)}$$

در این معادله x1 و x2 دو کلمه موجود در پایگاه دانش بوده، C1, C2 دو خوشه بدون ارتباط به یکدیگر و C0 خوشه ای که با هر دو خوشه قبلی ارتباط دارد.میزان شباهت x1 و x2 با معادله فوق محاسبه می‌شود.

امتیاز دهی به زنجیره‌های لغوی ایجاد شـــده: بـرای امتیـازدهی بـه زنجیـره‌هـای لغـوی و انتخـاب زنجیره با امتیاز بالا جهت استخراج جملات مربوط به آن و قرار دادن در خلاصه خروجی تحقیق مـورد بررسی از فرمول زیر استفاده می‌کند:

$$\text{Score (chain)} = \frac{\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\text{sim}(Wi,Wj)}{\frac{((n-1)*n)}{2}}$$

در این معادله n تعداد زنجیره‌های لغوی ایجاد شده می‌باشد.

۳-۴-خلاصه‌سازی متن و تئوری گراف

در این مرحلـه بـه معرفـی چنـد روش بکـار گرفتـه شـده در تئـوری گرافهـا در خلاصـه‌سـازی مـتن می‌پردازیم. یکی از این روشها الگوریتم Text Rank است. روش Text Rank روشی است کـه بـر پایـه تئوری گراف به متن امتیاز دهی می‌کند[۴]. این روش از یک متد غیر نظارتی بـرای اسـتخراج کلمـات کلیدی و جمله ارائه کرده است. امتیاز دهی به رئوس گراف در این روش براساس یـک سـری اطلاعـات سراسری و به صورت بازگشتی و بر روی تمام جملات انجام می‌شود. این روش یک گـراف ایجـاد کـرده که در آن جملات رئوس گراف بوده و این رئوس توسط ارتباط معناداری به هم مرتبط می‌شـوند. بـرای مرتبط کردن این رئوس و ایجاد یالهای گراف از یک سری معیارهای شـباهت اسـتفاده مـی‌شـود. ایـن ارتباط بین دو جمله را به نوعی می‌توان همان پروسه پیشـنهاد در نظـر گرفـت و جملـه‌هـای بـا مفهـوم خاصی در متن به خواننده پیشنهادی برای مراجعه به جمله دیگری در متن مـی‌دهـد کـه هـر دو جملـه مفهوم مشترکی در متن دارند و بنابراین می‌توان لینکی بین این دو جمله برقرار کرد[۳۲].

ایده اصلی در این الگوریتمها مانند رای دادن یا پیشنهاد دادن اسـت. در ایـن الگوریتمهـا وقتـی از یـک نود به نود دیگری لینک برقرار می‌شود به این معنی است که به آن نود رای می‌دهـد. بـه ایـن صـورت که هر چه درجه ورودی یک نود بیشتر شـود از اولویـت بـالاتری برخـوردار خواهـد بـود. در ایـن روش اولیت و اهمیت رای نیز وجود دارد. یعنی یـک نـود کـه بـه نودهـا بیشـتری رای داده باشـد و یالهـای

خروجی بیشتری داشته باشد نیز از اهمیت بالایی برخوردار است. درجه اهمیت یک نود در الگوریتم یاد شده از رابطه زیر محاسبه می‌شود. درجه اهمیت هر نود با معادله یاد شده به صورت بازگشتی محاسبه شده و این عمل برای تمام نودهای گراف انجام می‌شود و این عمل تا زمانی ادامه پیدا می‌کند که دیگر تغییری درمیزان امتیازات حاصل نشود.

$$S(V_a) = (1 - d) + d * \sum_{V_b \in \ln\ (V_a)} \frac{S(V_b)}{out(V_b)}$$

در این معادله در گراف G(V, E) رئوس و نودهای گراف V و E یالهای گراف بوده که تعداد یالها زیر مجموعه ای از V*V می‌باشند. به این معنی که کل نودها می‌توانند با هم ارتباط داشته باشند. مقدار in(Va) تعداد یالهای ورودی Va بوده و مقدار out(Va) تعداد یالهای خروجی از راس نامبرده می‌باشند.مقدار d نیز پارامتریذاست که بین صفر تا یک بوده و توسط ورودی مشخص می‌شود. نگارنده مدعی است که این الگوریتم را می‌توان بر روی گراف‌های غیر جهت دار نیز اعمال کرد که توضیحی در مورد چگونگی عملکرد آن نداده است. فقط به این موضوع اشاره کرده است که این عمل دارای مقدار تکرار زیادی بوده و هزینه زمانی بالایی بوده و خروجی نهایی آن نیز بسیار متفاوت تر از گراف‌های جهت دار می‌باشد. S(Va) مقدار امتیاز نود Va بوده و S(Vb) مقدار امتیاز راس Vb که با راس قبلی ارتباط داشته است می‌باشد.

الگوریتم با یک مقدار اختیاری برای هر یک از رئوس انتخاب می‌شود و محاسبات تکرار می‌شود تا به یک همگرایی پایینتر از یک حد خاصی برسد .بعد از اجرای الگوریتم امتیازی به هر یک از رئوس داده می‌شود که درجه اهمیت آن راس را بیان می‌دارد .مقدار اولیه ای که به رئوس داده می‌شود تاثیری روی امتیاز نهایی ندارد.

روش ارائه شده توسط M. Pourvali & M. Saniee Abadeh, 2012 : روش پیشنهاد شده در مقاله الگوی دیگری را برای خلاصه‌سازی و ارتباط بین جملات استفاده کرده است. برای ایجاد این ارتباط فقط از مشابهت بین ریشه اسم‌ها و فعلها استفاده کرده که موجب ایجاد چندین گراف شده که

هر گراف مربوط موضوعی خاص از متن می‌باشد[۳۱]. الگوریتم استفاده شده بـرای تشـخیص مشـابهت بین جملات با سایر روشهای موجود تفاوت دارد. موضوع اصلی متن موجب ایجاد گرافـی بزرگتـر شـده اما سایر موضوعات موجود در متن نیز در دسترس خواهد بود.

در الگوریتم ارائه شده در این مقاله ابتدا متن وارد شده پیش پردازش می‌شده و stop-wordهـا حـذف می‌شوند. سپس ریشه کلمات استخراج شده و اجزای جمله (POS) مشخص می‌شـوند. سـپس از بـین POSهای مشخص شده متن فقط از افعال و اسامی استخراج می‌شوند. الگوریتم از ابتدای مـتن شـروع کرده و تمامی افعال و اسامی هر جمله را انتخاب و ریشه آنها را استخراج می‌کند. حـال جمـلات دیگـر را نیز بررسی و این عمل را نیز برای آنان انجام و لغات مشابه را پیدا و بـین جمـلات شـامل آنهـا یـک لینک برقرار می‌کند. به این ترتیب متن به چندین گراف تبدیل شـده کـه هـر گـراف جمـلات شـامل کلمات شبیه به هم را در خود خواهد داشت. در هنگام ایجاد ایـن گرافهـا یالهـای آنـان نیـز وزن دهـی می‌شوند. وزن دهی یالها با توجه به فاصله کلمات بـین دو جملـه و بـا اسـتفاده از الگـوریتم Average Google normalized distance انجام می‌شـود. در ایـن الگـوریتم $S = \{S1, S2, \ldots, Sn\}$ مجموعـه جملات کل متن و $W = \{W11, W21, \ldots, Wij, Wkn\}$ مجموعـه کلمـاتی شـامل اسـامی و افعـال متن ورودی بوده که Wij نشان دهنده i امین کلمه در j امین جمله می‌باشد. در این روش نیـز رئـوس گراف جملات بوده و ارتباطات بین کلمات برای مقدار دهی به یالها استفاده می‌شود.

حال از بین این گرافها بزرگترین گراف شامل موضوع اصـلی مـتن بـوده و از جمـلات آن گـراف بـرای خلاصه‌سازی متن استفاده می‌شود. مانند سایر روش‌های خلاصه‌سازی متن با توجه بـه جمـلات امتیـاز دهی شده و بر اساسمیزان درصد متن خلاصه خواسته شده تعدادی از جملات در مـتن خلاصـه شـده قرار می‌گیرند. در این روش امکان ایجاد متن خلاصه شده با توجه به سایر موضوعات کم اهمیت تـر در متن و یا به صورت ترکیبی نیز وجود داشته که با انتخاب گرافهای کوچکتر و یا ترکیبی از دو یـا چنـد

گراف برای این کار استفاده می‌شود. معادله زیر برای انتخاب جملات موجود در بزرگترین گراف برای متن خروجی به کارمی رود:

$$Fi = ((Ii + Oi) / L) * \sum_{t=1}^{m} Wti$$

در این معادله Oi تعداد یالهای خروجی از i امین جمله و Ii نیز تعداد یالهای ورودی به آن می‌باشد. وزن کلمه Wti نیز به صورت زیر محاسبه می‌شود:

$$W\ ti = TF\ ti * ISF\ ti$$

که در اینجا نیز TF ti تعداد وقوع عبارت t در جمله Si بوده و ISF مربوط به هر جمله نیز به صورت زیر محاسبه می‌شود:

$$ISF = \log (N / Nt)$$

و در آن N تعداد کل جملات گراف مربوطه و Nt تعداد جملاتی که کلمه ti درآن وجود داشته است.

بررسی الگوریتم HITS: HITS[۲] یک الگوریتم تکرار شونده بوده که در ابتدا برای امتیاز دهی به صفحات وب و پیشنهاد شد (Kleinberg, 1999). این الگوریتم یک تمایز و برتری برای صفحاتی که تعداد لینکهای ورودی بیشتری دارند قائل می‌شود. معیار hubs نیز تعداد لینکهای خروجی یک صفحه وب است. می‌توان از این الگوریتم برای امتیاز دهی به جملات یک متن که در رئوس یک گراف قرار گرفته‌اند نیز استفاده کرد. به این صورت که تعداد لینکهای ورودی (HITS) و تعداد لینک‌های خروجی (hubs) آنرا محاسبه و از آنها برای امتیاز دهی به آن جمله استفاده کرد[۴].

$$HITS_a\ (Vi) = \sum_{Vj\ \epsilon\ in\ (Vi)} HITS\ h\ (Vj)$$

$$HITS\ h\ (Vi) = \sum_{Vj\ \epsilon\ out\ (Vi)} HITS\ a\ (Vj)$$

---

[۲] - Hyperlinked Induced Topic Search

۳-۵-خلاصه‌سازی متن با رویکرد خوشه بندی

رویکرد مهم دیگری که برای خلاصه‌سازی متون استفاده می‌شود روشهای مبتنی بر خوشه بندی متن می‌باشد. در این رویکرد با استفاده از الگوریتم‌های مختلف خوشه بندی، متن مورد نظر به توکن هایی تقسیم می‌شود. این توکن‌ها می‌توانند هر یک از اجزای جمله مانند کلمات، عبارات، جملات و حتی پاراگرافها باشند. بیشتر تحقیقات انجام شده در این رویکرد نیز از جملات به عنوان توکن‌های متن استفاده کرده اند. این نوع خلاصه‌سازی نیز از نوع استخراجی بوده و در آن سعی می‌شود که جملات با بیشترین شباهت درون خوشه هایی قرار گرفته و سپس با توجه بهمیزان اهمیت هر خوشه جملات مهم انتخاب و در متن خروجی قرار بگیرند. در این روشها نیز در ابتدا جملات با توجه به معیارهای معرفی شده در سایر رویکردها امتیاز دهی شده و درون خوشه هایی قراربگیرند. معیارهای شباهت جهت قرار دادن جملات درون خوشه‌های متفاوت معیارهای متفاوتی مانند فاصله اقلیدسی، معیار شباهت کارتزین، معیار شباهت کسینوسی وغیره باشد. در ادامه به معرفی و بررسی چند مقاله در خصوص این رویکرد پرداخته می‌شود[۳۲].

روش ارائه شده توسط A. Agrawal & U. Gupta, 2014: رویکرد بکار گرفته شده در این تحقیق یک تکنیک یادگیری بدون ناظر مبتنی بر خوشه بندی بوده و شامل مراحل زیر می‌باشد[۱۶]:

- تعیین مرز و جداسازی جملات متن
- تعیین و محاسبه امتیازهای هر جمله
- اعمال خوشه بندی مبتنی بر مرکز و استخراج جملات مهم متن

در این روش نیز مانند سایر روشهای خلاصه‌سازی به جملات موجود بر اساس معیارهایی امتیاز داده که معیارهای در نظر گرفته شده در این تحقیق Tf-IDF و طول جمله می‌باشند. روش یاد شده از الگوریتم خوشه بندی k-means استفاده کرده و از دو معیار معرفی شده برای تشخیص تعلق جمله مورد نظر به نزدیکترین خوشه استفاده می‌کند.

