مروری بر روش‌های خلاصه سازی خودکار متون

لیلا طالب علی[۱] ، نوشین ریاحی[۲]
۱-گروه مهندسی کامپیوتر، دانشگاه الزهرا (س) ، تهران
۲-گروه مهندسی کامپیوتر، دانشگاه الزهرا (س) ، تهران

چکیده

امروزه پردازش زبان طبیعی در زمینه های گوناگون نظیر خلاصه‌سازهای خودکار و مترجم‌های ماشینی[۱]، توجه زیادی را به خود جلب نموده‌اند. در زبان فارسی هم مانند سایر زبانهای دیگر دنیا تلاش‌هایی در زمینه ساخت ابزارهای خلاصه سازی صورت گرفته است. تمرکز محققان بر ارایه روشهایی متمرکز است که بتواند خلاصه هایی پر محتوا، سلیس و روان نسبت به روشهای خلاصه سازی پیشین ارایه دهند . خلاصه سازی یک مهارت نگارشی به شمار می رود، که هدف از ایجاد سیستم خلاصه ساز اتوماتیک تقلید کلیه مراحلی است که توسط عامل انسانی انجام می شود، بدین صورت که متن به طور کامل خوانده و فهمیده شود و با تشخیص و تفکیک قسمت های مهم و غیرمهم متن، نسخه خلاصه شده متن اصلی تولید گردد.

هدف از خلاصه سازی خودکار سند، تولید یک نسخه مختصرتر از سند اصلی توسط یک برنامه رایانه ای به نحوی که ویژگی ها و نکات اصلی سند اولیه حفظ شود.بنابر تعریف ارائه شده در استاندارد ISO ۲۱۵، خلاصه "یک بازگویی مختصر از سند" می باشد. روش های تولید خلاصه را با توجه به انواع دیدگاه های مختلف خلاصه سازی می توان به چندین دسته تقسیم بندی نمود، به عنوان مثال روش تولید خلاصه تک سندی و چند سندی، تک زبانه و چند زبانه، مبتنی بر تعامل با کاربر و غیر مبتنی بر تعامل با کاربر و... لیکن با توجه به اهمیت فاکتور خروجی در تولید خلاصه در این مقاله به بررسی روش های تولید خلاصه گزینشی(استخراجی) و چکیده ای(مفهومی) پرداخته می شود.

واژگان کلیدی:پردازش زبان طبیعی، خلاصه‌سازهای ماشینی، روابط کلامی ، تشابه معنایی، شبکه واژگان

۱. Machine Translator

ISBN: 978-600-7461-60-0

مقدمه

خلاصه سازی خودکار متن به عنوان هسته مرکزی طیف گسترده ای از ابزارهای پردازش گر متن مانند خلاصه سازهای ماشینی، سیستم های تصمیم یار، سیستم های پاسخ گو، موتورهای جستجو، از سال ها پیش مطرح شده و همواره به عنوان یک موضوع مهم مورد بررسی و تحقیق قرار گرفته است. (Mekeown et al,۲۰۰۳) آغاز فعالیت سیتسم های خلاصه سازی خودکار متن مربوط به سال ۱۹۵۰ می شود. به دلیل کمبود کامپیوترهای قدرتمندو مشکلات موجود برای پردازش زبانهای طبیعی (NLP)، کارهای اولیه بر روی مطالعه ظواهر متن مانند موقعیت جمله و عبارات اشاره متمرکز شده بود. سال ۱۹۷۰ تا ۱۹۸۰ هوش مصنوعی بکار آمد(Buyukkokten et al,۲۰۰۳)،(شهاب ،۱۳۸۱)و (Mazdak, N., Hassel,۲۰۰۴) . ایده ی AI، استخراج نمایش های دانش مانند فریمها یا الگوها برای شناسایی موجودیتهای مفهومی از متن و استخراج روابط بین موجودیت ها با مکانیزمهای استنتاج بود. از اوایل ۱۹۹۰ تا به حال هم روشهای بازیابی اطلاعات(IR) بکار گرفته شده است(Dalianis,۲۰۰۰) ،(کریمی و شمس فرد،۱۳۸۵)و(مشکی و آنالویی ،۱۳۸۸). بیشتر این روش ها بر روی سطح سمبولیک متمرکز بوده و وارد حوزه های معنایی نمی شوند.

Kupiec اولین الگوریتم را در این زمینه پیشنهاد داد(کریمی و شمس فر،۱۳۸۵). در این روش بر اساس مقادیر ویژگیهای یک جمله، احتمال حضور آن در خلاصه تخمین زده می شود. او عمل خلاصه سازی را به صورت یک مسئله دسته بندی، در نظر گرفت و دسته بندی کننده های بیزین را برای تعیین جملاتی که باید در خلاصه وارد شوند بکار برد.Chuangو Yang چندین الگوریتم مانند درخت تصمیم و دسته بندی کننده را برای استخراج قطعات جمله پیشنهاد داد.(اخوان و همکاران،۱۳۸۷) در سال ۱۹۹۷، Barzialy روشی برای تولید خلاصه با پیداکردن زنجیره های لغوی معرفی کرد که به توزیع کلمه و اتصالات لغوی بین آنها برای تقریب زدن محتوا و ارائه یک نمایش از ساختار لغوی بهم پیوسته متن اتکا می کرد.(بهره پور و همکاران،۱۳۸۷)از روش های آماری هم در خلاصه سازی متن بسیار استفاده شده است که از جمله آنها می توان به روش های مبتنی بر مدل موضوع و و روش های مبتنی بر گراف اشاره نمود.

از اوایل سال ۲۰۰۰ به تدریج بحث خلاصه سازی مبتنی بر کاربر و یاخلاصه سازی شخصی سازی شده مطرح شد و تا به امروز نیز مقالات بسیاری در این زمینه منتشر شده است. ایده اصلی خلاصه سازی شخصی سازی شده و یا مبتنی بر کاربر این است که کاربران مختلف با توجه به دانش و پس زمینه اطلاعاتی که دارند، دیدگاههای متفاوتی روی اسناد یکسان دارند.

علی رغم اینکه در زمینه خلاصه سازی متن کارهای زیادی از سالهای دور انجام شده است ولی همچنان شاهد مقالات متعددی هستیم که همه ساله در این زمینه منتشر می شوند و سعی در بهبود دقت روش های پیشین دارند .(R. Ferreira,۲۰۱۴)

جنبه ها و مدل های مختلف خلاصه سازی متن

روش های خلاصه سازی متن معمولا از دیدگاه های مختلفی تقسیم بندی می شوند. این دسته بندی را در چند قالب اصلی می توان تشریح کرد(ستوده و همکاران،۱۳۸۹).

