دانشکده مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

خلاصه سازی خودکار چندسندی مبتنی بر استخراج مفاهیم

تهیه و تنظیم:

آصف پورمعصومی

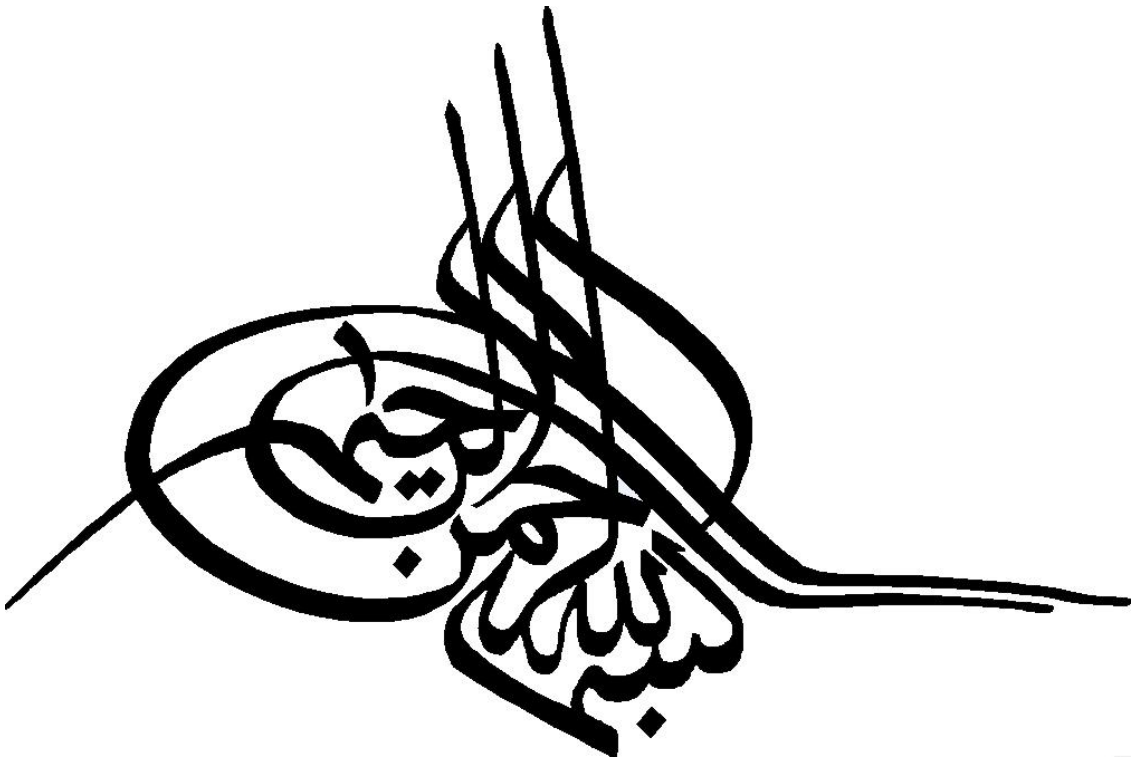
استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر نادر جهانگیری

تابستان ۹۰


بسم الله الرحمن الرحيم

تقدیم به:

کامل شود.

تقدیر و تشکر

کامل شود.

چکیده:

کامل شود.

کلمات کلیدی:

کامل شود.

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل اول:

مقدمه

۱– مقدمه

با گسترش روزافزون حجم اطلاعات موجود در وب، وجود سایت های خبری مختلف، انتشار کتب الکترونیکی متعدد و افزایش چشم گیر مقالات منتشر شده در زمینه های مختلف علمی، دسترسی درست و دقیق به اطلاعات مورد نیاز، همواره یکی از مشکلات محققان و پژوهشگران قرن ۲۱ بوده است. حجم انبوه منابع اطلاعات از یک سو و محدودیت زمان از سوی دیگر، منجر به گرایش محققان به موضوع جذاب خلاصه‌سازی خودکار متن شده است.خلاصه‌سازی خودکار متن به‌عنوان هسته‌ی مرکزی طیف گسترده‌ای از ابزارهای پردازش‌گر متن مانند خلاصه‌سازهای ماشینی، سیستم‌های تصمیم‌یار، سیستم‌های پاسخ گو، موتورهای جستجو و ... از سال‌ها پیش مطرح شده و همواره به عنوان یک موضوع مهم مورد بررسی و تحقیق قرار گرفته است. خلاصه‌سازی خودکار سند، یعنی تولید یک نسخه مختصرتر از سند اصلی توسط یک برنامه رایانه‌ای به نحوی که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود[Man-1999]. بنابر تعریف ارائه شده در استاندارد ISO215سال ۱۹۸۶، خلاصه "یک بازگویی مختصر از سند" می باشد. همانطور که اشاره شد خلاصه‌سازی خودکار توسط رایانه انجام می‌شود و به همین دلیل تفاوت‌های زیادی با خلاصه‌ای که توسط انسان تولید می-شود دارد. انسان‌ها با توجه به هوش و شعور ذاتی خود قادر به درک و فهم مفاهیم موجود در متن و ارتباط بین آنها می‌باشند و این در حالی است که انجام این عملیات توسط ماشین کار بسیار دشوار و پیچیده‌ای می-باشد. سیستم‌های خلاصه ساز در دنیای امروز کاربردهای فراوانی دارند. تولید خلاصه‌های کتب مختلف و مقالات علمی، تولید خلاصه اخبار و انتقال آن از طریق سیستم‌های نظیر تلفن همراه، نمایش خلاصه سند یافته شده توسط موتور جستجو ، تولید سیستم های پاسخ گویی به سوال و ... همگی از کاربردهای این سیستم می باشند.

۱–۱– ساختار پایان‌نامه

در این پایان نامه با استفاده از روش آنالیز روابط معنایی پنهان در متن و با بهره گیری ا زبرچسب زنی معنایی جملات و شبکه واژگان روشی جدید برای خلاصه سازی خودکار چندسندی متن ارائه شده است. این نوشتار شامل چهار فصل دیگر به شرح زیر می باشد:

فصل دوم (مرور ادبیات): در این فصل در ابتدا مروری بر تاریخچه خلاصه سازی، کاربردها و همچنین انواع مدل های خلاصه سازی شده است و سپس در ادامه به روش های مختلف خلاصه سازی متن اشاره شده است. مجموعه داده های استاندارد و همچنین روش ها و ابزارهای استاندارد برای ارزیابی خلاصه های سیستمی تولید شده در ادامه مورد بررسی قرار گرفته است. در پایان هم به سیستم های معروف خلاصه ساز و همچنین کارهای انجام شده در زبان فارسی اشاره شده است.

فصل سوم (روش پیشنهادی): در این فصل که شامل چند بخش اصلی می باشد به روش جدید پیشنهادی

برای خلاصه سازی متن اشاره شده است. در بخش اول مرحله پیش پردازش توضیح داده شده است. در بخش دوم به انتخاب جملات متناسب با زمینه اشاره شده است. در بخش سوم هم به محاسبه شباهت معنـایی بـین جملات و حذف افزونگی پرداخته شده است.

فصل چهارم (نتایج) : در این فصل نتایج ارزیابی روش پیشنهادی با استفاده از ابزار ارزیـابی ROUGE بـرروی مجموعه داده های DUC2006 و DUC2007 ارائه شده است. همچنین در ادامه نتایج پیاده سازی و ارزیـابی سیستم پیشنهاد شده برروی زبان فارسی تست ذکر شد.

فصل پنجم (نتیجه گیری): در این فصل به نتیجه‌گیری متدهای ارائه شده پرداخته شده و پیشنهادهایی بـرای کارهای آتی ارائه شده است.

فصل دوم:

مرور ادبیات

۲– مرور ادبیات

۱–۲– خلاصه سازی خوکار متن

خلاصه‌سازی خودکار متون به‌عنوان هسته‌ی مرکزی طیف گسترده‌ای از ابزارهای پردازش‌گر متن مانند خلاصه‌سازهای ماشینی، سیستم‌های تصمیم‌یار، سیستم‌های پاسخ گو، موتورهای جستجو و ... از سال‌ها پیش مطرح شده و همواره به عنوان یک موضوع مهم مورد بررسی و تحقیق قرار گرفته است. خلاصه‌سازی خودکار سند، یعنی تولید یک نسخه مختصرتر از سند اصلی توسط یک برنامه رایانه‌ای به نحوی که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود. همچنین خلاصه تولید شده باید از خوانایی[۱] و پیوستگی[۲] بالایی برخوردار بوده و فاقد اطلاعات تکراری[۳] باشد.روش‌های خلاصه سازی خودکار متن از دیدگاههای متفاوت دسته بندی های مختلفی دارند که در ادامه به آنها اشاره شده است.

۱–۱–۲– تاریخچه خلاصه سازی متن

آغاز فعالیت سیستم های خلاصه سازی خودکار متن مربوط به سال ۱۹۵۰ می شود.به دلیل کمبود کامپیوترهای قدرتمند و مشکلات موجود برای پردازش زبانهای طبیعی (NLP)، کارهای اولیه بروی مطالعه ظواهر متن مانند موقعیت جمله و عبارات اشاره متمرکز شده بود . سال 1970 تا 1980هوش مصنوعی بکارآمد [You 85][Sch 77] [Mck 95][Gra 81][Dej 79][Azz 99]. ایدهی AI استخراج نمایش‌های دانش مانند فریمها یا الگوها برای شناسایی موجودیتهای مفهومی از متن واستخراج روابط بین موجودیت‌ها بامکانیزمهای استنتاج بود . مشکل اصلی آن است که فریم ی االگوهای تعریف شده محدودیتهایی دارند و ممکن است به تحلیل کامل موجودیتهای مفهومی منجر نشود . از اوایل 1990 تا به حال هم روشهای بازیابی اطلاعات(IR) بکار گرفته شده است [Man 99] [Kup 95] [Hov 97] [Gon 01] [Gol 99] [Aon 97] [Yeh 02] [Teu 97] [Sal 97]. بیشتر این روش ها برروی سطح سمبولیک متمرکز بوده و وارد حوزه هایی معنایی نمی شوند.

Kupiec[Kup 95] اولین الگوریتم را پیشنهاد داد . در این روش بر اساس مقادیر ویژگیهای یک جمله، احتمال حضور آن در خلاصه تخمین زده میشود . او عمل خلاصه سازی را به صورت یک مسئله دسته بندی، درنظرگرفت و دسته بندی کننده‌های بیزین را برای تعیین جملات یکه باید در خلاصه وارد شود بکار برد. Chuang و Yang[Chu 00] چندین الگوریتم مانند درخت تصمیم و دسته بندی کننده را برای استخراج

---

۱ Readability

۲ Coherency

۳ Information Redundancy

قطعات جمله پیشنهاد دادند .این روش خلاصه سازی اسناد در یک حوزه خاص عملکرد خوبی دارد. گرچه برای یادگیری صحیح نیازمند مجموعه‌ها ی آموزشی بسیار بزرگی هستند. در سال ۱۹۹۷، [Bar Barzilay97]روشی برای تولید خلاصه با پیدا کردن زنجیره‌های لغوی معرفی کرد که به توزیع کلمه و اتصالات لغوی بین آنها برای تقریب زدن محتوا و ارائه یک نمایش ازساختارلغوی بهم پیوسته متن اتکا می کرد. Gongو Liu [Gon 01] دو روش را پیشنهاد دادند : یکی سنجش ارتباط به منظور رتبه دهی به جملات مرتبط و دیگری استفاده از تحلیل معانی پنهان برای شناسایی جملات مهم از نظر معنایی.

از روش های آماری و مبتنی بر ریاضیات هم در خلاصه سازی متن زیاد استفاده شده است کـه از جملـه آنهـا می توان به روش های مبتنی بر مدل موضوع ؟؟؟ و روش های مبتنی بر گراف ؟؟؟ اشاره نمود.

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

علی رقم اینکه در زمینه خلاصه سازی متن کارهای زیادی از سالهای دور انجام شده است ولی همچنان شـاهد مقالات متعددی هستیم که در همه ساله در این زمینه منتشر مـی شـوند سـعی در بهبـود دقـت روش هـای پیشین دارند.

۲-۱-۲- کاربردهای خلاصه سازی

زمینه های کاربردی خلاصه سازی خودکار متن بسیار گسترده است.با توجه به این که پایـه و اسـاس سیسـتم های خلاصه سازی استخراج مفاهیم و دانش موجود در اسناد می باشد، بنابراین می توان از آن در کاربردهـای مختلفی استفاده نمود. برخی از این کاربردها عبارتند از:

- خلاصه سازی متون خبری
- خلاصه سازی مقالات و کتب علمی
- خلاصه سازی اسناد سازمانی و اداری
- تولید سیستم های پاسخ دهنده به سوالات (question answering)
- تولید سیستم های تصمیمیار (decision making)
- تولید سیستم های توضیح دهنده روی خط (wiki های online)
- استفاده از خلاصه ساز در موتورهای جستجو (در حال حاضر بسیاری از موتورهای جستجو نظیرgoogle برای کمک به کاربر در انتخاب صفحات مورد نظر، از سیستم های خلاصه سازی روی خط استفاده می نمایند)

- تولید سیستمی جهت استخراج اسناد مشابه به صورت خودکار
- ارائه سیستمی جهت بررسی شباهت مقالات و یا گزارشات منتشر شده جدید در مجامع علمی با مقالات انتشار یافته قبلی و کشف تخلفات احتمالی که صورت می گیرد.
- تولید فایلpowerpoint از روی گزارشات به صورت خودکار و بلعکس
- استفاده برای سیستم هایی که با محدودیت حجم تبادل داده روبرو هستند نظیر تلفن های همراه
- تولید سیستم راستی آزمایی اسناد
- کابردهای خلاصه سازی در وب معناییو ...

با توجه به حوزه وسیع کاربردهای خلاصه سازی متن، اهمیت توجه به این موضوع بیش از پیش جلب توجه می کند.

۳-۱-۲- انواع خلاصه سازی متن

سیستم های خلاصه ساز متن معمولا از دیدگاه های مختلفی تقسیم بندی می شوند. این دسته بندی را در چند قالب اصلی می توان تشریح کرد.

- براساس منبع ورودی : بر اساس نوع منبع 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 و ... باشد روش بر خورد 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 همرا جمله های اول پاراگراف ها 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 چندزبانه داریم.روش خلاصه سازی بسته به اینکه بخواهیم متون تک زبانه و یا چند زبانه باشد ، متفاوت می باشد. چراکه 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 :متن خلاصه به چه منظوری تولید می شود. هدف خلاصه سازی ، هشدار ، پیش نمایش، آگاهی ، زندگی نامه و .... در تعیین روش خلاصه سازی تاثیرگذار می باشد. همچنین خلاصه ها می توانند عمومی با مبتنی بر پرسو 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.

در ادامه به شرح تعدادی از این دسته های مهم می پردازیم.

۱-۳-۱-۲- خلاصه سازی استخراجی[۱]

در روش خلاصه سازی استخراجی که عموما اکثر روش های خلاصه سازی هم از این نوع می باشند[???] قسمت هایی از متن به عنوان خلاصه انتخاب شده و سپس با یک چیدمان مناسب در کنار همدیگر قرار گرفته و به عنوان خلاصه تلقی می شود. اکثر این روش ها جملات را جدا کرده و به آنها امتیازدهی کرده و سپس جمله های با بالاترین امتیاز را به عنوان خلاصه انتخاب می کنند. بنابراین در این روش ساختار جمله ها تغییری نمی کند. مسلما پیچیدگی ها در این مدل بسیار کمتر از خلاصه سازی چکیده ای می باشد. مهمترین مسئله در خلاصه سازی استخراجی، انتخاب درست و صحیح جملات می باشد. جملات باید به نحوی انتخاب شوند که ضمن پوشش کامل محتوای متن، فاقد افزونگی بوده و همچنین دارای خوانایی و پیوستگی مناسبی باشد. تقریبا بیشتر روش های ارائه شده از این دست می باشند. روش پیشنهادی ما هم در این دسته قرار می گیرد.

۲-۳-۱-۲- خلاصه سازی چکیده ای[۲]

در روش خلاصه سازی چکیده ای علاوه بر انتخاب جملات مناسب، ساختار جملات هم می تواند عوض شود. در این مدل می توان جملاتی را حذف نمود یا اینکه تغییر داد و یا حتی جملات جدیدی تولید نمود. خلاصه سازی چکیده ای بسیار نزدیک به مدل ذهنی انسان می باشد و همانطور که یک انسان برای یک متن خلاصه تولید می کند، در این مدل هم خلاصه ساز باید قادر به اعمال تغییر در ساختار جملات باشد. به عنوان مثال پاراگراف "علی بلیط هواپیما را گرفت و سوار هواپیما شد. هواپیما از باند فرودگاه بلند شد و به سمت تهران حرکت کرد. هواپیما یک ساعت بعد در فرودگاه مهرآباد فرود آمد و علی از هواپیما پیاده شد" در روش چکیده ای به جمله های "علی به تهران رفت" یا "علی با هواپیما به تهران رفت" می تواند خلاصه شود. یادآور می شویم که این روش بسیار بسیاردشوار بوده و پیچیدگی های آن از موضوعاتی نظیر "ترجمه ماشینی[۳]" هم به مراتب بیشتر می باشد. به همین دلیل اساسا کار مهمی در این زمینه صورت نگرفته است[Bar 99] [Kni 02].

۳-۳-۱-۲- خلاصه سازی تک سندی

در مدل خلاصه سازی تک سندی[۴] ورودی سیستم خلاصه ساز، تنها یک سند می باشد. پیچیدگی خلاصه سازی تک سندی نسبت به خلاصه سازی چند سندی به مراتب کمتر می باشد چراکه در این مدل از خلاصه سازی تنها با یک سند روبرو هستیم که احتمالا در مورد یک موضوع و به صورت پیوسته صحبت می کند و

---

۱ extractive summarization

۲ abstractive summarization

۳ machine translation

۴ Single text summarization

فاقد زیر موضوعات ضد و نقیض خواهد بود. تاکنون روش های زیادی برای خلاصه سازی تک سندی ارائه شده اند که تعدادی از آنها را در منابع ؟؟؟ می توان یافت.

۴-۳-۱-۲- خلاصه سازی چند سندی

در خلاصه سازی چند سندی[۱]ورودی خلاصه ساز چندین سند می باشد.خلاصه‌سازی چندسندی، ارتباط تنگاتنگی با مباحث سیستم‌های پاسخگو[۲]و خلاصه‌سازی مبتنی بر پرس‌وجو دارد[Hir۰۱]. در حقیقت خلاصه سازی چندسندی برروی اسنادی انجام می شود که در ارتباط با یک موضوع هستند ولی جهت دید آنها متفاوت از یکدیگر است. به عنوان مثال موضوع "مشکل جهانی کمبود آب" را در نظر بگیرید. در ارتباط با این موضوع اسناد مختلفی می توانیم داشته باشیم که از دیدگاههای متفاوت به این موضوع پرداخته باشند. مثلا یکی در مورد "کمبود آب در ایران" و دیگری در خصوص "کمبود آب در پاکستان" باشد.هر دوی این سندها به موضوع کمبود آب پراخته اند ولی از دیدگاههای مختلف.

در خلاصه‌سازی چندسندی با پیچیدگی‌های بیشتری نسبت به خلاصه سازی تک‌سندی رویرو هستیم. مهمترین چالش‌های پیش رو در این دسته از خلاصه سازها عبارتند از[wan 07]:

- چون اسناد با دیدگاههای متفاوت به شرح یک موضوع می پردازند که گاها متناقض با یکدیگر هم بوده، بنابراین تولید خلاصه‌ای با خوانایی بالا امری دشوار خواهد بود.
- با توجه به این که همه اسناد در ارتباط با یک موضوع کلی می باشند، بحث همپوشانی جمله¬های انتخاب شده از این اسناد یا همان افزونگی مشکل مهمی می باشد.
- استخراج دیدگاههای مختلف پنهان در اسناد مختلف و پوشش دادن تمامی آنها نیاز به دقت و توجه فراوان دارد.

افزایش قارچ گونه اسناد الکترونیکی در موضوعات شبیه به یکدیگر و مشکلات کمبود زمان برای عموم کاربران، منجر به استقبال روز افزون دانش پژوهان جهت تحقیق در زمینه خلاصه‌سازی خودکار چندسندی شده است. با توجه به ماهیت این مدل از خلاصه سازی، در حال حاضر بیشتر توجهات برروی این مدل از خلاصه سازها می باشد.

۵-۳-۱-۲- خلاصه سازی مبتنی بر پرس و جو

در این مدل متناسب با عبارت پرس و جوی[۳]کاربر، باید از اطلاعات موجود خلاصه ای استخراج شده و

---

۱ Multi-document summarization

2 - Question-answering

۳ Query

برگردانده شود. این مدل از خلاصه سازها به دو دسته کلی زیر تقسیم می شوند:

- اشاره کننده[۱]: در این مدل خلاصه حاوی اطلاعاتی است که به کاربر کمک می کند که تصمیم بگیرد آیا سند مطالعه کند یا خیر!
- اطلاع رسان[۲]: در این مدل خلاصه تمامی اطلاعات لازم برای نمایش و اطلاع رسانی مناسب سند اصلی را در بر می گیرد.

روش های زیادی در این زمینه ارائه شده است. از جمله آنها می توان به روش های مبتنی بر گراف سند نظیر [Bos 05][Moh 06] و روش های مبتنی بر تکنیک های زبان شناسی نظیر [Con 05][Koy 09] و همچنین تکنیک های مبتنی بر یادگیری ماشین نظیر [Sch 08][Jag 07]اشاره نمود.

۶-۳-۱-۲- خلاصه سازی مبتنی بر کاربر یا شخصی سازی شده

به دلیل افزایش حجم انبوه اطلاعات موجود، خلاصه سازی متن طی سال های اخیر بسیار مورد توجه قرار گرفته است. کاربران سیستم های خلاصه سازی متن بسیار زیاد بوده و از افرادی که کارشان وب گردیست تا افرادی که دانشمند و محقق هستند، همگی کاربر اینترنت می باشند. در سیستم های قدیمی بین کاربران تفاوتی فرض نمی شد و همه کاربران با یک دید نگریسته می شدند، به این معنا که یک متن بدون در نظر گرفتن ویژگی ها و اطلاعات پس زمینه ی خواننده آن خلاصه می شد. و این در حالی است که هنگامیکه انسان ها خلاصه سازی می کنند، اطلاعاتی را که مرتبط با علایق خودشان باشند را به عنوان خلاصه بر می گزینند. به عبارتی دیگر، خلاصه سازی فقط تابعی از اسناد ورودی نمی باشد، بلکه پارامتر دومی هم که یک مدل از رفتار و دانش فرد می باشد را نیز شامل می شود. به عنوان مثال فردی که سال‌ها در زمینه ی خاصی مثلی شبکه‌های بیسیم به تحقیق و مطالعه پرداخته است زمانی که بخواهد یک خلاصه از یک کتاب در این زمینه برای خود تهیه کند، مطمئنا مفاهیم اولیه شبکه‌های بیسیم را در این خلاصه لحاظ نخواهد کرد و این در حالی است که یک فرد که به تازگی در این زمینه تحقیق می کند احتمالا این مفاهیم اولیه را در خلاصه مخصوص خودش قرار می دهد. البته در ارتباط با اثبات این گفته ها، تحقیقات علمی هم صورت گرفته است. به عنوان مثال Marcu در سال ۱۹۹۷ [Mar۹۷] نشان داد که میزان تفاهم یا تطابق ۱۳ نفر از جامعه علمی کشور ایلات متحده برروی ۵ متن انتخاب شده ۷۱ درصد می باشد. Rath در [Rat61] نشان داد که خلاصه های تولید شده توسط چهار فرد متفاوت، تنها ۲۵ درصد با هم همپوشانی دارند. Salton هم در سال ۱۹۹۷ [Sal97] دریافت که مهمترین ۲۰ پاراگرافی که توسط دو کاربر انتخاب شد، تنها ۴۶ درصد با هم همپوشانی دارند. این نتایج بیانگر این مطلب است که افراد مختلف دیدگاههای متفاوتی روی متن‌های مشابه دارند و

---

۱ Indicative

۲ Informative

زمانی که افرادی با پس زمینه اطلاعاتی و تخصص های متفاوت، اسناد مشابه‌ای را خلاصه می‌کنند، محتواهای متفاوتی را بر می‌گزینند که انعکاس دهنده‌ی پس زمینه ی اطلاعاتی آنها می باشد. با توجه به موارد ذکر شده، مدل کردن رفتار کاربر کاملا مورد نیاز می باشد. اما مدل کردن رفتار کاربر کار بسیار پیچیده و دشواری می باشد و شاید حتی مدل کردن ۱۰۰ درصد کاربران امری غیر ممکن باشد. اما به هر حال هدف کسانی که در این زمینه فعالیت می کنند این است که با مدل کردن بخشی از رفتار کاربران، خلاصه هایی با کیفیت بیشتر نسبت به خلاصه هایی که بدون در نظر گرفتن دانش کاربر تولید می شد ایجاد نمایند. به عنوان مثال، بنا برتحقیقات صورت گرفته در [Nie 03] در سال ۲۰۰۳، چناچه بتوانیم با استفاده از شخصی سازی مناسب زمانی که کاربران برای جستجو در موتور جستجوی گوگل صرف می کنند را تنها ۱ درصد کاهش دهیم، بیشتر از ۱۸۷۰۰۰ انسان-ساعت که معادل ۲۱ سال است صرفه جویی خواهد شد. در همین راستا فعالیت های زیادی برای مدل کردن رفتار کاربران صورت گرفته که به صورت کلی می توان آنها را به چهار دسته زیر تقسیم بندی نمود :

- مدل کردن بر اساس تاریخچه کوئری های کاربر[1][Rad 05][Liu 02][Spe 05]
- مدل کردن بر اساس داده های کلیک[2][Joa 02][Dup 07]
- مدل کردن بر اساس زمان توجه کاربران[3][Kel 01][Kel 04]
- مدل کردن براساس سایر عکس العمل های ضمنی[4]که از کاربران گرفته می شود[Fox05] [Lv 06]

اکثر روش های خلاصه سازی مبتنی بر کاربر را می توان در یکی از این چهار دسته کلی اشاره شده قرار داد.