الگوریتم خوشه بندی k-means بکار گرفته شده به صورت زیر عمل می‌کند:

- ابتدا تعداد خوشه‌ها را مشخص کرده و سپس به تعداد خوشه‌های مشخص شده و به صورت رندم از جملات موجود در متن انتخاب و هر کدام به عنوان یک مرکز خوشه در نظر گرفته می‌شوند. سپس سایر جملات موجود با توجه به معیارهای ذکر شده به نزدیکترین مرکز ملحق شده و تشکیل خوشه ای اولیه را می‌دهند.
- در این مرحله مجددا مرکز هر خوشه محاسبه شده و دوباره جملات موجود در متن به نزدیکترین خوشه نسبت داده می‌شوند. چون با اضافه کردن هر نمونه به خوشه مرکز آن خوشه تغییر می‌کند دوباره مرکز خوشه را محاسبه می‌شود.
- اعمال دو مرحله قبل آنقدر تکرار شده تا نمونه‌های درون هر خوشه تغییری نکند.
- معادله زیر نحوه قرار دادن جملات متن را درون خوشه‌های مربوطه نشان می‌دهد:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} |x_i^{(j)} - c_j|^2$$

در این معادله k تعداد خوشه ها، n تعداد جملات موجود در متن، x جمله موجود در متن، c مرکز انتخاب شده و $x_i - c_j$ فاصله جمله تا مرکز مربوطه می‌باشد.

در این مقاله دو ایده مطرح شده که اولین آن عبارت است از نحوه محاسبه امتیاز هر جمله جهت مقایسه با جمله انتخاب شده برای مرکز هرحوشه و ایده دوم روشی برای بدست آوردن تعداد خوشه‌های بهینه می‌باشد. محاسبه امتیاز هر جمله موجود در متن با توجه به دو معیار معرفی شده و با استفاده از **معیار فاصله کارتزین** به صورت زیر می‌باشد:

$$\text{Score}(x) = \sum_t tf - idf_t \ / \ |x|$$

در این معادله مقدار بدست آمده توسط معیار tf-idf بر طول جمله تقسیم می‌شود. پس از خوشه بندی متن، بزرگترین خوشه به عنوان موضوع اصلی متن در نظر گرفته شده و جملات قرار گرفته شده در این خوشه به عنوان متن خلاصه شده پذیرفته می‌شوند. در صورت نیاز به متن با درصد

خلاصه‌سازی کمتر از تعداد خوشه‌های بیشتری و به ترتیب سایز خوشه‌ها جملات موجود را به متن خروجی برد.

یکی از مهمترین مسائل در خوشه بندی k-means تعیین تعداد خوشه‌های بهینه بوده و با توجه به اینکه در این روش اغلب بزرگترین خوشه به عنوان متن خروجی در نظر گرفته می‌شود، اندازه متن ورودی تاثیر مستقیمی در تعیین تعداد خوشه‌ها دارد. به عنوان مثال اگر برای یک متن مقدار k بزرگی در نظر گرفته شود خوشه‌های ایجاد شده بسیار متفرق بوده و متن خلاصه شده بسیار کوچک و دارای وابستگی بسیار کم بین جملات خواهد بود. برعکس اگر مقدار k کوچک در نظر گرفته شود خوشه‌های ایجاد شده بزرگ و چگال خواهند شد. نگارنده به صورت سعی و خطا با شرط زیر تعداد بهینه خوشه‌های پیشنهادی را بدست آورده است. در این شرط k تعداد بهینه خوشه و N تعداد جملات متن ورودی است.

```
if N <= 20
k = N – 4;
else
k = N – 20;
```

روش پیشنهادی برای ارزیابی کار خود، متن خلاصه ایجاد شده توسط سیستم را با متن خلاصه شده توسط فرد خبره مقایسه می‌کند.

بررسی مقاله Anjali R. Deshpande & Lobo L. M. R. J, 2013: روش ارائه شده در این تحقیق یک رویکرد خلاصه‌سازی مبتنی بر پرسش کاربر (کوئری بیس) و استخراجی چند سندی با استفاده از الگوریتم‌های خوشه بندی می‌باشد[۱۷]. روش خوشه بندی اعمال شده در این تحقیق دو سطحی بوده، به این صورت که ابتدا متنهای مورد نظر خوشه بندی شده و سپس در مرحله دوم هر یک از متون مجددا به خوشه هایی متشکل از جملات شبیه به هم می‌شوند. معیار شباهتی که برای

تعیین شباهت اعضای موجود و قرار دادن آنها در خوشه استفاده می‌شود معیار شباهت کسینوسی و با معادله زیر می‌باشد:

$$\text{Similarity} = \cos(\Theta) = \frac{A.B}{|A|*|B|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 * \sqrt{\sum_{i=1}^{n} B_i^2}}}$$

الگوریتم استفاده شده در این مقاله به صورت زیر است:

- ابتدا مجموعه متن‌های مورد نظر به همراه پرسش کاربر از ورودی دریافت می‌شود.
- عملیات پیش پردازشی متن بر روی کوئری مورد نظر اعمال و سپس به توکن هایی تقسیم می‌شود.
- کلیه جملات موجود در تمام متن‌های مورد نظر امتیاز دهی می‌شوند. معیارهای امتیاز دهی در این تحقیق عبارتند از: اسامی خاص، عبارات نشانه ای، طول جمله، داده‌های عددی، موقیت مکانی جمله در متن، شباهت جمله با سایر جملات متن، کلمات با حروف بزرگ،میزان شباهت جمله با کوئری کاربر و معیار tf-idf .
- ابتدا متن‌های مورد نظر خوشه بندی می‌شوند.
- سپس جملات هر متن به صورت جداگانه و بر اساسمیزان شباهت درون متنی خوشه بندی می‌شوند.
- خوشه‌های ایجاد شده در هر متن و بر اساس اندازه خوشه امتیاز دهی و به صورت نزولی مرتب می‌شوند.
- با توجه بهمیزان خلاصه‌سازی خواسته شده و تعداد جملات درخواستی در متن خروجی از درون هر متن و از با امتیازترین خوشه تعدادی جمله انتخاب و در متن خلاصه شده قرار می‌گیرند.

در تحقیق ارائه شده هیچگونه توضیحی در مورد نحوه خوشه بندی متون نیامده است. اما آنچه که به نظر می‌آید این است که برای خوشه بندی متون موجود در مجموعه وارد شده هر متن را به صورت

یک جمله بزرگ در نظر گرفته و کلیه مراحلی را که برای خوشه بندی جملات موجود در هر متن بکار گرفته شده است بر روی متون تک جمله ای در نظر گرفته شده اعمال کرده است.

همچنین روش استفاده شده و نوع الگوریتم خوشه بندی نیز مشخص نگردیده است. اما می‌توان پس از انجام مراحل پیش پردازش انجام شده و با توجه به معیار شباهت پیشنهادی (کسینوسی) از هر الگوریتم خوشه بندی برای قرار دادن جملات و متن‌ها در دسته‌های مجزا استفاده کرد.

در [۱۸] از یک تکنیک ساختاری برای ایجاد یک سیستم خلاصه‌سازی کوئری بیس با استفاده از روشهای خوشه بندی استفاده شده است. الگوریتم خوشه بندی بکار گرفته شده Expectation Maximization Clustering (EM) بوده و روش پیاده سازی شده به دو بخش کلی تقسیم شده است. در بخش اول الگوریتم EM اجرا شده و در بخش دوم خلاصه‌سازی کوئری بیس انجام می‌شود. در روش ارائه شده از Word net نیز استفاده شده است. در متن ورودی پس از اعمال عملیات پیش پردازشی، جملات جداسازی شده و هر جمله به عنوان یک نود گراف در نظر گرفته می‌شود. معیار امتیاز دهی هر جمله مقایسه این نودها با هم و با کمک word net انجام می‌شود. پس از پیش پردازش‌های انجام شده عمل خوشه بندی EM انجام می‌شود.

عملیات پیش پردازش مورد استفاده در روش یاد شده عبارتند از:

- جداسازی جملات
- حذف Stop word ها
- Tokenization
- برچسب گذاری اجزای جمله (POS)
- مشخص کردن سایر عبارات گرامری مانند عبارات اسمی، عبارات اسمی و غیره
- پارس کردن جمله که عبارت است ازدرخت پارس برای هر جمله

کلیه عملیات ذکر شده با توجه به اگوریتم‌ها و روشهای ارائه شده در سایر روشهای NLP انجام می‌شود. در این مرحله ماتریس فاصله ایجاد شده و از آن در خوشه بندی جملات با روش EM استفاده می‌شود. انتخاب جملات مهم برای انتقال به متن خروجی مانند سایر روشهای خلاصه‌سازی مبتنی برخوشه بندی بوده که خوشه بزرگتر و با تراکم بیشتر به عنوان موضوع اصلی متن در نظر گرفته شده و جملات موجود در آن در متن خلاصه قرار می‌گیرند.

۳-۶-خلاصه‌سازی متن با رویکرد فازی

در اکثرروشهای ارائه شده که از منطق فازی برای خلاصه‌سازی متن استفاده کرده‌اند فقط بخش تشخیص و استخراج جملات مهم را توسط منطق فازی انجام داده اند. تفاوت روشهای متفاوت در استخراج ویژگی‌های متفاوت از جملات، ایجاد قوانین فازی متفاوت، متغیرهای زبانی ایجاد شده، توابع عضویت، روشهای fuzzification و defuzzification مختلف است. در برخی از روشها تغییرات کوچکی در مقادیر عددی بدست آمده از معیارهای امتیاز دهی به جملات مهم ایجاد گردیده که تبدیل این مقادیر به متغیرهای زبانی راحت تر و اعمال آنان به سیستم استنتاج فازی بهتر کرده است[۳۲]. در ادامه چند مقاله که عمل خلاصه‌سازی متن را با روشی فازی انجام داده‌اند بررسی می‌شوند:

روش ارائه شده در [۲۴] دارای دو مرحله است. در مرحله اول پیش پردازش متن و مرحله دوم تحلیل فازی انجام شده بر روی متن بوده که در مرحله اول پس از عملیات پیش پردازش انجام شده همانند سایر رویکردها، ویژگی‌های زیر را از هر جمله استخرج می‌کند:

- تعداد کلمات مشترک بین جمله مربوطه و عنوان متن
- تعداد کلمات جمله (جملات بسیار کوتاه از اهمیت بسیار پایینی برخوردار هستند).
- تعداد کلمات موضوعی مشابه و هم معنی با کلمات کلیدی

- میزان تشابه با جملات اول و آخر پاراگراف مربوطه (همیشه جملات اول و آخر پاراگراف مهم در نظر گرفته می‌شوند).

مقادیر بدست آمده از معیارهای فوق به اعدادی بین صفر و یک نرمال شده و به عنوان پارامترهای ورودی به سیستم استنتاج فازی داده می‌شوند. مقادیر نامبرده به متغیرهای زبانی و به چند مجموعه توابع عضویت مثلثی تبدیل می‌شوند. مقاله مذکور دلیل استفاده از توابع عضویت مثلثی را سادگی عملیات ریاضی مورد نیاز برای محاسبه خروجی عددی ذکر کرده است.

مراحل عملیات انجام شده به ترتیب زیر می‌باشد:

۱. ایجاد متغیرهای زبانی برای تولید مجموعه قوانین فازی
۲. فازی کردن ورودی‌ها با اعمال توابع عضویت مثلثی ورودی
۳. ترکیب ورودی‌های فازی شده و ایجاد قوانین فازی
۴. تعیین نتایج خروجی قوانین و ترکیب آنان با قوانین فازی ایجاد شده برای ایجاد تابع عضویت خروجی
۵. دفازی کردن خروجی‌ها برای ایجاد خروجی عددی با استفاده از روش دفازی COA

در [۲۵] عملیات انجام شده جهت خلاصه‌سازی به مراحل زیر تقسیم شده است:

۱. پیش پردازش متن ورودی
۲. استخراج ویژگی‌های لازم جهت امتیاز دهی به جملات مهم
۳. امتیاز دهی به جملات توسط منطق فازی
۴. انتخاب جملات مهمتر و ایجاد متن خلاصه

عملیات پیش پردازش انجام شده بر روی متن ورودی در تحقیق مورد بحث همانند سایر روشهای پیشین و شامل مراحل زیر می‌باشد:

۱. جدا سازی جملات

۲. حذف Stop word ها

۳. جدا سازی کلمات (Tokenization)

۴. ریشه یابی کلمات

در مرحله استخراج ویژگی‌های جملات، ویژگی هایی از جملات استخراج گردیده است همانند ویژگی‌های سایر روشهای خلاصه‌سازی می‌باشند. در این مرحله سعی گردیده است که مقادیر عددی بدست آمده در محدوده صفر و یک باشند. این ویژگی‌ها عبارتند از:

کلمات مشترک با عنوان مقاله: این ویژگی به صورت زیر محاسبه می‌شود:

F1 = تعداد کلمات مشترک با عنوان جمله / تعداد کلمات کل جمله

طول جمله: این ویژگی به صورت زیر محاسبه می‌شود:

F2 = تعداد کلمات جمله / تعداد کلمات بلندترین جمله

موقعیت جمله در پاراگراف: با توجه به اینکه جملات اول و آخر هر پاراگراف از اهمیت بیشتری نسبت به سایر جملات برخوردار است نحوه محاسبه این ویژگی به صورت زیر است:

$F3\ (S1) = n / n$

$F3\ (s2) = (n - 2) / n$

$F3\ (s3) = (n - 3) / n$

…..

$F3\ (sn) = (n - 1) / n$

در این معادلات Si شماره i امین جمله در پاراگراف و n تعداد جملات پاراگراف است.