ISBN: 978-600-7461-60-0

- بر اساس منبع ورودی: براساس نوع منبع ورودی خلاصه سازها به دسته های کلی تک سندی و چند سندی تقسیم بندی می شوند. همچنین بسته به این که نوع متن، اخبار، علمی، گزارش، عمومی و... باشد، روش برخورد هم متفاوت می باشد. به عنوان مثال، در خلاصه سازهای اخبار یکی از ساده ترین روش هایی که استفاده می شود این است که تیتر متون به همراه جمله های اول پاراگراف به عنوان خلاصه استفاده می شود و این در حالی است که این روش برای متون علمی جوابگو نمی باشد.همچنین بر اساس زبان متن هم دو دسته خلاصه ساز تک زبانه و چند زبانه داریم. روش خلاصه سازی بسته به اینکه متون تک زبانه و یا چند زبانه باشد، متفاوت می باشد. چرا که زبان های مختلف با یکدیگر تفاوت داشته و بسیاری از ویژگی ها که در یک زبان صادق است در زبان دیگر ممکن است صادق نباشد.
- بر اساس هدف:در این تقسیم بندی باید مشخص شود که متن خلاصه به چه منظوری تولید می شود. هدف خلاصه سازی می تواندهشدار، پیش نمایش، آگاهی، زندگی نامه و... باشد و این موضوع در تعیین روش خلاصه سازی تاثیرگذار می باشد.همچنین خلاصه ها می توانند عمومی یا مبتنی بر پرس و جو باشند و یا اینکه مبتنی بر کاربر و یا کلی باشند.
- بر اساس خروجی: بر این اساس خلاصه ها به دو دسته عمده خلاصه سازهای چکیده ای و استخراجی تقسیم بندی می شوند.

دسته‌بندی الگوریتم‌های خلاصه سازی

به طور کلی چهار دسته روش برای روند خلاصه سازی وجود دارد که عبارتند از:

۱. روش های آماری سطحی: خلاصه سازهای سطحی روش های آماری، مکانی و تحلیل های مکانی- آماری انجام می دهند. این روشها از یک سری خصوصیات ساده متن برای نمایش اطلاعات محتوا استفاده می نماید. از جمله این خصوصیات می توان از خصوصیات کمی متن، فرکانس عبارت، خصوصیات محلی مانند: مکان جمله در متن، مکان محلی پاراگراف، خصوصیات زمینه‌ای مانند: وجود کلمات عنوان، کلمات و عبارات خاص نام برد. (عبارات خاص را می‌توان خلاصه‌هایی که در خود متن ورودی موجود است دانست.)

۲. روش های درک متن: تفاوت ایده اصلی این روش با ایده روش سطحی آن است که یک نمایش داخلی از متن ورودی با مدل کردن متن و ارتباطات بین آن‌ها ایجاد می‌کند. ارتباط بین موجودیت‌ها شامل شباهت، هم‌معنی، ارتباط نحوی و ارتباط معنایی مانند: تضاد، سازگاری متن، ارتباط لغوی و هم‌مکانی کلمات می‌باشد.روش‌های بر اساس زبان شناسی با تجزیه کلمه و مشخص کردن نحو کلمه آغاز می‌شود و از پایگاه‌داده لغوی استفاده می‌نماید تا ارتباط بین جملات را استخراج نماید و جملات مهم بر اساس معنایشان انتخاب می‌شوند.

۳. روش های مبتنی بر یادگیری نظارتی: یادگیری نظارتی گونه‌ای از یادگیری است که در آن به عامل یادگیرنده تعدادی جنبه و مقدار متناظر برای خصوصیتی که بایستی یادگرفته شود، داده شده و از او خواسته می‌شود تا در موارد مشابه بر اساس آنچه در ابتدا به او داده شده است عمل نماید. در این روش، با مجموعه‌های از سندهای آموزش و خلاصه‌های استخراجی آن‌ها، فرآیند خلاصه‌سازی به عنوان یک مسئله طبقه‌بندی مدل می‌شود: جملات به "جملات خلاصه" و "جملات غیر خلاصه" براساس مشخصه‌هایی که دارا هستند، طبقه‌بندی می‌شوند.

ISBN: 978-600-7461-60-0

۴. روش های ساختار کلامی: روش‌هایی از ساختار کلامی برای تشخیص بخش‌های مهم متن استفاده می‌نمایند، این سطح پردازش متن ساختار عمومی متن و ارتباط آن‌ها را مدل می‌نماید. این ساختار شامل هدف سند، موضوعاتی که در متن آمده است و همین طور شامل ساختار کلامی متن می‌باشد. این مسئله که چگونه بتوان یک ساختار را به صورت اتوماتیک استخراج نمود و چگونگی استفاده از آن در تبدیل نحوی اطلاعات مسئله مهمی است.

روش های آماری سطحی

مزیت اصلی این گونه روش‌ها سادگی آن‌ها و همچنین امکان اعمال بر روی هر نوع متن ورودی می‌باشد. همچنین از آنجایی که هیچ تلاشی در آن برای فهم متن صورت نمی‌پذیرد، کارایی و رواج بالایی دارند.از آنجایی که در این دسته از روش‌ها ارتباط بین کلمات در نظر گرفته نمی‌شود، عدم انسجام و پیوستگی در خلاصه‌های تولیدی به وضوح ملاحظه می‌شود، که می‌توان این مشکل را عیب مهم این دسته از روش‌ها دانست.

- ویژگی‌های ادمونسونی[۱]:احتمالاً تاریخچه اولین گروه از روش‌های خلاصه‌سازی به زمان ادمونسون باز می گردد. زیرا بسیاری از روش‌هایی که پس از آن تولید شدند، تلاش نمودند جنبه‌های گوناگون این سیستم و نحوه تأثیر آن‌ها را به گونه‌ای مختلف تغییر دهند. برخی از این ویژگی‌ها جهت امتیازدهی به جملات مورد استفاده قرار می گیرد، که عبارتند از(بازقندی و تدین تبریزی،۱۳۹۱):
- فراوانی واژه‌ها[۲]: مهم‌ترین و پرکاربردترین جنبه در میان ویژگی‌های ادمونسونی برای امتیازدهی، میزان تکرار کلمات گوناگون موجود در متن یا متون ورودی است. بکار گیری این جنبه بر پایه فرض واژه‌های موضوعی[۳] است. این فرض بیان می‌دارد که واژه‌هایی که به نسبت در یک متن بیشتر بکار رفته‌اند، دارای بار معنایی بیشتری هستند. از آنجایی که در یک متن با موضوعی خاص، واژه‌های زیادی مرتبط با آن موضوع بکار می‌روند، این فرض منطقی به نظر می‌رسد. البته بکارگیری این ایده به صورت خام کارایی چندانی نخواهد داشت؛ زیرا واژه-هایی مانند حروف اضافه، حروف ربطی و ضمایر اشاره در هر متنی به صورت پرتکرار وجود دارند، اما بار معنایی مهمی ندارند. یک ایده برای جلوگیری از اثر مخرب کلمات پرکاربرد و کم اهمیت این است که میزان اهمیت یک کلمه را در یک متن، علاوه بر تعداد تکرارهای آن، به صورت معکوس به تعداد اسنادی که در یک پیکره دارای این واژه می‌باشند، وابسته نمود. به این نحوه امتیاز دهی به هر کلمه، TF/IDF[۴] گفته می‌شود، که معادل فرکانس کلمه – معکوس فرکانس سند می‌باشد، و برای هر واژه به صورت زیر محاسبه می گردد:

$$tf.idf(term) = tf(term).\log(\frac{NumDoc}{NumDoc(term)}) \qquad (۱)$$

که در آن NumDoc تعداد کل اسناد موجود در پیکره، NumDoc(term)، تعداد اسنادی که دارای واژه term هستند و tf(term)، تعداد تکرارهای واژه term در سند یا اسناد ورودی می‌باشد. حال در صورتی که

۱. Edmonsonian Features
۲. Term Frequency
۳. Thematic Term Assumption
۴. Term Frequency- Inverse Document Frequency

ISBN: 978-600-7461-60-0

متن ورودی تنها حاوی یک سند به جای چندین سند باشد، این معیار به صورت فرکانس کلمه – معکوس فرکانس جمله یا TF_ISF(w,s) بیان می‌گردد. فرکانس جمله تعداد جملات سند است که حاوی کلمات هستند. این مشخصه برای تمام کلمات هر جمله محاسبه می‌شود(Barzilay et al,۲۰۰۲).