۷-۳-۱-۲- سایر پارامترهای تاثیر گذار در دسته بندی خلاصه سازها

به صورت کلی روش‌های خلاصه‌سازی را می‌توان به دو دسته نظارت محور[5]و غیرنظارت‌شونده[6] تقسیم نمود. در روش های نظارت محور یک مجموعه از متن هایی که قبلا توسط انسان خلاصه هایی برای آنها تولید شده است موجود می باشد. این مجموعه متون بهمراه خلاصه های انسانی متناظرشان به عنوان داده های اولیه تست به سیستم داده می شوند و بعد از اینکه عملیات یادگیری برروی آنها انجام شد، سیستم قادر خواهد بود تا متون جدید را خلاصه نماید. بر این دسته از روش ها چند ایراد عمده وارد است که مهمترین آنها مبتنی بر مدل بودن آنهاست. به عبارت دیگر این دسته از خلاصه ها به شدت وابسته به خلاصه های اولیه که توسط یک سری انسان تولید می شوند می باشد و با توجه به بحث مربوط به خلاصه سازی شخصی سازی شده،

---

۱ query history

۲ data click

۳ attention time

۴ implicit user feedbacks

5 - supervised

6 - unsupervised

قطعا نظرات و اطلاعات اولیه این افراد در داده های تست تاثیر گذار می باشد.

در روش های غیر نظارت شونده نیازی به خلاصه های انسانی اولیه نمی باشد و این روش ها با استفاده از اطلاعات موجود خلاصه مورد نظر را تولید می نمایند[???].

۲-۲- تعاریف پایه زبان شناسی

۱-۲-۲- بازیابی اطلاعات

بازیابی اطلاعات[۱] به فن‌آوری و دانش پیچیدهٔ جستجو و استخراج اطلاعات، داده‌ها، فراداده‌هادر انواع گوناگون منابع اطلاعاتی مثل بانک اسناد، مجموعه‌ای از تصاویر، و وبگفته می‌شود.

با افزایش روز افزون حجم اطلاعات ذخیره شده در منابع قابل دسترس و گوناگون، فرایند بازیابی و استخراج اطلاعات اهمیت ویژه‌ای یافته است .اطلاعات مورد نظر ممکن است شامل هر نوع منبعی مانند متن، تصویر، صوتو ویدئوباشد. بر خلاف پایگاه داده‌ها، اطلاعات ذخیره شده در منابع اطلاعاتی بزرگ مانند وبو زیرمجموعه‌های آن مانند شبکه‌های اجتماعی از ساختار مشخصی پیروی نمی‌کنند و عموما دارای معانی تعریف شده و مشخصی نیستند. هدف بازیابی اطلاعات در چنین شرایطی، کمک به کاربر برای یافتن اطلاعات مورد نظر در انبوهی از اطلاعات ساختارنایافته است.جستجوگرهای گوگل، یاهوو بینگسه نمونه از پراستفاده‌ترین سیستم‌های بازیابی اطلاعات هستند که به کاربران برای بازیابی اطلاعات متنی، تصویری، ویدئویی و غیره کمک می‌کنند.

۲-۲-۲- استخراج اطلاعات

استخراج اطلاعات نوعی از بازیابی اطلاعات می باشد که هدف آن استخراج خودکار اطلاعات ساخت یافته از میان اسناد غیر ساخت یافته یا شبه ساخت یافته قابل خواندن توسط ماشین می باشد.

۳-۲-۲- کلمات ایست (Stop words)

کلمات ایست لغاتیهستند که در جملات کاربرد ربطی دارند و زیاد استعمال می شوند مثل " اگر" ،" و"، "or"،"the". این کلمات علی رغم اینکه بسیار استفاده می شوند اما از لحاظ معنایی دارای اهمیت کمی بوده و بهمین دلیل عموما در فعالیت های مربوط به حوزه پردازش زبان طبیعی در فاز پیش پردازش حذف می شوند. برای حذف این کلمات عموما لیستی از این کلمات از پیش تهیه می شود و سپس در صورت رخداد این کلمات در متن، از سند حذف می شوند. برای زبان انگلیسی چندین لیست از این کلمات منتشر شده است که

---

۱ Information Reterival

بطور میانگین شامل ۵۰۰ کلمه می باشند.

۴-۲-۲- ریشه یابی

ریشه یابی به فرآیند کاهش دادن لغات به ریشه های آنها اطلاق می گردد. بنابراین "computer" و "compute"و "computing" به"compute" که ریشهٔ اصلی است کاهش می یابند. لازم به ذکر است که منظور از ریشه در این بخش، دقیقا ریشه کلمات که در زبان شناسی استفاده می باشد نمی باشد. بلکه منظور از ریشه یک نماینده برای کلماتی است که از لحاظ معنایی و نحوی در یک حوزه قرار می گیرند. تمامی سیستمای بازیابی اطلاعات نوع یکسانی از "ریشه یاب" را مورد استفاده قرار نمی‌دهند. در انگلیسی معروف ترین ریشه یاب، الگوریتم ریشه یاب "مارتین پورتر"[Por 80] است.

۵-۲-۲- برچسب زنی نحوی (POS)

برچسب گذاری اجزای واژگانی کلام[۱]عمل انتساب برچسب های واژگانی به کلمات و نشانه های تشکیل دهنده یک متن است به صورتی که این برچسب ها نشان دهنده نقش کلمات و نشانه ها در جمله باشد. درصد بالایی از کلمات از نقطه نظر برچسب واژگانی دارای ابهام هستند زیرا کلمات در جایگاههای مختلف برچسب های واژگانی متفاوتی دارند. بنابراین برچسب گذاری واژگانی عمل ابهام زدایی از برچسب ها با توجه به زمینه (متن) مورد نظر است. برچسب گذاری واژگانی عملی اساسی برای بسیاری از حوزه های دیگر پردازش زبان طبیعی (NLP) از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می باشد.

در زیر زیر نمونه ای فرضی از یک مجموعه برچسب (Tagset) برای زبان فارسی معرفی شده است:

ADV	Adverb	قید
AJ	Adjective	صفت
CL	Classifier	شاخص
CONJ	Conjunction	حرف ربط
DET	Determiner	حرف تعریف
INT	Interjection	حرف صوت
N	Noun	اسم
NUM	Number	عدد
P	Preposition	حرف اضافه
POSTP	Postposition (را)	حرف اضافه پسین
PRO	Pronoun	ضمیر
PUNC	Punctuation	جدا کننده
RES	Residual:Arabic and Latin words, etc	متفرقه
V	Verb	فعل

شکل ۱- خروجی POS

[۱] Part of Speech tagging

۶-۲-۲- برچسب زنی نقش معنایی کلمات (SRL)

برچسب زنی معنایی کلمات[۱] مشابه برچسب گذاری اجزای واژگانی کلام بوده با این تفاوت که عمیق تر و پیچیده تر از آن می باشد. برچسب زنی معنایی وظیفه استخراج نقش های معنایی جملات نظیر فاعل، مفعول مستقیم، مفعول غیر مستقیم، فعل و ... را بر عهده دارد. برچسب زنی معنایی کلمات هم عملی اساسی برای بسیاری از حوزه های دیگر پردازش زبان طبیعی (NLP) از قبیل ترجمه ماشینی، خطایاب و شباهت معنایی می باشد.در زیر یک مجموعه برچسب معرفی شده است:

ARGUMENTS	
A0	subject
A1	object
A2	indirect object
ADJUNCTS	
AM-ADV	adverbial modification
AM-DIR	direction
AM-DIS	discourse marker
AM-EXT	extent
AM-LOC	location
AM-MNR	manner
AM-MOD	general modification
AM-NEG	negation
AM-PRD	secondary predicate
AM-PRP	purpose
AM-REC	recipricol
AM-TMP	temporal

شکل ۲- خروجی SRL

۷-۲-۲- شبکه واژگان

WordNet فرهنگی از واژگان است که براساس تئوری‌های زبانی-روانی[۲] بوده و مدل‌ها و معانی کلمات را تعریف می‌کند. در تعریف مدل‌های کلمات، WordNet نه تنها تداعی معانی واژگان را شامل می‌شود، بلکه تداعی معنی-معنی را نیز در بر می‌گیرد. WordNet بیشتر بر معنی کلمات تکیه دارد تا فرم کلمات، البته در WordNet ریخت‌شناسی صرف افعال نیز مد نظر قرار گرفته است.WordNetشامل سه پایگاه داده می‌باشد: یکی برای اسامی، یکی برای افعال و یکی نیز مشترکاً برای صفات و قیود. WordNet شامل مجموعه‌ی مترادف‌های کلمات می‌باشد که از آن به عنوان ''Synsets'' یاد می‌شود. هر Synset یک مفهوم و یا یک معنی از گروهی از کلمات، را شامل می‌شود. Synset- ها روابط معنایی متفاوتی چون مترادف،[۳] متضاد،[۱] ابرمفهوم،[۲]

۱- Semantic Role Labeling

۲- Psycholinguistic

۳Synonymy (Similar)

زیرمفهوم [۳](IS-A)،جزئیت[۴](Part of) ، شمول[۵](Has-A) را دربر می‌گیرند. روابط معنایی بین Synset- ها با توجه به طبقه‌بندی‌های گرامری- همانند آنچه در جدول ۲-۷ دیده می‌شود- متفاوت است [LIN08]. WordNet هم‌چنین تعاریف متنی از مفاهیم را فراهم می‌سازد (Glossary) که شامل تعاریف و مثال‌ها می‌باشد. WordNet را می‌توان به عنوان یک مجموعه‌ی مرتب جزئی[۶] از منابع عبارات مترادف، برشمرد.

رابطه‌ی معنایی	دسته‌ی نحوی	مثال
ترادف (Synonymy)	فعل اسم صفت قید	Rise, ascend Pipe, tube Sad, unhappy Rapidly, speedly
تضاد (Antonymy)	فعل اسم صفت قید	Rise, fall Top, bottom Wet, dry Rapidly, slowly
زیرمفهوم (Hyponymy)	اسم	Sugar, maple, maple maple, tree tree, plant
جزئیت (Meronymy)	اسم	Brim, hat gin, martini ship, fleet
حالت انجام فعل (Troponymy)	فعل	Match, walk whisper, speak
Entailment	فعل	Drive, ride

[۱] Antonymy (Opposite)

[۲] Hypernymy (Superconcept)

[۳] Hyponymy (Subconcept)

[۴] Meronymy

[۵] Holonymy

[۶] Partial Ordered

		divorce, marry
اشتقاق (Derivation)	قید صفت	simply, simple Magnetic, magnetism

جدول ۱ - روابط معنایی در WordNet

مشابهت معناییِ مبتنی بر WordNet بصورت گسترده در پردازش زبانهای طبیعی (NLP)و بازیابی اطلاعات (IR) مورد بررسی قرار گرفته است.

روش‌های بسیاری برای محاسبه‌ی مشابهت معنایی بین دو کلمه و براساس WordNet ارائه شده است. معیارهای تشابه بر روی اسم‌ها و فعل‌ها بوده و نیز اکثراً بر روابط IS-A در WordNet اعمال شده‌اند. علت این امر آن است که نزدیک ۸۰ درصد از رابطه‌ها و لینک‌های بین مفاهیم را روابط ابرمفهوم/ زیر مفهوم تشکیل می‌دهند. با این‌حال به هنگام بررسییک رابطه معنایی در سطح مفاهیم، چندین نوع رابطه‌ی بالقوه را می‌توان متصور شد: مترادف، رابطه‌ی ابرمفهمومی/ زیرمفهومی (IS-A)، جزییت/شمول (Part of)، علت و معلولی،Material-Product، Event-Role و... . در این میان سه رابطه‌ی اول سهم بزرگتری از روابط بین مفاهیم را تشکیل می‌دهند.در ضمن روابط ویژگی‌های سلسله‌مراتبی برای صفات و قیود موجود نمی‌باشد. روش‌های تشابه معنایی به چهار دسته‌ی اصلی طبقه‌بندی می‌شوند:

- روشهای مبتنی بر شمارش یالهای مسیر: برای اندازه‌گیری تشابه معنایی بین دو کلمه، فاصله‌ی (یالهای مسیر بین دو کلمه در گراف) بین موقعیت دو کلمه را در درخت سلسله‌مراتب کلمات، اندازه‌گیری می‌شود. ایده‌ی این روشها از مدل حافظه‌ی معنایی کوئیلیان و براساس یافتنکوتاهترین فاصله بین مفاهیم در شبکه‌ی سلسله‌مراتبی، نشات گرفته است. عملکرد اصلی این روشها چنان است که هرچه فاصله‌ییک گره تا گره‌ی دیگر کمتر باشد، آنگاه کلمات متناظر با آن دو گره به یکدیگر شبیه‌تر هستند [۹۸,LEA۰۴,SU۹۴WU]. یکی از مشکلاتی که روشهای مبتنی بر شمارشیالها در محاسبه‌ی میزان شباهت مفاهیم دارند، این است که در این روشها وزن‌دهییکسان به تمامییالها (لینکها) در تمامی گراف، مشابهت دو مفهوم را به صورت دقیق نمی‌رساند، چراکه هر دو مفهوم همسایه دارای فاصله‌ییکسانی هستند و این روش تفاوت بینیالهایی که بینمفاهیم وجود دارد، را نادیده می‌گیرد. در [۹۸JIA] و [۰۵YAN] فاکتورهای دیگری چون عمق گره‌ی مفاهیم، چگالی سلسله‌مراتب و نوع یال (لینک)، به منظور بهبود روش اندازه‌گیری مسیر، اضافه شده است. هم‌چنین در [۰۶GAN] و [۰۲ZHO] برای محاسبه‌ی تشابه بین مفاهیم به‌جای وزن‌دهی به یالها به گره‌ها(مفاهیم) وزن اختصاص داده شده است.
- روشهای آماری مبتنی بر اطلاعات. این روش تفاوت بین محتوای اطلاعاتی بین دو کلمه را به صورت تابعی از احتمال حضور آن دو کلمه در یک انبوه اسناد ، اندازه‌گیری

می‌کند[MIL۹۹RES][LIN۱۹۹۱][JIA۱۹۹۸] [RES۹۸] .[۹۹] یک روش آماری مبتنی بر اطلاعات ارائه داده است و بیان می‌دارد که هرچه اطلاعات مشترک بین دو مفهوم بیشتر باشد، آن دو مفهوم به یکدیگر شبیه‌تر هستند. رزنیک بیان می‌دارد که می‌توان میزان شباهت دو مفهوم را از روی محتوای اطلاعاتی نزدیکترین گره‌ی مشترک (NCN) و با استفاده از آمار رخداد که از انبوه متن استخراج می‌شود، بدست آورد. البته واضح است که یکی از مشکلات روش فوق این است که NCN برای تمامی جفت گره‌هایی که پدر مشترک دارند، یکسان است. در [۰۴SEC] از WordNet به عنوان منبع آماری برای محاسبه‌ی احتمال حضور کلمات استفاده شده است؛ هرچه کلمات عمومی‌تر باشند (در ساختار گراف آنتولوژی بالاتر باشند) و کلمات فرزند (زیرمفهوم) بیشتری داشته باشند، آنگاه این کلمات محتوای اطلاعاتی کمتری نسبت به کلمات جزئی‌تر (در ساختار گراف آنتولوژی پایین‌تر) و با کلمات فرزند کمتر دارند. این روش مستقل از انبوه اسناد بوده و همچنین تضمین می‌کند که محتوای اطلاعاتییک کلمه از محتوای اطلاعاتی فرزندانش کمتر می‌باشد. این قید برای تمامی روشهای مبتنی بر محتوای اطلاعاتی صدق می‌کند.

- روشهای مبتنی بر ویژگی‌ها. در این دسته از روش‌ها میزان شباهت بین دو کلمه به عنوان تابعی از ویژگی‌های (بطور مثال تعاریفGlossary کلمات) آن کلمات در WordNet و یا براساس رابطه‌ی آنها با کلمات مشابه در ساختار سلسله‌مراتب کلمات، محاسبه می‌شود. ویژگی‌های مشترک باعث افزایش میزان مشابهت شده و برعکس ویژگی‌های غیرمشترک باعث کاهش میزان مشابهت مفاهیم خواهد شد [۹۷TVE]. هم‌چنین در [۰۷LIU] یک روش محاسبه‌ی شباهت معنایی براساس چندین ویژگی ارائه شده است. یکی از خوبی‌های این روش آن است که بسیار به روش تمایز مفاهیم توسط انسان نزدیک است ولی از سویی دیگر در این روش‌ها به توصیف فراگیر و در سطح جزئی از مفاهیم، نیاز است.
- روشهای ترکیبی .این روشها از تلفیقی از متدهای فوق‌الذکر استفاده می‌کنند [ROD03][PET06] : مشابهت کلمات با انطباق‌دادن مترادف‌ها، همسایگان کلمات و ویژگی‌های کلمات صورت می‌گیرد. ویژگی‌های کلمات سپس به اجزا، توابع و مشخصات شناسایی شده و همانند آنچه که در [TVE97] صورت گرفته است، تطابق آنتولوژی صورت می‌گیرد.

۱-۷-۲-۲- روشهای مبتنی بر شمارش یالها

در اینجا فرض می‌کنیم که دو مفهوم (برای مثال M وB) رابطه‌ی ابرمفهوم/زیرمفهوم ندارند. در [WU94] مشابهت دو مفهوم براساس مفاهیم مشترک ونیز استفاده از مسیر محاسبه شده است(رابطه‌ی ۲-۱۱).

$$sim(C_1, C_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \qquad (۱۱-۲)$$

بطوری‌که C_3 پایینترین ابرمفهوم مشترک C_1 و C_2 می‌باشد. N_1 تعدادگره‌های مسیر C_1 تا C_3 می‌باشد. N_2 تعداد گره‌های مسیر C_2 تا C_3 می‌باشد. N_3 نیز تعداد گره‌های مسیر C_3 تا گره‌ی ریشه می‌باشد.

[RES95] گونه‌ای از روش مبتنی بر شمارش یال را ارائه کرده است، بطوری که توانسته است با تفاضل طول مسیر از ماکزیمم طول مسیر ممکن، فاصله را به میزان شباهت تبدیل کند(رابطه‌ی ۲-۱۲).

$$sim_{edge}(w_1, w_2) = (2 \times MAX) - [\min_{c_1,c_2} len(c_1, c_2)] \quad (۱۲-۲)$$

بطوری‌که $s(w_1)$و $s(w_2)$ بیانگر مجموعه‌ی مفاهیمی است که دربرگیرنده‌یتعاریف(Sense)w_1و w_2 هستند. c_1متعلق به مجموعه‌ی $s(w_1)$ و c_2متعلق به مجموعه‌ی $s(w_2)$می‌باشد. MAXماکزیمم عمق سلسله‌مراتب (آنتولوژی) می‌باشد، هم‌چنین $len(c_1, c_2)$ طول مسافت کوتاهترین مسیر از c_1به c_2 می‌باشد.

[LEA98] از روش رزنیک استفاده کرده و میزان مشابهت دو مفهوم را بر طبق رابطه‌ی زیر پیشنهاد داده است(رابطه‌ی ۲-۱۳):

$$Sim(w_i, w_j) = Max\left[-\log \frac{Dist(c_i, c_j)}{2 \times D}\right] = Max\left[\log 2D - \log Dist(c_i, c_j)\right] \quad (۱۳-۲)$$

به‌طوری که Dماکزیمم عمق در سلسله‌مراتب WordNet می‌باشد.[1]

روش دیگر متعلق به [SUS93] می‌باشد که براساس تمامی لینک‌های ممکن می‌باشد. برای هر رابطه‌ی r، وزن wرا برای رابطه‌ی ($c_i \xrightarrow[r]{} c_j$) در بازه‌ی $[min_r;max_r]$ تعریف می‌کنیم. این وزن براساس چگالی محلی که وابسته به تعداد رابطه‌های نوع rاست که از c_iخارج می‌شود، محاسبه می‌گردد(رابطه‌ی ۲-۱۴):

$$w(c_i \xrightarrow[r]{} c_j) = \max_r - \frac{\max_r - \min_r}{n_r(c_i)} \quad (۱۴-۲)$$

که$n_r(c_i)$تعداد رابطه‌ی نوع r می‌باشد که از c_iخارج می‌شود. سوسنا فاصله‌ی بین دو مفهوم همسایه را تعریف می‌کند. این لینک در ارتباط با روابط r و معکوسهای آنها یعنی r'می‌باشد(رابطه‌ی ۲-۱۵):

$$dist(c_i, c_j) = \frac{w(c_i \xrightarrow[r]{} c_j) + w(c_i \xrightarrow[r']{} c_j)}{2 \times \max[len(c_i, root), len(c_j, root)]} \quad (۱۵-۲)$$

رابطه‌ی بالا تنها فاصله‌ی بین دو گره‌ی مجاور را در آنتولوژی محاسبه می‌کند. برای محاسبه‌ی فاصله‌ی بین دو مفهوم، بایستی فاصله‌ی بین تمامی مفاهیمی که در سرراه کوتاهترین مسیر بین دو مفهوم هستند، را باهم جمع کرد. در این روش فاصله‌ی بین دو مفهوم به سه پارامتر وابسته می‌باشد: کوتاهترین فاصله‌ی بین c_i و $c_j(p_1)$ ، چگالی مفاهیم در همین مسیر (p_2) و کوتاهترین مسیر بین ریشه تا کمترین پدر مشترکِ c_iو

---

[1] اگر ما فرض کنیم که همه‌ی مفاهیم در WordNet یک ریشه‌ی مشترک دارند، آنگاه *D*برابر ۱۶ خواهد بود.

$c_j(LCS)$ (p_3).

۲-۷-۲-۲- روشهای آماری مبتنی بر اطلاعات

رزنیک[۲] یک روش آماری مبتنی بر اطلاعات ارائه کرده است [RES99]. درابتدا احتمال وجود مفاهیم (در یک انبوه متن بزرگ) در سلسله‌مراتب محاسبه می‌شود، سپس از تئوری اطلاعات[۳] استفاده می‌شود. تئوری اطلاعات بیان می‌دارد که محتوای اطلاعاتی هر مفهوم برابر است با منفی لگاریتم احتمال حضور مفهوم در انبوه متن ℓ .مجموعه‌ی مفاهیم در ساختار سلسله مراتبی می‌باشد. تشابه دو مفهوم به اندازه‌ی اطلاعات محتوای مفهوم خاصی است که هردوی این مفاهیم به صورت مستقیم یا غیرمستقیم فرزندان آن هستند. تابع Pرا تعریف می‌کنیم: $P:\ell\to[0,1]$ به طوری‌که به ازای هر $c\in\ell$ ، $P(c)$ احتمال مواجه شدن با مفهوم cباشد. اگر در ساختار آنتولوژی تنها یک گره‌ی بالایی داشته باشیم، آنگاه به آن احتمال ۱ می‌دهیم. محتوای اطلاعاتی مفهوم c برابر است با $-\log P(c)$. بنابراین:

$$sim(c_1,c_2)=\max_{c\in S(c_1,c_2)}[-\log P(c)] \quad (۲-۱۶)$$

به‌طوری که :

$$P(c)=\frac{\text{تعداد مرتبه‌ای که مفهوم } c \text{ در انبوه متن ظاهر شده}}{\text{تعداد کل انبوه}} \quad (۲-۱۷)$$

و $s(w_1)$و $s(w_2)$ بیانگر مجموعه‌ی مفاهیمی است که دربرگیرنده‌یتعاریفw_1(Sense)و w_2 هستند. c_1متعلق به مجموعه‌ی $s(w_1)$ و c_2متعلق به مجموعه‌ی $s(w_2)$ می‌باشد.