تعداد داده‌های عددی در جمله: امتیاز جملاتی که دارای عدد هستند به صورت زیر محاسبه می‌شود:

F4 = تعداد داده‌های عددی موجود در جمله / طول جمله

تعداد کلمات موضوعی: مقدار عددی این معیار نیز با کسر زیر بدست می‌آید:

F5 = تعداد کلمات موضوعی موجود در جمله / ماکزیمم تعداد کلمات موضوعی در یک جمله متن

شباهت دو جمله به هم: این معیارمیزان شباهت دو جمله را به هم مشخص می‌کند که قطعا جمله ای که با جملات بیشتری شبیه است مهم می‌باشد. برای تعیین مقدار این معیار یک ماتریس n در n ایجاد شده که n تعداد جملات موجود در متن می‌باشد. قطر اصلی ماتریس برابر صفر قرار داده شده زیرامیزان شباهت یک جمله به خودش مد نظر نمی باشد. حال درایه‌های ماتریس بامیزان تعداد کلمات مشابه جمله i ام با جمله j هم تقسیم بر بالاترینمیزان تشابه دو جمله موجود در متن به صورت زیر مشخص می‌گردد:

$$F6 = \frac{sim\ (Si,Sj)}{m\ \ (sim\ (Si,Sj))}$$

تعداد اسامی خاص: تع‌داداسامی خاص در جمله موجب افزایش امتیاز آن جمله شده که مقدار عدی آن نیز به صورت زیر محاسبه می‌شود:

F7 = تعداد کلمات خاص موجود در جمله / تعداد کلمات جمله

مقادیر عددی بدست آمده از معیارهای بالا ابتدا به متغیرهای زبانی تبدیل شده و پس از ایجاد fuzzy ruleها به عنوان ورودی سیستم استنتاج فازی مورد استفاده قرار می‌گیرند. برای خروجی‌های ایجاد شده که جملات مهم متن می‌باشند نیز متغیرهای زبانی مانند بسیار مهم، مهم، کم اهمت، ... ایجاد شده که با توجه به اندازه خلاصه خوسته شده تعدادی جمله در متن خلاصه شده خروجی قرار می‌گیرند. در تحقیق مورد بررسی هیچگونه اطلاعاتی در مورد شکل توابع عضویت و روش دفازی داده نشده است.

در [۲۶] نیز از روش منطق فازی استفاده شده است. روش ارائه شده در این تحقیق یک روش ترکیبی با استفاده از الگوریتم ژنتیک (GA) و Genetic Programming (GP) برای بهینه سازی مجموعه

Rule setها و توابع عضویت فازی استفاده شده که GP برای بهینه سازی در بخش ساختاری و GA در بخش متنی (توابع عضویت فازی) بکار گرفته شده است.

نوع خلاصه‌سازی مورد مطالعه در این تحقیق نیز استخراجی بوده و مانند سایر روشهای ارائه شده در سایر مقالات ابتدا جملات متن از هم جدا شده و سپس ویژگی‌های زیر برای هر جمله از آن استخراج و آماده اعمال به سیستم فازی می‌شود. معیارهای امتیاز دهی به جملات در این تحقیق نیز مانند سایر روشها بوده و به شرح زیر می‌باشند:

۱. تعداد کلمات مشابه با عنوان مقاله
۲. تعیین جمله اول یا آخر متن بودن جمله مورد نظر
۳. تعداد کلمات موجود در جمله
۴. تعداد کلمات موضوعی موجود در متن
۵. تعدادکلمات تاکیدی موجود در متن

ویژگی‌های استخراجی فوق به عنوان ورودی‌های سیستم استنتاج فازی مورد نظر بکار می‌روند. تابع عضویت ورودی استفاده شده در این سیستم Bell و ذوزنقه ای بوده و توابع عضویت خروجی نیز از نوع Bell هستند که به سه تابع عضویت غیر مهم، میانگین و مهم تقسیم شده اند. معیارهای تصمیم گیری در نظر گرفته شده‌اند که از قرار زیر می‌باشند:

۱. جملات ابتدا و انتهای پاراگراف از اهمیت بالایی برخوردارند.
۲. جملات حاوی کلمات موضوعی دارای اهمیت بالایی می‌باشند.
۳. جملات با تعداد کلمات متوسط با شرط دارا بودن سایر معیارها انتخابی مناسب برای خروجی هستند.

مهمترین بخش در یک سیستم فازی خلاصه‌سازی متن بخش ایجاد قوانین if _ then بوده و در سیستم طراحی شده فوق ۷۲۹ قانون ایجاد شده است. با توجه به اینکه تعداد ruleهای سیستم

طراحی شده زیاد بوده و سیستم خلاصه ساز ایجاد شده برای متون تحت وب استفاده می‌شود، روش ارائه شده از ترکیب GA و GP برای یافتن ruleهای بهینه استفاده می‌کند. روشهای بهینه سازی GA و GP بکار گرفته شده از حوزه مورد بررسی این بخش خارج بوده که در جای خود مورد بررسی قرار خواهد گرفت.

چند مقاله و تحقیق دیگر در مورد خلاصه‌سازی متن با بکارگیری منطق فازی مورد مطالعه و بررسی قرار گرفت که در مجموع تمامی روشهای ارائه شده در بخش استخراج ویژگی‌ها با تفاوت بسیار اندک شبیه هم بوده و ویژگی‌های استخراجی به عنوان ورودی‌های سیستم فازی در نظر گرفته می‌شوند. تفاوت روشهای مطرح شده صرفا در انتخاب توابع عضویت متفاوت، نوع موتور فازی انتخابی و روش دفازی کردن اطلاعات خروجی می‌باشد. برخی از روشها نیز از الگوریتم‌های دیگری مانند شبکه‌های عصبی، الگوریتمهای ژنتیک، PSO و نظایر آن برای بهینه سازی ویژگی‌های استخراجی و rule baseهای ایجاد شده استفاده کرده اند. از آن جمله می‌توان به روش ارائه شده توسط Ladda Suanmali, 2009 [۲۷] اشاره کرد که فقط تفاوت در انتخاب استفاده از توابع عضئویت مثلثی، Mohammed Salem Binwahlan, 2009 [۲۸] بکار گیری الگوریتم ازدحام ذرات (PSO) در بهینه سازی و S. Santhana Megala, 2014 [۲۹] در استفاده از شبکه‌های عصبی در انتخاب جملات مهمتر پس از امتیاز دهی فازی در خروجی استفاده کرده اند.

۳-۷-چالشهای زبان فارسی در خلاصه‌سازی متن

زبان فارسی زبان رسمی سه کشور ایران، افغانستان و تاجیکستان بوده و زبان مادری بیش از یکصد میلیون نفر می‌باشد. شکل نوشتاری این زبان برخلاف اکثر زبانهای دیگر از راست به چپ بوده اما برخلاف زبان انگلیسی و سایر زبانهای با حروف لاتین که در آن هم حروف و هم لغات کاملا از هم متمایزند، در زبان فارسی پیوستگی می‌ان بیشتر کاراکترهای لغات وجود داشته و علاوه بر آن کلمات نیز در نگارش متنوع می‌باشند. همچنین ریشه یابی فعل که یکی از مراحل مهم پیش پردازش متن

برای خلاصه‌سازی می‌باشد، در زبان فارسی چالشهای خاص خود را دارد. به عنوان مثال در یک لغت به هم پیوسته می‌توان ریشه فعل، شناسه، علامت زمان فعل و حتی شناسه‌های مفعولی نیز می‌توان داشت که کار پردازش لغات را پیچیده تر می‌کند، به طوری که نمیتوان از دانش، تجربه و نرم افزارهای موجود در این زمینه استفاده نمود و تولید نرم افزاری که بتواند تمامی این پیچیدگی‌ها را حل کند فرایندی زمانبر بوده و نیازمند تلاش فراوان می‌باشد. تفاوتهای ذاتی زبانهای گسسته مانند انگلیسی با زبانهایی مانند فارسی، عربی، چینی و غیره که ساختاری پیوسته داشته و دارای تفاوتهای بنیادی در قواعد دستوری هستند، منجر به آن شده است که ادعای اعمال تغییرات در ساختار یک نرم افزار انگلیسی و به کار بردن آن برای زبان فارسی را مورد سئوال قرار داده که مستلزم آزمایشهای فراوان برای اثبات صحت آن خواهد بود. با توجه به اینکه ایجاد خلاصه‌های چکیده ای نیازمند بسیاری از اعمال پیش پردازشی مانند تشخیص مرز کلمات، ریشه یابی اسامی و افعال و بسیاری دیگر از مفاهیم درک زبان طبیعی می‌باشد و با در نظر مشکلات و چالشهای ذکر شده ایجاد خلاصه‌های چکیده ای در زبان فارسی بسیار مشکل خواهد بود. از این رو ایجاد خلاصه‌های استخراجی که در آن سیستم خلاصه ساز فقط درگیر با یافتن مرز جملات و استفاده از روشهای آماری برای تشخیص جملات مهم بوده بسیار ساده تر و کاراتر خواهد بود. در ادامه برخی از این چالشها بررسی می‌شوند.

۳-۷-۱-چالشهای نگارش زبان فارسی

برخی موارد و خصوصیات نگارشی زبان فارسی مشکلاتی را در بخش پیش پردازش اتوماتیک خلاصه‌سازی متون را دچارمشکل می‌کند. برخی از این موارد را در از قرار زیر می‌باشند[۹]:

- تفاوت در ترکیب دو کلمه با هم با فاصله و نیم فاصله. در این حالات اگر دو کلمه با فاصله از هم جدا شوند تبدیل به دو کلمه و اگر با نیم فاصله از هم جدا شوند تبدیل به یک کلمه خواهند شد. به عنوان نمونه می‌توان سیب زمینی را نام برد. این مسئله در انگلیسی بندرت اتفاق می‌افتد.

- انواع مختلف نگارش برای یک کلمه با استفاده از کاراکترهایی با تلفظ یکسان و نگارش متفاوت مانند اتاق و اطاق
- نگارش متفاوت یک کلمه با استفاده همزه یا بدون استفاده از همزه مانند مسئول و مسوول
- استفاده از کلمات ترجمه نشده انگلیسی با حروف لاتینیا فارسی مانند سورس وsource
- استفاده از ا و آ به جای هم مانند کلمه فرایند و فرآیند
- بازگذاشتن دست نویسندگان در فاصله گذاری می‌ان کلمات.
- عدم وجود یک استاندارد و دستورالعمل قطعی برای استفاده از نیم فاصله.
- عدم وجود یک قاعده ثابت و مشخص برای فاصله گذاری
- استفاده از پیشوند استمرای افعال می‌به صورت چسبان و غیر چسبان مانند کلمات می‌تواند و می‌تواند
- استفاده از پسوند جمع‌ها به دو صورت چسبان و غیر چسبان مانند کلمات آنها و آن ها
- استفاده از "ی" برای اتصال صفت و موصوف به دو صورت نگارش یا عدم نگارش آن مانند کلمات خانه سبز و خانه ی سبز

در نظر نگرفتن موارد نگارشی بیان شده موجب می‌شود تا تشخیص و تفکیک لغات لازم و غیر لازم در متن مشکل تر شده، تشخیص ریشه لغات سخت و تشخیص لغات شبیه به هم سخت تر شده و در امتیاز دهی به لغات در مراحل پردازش متن تاثیر خواهد گذاشت.

۳-۷-۲-چالشهای زبان فارسی در بخش ریشه یابی

با توجه به مطالعه و بررسی مقالات مختلف در مورد خلاصه‌سازی متن یکی از مراحل مهم در پیش پردازش متن ریشه یابی افعال متن می‌باشد. ریشه یابی متن در زبان فارسی دارای چالشهایی بوده که برخی از آنها از قرار زیر می‌باشند[۱۰]:

- بی قاعده بودن برخی از افعال و ابهام در تولید ریشه آنها (بن مضارع) مانند **رفتن** و **رو**

- عدم وجود اعراب در زبان فارسی و از آن جایی که یک کلمه بدون اعراب می‌تواند معانی مختلفی داشته باشد، تشخیص فعل از اسم را مشکل می‌نماید.
- پسوندهای چسبیده به فعل در اگثر اوقات موجب ابهام می‌شوند. به عنوان مثال در افعال **گردید** ، **گشتید**، پسوند **ید** که به ریشه فعل متصل گردیده است موجب ابهام در تشخیص بین گذشته ساده سوم شخص و حال استمراری دوم شخص می‌شود.
- حروف التزامی پیشوند که به ریشه فعل متصل شده و با حرف اول فعل یکسان است و موجب ابهام می‌شود، مانند **ببر** از مصدر **بردن**.
- اتصال ضمیر موصولی به انتهای فعل موجب ابهام شده و تشخیص این ضمیر موصولی واقعا برای یک نرم افزار دشوار خواهد بود. مانند **خوردندش** که به معنی **آن را خوردند** می‌باشد.
- مختصر کردن بعضی از افعال و اتصال آنان به کلمات قبل خود که فعل نیستند تشخیص فعل را برای نرم افزار کاملا مشکل می‌کنند. مانند **بیدارند** ، **خوشحالند**.
- تعداد زیاد افعال فارسی که عملا کار ریشه یابی این همه افعال مشکل است.
- حذف قرینه ای بخشی از افعال دو قسمتی در جمله که تعداد آنها در جمله بیشتر از یکی است. مانند حذف **ام** در فعل اول جمله **من به خانه رفته و غذا خورده ام**.
- نبود یک قاعده مشخص در ریشه یابی افعال.