- مکان واحدهای متنی:در(Chuang and Yang,۲۰۰۰) نشان داده شده است که جملات مهم در ۸۵ درصد موارد در ابتدای پاراگراف و در ۷ درصد موارد در انتهای آن قرار دارند. همچنین یک سیاست بهینه برای مکان یابی جملات مهم ارائه شده است. بر اساس مطالعه‌ای در ۱۳۰۰۰ خبر از مجموعه Ziff-Davis در رابطه با محصولات کامپیوتری، عنوان هر مقاله و پس از آن جمله اول از هر پاراگراف دارای بیشترین بار معنایی می‌باشند. از سوی دیگر در خبرهای مجله وال استریت، پس از عنوان جملات موجود در پاراگراف اول دارای بار معنایی بیشتری می‌باشند. نتایج این تحقیق نشان می‌دهد که مکان جملات مهم بسته به منبع آن جمله، متفاوت است، که این خود یک چالش برای این مبحث می‌باشد.
- عبارات توضیحی: این عبارت‌ها وابسته به یک دامنه خاص می‌باشند و بر اساس تناوب تکرار از یک پیکره استخراج می‌شوند. این عبارت‌ها را می‌توان به دو دسته عبارات تنبیهی و تشویقی تقسیم نمود. عبارت تشویقی احتمال آمدن واحدهای متنی پر ارزش را در خلاصه افزایش می‌دهد و عبارت‌های تنبیهی این احتمال را کاهش می‌دهد. به عنوان مثال عبارت"نتیجه می گیریم که" یک عبارت تشویقی و عبارت"وی گفت" یک عبارت تنبیهی می‌باشد.
- امضاء سرفصل: این روش نیز استراتژی دیگر برای یافتن جملات پر اهمیت موجود در متن و امتیازدهی به آن‌هاست. در این روش آماری بدون استفاده از پایگاه‌های دانش و یا پارسرها جملات پر اهمیت استخراج می گردد. ایده اصلی این روش این گونه است که واژگانی که از نظر معنایی با یکدیگر مشابهت دارد در محیط‌های یکسانی ظاهر می شوند. به عبارت دیگر این روش بیان می‌دارد که اگریک ایده یا اتفاق خاص مهم و کلیدی در یک متن وجود داشته باشد، مجموعه‌هایی از واژگان با ساختاری قابل پیش‌بینی در متن وجود خواهد داشت که یک موضوع را به نحوی یکتا شرح می‌دهند. امضاء سرفصل با معرفی یک زوج متشکل از یک مبحث(موضوع) و مجموعه‌ای از واژگان مرتبط با آن، این ساختارها را به صورتی رسمی بیان می کند. برای یافتن این امضاءهای سرفصل معمولاً از روش‌های باهم‌آیی استفاده می‌شود. در(هنر پیشه ،۱۳۸۶) مدلی بهبود یافته از امضاءهای سرفصل ارائه شده است. در این مدل علاوه بر پیدا نمودن امضاءهای سرفصل، ارتباط میان واژه‌های موجود در امضاءهای سرفصل نیز استخراج می‌شود. این ارتباطات یا به صورت ارتباطات نحوی و یا به صورت باهم‌آیی می‌باشد .
- روش باهم‌آیی :باهم‌آیی کلمات بر اساس این ایده عمل می کند که کلماتی که در یک زمینه مشترک با هم مشاهده می-شوند، با یکدیگر در ارتباط باشند. چند- وزنی ترتیبی از چند کلمه است. چند- وزنی زمانی در یک سند وجود دارد که این کلمات به یک ترتیب پشت سرهم در سند ظاهر شود. در اکثر مراجع این توالی کلمات را با در نظر گرفتن ریشه کلمات بدون در نظر گرفتن کلمه‌های توقف محاسبه می‌نمایند.

ISBN: 978-600-7461-60-0

روش های درک متن

این گونه روش‌ها، روش‌های پرهزینه‌ای هستند زیرا علاوه بر نیاز به منبع دانش، نیازمند تفسیر متن پیچیده می‌باشد. ولی عملکرد بهتری نسبت به روش اول دارند. از آنجایی که ایده اصلی این نوع خلاصه‌سازی‌ها بر پایه درک متن می‌باشد نیاز به خط مشی‌های بر پایه دانش و مقایسه بخش‌هایی از متن با پایگاه داده موجود دارد. تحلیل نحوی و معنایی دو بخش مهم این نوع خلاصه‌سازی می‌باشد.

یک نمونه از اطلاعاتی که در بسیاری از کاربردهای متن کاوی اهمیت فراوان دارد، وابستگی بین اجزای متن است. دو دسته مهم از وابستگی‌های متنی عبارتند از: همبستگی[۱] و ارتباط معنایی[۲]. پدیده همبستگی، معادل با این واقعیت است که بعضی از عناصر متنی مانند واژه‌ها تمایل دارند که در کنار هم ظاهر شوند، در حالی که پدیده ارتباط معنایی بر این حقیقت اشاره دارد که یک ارتباط هوشمندانه بین جملات متن وجود دارد. ارتباط معنایی یک ارتباط سطح بالاتر از همبستگی است. هر دو پدیده همبستگی و ارتباط معنایی پدیده‌هایی ذهنی[۳] هستند، به همین دلیل شناسایی آن‌ها با چالش‌هایی مواجه است. شناسایی همبستگی ساده تر از ارتباط معنایی است، زیرا با بررسی فراوانی واژه‌ها و وقوع هم‌زمان آن‌ها قابل شناسایی است(ارژنگ ،۱۳۸۳).

- روش مبتنی بر گراف(R. Ferreira,۲۰۱۴):

در این گونه روش‌ها بعد از گام‌های پیش پردازش برای مشخص کردن عناوین (اسامی مهم و...) در متن، سند به شکل گرافی غیر جهت‌دار که جملات گره‌های تشکیل دهنده‌ی آن هستند ارائه می‌شود و تئوری گراف می‌تواند به راحتی برای تجسم شباهت بین سندی و درون سندی به کار رود (Hovy and Marcu,۱۹۹۸). یکی از الگوریتمهایی که برای پیدا کردن جملات با ارزش مورد استفاده قرار می گیرد الگوریتم TextRank می باشد. الگوریتم Page Rank اولین بار در سال ۱۹۹۲ توسط Larry Page و Sergey Brin در دانشگاه استنفورد ارائه شده است، این الگوریتم یک روش مستقل از پرس و جو می باشد. الگوریتم رتبه بندی صفحات مهم ترین مهمترین عامل در اندازه گیری اعتبار یک سایت می باشد. امروزه در دنیای اقتصادی وب سایت هایی که رتبه PR بالاتری دارند با مبالغ هنگفتی معامله می شوند و همچنین برای موتور های جستجو عامل مهم برای دادن پاسخ به درخواست کاربر ها همین معیار رتبه بندی است. این روش یک بار به هر سند وب امتیاز اختصاص داده و از این امتیاز، با در نظر گرفتن معیاری با توجه به پرس و جوی کاربر جهت رتبه بندی اسناد استفاده می کند. این الگوریتم رتبه هر صفحه را با اختصاص وزن به پیوندی که به آن صفحه داده شده است به دست می آورد که مقدار این وزن به کیفیت صفحه هایی که پیوند در آن قرار گرفته، بستگی دارد. در این صورت پیوندهای صفحات مهمتر وزن بیشتری میگیرند. جهت مشخص کردن کیفیت صفحه های رجوع کننده، در Page Rank از رتبه آن صفحه که به صورت بازگشتی تعیین و مقدار اولیه آن اختیاری است، استفاده می شود.