[LIN98] روش رزنیک را گرفته و آنرا بهبود داده است. او مشابهت دو مفهوم را به‌صورت نسبت مقدار اطلاعاتی که نیاز است تا اشتراک بین دو مفهوم توصیف شود به مقدار اطلاعاتی که نیاز است تا هریک به صورت کامل توصیف شوند، تعریف کرده است. در واقع کاری که لین کرده است بدان معنی است که شباهت دو مفهوم تنها به NCN آن‌دو بستگی ندارد بلکه به محتوای اطلاعاتی خود آن دو نیز وابسته است (رابطه‌ی ۲-۱۸

$$sim(x_1,x_2)=\frac{2\times\log p(c_0)}{\log p(c_1)+\log p(c_2)} \quad (۲-۱۸)$$

Least Common Subsumer

Resnik

Information Theory

به‌طوری که $x_1 \in c_1$ و $x_2 \in c_2$. c_0 پائین‌ترین کلاس مشترکی است که هر دو مفهوم c_1 و c_2 از آن مشتق شده‌اند. ابزار انطباق آنتولوژی RiMOM[4] روش لین را پیاده سازی کرده است.

۳-۷-۲-۲- روشهای ترکیبی

جیانگ و کنارث[5]مدل ترکیبی ارائه کرده‌اندکه براساس مدل شمارش یالها بوده و از محتوای اطلاعاتی به عنوان یک فاکتور تصمیم استفاده کرده است [JIA97]. محتوای اطلاعاتی[6](IC) مفهوم cرا می‌توان با مقدار $-\log P(c)$ بیان داشت. قدرت لینک[7](LS) هر یال برابر است با تفاضل محتوای اطلاعاتی مفهوم فرزند و مفهوم پدر در ساختار سلسله‌مراتبی (رابطه‌ی ۲-۱۹).

$$LS(c_i, p) = -\log(P(c_i \mid p)) = IC(c_i) - IC(p) \qquad (۲-۱۹)$$

به طوری‌که مفهوم فرزند c_i زیرمجموعه‌ای از مفهوم پدر خود p می‌باشد.پس از درنظرگرفتن سایر فاکتورها ازجمله چگالی محلی، عمق گره‌ی مفهوم و نوع یال (لینک)، تابع فاصله به صورت رابطه‌ی ۲-۲۰ بیان می‌شود:

$$Dist(w_1, w_2) = IC(c_1) + IC(c_2) - 2 \times IC(LSuper(c_1, c_2)) \qquad (۲۰-۲)$$

به‌طوری‌که $LSuper(c_1, c_2)$ پائین‌ترین ابرمفهوم مشترک c_1 و c_2 می‌باشد.

رودریگز[8]روش دیگری برای تعیین نهادهای مشابه را براساس WordNet ارائه کرده است [ROD03]. برای مثال در آن، روابط زیرمفهوم/ابرمفهوم، جزئیت/شمولیت درنظر گرفته شده است. اندازه‌ی مشابهت براساس نرمال‌سازی مدل توِرسکی[9]و توابع اشتراک $|A \cap B|$وتفاضل $|A / B|$و برطبق رابطه‌ی ۲-۲۱ صورت می‌گیرد:

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)|A / B| + (1-\alpha(a,b))|B / A|} \qquad (۲-۲۱)$$

به طوری‌که aو bنهادهای کلاس بوده، Aو Bمجموعه‌ی توضیحات aو b (یعنی مجموعه‌ی مترادف‌ها، روابط IS-Aو یا Part-Whole)و α تابعی است که اهمیت نسبی مشخصات غیرمشترک را تعریف می‌کند. برای سلسله‌مراتب IS-A، αبرحسب عمق نهادهای کلاس تعریف می‌شود (رابطه‌ی ۲-۲۲).

---

[4] "*Risk Minimization based Ontology Mapping*": http://keg.cs.tsinghua.edu.cn/project/RiMOM/

[5] Jinag and Conarth

[6] Information Content

[7] Link Strength

[8] Rodriguez

[9] Tversky

$$\alpha(a,b)=\begin{cases}\dfrac{depth(a)}{depth(a)+depth(b)} & if\ depth(a)\le depth(b)\\ ۱-\dfrac{depth(a)}{depth(a)+depth(b)} & if\ depth(a)>depth(b)\end{cases} \quad (۲۲-۲)$$

پتراکیس و دیگران در [PET06] براساس روش رودریگز روش مشابهت X-Similarity را پیشنهاد داده‌اند که مبتنی بر Synset- ها و نیز مجموعه‌های توصیفی می‌باشد. تساوی ۱-۲۱ با میزان شباهت مجموعه‌ای S جایگزین می‌شود که در آن Aو Bدرواقع Synset ها و یا مجموعه‌های توصیفی کلمات هستند (رابطه‌ی ۲-۲۳).

$$S(a,b)=\max\frac{A\cap B}{A\cup B} \quad (۲-۲۳)$$

تشابه بین همسایگان کلمه $S_{neighborhoods}$ نیز به ازای هر نوع رابطه(برای مثال IS-A ویا Part-Of) محاسبه می‌شود (رابطه‌ی ۲-۲۴):

$$S_{neighborhoods}(a,b)=\max\frac{A_i\cap B_i}{A_i\cup B_i} \quad (۲-۲۴)$$

به‌طوری که iبیانگر نوع رابطه می‌باشد. نهایتاً:

$$Sim(a,b)=\begin{cases}۱ & if\ S_{synsets}(a,b)>۰\\ \max(S_{neighborhoods}(a,b),S_{descriptions}(a,b)) & if\ S_{synsets}(a,b)=۰\end{cases} \quad (۲-۲۵)$$

به‌طوری که $S_{descriptions}$ بیانگر تطابق مجموعه‌های توصیف کلمه می‌باشد. $S_{descriptions}$ و $S_{synsets}$ برطبق تساوی ۲-۲۴ محاسبه می‌شوند.

در [BAC04] از معیار جارو- وینکلر(JW)برای تجمیع WordNet یا EuroWordNet در فرآیند تطابق آنتولوژی استفاده کرده است. شباهت نام(NS)دو نام $N_۱$و $N_۲$از دو کلاس Aو B (هرنام مجموعه‌ای از توکن‌ها است، $N=\{n_i\}$) به‌صورت رابطه‌ی ۲-۲۶ تعریف می‌شود:

$$NS'(N_۱,N_۲)=\frac{\sum_{n_۱\in N_۱}MJW(n_۱,N'_۲)+MJW(n_۲,N'_۱)}{|N_۱|+|N_۲|} \quad (۲۶-۲)$$

به‌طوری‌که:

Jaro-Winkler

Name Similarity

$MJM(n_i, N) = \max_{n_j \in N} JW(n_i, n_j)$ ، $N_i' = N_i \cup \{n_k \mid \exists n_j \in N_i \cap n_k \in synset(n_j)\}$ و $NS'(A,B) = NS'(N_1, N_2)$.

synset(n_j) مجموعه‌ی مترادف‌های کلمه‌ی n_j است و

۴-۷-۲-۲- ارزیابی روش‌های تشابه معنایی مبتنی بر WordNet

WordNet-Similarity چندین متد محاسبه‌ی مشابهت مبتنی بر WordNet مانند [LEA98]،[JIA97]، [RES95]، [LIN98]، [HIR97]، [WU94] و [PAT03] در یک بسته‌ی Perl گنجانده است.

در [PET06] «سیستم تشابه معنایی» پیاده‌سازی شده و چندین اندازه‌ی تشابه معنایی شامل : ،[RAD89]، [WU94]، [LI03]،[LEA98]، [RIC94]، [RES99]، [LIN98]، [PWL02] ، [JIA97]، [PET06] و [ROD03] ارزیابی شده است. ارزیابی آنها در یک آنتولوژی براساس [MIL91] و در مقایسه با نتایج پرسش‌شده از انسان بوده است. هرچه نتایج یک روش همبستگی بیشتری با نتایج پرسیده‌شده از انسان داشته باشد، آن روش بهتر می‌باشد (یعنی یه نتایج قضاوت انسانی نزدیک‌تر است). همچنین آنها روشهای [ROD03] و X-Similarity[PET06] را در حالتی که مقایسه بین آنتولوژی‌های متفاوتی بوده است، مقایسه کرده‌اند. در بسته‌ی SimPack[SIM08] نیز چندین روش مانند [JIA97]،[LIN98] و [RES99] را پیاده‌سازی کرده است. این روش‌ها توسط [BUD06] ارزیابی شده‌اند.

۳-۲- مروری بر روش‌های خلاصه سازی

با توجه به حوزه کاربرد خلاصه سازی خودکار متن و اهمیت آن، تاکنون روش های بسیار زیادی برای خلاصه سازی متن ارائه شده است. از تکنیک های ساده ی آماری گرفته تا تکنیک های مبتنی بر هوش مصنوعی، از روشهای مبتنی بر زبان شناسی تا روش های مبتنی بر ریاضیات، همه و همه در خلاصه سازی متن بکار رفته اند. بر همین اساس دسته بندی های مختلفی برای روش های خلاصه سازی متن ارائه شده است. در این پایان‌نامه به یک دسته بندیها اشاره کرده و به بعضی از این روش ها صورت مختصر اشاره می کنیم. بر اساس دسته بندی ارائه شده توسط ؟؟؟، روش های خلاصه سازی متن به سه دسته عمده زیر تقسیم بندی می شوند:

- روش های مبتنی بر تکنیک های آماری
- روش های مبتنی بر تکنیک های معنایی سطوح بالاتر
- روش های مبتنی بر تکنیک های هوش مصنوعی

در ادامه به شرح مختصر هر یک از این دسته ها و روش های موجود در این دسته ها می پردازیم.

۱-۳-۲- روش های خلاصه سازی آماری

روشهای آماری جزء اولین روشهایی هستند که در خلاصه سازی متن استفاده شده اند. در این روش ها عموما سعی می شود با انجام محاسباتی به وزن داده می شود. سپس جملاتی که بیشترین وزن را دارند به عنوان خلاصه برگردانده می شوند. در ادامه به برخی از این روش ها اشاره شده است.

۱-۱-۳-۲- روش فرکانس کلمه

این روش برای اولین بار توسط آقای Luhn در سال ۵۸ مورد استفاده قرار گرفت[Luh 58] و جزء اولین روش هایی است که برای خلاصه سازی استفاده شده است و البته هنوز هم به صورت ترکیبی با سایر روش ها استفاده می شود. در این روش در ابتدا با بهره گیری از یکی از روش های وزن دهی به کلمات نظیر Tf-Idf وزن داده می شود. سپس این وزن در واحد جمله توزیع می شود. در ادامه هم با توجه به میزان فشرده سازی، جملاتی که بیشترین وزن را دارند برای خلاصه انتخاب می شوند. در ادامه به برخی از روش های وزن دهی به کلمات اشاره شده است. ادعای این روش این است که کلماتی که در متن بسیار پرکاربرد هستند، کلماتی هستند که به موضوع اشاره می کنند. این روش عموما با حذف کلمات ایست و ریشه یابی انجام می شود. در جدول(۲) تعدادی از روش های وزن دهی که مورد استفاده قرار می گیرده آورده شده است:

Method	Local weight	Global weight
TF-IDF	$f_{i,j}$	$\log \frac{N}{n_i}$
Log-IDF	$1+\log f_{i,j}$ if $f_{i,j}>0$ 0 if $f_{i,j}=0$	$\log \frac{N}{n_i}$
GF-IDF	$\frac{gfi}{ni}$	$\log \frac{N}{n_i}$
No-weight	$f_{i,j}$	None

جدول ۲- تعدادی از روش های وزن دهی کلمات

مشکل این روش این هست که لزوما کلمات پر کاربرد، کلمات مهم نبوده و همچنین عدم توجه به کلمات هم خانواده، مشکل هم آوایی ها و همچنین خوانایی و پیوستگی متن از دیگر مشکلات این روش می باشد.

۲-۱-۳-۲- روش مبتنی بر موقعیت یا جایگاه

ایده اصلی این روش این است که جملات مهم متن همواره در جاهای خاص و مشخصی از متن قرار می گیرد[Edm 69][Hov 97]. اینکه کدام قسمت از متن دارای اهمیت بیشتری می باشد بستگی به نوع متن دارد. به عنوان مثال برای متون خبری انگلیسی ثابت شده است که در اکثر موارد اولین پاراگراف مهمترین

پاراگراف و اولین جمله مهمترین جمله می باشد. یا برای متون علمی و مقالات، مشخصا قسمت چکیده و نتیجه گیری همواره مهمترین قسمت مقاله می باشد. بر همین اساس به جملات بر اساس موقعیتشان در پاراگراف وزن می گیرند. پاراگراف ها هم براساس موقعیتشان در متن وزن می گیرند و نهایتا جملات براساس امتیازشان به عنوان خلاصه انتخاب می شوند. سیستم swesum[Swe 00]که یک سیستم چندزبانه برای خلاصه سازی اخبار می باشد، بر همین اساس عمل می کند. لازم به ذکر است این قانون برای تمامی زبان ها درست نمی باشد. به عنوان مثال با آزمایشات ساده ای که با همکاری گروه زبان شناسی دانشگاه فردوسی مشهد انجام شد، مشخص شد که برای متون خبری نوشته شده به زبان فارسی این قانون صادق نمی باشد. به عبارت دیگر در اکثر موارد در متون خبری نوشته شده به زبان فارسی، اولین جمله مهمترین جمله نمی باشد. البته این موضوع مستقل از ویژگی های زبان فارسی بوده و بیشتر به سبک نوشتاری خبرگزاری های فارسی زبان وابسته می باشد.

۳-۱-۳-۲- روش مبتنی بر عنوان

این روش هم جزء اولین روش های خلاصه سازی متن می باشد[Edm 69] و ایده اصلی آن این است که همواره موضوع و عنوان متن، بیانگر محتوای آن می باشد و بر اساس همین اصل، جملات مهم با توجه به عنوان متن باید انتخاب شوند. با توجه به این اصل، در این روش ها سعی می شود از کلمات موضوع متن به عنوان کلمات کلیدی استفاده شود و از آنها برای استخراج جملات مهم استفاده شود.مشکل اصلی این روش این است که بسیاری از متون فاقد عنوان بوده و یا اینکه امکان استخراج عنوانشان به صورت خودکار فراهم نمی باشد. همچنین ممکن است عنوانشان بسیار کلی باشد مثل عنوان سرفصل های یک کتاب که به صورت کلی به محتوای آن اشاره می کند. با این وجود لازم به ذکر است که این روش ساده و بدیهی اساس و اصل بسیاری از روش های امروزی برای خلاصه سازی متن می باشند. در بسیاری از روش های جدید سعی می شود تا در ابتدا موضوعات یا زیر موضوعات پنهان موجود در متن استخراج شده و سپس بر اساس آنها جملات مهم انتخاب شوند. از جمله این روش ها می توان به روش های مبتنی بر مدل موضوع[۱] اشاره نمود. در ادامه سعی می شود به این روش ها اشاره مختصری شود.

۴-۱-۳-۲- روش مبتنی بر عبارات اشارت

بحث اصلی در این روش، عبارات خاصی هست که در تمامی زبان ها وجود دارند و بر اهمیت جمله های بعدشان می افزایند. به این عبارات، عبارات اشاره یا "cue phrase" می گویند. از جمله این کلمات یا عبارات برای زبان انگلیسی به عبارات "The main aim" ، "as result" و "In this report"می توان اشاره کرد.این

---

۱ Topic model summarizaion

عبارات در تمامی زبان ها موجود می باشند. در زبان فارسی هم عباراتی نظیر "در این گـزارش" ، "بـه عنـوان نتیجه"، "هدف از" ،"به عنوان خلاصه"و"در پایان" جزء عبـارات اشـاره مـی باشـند. در زیـر برخـی از ایـن کاربردها برای زبان انگلیسی نشان داده شده است:

"The main aim of the present paper is to describe…" (IND)
"The purpose of this article is to review…" (IND)
"In this report, we outline…" (IND)
"Our investigation has shown that…" (INF)

در این دسته از روشهای خلاصه سازی متن، در ابتدا باید کلیه این عبارات و کلمـات اسـتخراج شـوند. سـپس باید گرامری که این کلمات و عبارات در آن صدق می کنند استخراج شوند.آنگاه برای خلاصه سازی یک مـتن، آن را تجزیه کرده و سپس بررسی می کنیم که آیا در گرامر از پـیش تعیـین شـده صـدق مـی کنـد یـا خیـر. جملاتی که دارای عبارات اشاره بوده و در گرامر صدق کننـد وزن بیشـتری بـه خـود اختصـاص خواهنـد داد. جملاتی که بیشترین وزن را کسب کنند نهایتا به عنوان جملات برتر برای خلاصه انتخـاب مـی شـوند. آقـای لمن در سال ۹۷ این روش را برای زبان فرانسه پیاده سـازی کـرد. وی در ابتـدا دیکشـنری<cue,weight> را تشکیل داد و سپس گرامر این زبان را برای زبان فرانسه استخراج کرد و بـه جمـلات بـر اسـاس دیکشـنری از پیش تعریف شده وزن اختصاص داد.

از این روش در بسیاری از سیستم های خلاصه سازی امروزی به صورت ترکیبی با سایر روش ها اسـتفاده مـی شود.

۵-۱-۳-۲- روش ترکیبی

روش دیگری که می توان پیشنهاد کرد، ترکیب خطی این چهار روش می باشد. به این ترتیب که به هرکدام از روش های بالا وزنی اختصاص داده می شود. این وزن هم معمولا با استفاده از یـادگیری ماشـینی بدسـت مـی آید. آقای ادمونسون این روش ها را با هم ترکیب کرده انـد و بهتـرین حـالتی کـه گـزارش دادنـد اسـتفاده از ترکیب عبارات اشاره + عنوان + جایگاه می باشد.

۶-۱-۳-۲- روش مبتنی بر دسته بندی کننده های بیزین

در این روش ها از قانون احتمال شرطی اولیه بیزین برای خلاصه سازی استفاده می شود[Kup 97]. بر همـین اساس احتمال اینکه جمله ی s در خلاصه S قرار بگیرد با استفاده از فرمول زیر محاسبه می گردد :

$$P(s \in S \mid F_1, F_2, ...F_k) = \frac{P(F_1, F_2, ...F_k \mid s \in S)P(s \in S)}{P(F_1, F_2, ...F_k)}$$

که در آن $F_1, F_2, \ldots, F_k$ ویژگی ها ۱ تا k بوده و می توانند به عنوان مثال هر کدام از روش های قبلی باشند. بر اساس قانون بیزین احتمال اینکه جمله s در خلاصه S قرار بگیرد با فرض اینکه ویژگی های $F_1, F_2, \ldots, F_k$ برقرار باشد به صورت زیر محاسبه می گردد:

$$P(s \in S \mid F_1, F_2, ...F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

که همان قانون احتمال شرطی بیزین می باشد. پس از محاسبه این احتمال برای تمامی جملات، جمله هایی که بیشترین احتمال را داشته باشند در خلاصه قرار می گیرند.

۷-۱-۳-۲- روش مبتنی بر زنجیره های مارکوف

زنجیره مارکوف، دنباله‌ای از متغیرهای تصادفی است که همگی این متغیرهای تصادفی دارای فضای نمونه‌ای یکسان هستند اما، توزیع احتمالات آنها می‌تواند متفاوت باشد و در ضمن هر متغیر تصادفی در یک زنجیره مارکوف تنها به متغیر قبل از خود وابسته است. از زنجیره های مارکوف در علوم مختلفی استفاده شده است. در مدیریت و برنامه ریزی و بطور کلی مسائل تصمیم گیری، در شبکه های کامپیوتری و همچنین در بحث خلاصه سازی متن از زنجیره مارکوف استفاده شده است.

اولین بار آقای کونری از زنجیره های مارکوف در خلاصه سازی متن استفاده کرده است[Con 01]. وی با استفاده ازیکسری ویژگی های داده شده، احتمال حضور هر جمله در خلاصه را محاسبه کرده است. در این مدل نسب به روش بیزین[Kup 97]، فرض کمتری برای استقلال متغیرهای تصادفی شده است. بعبارت بهتر در مدل مارکوف برای خلاصه سازی، فرض نمی شود که احتمال حضور جمله i ام در خلاصه نهایی مستقل از حضور جمله i-1 در خلاصه باشد (برخلاف روش بیزین). ویژگی هایی که در روش کونری برای توسعه خلاصه سازی مبتنی بر زنجیره های مارکوف ارائه شده است شامل موارد زیر می باشد:

- موقعیت جملات در سند
- تعداد کلمات موجود در جملات
- میزان شباهت کلمات جمله

بر همین اساس در این مدل انتظار می رود که احتمال حضور جمله بعدی در خلاصه بسته به اینکه جمله فعلی در خلاصه حضور داشته باشد یا خیر متفاوت خواهد بود. مدل زنجیره های ماکوف وابستگی موقعیتی، وابستگی ویژگی ها و مارکوفیتی را محاسبه می کند.

۲-۳-۲- روش های خلاصه سازی مبتنی بر تکنیک های هوش مصنوعی

با توجه به اهمیت موضوع خلاصه سازی متن، تقریبا اکثر تکنیک های هوش مصنوعی در خلاصه سازی بکار برده شده است[You 85][Sch 77] [Mck 95][Gra 81][Dej 79][Azz 99]. در مقالات بسیار زیادی شاهد استفاده از روش های فازی برای خلاصه سازی هستیم[Kia 06][Kyo 08][Sua 09]. شبکه های عصبی[Svo Kia 04][07] هم در مقالات زیادی استفاده شده است. الگوریتم ژنتیک هم در خلاصه سازی زیادی استفاده شده است[Qaz 08][Sil 04] و البته ترکیب این روش ها با همدیگر را هم در مقالات زیادی می توان مشاهده کرد. به هرحال با توجه به اینکه تعداد این روش ها بسیار زیاد و خارج از حوصله این بحث می باشد بنابراین از ذکر جزییات آنها خودداری می کنیم.

۳-۳-۲- روش های خلاصه سازی مبتنی بر روش های معنایی سطوح بالاتر

این دسته از روش ها برای غلبه بر مشکلات روش های پیشین معرفی شدند و سعی دارند تا با فهم عمیق تر متن و استخراج روابط معنایی پنهان موجود در متن، جملات مهم را با دقت بیشتری انتخاب نمایند. از جمله این روش ها می توان به روش مبتنی بر ساختارهای معنایی/ نحوی سطح بالا، روشهای مبتنی بر شبکه یا گراف و روش های مبتنی بر آنالیز روابط معنایی پنهان موجود در متن اشاره نمود.در ادامه به هریک از این دسته از روشها اشاره می کنیم.

۱-۳-۳-۲- روش مبتنی بر ساختارهای معنایی/ نحوی سطح بالا

ادعای این دسته از روش ها این است که جملات/ پاراگراف های مهم، نهاهای مرتب به همدیگر در ساختار معنایی متن می باشند. این روش را می توان به دسته های زیر تقسیم بندی نمود :

- روش مبتنی بر شباهت لغوی (زنجیره های لغوی ، مبتنی بر شباهت شبکه واژگان)
- روش مبتنی بر استخراج کلمات هم رخداد
- روش مبتنی بر کلمات هم ارجاع
- روش ترکیبی

روش مبتنی بر شباهت لغوی : در قسمت معرفی شبکه واژگان، اشاره جامع و کافی ایی به روشهای محاسبه شباهت معنایی با استفاده از شبکه واژگان برای دو کلمه ارائه کردیم. تمامی این روش ها را می توان برای خلاصه سازی متن هم مورد استفاده قرار داد. کافی است تا این شباهت معنایی بین کلمات را برروی

جملات بسط داده و شباهت بین جملات را محاسبه کرد. یعنی برای محاسبه شباهت بین دو جمله مثلا می توان شباهت معنایی مبتنی بر شبکه واژگان بین تمامی کلمات دو جمله را محاسبه کرد و پس از محاسبه شباهت بین جملات به صورت ترکیبی از سایر روش ها برای انتخاب بهترین جمله ها استفاده نمود.

روش زنجیره های لغوی از جمله روش های معروف در خلاصه سازی متن می باشد که توسط افراد مختلفی مورد استفاده قرار گرفته است[???]. برای توضیح بیشتر این روش در ابتدا تعاریف زیر را شرح می دهیم:

پیوستگی لغوی[۱](تعریف ۱) :پیوستگی لغوی را می توان در قالب شباهت مبتنی بر شبکه واژگان تعریف کرد. پیوستگی لغوی میزان پیوستگی و ارتباط بین دو کلمه را مشخص می کند و شامل دو دسته اصلی می شود:

- تصریحی[۲]: که شامل مواردی نظیر هم خانواده ها،[۳] متضادها،[۴] شمول[۵] و ... می باشد.
- هم رخدادی ها[۶]: که شامل کلماتی است که با هم در متن رخ می دهند نظیر کلمات "معلم" و "مدرسه" در جمله "او به عنوان معلم درمدرسه کار می کند".