۳-۷-۳-چالش‌های ذاتی زبان فارسی

زبان فارسی در ساختار و سمانتیک با زبان انگلیسی متفاوت می‌باشد. برخی از مشکلات ذاتی مربوط به متون فارسی به شرح زیر می‌باشد[۹]:

- نبود نکات گرامری تعریف شده همانند آنچه که در زبان انگلیسی وجود دارد .این مسئله در ریشه یابی و پیش پردازش متن تاثیر گذار خواهد بود

- وجود لغات ترکیبی چند جزئی همانند آب سرد کن
- در زبان فارسی گاهی اوقات ضمیرهای مفعولی و نشانه‌های جمع که به لغات متصل می‌شوند شکل یکسانی داشته که موجب پیچیدگی تفکیک این لغات برای نرم افزار می‌شوند. به عنوان مثال ان در درختان نشانه جمع بوده در حالی که ان در خوران نشانه صفت فاعلی می‌باشد.
- در زبان فارسی قاعده خاصی تبدیل ریشه فعل حال به ریشه فعل گذشته و بالعکس به صورت قانونمند وجود ندارد، در نتیجه نرم افزار برای تشخیص آنان هیچ راه مشخصی را ندارد. به عنوان مثال می‌توان همان کلمه خوران مثال قبل را ذکر کرد.
- ابهام‌های ساختاری در کلمات با چند معنی متفاوت مانند شیر با معانی شیر حیوان، شیر آب، شیر خوراکی.
- عدم وجود روشی خاص برای تشخیص اسامی و مکانهای خاص مانند استفاده از حروف بزرگ در ابتدای آنها در زبان انگلیسی.
- ابهام در معنی کلمه به علت نبود اطلاعات آوایی مربوط به حروف صدادار مانند مرد به عنوان انسان مذکر و مرد به معنی فوت کردن.
- عدم وجود یک روش مشخص و قطعی برای استفاده از نیم فاصله و فاصله گذاری در کلمات ترکیبی.
- عدم تشخیص نرم افزار در تشخیص نقش کلمه در یک جمله به دلیل قادر نبودن در تشخیص مرز کلمات. به عنوان مثال می‌توان آیین نامه نوشتن و آیین نامه رانندگی را نام برد.
- وجود کلمه‌های ترکیبی زیاد و امکان در نظر گرفته شدن دو کلمه مجزا مانند سیب زمینی.
- امکان نگارش کلمات با کاراکترهای متفاوت و نیاز به یکسان سازی آنان برای پیش پردازش مانند آئین نامه و آیین نامه.
- دشواری تشخیص کلمات اختصاری از سایر کلمات. در زبان انگلیسی کلمات اختصاری با حروف بزرگ نوشته شده که در زبان فارسی مفهومی به نام حروف بزرگ و کوچک وجود ندارد.

- نبـود یـک پایگـاه داده لغـوی ماننـد WordNet اعمـال کامـل خصوصـیات زبـان شـناختی بـرای خلاصه‌سازی را مشکل می‌کند.
- وجود کاراکترهایی که از زبان عربی وارد فارسی شده اند. زبان فارسی دارای ۳۲ حرف بـوده کـه ۲۸ تای آن با زبان عربی مشترک می‌باشد. علاوه بر این حروف کاراکترهایی دیگر نیز وجـود داشـته کـه منحصرا از زبان عربی وارد شده که در بسیاری از حالات مـی‌تـوان آنـان را در نظـر نگرفـت. از ایـن نمونه می‌توان به تشدید و تنوین اشاره کرد که در بسیاری از نگارشها درج نمی شوند.
- هوموگراف‌ها یا کلمات همشکل از دیگـر ابهامـات مـی‌باشـند. ایـن گونـه کلمـات دارای یـک شـکل نگارشی اما معانی کاملا متفاوت هستند. به عنوان نمونه کلمه **اشکال** را می‌تـوان مثـال زد کـه اگـر با فتحه تلفظ شود جمع کلمه شکل بوده و اگر با کسره تلفظ شود به معنی مشـکل و نقـص خواهـد بود.
- وجود حروفی که تلفظ شده اما نوشته نمی شود و نقش مهمی در مفهـوم کلمـات در جملـه دارد. از آن جمله می‌توان به کسره ارتباط دهنده بین صفت و موصوف اشاره کرد. به عنوان مثال در عبـارت **سیب قرمز** حرف **ب** دارای کسره بوده که موجب اتصال صفت و موصوف شـده و در جایگـاه S‘ در انگلیسی قرار دارد و یا **ی** مالکیت در عبارت **خانه ما** که نوشته نشده اما تلفظ می‌شود.
- حذف برخی از کلمات درجمله که می‌تواند ابهامی برای نرم افـزار ایجـاد کنـد. بـه عنـوان مثـال در جمله **علی خانه نیست** کلمه **در** از جمله حذف شده که می‌توانـد ایـن ابهـام را ایجـاد کـرده کـه **علی** را با **خانه** مقایسه کرده است.

۳-۸-مروی بر سابقه تحقیقات خلاصه‌سازی متن در زبان فارسی

استفاده از تئوری گراف در خلاصه‌سازی متن در زبان فارسی سابقه چندانی نداشته و منحصرا به چنـد تحقیق و مقاله خلاصه می‌شود. مهمترین سیستم هایی که برای خلاصـه‌سـازی مـتن در زبـان فارسـی شامل موارد زیر می‌باشد:

۳-۸-۱- سیستم خلاصه ساز فارسی FarsiSum

FarsiSum یک خلاصه ساز مبتنی بر وب است که SweSum را برای زبان فارسی مدل می‌کند. این سیستم قادر به خلاصه کردن متون روزنامه های فارسی با قالب HTML و متن کد شده با فرمت یونیکد بوده ، تحت ویندوز و به زبان پرل نوشته شده است. از ساختاری مشابه SweSum استفاده می‌کند با این تفاوت که چند ماژول آن برای کار با محتوای یونیکد و رمزگذاری UTF_ 8 تغییر داده شده است. خلاصه ساز SweSum از نوع استخراجی بوده که برای زبان های سوئدی ، فرانسوی ، آلمانی ، انگلیسی، دانمارکی و اسپانیایی پیاده سازی شده است.

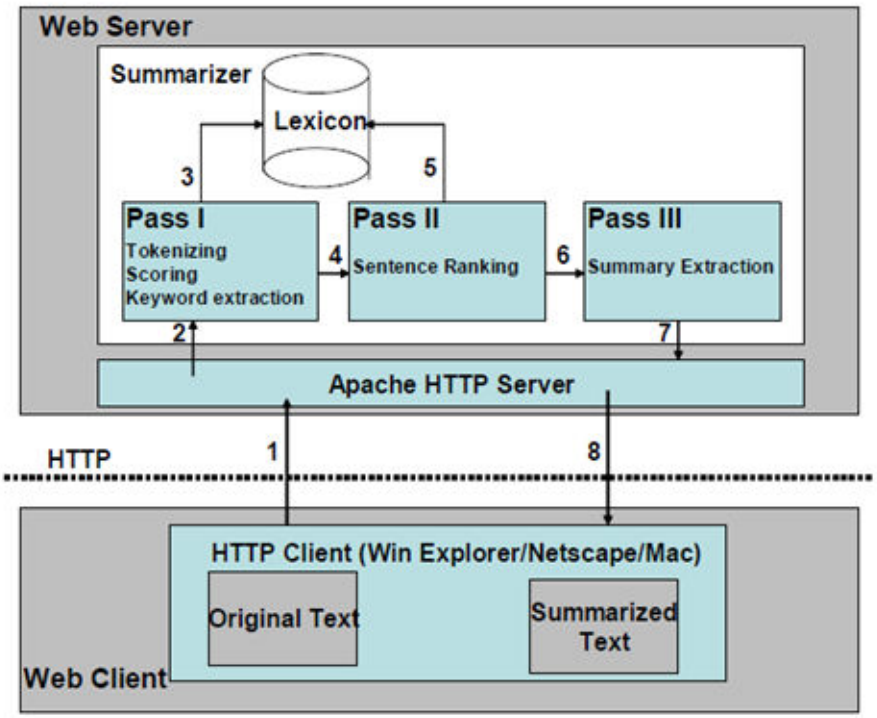
خلاصه ساز فارسی FarsiSum یک خلاصه ساز مطرح بوده که بر پایه خصوصیات آماری متن کار می‌کند و خصوصیات زبان شناختی متن را به جز stop listها را برای پیش پردازش در نظر نمی گیرد. خلاصه ساز متنی با فرمت html یا text را به عنوان ورودی از روزنامه یا گزارش دریافت کرده و خلاصه آن را تولید می کند. به این منظور ابتدا محدوده جملات و کلمات را تعیین کرده و کلمات مخفف شده برچسب می خورند. سپس فرکانس ریشه ی کلمات بازمانند اسامی، افعال و صفات با استفاده از یک لیست واژگان ایستا محاسبه شده و به جملات متن امتیاز می‌دهد[۶][۷].

در farsisum برخی از ماژولها برای کار با اسناد محتوی Unicode و رمزگذاری UTF- 8 و نگارش راست به چپ اصلاح شده و به دلیل عدم استفاده از واژگان فارسی به جای استفاده از لیست کلمات و مخفف ها تنها با تهیه لیست کلمات غیر مفید (۲۰۰ کلمه از پرتکرارترین کلمات فارسی شامل افعال، ضمایر، ، قیدها حروف ربط حروف اضافه و تعریف حروف) عمل پردازش را انجام می دهد. در ارزیابی این سیستم از مقایسه روش با خلاصه مرجع دستی با تعداد پایین شرکت کننده ها و متن های با رمزگذاری UTF-8 استفاده شده است. Farsisum از روش کلاسیک بدون توجه به معنی کلمات استفاده می‌کند. در کل خلاصه سازهایی مانند farsisum که مدعی هستند وابسته به حوزه

خاصی نمی باشند تنها می‌توانند متون خبری را خلاصه کرده و در خلاصه‌سازی متـون علمـی کـارایی پایینی دارند. جهت تشخیص شکل و معماری FarsiSum باید به معماری SweSum توجه کرد.

معماری SweSum: خلاصه ساز SweSum یک سیستم خلاصه کننده متـون تحـت وب بـوده کـه در موسسه Royal Institute of Technology (KTH) سوئد ایجاد شده اسـت. ایـن سیسـتم تـگ‌هـای HTML شامل متون خبری را با استفاده از متدهای آماری و زبان شناسی و با کمک توابـع مکاشـفه ای خلاصه می‌کند. سیستم یاد شده برای زبانهای سوئدی، نروژی، دانمارکی، اسپانیایی، آلمـانی ، انگلیسـی و فرانسوی نیز وجود داشته که نوع تکمیل شده فارسـی آن نیـز تحـت نـام farsisum ایجـاد گردیـده است. SweSum از بسیاری از تکنیک‌های خلاصه‌سازی موجود استفاده نکرده که از آن جمله مـی‌تـوان به تکنیک زنجیره‌های لغوی اشاره کرد. اسـتفاده از تکنیـک زنجیـره‌هـای لغـوی در یـک خلاصـه سـاز نیازمند داشتن و استفاده از WordNet بوده که SweSum توانایی استفاده از آن را ندارد.

معماری SweSum در تصویر زیر نمایش داده شده اسـت. برنامـه خلاصـه سـاز در سـمت سـرور قـرار داشته که کاربر توسط پروتکل HTTP با آن در تعامل بوده و متنی که باید خلاصه شـود بـه آن ارسـال و خلاصه را دریافت می‌کند.



مراحل اجرا شامل سه بخش است در بخش اول:

Tokenizer :

در این قسمت اعمال زیر انجام می‌شود:

- حذف کاراکترهای انتهای خط (\n)
- یافتن مرزهای جملات
- یافتن مرزهای کلمات
- استخراج کلمات کلیدی (SweSum نامها، صفات و قیدهای زمان را به عنوان کلمات کلیدی در نظر می‌گیرد)

Scoring : در این مرحله به خطوط HTML موجود امتیاز داده می‌شود که در سیستم یاد شده به خطوطی که شامل تگهای html بوده هیچ امتیازی داده نشده و به خطوط متنی با توجه به قواعدی امتیاز داده می‌شود. این قواعد در مورد خطوط متفاوت فرق داشته که به عنوان مثال خطوط دارای

موقعیت مکانی خاص (خطوط اول هر پاراگراف) ، خطوط دارای کلمات کلیدی ، خطـوط شـامل مقـادیر عددی و خطوط بولد دارای امتیاز بالاتری هستند.

بخش دوم: در این بخش به کلمات موجود در هر جمله امتیازی تعلق می‌گیرد که این امتیاز دهـی هـم دارای شرایطی بوده که به عنان مثال کلمات کلیدی ، اعداد و تاریخها و کلمات با فرکانس بـالا در مـتن از امتیاز بیشتری برخوردار خواهند بود و کلماتی مانند حروف اضافه هیچ امتیازی ندارند.