اگر n سند در دسترس باشد، مقدار اولیه رتبه سند را می توان برابر ۱/n در نظر گرفت.

$$PR(A) = (۱ - d) + d.\sum \frac{PR(Ti)}{C(Ti)} \qquad (۲)$$

۱. Cohesion
۲. Coherence
۳. Subjective

ISBN: 978-600-7461-60-0

فرمول فوق جهت محاسبه رتبه یک صفحه مانند A استفاده می شود. در این فرمول،$PR(T_i)$ رتبه صفحه T_i ، (T_iصفحاتی میباشند که به صفحه A اتصال دارند) است. $C(T_i)$، تعداد لینکهای خروجی صفحه T_i و dضریب اثر میباشد)که دارای مقداری بین ۰ و ۱ است(فتوحی،۱۳۹۲).الگوریتم TextRank در واقع کاربرد الگوریتم PageRank در زمینه پردازش زبان طبیعی است. به وسیله این الگوریتم می توان کلمات مهم یا همان کلمات کلیدی را استخراج نمود. همچنین می توان جملات مهم را بر اساس میزان ارتباط با سایر جملات یا بقیه جملات با آن تعیین نمود.

- روش زنجیره لغوی :

زنجیره‌های لغوی به عنوان خوشه‌هایی از کلمات از نظر معنایی مرتبط با هم تعریف شده‌اند. به عنوان مثال "خانه، سرا، اتاق" یک زنجیره است، در حالی که خانه و سرا هم معنی هستند ولی اتاق بخشی از خانه است.

یک زنجیره لغوی[۱] مجموعه‌ای از لغات مرتبط با یکدیگر است که واحدی موضوعی از متن را پوشش می دهد. (ریاحی و عزالی،۱۳۹۳) به عبارتی دیگر اگر زیر بخش‌های موضوعی در ارتباط با یک مطلب خاص در متنی وجود داشته باشد، یک زنجیره لغوی مرتبط، آن موضوع را متمایز می‌سازد. نخستین روش محاسباتی استخراج زنجیره‌های لغوی، برای استفاده در سیستم‌های خلاصه‌ساز ارائه گردید(حافظی و همکاران،۱۳۸۵). این روش با بکار گیری مجموعه‌ای از ارتباطات میان لغات، مانند تکرار، ترادف[۲]، زیر مجموعه[۳] (به عنوان مثال زیر مجموعه بودن پراید در گروه ماشین‌ها)، تضاد[۴] و زیر بخشی (به عنوان مثال زیر بخش بودن برگ برای درخت)، زنجیره-های لغوی را تولید می‌نماید.نحوه ساخت زنجیره لغوی به این صورت است که ابتدا زیر مجموعه‌ای از لغات پرتکرار موجود در متن انتخاب می‌شوند. به ازای هریک از لغات، اگر اتباطی با یکی از زنجیره‌ها موجود بود، آن لغت با ارتباطش به زنجیره متصل می‌شود؛ در غیر این صورت برای آن لغت یک زنجیره مجزا تشکیل داده می‌شود. اگر یک لغت دارای چندین معنا باشد برای هریک از معناهای آن گروه زنجیره‌ای متفاوتی تشکیل می‌دهد. از میان گروه زنجیره‌های متفاوت گروه زنجیره‌ای انتخاب می‌شود که دارای ارتباطات بیشتری میان لغت‌هایش باشد. در پایان بر اساس گروه زنجیره انتخابی به هر یک از جملات امتیازی نسبت داده می‌شود. نحوه امتیازدهی به جملات بر اساس لغات موجود در آن‌هاست؛ به این ترتیب که به هر لغت به اندازه تعداد ارتباط‌هایش با دیگر لغات در زنجیره امتیاز داده می‌شود.

مشکل مهم محاسبه زنجیره‌ها با استفاده از شبکه لغات، تعداد زیاد معنای کلمات در زنجیره لغات است که باعث رشد نمایی تعداد گروه زنجیره‌ها می‌گردد. در صورتی که تعداد گروه زنجیره‌ها از یک حد آستانه بزرگ‌تر شود، در هر مرحله گروه زنجیره‌هایی انتخاب می‌شوند که امتیاز بالاتری داشته باشد. دو الگوریتم مختلف برای ایجاد زنجیره‌ها وجود دارد: رویه غیرمبهم حریصانه که زنجیره یک کلمه فقط به وسیله زنجیره‌های کلمات قبل

۱. Lexical Chain
۲. Synonimy
۳. Hypernymy
۴. Antonymy

ISBN: 978-600-7461-60-0

ازآن در متن مشخص می‌شود به این صورت که بر اساس روابط تعیین شده اگر زنجیره مرتبط با کلمه در زنجیره-های از قبل موجود یافت شود، آن را در همان زنجیره درج می‌کند و در غیر این صورت برای آن زنجیره جدید ساخته می‌گردد. در مقابل الگوریتم غیرحریصانه تا هنگامی که همه کلمات در متن پردازش شود منتظر می‌ماند سپس با توجه به تمام لغات متن زنجیره هر کدام را یافته یا ایجاد می‌کند.

به طور کلی اکثر خلاصه‌سازهای بر اساس زنجیره لغوی روش یکسانی را در مرحله پردازش متن به صورت ذیل دنبال می‌کنند:

الف- تولید زنجیره‌های لغوی

ب- امتیاز دادن به زنجیره‌ها

ج- یافتن قویترین این زنجیره‌ها جهت ایجاد خلاصه متن

روش استفاده از رویدادها [۱]:

اخیراً رویدادها به عنوان یکی از راه‌های مدل سازی متن مورد توجه قرار گرفته‌اند. رویداد در هر حوزه تعریف مخصوص به خود را دارد. زبان شناسی به تعریف معنایی از رویداد و تعریف معنایی ساختار افعال می‌پردازد. بازیابی اطلاعات در بخش یافتن و ردیابی موضوع به رویدادها به عنوان زیر موضوعاتی نگاه می‌کند که می‌تواند از آن‌ها برای دسته بندی استفاده نماید. در این تعریف رویدادها می‌توانند رخدادها و دلایل آن‌ها، نتایج و هر گونه تأثیر اتفاق‌ها را شامل شوند. استخراج اطلاعات رویدادها را قالب‌هایی از پیش تعیین شده و ساخت‌یافته در نظر می‌گیرد که یک عمل را به انجام دهندگان، زمان و دیگر موجودیت‌های شرکت کننده مرتبط می‌سازد. این تعریف همان تعریف اتفاق‌های ساده در زبان شناسی می‌باشد. به طور کلی اتفاق‌ها مجموعه‌ای از فعالیت‌ها همراه با موجودیت‌های مرتبط می‌باشد.