زنجیره لغوی[۷](تعریف ۲) :زنجیره لغوی به توالی ای از کلمات اتلاق می گردد که با یکدیگر پیوستگی لغوی دارند.

برای خلاصه سازی با استفاده از زنجیره های لغوی گام های زیر انجام می شود:

- در ابتدا کلمات کاندید انتخاب می شود. عموما در مقالات اسامی را به عنوان کلمات کاندید در نظر می گیرند[???]. بنابراین در ابتدا در فاز پیش پردازش از عملیات برچسب جزئی سخن[۸]استفاده می شود تا نوع کلمات مشخص شود. سپس اسامی به عنوان مجموعه کلمات کاندید در نظر گرفته می شوند.
- برای هر یک از کلمات کاندید زنجیره مناسب را بر اساس میزان ارتباط بین کلمات زنجیر و کلمه کاندید یافته می شود.
- اگر چنین زنجیری پیدا شد که کلمه کاندید را در آن قرار می گیرد، در غیر اینصورت کلمه در یک زنجیر جدید قرار می گیرد.

---

۱ Lexical cohesion

۲ Reiteration

۳ Synonym

۴ Antonym

۵ Hyperonym

۶ Co occurance

۷ Lexical chain

۸ Part of speech tagging

- سپس با استفاده از روش هایی تعریف شده قوی ترین زنجیرها[۱]را استخراج می گردد.
- در ادامه برای تولید خلاصه از راه حل های مختلفی می توان استفاده نمود. به سه راه حل در زیر اشاره شده است.
  - یک راه حل این است که اولین جمله ای که یکی از کلمات یکی از زنجیره های قوی را دارد به عنوان نماینده آن زنجیر انتخاب شود. این روش مشکلاتی دارد که در مثال زیر به آن اشاره شده است:

Chain: AI=2 ; Artificial Intelligence =1 ; Field=7 ; Technology=1 ; Science=1

همانطور که در تصویر هم مشخص است کلمه Science دارای تعداد تکرار کمتری نسبت به کلمات دیگر می باشد و صرف این که یک جمله این کلمه را داشته باشد نمی تواند معیار مناسبی برای انتخاب آن جمله به عنوان نماینده زنجیر باشد.

  - راه حل دوم این است که برای تولید خلاصه، اولین جمله ای را انتخاب نماییم که شامل کلمه نماینده زنجیر (مثلا کلمه ای که دارای فرکانس بالایی هست) باشد.
  - راه حل سوم تعیین قسمت هایی از متن می باشد که چگالی زنجیره (نسبت کلمات زنجیره در متن) در آنجا زیادتر می باشد.

در تصویر زیر یک نمونه متن و زنجیره های تشخیص داده شده آورده شده است.

Mr. Kenny is the**person** that invented the anesthetic **machine** which uses **micro-computers** to control the rate at which an anesthetic is pumped into the blood. Such **machines** are nothing new. But his **device** uses two **micro-computers** to achineve much closer monitoring of the **pump** feeding the anesthetic into the patient.

شکل ۳- نمونه ای از زنجیره های لغوی

[۱] Strong chain

همانطور که در شکل (۳) مشاهده می شود"person" و "Mr. Kenny" به درستی در یک زنجیره و"machine"، "micro-computers"، "machines"، "device"، "micro-computers" و "pump" در یک زنجیره قرار گرفته اند.

روش مبتنی بر استخراج کلمات هم رخداد: کامل شود.

روش مبتنی بر کلمات هم ارجاع : کامل شود.

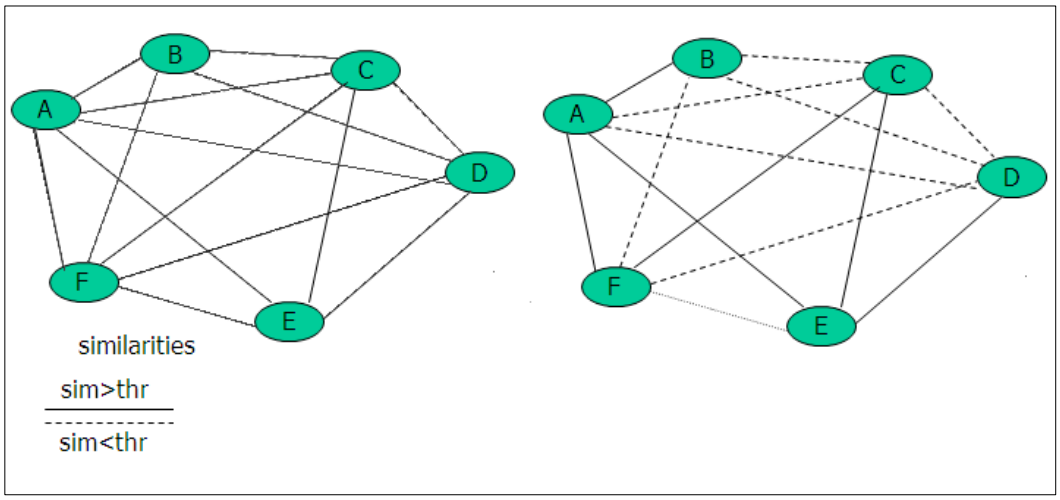
روش ترکیبی : کامل شود.

۲-۳-۳-۲- روشهای مبتنی بر شبکه یا گراف

در این دسته از روشها، واحدهای متن نظیر جمله و یا پاراگراف به صورت برداری از کلمات وزن دار $D_i = (d_{i1}, d_{i2}, \dots, d_{in})$ نمایش داده می شوند. سپس شباهت بین جملات با استفاده از روش های اندازه گیری شباهت برداری نظیر ضرب داخلی یا فاصلی کسینوسی محاسبه می گردد:

$$sim(D_i, D_j) = \sum d_{ik}.d_{jk}$$

برای ساخت گراف وزن دار، واحدهای متن نظیر جمله یا پاراگراف به عنوان گره های گراف در نظر گرفته می شود. برای وزن دهی به یال های بین دو گره ji , از میزان شباهت بین دو گره یا همان $sim\ (D_i, D_j)$استفاده می شود.پس از ساخت گراف اولیه، یال هایی که اهمیت کمتری دارند حذف می شوند. برای همین منظور یال هایی که وزنشان از مقدار حد آستانه α (که توسط الگوریتم تعیین می شود) کمتر باشند حذف شده و گراف نهایی ساخته می شود (شکل ۴).



شکل ۴- نمایش گراف جملات

در ادامه خلاصه از روی گراف نهای ساخته می شود. برای ساخت خلاصه از روی گراف هم راهکارهای مختلفی پیشنهاد شده است که در ذیل به چند نمونه از آنها اشاره می کنیم:

- روش انبوه ترین مسیر:[۱] در این روش برای انتخاب جملات مهم از گره ای شروع می شود که بیشتری ارتباط یا لینک را با سایر گره ها دارد. سپس گره دومی که بیشترین ارتباط را با سایر گره ها دارد به عنوان جمله یا پاراگراف دوم به خلاصه اضافه می شود. این عملیات آنقدر ادامه پیدا می کند که میزان خلاصه مورد نظر تولید شود. سپس جملات بر اساس موقعیتشان در متن دوباره مرتب می شوند. در حقیقت ایده اصلی این روش این است که واحد هایی از متن مهم و برجسته هستند که باسایر قسمت ها ارتباطات زیادی دارند.
- روش اول عمق:[۲] در این مدل سعی شده است تا جملات خلاصه ارتباطات بالایی داشته باشند و علاوه بر آن خلاصه انتخاب شده دارای پیوستگی مناسبی هم باشد. به همین دلیل برای انتخاب بهترین جمله ها، در ابتدا گره ای که بیشترین لینک را دارد انتخاب می شود. سپس از بین گره هایی که به گره انتخاب شده متصل هستند، گره ای که بیشترین لینک را دارد به عنوان جمله دوم انتخاب می شود و به همین ترتیب الگوریتم به صورت اول – عمق جلو می رود. به عنوان مثال در تصویر ؟؟؟ ابتدا A انتخاب می شود چون دارای ۳ لینک بوده و بیشترین لینک را دارد. سپس از بین گره هایی که به A متصل هستند یعنی گره های B، E و F گره ایی که بیشترین لینک را دارد (گره E) انتخاب می شود و به همین ترتیب عملیات دنبال می شود.

با توجه به ویژگی های روش های مبتنی بر گراف، مسلما این دسته از روش های خلاصه سازی را به راحتی می توان با سایر روش های خلاصه سازی ادغام کرده و نتایج مناسبی حاصل نمود. مثلا بجای نمایش برداری جملات و استفاده از فاصله کسینوسی برای وزن دهی به یال ها، می توان از شباهت های مبتنی بر زنجیره لغوی استفاده نمود و سایر عملیات را دنبال کرد.[؟؟؟]مقالات زیادی منتشر شده اند که در آنها از روش های گرافی به صورت ترکیبی با سایر روش ها استفاده شده است.

۳-۳-۳-۲- روش های مبتنی بر آنالیز روابط معنایی پنهان در متن

در سال‌های اخیر مقالات زیادی از LSA[۳][Yu 09][Yeh 05][Gon 01][Ste 05]برای خلاصه‌سازی استفاده کرده‌اند. LSAبرای حل مشکل تنک[۴]بودن داده‌ها ارائه شده است[Dee 90]. این روش با نمایش داده‌ها در فضای معنایی کوچکتر و در حقیقت کاهش ابعاد، تا حد زیادی مشکل داده‌های تنک را حل می کند. این

---

۱ Bushy path

۲ Depth first path

۳ Latent semantic analysiss

4 -Sparse

کاهش ابعاد منجر به بهبود کارایی در بسیاری از کاربردها شده است.[Dee 90][Pap 00] در تمامی روش‌هایی که از LSA برای خلاصه سازی استفاده می کنند بردار جملات را بااستفاده از اعمال SVD[1]برروی ماتریس کلمه- جمله می‌سازند. LSA جملات را مستقیما با استفاده از این ماتریس و با توجه به ویژگی‌های معنایی آنها مرتب و دسته‌بندی می کند

- اولین بار Gong در سال ۲۰۰۱ از LSA در خلاصه‌سازی استفاده نمود[Gon 01] وی در ابتدا ماتریس کلمه-جمله را با محاسبه معیارTFISF تشکیل داد. TFISF معیار محاسبه وزن در واحد جمله بوده و به صورت زیر محاسبه می شود

$$(TF-ISF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) Log_2\left(\frac{|S|}{|\{s: t_i \in s\}|}\right)$$

$tf_{i,j}$ فرکانس نسبی کلمه i در جمله j و $|S|$ تعداد جملات موجود در پیکره می باشد. پس از محاسبه معیار $TFISF$ ماتریس کلمه-جمله تشکیل داده شده و سپس با بهره‌گیری از LSAمهمترین جملات استخراج می شود. مشکل اصلی این روش این است که معیارهای وزن دهی نظیر $TFISF$ قادر به بیان مفاهیم اصلی پنهان از دیدگاه کلی نمی باشد چراکه در معیار وزن دهی $TFISF$، فرکانس کلمات در سطح جمله مورد بررسی قرار می گیرد و تاثیر جملات بر یکدیگر در سطوح بالاتر لحاظ نمی شود.

- آقای Steinberger [Ste 05] با دخیل کردن روش ادغام ضمایر[2]در LSA، کارایی روش Gong را افزایش داد. ایده آقای Steinberger ادغام اطلاعات لغوی با روش SVD بود. او در این مقاله دو روش جانشینی و روش افزایشی را پیشنهاد کرد. در روش اول از ادغام ضمایر به عنوان یکی از مراحل پیش پردازشی استفاده می شود. یعنی در ابتدا ضمایر با کلمات اصلی جایگزین شده و سپس ماتریس ورودی SVD برروی متن اصلاح شده ساخته می شود. البته نتایج بررسی های منتشر شده در [Ste 05] بیانگر این است که این روش نه تنها موجب افزایش دقت چندانی در کارایی نمی شود، بلکه در مواردی هم باعث کاهش دقت می شود. در روش دوم یا همان روش افزایشی، به ماتریس ورودی SVD یک بخش دیگر که شامل اطلاعات لغوی مربوط به زنجیره های آنافوریک می باشد هم اضافه می شود.وی در این مقاله ثابت کرد که ادغام ضمایر با استفاده از روش افزایشی، دقت خلاصه-سازی روش Gong را افزایش می دهد.

---

[1] - Singular value decomposition

[2] Anaphora resolution

- Yeh[Yeh 05]در سال ۲۰۰۵، از ترکیب نگاشت رابطه ای متن و آنالیز روابط معنایی پنهان در متن بری تولید خلاصه استفاده کرد. وی با استفاده توامان از LSA و T.R.M(نگاشت رابطه ای متن) ساختار معنایی پنهان موجود در سند را استخراج کرده و خلاصه را بر اساس این ساختار پنهانی استخراج شده تولید نمود. این ترکیب منجر به افزایش دقت در تولید خلاصه شده است.
- Yu[Yu 09] در سال ۲۰۰۹ روشی برای خلاصه‌سازی اخبار مبتنی بر LSA ارائه نمود. وی در ابتدا با استفاده از LSA شباهت معنایی بین جملات را محاسبه کرده و سپس با استفاده از روش خوشه بندی مبتنی بر روش k-means جملات را دسته‌بندی کرده و سپس از هر کلاستر، جمله‌ای که بیشترین شباهت را به موضوع داشته باشد به عنوان نماینده آن خوشه بر می‌گرداند. وی در حقیقت این روش را برای بهبود یکی از مشکلات روش Gong که در ادامه به آن اشاره می کنیم ارائه کرده است.

در تمامی این روش‌ها با محوریت قرار دادن جمله و ساخت ماتریس کلمه- جمله، از LSA استفاده شده است. ایراد تمامی این روش‌ها عدم توجه به مفاهیم اصلی پنهان درکل سند می‌باشد زیرا مبنای کار این روش ها استفاده از ماتریس کلمه-جمله می باشد. و این در حالی است که یک خلاصه خوب در خلاصه‌سازی چند-سندی، باید ضمن بیان موضوع و زمینه اصلی موجود در اسناد، دیدگاههای مختلف مرتبط با زمینه را نشان دهد.

همچنین در روش Gong بین بردارهای مختلف ارزش یکسان قائل می شوند و برای هر بردار از ماتریس Vیک جمله به عنوان نماینده آن مفهوم انتخاب می شود و این در حالی است که ممکن است ارزش یک مفهون به قدری زیاد باشد که بایدچندین جمله از متن را به عنوان نماینده آن انتخاب کرد تا خلاصه تولید شده، جامعیت کافی را داشته باشد. از طرف دیگر ممکن است بعضی از مفاهیم بدست آمده آنقدر کم ارزش باشند که انتخاب جمله ای بعنوان نماینده آنها کار اشتباهی باشد.

۴-۳-۳-۲- روش های مبتنی بر آنتولوژی

کامل شود

۴-۲- روش های ارزیابی و مجموعه داده های استاندارد

۱-۴-۲- روش های ارزیابی خلاصه سازی

در ارائه یک سیستم خلاصه ساز مطلوب باید به موارد زیر توجه نماییم :

الف) اطلاع رسانی[۱]مناسب : سیستم خلاصه ساز باید قادر باشد تا حداکثر اطلاعات مفید در متون اصلی را منتقل نماید.

ب) رعایت اختصار : یک سیستم خوب باید قادر باشد تا کنه مطلب را با حداقل جملات منتقل نماید.

ج) پیوستگی[۳]مطالب : یکی از مشکلات سیستم های خلاصه ساز عدم توجه آنها به پیوستگی کلی مطالب انتخاب شده برای خلاصه می باشد. خلاصه های سیستم های خلاصه ساز باید از پیوستگی مناسبی برخوردار باشد.

د) خوانایی مطالب[۴]: در خلاصه های تولید شده باید تکلیف ضمایر مشخص شده و ارجاعات سرگردان وجود نداشته باشد.

روش‌های ارزیابی خلاصه‌سازی از یک دیدگاه کلی به دو مدل روشهای ارزیابی درونی[۵]و برونی[۶]تقسیم می شوند. در ادامه به معرفی این روشها پرداخته شده است.

۱-۱-۴-۲- روش ارزیابی بیرونی

روش های ارزیابی بیرونی بیشتر بر استفاده از خلاصه ها در تسک های خاص نظیر اجرای دستورات، بازیابی اطلاعات، پاسخ دهی به سوالات و ارزیابی ارتباط بین خلاصه و سند اصلی متمرکز می باشند.

تعدادی از روش های ارزیابی بیرونی عبارتند از :

الف) بازی سوال : هدف از بازی سوال، آزمایش میزان فهم خواننده از خلاصه و توانایی آن برای نقل وقایع کلیدی مقاله منبع است.این عمل ارزیابی در دو مرحله انجام می شود.ابتدا آزمایشگر مقاله های اصلی را می خواند و بخشهای مرکزی آن را علامت گذاری می کند. سپس از عبارات مهم بخشهای مرکزی متن، سوالاتی طرح می کند. در مرحله بعد، ارزیاب سوالات را سه مرتبه پاسخ می دهد؛ یکبار بدون مشاهده هیچ متنی (baseline 1)، یکبار پس از مشاهده یک خلاصه ساخته شده توسط سیستم و درانتها پس از مشاهده متن اصلی (baseline 2). خلاصه ایی که به خوبی وقایع کلیدی مقاله را نقل کرده باشد، باید قادر به پاسخگویی به بیشتر سوالات (با نزدیکتر بودن به baseline2 نسبت به baseline1) باشد.

ب) بازی دسته بندی : بازی دسته بندی با دسته بندی اسناد منبع (آزمایشگرها) و متون خلاصه (اطلاع دهنده ها)، سعی درمقایسه قابلیت دسته بندی آنها به یکی از N دسته دارد. سپس مطابقت دسته بندی

---

۱ Informativeness

2- Conciseness

3 - Coherency

4- Readability

۵ Intrinsic

۶ - Extrinsic

خلاصه ها به متون اصلی سنجیده می شود. یک خلاصه کاربردی باید در همان دسته ی سند منبع اش قرارگیرد.

ج) بازی Shannon : بازی Shannon که نوعی از معیارهای سنجش Shannon در تئوری اطلاعات است، تلاشی برای تعیین کیفیت محتوی اطلاعات بوسیله حدس لغت بعدی (حرف یا کلمه) می باشد، و به این ترتیب متن اصلی را مجددا ایجاد می کند. این ایده از معیارهای Shannon از تئوری اطلاعات اقتباس شده است، که در آنجا از سه گروه مخبر خواسته می شود قطعات مهم از مقاله منبع را (با مشاهده متن کامل، یک خلاصه تولید شده و یا حتی هیچ متنی) به صورت حرف به حرف یا کلمه به کلمه مجددا تولید کنند. سپس معیار حفظ اطلاعات با تعداد ضربه های کلیدی که برای ایجاد مجدد قطعه اصلی طول می کشدسنجیده می شود. Hovey و Marcu نشان دادند که اختلاف زیادی در این سه سطح (در حدود فاکتور ۱۰ در بین هر گروه) وجود دارد. مشکل روش Shannon این است که به فردی که عمل حدس زدن را انجام می دهد وابسته است و درنتیجه بطورضمنی مشروط به دانش خواننده است. معیار اطلاعات با دانش بیشتر از زبان و حوزه و ... کاهش می یابد.

۲-۱-۴-۲- روش ارزیابی درونی

تمرکز این دسته از روش های ارزیابی بیشتر برروی پیوستگی مطالب و میزان اطلاع رسانی آنها می باشد و معمولا شامل مقایسه کیفیت خلاصه های تولید شده توسط سیستم و خلاصه های مرجع(خلاصه های ایده آل که توسط انسان تولید شده اند) می باشد. روش های ارزیابی درونی انسانی، کیفیت اسناد خلاصه را با ارزیابی معیارهای دقت، سلاست و وضوح تعیین می نمایند. معیارهای ارزیابی درونی خودکار نظیر [Lin 03]ROUGEیا[Win 02] MEADEval از مقایسهی رشته ای برای ارزیابی خلاصه های استفاده می کنند.

با توجه به بررسی های انجام شده و بررسی ابزارهای مختلف، ابزار استاندارد ROUGE که پر کاربردترین معیار ارزیابی در خلاصه سازی متون می باشد استفاده شده است. به همین منظور به شرح مختصر این ابزار می پردازیم.

۳-۱-۴-۲- ابزار ارزیابی ROUGE

ابزار Rouge معروفترین ابزار برای ارزیابی در خلاصه سازی خودکار می باشد که البته از آن در دیگر کاربردهای پردازش زبان طبیعی[1] و بازیابی اطلاعات[2]هم استفاده شده است. Rouge مخفف جمله ی "Recall-Oriented Understudy for Gisting Evaluation" به معنای "ارزیابی مبتنی بر یادآوری برای خلاصه" می باشد. این ابزار شامل معیارهایی برای تعیین کیفیت خلاصه ها به صورت خودکار، از طریق

[1] - NLP(Natural Language Processing)
[2] - IR(Information Retrieval)

مقایسه آنها با خلاصه های تولید شده توسط انسان (خلاصه های ایده آل) می باشد. این معیار ها تعداد واحدهایی که بین خلاصه های سیستمی و خلاصه های انسانی هم پوشانی دارند نظیر n تایی ها[1] رشته ی کلمات[2] و جفت کلمات را محاسبه می نمایند. از جمله این معیار ها به ROUGE-N، ROUGE-L، ROUGE-W و ROUGE-S می توان اشاره کرد. در ادامه به این معیار ها اشاره می کنیم.

- ***معیار ارزیابی ROUGE-N:*** معیار ROUGE-N ، روشی است که مبتنی بر فراخوانی n تایی ها بین یک خلاصه سیستمی و مجموعه ای از خلاصه های انسانی می باشد. ROUGE-N توسط فرمول زیر محاسبه می شود:

$$ROUGE - N = \frac{\sum_{S\epsilon \{Reference\ Summaries\}} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{S\epsilon \{Reference\ Summaries\}} \sum_{gram_n} Count(gram_n)}$$

در این معادله، n بر گرفته شده از طول n تایی بوده $(gram_n)$ و $Count_{match}(gram_n)$ همحداکثر تعداد n تایی هایی است که هم در خلاصه ی تولید شده توسط سیستم و هم در خلاصه مرجع (تولیده شده توسط انسان) رخداد است. پر واضح است که معیار ROUGE-N یک معیار مبتنی بر فراخوانی[3] می باشد چراکه مخرج کسر معادله، مجموع کل تعداد n تایی های است که در خلاصه های مرجع وجود دارد. معیار مشابه BLEU که در ترجمه ماشینی مورد استفاده قرار می گیرد یک روش مبتنی بر دقت[4] می باشد. این معیار میزان انطباق یک ترجمه ماشینی را با تعدادی از ترجمه های انسانی، از طریق محاسبه ی میزان درصد n تایی هایی که بین دو ترجمه مشترک هستند ارزیابی می کند.

لازم به یادآوری است که در محاسبه ROUGE-N هرچه تعداد خلاصه های مرجع بیشتر شود ، تعداد n تایی ها هم در مخرج کسر معادله بیشتر خواهد شد که این امر معقول می باشد چراکه ممکن است چندین خلاصه خوب موجود باشد. هر زمان که تعدادی خلاصه مرجع به مجموعه خلاصه های ایده آل افزوده شود، در حقیقت فضای خلاصه های جایگزین و مطلوب افزوده خواهد شد. در صورتی که از چندین مجموعه مرجع استفاده شود، ROUGE-N بین هر جفت خلاصه ی سیستمی و هر یک از اعضای مجموعه خلاصه های انسانی، محاسبه خواهد شد و سپس بیشترین امتیازی که بدست آمده باشد به عنوان امتیاز نهایی لحاظ خواهد شد. این موضوع به شکل زیر هم بیان می شود:

$$ROUGE - N_{multi} = argmax_i\ ROUGE - N(r_i, s)$$

---

[1] - N- gram

[2] - Words sequence

[3]- recall-related measure

[4]- precision-based measure

در پیاده سازی ROUGE از عملیات Jackknifing استفاده می شود.بدین ترتیب که ابتدا M تا مجموعه ی M-1 عضوی تشکیل داده می شود و سپس به عنوان ورودی به ROUGE داده می شود. میانگین امتیازاتی که به هر کدام از این مجموعه ها داده می شود به عنوان امتیاز نهایی در نظر گرفته می شود. استفاده از عملیات Jackknifing باعث می شود که بتوانیم انسان هایی که در تولید خلاصه شرکت کرده اند را ارزیابی کنیم.