بخش سوم: در این مرحله فایل نهایی خلاصه شده در قالب html ایجاد می‌شود. این فایل شامل مـوارد زیر است:

- تمام خطوط غیر متنی و تگهای html
- خطوط متنی با امتیاز بالا
- سایر اطلاعات آماری در مورد متن خلاصه شده مانند تعداد جملات ، تعداد کلمات ، تعداد خطـوط و کلمات کلیدی با فرکانس بالای موجود در متن

۳-۸-۲- سیستم Automatic Persian Text Summarization

در این سیستم از یک روش ترکیبی برای خلاصه‌سازی متن استفاده می‌شود. تکنیک هایی که در ایـن سیستم بکار گرفته شده‌اند زنجیره لغوی و تئوری گراف هستند که جملات مهم متن را بر پایه کلمـات و عبارات نشانه ای، تعداد جملات مشابه، تشابه بین جملات و تشابه بـین جمـلات و جملـه موضـوعی استفاده کرده است.

۳-۸-۳- سیستم Hybrid Farsi Text Summarization

در یکی از تحقیقات انجام شده از برخی از معیارهای امتیاز دهی به جملات که در مراحـل قبـل ذکر شد استفاده شده اما نوآوری آن در این است که در این تحقیق علاوه بر معیـار جمـلات بـا

اهمیت از معیار ارتباطات با اهمیت بین جملات نیز استفاده شده است[۵]. در بخش زیر به بررسی این تحقیق می‌پردازیم.

در تحقیق یاد شده خلاصه‌سازی متن از نوع استخراجی می‌باشد. نوع گراف بکار گرفته شده غیر جهت دار بوده که در آن هم مانند سایر روشهای خلاصه‌سازی بر پایه گراف جملات در جایگاه نودهای گراف و ارتباط بین این جملات نیز در جایگاه یالهای گراف قرار می‌گیرند. در این روش علاوه بر نودها به یالها نیز امتیاز داده می‌شود. معیارهایی که برای امتیاز دهی به نودها در این روش استفاده می‌شود پنج مورد از معیارهایی است که در سایر روشهای خلاصه‌سازی متن مبتنی بر گراف انجام شده اند. این معیارها عبارتند از:

- تعداد یالهای متصل به هر نود
- فرکانس کلمات در یک جمله
- تعداد کلمات کلیدی موجود در جمله
- میزان انحرافی که بین جمله مربوطه با جمله موضوعی وجود دارد
- تعداد معیارهایی که به طور همزمان در یک جمله وجود دارند. در این روش علاوه بر جمع جبری امتیازهای موجود در جمله به خود عدد تعداد معیار نیز امتیاز داده می‌شود.
- این پنج معیار برای هر جمله با معادله زیر محاسبه شده که مجموع امتیاز آن جمله را مشخص می‌کند. T_w مجموع امتیاز جمله، W_i امتیاز هر معیار و C_i نیز ضریبی بوده که به صورت تجربی برای هر نوع متن بر اساس حوزه آن در نظر گرفته می‌شود.
- $T_w = \sum_{i=1}^{5} Ci.Wi$
- سپسمیزان انحراف هر جمله با جمله موضوعی متن بدست به صورت زیر بدست می‌آید:
- $Diff = (N - I) / D$
- در این معادله Diffمیزان انحراف، N تعداد کلمات موجود در جمله، I تعداد کلمات مشترک بین جمله مورد نظر با جمله موضوعی و D تعداد تنوع کلمات در کل متن می‌باشد. امتیاز اصلی نود

حاوی جمله با توجه به معادله زیر محاسبه شده که در آن S مقدار امتیاز نـود و P تعـداد کلمـات جمله موضوعی می‌باشد.

- $S = (P / D) - Diff$
- همانطور که قبلا اشاره شد این تحقیق علاوه بر امتیاز دهی به نود ها، به یالهای گـراف نیـز امتیـاز دهی می‌کند. معیارهای امتیاز دهی به یک یال متصل کننده دو نود از قرار زیر می‌باشد:
- تعداد کلمات مشترک بین دو جمله
- تعداد کلمات کلیدی مشترک بین دو جمله
- تعداد کلمات انگلیسی شمترک موجود در دو جمله (در جملات فارسی)
- وجود کلمات مشترکی که توضیحاتی در پاورقی دارند
- وجود دو جمله مشابه در یک پاراگراف

حال با توجه به معیارهای مربوط به لبه‌هامیزان امتیاز هر جمله با در نظـر گـرفتن هـر ده معیار محاسبه می‌شود:

$$T_w = \sum_{i=1}^{10} Ci.Wi$$

در این مقدار C_i برای iهای یک تا پنج بیشتر در نظر گرفته می‌شـود. دلیـل آن مـی‌تـوانـد ایـن باشد که اهمیت امتیاز جملات بیشتر از اهمیت امتیازات رابطه بین آنان می‌باشد.

روش یاد شده از دو معیار برای ارزیابی دقت خلاصه‌های تولید شـده اسـتفاده مـی‌کنـد. ایـن دو معیار Precision و Recall بوده و برای ارزیـابی سیسـتمهای خلاصـه‌سـازی اسـتخراجی بکـار می‌روند. معیار Recall به صورت زیر محاسبه می‌گردد:

$$Recall = Correct / (Correct + Missed)$$

در این معیار مقدار Correct برابر با تعداد جملاتی است که در هر دو متن خلاصه شده توسط فرد خبره و ماشین وجود داشته و مقدار Missed تعداد جملاتی است که در متن خلاصه شده توسط فرد خبره وجود داشته اما در متن خلاصه شده توسط سیستم موجود نمی باشد.

معیار Precision نیز به صورت زیر محاسبه می‌شود:

Precision = Correct / (Correct + Wrong)

در این معیار نیز مقدار Wrong نیز تعداد جملاتی است که در متن خلاصه شده توسط نرم افزار وجود داشته اما در متن خلاصه شده توسط فرد خبره وجود ندارد. در نهایت هر دو معیار با هم ترکیب شده و برای تعیینمیزان دقت خلاصه تولید شده از آن استفاده می‌شود:

F1 = (2 * (Precision * Recall)) / (Precision + Recall)

فصل چهارم

مدل پیشنهادی

۴-۱- مقدمه

در این فصل به تشریح مدل پیشنهادی پرداخته و نتایج را بررسی خواهیم کرد.

۴-۲-مدل پیشنهادی

مدل پیشنهادی این پایان‌نامه، یک مدل از نوع استخراجی است. بدین معنا که چندین جمله از خود متن انتخاب شده و بعنوان خلاصه در نظر گرفته می‌شود. این مدل شامل مراحل زیر است:

۱. بخش‌بندی:
۲. یافتن ریشه کلمات
۳. استخراج کلمات کلیدی
۴. وزن‌دهی به کلمات
۵. وزن‌دهی به جملات براساس کلمات کلیدی
۶. استخراج جملاتی که وزن بیشتری دارند
۷. انتخاب جملاتی که پیوستگی بیشتری دارند از بین جملات مرحله قبل

مراحل (۱) تا (۶) تقریبا مراحلی است که سایرین نیز استفاده کرده‌اند. در واقع نوآوری این پایان‌نامه، مرحله (۷) بوده که باعث استخراج جملات دقیق‌تری خواهد شد.

مراحل فوق را برروی یک نمونه که از مجموعه داده همشهری انتخاب شده است، اجرا می-کنیم. این نمونه که در شکل (۴-۱) نشان داده شده است، از نوع "ادب و هنر" می‌باشد.

جاودانگی در زندگی گروهی از طریق هنر: نگاهی به نمایشگاه آثار هنری احمد طباطبایی موضوع آثار طباطبایی مورچگان هستند ولی در باطن چنین ظاهری، این انسانهاهستند که در هیبتی حشره گونه در تابلوهای نقاشی نمایشگر گوشه هایی از زندگی خود هستند. مورچه ای را دیده ایم که بار سنگین خودرا به دوش می کشد و در بین راه خسته می شود یا در نتیجه پیش آمدن یک حادثه اتفاقی بار از دستش رها شده و برای گریز از خطر آن را می گذارد و فرار می کند. اما ساعتی بعد به همان نقطه برمی گردد و بار را دوباره به دوش کشیده نفس زنان و عرق ریزان آن را به منزل او می رساند می داند که اگر در میان راه از بین رفت، یک مورچه دیگر با همان استعداد غریزه عملی را که او شروع کرده به پایان می رساند. پس او می داند که بعد از خودش نسل آینده ای وجود دارد که باید از مزایای آن استفاده نماید. چه داستان درازی، انتها و پایانی هم میلیونها ندارد سال گذشته، میلیاردها سال دیگر خواهد گذشت وهمین برنامه میلیونها بار دیگر تجدید می شود. پس باید گفت سیر حیات و تکامل پایان ناپذیر است. اسرار جهان پایانی ندارد. پس اگر بخواهیم به یک چیز پایان ناپذیر برسیم چون انتهایی در بین نیست، به جایی نمی رسیم. (از کتاب مورچگان، نوشته موریس مترلینگ)

شکل (۴-۱) نمونه ای از مجموعه داده همشهری

۴-۲-۱- بخش‌بندی

اولین مرحله در مدل پیشنهادی، جداسازی جملات و سپس کلمات می‌باشد. برای این منظور ابتدا با استفاده از نشانگرهای ؟ و ! و . جملات را از یکدیگر جدا می‌کنیم. بدین-ترتیب که متن ورودی را دریافت کرده و سپس با دیدن هر یک از این سه نشانگر، جملات استخراج می‌شوند. بعد از انجام فرآیند بخش‌بندی برروی مجموعه داده فوق، هفت جمله به صورت زیر استخراج شدند:

جاودانگی در زندگی گروهی از طریق هنر: نگاهی به نمایشگاه آثار هنری احمد طباطبایی موضوع آثار طباطبایی مورچگان هستند ولی در باطن چنین ظاهری، این انسانهاهستند که در هیبتی حشره گونه در تابلوهای نقاشی نمایشگر گوشه هایی از زندگی خود هستند

مورچه ای را دیده ایم که بار سنگین خودرا به دوش می کشد و در بین راه خسته می شود یا در نتیجه پیش آمدن یک حادثه اتفاقی بار از دستش رها شده و برای گریز از خطر آن را می گذارد و فرار می کند

اما ساعتی بعد به همان نقطه برمی گردد و بار را دوباره به دوش کشیده نفس زنان و عرق ریزان آن را به منزل او می رساند می داند که اگر در میان راه از بین رفت، یک مورچه دیگر با همان استعداد غریزه عملی را که او شروع کرده به پایان می رساند

پس او می داند که بعد از خودش نسل آینده ای وجود دارد که باید از مزایای آن استفاده نماید

چه داستان درازی، انتها و پایانی هم میلیونها ندارد سال گذشته، میلیاردها سال دیگر خواهد گذشت وهمین برنامه میلیونها بار دیگر تجدید می شود

پس باید گفت سیر حیات و تکامل پایان ناپذیر است. اسرار جهان پایانی ندارد

پس اگر بخواهیم به یک چیز پایان ناپذیر برسیم چون انتهایی در بین نیست، به جایی نمی رسیم. (از کتاب مورچگان، نوشته موریس مترلینگ)

شکل (۴-۲) نمونه ای از مجموعه داده همشهری

۴-۲-۲-یافتن ریشه کلمات

این مرحله یک کار پردازشی دستوری است که باعث بهبود کارایی الگوریتم می‌شود. نمونه این کار در زبان انگلیسی انجام شده و الگوریتم‌های آن نوشته شده است. برای مثال کتابخانه NLTK در زبان پایتون، این مورد را پیاده‌سازی کرده است. برای مثال حذف ing از انتهای کلمات. در زبان فارسی برای یافتن ریشه کلمات ناچار هستیم از یک دیکشنری استفاده کنیم. برای مثال تمام شکلهای رفتم، رفتی، رفت، می‌روند و ... به مصدر از مصدر رفتن می‌باشد. ما در این پایان نامه برای تعداد محدودی از افعال، این دیکشنری را ایجاد کرده‌ایم.