در این روش ابتدا می‌بایست مجموعه موجودیت‌های مهم (مانند: فرد، زمان و مکان‌ها) را که رابطه نامیده می‌شوند، استخراج نمود، سپس ارتباط میان این رابطه‌ها را که همان فعل‌ها و اسم‌های فاعلانه می‌باشند و به آن‌ها ارتباط دهنده گفته می‌شود، استخراج کرد. برای امتیاز دهی به جملات در این روش از میزان باهم‌آیی رابطه‌ها استفاده می‌شود(حسامی فردو همکاران،۱۳۸۴).

روش های مبتنی بر یادگیری نظارتی

تاکنون روش‌های یادگیری نظارتی و بدون نظارتی گوناگونی برای بدست آوردن یا تقریب زدن مدلی که انسان‌ها برای خلاصه‌سازی استفاده می‌نمایند، ارائه شده است. اما در رابطه با این که به طور کلی کدامیک از روش‌ها بهتر عمل می‌نماید، اختلاف نظر وجود دارد. از دلایل این امر می‌توان به اختلاف نظر در اهمیت موضوعات گوناگون در متون و هزینه زیاد تولید یک پیکره با اندازه مناسب اشاره نمود(Porter,۱۹۸۰).

برای تعیین میزان اهمیت واحدهای متنی، می‌توانیم هریک از جنبه‌های مطرح پیشین برای امتیازدهی را بکار بریم. این که امتیاز ارائه شده توسط هر یک از روش‌ها چه مقدار در امتیاز نهایی آن‌ها می تواند تأثیر داشته باشد، توسط روش‌های

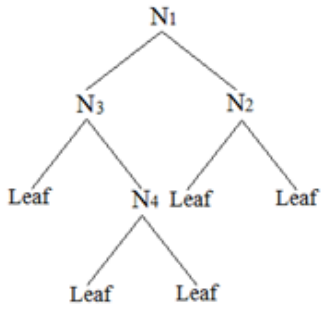
[۱] .Event

ISBN: 978-600-7461-60-0

یادگیری ماشین تعیین می‌شود. دو نمونه از الگوریتم‌های یادگیری نظارتی را که در خلاصه‌سازی بکارگرفته شده‌اند، مورد بررسی قرار خواهد گرفت.

- روش‌های مبتنی بر درخت‌های تصمیم[۱]:

در (عربی نرئی و همکاران،۱۳۸۶) یک سیستم خلاصه‌سازی بر پایه الگوریتم‌های C۴,۵ برای استخراج قوانین با استفاده از درخت‌های تصمیم ارائه شده است. البته برای بکار گیری این الگوریتم در خلاصه‌سازی چند تغییر در آن اعمال شده است. یکی از این تغییرات، استفاده از یک روش بهینه سازی، با بکار گیری هرس نمودن، بر پایه اصل کوتاه‌ترین توضیح می‌باشد. یکی دیگر از این تغییرات، تغییر درخت بگونه ای است که به جای آن که مشخص کند یک واحد متن می‌بایست در خلاصه قرار بگیرد یا خیر، احتمال انتخاب آن را برای قرار گرفتن در خلاصه تعیین می‌نماید. این الگوریتم برای تولید خلاصه از جنبه‌هایی چون طول، مکان و مجموعه‌هایی از ویژگی-های زبانی استفاده می‌کند. شکل (۱) نمایش نمونه‌ای از یک درخت می‌باشد.



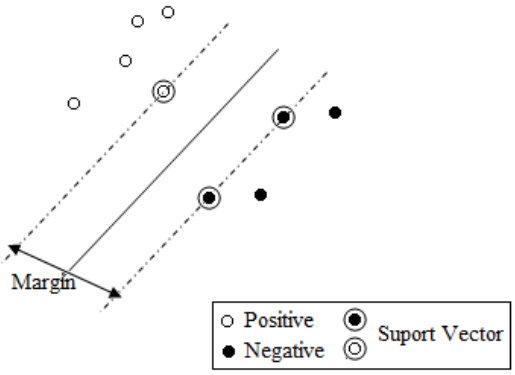
شکل (۱)- نمونه‌ای از یک درخت تصمیم`

- بردارهای پشتیبان[۲]:

در (ایران پور،۱۳۸۶) پیشنهاد شده است که با در نظر گرفتن مجموعه‌ای از جنبه‌ها، چون مکان و طول برای جملات، یک ماشین بردار پشتیبان را بر روی یک پیکره خلاصه‌سازی آموزش داده و سپس به جملات بر اساس میزان فاصله آن‌ها از صفحه بردار پشتیبان امتیاز داده شود.در شکل (۲) نمونه ای از بردارهای پشتیبان نشان داده شده است.

۱ . Decision tree
۲ . Support Vector

ISBN: 978-600-7461-60-0



شکل(۲) مثالی از بردارهای پشتیبان

روش‌های ساختار کلامی

تئوری ساختار کلامی توسط تامسون[۱] و مان[۲] در سال ۱۹۸۸ ارائه گردید. ارتباط موجود بین دو بخش غیر هم پوشان متن به نام بخش پایه[۳] و پخش پیرو[۴] بیان می‌شود(مدرس خیابانی،۱۳۸۵). تفاوت بین پایه و پیرو در این است که بخش پایه هدف اصلی نویسنده را بیان می‌دارد. مسئله دیگر آن است که بخش پایه و بخش پیرو از نظر معنایی و فهمی از یکدیگر مستقل می‌باشند. تئوری ساختار کلامی شامل ساختن درخت دودویی است که ارتباط انسجامی که در متن موجود است را بیان می‌دارد. ارتباط بین واحدهای بیانی تشخیص و برچسب گذاری می‌شود. دو هدف برای تحلیل ساختار کلامی وجود دارد. اول آن که بخش غیر مهم جمله‌ها حذف شود و دوم آن که ارتباط کلامی جمله‌های باقیمانده حفظ شود تا خلاصه منسجم ایجاد شود. سه سطح ارتباط کلامی پاراگراف، جمله وبخشی از جمله وجود دارد. ارتباط کلامی بین پاراگراف‌های همسایه به این صورت مشخص می‌شود که آیا آن‌ها یک موضوع را بر اساس ضریب باهم آیی اش دنبال می‌نمایند یا خیر. همچنین ارتباط کلامی بین جمله‌ها بر اساس استفاده از باهم‌آیی معنایی می‌باشد.

بر اساس تئوری ساختار زبانی یک متن منسجم می‌تواند به صورت یک درخت ساخت‌یافته در نظر گرفته شود که گره‌های میانی آن ارتباطات کلامی و برگ‌ها گزاره‌هایی می‌باشند که به وسیله بخش‌های موجود در متن توضیح داده می‌شوند. تجزیه کننده‌های تئوری ساختار زبانی وجود دارند که درخت تئوری ساختار زبانی مناسبی برای متن ایجاد می‌نمایند. مطابق با قاعده تئوری ساختار زبانی، تمامی واحدهای گزاره‌ای که در متن موجود می‌باشند، می‌بایست به وسیله ارتباط کلامی- بیانی[۵] به یکدیگر متصل باشند تا متن پیوسته باشد. اتصال واحدهای گزاره‌ای متن ساختار کلامی- بیانی را ایجاد می‌نماید. تئوری ساختار زبانی عموماً به وسیله درخت نشان داده می‌شود که هر ارتباطی در آن برای متصل کردن زیر درخت‌ها به کار می‌رود که خود می‌توانند درخت دیگری باشند.