- ***معیار ارزیابی ROUGE-L: Longest Common Subsequence:*** در این معیار ارزیابی از الگوریتم های محاسبه طولانی ترین زیر رشته مشترک بین دو رشته[۱] استفاده می شود. اگر فرض کنیم که $\boldsymbol{X}$ یک جمله ی خلاصه مرجع با طول $\boldsymbol{m}$ و $\boldsymbol{Y}$ هم یک جمله از خلاصه ی سیستمی با طول $\boldsymbol{n}$ باشد فاکتورهای ارزیابی به صورت زیر محاسبه خواهند شد:

$$R_{\text{lcs}} = \frac{LCS(X,Y)}{m}$$

$$P_{\text{lcs}} = \frac{LCS(X,Y)}{n}$$

$$F_{\text{lcs}} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

$$\beta = \frac{P_{lcs}}{R_{lcs}} When \quad \frac{\partial F_{\text{lcs}}}{\partial P_{\text{lcs}}} = \frac{\partial F_{\text{lcs}}}{\partial R_{\text{lcs}}}$$

که در آن $LCS(X,Y)$ طول بزرگترین زیر رشته مشترک بین X و Y می باشد. در ارزیابی های DUC مقدار β برابر با ∞ در نظر گرفته شده است بنابراین فقط R_{lcs} در نظر گرفته شده است. معیار F_{lcs} مبتنی بر LCS، را ROUGE-L می نامند. زمانی که X و Y برابر باشند ROUGE-L برابر یک و زمانی که $LCS(X,Y)=0$ باشد، ROUGE-L صفر خواهد بود.یکی از مزایای ROUGE-L این است که نیازی به محاسباتی انطباق متوالی ندارد. مزیت دوم این روش این است که به صورت اتوماتیک طولانی ترین زیر رشته ی nتایی را در نظر می گیرد و بنابراین نیازی به تعیین طول n تایی پیش فرض نمی باشد. همانطور که قبلا هم اشاره شد ROUGE-L از F-Measure برای ارزیابی استفاده می کند و این در حالی است که در ROUGE-N از معیار Recallاستفاده می شود. Recall در ROUGE-N میزان انطباق کلمات جمله های خلاصه ی مرجع (ایده آل) در خلاصه ی سیستمی را محاسبه می کند. Precision بر عکس Recall بوده و میزان انطباق کلمات خلاصه های سیستمی در خلاصه مرجع را محاسبه می کند. هر دو فاکتور دقت و فراخوانی، به ترتیب بین کلمات توجهی نمی کنند و این یک نقطه ضعف برای معیار ROUGE-N

۱- LCS (Longest Common Subsequence)

می باشد. این موضوع در ROUGE-L در نظر گرفته می شود. به مثال زیر توجه نمایید

مثال : فرض کنید که سه جمله S1 و S2وS3 به صورت زیر موجود باشند.

S1. police killed the gunman
S2. police kill the gunman
S3. the gunman kill police

فرض می کنیم S1 به عنوان مرجع بوده و S2 و S3 هم جملات خلاصه های سیستمی باشد. S2 و S3 امتیازات یکسانی را در ROUGE-2 کسب می کنند چراکه هر ۲ جمله شامل یک ۲تایی مشترک "the gunman" با جمله مرجع می باشند و این در حالی است که معنای این دو جمله کاملا متفاوت از همدیگر می باشد. در ارزیابی با ROUGE-L ، جمله S2 امتیاز 3/4=0.75 و جمله S3 امتیاز 2/4=0.5 را کسب می کنند (با فرض $\beta = 1$). بنابراین در این مثال با ارزیابی ROUGE-L ، جمله دوم امتیاز بیشتری نسبت به جمله سوم کسب خواهد کرد.به هر حال LCS دارای یک مشکل هم می باشد و آن هم این است که فقط به بزرگترین زیر رشته توجه می کند و به سایر زیررشته ها توجهی ندارد.به عنوان مثال در جمله

S4. the gunman police killed

الگوریتم LCS ، یکی از زیر رشته های "the gunman" یا "police killed" و نه هر دو را در نظر گرفته و بر همین اساس امتیاز جمله ۴ و جمله ۳ در ROUGE-L یکسان می شود. در ROUGE-N جمله ۴ به جمله ۳ ترجیح داده می شود.

در قسمت قبل به محاسبه ROUGE-L در سطح جمله اشاره شد. برای محاسبه ROUGE-L در سطح خلاصه هم مانند قسمت قبل از LCS استفاده می شود. اگر خلاصه مرجع شامل u جمله و در مجموع m کلمه باشد و خلاصه سیستمی شامل v جمله و n کلمه باشد معیار ارزیابی ROUGE-L به صورت زیر محاسبه خواهد شد

$$R_{\text{lcs}} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m}$$

$$P_{\text{lcs}} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m}$$

$$F_{\text{lcs}} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

همانطور که قبلا هم اشاره شد در مجموعه DUC فقط از R_{lcs} استفاده می شود. $LCS_{\cup}(r_i, C)$ ، امتیاز LCS ی اجتماع طولانی ترین زیررشته مشترک بین جمله مرجع r_i و خلاصه C می باشد. به

عنوان مثال اگر $ri = w1\ w2\ w3\ w4\ w5$ و C(خلاصه سیستمی) هم شامل دو جمله ی $c1 = w1$ $w2\ w6\ w7\ w8$و $c2 = w1\ w3\ w8\ w9\ w5$ باشد، آنگاه طولانی ترین زیر رشته ی مشترک ri و $c1$ ، زیر رشته ی "$w1\ w2$" و طولانی ترین زیر رشته ی مشترک بین ri و $c2$ ، "$w1\ w3\ w5$" می باشد. اجتماع زیر رشته های مشترک ri و $c1$ و $c2$، زیر رشته ی "$w_1\ w_2\ w_3\ w_5$" و $LCS_U(r_i, C) =$ 4/5 خواهد بود.

- ***معیار ارزیابی ROUGE-W: Weighted Longest Common Subsequence*** ویژگی LCS های جذابی دارد که در قسمت قبل به آنها اشاره کردیم. متاسفانه LCS مشکل دیگری هم دارد و آن عدم در نظر گرفتن فاصله قرار گیری بین کلمات می باشد. به عنوان مثال جمله مرجع ***X*** و جملات خلاصه ***Y1*** و ***Y2*** را به صورت زیر در نظر بگیرید:

X: [**ABCDE F G**]

Y_1: [**ABCDH I K**]

Y_2: [**AHBKCID**]

با معیار ROUGE-L ، $Y1$ و $Y2$ هر دو به طور یکسان امتیاز می گیرند. در حالی که $Y1$ باید امتیاز بیشتری کسب نماید. ROUGE-W با بخاطر سپردن طول کلمات متوالی این مشکل را حل می کند.

- ***معیار ارزیابی ROUGE-S: Skip-Bigram Co-Occurrence Statistics*** : به هر جفت کلمه در (با حفظ ترتیب) جمله، Skip-bigram گفته می شود. ROUGE-S با اندازه‌گیری تعداد Skip-bigramهای مشترک بین خلاصه‌های سیستم و خلاصه‌های مرجع محاسبه می‌شود. به عنوان مثال جملات زیر را در نظر بگیرید :

> *S1. police killed the gunman*
> S2. police kill the gunman
> S3. the gunman kill police
> S4. the gunman police killed

هر جمله ای $C(4,2)= 6$ [1] تا Skip-bigram دارد.

S1 = (“*police killed*”, “*police the*”, “*police gunman*”, “*killed the*”, “*killed gunman*”,

[1]- C(4,2) = 4!/(2!*2!) = 6

"*the gunman*")

با محاسبه تعداد انطباق ها در خلاصه های مرجع و سیستمی این معیار محاسبه می شود. از این معیار بیشتر در ارزیابی ترجمه ماشینی استفاده می شود.

- ***معیار ارزیابی ROUGE-SU: Extension of ROUGE-S :*** مشکل اصلی روش ROUGE-S این است که برای جملات کاندیدی (جمله های خلاصه سیستمی) که هیچ Skip-bigram مشترکی با خلاصه مرجع نداشته باشند هیچ امتیازی در نظر نمی گیرد. به عنوان مثال به جمله زیر امتیاز صفر می دهد.

S5. gunman the killed police

جمله ۵ دقیقا عکس جمله ۱ می باشد و هیچ Skip-bigram مشترکی بین آن دو وجود ندارد. در حقیقت باید بین جمله ۵ و جمله هایی که هیچ کلمه مشترکی با جمله ۱ ندارند تفاوت گذاشته شود. برای رسیدن به این هدف، ROUGE-SU با در نظر گرفتن یکتایی ها به عنوان واحد شمارش محاسبه می شود.

۲-۴-۲- مجموعه داده های استاندارد برای خلاصه سازی

یکی از چالش های مهم در امر خلاصه سازی متون، بحث ارزیابی روش های ارائه شده می باشد. برای یک ارزیابی مناسب و دقیق احتیاج به یک مجموعه داده ی مناسب و استاندارد داریم. در مقالات مختلف از داده های مختلفی تاکنون استفاده شده است که از جمله آنها می توان به مجموعه داده های خبری CNN،BBC، TREC[۱]، CASTcorpus[۲]، DUCcorpus[۳]اشاره نمود. با توجه به بررسی های انجام شده، مجموعه داده های DUC(DocumentUnderstandingConferences) انتخاب شده اند. در ذیل مختصرا این مجموعه داده ها شرح داده شده است.

۱-۲-۴-۲- داده های استاندارد DUC

کنفرانس DUC از سال ۲۰۰۱ زیر نظر NIST[۴] شروع به انتشار داده های مورد نیاز برای خلاصه سازی متون کرده است و تاکنون ۷ مجموعه از داده ها را تحت عنوان DUC2001 تا DUC2007 ارائه نموده است. هر کدام از این مجموعه‌ها با اهداف خاصی انتشار یافته اند. هدف اصلی این کنفرانس کمک در ارزیابی روش‌های خلاصه سازی خودکار متون و بررسی روش های ارزیاب خلاصه سازی می باشد. مجموعه داده های

۱- http://trec.nist.gov

۲ http://clg.wlv.ac.uk/projects/CAST/corpus/

۳- http://duc.nist.gov

۴- national institute of standards and technology

DUC2001 تا DUC2004 برای خلاصه سازی تک سندی و چند سندی تولید شده اند. مجموعه داده های DUC2005 تا DUC2007 هم فقط برای خلاصه سازی چند سندی تولید شده اند. با توجه به اینکه مجموعه DUC2007 اخرین مجوعه از این داده ها و کاملترین آنها می باشد، بنابراین از این مجموعه از داده ها استفاده می کنیم.

داده‌های DUC2007 در مجموع شامل ۴۵ موضوع بوده که هر کدام شامل ۲۵ سند می باشد. ۱۰ نفر از اعضای NIST وظیفه نوشتن خلاصه های دستی برای این مجموعه را بر عهده داشته اند به طوری که برای هر موضوع ۴ نفر به صورت تصادفی انتخاب شده و خلاصه های چکیده ای تولید کردند. ۳۲ سیستم هم در این پروژه دخیل بوده و برای هر موضوع خلاصه خودکار تولید کردند. سپس این خلاصه ها با استفاده از ابزار ROUGE با خلاصه های انسانی مقایسه شده و سیستم ها بر اساس نتایج بدست آمده مرتب می شوند.

همچنین اعضای عضو تیم ارزیاب، سیستم ها را از دید قواعد زبان شناسی بررسی کرده و با در نظر گرفتن فاکتور هایی نظیر پیوستگی و خوانایی ، دقت، ساختار گرامری و ... به آنها امتیاز داده اند. در جدول زیر اطلاعات کلی از این مجموعه داده ها آورده شده است :

DataSet	Num of Topics	Task of Dataset	# document	Summary Length	Num Of Systems
DUC 2001	30	single-multi	303	100w	12
DUC 2002	59	single-multi	~600	100 w	12
DUC 2003	60	single-multi	600	75 byte	23
DUC 2004	50	single-mlti	500	665 byte	-
DUC 2005	۵۰	multi	1300	250w	۳۲
DUC 2006	۵۰	multi	1225	250w	35
DUC 2007	۴۵	multi	1125	250w	۳۲

جدول 3- اطلاعات مربوط به مجموعه داده های DUC

۵-۲- سیستم های معروف خلاصه سازی

۱-۵-۲- سیستم MEAD

یکی از معروف ترین خلاصه سازی می باشد که در تعدادی از مقلات مورد استفاده قرار گرفته است. MEADخلاصه‌ساز چند سندی می‌باشد. ورژن ۱ و ۲ این خلاصه ساز در دانشگاه میشیگان در سال ۲۰۰۰ - ۲۰۰۱ پیاده سازی شده است. این خلاصه ساز با زبان پرل پیاده سازی شده است. این خلاصه ساز طوری پیاده سازی شده است که برای تمامی زبان ها ، قابل اجرا باشد و در حال حاضر نسخه های انگلیسی و چینی و ژاپنی و هلندی آن موجود می باشد. اطلاعات کامل در مورد نحوه ی عملکرد این خلاصه ساز موجود می باشد.

نسخه هلندی این خلاصه ساز بهمراه اطلاعات و مستندات کامل درباره این خلاصه ساز در سایت به نشانی http://www.summarization.com/mead موجود می باشد.

۲-۵-۲- سیستم SweSum

یکی از معروف ترین سیستم های بر خط خلاصه سازی می باشد که به زبان های مختلف عمل خلاصه سازی را انجام می دهد. اولین خلاصه ساز تولید شده برای زبان سوئدی می باشد. دقت آن برای متون خبری برای زبان سوئدی ، برای خلاصه با میزان ۴۰ درصد ، برابر ۸۴ درصد اندازه گیری شده است(با طول متوسط ۱۸۱ کلمه). نسخه نهایی این خلاصه ساز در حال حاضر برای زبان های انگلیسی ، دانمارکی ، نروژی و سوئدی تولید شده و نسخه آزمایشی آن هم برای زبان های فارسی ، فرانسوی ، آلمانی ، ایتالیایی، اسپانیایی و یونانی تحت بررسی و آزمایش است. این سیستم توسط مارتین حصل؟؟؟ و گروه تحقیقاتی زبانشناسی رایانه ای دانشگاه استکهلم پیاده سازی شده است.

نسخه تحت وب این خلاصه ساز در حال حاضر موجود بوده که در سایت به نشانی http://swesum.nada.kth.se/index-eng.html قرار دارد

۳-۵-۲- سیستمPERSIVAL

PERSIVAL مخفف Personalized Retrieval and Summarization of Image , Video and Language می باشد. سیستم، خلاصه‌های از مقالات پزشکی مخصوصیک بیمار را از کتابخانه ی دیجیتالی توزیع شده چندرسانه ای مراقبت از بیماران، تهیه می کند.

سیستم شامل تفاسیر و سازمان مجموعه بزرگی از داده‌های ویدئویی است. اسناد ویدئویی قطعه بندی می-شوند و یک خلاصه storyboard تولید می‌شود. ویدئو ها در سطوح نحوی و معنایی شاخص بندی می‌شوند.

یک مجموعه ابزارهای مبتنی بر محتوا برای جستجوی ویدئو توسعه داده شده‌اند. این سیستم از DEFINDER برای جستجوی تعاریف استفاده می‌کند.

۴-۵-۲- سیستم SUMMARIST

این سیستم از نوع تک سندی می باشد. سیستم شامل سه مرحله پردازش است : شناسایی موضوع ، تفسیر و تولید خلاصه. شناسایی موضوع، به کمک امضاهای موضوع که قبلا تهیه شده انجام می‌شود. امضاهای سند، چندتایی هایی به فرم <topic, signature> است که امضا لیستی از کلمات وزن دار بفرم <tn,wn> و ... و <t2,w2> و <t1,w1> است. امضاهای سند می‌توانند به صورت خودکار یاد گرفته شوند. پس شناسایی موضوع شامل قطعه بندی متن (با کاشی کاری متن) و مقایسه span های متن با امضاهای موضوع موجود است. موضوع شناسایی شده در طول پروسه تفسیر ترکیب می‌شود. سپس موضوعات ترکیب شده مجددا فرمول بندی می‌شوند. مرحله آخر هم طبق معمول استخراج است.

این خلاصه ساز برای زبان های مختلفی عمل کرده و اساسا طوری نوشته شده که تمام زبان ها را پشتیبانی کند مثل : انگلیسی ، ژاپنی ، اسپانیایی ، عربی ، اندونزیایی ، کره یی.

۵-۵-۲- سیستم های تجاری موجود

در سال‌های اخیر تعداد مقالات زیادی در خصوص خلاصه سازی چند سندی ارائه شده است. بسیاری از کنفرانس ها نظیر DUC هم تمرکز اصلی خود را برروی این مدل از خلاصه سازی معطوف کرده اند. با توجه به ماهیت خلاصه سازی چندسندی و نزدیک بودن این مبحث به موضوعاتی نظیر موتورهای جستجو و سیستم های پرسش و پاسخ، سیستم های تجاری زیادی در این زمینه ارائه شده است. در ذیل به سه مورد از این سیستم ها اشاره می کنیم.

۱-۵-۵-۲- سایت[۱] Newsfeedresearcher

یک پرتال خبری می باشد که خلاصه سازی چندسندی انجام می دهد. هفت دسته کلی خبر دارد که با انتخاب هر کدام از این دسته ها مهمترین خبرهای روز را در زمینه دسته انتخاب شده می توان مشاهده کرد. این وب سایت در درون خود از جستجوگر گوگل استفاده کرده و خلاصه سازی را روی نتایج حاصل از این جستجوگر اعمال می نماید.

---

۱ http://newsfeedresearcher.com/

۲-۵-۵-۲- سایت Iresearch-reporter[۱]

سیستم تجاری استخراج متن و خلاصه ساز متن می باشد. نسخه دموی روی خط آن، پرس و جوی کاربر را به عنوان ورودی دریافت کرده و توسط گوگل جستجو می کند و سپس اسناد برگردانده شده را پردازش کرده و بهترین ها را انتخاب کرده و عملیاتهایی نظیر دسته بندی، استخراج نهادها، استخراج روابط و رویدادها و پردازش های زبانی و خلاصه سازی را روی آنها اعمال می کند.

۳-۵-۵-۲- سایت Shablast[۲]

این سایت در حقیقت یک موتور جستجوی عمومی می باشد که برای پرس و جوی داده شده توسط کاربر خلاصه چند سندی بر می گرداند. برای جستجو هم از نتایج جستجوگر بینگ مایکروسافت[۳] استفاده می کند.

۶-۵-۲- سایر سیستم ها

در قسمت‌های قبلی تعدادی از سیستم‌های خلاصه سازی را معرفی نمودیم. این سیستم ها جزء معروفترین سیستم های خلاصه سازی موجود می باشند. اما سیستم های خلاصه سازی دیگری هم وجود دارند که برای زبان های مختلف ارائه شده اند. در ذیل نام تعدادی از این سیستم ها آورده شده است:

- GLEANS[Dau 00]
- SumUM [Sag 02]
- RIPTIDES[Whi 01]
- NTT[Hir 03]
- NeATS , iNeATS[Lin 02]
- GISTSumm[Par 01]
- GISTexter[Lac 06]
- DiaSumm[Zec 00]

و ...

۶-۲- خلاصه سازی در زبان فارسی

از اوایل سال ۱۹۵۷ که اولین سیستم خلاصه سازی برای زبان انگلیسی معرفی شد تا به امروز، بیش از صدها مقاله در این زمینه ارائه شده است و این در حالی است که متاسفانه مقالاتی که برای زبان فارسی ارائه شده است بسیار اندک می باشد که دلایل مختلفی را می توان برای این موضوع متصور شد. عدم توجه محققان و

---

۱ http://iresearch-reporter.com/

۲ http://shablast.com/

۳ http://www.bing.com/

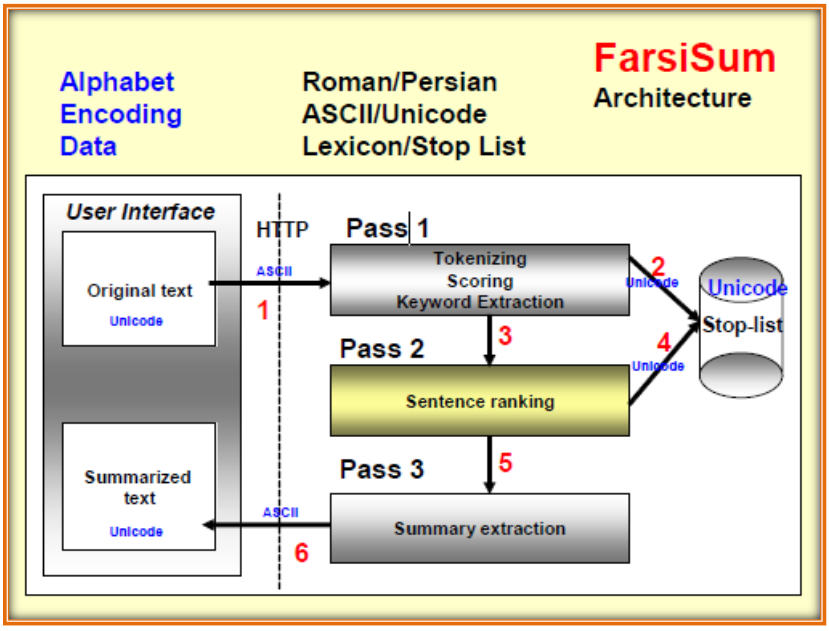
دانشگاههای داخل کشور، عدم وجود ابزارهای پیش پردازش مورد نیاز و همچنین پیچیدگی های زبان فارسی از جمله دلایل کمبود مقالات در این زمینه می باشد. تنها سیستمی که در حال حاضر به صورت یک ابزار قابل استفاده برای کاربران می باشد سیستم خلاصه سازی FarsiSum هست که به صورت روی خط قابل دسترس می باشد.

۱-۶-۲- سیستم FarsiSum

FarsiSum یک خلاصه ساز مبتنی بر وب است که SweSum را برای زبان فارسی مدل میکند. این سیستم قادر به خلاصه کردن متون روزنامههای فارسی با قالب HTML و متن کد شده با فرمت یونیکد میباشد. این سیستم تحت ویندوز و به زبان پرل نوشته شده است . این سیستم توسط نیما مزدک در سال ۲۰۰۴ ارائه شده و معروف ترین سیستم خلاصه ساز فارسی می باشد.

پروسه تولید خلاصه در این خلاصه ساز را در تصویر ۱۱ مشاهده می کنید :

- مرورگر درخواست خلاصه را بهمراه سند ، از طریق سرور HTTP به سروری که FarsiSumدر آن قرارگرفته می فرستد.
- سند درطی سه مرحله با استفاده از لیست توقف فارسی ، خلاصه میشود.
- متن خلاصه تهیه شده از طریق HTTP به کاربر باز می گردد.
- سپس مرورگر متن خلاصه را در صفحه پردازش میکند



شکل 5 - ساختار کلی FarsiSum

۲-۶-۲- سایر کارهای انجام شده

در [کری ۸۵] ، یک روش خلاصه سازی تک سندی دیگر پیشنهاد شده است .این بر مبنای روش مانند FarsiSum بر مبنای گزینش جمله ها کار می کند .همچنین، محتوی خلاصه می تواند کلی یا بر اساس پرس وجوی کاربر باشد .ایده ی بکار رفته در گزینش جمله ها در این خلاصه ساز، ترکیبی از دو روش زنجیره ی لغوی و نظریه ی گراف است کار دیگر نیز سیستم خلاصه سازی PARSUMIST [sha 09] می باشد که تکمیل شده پروژه قبلی آنها است که یک سیستم خلاصه ساز چند سندی است و ترکیبی از دو روش زنجیره لغوی و تئوری گراف می‌باشد.

در [Zam 08] هم روشی ترکیبی برای خلاصه سازی فارسی ارائه شده است که از ویژگی هم رخدادی کلمات و همچنین ویژگی های مفهومی زبان فارسی استفاده می کند.

۲-۷- خلاصه

در این فصل در ابتدا مروری بر مفهوم خلاصه سازی خوکار متن، تاریخچه خلاصه سازی خودکار، کاربردهای خلاصه سازی و انواع مختلف مدل های خلاصه سازی شد و سپس اطلاعات پایه و اولیه زبان شناسی مورد نیاز شرح داد شد. در ادامه به انواع روشهایی که برای خلاصه سازی متن در گذشته مورد استفاده قرار قرار گرفته اند اشاره شد. سپس به روشهای ارزیابی خودکار خلاصه سازی و ابزارهای موجود در این زمینه و همچنین مجموعه داده های استاندارد برای خلاصه ساز خودکار متن اشاره شد. در پایان هم سیستم های معروف خلاصه سازی خودکار متن معرفی شدند. خلاصه سازی بر روی زبان فارسی هم در ادامه مورد بحث قرار گرفت و به روش هایی که تاکنون برروی زبان فارسی ارائه شدند اشاره شد.

در فصل بعد روش پیشنهادی برای خلاصه سازی خودکار چند سندی شرح داده شده است.