جدول (۴-۱) استخراج کلمات بکار رفته در متن

واژه	فراوانی	واژه	فراوانی	واژه	فراوانی	واژه	فراوانیت
جاودانگی	۱	باطن	۱	کشیدن	۲	آن	۳
در	۸	چنین	۱	و	۸	گذاشتن	۱
زندگی	۲	ظاهری	۱	بین	۳	فرار	۱
گروهی	۱	این	۱	راه	۲	کردن	۲
از	۷	انسان	۱	خسته	۱	اما	۱
طریق	۱	که	۶	شدن	۲	ساعتی	۱
هنر	۱	هیبتی	۱	یا	۱	بعد	۲
نگاهی	۱	حشره	۱	نتیجه	۱	همان	۲
به	۸	تابلو	۱	پیش	۱	نقطه	۱
نمایشگاه	۱	نقاشی	۱	آمدن	۱	برگشتن	۱
آثار	۲	نمایشگر	۱	یک	۳	دوباره	۱
هنری	۱	گوشه	۱	حادثه	۱	نفس	۱
احمد	۱	خود	۲	اتفاقی	۱	زنان	۱
طباطبایی	۲	را	۶	دست	۱	عرق	۱
موضوع	۱	دیدن	۱	رها	۱	ریزان	۱
مورچه	۴	بار	۴	برای	۱	منزل	۱
هستند	۳	سنگین	۱	گریز	۱	او	۳
ولی	۱	دوش	۲	خطر	۱	رساندن	۲
دانستن	۲	پایان	۳	نسل	۱	چه	۱
اگر	۲	پس	۳	آینده	۱	داستان	۱
میان	۱	خودش	۱	وجود	۱	درازی	۱
رفتن	۱	غریزه	۱	نماید	۱	انتها	۱
با	۱	عملی	۱	داشتن	۱	پایانی	۲

استعداد	۱	شروع	۱	استفاده	۱	هم	۱
مزایای	۱	باید	۲	میلیون	۲	گذشته	۱
گذشت	۱	نداشتن	۲	سال	۲	گفت	۱
میلیارد	۱	دیگر	۲	خواهد	۱	سیر	۱
همین	۱	برنامه	۱	تجدید	۱	حیات	۱
تکامل	۱	ناپذیر	۲	است	۱	اسرار	۱
جهان	۱	خواستن	۱	چیز	۱	رسیدن	۱
چون	۱	نیست	۱	جایی	۱	انتهایی	۱
نرسیدن	۱	کتاب	۱	نوشته	۱	موریس	۱

۴-۲-۳-استخراج کلمات کلیدی

در این مرحله با استفاده از الگوریتم RAKE ، کلمات کلیدی را استخراج خواهیم کرد. این الگوریتم نیازمند استخراج تمامی کلمات و سپس وزن‌دهی به آنها می‌باشد. بدین منظور ابتدا تمامی کلماتی که در متن فوق وجود دارد می‌بایست استخراج گردد. جدول (۴-۱) کلمات استخراج شده، به همراه فرکانس هر یک را نشان می‌دهد.

۴-۲-۴-وزن‌دهی به کلمات کلیدی

وزن‌دهی به کلمات براساس الگوریتم RAKE انجام می‌شود. بدین ترتیب که یک ماتریس تشکیل شده و درایه‌های آن، کلمات موجود در متن می‌باشد. قطر اصلی این ماتریس، تعداد تکرار یک کلمه را نشان داده و دیگر درایه‌ها، تکرار کلمات به همراه سایر کلمات را نشان می‌دهد. رابطه (۴-۱) نحوه بدست آوردن هر درایه را نشان می‌دهد. بخشی از ماتریس ایجاد شده با استفاده از الگوریتم RAKE برای متن شکل (۴-۱) در جدول (۴-۲) نشان داده شده است

جدول (۴-۲) اعمال الگوریتم RAKE برروی بخشی از متن داده شده

	جاودانگی	زندگی	گروهی	هنر	نگاهی	نمایشگاه	آثار	هنری	احمد	طباطبایی	موضوع	مورچه	هستند	ولی	دانستن	خود
جاودانگی	۱															
زندگی		۲	۱													۱
گروهی			۱													
هنر				۱					۱							
نگاهی					۱											
نمایشگاه						۱	۱									
آثار							۲	۱		۱						
هنری								۱	۱							
احمد									۱	۱						
طباطبایی										۲	۱	۱				
موضوع							۱				۱					
مورچه												۴	۱			
هستند													۳	۱		
ولی														۱		
دانستن															۲	
خود													۱			۲

$$w = \deg(word_i) + fre(w\square\ \ _i)/fre(word_i) \qquad (۱-۴)$$

ماتریس نشان داده شده در جدول (۴-۳) وزن‌دهی براساس الگوریتم RAKE را نشان می‌دهد. در این ماتریس کلمات پرتکرار حدف شده‌اند. برای این منظور، ابتدا لازم است تا آستانه‌ای تعیین شده و سپس کلماتی که تکراری بالاتر از این آستانه دارند حدف شوند. ما این آستانه برای متن فوق را برابر با ۴ در نظر گرفتیم. لذا کلماتی که تعداد تکرار آنها بیش‌تر از ۴ بوده است، از جدول فوق حدف شده‌اند. همچانکه در جدول (۴-۳) نشان داده شده است، کلماتی مانند "از، که و ..." حدف شده‌اند.

۴-۲-۵-وزن‌دهی به جملات براساس کلمات کلیدی

در این مرجله، با توجه به وزن کلمات (جدول (۴-۳))، جملات موجود در متن وزن‌دهی می‌شوند. جدول (۴-۴) وزن هفت جمله موجود در شکل () را نشان می‌دهد.

۴-۲-۶- استخراج جملاتی که وزن بیشتری دارند

در نهایت از بین این هفت جمله، آنهایی که وزن بیشتری دارند، انتخاب شده و از کنار هم قرار دادن آنها، خلاصه نهایی شکل می‌گیرد. جدول (۴-۴) وزن‌دهی جملات را نشان می‌دهد. براین اساس سه جمله اول، دوم و سوم بعنوان کاندید برای خلاصه انتخاب خواهند شد شکل (۴-۳) خلاصه این سه جمله را نشان می‌دهد.

جاودانگی در زندگی گروهی از طریق هنر: نگاهی به نمایشگاه آثار هنری احمد طباطبایی موضوع آثار طباطبایی مورچگان هستند ولی در باطن چنین ظاهری، این انسانهاهستند که در هیبتی حشره گونه در تابلوهای نقاشی نمایشگر گوشه هایی از زندگی خود هستند

مورچه ای را دیده ایم که بار سنگین خودرا به دوش می کشد و در بین راه خسته می شود یا در نتیجه پیش آمدن یک حادثه اتفاقی بار از دستش رها شده و برای گریز از خطر آن را می گذارد و فرار می کند

اما ساعتی بعد به همان نقطه برمی گردد و بار را دوباره به دوش کشیده نفس زنان و عرق ریزان آن را به منزل او می رساند می داند که اگر در میان راه از بین رفت، یک مورچه دیگر با همان استعداد غریزه عملی را که او شروع کرده به پایان می رساند

شکل (۴-۳) سه جمله کاندید برای انتخاب خلاصه

تا اینجا روشی بکار برده شد، که قبلا توسط دیگران پیشنهاد شده است. تنها نوآوری این پایان‌نامه ایجاد یک دیکشنری به منظور یافتن ریشه بخشی از کلمات (افعال) بوده است. در اینجا می‌خواهیم با استفاده از روشی، جملات با وزن بیشتر که پیوستگی بالایی دارند، بعنوان خروجی نهایی انتخاب شود این امر باعث می‌شود تا جملات انتخاب شده، خلاصه‌ای بهتری (از لحاظ محتوا) استخراج شود و می‌تواند بعنوان نوآوری این پایان نامه محسوب شود. پگونگی انتخاب این جملات در ادامه توضیح داده می‌شود.

جدول (۴-۳) وزن‌دهی کلمات متن داده شده براساس الگوریتم RAKE

واژه	وزن	واژه	فراوانی	واژه	فراوانی	واژه	فراوانیت
جاودانگی	۱	باطن	۱	کشیدن	۱.۵	آن	۱.۵
زندگی	۲	چنین	۱	بین	۱.۶۷	گذاشتن	۱
گروهی	۱	ظاهری	۱	خسته	۱	فرار	۱
طریق	۱	انسان	۱	ساعتی	۱	اما	۱
هنر	۱	شدن	۱.۵	یا	۱	راه	۱.۵
نگاهی	۱	هیبتی	۱	همان	۲	بعد	۱
حشره	۱	نتیجه	۱	پیش	۱		
نمایشگاه	۱	تابلو	۱	آمدن	۱	نقطه	۱
آثار	۲	نقاشی	۱	یک	۲	برگشتن	۱
هنری	۱	نمایشگر	۱	حادثه	۱	دوباره	۱
احمد	۱	گوشه	۱	اتفاقی	۱	نفس	۱
طباطبایی	۲	خود	۲	عرق	۱	زنان	۱
موضوع	۱	دست	۱	رها	۱		
مورچه	۱.۷۵	دیدن	۱	برای	۱	ریزان	۱
هستند	۱.۵	بار	۱.۵	گریز	۱	منزل	۱
ولی	۱	سنگین	۱	خطر	۱	او	۲
دانستن	۱	دوش	۲	نسل	۱	رساندن	۱.۵
اگر	۱.۵	پایان	۲	آینده	۱	چه	۱
میان	۱	پس	۲	وجود	۱	داستان	۱
رفتن	۱	خودش	۱	نماید	۱	درازی	۱
با	۱	غریزه	۱	داشتن	۱	انتها	۱
استعداد	۱	عملی	۱	استفاده	۱	پایانی	۲
مزایای	۱	شروع	۱	میلیون	۲	هم	۱
گذشت	۱	باید	۱.۵	سال	۲	گذشته	۱

میلیارد	۱	نداشتن	۱.۵	خواهد	۱	گفت	۱
همین	۱	دیگر	۲	تجدید	۱	سیر	۱
تکامل	۱	برنامه	۱	است	۱	حیات	۱
جهان	۱	ناپذیر	۲	چیز	۱	اسرار	۱
چون	۱	خواستن	۱	جایی	۱	رسیدن	۱
نرسیدن	۱	نیست	۱	نوشته	۱	انتهایی	۱
مترلینگ	۱	کتاب	۱	کردن	۱	موریس	۱

جدول (۴-۴) وزن‌دهی جملات براساس وزن کلمات تشکیل دهنده آن

	جمله اول	جمله دوم	جمله سوم	جمله چهارم	جمله پنجم	جمله ششم	جمله هفتم
وزن	۳۱.۲۵	۲۸.۶۷	۳۰	۱۱.۵	۱۷	۱۲.۵	۱۲

۴-۲-۷- استخراج جملات که پیوستگی بیشتری دارند

در اینجا ابتدا می‌بایست معیاری به منظور اندازه‌گیری پیوستگی جملات یافت شود. برای این منظور ما از روش خوشه‌بندی استفاده خواهیم کرد. بدین ترتیب جملاتی که شبیه به یکدیگر هستند، در یک خوشه قرار گرفته و سپس جملاتی که وزن بیشتری دارند و در یک خوشه قرار دارند، انتخاب خواهند شد.

الگوریتم‌های خوشه‌بندی متفاوتی وجود دارد که به ظور کلی در دو دسته قرار خواهند گرفت:

۱. الگوریتم‌های خوشه‌بندی که تعداد خوشه‌ها از قبل مشخص است
۲. الگوریتم‌های خوشه‌بندی که تعداد خوشه از قبل مشخص نیست

الگوریتم‌هایی که در دسته اول قرار می‌گیرند (مانند kmeans، fuzzy kmeans و ...) پیاده‌سازی ساده‌تری داشته کارایی مناسبی دارند. مشکل آنها مشخص بودن تعداد خوشه‌ها

است که می‌بایست بعنوان یه پارامتر به الگوریتم داده شود. از طرف دیگر الگوریتم‌های نوع دوم که سلسله مراتبی نیز نامیده می‌شوند، نیاز به مشخص بودن تعداد خوشه در ابتدا ندارند. در عوض آنها پیچیده‌گی بیشتری داشته و نیازمند تعیین دقیق پارامترهای خود می‌باشند (برای مثال در مدل تخمین بیزین، میانگین و واریانس توزیع نمونه‌ها می‌بایست به دقت مشخص گردد. در مدل birch، می‌بایست آستانه تجمیع خوشه‌ها مشخص گردد).

در این پایان‌نامه از مدل خوشه‌بندی kmeans استفاده شده است.

به منظور خوشه‌بندی جملات، می‌بایست پیش‌پردازش‌هایی برروی آنها صورت گیرد. یکی از این پیش‌پردازش‌ها استخراج ویژگی[۳] می‌باشد. یکی از روشهای استخراج ویژگی متداول، ماتریس TF-IDF است. در این روش یک ماتریس شکل گرفته که هر ستون آن یک جمله، و هر سطر آن یک کلمه[۴] (عبارت) می‌باشد. سپس درایه‌های این ماتریس یک در صورت وجود آن کلمه در متن، و صفر در صورت عدم وجود آن، می‌باشد. جدول (۴-۵) ماتریس TF-IDF برای جمله‌های استخراجی از متن شکل (۴-۱) را نشان می‌دهد.