در تئوری ساختار کلامی از یک سری نشانه‌ها همانند "اگرچه"، "بنابراین" و "در نتیجه" برای پیدا کردن ارتباط بین بخش‌های مختلف جمله استفاده می‌نمایند. ابتدا از یک تجزیه کننده استفاده می گردد تا مشخص شود چگونه زیردرخت‌ها به یکدیگر متصل شوند. ۲۵ ارتباط زبانی در تئوری ساختار زبانی وجود دارد که از آن جمله می توان به نتیجه و تصدیق

---

۱. Thompson
۲. Mann
۳. Nucleus
۴. Satellite
۵. Rhetorical

ISBN: 978-600-7461-60-0

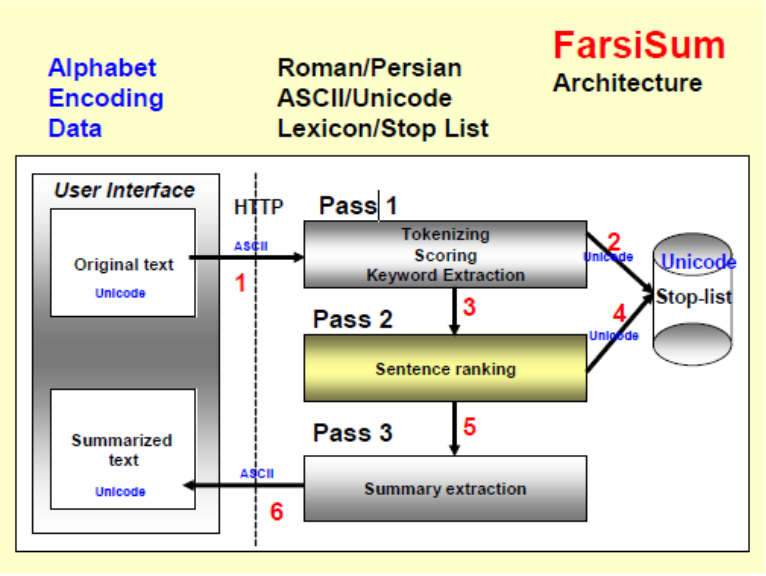
اشاره نمود.تمام روش‌هایی که بر اساس تئوری ساختار زبانی عمل می‌نمایند از روش سطحی بهتر هستند. این تئوری به میزان زیادی در محاسبات زبان شناختی استفاده می‌شود و دیگر کارهای انجام شده در حوزه پردازش زبان همانند تفکیک ضمایر و امتیاز دهی به سندها را تحت تأثیر قرار داده است(مشکی و آنالویی،۱۳۸۷).

علاوه بر مشکل بودن روش‌هایی که بر اساس تحلیل ساختار متن عمل می‌نمایند، از مهم‌ترین معایب این روش‌ها می‌توان به پردازش‌های پر هزینه زبان شناختی اشاره نمود، این امر باعث می‌شود در مواردی نتیجه دلخواه ایجاد نشود. (Mani and Maybury,۲۰۱۴)

لازم به ذکر است که کلیه روش‌های فوق الذکر را می‌توان به صورت ترکیبی جهت خلاصه سازی متون مورد استفاده قرار داد.

مهم‌ترین پژوهش‌های انجام‌شده در خلاصه سازی در زبان فارسی

در این بخش به فعالیت‌های انجام‌شده در زبان فارسی پرداخته می‌شود.در(Hovy and Marcu,۱۹۹۸) یک خلاصه ساز مبتنی بر وب به نام FarsiSum معرفی شده است که نسخه تغییر یافته یک خلاصه ساز سوئدی به نام SweSum برای پوشش زبان فارسی است. این سیستم در حال حاضر به صورت یک ابزار قابل استفاده برای کاربران می باشد.پروسه تولید خلاصه در این سیستم که در شکل (۳) نیز نشان داده شده است شامل سه مرحله ی ارسال درخواست خلاصه از طریق سرور HTTP، خلاصه سند در طی سه مرحله با استفاده از لیست توقف فارسی و بازگرداندن متن خلاصه شده از طریق HTTP به کاربر و پردازش متن خلاصه توسط مرورگر می باشد.



شکل (۳)-FarsiSum

در مقاله(Barzilay et al,۲۰۰۲) روشی برای ساخت سیستم خلاصه سازی خودکار متون فارسی مطرح کردند. کل روش به صورت نحوی/ معنایی عمل می کند و ترکیبی از دو روش زنجیره‌های لغوی و خلاصه‌سازی مبتنی بر گراف است که از پنج معیار میزان شباهت جملات با یکدیگر، شباهت جملات با کلمات کلیدی کاربر، شباهت جملات با عنوان، تعداد