فصل سوم :

روش پیشنهادی

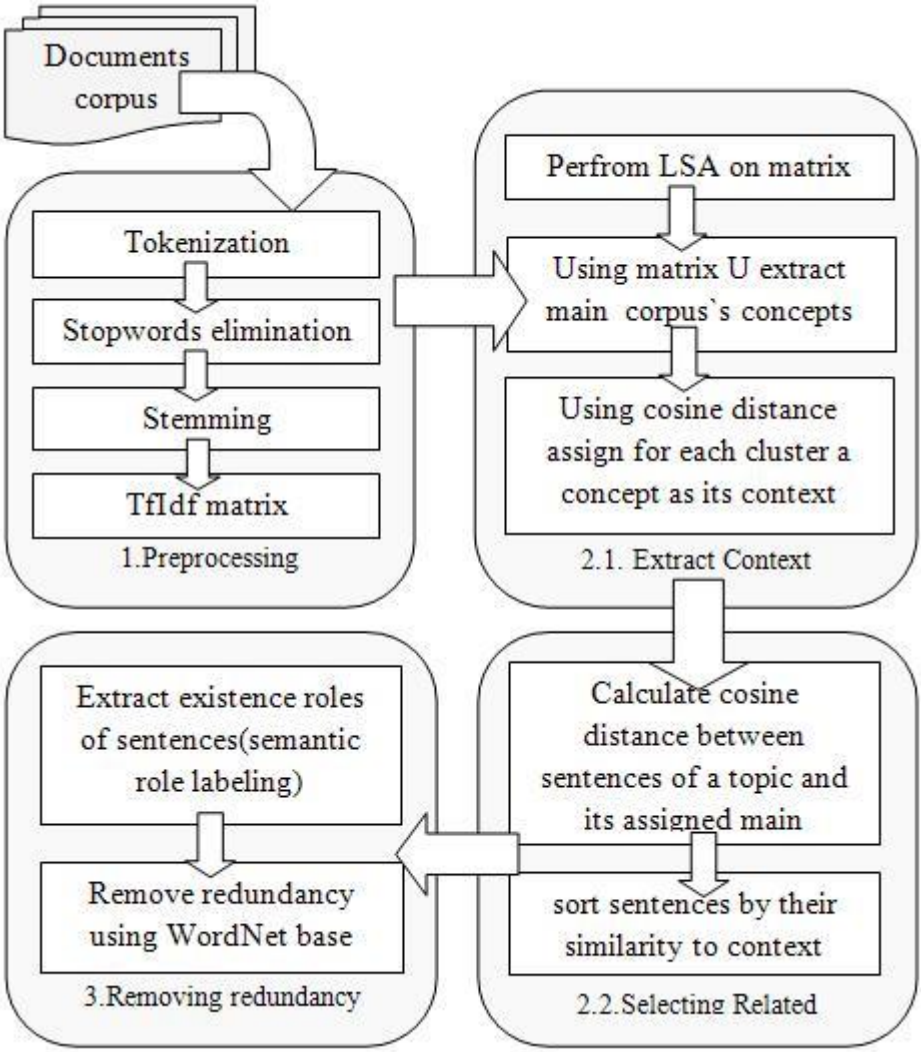
۳- روش و الگوریتم پیشنهادی برای خلاصه سازی چند سندی

در این بخش روش کلی برای خلاصه سازی خودکار چند سندی با استفاده از استخراج زمینه و محاسبه شباهت معنایی بین جملات توضیح داده شده است. در ابتدا باید مجموعه ای از متون استاندارد برای خلاصه سازی چندسندی بهمراه خلاصه های انسانی از پیش تهیه شده جهت مقایسه فراهم آوریم. بهمین منظور از داده های خبری استاندارد DUC2007 تهیه شده توسط موسسه استاندارد ملی آمریکا که از سه خبرگذاری معروف شینهوا، آسوشییت پرس[۱] و نیویورک تایمز[۲] گردآوری شده استفاده می نماییم. مجموعه داده های DUC استاندارترین و معروفترین و پرکاربردترین مجموعه داده های موجود برای خلاصه سازی تک سندی و چند سندی در وب می باشد که در این مجموعه DUC2007 آخرین مجموعه داده ایی می باشد که تولید شده است. همانطور که پیش تر هم اشاره شد این مجموعه داده ها شامل ۴۵ موضوع بوده که برای هر موضوع ۲۵ سند از خبرگذاری های مختلف گردآوری شده است. بعد از انتخاب مجموعه داده ها عملیات پیش پردازش برروی داده ها شامل استخراج جملات، حذف کلمات ایست، ریشه یابی و ساخت ماتریس کلمه-سند انجام می شود. البته این فاز در قالب دو آزمایش مبتنی بر پیکره و عمومی انجام شده است. سپس ماتریس ساخته شده به عنوان ورودی به LSA داده شده و از خروجی مارتیس U برای استخراج مفاهیم استفاده می شود. با استخراج بردار مفاهیم داده ها، جملات مرتبط با این مفاهیم با استفاده از فاصله کسینوسی محاسبه می شود. سپس جملات بر اساس میزان شباهتشان با بردار مفهوم مرتب می شوند. در گام بعدی با استفاده از محاسبه شباهت معنایی با استفاده از شبکه واژگان در نقش‌های همسان، جملات شبیه به هم بدست آمده وبا حذف افزونگی جملات تکراری عملا حذف شده و دیدگاههای مختلف موجود در متن پوشش داده می شوند. در مرحله پایانی هم جملات را بر اساس ؟؟؟ مرتب می نماییم.

در شکل ؟؟؟ روند کلی روش پیشنهادی توضیح داده شده است.

۱- ??

۲ ??



شکل ۶- معماری کلی سسیستم پیشنهادی

مراحل کلی انجام تحقیق را به چهار فاز کلی می توان تقسیم کرد

الف) فاز پیش پردازش مجموعه دادههای موجود

ب) فاز استخراج جملات مرتبط با زمینه و عنوان

ج) فاز استخراج دیدگاه و حذف افزونگی در جملات خلاصه

د) فاز مرتب سازی جملات

در ادامه به تفصیل به شرح هر یک از مراحل روش پیشنهادی می پردازیم.

۱-۳- فاز پیش پردازش

در این گام باید داده های اولیه DUC را پردازش کرده و آماده استفاده برای مراحل بعدی کنیم. این داده ها

در قالب فایل متنی و با برچسب های مشخص مبتنی بر XML می باشند. بنابراین در گام اول باید این برچسب‌ها تجزیه شده و متن مربوط به هر خبر استخراج شود. برای این کار می توان از روش و ابزارها و زبان-های برنامه نویسی مختلفی استفاده نمود. در این پایان‌نامه از کلاس parseXML زبان برنامه‌نویسی PHP استفاده شده است. پس از استخراج متن خبر پردازشهای زیر به ترتیب باید انجام شود:

۱) ***استخراج جملات و کلمات تشکیل دهنده آنها :*** در اینگام باید در ابتدا جملات را استخراج نماییم. سپس این جملات را به کلمات تشکیل دهنده اش تجزیه می کنیم. برای اینکار ابزارهای پردازش متن متنوعی ارائه شده است که از جمله آنها می توان به ابزارهای معروفGATE و همچنین OpenNLP اشاره کرد. با توجه به فرمت مجموعه داده های DUC و برای یکپارچه شدن این فاز بامرحله قبلی از قطعه کد ساده ای به زبان PHP برای استخراج جملات و همچنین استخراج کلمات استفاده شده است.

۲) ***حذف کلمات بی اهمیت :*** در گام بعدی باید کلمات ایست را حذف نماییم. کلمات ایست به کلمات پرکاربردی که اهمیت چندانی ندارند (stopword) اتلاق می شود. برای این منظور هم می توان از ابزار متن کاویGateیاOpenNLP استفاده نمود. در این پایان‌نامه از لیست ۶۳۷ تایی کلمات ایست برای زبان انگلیسی استفاده شده است و با استفاده از زبان برنامه‌نویسی PHP برنامه ای برای تشخیص و حذف این کلمات نوشته شده است.

۳) ***ریشه یابی :*** در گام بعدی عملیات ریشهیابی برروی کلمات استخراج شده را انجام می دهیم. هدف از انجام ریشهیابی انتخاب یک نماینده برای کلماتی که از نظیر معنایی به هم نزدیک و از یک ریشه می باشند، هست. مثلا کلمات stemmer، stemming و stems همگی در یک حوزه معنایی قرار داشته و نماینده آنها هم کلمه stem می باشد. الگوریتم‌های ریشهیابی زیادی تاکنون ارائه شده است که معروفترین آنها الگوریتمporter می باشد. پرتر یک الگوریتم مبتنی بر نحو بوده و نیازی به فرهنگ لغات برای استخراج ریشه ندارد. ابزار Gateهم عملیات ریشهیابی را پوشش می دهد. در این پایان‌نامه از الگوریتم ریشه یابی پرتر نوشته شده به زبان PHP استفاده شده است.

۴) ***ساخت ماتریس کلمه-سند :*** پس از انجام مراحل بالا ماتریس کلمه – سند ساخته می شود. بر خلاف روشهای قبلی که از LSA برای خلاصه سازی استفاده کرده‌اند[Gon 01][Ste 05]، در روش پیشنهادی که در این پایان‌نامه ارائه شده به جای استفاده از ماتریس کلمه–جمله و سپس انتخاب مستقیم جملات توسط LSA، از ماتریس کلمه– سند استفاده شده و در طی دو مرحله جملات مهم استخراج می‌گردد که با توجه به ماهیت خلاصه سازی چند سندی این تغییر به ظاهر ساده، تاثیر چشم گیری در افزایش دقت خلاصه سازی داشته است که در قسمت

نتایج هم به آن اشاره شده است. دلیل این امر را می توان در ماهیت خلاصه سازی چندسندی جستجو کرد. همانطور که پیش‌تر هم اشاره شد، در خلاصه سازی چند سندی، چندین سند در ارتباط با یک موضوع وجود دارند که همگی در ارتباط با این موضوع بوده ولی از دیدگاههای مختلف به آن اشاره می کنند. بنابراین زمینه اصلی همه این اسناد یکی می باشد. در روش های پیشین که از ماتریس کلمه-جمله استفاده می کنند، با توجه به اینکه مفاهیم بر مبنای جملات بدست می آید و نه بر اساس اسناد موجود در کلاستر، بنابر این جملات مهم با دید محلی استخراج شده و نه با دید عمومی و همین موضوع باعث می شود که جملات اصلی و مهم در ارتباط با زمینه متن استخراج نگردد. در روش پیشنهادی تاثیر همه اسناد با هم در نظر گرفته می شود و جملات بر اساس شباهتشان با مفهوم اصلی استخراج می گردند.همانطور که پیش‌تر هم اشاره شد روش پیشنهادی در قالب دو مدل مبتنی بر پیکره و عمومی پیاده سازی شده است. در روش مبتنی بر پیکره، ماتریس اولیه از داده های کل پیکره ساخته می شود که شامل ۴۵ موضوع و ۲۵ سند در هر موضع (۱۱۲۵ سند در مجموع) و در حدود ۲۰۰۰۰ کلمه می باشد. در روش عمومی این ماتریس برروی هر موضوع جداگانه باید ساخته شود. در جدول ؟؟؟ اطلاعات کلی مربوط به این دادهها آورده شده است.

مشخصه پیکره DUC2007	مقدار
تعداد موضوعات	۴۵
تعداد اسناد موجود در هر موضوع	۲۵
تعداد کلمات	۵۳۱۱۷۴
تعداد کلمات بدون با حذف کلمات ایست و ریشه یابی	۲۰۰۵۷
تعداد سیستم های خلاصه ساز	۳۲
معیار ارزیابی	ROUGE2-ROUGESU4

جدول 4- مشخصات کلی پیکره دادههای DUC2007

مشخصه موضوع 703	مقدار
موضوع	steps toward introduction of the Euro
تعداد جملات	۱۶۷
تعداد کلمات پس از ریشه یابی	۸۱۴
تعداد کلمات	۵۳۱۱۷۴

جدول 5- مشخصات موضوع 703 موجود در پیکره (مدل عمومی)

مشخصه پیکره DUC2006	مقدار
تعداد موضوعات	۵۰
تعداد اسناد موجود در هر موضوع	۲۵
تعداد کلمات	۵۷۹۲۰۱
تعداد کلمات بدون با حذف کلمات ایست و ریشه یابی	۲۱۱۳۵
تعداد سیستم های خلاصه ساز	۳۵
معیار ارزیابی	ROUGE2-ROUGESU4

جدول ۶- مشخصات کلی پیکره دادههای DUC2006

برای ساخت ماتریس از روشهای وزن‌دهی مختلفی می‌توان استفاده نمود. در این پایانه‌نامه از روش وزن‌دهی tf-idf که ترکیبی از روشهای وزن‌دهی سراسری[۱] و روشهای وزن‌دهی محلی می باشد استفاده شده است که به صورت زیر محاسبه می شود:

$$(TF-IDF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) Log_2\left(\frac{|D|}{|\{d: t_i \in d\}|}\right)$$

$tf_{i,j}$تعداد دفعات تکرار کلمه i در سند j و $|D|$کل اسناد موجود در پیکره می‌باشد. البته از روش های وزن دهی دیگر هم می توان استفاده نمود. اطلاعات کلی در مورد روش های وزن دهی به کلمات در مرجع [Eri 99] آورده شده است.

۲-۳- استخراج زمینه مرتبط با موضوع و جملات مرتبط با زمینه

همانطور که پیش‌تر هم اشاره کردیم در خلاصه سازی چندسندی زمینه کلی اسناد در ارتباط با یک موضوع بوده ولی دیدگاهها متفاوت از یکدیگر می باشد. مثلا در مورد موضوع "یورو" می توانیم چندین سند داشته باشیم که همه در ارتباط با یورو بوده ولی این موضوع را در کشورهای مختلف بررسی کرده باشند. بنابراین زمینه اصلی همه در ارتباط با یورو و مسائل مرتبط با آن می باشد. بر همین اساس در این فاز از پروژه سعی می شود تا زمینه اصلی مرتبط با هر موضوع بدست آید. پس از استخراج زمینه مرتبط با هر موضوع، جملات

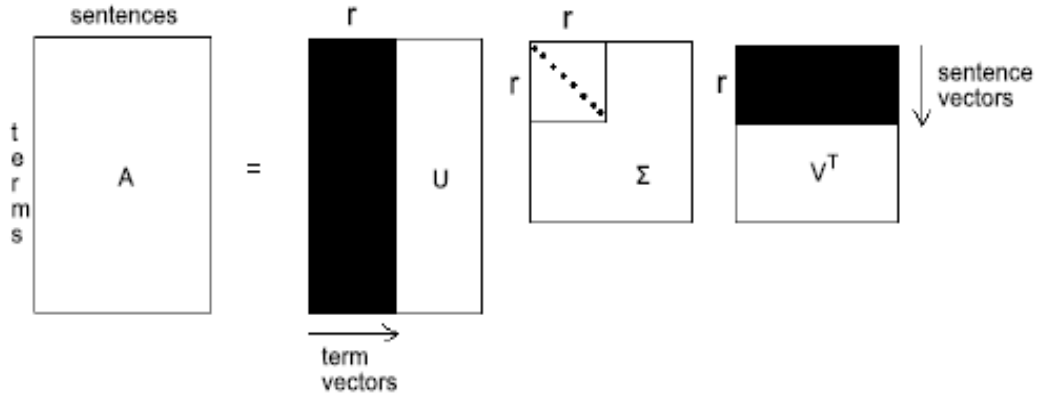
---

۱ Global weighting methods

بر اساس میزان شباهتشان با زمینه متن مرتب می شوند. نتایج بدست آمده نشانگر آن است که در این فاز بخوبی توانسته ایم زمینه مرتبط با هر موضوع را استخراج کنیم. در ادامه مختصرا به این نتایج و آزمایشات انجام شده می پردازیم.

۱-۲-۳- استخراج زمینه

زمینه متن را می توان مجموعه ای از کلمات در نظر گرفت که در کنار یکدیگر مفهوم اصلی و منظور متن را اطلاع رسانی می کنند. در این بخش با بهره‌گیری از الگوریتم LSA، زمینه مرتبط با موضوع متن استخراج می شود. LSA در ابتدا برای حل مشکل هم خانواده‌ها و هم آوایی‌هادر IR معرفی شد[Dee 90]. از آن پس LSA در کانون توجه محققان در زمینه پردازش متن قرار گرفت و در بسیاری از مقالات به صورت تئوری و عملی آنالیز گردید[Dee 90][Pap 00]. LSA در درون خود از روش SVD[۲] استفاده می‌کند. SVD ماتریس A با ابعاد m×n را به سه ماتریس $A = USV^T$ تجزیه می‌کند که $U = [u_{ij}]$ ماتریس ارتونرمال‌ستونی[۳] m×m بوده و ستون‌های آن بردارهای یکه سمت چپ نامیده می شوند؛ $S = diagonal(\sigma_1, \sigma_2, \dots, \sigma_n)$ ماتریس قطری n×n است که عناصرقطر اصلی آن مقادیر ویژه غیرمنفی می‌باشند که به صورت نزولی مرتبط شده‌اند و$V = [v_{ij}]$ هم ماتریس ارتونرمال سطری بوده و ستون‌های آن بردارهای یکه سمت راست نامیده می‌شوند.



شکل ۷- خروجی SVD برای ماتریس کلمه-جمله

اگر rank(A)= r باشد آنگاه حالت زیر برقرار خواهد بود :

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \sigma_{r+1} = \dots = \sigma_n = 0$$

از دیدگاه ریاضی، استفاده از SVD برروی یک ماتریس منجر به نگاشت فضای m بعدی به یک فضای r بعدی

---

[۱] polysemy

[۲]- singular value decomposition

[۳] orthonormal

می شود و در بسیاری از کاربردها از آن برای کاهش ابعاد در ماتریس‌های تنکاستفاده می شود. از دیدگاه پردازش زبان طبیعی، SVD باعث استخراج روابط معنایی پنهان در ساختار داده می‌شود.

از LSA در خلاصه سازی متن هم استفاده شده است. اولین بار Gong در سال ۲۰۰۱ از LSA در خلاصه-سازی استفاده نمود[Gon 01]. در روش Gong و سایر روش‌های اشاره شده، بر ویژگی‌های جمله به عنوان عنصر محوری تمرکز شده است و از آن برای استخراج زیر موضوعات و جملات متناسب با انها استفاده شده است. در این روش‌ها، با تولید بردار فرکانس نسبی جملات از طریق محاسبه معیارهای وزن‌دهی در واحد جمله نظیر $TFISF$، ماتریس کلمه-جمله تولید می‌شود. سپس SVD برروی این ماتریس اعمال شده و با استخراج ماتریس V که می‌توان آن را ماتریس مفهوم- جمله نامید، بیشترین مقادیر موجود در هر ستون (مفهوم) به عنوان بهترین جمله‌ای که نشانگر آن مفهوم است، انتخاب و در خلاصه قرار داده می‌شود. این روش ها دارای دو ایراد اساسی می باشند :

۱. ایراد تمامی این روش‌ها عدم توجه به مفاهیم اصلی پنهان درکل سند می‌باشد زیرا مبنای کار این روش ها استفاده از ماتریس کلمه-جمله می باشد. و این در حالی است که یک خلاصه خوب در خلاصه‌سازی چندسندی، باید ضمن بیان موضوع و زمینه اصلی موجود در اسناد، دیدگاههای مختلف مرتبط با زمینه را نشان دهد. محوریت قرار گرفتن جمله به عنوان محور محاسبات باعث می شود که قادر به کشف ارتباط دقیق بین کلمات نباشیم.
۲. در روش Gong بین بردارهای مختلف ارزش یکسان قائل می شوند و برای هر بردار از ماتریس Vیک جمله به عنوان نماینده آن مفهوم انتخاب می شود و این در حالی است که ممکن است ارزش یک مفهون به قدری زیاد باشد که بایدچندین جمله از متن را به عنوان نماینده آن انتخاب کرد تا خلاصه تولید شده، جامعیت کافی را داشته باشد. از طرف دیگر ممکن است بعضی از مفاهیم بدست آمده آنقدر کم ارزش باشند که انتخاب جمله ای بعنوان نماینده آنها کار اشتباهی باشد.

در این بخش با ارائه روشی جدید سعی شده تا این مشکلات برطرف شوند. به همین منظور در ابتدا به جای استفاده از $TFISF$، وزن $TFIDF$ برای کلمات در *واحد سند* محاسبه شده تا اثر تمامی اسناد لحاظ شود. سپس ماتریس کلمه– سند ایجاد می‌شود. در گام بعدی SVD برروی این ماتریس اعمال شده و ماتریس بردارهای ویژه استخراج می‌شود. در مرحله بعد، از ماتریس ستونی U استفاده می‌شود. ماتریس U در حقیقت ماتریس کلمه-مفهوم می باشد. چون در نظر داریم که مفاهیم موجود در متون را استخراج کنیم بنابراین به جای استفاده از ماتریس در ماتریس Vاز ماتریس Uاستفاده می کنیم. در ماتریس U، بردارهای ستونی مستقل خطی هستند. این بردارهای ستونی از دید پردازش زبان طبیعی همان مفاهیم مستقل از هم می‌باشند. بنابراین از دید معنایی، ماتریس U، ماتریس کلمه- مفهوم است. با توجه به اینکه روش پیشنهادی در دو مدل مبتنی بر پیکره و مدل عمومی ارائه شده است بنابراین ادامه بحث را در قالب این دو مدل تشریح می کنیم.

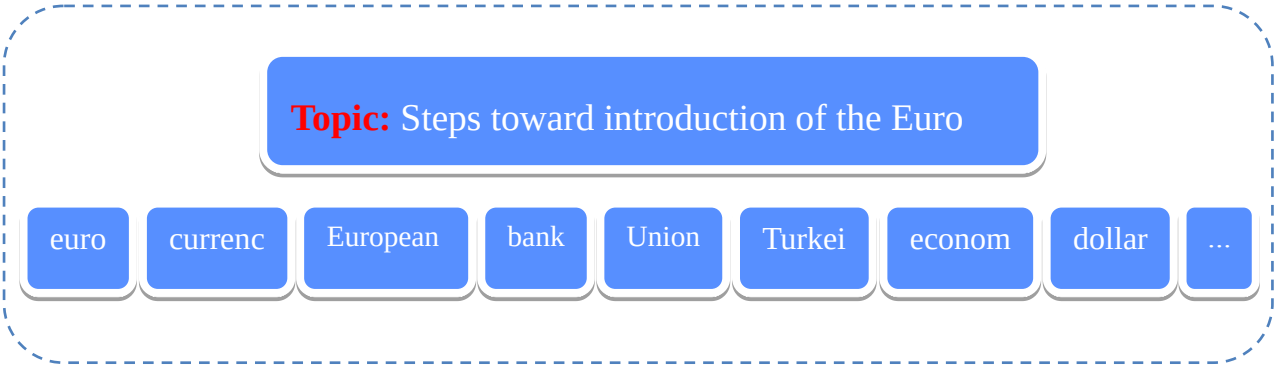
۱-۱-۲-۳- مدل مبتنی بر پیکره

در مدل مبتنی بر پیکره همانطور که گفتیم ماتریس کلمه-سند برروی کل پیکره ساخته می شود. سپس SVD برروی آن اعمال شده و ماتریس U بدست می آید. در این مرحله باید مفاهیم را به موضوعات مربوطه منتسب نمایم.برای اینکار فاصله کسینوسی بین بردارهای مفاهیم $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ و بردارهای اسناد $D_j = [d_{j1}, d_{j2}, \dots, d_{jm}]$ محاسبه می‌شود. فاصله کسینوسی بین این دو بردار به صورت زیر محاسبه می شود :

$$cos(C_i, D_j) = \frac{\sum_k c_{ik} d_{jk}}{\sqrt{\sum_k (c_{ik})^2} \times \sqrt{\sum_k (d_{jk})^2}}$$

نحوه انتخاب ابعاد کامل شود.

می شود. بر همین اساس نزدیکترین بردار به بردار کلمات هر موضوع را به عنوان نماینده آن موضوع انتخاب می کنیم. در تصویر زیر نمونه ای از این بردارها را که به موضوع "یورو" از مجموعه داده های DUC2007 تخصیص داده شده است، مشاهده می کنید:



شکل 8- بردار کلمات مفهوم موضوع 703

کلماتی که در تصویر ؟؟؟ مشاهده می کنید، بخشی از بردار مفهوم انتساب داده شده به موضوع "گام هایی به سوی معرفی یورو" می باشند که بیشترین وزن را در بین سایر کلمات بردار دارند. همانطور که در تصویر هم این کلمات بسیار در ارتباط با حوزه یورو می باشند و به خوبی می توانند زمینه مربوط به این موضوع را نمایش دهند.

تاثیر مقادیر ویژه در انتخاب مفاهیم لحاظ شود؟؟؟ کامل شود.