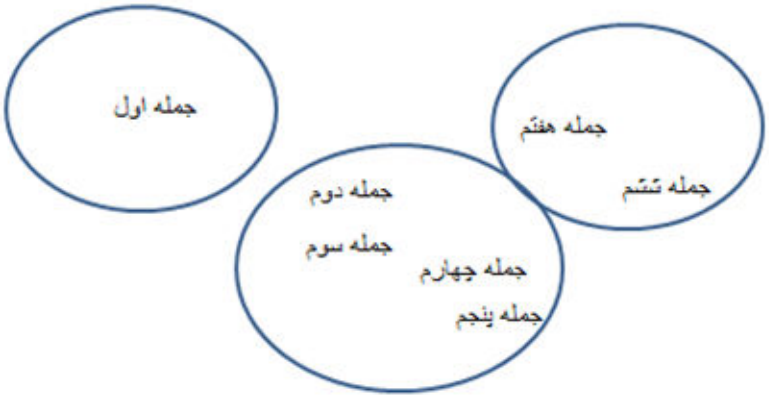
بعد از بدست آوردن جدول فوق، آنرا بعنوان ورودی به الگوریتم kmeans می‌دهیم. الگوریتم kmeans تعداد خوشه‌ها را بعنوان پارامتر از ورودی دریافت می‌کند. تعداد خوشه‌ها را برای نمونه، سه انتخاب نمودیم. خروجی الگوریتم به صورت شکل (۴-۳) بود. همانطور که در شکل فوق نشان داده شده است، جمله‌های دوم، سوم، چهارم و پنجم بسیار شبیه یه یکدیگر بوده و لذا آنها در یک خوشه قرار خواهند گرفت. دلیل ترسیم شکل فوق بدین منظور ترسیم شده است که شباهت جمله‌ها از خوشه‌های متفاوت را نشان دهد. برای مثال جمله اول شبیه به جمله دوم بوده، اما این شباهت انقدری نبوده که الگوریتم آنها را در یک خوشه قرار دهد.

---

[۳] - Feature
[۴] - Term

حال با توجه به جدول (۴-۴) و نیز شکل (۴-۳) جمله اول بعنوان خلاصه متن فوق استخراج می‌شود. بدین دلیل که براساس جدول (۴-۴) جمله اول وزنی بالاتری نسبت به بقیه جملات دارد و نیز بدین دلیل که جمله اول با هیچ جمله‌ای در یک خوشه قرار نمی-گیرد به تنهایی به عنوان متن استخراجی انتخاب می‌شود. این در حالی است که اگر تنها از وزن جملات برای استخراج متون استفاده می‌کردیم، بین سه جمله اول، دوم و سوم چالشی بوجود می‌آمد. بدین دلیل که وزن این جملات بسیار به یکدیگر نزدیک است.

دقت یه این موضوع اهمیت دارد که مثال فوق، نمی‌تواند برتری روش فوق را به خوبی نشان دهد بدین دلیل که تنها جند جمله استفاده شده است. روش پیشنهادی از یک مدل خوشه-بندی استفاده می‌کند که نیازمند داده‌های بیشتری به منظور خوشه‌بندی بهتر می‌باشد. این مثال تنها بیان می‌کند که این مدل به مراتب کاراتر از مدل قبلی (بدون در نظر گرفتن خوشه‌بندی) خواهد بود.



شکل (۴-۴) خوشه‌بندی جملات استخراجی

جدول (۴-۵) ماتریس IF-IDF

	جمله اول	جمله دوم	جمله سوم	جمله چهارم	جمله پنجم	جمله ششم	جمله هفتم
بین		۱	۱				۱
خسته		۱					
راه		۱	۱				
پیش		۱					
آمدن		۱					
یک		۱	۱				۱
حادثه		۱					
اتفاقی		۱					
دست		۱					
رها		۱					
شدن		۱			۱		
گریز		۱					
خطر		۱					
گذاشتن		۱			۱		
فرار		۱					
یا		۱					
کردن		۱	۱				
ساعتی			۱				
بعد			۱	۱			
همان			۱				
نقطه			۱				
برگشتن			۱				
دوباره			۱				

نفس			۱				
زنان			۱				
عرق			۱				
ریزان			۱				
منزل			۱				
او			۲	۱			
رساندن			۲				
دانستن			۱	۱			

ارزیابی روش پیشنهادی

به منظور ارزیابی روش پیشنهادی، یک مجموعه داده ۲۰۰ نمونه‌ای جمع‌آوری گردید. این مجموعه داده شامل مطالب روزنامه همشهری و مجله افتاب در زمینه‌های ادب و هنر، اقتصادی، فرهنگی و ... می‌باشد. نمونه‌های از این مجموعه داده در شکلهای (۴-۴) و (۴-۵) آمده است. به منظور ارزیابی روش پیشنهادی از سه معیار زیر استفاده شده است:

- مناسب بودن: آیا در کل خلاصه‌گیری حاصل شده مناسب می‌باشد
- انسجام: آیا جملات استخراج شده دارای انسجام هستند
- جامعیت: آیا خلاصه استخراج شده، تمام موضوع متن را پوشش می‌دهد.

معیارهای فوق، معیار کمی نیستند و لذا به منظور اندازه‌گیری آنها به ناچار از عامل انسانی استفاده شده است بدین صورت که بعد از اعمال مدل فوق برروی مجموعه داده جمع‌آوری شده، خلاصه آنها را بدست آورده و سپس آنها را به ۲۰ نفر از خبرنگاران، خوانندگان روزنامه و سردبیر رونامه داده و از آنها خواسته شد تا عددی بین ۰ تا ۱۰ را به خلاصه‌های استخراج شده نسبت بدهند. این روش آزمون، کارایی مدل پیشنهادی را از نظر کیفی بررسی نموده و معیار خوبی به منظور ارزیابی آن خواهد بود.

تحقیقات باستان‌شناسی اخیر نشان داده است که انسان عصر حجر که تازه از کوه فرود آمده و در دشت سکونت گزیـده بود بر روی مسیر کمانی شکل اطراف کویر نمک استقرار یافته بود . به گزارش روابط عمومی پژوهشگاه میراث فرهنگی و گردشگری، بهرام عنانی سرپرست هیات بررسی و باستان‌شناسی دهستان مرکزی شهرستان طبس با اعـلام ایـن مطلـب افزود: فلات مرکزی ایران که بخشی از آن را کویر تشکیل می‌دهد در باستان‌شناسی پـیش از تـاریخ از اهمیـت بـالایی برخوردار است و محوطه‌های بسیار مهمی چون سیلک کاشان، استقرارهای دشت قزوین و قم در آن قرار دارنـد. وی در رابطه با پیشینه فعالیت‌های باستان‌شناسی در طبس گفت: شهرستان طبس به دلیل جابجایی بین استان خراسان رضـوی، جنوبی و یزد و نیز دوری از مرکز استانی به هیچ وجه مورد مطالعات باستان‌شناسی قرار نگرفته است. بـه گفتـه عنـانی نخستین برنامه جامع در اوایل امسال در بخش دستگردان طبس به سرپرستی خود او انجام گرفت که نتایج آن بـزودی منتشر خواهد شد. این باستان‌شناس افزود: بررسی باستان‌شناسی دهستان منتظریه دومین برنامه جـامع باسـتان‌شناسـی است که در این شهرستان در حال انجام است. او ابراز امیدواری کرد که با همکاری پژوهشکده باستان‌شناسی و اداره کل میراث فرهنگی، صنایع دستی و گردشگری استان این برنامه در کل شهرستان ادامه یابد و نقشه باستان‌شناسی کلـی آن آماده شود.عنانی افزود: درحال حاضر دهستان منتظریه با حجم عظیمی از کویر لوت و نمک در حال پیمایش توسط هیات بررسی باستان‌شناسی برای کشف آثار باستانی است. او با ابراز امیدواری برای تبیین الگوهای استقراری دوره‌های مختلف در کل شهرستان طبس با بررسی این منطقه و سایر بخشها افزود: منطقه شرق ایران از نظر مطالعات باسـتان‌شناسـی در قیاس با مناطق غربی ایران بسیار ناشناخته است.این باستان‌شناس گفت: کندوکاو در تاریخ ما را به حضور در هـزاره اول قبل از میلاد مسیح یعنی زمان مادها و هخامنشیان و جنگ و گریز اسکندر با داریوش سوم آخـرین پادشـاه هخامنشـی رهنمون می‌کند.

تحقیقات باستان‌شناسی اخیر نشان داده است که انسان عصر حجر که تازه از کوه فرود آمده و در دشت سکونت گزیـده بود بر روی مسیر کمانی شکل اطراف کویر نمک استقرار یافته بود . به گزارش روابط عمومی پژوهشگاه میراث فرهنگی و گردشگری، بهرام عنانی سرپرست هیات بررسی و باستان‌شناسی دهستان مرکزی شهرستان طبس با اعـلام ایـن مطلـب افزود: فلات مرکزی ایران که بخشی از آن را کویر تشکیل می‌دهد در باستان‌شناسی پـیش از تـاریخ از اهمیـت بـالایی برخوردار است و محوطه‌های بسیار مهمی چون سیلک کاشان، استقرارهای دشت قزوین و قم در آن قرار دارند.

شکل (۴-۵) نمونه‌ای از مجموعه داده جمع‌آوری شده. قسمت بعد از خط افقی خلاصه بدست آمده را نشان می‌دهد.

یک پژوهش جدید نشان می‌دهد حدود نیمی از دانش‌آموزان آمریکایی در رده راهنمایی و دبیرستان که هرگز سیگار نکشیده‌اند، در معرض دود محیطی زیان‌بار سیگار یا به‌اصطلاح دود سیگار دست‌دوم قرار دارند. به گزارش سی‌ان‌ن این یافته‌ها از نظرخواهی ملی دخانیات در جوانان در سال ۲۰۱۳ که به‌وسیله مراکز پیشگیری و کنترل بیماری‌ها انجام شد، به دست آمده است که شامل داده‌های مربوط به بیش از ۱۸۰۰۰ دانش‌آموز راهنمایی و دبیرستان (کلاس ۶ تا ۱۲) در مدارس عمومی و خصوصی در سراسر آمریکا می‌شد. این دانش‌آموزان به پرسشنامه‌هایی درباره اینکه آیا دخانیات ازجمله سیگار، سیگار الکترونیکی، پیپ، قلیان یا آدامس نیکوتین‌دار مصرف کرده‌اند، پاسخ دادند. این پرسشنامه‌ها همچنین با پرسش درباره اینکه آیا در هفت روز گذشته کسی در اطراف آنان در خانه‌هایشان کسی سیگار کشیده است، به ارزیابی قرار گرفتن آنان در معرض دود سیگار محیطی پرداختند. در میان نوجوانانی که گزارش کرده بودند هرگز دخانیات مصرف نکرده‌اند، ۴۸ درصدشان قرار گرفتن در معرض دود سیگار محیطی را ذکر کردند. به‌علاوه، این بررسی نشان داد قرارگیری در معرض دود سیگار محیطی در میان نوجوانانی که هرگز سیگار نکشیده‌اند و در خانه یا ماشینشان قواعدی برای سیگار نکشیدن حاکم نیست، در مقایسه با نوجوانانی که در محیط‌های صد در صد عاری از دود سیگار زندگی می‌کنند، ۹ برابر بیشتر است. سرپرست این بررسی، برایان کینگ، معاون رئیس بخش تفسیر پژوهش‌ها در اداره سیگار کشیدن و سلامت در مراکز کنترل و پیشگیری بیماری‌های آمریکا گفت» :این یافته‌ها این نظر را که دود سیگار محیطی هنوز مشکلی در کشور است، تقویت می‌کند. نیمی از دانش‌آموزان راهنمایی و دبیرستان در معرض چیزی قرار دارند که ما می‌دانیم برای سلامتشان زیان‌بار است و کاملاً هم قابل‌پیشگیری است.

یک پژوهش جدید نشان می‌دهد حدود نیمی از دانش‌آموزان آمریکایی در رده راهنمایی و دبیرستان که هرگز سیگار نکشیده‌اند، در معرض دود محیطی زیان‌بار سیگار یا به‌اصطلاح دود سیگار دست‌دوم قرار دارند. به گزارش سی‌ان‌ن این یافته‌ها از نظرخواهی ملی دخانیات در جوانان در سال ۲۰۱۳ که به‌وسیله مراکز پیشگیری و کنترل بیماری‌ها انجام شد، به دست آمده است که شامل داده‌های مربوط به بیش از ۱۸۰۰۰ دانش‌آموز راهنمایی و دبیرستان (کلاس ۶ تا ۱۲) در مدارس عمومی و خصوصی در سراسر آمریکا می‌شد. سرپرست این بررسی، برایان کینگ، معاون رئیس بخش تفسیر پژوهش‌ها در اداره سیگار کشیدن و سلامت در مراکز کنترل و پیشگیری بیماری‌های آمریکا گفت» :این یافته‌ها این نظر را که دود سیگار محیطی هنوز مشکلی در کشور است،

شکل (۴-۶) نمونه‌ای دیگری از مجموعه داده جمع‌آوری شده. قسمت بعد از خط افقی خلاصه بدست آمده را نشان می‌دهد.

جدول (۴-۶) نمره‌دهی به ۲۰ نمونه خلاصه‌های استخراج شده را نشان می‌دهد. نتایج نشان داده شده در جدول فوق نشان می‌دهد که خلاصه‌های استخراج شده، خلاصه‌های مناسبی می‌باشند.