ISBN: 978-600-7461-60-0

جملات مشابه هر جمله و وجود کلمات اشاره در جمله برای امتیازدهی به جملات استفاده نموده و جملات با بیشترین امتیاز را به عنوان خروجی سیستم انتخاب می کند. در واقع با محاسبه ارتباط بین جملات (گره‌های گراف) بر اساس زنجیره لغوی، شباهت بین جملات را بر اساس مفهوم آن‌ها در نظر گرفته‌اند. سیستم پیاده‌سازی شده براساس این روش با خلاصه‌های مرجعی که به صورت دستی تهیه شده‌اند، مورد مقایسه قرار گرفته اند. ارزیابی سیستم پیاده سازی شده به صورت دستی انجام گردیده‌است. به علت عدم دستیابی به منبعی شامل متون و خلاصه مرجع در زبان فارسی، خلاصه مرجع ۱۰ متن از روزنامه همشهری با نسبت فشردگی ۳۰ درصد توسط چند نفر از دانشجویان تولید شده است، که شباهت خلاصه مرجع با خلاصه سیستم در این متون به صورت متوسط ۵۶/۶۶ ٪ است. در(مشکی ،۱۳۸۸) یک روش مبتنی بر خوشه بندی برای خلاصه سازی چند سندی متون فارسی پیشنهاد گردیده است. در این روش، پس از پیش پردازش متن، در مرحله خلاصه-سازی، ابتدا جمله‌ها خوشه‌بندی می شود و سپس به ازای هر خوشه جمله‌ای که بیشترین ارتباط با سایر جمله‌ها را دارد، گزینش می‌شود. در آخرین مرحله خلاصه سازی، جمله‌ها با توجه به ترتیب زمانی متن‌ها در خلاصه نهایی درج می گردند. نتایج پیاده‌سازی نشان می دهند که در بیشتر موارد خروجی سامانه خلاصه‌سازی پیشنهادی خلاصه قابل قبولی را تولید می کند. جهت ارزیابی نتایج این روش پیشنهادی نیز از شیوه قضاوت انسانی استفاده شده‌است. بدین صورت که برای خلاصه هر یک از ۱۰ مجموعه، ۳ داور انسانی رای خود را به صورت ۳۳ ٪ خوب، ۵۱ ٪ متوسط، ۱۶ ٪ ضعیف ارائه شده است.سیستم دیگری که سعی کرده معایب سیستم های مشابه موجود را برطرف نماید سیستم(شهاب،۱۳۸۱) می باشد که ترکیبی از روش های مبتنی برگراف و الگوریتم ژنتیک است که پس از وزن دهی جملات و تشکیل ماتریس شباهت برای جملات سند، گراف جهت داری بوجود می آورد. پس از آن سه فاکتور شباهت با عنوان، پیوستگی و قابلیت خوانایی برای جملات موجود در خلاصه محاسبه می‌شود و در مرحله بعد با استفاده از تابع برازندگی و الگوریتم ژنتیک، جملات خلاصه انتخاب می‌شوند. جهت ارزیابی این سیستم از ده متن در مقوله‌های مختلف علمی و خبری استفاده شده است. این متن‌ها توسط افراد مختلف با نسبت فشردگی مختلف خلاصه شده‌است. با مقایسه خلاصه سیستم و خلاصه‌های دستی مشاهده شده است که؛ به طور میانگین در ۶۰/۲۳ درصد موارد جملات خلاصه مانند جملات خلاصه دستی انتخاب شده‌اند.سیستم خلاصه ساز دیگری که در مقاله(مشکی ،۱۳۸۸) ارائه گردیده است با الهام از شیوه انسان در خلاصه سازی متن، روشی جدید برای خلاصه سازی متن براساس گزینش ارائه می دهد. مبنای روش پیشنهادی یافتن زنجیره‌ای از جمله‌ها می‌باشد که قوی‌ترین همبستگی را با هم داشته باشند. به این منظور متن به دو دید بیرونی و دید داخلی تقسیم بندی شده است. در دید بیرونی متن به پاراگراف‌ها تقسیم بندی می‌شود و در دید داخلی هر پاراگراف به مجموعه‌ای از جملات تقسیم می‌شود و اعضای داخلی هر پاراگراف را تشکیل می‌دهند. انتخاب پارامترها نقش مهمی را درمشخص کردن جمله‌های برجسته برای قرار گرفتن در خلاصه، بازی می کند. به همین دلیل در این روش ارائه شده از میزان نزدیکی جملات به اول یا آخر پاراگراف، تعداد کلمه‌های مشترک بین اجزای جمله‌ها، میزان فاصله پاراگراف‌ها از یکدیگر و تعداد کلمه‌های کلیدی مشترک بین اجزای جمله‌ها استفاده شده است. پس از انتخاب پارامتر میزان همبستگی بین جملات محاسبه شده و در نهایت زنجیره‌ای از جمله‌هایی را که قوی‌ترین ارتباط را با یکدیگر دارند پیدا می کند. در این زنجیره از هر پاراگراف فقط یک جمله انتخاب می‌شود و شرط انتخاب آن جمله قرار گرفتن در زنجیره‌ای می‌باشد که حاصل جمع همبستگی اجزای آن حداکثر باشد.

ISBN: 978-600-7461-60-0

شروع زنجیره از پاراگراف اول می‌باشد و به دلیل این که الزاماً حق انتخاب یک جمله از هر پاراگراف وجود دارد، زنجیره نهایی از هر پاراگراف فقط یک عضو خواهد داشت. در واقع زنجیره حاصل که حاصل جمع همبستگی اجزای آن حداکثر است، همان خلاصه متن ورودی خواهد بود. به عبارت دیگر پس از پیدا کردن زنجیره مناسب به ازای هر عضو آن، جمله‌های متناظر انتخاب می‌شوند و خلاصه نهایی را تشکیل می‌دهند.برای ارزیابی نتایج حاصل از روش ذهنی استفاده شده است. در این نوع ارزیابی از افراد خبره به عنوان داور برای قضاوت در مورد خلاصه تولید شده استفاده می‌شود.برای این منظور ۴ داور از افراد خبره انتخاب شده و نتایج خلاصه‌سازی به همراه اصل متون را بررسی کرده–اند و قضاوت خود را در رابطه با نتایج بدست آمده در قالب یکی از کلمات ضعیف، قابل قبول و خوب بیان کرده‌اند. که ۴۰ ٪ رای خوب ۴۵ ٪ قابل قبول و ۱۵ ٪ ضعیف داده شده است.

نتیجه‌گیری

به‌طورکلی روش‌های خلاصه سازی در چهار دسته قابل تقسیم‌بندی هستند که شامل روش های آماری سطحی، درک متن، مبتنی بر یادگیری نظارت شده و ساختار کلامی هستند. روش های آماری سطحی تنها از ویژگی های آماری از قبیل فراوانی واژه ها، مکان واحد های متنی، عبارات توضیحی، امضاء سرفصل و با هم آیی برای تولید خلاصه استفاده می کنند (Barzilay et al,۲۰۰۲). مزیت اصلی این گونه روش ها این است که بسیار ساده اند و بر روی هر نوع متن ورودی می توان آن ها را اعمال کرد و عیب مهم این دسته از روش ها عدم در نظر گرفتن ارتباط بین کلمات است و به همین دلیل عدم انسجام و پیوستگی در خلاصه های تولیدی به وسیله ی این روش ها به وضوح قابل مشاهده است.

دسته دوم از روش های خلاصه سازی، روش های درک متن می باشند. اینگونه روشها علاوه بر نیاز به منبع دانش نیازمند تفسیر پیچیده متن می باشند ولی عملکرد بهتری را نسبت به روش های آماری ارائه می دهند. در این روش ها از ارتباط میان موجودیت ها شامل شباهت، هم معنی، ارتباط نحوی و ارتباط معنایی ، ارتباط لغوی و هم مکانی استفاده می کنند.

دسته سوم روش های مبتنی بر یادگیری نظارتی بودند که گونه ای از یادگیری است که در آن به عامل یادگیرنده تعدادی جنبه و مقدار متناظر برای خصوصیتی که بایستی یاد گرفته شود داده شده و آموزش لازم جهت عمل نمودن در موارد مشابه داده می شود و دسته ی آخر روش های ساختار کلامی بودند که برای تشخیص بخش های مهم متن استفاده می شدند. توجه به این روش باعث انسجام خلاصه های تولیدی می شود.

با توجه به موارد ذکر شده در این پژوهش، توجه به ترکیب روش های مطرح شده برای تولید خلاصه های باکیفیت تر بسیار اهمیت دارد. توجه به یکی از این روشها به تنهایی باعث می شود که خلاصه تولیدی از لحاظ پیوستگی و انسجام و یا انتخاب جملات دچار مشکل گردد. به طور مثال در روش های آماری تنها تکیه بر مباحث آماری شده است و در صورتی که با روش های درک متن و ارتباط کلامی ترکیب گردد به نتایج قابل قبول تری می رسد. همچنین می توان یک ضریبی را تعیین نمود و مقدار تاثیر هر کدام از روش ها را با استفاده از آن ضریب تنظیم نمود و نقطه طلایی سیستم را که در آن حالت خلاصه های بهتری تولید می گردد را تعیین نمود. کارهایی که در زبان فارسی انجام گرفته بیشتر به روش های آماری و مبتنی بر درک متن (شامل روشهای خوشه بندی و مبتنی بر گراف) متکی است حال آنکه از طریق ترکیب با روش های روابط کلامی می تواند به نتایج بهتری رسید.

ISBN: 978-600-7461-60-0

منابع و مراجع

شهاب،امیرشهاب ، چکیده سازی متون فارسی ، دومین کنفرانس علوم شناختی، تهران ، ص ۵۶،۱۳۸۱.