۲-۱-۲-۳- مدل عمومی

در مدل عمومی بجای ساخت ماتریس برروی کل ۴۵ تا موضوع مجموعه داده های DUC2007، برای هر یک

از موضوعات بصورت جداگانه این ماتریس را ساخته و مراحل را دنبال می کنیم. تنها تفاوت پیاده سازی این مدل و مدل قبلی در این است که دیگر نیاز به محاسبه فاصله کسینوسی بین بردار مفاهیم و بردار کلمات کلاستر نمی باشد چراکه ماتریس مورد نظر فقط بروی همان کلاستر ساخته شده است بنابراین مفاهیم استخراج شده منتسب به همان کلاستر موضوعی خواهد بود. بنابراین در این مدل، در ابتدا ماتریس کلمه-سند برروی هر موضوع ساخته می شود. سپس ماتریس ساخته شده به عنوان ورودی به الگوریتم SVD داده شده تا پردازش شود و ماتریس ماتریس ستونی U ساخته شود. اولین بردار ستونی از این ماتریس مهمترین و اصلیترین بردار و در حقیقت زمینه اصلی این اسناد می باشد. این موضوع هم بصورت شهودی و هم با بررسی تفاوت میزان بزرگی مقادیر ویژه بردار اول با سایر بردارها قابل توجیه می باشد.

اگر به صورت شهودی بخواهیم به این قضیه نگاه کنیم، با توجه به اینکه در حوزه خلاصه سازی چندسندی، زمینه اصلی اسناد مرتبط با یک موضوع یکسان می باشد، بنابراین تعداد زیادی از جملات شبیه به هم و در یک حوزه معنایی هستند. در نتیجه تعدادی از کلمات که در حوزه مربوط به زمینه هستند رخداد همزمانشان در متن بسیار بیشتر از بقیه کلمات می‌شود. با توجه به این نکته که وظیفه LSA هم استخراج کلمات هم رخداد می باشد، بنابراین بدیهی است که همیشه اولین بردار مفهوم استخراجی مهمترین و با اهمیت ترین بردار مفهوم که همان زمینه اصلی می باشد هست (مقادیر ویژه در ماتریس وسطی به صورت نزولی مرتب می باشد).

بررسی نموداری

نمودار تفاوت بزرگی مقادیر ویژه بردار اول با سایر بردار ها؟؟؟ کامل شود..

۲-۲-۳- مرتب سازی جملات براساس شباهتشان با زمینه متن

پس از استخراج زمینه اصلی مرتبط با هر موضوع، جملات را بر اساس میزان شباهتشان با زمینه اصلی موضوع مرتب می نماییم. برای این قسمت هم از فاصله کسینوسی استفاده شده است. در ابتدا بردار کلمات هر جمله را می سازیم.سپس مجددا از فاصله کسینوسی بین بردار جملات $S_j = [s_{j1}, s_{j2}, \dots, s_{jm}]$و بردار مفهوم $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ برای مرتب کردن جملات استفاده می‌شود.

$$cos(C_i, S_j) = \frac{\sum_k c_{ik} s_{jk}}{\sqrt{\sum_k (c_{ik})^2} \times \sqrt{\sum_k (s_{jk})^2}}$$

پس از محاسبه فاصله کسینوسی بردارهای تمامی جملات با بردار زمینه اصلی، جملات براساس مقدار شباهت و به صورت نزولی مرتب می شوند. البته با توجه به اینکه ما شباهت جملات را با زمینه اصلی بدست آوردیم بنابراین طبیعتا تعدادی از این جملات شبیه به یکدیگر خواهند بود که در فاز بعدی این افزونگی را برطرف

خواهیم ساخت.نتایج بدست آمده از ارزیابی خلاصه سازی حاکی از دقت مناسب این مرحله مـی باشـد کـه در قسمت نتایج به آن اشاره خواهیم کرد. البته برای این که کارایی این مرحله بیشتر آشکار شود مـا دو آزمـایش دیگر هم ترتیب داده ایم که در آن کیفیت انتخاب جملات را بررسی خواهیم کرد.

۱. ***آزمایش اول :*** کنفرانس DUC علاوه بر انتشار مجموعه داده های مورد نیاز برای خلاصه سازی متن، نتایج حاصل از خلاصه سازی تعدادی از سیستم های معروف برروی این داده ها را هم منتشر کرده است. در DUC2001 حدود ۱۲ سیستم شرکت داشته اند که در DUC2007 تعداد آنها به ۳۲ سیستم افزایش پیدا کرده است. تعدادی از این ۳۲ سیستم، از سیستم های معروف و بسیار قدرتمند در زمینه خلاصه سازی متن می باشند که از تکنیک های بسیار متنوعی برای خلاصه سازی متن استفاده می کنند. بعضی از این سیستم ها گام هایی برای خلاصه سازی چکیده ای برداشته اند و علاوه بر استخراج جملات مهم، بخش هایی از جمله که غیرضروری می باشد را حذف می نمایند. دقت این سیستم ها بقدری بالاست که عموما در مقالات مختلف نتایج را با میانگین این سیستم ها ارزیابی می کنند. بهمین دلیل در این آزمایش برای بررسی میزان دقت جملات انتخاب شده، در ابتدا ؟؟؟ تا موضوع را به صورت تصادفی انتخاب کردیم. سپس مراحل پیش پردارش و انتخاب زمینه و جملات متناسب با زمینه را اعمال کرده و جملات را بر اساس میزان اهمیتشان مرتب می کنیم. در این آزمایش می خواهیم بررسی کنیم که چه میزان از جملات با اهمیت تولید شده توسط سیستم پیشنهادی، در خلاصه تولید شده توسط بهترین سیستم های DUC2007 موجود می باشد. بهمین دلیل خلاصه های ۱۰ عدد از بهترین سیستم های موجود در DUC2007 که امتیازشان بالای میانگین بود را انتخاب نمودیم. سپس حدود ۱۵ درصد از جملات را از لیست جملات مرتب شده بر اساس اهمیت را انتخاب کردیم. با بررسی خلاصه های ۱۰ سیستم برتر DUC مشخص شد که بیش از ۸۴ درصد از این جملات در خلاصه های تولید شده توسط این سیستم ها موجود می باشد. این درصد بیانگر این موضوع است که لیست جملات مرتب شده دارای دقت مناسبی می باشد.

۲. ***آزمایش دوم :*** در آزمایش دوم میزان دقت جملات انتخاب شده بصورت مجزا بررسی شده است. برای این منظور یک موضوع به صورت تصادفی انتخاب شد. سپس فازهای ۱ و ۲ سیستم پیشنهادی برروی آن اعمال شد و جملات براساس اهمیتشان مرتب شدند. سپس ۳۰ جمله اول انتخاب شدند. برای بررسی دقت جملات انتخاب شده، تمامی حالاتی که از ۳۰ جمله می توان ۱۰ جمله(میزان فشرده سازی تقریبی برای داده های DUC2007) انتخاب کرد محاسبه شد که حدود ۳۰ میلیون حالت می شود. بنابراین ۳۰ میلیون فایل که هرکدام حاوی ۱۰ جمله از ۳۰ جمله تولید شده توسط سیستم پیشنهادی می باشد تولید شده و به عنوان ورودی به ابزار ارزیابی ROUGE داده شد (با توجه به توانایی رایانه مورد استفاده و همچنین حجم بالای عملیات ورودی/خروجی این عملیات

نزدیک به یک ماه زمان به خود اختصاص داد).با توجه به نتایج بدست آمده مشخص شد که در بین این حالات، حالت هایی وجود دارد که نتایجش بهتر از بهترین سیستم DUC2007 می باشد. این نتایج را در جدول زیر مشاهده می کنید:

System ID	ROUGE-2	Compresion Size (word)
1-2-3-4-8-9-16-20-21-29	0.14273	۲۸۸
1-2-3-4-8-9-20-22-25-29	0.13982	۲۷۶
1-2-3-4-8-9-14-16-20-29	0.13885	۲۶۹
1-2-3-4-8-10-16-20-28-29	0.13789	۲۷۱
1-2-3-4-8-14-16-17-25-29	0.13787	۲۵۲
1-2-3-4-8-9-16-20-26-29	0.13787	۲۶۱
1-2-3-4-8-9-21-22-26-29	0.13691	۲۷۳
1-2-4-8-9-16-20-21-28-29	0.13691	۲۵۵
1-2-3-4-8-9-17-20-22-29	0.1369	۲۵۹
1-2-3-4-8-9-20-21-29-30	0.1369	25۲
BEST DUC (15)	0.13599	۲۵۰

جدول ۷- نتایج ارزیابی آزمایش دوم با معیار ROUGE-2

System ID	ROUGE-SU4	Compresion Size (word)
1-2-3-4-8-9-14-22-28-29	0.21631	۲۹۱
1-2-3-4-8-9-14-16-28-29	0.21518	۲۸۵
1-2-3-4-6-8-9-22-28-29	0.20931	276
1-2-4-8-9-14-17-22-28-29	0.20899	281
1-2-4-8-9-14-16-23-28-29	0.20801	272
1-2-3-4-8-9-20-27-28-29	0.20591	265
1-2-3-4-8-12-14-22-23-28	0.20117	259

1-2-4-5-9-14-16-19-28-29	0.19971	261
1-2-4-6-9-12-13-14-28-29	0.19955	258
1-2-4-7-8-14-16-20-23-29	0.19709	254
BEST DUC (15)	0.18899	۲۵۰

جدول ۸- نتایج ارزیابی آزمایش دوم با معیار ROUGE-SU4

در جداول (۷) و (۸) ستون SystemID شماره ۱۰ جمله انتخاب شده از بین ۳۰ جمله اول را نشان می دهد. ستون های ROUGE-2 و ROUGE-SU4 هم مربوط به معیار ارزیابی خلاصه ها می باشد که هم در فصل پیش و هم در بخش نتایج شرح داده شده اند. همانطور که در جداول هم مشخص می باشد حالت های زیادی را می توان یافت که نتایجی بهتر از بهترین سیستم DUC2007 کسب کرده باشند. البته در این جداول تنها بخشی از حالت ها نشان داده شده است. بنابراین با توجه به این نتایج می توان به صورت شهودی دریافت که جملات بر گردانده شده در این فاز جملات خوب و مناسبی می باشند.

۳-۳- حذف افزونگی با محاسبه شباهت معنایی مبتنی بر شبکه واژگان جملات

در مرحله قبلی جملات بر اساس میزان ارتباطشان با زمینه متن یا همان مفهوم اصلی مرتب شدند. طبیعتا بدلیل ماهیت چندسندی بودن پیکره، بسیاری از این جمله‌ها شبیه به هم هستند. در اینمرحله باید جملات تکراری از لحاظ معنا حذف شده و دیدگاه‌های مختلف استخراج شوند. در این فاز برخلاف فاز قبلی، استفاده از فاصله کسینوسی نمی تواند به خوبی شباهت بین جملات را نشان دهد. برای روشن‌تر شدن موضوع، جملات زیر را در نظر بگیرید:

S1 = *United StatesArmy successfully tested an anti-missile defense system.*

S2 = *U.S.militaryprojectile interceptor streaked into space and hit the target.*

S3 = *Iran's weekend test of a long-range missile underscored the need for a U.S. national missile defense system.*

جملات S1 و S2 تقریبا بیانگریک خبر هستند ولی چون از لحاظ نحوی کلمه مشترکی ندارند فاصله کسینوسی آنها صفر می‌شود (از دید فاصله کسینوسی شبیه هم نیستند چون در محاسبه فاصله کسینوسی بردارها در کلمات یکسان ضرب داخلی می شوند و چون کلمات یکسانی از لحاظ نحوی وجود ندارد بنابراین ضرب داخلی آنها صفر شده و در نتیجه فاصله کسینوسی آنها صفر می شود) : اما با احتساب فاصله کسینوسی، جملات S1 و S3 چون دارای کلمات مشترک هستند، شبیه به هم می‌باشند که البته این تشخیص اشتباه است.

بنابراینبا توجه به این که فاصله کسینوسیصرفا به شباهت نحویتوجه می کند و موقعیت و نقش کلمات در جمله را در نظر نمی گیرد نمی‌تواند به خوبی شباهت بین جملات را محاسبه نماید. استفاده از فاصله کسینوسی در فاز قبل این مشکل را نداشت. چراکه در فاز قبلی، فاصله کسینوسی بین بردار جملات و بردار مفاهیم (که در برگیرنده همه کلمات موجود در متن با وزن خاص خود هست) محاسبه شد و هدف صرفا استخراج جملات مرتبط با زمینه متن بود. برای حل این مشکل، شباهت معنایی بین جملات محاسبه می-شود.

۱-۳-۳- شباهت معنایی مبتنی بر شبکه واژگان جملات

در بخش قبل به مشکلات فاصله کسینوسی برای اندازه‌گیری میزان شباهت جملات به همدیگر اشاره شد. در این بخش سعی شده تا با در نظر گرفتن مشکلات روشهای محاسبه میزان شباهت بین جملات روش جدیدی برای افزایش دقت محاسبه شباهت ارائه شود. به طور کلی روش های اندازه گیری شباهت بین جملات در سه دسته کلی می توان قرار داد : ۱) دسته اول روش های مبتنی بر شباهت لغوی می باشد[Kon 05]. در این دسته میزان انطباق لغوی بین کلمات دو جمله، ملاک شباهت دو جمله می باشد. عموما در این روش ها از الگوریتم محاسبه طولانی ترین زیر رشته مشترک [All 86]استفاده می شود. این دسته از روش ها قادر به تشخیص شباهت معنایی بین جملات نمی باشند. ۲) دسته دوم روش های مبتنی بر پیکره می باشند که شباهت جملات را براساس روش های آماری-مبتنی بر پیکره اندازهگیری شباهت بین کلمات که در فصل قبل اشاره شده، محاسبه می کنند[Jia 97]. ۳) دسته سوم هم روش های مبتنی بر[Mih 06]شباهت معنایی می باشندکه عموما از شبکه واژگان برای اندازه گیری شباهت بین جملات استفاده می کنند.

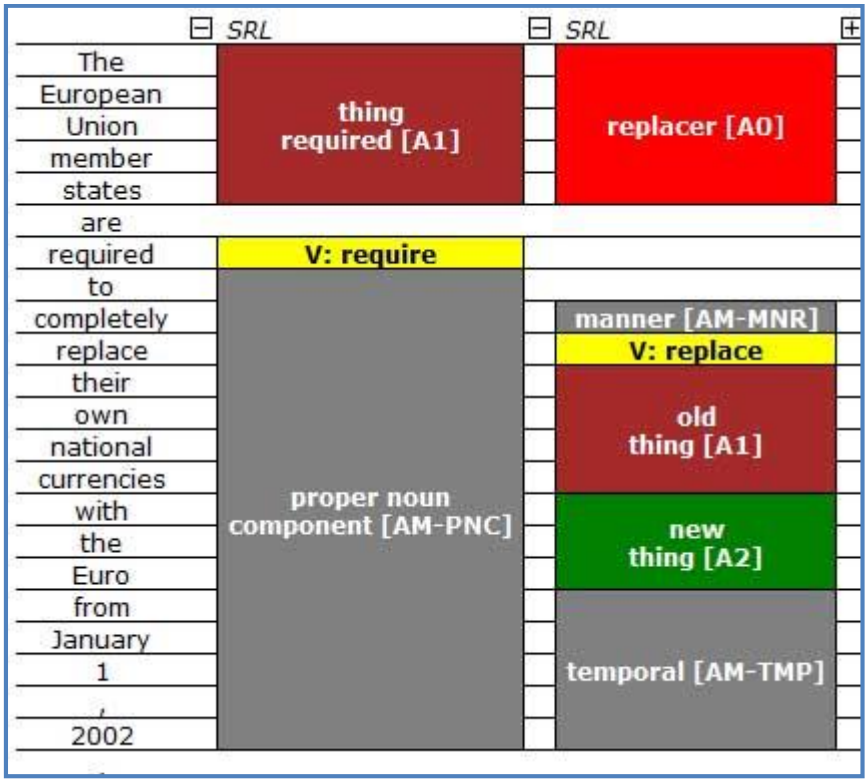
در این بخش روشی برای اندازه گیری شباهت بین جملات ارائه شده است که مبتنی بر شبکه واژگان و بر چسب زنی معنایی می باشد. برای اندازه‌گیری شباهت معنایی بین دو جمله، ابتدا اجزای آنها با استفاده از ابزار برچسب زنی معناییجدا شده و سپس با استفاده از شبکه واژگان ارتباط بین کلمات در نقش‌های معنایی یکسان محاسبه می‌شود. نظیر باشند، آنگاه از لحاظ معنایی مرتبط با هم خواهند بود. همچنین در روش پیشنهادی از موقعیت کلمات در عبارات اسمی و فعلی به عنوان معیاری جدید برای تعیین اهمیت کلمه استفاده شده است. در ادامه به جزییات این مدل ارائه شده است.

کامل شود؛ حتی شکل هم می توان کشید :

۱-۱-۳-۳- استخراج نقش های معنایی جملات

نقش معنایی نقشی است که یک جز در ارتباط با فعل جمله ایفا می کند. در این پایان نامه از ابزار برچسب زنی معنایی دانشگاه Illinois At Urbana Champaignکه از قوی ترین ابزارهای برچسب زنی معنایی می باشد استفاده شده است. با استفاده از این ابزار، نقش‌های معنایی جمله نظیر فاعل، مفعول مستقیم و غیر مستقیم و ... در ارتباط با فعل جمله مشخص می‌شود. بعبارت دیگر هر فعل در جمله با نقش های تکمیل کننده اش برچسب زده می شود که در این پایان نامه آن را سطح می نامیم. بنابراین به تعداد فعل های جمله، سطوح مختلفی داریم.در تصویر ؟؟؟ نمونه ای از خروجی این ابزار برای جمله ورودی زیر نشان داده شده است.

Example 1:The European Union member states are required to completely replace their own national currencies with the Euro from January 1, 2002.



شکل ۹- نمونه از خروجی ابزار برچسب زنی دانشگاه ایلینویز

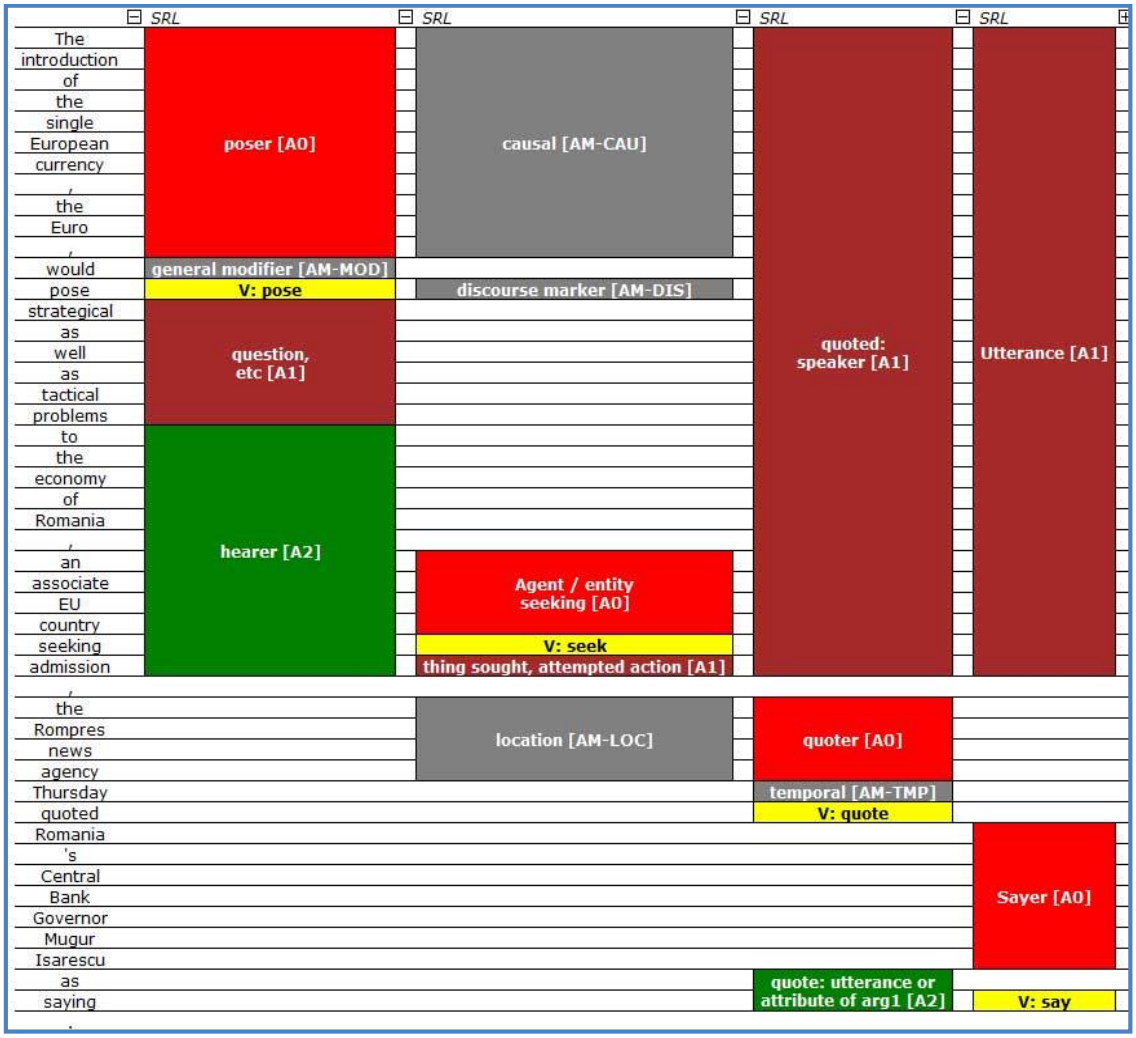
در جمله داده شده دو فعل replace و require وجود دارد و بهمین دلیل دو سطح هم متناسب با هر فعل

تعریف شده اس. در سطح اول که متناسب با فعل require می باشد سه نقش معنایی تشخیص داده شده است. "The European union member states" به عنوان مفعول مستقیم، "required" به عنوان فعل و "to completely replace their own national currencies with the ero from January 1, 2002" به عنوان نقش قیدی. در سطح دوم هم متناسب با فعل replace پنج نقش معنایی تشخیص داده شده است.

لازم به ذکر است که در مواردی که عموما طول جملات طولانی تر می شود، تعداد این سطوح هم افزایش می یابد. به عنوان مثال به جمله طولانی زیر را درنظر بگیرید:

Example2: The introduction of the single European currency, the Euro, would pose strategical as well as tactical problems to the economy of Romania, an associate EU country seeking admission, the Rompres news agency Thursday quoted Romania's Central Bank Governor Mugur Isarescu as saying.

خروجی برچسب زده آن به صورت زیر می باشد که شامل چهار سطح متناسب با چهار فعل آن می باشد.



شکل ۱۰- نمونه خروجی ابزار SRL برای جملات طولانی

اما همانطور که پیش تر هم اشاره شد، برای محاسبه میزان شباهت جملات، شباهت معنایی کلمات آنها را در نقش های معنایی یکسان، توسط شبکه واژگان اندازه گیری می کنیم. مهمترین چالشی که در این قسمت با آن مواجه هستیم وجود سطوح برچسب زده متفاوت از هم می باشد که هرکدام شامل چندین نقش مجزا از یکدیگر هستند. به عنوان نمونه در جمله مثال ۲ چهار سطح وجود دارد که هر کدام برچسب "A0" یا همان فاعل را دارند. چالشی که در این جا مطرح می شود این است که کدام یک از این برچسب ها را برای مقایسه در نقش فاعلی باید انتخاب نمود. در کار مشابه ای (البته مشابهت صرفا از لحاظ استفاده از برچسب زنی معنایی می باشد) که در [Aks 09] انجام گرفته است، سطحی که بیشترین کلمات را پوشش می دهد به عنوان سطح اصلی در نظر گرفته شده و مقایسه ها تنها در نقش های این سطح صورت می گیرد. با توجه به این روش در جمله مثال ۲، سطح چهار یا همان آخرین سطح به عنوان اصلی ترین سطح در نظر گرفته می شود. این شیوه مقایسه دارای ایرادات مهمی می باشد که در ذیل به آنها اشاره شده است:

۱) ***عدم در نظر گرفتن حالت های مختلف جمله:*** پر واضح است که جملات را می توان به شیوه های مختلفی بیان نمود. گاه با حفظ معنا ساختار جملات به کلی عوض می شود و گاه هم با تغییرات جزیی می توان جملات جدید تولید نمود. مثلا جمله مثال ۱ را به شکل زیر هم می توان بیان کرد:

> Example۳: The European Union member states completely replace their own national currencies with the Euro from January 1, 2002.

این دو جمله دقیقا یک خبر و مفهوم را می رسانند. حال اگر با توجه به روش پیشنهادی در [Aks ۰۹] طولانی ترین سطح را برای جمله ۱ انتخاب کنیم باید سطح اول را به عنوان سطح اصلی در نظر بگیریم. با توجه به خروجی برچسب زده جمله ۳ که در زیر نشان داده شده است، انتخاب سطح اول برای جمله ۱ به عنوان سطح اصلی کاملا اشتباه می باشد.