جدول (۴-۶) ارزیابی مدل پیشنهادی براساس سه معیار مناسب بودن، انسجام و جامعیت

	مناسب بودن	انسجام	جامعیت
۱	۸	۹	۷
۲	۹	۹	۸
۳	۹	۹	۸
۴	۸	۸	۸
۵	۷	۸	۸
۶	۹	۹	۸
۷	۹	۹	۸
۸	۷	۸	۸
۹	۷	۷	۶
۱۰	۸	۸	۷
۱۱	۹	۹	۸
۱۲	۶	۷	۷
۱۳	۸	۸	۸
۱۴	۹	۹	۸
۱۵	۷	۸	۷
۱۶	۸	۸	۷
۱۷	۷	۷	۶
۱۸	۹	۹	۹
۱۹	۷	۷	۶

	مناسب بودن	انسجام	جامعیت
۲۰	۹	۹	۹
میانگین	**۸**	**۸.۲۵**	**۷.۵۵**

فصل پنجم

نتیجه‌گیری و پیشنهادات

۵-۱- نتیجه‌گیری

خلاصه سازی عبارت است از فشرده سازی متن (متون) منبع و تولید یک نسخه کوتاه تر از آن به نحوی که محتوای اطلاعاتی آن حفظ شود. از کابردهای خلاصه سازی می توان به خلاصه سازی اتوماتیک اخبار و ارسال آن ها از طریق پست الکترونیکی یا پیامک اشاره نمود خلاصه‌سازی متون به دو صورت چکیده‌ای و استخراجی قابل انجام است. روش چکیده‌ای شبیه به روشی است که توسط انسان بکار برده می‌شود. این مورد به دلیل اینکه نیاز به درک معنی دارد، پیچیده می‌باشد از اینرو بیشتر روش‌های ارائه شده برای خلاصه‌سازی متون، از نوع استخراجی می‌باشد. در روش استخراجی، جملاتی از متن براساس متدهایی بعنوان چیکیده انتخاب می‌شوند.

در این پایان‌نامه یک مدل آماری به منظور استخراج جملات متون فارسی پیشنهاد شده است. مراحل این مدل به ترتیب عبارتند بودند از بخش‌بندی: یافتن ریشه کلمات، استخراج کلمات کلیدی، وزن‌دهی به کلمات، وزن‌دهی به جملات براساس کلمات کلیدی، استخراج جملاتی که وزن بیشتری دارند، انتخاب جملاتی که پیوستگی بیشتری دارند.

در مرحله آخر، که نوآوری مفاله می‌باشد، جملات با وزن بالا که پیوستگی بیشتری دارند، انتخاب می‌شوند. پیوستگی بین جملات براساس مدل خوشه‌بندی kmeans صورت گرفته است. این مدل خوشه‌بندی باعث قرار گرفتن جملات شبیه بهم در کنار یکدیگر شده و لذا کمک زیادی در انتخاب جملات استخراجی خواهد نمود.

همچنین در این پایان‌نامه یک دیکشنری به منظور استخراج مصدر برخی از افعال ایجاد گردید و مورد استفاده قرار گرفت.

در این پایان‌نامه برای تشریح مدل پیشنهادی، از یک نمونه مجموعه داده همشهری استفاده شد. استفاده از شباهت‌سنجی بین متون باعث انتخاب جمله بهتری به منظور خلاصه‌سازی گردید. که این بیانگر برتری مدل پیشنهادی در مقایسه روش RAKE می‌باشد.

بـــه منظـــور ارزیـــابی مـــدل پیشـــنهادی، یـــک مجموعـــه داده کـــه شـــامل ۲۰۰ نمونـــه مطلـــب از روزنامـــه‌هـــای همشـــهری و همچنـــین مجلـــه نســـیم آفتـــاب بـــود، جمـــع‌آوری گردیـــد. ســـپس مـــدل پیشـــنهادی بـــرروی آن اجـــرا شـــده و خلاصـــه‌هـــای اســـتخراج شـــده را براســـاس ســـه معیـــار مناســـب بودن، انسـجام و جامعیـت ارزیـابی گردیـد. ارزیـابی بـدین صـورت بـود کـه مجموعـه داده بـه همـراه حلاصـــه‌هـــا را بـــه عامـــل انســـانی داده و از آنهـــا خواســـته شـــد کـــه عـــددی بـــین ۰ تـــا ۱۰ بـــه آن خلاصـــه‌هـــا انتســـاب نماینـــد. جـــدول (۴-۶) نشـــان داده شـــده در فصـــل قبـــل بیـــان مـــی‌کنـــد کـــه خلاصه‌های بدست آمده کارایی مناسبی داشته‌اند.

۵-۲-پیشنهاد آینده

در ایـــن پایـــان‌نامـــه روشـــی پیشـــنهاد گردیـــد کـــه بـــا اســـتفاده از روش‌هـــای آمـــاری، خلاصـــه مـــتن را بدســـت مـــی‌آورد. مشـــکلی کـــه در روش‌هـــای آمـــاری وجـــود دارد، ایـــن اســـت کـــه آنهـــا از اطلاعـــات موجـــود در مـــتن و ســـایر اطلاعـــاتی کـــه مـــی‌توانـــد بعنـــوان مکاشـــفه یکـــار رود، اســـتفاده نمـــی‌کننـــد. برای مثال بـــرای حـــذف لغـــات بـــی‌معنـــی، تنهـــا متکـــی بـــر لغـــات پرتکـــرار هســـتند و امیـــدوار هســـتند حـــذف لغـــات پرتکـــرار بتوانـــد آنهـــا را نیـــز حـــذف کننـــد. در صـــورتیکه در مثـــالی کـــه در بخـــش قبلـــی تشریح گردید، لغات بـــی‌معنـــی وجـــود داشـــتند کـــه از ایـــن طریـــق حـــذف نشـــدند. لـــذا پیشـــنهاد مـــی-شود که بتوان از روش-هـــای مکاشـــفه‌ای نیـــز بهـــره بـــرد. بـــرای مثـــال مـــی‌تـــوان لیســـتی از لغـــات بـــی-معنی نشکیل داده و متون را براساس آن، پردازش نمود.

مراجع

[1] H. Luhn, “The automatic creation of literature abstraction”, IBM Journal of research and development, 2-159-165, 1958

[2] V. Gupta, G. Lehal, “A Survey of Text Summarization Extractive Techniques”, Journl of emerging technologies in web intelligence, VOL. 2, NO. 3, 2010.

[3] V. Gupta, G.Sl Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 2009

[4] R. Mihalcea, “Graph-based ranking algorithms for sentence extraction, applied to text summarization,” in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Barcelona, Spain, 2004.

[5] Hassan Shakeri, Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Fatemeh Ghadamyari, “A New Graph-Based Algorithm for Persian Text Summarization”, Computer Science and Convergence, Lecture Notes in Electrical Engineering 114, DOI: 10.1007/978-94-007-2792-2_3, ! Springer Science+Business Media B.V. 2012.

[6] Nima Mazdak , “A Persian text summarizer”, Master Thesis, Department of Linguistics Stockholm University January 2004.

[7] Martin Hassel, Nima Mazdak, “FarsiSum - A Persian text summarizer”, Royal Institute of Technology 100 44 Stockholm, Sweden

[8] S Gholamrezazadeh, M Amini, B Gholamzadeh, “A Comprehensive Survey on Text Summarization Systems”, 978-1 4244- 4946- 0/09/$25.00 , IEEE, 2009.

[9] M. Shamsfard, “Challenges and Open Problems in Persian Text processing”, 2011.

[10] M. Iranpour, B. Minaei, “Verb Detection in Persian Corpus”, International Journal of Digital Content Technology and its Application, 3[1], pp. 58-65, 2009.

[11] Makbule Gulcin Ozsoy, “Text Summarization using Latent Semantic Analysis”, Master Of Science Degree in Computer Engineering, Middle East Technical University, 2011.

[12] Gong, Yihong, and Xin Liu. "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'01). ACM, 19- 25, 2001.

[13] R. Mohammed Badry, A. Sharaf Eldin, D. Saad Elzanfally, “Text Summarization within the Latent Semantic Analysis Framework: Comparative Study”, International Journal of Computer Applications (0975 – 8887) Volume 81 – No.11, November 2013.

[14] M. Gulcin Ozsoy, I. Cicekli, F. Nur Alpaslan, “Text Summarization of Turkish Texts using Latent Semantic Analysis”, 23rd International Conference on Computational Linguistics, pages 869–876, , August 2010.

[15] J. Steinberger, K. Jezek, “Using Latent Semantic Analysis in Text Summarization and Summary Evaluation”, Proceedings of ISIM '04, pages 93-100, 2004.

[16] A. Agrawal , U. Gupta, “Extraction based approach for text summarization using k-means clustering”, International Journal of Scientific and Research Publications, Volume 4, Issue 11, 1 ISSN 2250-3153, 2014.

[17] Anjali R. Deshpande, Lobo L. M. R. J, “Text Summarization using Clustering Technique", International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8, 2013.

[18] N.I.Meghana, M.S.Bewoor, S.H.Patil, “Text Summarization using Expectation Maximization Clustering Algorithm”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 4, pp.168-171, 2012.

[19] R. Barzilay, and M. Elhadad, “Using Lexical Chains for Text Summarization”, the MIT Press, pp. 111-121, 1999.

[20] M. Pourvali and S. Abadeh Mohammad, "Automated text summarization base on lexical chain and graph using of word net and Wikipedia knowledge base", IJCSI International Journal of Computer Science Issues, No. 3, vol. 9, 2012.

[21] C. Sofia Oliveira Santos, "ALEXIA - Acquisition of Lexical Chains for Text Summarization", A Thesis submitted to the University of Beira Interior to require the Degree of Master of Computer Science Engineering, 2006.

[22] Brants, T. "TnT - a statistical part-of-speech tagger", In Proceedings of the 6th Applied NLP Conference, ANLP, 2000.

[23] Cleuziou, G., Martin, L. and Vrain, C. "PoBOC: an Overlapping Clustering Algorithm, Application to Rule-Based Classification and Textual Data", In Proceedings of the 16th biennial European Conference on Artificial Intelligence (ECAI'04), pp. 440-444, 2004.

[24] F. Kyoomarsi, H. khosravi, E. Eslami, M. Davoudi, "Extraction-based Text Summarisation using Fuzzy Analysis", Iranian Journal of Fuzzy Systems Vol. 7, No. 3, pp. 15-32, 2010.

[25] Ms.Pallavi D.Patil, Prof.N.J.Kulkarni, "Text Summarization Using Fuzzy Logic", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 3, ISSN: 2278-2311, p 42-45, 2014.

[26] A. Kiani, M. R. Akbarzadeh, "Automatic Text Summarization Using: Hybrid Fuzzy GA-GP", IEEE International Conference on Fuzzy Systems Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006.

[27] L. Suanmali1, N. Salim, M. Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

[28] M. Salem Binwahlan, N. Salim, L. Suanmali, "Fuzzy Swarm Based Text Summarization", Journal of Computer Science 5 [5]:338-346, ISSN 1549-3636, 2009.

[29] S. Santhana Megala, Dr. A. Kavitha, Dr. A. Marimuthu, "Enriching Text Summarization using Fuzzy Logic", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 [1] , p. 863-867, 2014.

[30] V. Gupta, G. S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, 2009.

[31] M. Pourvali & M. Saniee Abadeh, “A new graph based text segmentation using Wikipedia for automatic text summarization”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 1, 2012.

[32] N. Alami, M. Meknassi, N. Rais, “Automatic Texts Summarization: Current State of the Art”, Journal of Asian Scientific Research, 5[1]: 1-15, 2015.

[33] H. Edmundson, “New Methods in Automatic Extracting”, Journal of the Association for Computing Machinery, 16 [2], 264-285,1969.

Summarization of Persian texts using artificial intelligence techniques

Abstract

Objective One of the methods proposed for encapsulating texts is the statistical method. In such methods, all of the information is drawn from the text itself and there is no need to possess any basic knowledge. The purpose of this thesis is to propose a statistical method to encapsulate Persian texts.

Methods: Statistical method to encapsulate Persian texts, which combines a statistical algorithm and a clustering model, initially extracts the key words and then allocates a specific rhythm to each single word and finally makes the sentences rhythmic according to them. What is next, using a clustering model, the proposed model categorizes similar sentences into a cluster and ultimately extracts the best sentences according to their rhythm and similarity. Using the similarity of the sentences and their rhythm causes more significant sentences, which are also related to one another, to be extracted.

Results: In this study in order to evaluate the proposed model, a set of data included 200 sample texts from Hamshahri newspaper and also Nasim Aftab magazine was collected. The proposed model was applied on collected data set therefore extracted abstracts evaluated based on three criteria which were being fit, coherency and comprehensiveness. In the evaluation process the data set was given to the human factors and they have being asked to assign a number between 0 and 10 to the abstractions. According to the table 4-6 in context, the abstractions have got appropriate efficiency.

Conclusion: In this thesis in order to extract sentences of Persian texts, a statistical model has been proposed. This model includes some stages specially weighting the words, weighting sentences based on key words and selecting sentences based on more continuity. In the last one which is our novelty the sentences with more continuity has been selected. To do so, K-means model has been chosen. This model puts similar sentences next to each other so would be a great help to select the extracted sentences. Also in order to extract infinitive from some verbs a dictionary has been made and applied. Using similarity between contexts led to choose better sentence for summarization that expresses the excellence of our model in comparison with RAKE method.

Keywords: Encapsulation, Persian texts, statistical model, clustering, abstraction, summarization, similarity, K-means, RAKE

Faculty of Virtual Educations
Computer Engineering and Information Technology Department
M.S. Thesis in Computer Engineering – Artificial Intelligence

Summarization of Persian texts usin artificial intelligence techniques

By: Aazam Mohammadzadeh

Supervisor:
Dr. Morteza Zahedi

February 2017