کریمی، زهره ، شمس فرد، مهرنوش، سیستم خلاصه سازی خودکار متون فارسی ،.دوازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، تهران ، دانشگاه شهید بهشتی، ۱۳۸۵.

مشکی محسن، آنالویی، مرتضی ،خلاصه سازی چند سندی متون فارسی با استفاده از یک روش مبتنی بر خوشه بندی ،اولین کنفرانس ملی مهندسی نرم افزار، دانشگاه آزاد رودهن،۱۳۸۸.

اخوان، تارا.، شمس فرد، مهرنوش. عرفانی جورابچی، مونا،PARSUMIST خلاصه ساز تک سندی و چند سندی متون فارسی ،.چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، تهران، دانشگاه صنعتی امیر کبیر،۱۳۸۷.

بهره پور، مجید.، مهدی پور، الهام، کامل، آزاده، امیری، ملیحه، طهماسبی، آیدا.،اکبرزاده توتونچی ،محمدرضا ،سیستم خلاصه سازی خودکار متن های فارسی،.چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، تهران ، دانشگاه صنعتی امیرکبیر، ۱۳۸۷.

ستوده، حمیدرضا، اکبرزاده توتونچی، محمدرضا، تشنه لب ،محمد ،خلاصه سازی متن بر اساس گزینش با استفاده از رویکرد انسان شناختی ،.هجدهمین کنفرانس مهندسی برق ایران، اصفهان، دانشگاه اصفهان، صفحه ۲۱-،۲۳،۱۳۸۹.

بازقندی.مهدی.، تدین تبریزی، قمرناز.، وفایی جان، مجید.، بازقندی، علی،خلاصه سازی گزینشی متون فارسی مبتنی بر خوشه بندی PSO ،اولین کنفرانس بین المللی پردازش خط و زبان فارسی، ایران، دانشگاه سمنان،۱۳۹۱.

هنرپیشه، محمد علی،طراحی و پیاده سازی یک سیستم خلاصه ساز متون فارسی ، پایان نامه دوره کارشناسی ارشد، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، ۱۳۸۶.

ارژنگ، غلامرضا، زبان و ادب فارسی، تهران: نشر قطره،۱۳۸۳.

ریاحی، نوشین.، غزالی، فاطمه.، محمد علی ،بهبود کارایی سیستم خلاصه ساز متون فارسی با استفاده از الگوریتم هرس در شبکه های عصبی ، اولین کنفرانس بین المللی پردازش خط و زبان فارسی، ایران، دانشگاه سمنان،۱۳۹۳.

حافظی، محمد مهدی.، ثامتی، حسین.، منصوری، نیلوفر.، منتظری، نیلوفر.، بحرانی، محمد.، موثق، حامد،ارائه یک مدل دستوری برای بهبود دقت سیستم های بازشناسی گفتار پیوسته فارسی ،.دومین کارگاه پژوهشی زبان فارسی و رایانه، تهران ، دانشگاه تهران،۱۳۸۵.

حسامی فرد، رضا.، قاسم ثانی، غلامرضا،طراحی و پیاده سازی یک الکوریتم ریشه یابی برای زبان فارسی ،.یازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، تهران، ص ۵۱۵-۵۱۹،۱۳۸۴.

عربی نرئی، سمیه.، وحیدی اصل، مجتبی.، مینایی بیدگلی، بهروز،استخراج کلمات کلیدی جهت طبقه بندی متون فارسی ،.اولین کنفرانس داده کاوی ایران، تهران، دانشگاه صنعتی امیرکبیر،۱۳۸۶.

ایران پور مبارکه، مهدی،بررسی مشکلات تعیین حدود جمله و کلمه ،.سمینار کارشناسی ارشد، دانشگاه علم و صنعت،۱۳۸۶.

مدرس خیابانی، شهرام.، قیومی، مسعود ،نقش پیکره های زبانی در با هم آیی: رویکردی مقایسه ای ،دومین کارگاه پژوهشی زبان فارسی و رایانه، تهران، ص ۵۵ تا ۵۶،۱۳۸۵.

مشکی، محسن.، آنالویی، مرتضی،یک روش آماری مبتنی بر پیکره برای جداسازی واژه های به هم چسبیده ،.دومین کنگره مشترک سیستم های فازی و هوشمند ایران،۱۳۸۷.

مشکی، محسن،خلاصه سازی گزینشی چند سندی متون فارسی،پایان نامه دوره کارشناسی ارشد، دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر،۱۳۸۸.

ISBN: 978-600-7461-60-0

فتوحی، رضا،ارزیابی کارایی الگوریتم های رتبه بندی صفحه برای اسخراج صفحات وب، هشتمین سمپوزیوم مهندسی کامپیوتر و توسعه پایدار با محوریت شبکه های کامپیوتری ، مدل سازی و امنیت سیستم ها، موسسه آموزش عالی خاوران ، مشهد ۱۳۹۲.

MeKeown, k., Barzilay, R., Chen, J., Elson, D., Evans, D., Klavans, J., Nenkova, A., Schiffman, B., and Sigelman, S. (۲۰۰۳). "Columbia'S News Blaster: New Features and Future Directions". In North American Chapter of the Association for Computational Linguistics on Human Language Technology, Pages ۱۵-۱۶.

Buyukkokten, O., Gacia – Molina, H., and Paepke A. (۲۰۰۱). "seeing the Whole in Parts: Text Summarization for web Browsing on Handheld Devices". In ۱۰th international www Conference. Hong Kony, China.

Elhadad, N., Kan, M., Klavans, J., and McKeown, K.(۲۰۰۵) "Customization in a unified framework for summarizing medical literature". In Journal Artificial Intelligence in Medicine, Volume ۳۳, Pages ۱۷۹-۱۹۸.

Mazdak, N., Hassel, M. (۲۰۰۴). "FarsiSum-a persian text summarizer", Master thesis, Department of linguistics, Stockholm University .

Dalianis, H. (۲۰۰۰). "SweSum - A Text Summarizer for Swedish, Technical report ."In Interaction and Presentation Laboratory, TRITA-NA-P۰۰۱۵, IPLab-.۱۷۴

Barzilay, R., Elhadad, N., and McKeown, K. (۲۰۰۲). "Inferring Strategies for Sentence Ordering in Multidocument News Summarization". In Journal of Artificial Intelligence Research, Pages ۳۵-.۵۵

Chuang, W., Yang, J. (۲۰۰۰). "Extracting Sentence Segments for Text Summarization: A machin Learning Approach". In ۲۳rd annual international Conference ACM SIGIR, Athens, Greece.

Porter, M. F. (۱۹۸۰). "Readings in information retrieval". Morgan Kaufmann Publishers Inc. San Francisco, Pages ۳۱۳-۳۱۶.

Mani, I., and Maybury, M. (۲۰۱۴). Advances in Automatic Text Summarization .Cambridge, MA: The MIT Press .

Hovy, E., and Marcu, D. (۱۹۹۸). "Automated Text Summarization". Tutorial at OLING/ACL'.۹۸

R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. de França Silva, S. J. Simske, et al. ۲۰۱۴, "A multi-document summarization system based on statistics and linguistic treatment," Expert Systems with Applications, vol. ۴۱, pp. ۵۷۸۷-۵۷۸۰.

ISBN: 978-600-7461-60-0