شکل 11- خروجی برچسب زده شده جمله ۳

۲) ***دقت پایین:*** با توجه به اینکه همیشه طولانی ترین سطح به عنوان سطح اصلی در نظر گرفته می شود و عموما در طولانی ترین سطح نقش ها کلی می باشد، بنابراین امکان اندازه گیری شباهت دقیق وجود ندارد. به عنوان نمونه همانطور که پیش تر اشاره شده در جمله مثال ۲ سطح چهارم به

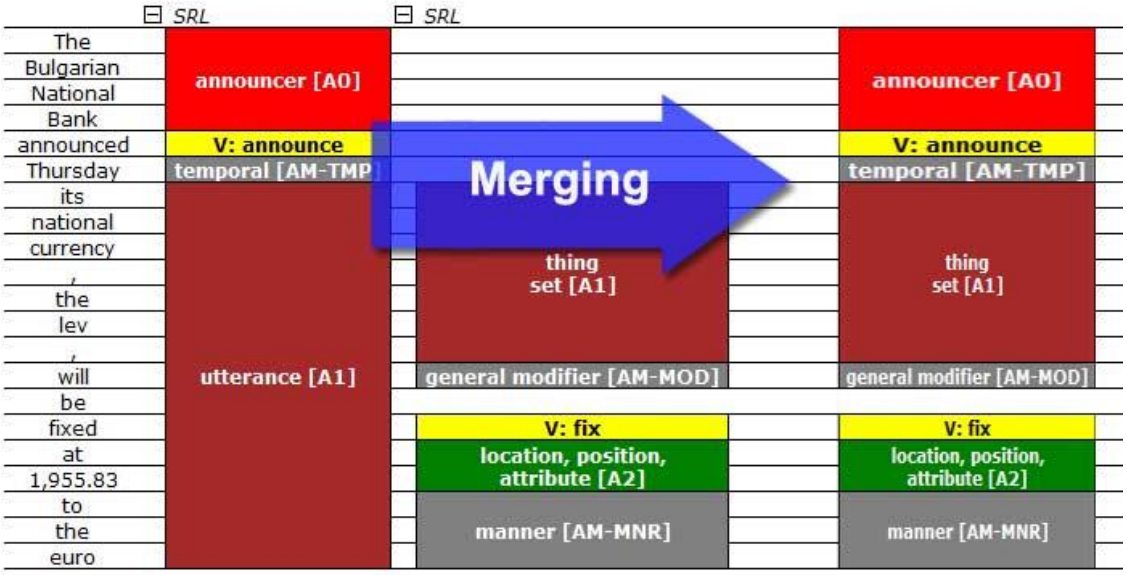
عنوان سطح اصلی در نظر گرفته شده است. همانطور که در تصویر ؟؟؟ مشخص است، این سطح دارای سه نقش اصلی فاعلی "A0" ، مفعولی "A1" و فعلی "V" می باشد و این درحالی است که نقش مفعولی " The introduction of the single European currency, the Euro, would pose strategical as well as tactical problems to the economy of Romania, an associate EU country seeking admission" در حقیقت یک جمله دیگر بوده و شامل برچسب های مربوط به خود می باشد.بنابر این مقایسه در این سه نقش برای این مثال، تفاوت چندانی با روش های پیش نخواهد داشت چراکه نقش ها بسیار کلی می باشند.

به مشکلات مربوط به تعیین سطح در قسمت قبل اشاره شد. در این قسمت ۲ راه کار برای حل این مشکل پیشنهاد شده که با پوشش مشکلات مربوط به انتخاب سطح باعث افزایش دقت خواهد شد:

۱) ***مقایسه در تمامی سطوح :*** بر خلاف روش [Aks 09] که تنها از یک سطح برای مقایسه استفاده می کرد، در روش جدید تمامی سطوح جمله اول را با تمام سطوح جمله دوم در نقش های یکسان مقایسه کرده و بیشترین مقدار را به عنوان میزان شباهت دو جمله در نظر می گیریم. بدین ترتیب مشکل مربوط به ساختار های مختلف یک جمله بر طرف می شود. به عنوان نمونه برای جملات مثال ۱ و ۳، سطوح ۱ و ۲ جمله مثال ۱ را با تنها سطح جمله مثال ۳ مقایسه می شود و چون ماکزیمم این دو مقایسه که همان شباهت سطح ۲ جمله مثال ۱ و سطح ۱ جمله مثال ۳ می باشد، به عنوان میزان شباهت این دو جمله بر گردانده می شود.

۲) تولید سطوح جدید: یکی از مشکلات روش [Aks 09]دقت پایین محاسبه شباهت بدلیل عدم در نظر گرفتن نقش های جزیی تر می باشد. برای حل این مشکل در روش پیشنهادی سطوحی که در حقیقت جزی از سطوح دیگر می باشند با آنها ادغام شده و به عنوان سطوح جدید جهت مقایسه در نظر گرفته می شوند.در مثال زیبر با ادغام سطوح ۱ و ۲، سطح جدید ۳ را به مجموعه سطوح اضافه می کنیم:

Example۴:The Bulgarian National Bank announced Thursday its national currency, the lev, will be fixed at 1,955.83 to the euro.



شکل ۱۲- تولید سطوح جدید

الگوریتم نحوه ادغام سطوح فعلی و تولید سطوح جدید در ضمایم آورده شده است. کامل شود.

۲-۱-۳-۳- اندازه گیری شباهت بین کلمات بر اساس شبکه واژگانی

بعد از استخراج نقش های معنایی باید شباهت کلماتی که در این نقش ها قرار می گیرند را با استفاده از شبکه واژگان محاسبه نمایند. در فصل قبل به روابطی که در شبکه واژگان تعریف می شوند اشاره شد. در ادامه سه دسته اصلی از این روابط مورد استفاده قرار می گیرد.

- نوع یال یا رابطه

مهمترین فاکتور موثر در اندازه گیری شباهت بین کلمات توسط شبکه واژگان، نوع یالها می‌باشد. در WordNet مهمترین روابط به ترتیب عبارتند از روابط ابرمفهوم/زیرمفهوم (IS-A)شامل شمول[۱] و مشمول[۲]، رابطه‌ی هم معنایی (Synonymy)و رابطه‌ی جزئیت (Part-Of) شامل روابط جزئی از کل[۳] و کل مربوط به جزء.[۴] بقیه‌ی انواع روابط مانند علت و معلول، material-productو event-roleدرصد کمی از روابط WordNet را تشکیل می‌دهند.

انواع مختلف لینک در WordNet با توجه به تابع تشابه، وزن‌های متفاوتی می‌گیرند. مهم‌ترین رابطه، یال

---

۱ Hyperrnym

۲ Hyponym

۳ Meronym

۴ Holonym

ترادف (Synonymy)می‌باشد که نشان می‌دهد که دو کلمه در دو انتهای یال یکسان هستند. علاوه بر این وزن شباهت لینک IS-Aاز وزن شباهت لینک Part-Of بیشتر می‌باشد. بنابراین ما وزن رابطه بین دو کلمه را براساس نوع آن به صورت زیر تعریف می‌کنیم (رابطه‌ی ؟؟؟ ۳-۱۳):

$$Sim_{wordnet}(w_1, w_2) \propto \begin{cases} 1, & type(w_1, w_2) = synonymy \ or \ w_1 = w_2 \\ \circ/95 & type(w_1, w_2) = IS - A \\ \circ/85 & type(w_1, w_2) = Part - Of \end{cases} \qquad (۱۳-۳)$$

- ویژگی‌های گره‌های شبکه واژگان

در سلسله‌مراتب آنتولوژی WordNet ویژگی‌های تمامی گره‌ها به تفصیل تشریح شده است. برای ویژگی تعاریف مختلفی می‌توان متصور شد. در اینجا از تعریف [???PET06] در بیان X-Similarityاستفاده کرده‌ایم. در اینجا ویژگی‌ها به معنای مجموعه‌های تعاریف[۱]کلمه‌ای ویا Synset ها می‌باشد. مجموعه‌های تعریف کلمه‌ای شامل کلماتی است که از تعاریف مفاهیم[۲]در WordNet استخراج شده است. بنابراین می‌توان گفت که دو کلمه از لحاظ ویژگی‌ها با یکدیگر تشابه دارند هرگاه Synset ها و یا مجموعه‌های تعریفی خودشان مشابه باشند. یک نمونه از این تعاریف در مثال زیر آورده شده است:

Example 5:

1. dollar -- (the basic monetary unit in many countries; equal to 100 cents)

2. dollar, dollar bill, one dollar bill, buck, clam -- (a piece of paper money worth one dollar)

3. dollar -- (a United States coin worth one dollar; "the dollar coin has never been popular in the United States")

4. dollar, dollar mark, dollar sign -- (a symbol of commercialism or greed; "he worships the almighty dollar"; "the dollar sign means little to him")

به ترتیب اشتراک ویژگی‌هایدو گره را تعریف می‌کنیم (رابطه‌ی ۳-۲۰):

$$Sim_f(w_1, w_2) = \frac{|A \cap B|}{|A \cup B|} \qquad (۲۰-۳)$$

به‌طوری که Aو Bمجموعه‌ی کلمات توضیحی (بعد از حذف کلمات ایست) و یا Synset- های دو کلمه w_1 و w_2 هستند.

۱- Description Sets

۲- Glosses

- سایر معیارهای مورد استفاده در شباهت کلمات

معیارهای دیگری نیز برای اندازه گیری شباهت بین کلمات وجود دارد که از جمله آنها می توان به عمق گره در ساختار شبکه واژگان، چگالی گره ها و درجه خوشه ها در آنتولوژی شبکه واژگان اشاره کرد. با بررسی های انجام شده مشخص شد که استفاده از این ویژگی ها نه تنها در اندازه گیری شباهت جملات مفید نیست بلکه ممکن است اثر منفی هم داشته باشد. به عنوان نمونه در مثال ۶، در دو جمله ۱ و۲، کلمات "kenyan" و "pakistan" دارای پدر مشترک بوده و هر دو جزء "country" می باشند. اما در خلاصه سازی چند سندی این دو جمله را باید به عنوان جملات مجزا از هم که هر کدام دیدگاهی را در ارتباط با موضوع داده شده بیان می کنند لحاظ کرد و بنابراین نباید این دو جمله را شبیه به هم در نظر بگیریم. به همین ترتیب دو پارامتر دیگر اشاره شده دیگر هم نمی تواند برای این بخش مفید فایده باشد.

Example 6:

Water shortage in the Pakistani capital of Islamabad is expected to deteriorate as water reservoir can meet demand of the consumers just for 12 days.

The acute water shortage in the Kenyan capital Nairobi entered its third month Tuesday with water-borne diseases imminent.

۳-۱-۳-۳- وزن دهی به کلمات بر اساس موقعیتشان در درخت تجزیه جمله

در بخش های قبلی کلمات هر دو جمله را در نقش های معنایی یکسان با هم مقایسه کردیم. اما نوع و موقعیت این کلمات را در مقایسه دخیل نکردیم. با کمی دقت و بررسی مشخص می شود که موقعیت کلمات و نوعشان (اسم، صفت، فعل، قید زمان، قید مکان و ...) هم می تواند تاثیر بسزایی در اندازه گیری شباهت بین دو جمله داشته باشد. برای پی بردن به اهمیت موضوع مثال ۷ آورده شده است:

Example 7:European Central Bank President, Wim Duisenberg said Thursday that Europe 's new single currency , the euro , wo n't compete with the U.S. dollar as the currency of choice for foreign reserves.

خروجی بر چسب گذاری شده این جمله هم در تصویر ؟؟؟ آورده شده است.همانطور که در تصویر هم مشخص است ابزار برچسب زنی ایلنویز در خروجی خود نمایشی از درخت تجزیه جمله را هم در لبه "charniak" ارائه می دهد. این خروجی علاوه بر درخت تجزیه، جزییات بیشتری از جمله نظیر نوع کلمات (اسم، صفت، فعل) هم ارائه می دهد.نمایش درختی این اطلاعات هم در تصویر ؟؟؟ آورده شده است. از این

اطلاعات در ادامه استفاده شده است.

در خروجی ابزار برچسب زنـی، عبـارت "European Central Bank President, Wim Duisenberg" در نقش فاعلی ظاهر شده است و شامل ۶ کلمه می باشد. در روش های معمول محاسبه شباهت معنایی بین ایـن کلمات اهمیت قائل نمی شویم. اما با توجه به خاصیت موصوف و موصوف صـفت و مضـاف و مضـاف الیـه مـی توان این نکته را برداشت کرد که همیشه اهمیت موصوف بیشتر از موصوف صفت و اهمیـت مضـاف بیشـتر از مضاف الیه می باشد و بنابراین در عبارات اسمی و فعلی اهمیت کلمات از چپ به راست افزایش می یابـد.یعنی اهمیت "Duisenberg" از "Wim" بیشتر و اهمیت"president" بیشتر از"Bank" و اهمیـت "Bank" هـم بیشـتراز "Central" مـی باشـد. در عبـارت اسـمی "new single currency" هـم اهمیـت"currency" از "single" و "new" بیشتر می باشد.

⊟ SRL	⊟ SRL	⊞ Charniak
European	Sayer [A0]	(S1 (S (NP (NNP European)
Central		(NNP Central)
Bank		(NNP Bank)
President		(NNP President)
Wim		(NNP Wim)
Duisenberg		(NNP Duisenberg))
said	V: say	(VP (VBD said)
Thursday	temporal [AM-TMP]	(NP (NNP Thursday))
that	Utterance [A1]	(SBAR (IN that)
Europe	competitor [A0]	(S (NP (NP (NP (NNP Europe)
's		(POS 's))
new		(JJ new)
single		(JJ single)
currency		(NN currency))
,		(, ,)
the		(NP (DT the)
euro		(NNP euro))
,		(, ,))
wo	general modifier [AM-MOD]	(VP (MD wo)
n't	negation [AM-NEG]	(RB n't)
compete	V: compete	(VP (VB compete)
with	opponent [A1]	(PP (IN with)
the		(NP (DT the)
U.S.		(NNP U.S.)
dollar		(NN dollar)))
as	prize [A2]	(PP (IN as)
the		(NP (NP (DT the)
currency		(NN currency))
of		(PP (IN of)
choice		(NP (NN choice)))
for		(PP (IN for)
foreign		(NP (JJ foreign)
reserves		(NNS reserves))))))))))
.		(. .)))

شکل ۱۳- خروجی برچسب زده شده جمله مثال ۷

بنابراین با توجه به این موضوع و در نظر گرفتن درخت تجزیه فاکتورهـای زیـر را بـرای افـزایش دقـت انـدازه گیری شباهت اعمال می کنیم:

- موقعیت کلمات در عبارات اسمی و فعلی: در هر عبارت اسمی[۱] و فعلی[۲] بر اساس موقعیت کلمات به آنها و به صورت زیر به آنها وزن داده می شود:

$$weigth_{position}(w_i) = 1 + \frac{position_i}{\alpha * N_{phrase}}$$

که N_{phrase} تعداد کلمات موجود در عبارت اسمی یا فعلی، $position_i$ موقعیت کلمه i ام در عبارت و α هم ضریت تاثیرگذاری می باشد که برابر ۲ در نظر گرفتیم.

- نوع کلمه :بر اساس اینکه کلمه مورد نظر اسم، فعل و یا صفت باشد اهمیت آن فرق می کند و بر همین اساس نیز به کلمات وزن می دهیم که بصورت زیر محاسبه می شود:

$$weight_{type1} = \begin{cases} 1.2 & type = noun \\ 1.15 & type = verb \\ 1.1 & type = adjective \\ 1 & type = other \end{cases}$$

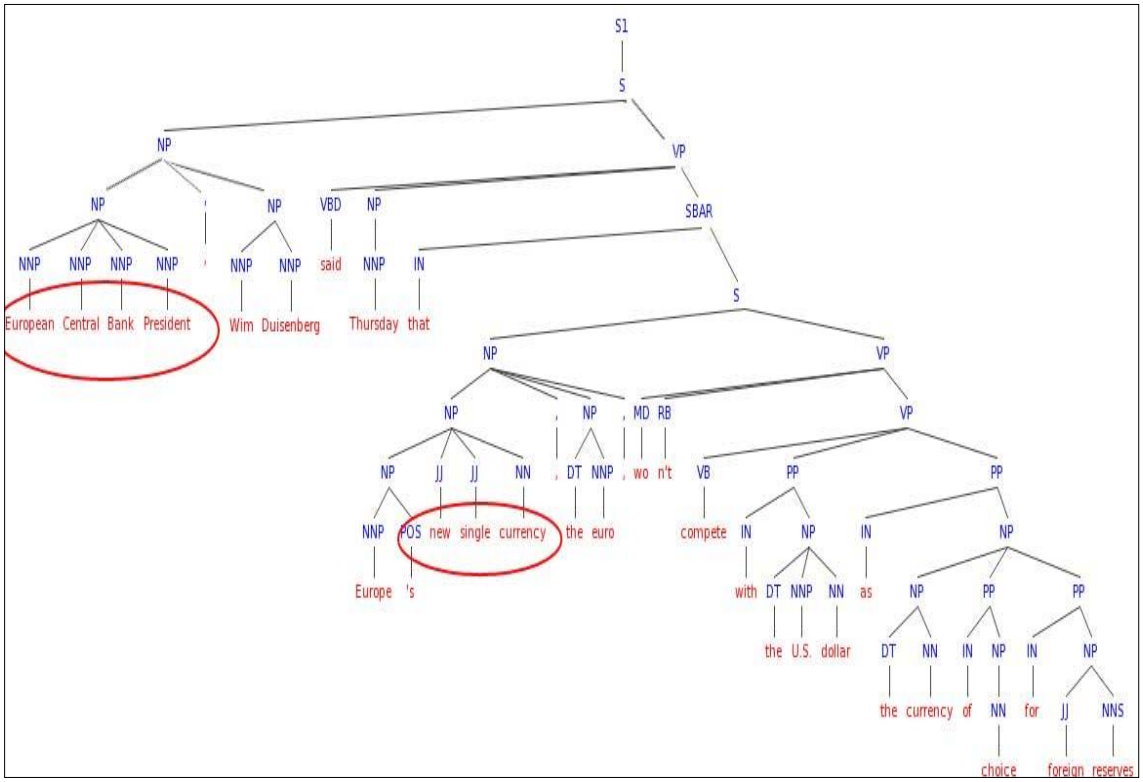
- اسامی خاص و اسم مکان و زمان : اسامی خاص و همچنین اسامی اماکن و اسم زمان هم همواره از اهمیت بالایی برخوردار می باشند. به همین دلیل معیاری هم برای برجسته کردن این نوع کلمات در نظر گرفته شده است:

$$weight_{type2} = \begin{cases} 1.3 & type = proper\ noun \\ 1.2 & type = location\ or\ temporal \\ 1 & type = other \end{cases}$$

---

۱ Noun phrase

۲ Verb phrase



شکل ۱۴- نمایش درخت تجزیه جمله مثال ۷

۴-۱-۳-۳- الگوریتم محاسبه شباهت جملات و حذف افزونگی

پس از توضیح قسمت های مختلف محاسبه شباهت بین جملات در این بخش الگوریتم کامل محاسبه جملات شبیه به هم و حذف افزونگی توضیح داده شده است. گام های محاسبه شباهت به صورت مختصر در ذیل آورده شده است:

۱) مشخص کردن اهمیت و موقعیت کلمات : همانطور که توضیح دادیم نوع و موقعیت کلمات می تواند بر ارزش آنها در مقایسه بیفزاید. بنابر این وزن نهایی هر کلمه را به صورت ترکیبی و به صورت زیر محاسبه می کنیم :

$$weight\ (W_i) = \ (\rho_1 \times weigth_{position}(w_i)) + (\rho_2 \times weight_{type1}) + (\rho_3 \times weight_{type1})$$

که ρ_1 ، ρ_2 و ρ_3 برابر با ۱/۳ در نظر گرفته شده است.

۲) پس از مشخص شدن اهمیت کلمات میزان شباهت دو کلمه با استفاده از شبکه واژگان به صورت زیر محاسبه می شود:

$$wsim(W_i, W_j) = \left(\frac{weight\ (W_i) + \ weight\ (W_j)}{2}\right) \times \ ((Sim_{wordnet}(W_i, W_j) + Sim_f(W_i, W_j))$$

۳) نحوه محاسبه شباهت کلمات بین دو سطح مختلف از خروجی برچسب زده شده دو جمله A و B در ذیل آورده شده است. فرض کنید که $R_{L_i^A} = \{r_1, r_2, \dots, r_n\}$ مجموعه نقش های موجود در سطح L_iام جمله A و $R_{L_j^B} = \{r_1, r_2, \dots, r_m\}$ مجموعه نقش های موجود در سطح L_j ام از جمله Bباشد. مجموعه $CR_{LiLj} = \{r_1, r_2, \dots, r_{intersectNum}\}$ را اشتراک نقش های معنایی بین دو سطح در نظر می گیریم. مجموعه$Terms_r(L_i^A) = \{t_1, t_2, \dots, t_{termNum}\}$ هم کلمات موجود در نقش r ام از سطح L جمله A می باشد (با این فرض که $|Terms_r(L_j^B)| < |Terms_r(L_i^A)|$). شباهت در نقش های مشترک بین سطح i ام جمله A و سطح j ام جمله B به صورت زیر محاسبه می شود::

$$lsim(L_i^A, L_j^B) = \frac{\sum_{r=1}^{intersectNum}(roleImpact \times \sum_{i=1}^{termNum} maxSim_r(w_i))}{|L_i^A|}$$

$$maxSim_r(w_A) = \max_{w_B \in Terms_r(L_i^B)} wsim(w_A, w_B)$$

که تابع $maxSim_r(w_A)$ ماکزیمم شباهت کلمه w_Aدر نقش معنایی r ام از جمله Aرا با تمامی کلمات موجود در همین نقش ولی در جمله B محاسبه می کند.$roleImpact$ هم پارامتری جهت مشخص کردن اهمیت نقش‌ها می باشد. مسلما شباهت در نقش فاعلی با شباهت در نقش قیدی اهمیت یکسانی ندارند. میزان این پارامتر به صورت زیر در نظر گرفته شده است:

$$roleImpact = \begin{cases} 1 & type: A0 \\ 0.95 & type: A1\ or\ A2 \\ 0.85 & type: others \end{cases}$$

۴) شباهت بین دو جمله A و B به صورت زیر محاسبه می گردد:

$$ssim(A, B) = \max_{i\ \in levels\ of\ A,\ \ j \in levels\ of\ B} lsim(L_i^A, L_j^B)$$

و برابر ماکزیمم شباهت بین جفت سطح های جملات A و B می باشد.

۵) پس از مشخص شدن فرمول محاسبه شباهت بین جملات، باید جملاتی که از لحاظ معنایی شبیه به هم می باشند از لیست جملات مرتب شده در فاز قبلی حذف گردند. پس جمله اول لیست جملات مرتب شده (لیستی که در فاز قبلی بر اساس میزان اهمیت جملات با زمینه متن مرتب شده است) انتخاب کرده و آن را در لیست جملات خلاصه $list_{summary}$قرار می دهیم. سپس شباهت جمله

دوم را با تنها جمله موجود در لیست $list_{summary}$ محاسبه می کنیم. اگر میزان شباهت از مقدار آستانه $Threshold_{sim}$ کمتر بود آن را به $list_{summary}$ اضافه می کنیم در غیر این صورت اینکار را با جملات بعدی تکرار می کنیم تا اندازه $list_{summary}$ به میزان فشرده سازی مد نظر برسد (برای مجموعه داده های DUC2006 و DUC2007 میزان فشرده سازی ۲۵۰ کلمه بدون حذف کلمات ایست می باشد). لیست $list_{summary}$ شامل جملات خلاصه می باشد.

۴-۳- مرتب سازی جملات

با توجه به تعریفی که از خلاصه سازی چندسندی ارائه کردیم، خلاصه تولید شده در این روش می تواند شامل جملاتی از قسمتهای مختلف سندهای مختلف باشد. بنابراین طبیعی است که ترتیب جملات در خلاصه تولید شده باید بررسی شود تا متن از خوانایی مناسبی برخوردار گردد. البته باید یادآور شویم که برای بررسی میزان و کیفیت پیوستگی و خوانایی متن، ابزار مناسبی تاکنون ارائه نشده است و معمولا این ارزیابی بصورت دستی و انسانی انجام می شود و بهمین دلیل چون قابلیت مقایسه با سایر روش ها وجود ندارد، عموما در مقالات توجه چندانی به آن نمی شود. البته نتایج ارزیابی کیفیت خلاصه های سیستم های موجود در مجموعه DUC موجود می باشد ولی از آنجا که امکان دسترسی به افرادی که این سیستم ها را ارزیابی کرد اند نمی باشد، بنابراین قادر به بررسی میزان پیوستگی و خوانایی روش پیشنهادی نمی باشیم. از طرف دیگر در بحث خلاصه سازی چند سندی مرتب سازی جملات اساسا کار دشواری بوده و در عین حال ممکن است حالت های زیادی پیدا کرد که در آن متن از خوانایی مناسبی بر خوردار باشد. به هر حال در ادامه راه کار پیشنهادی برای چیدن مناسب جملات خلاصه آورده شده است.

- ارجاع ضمایر: در ابتدا باید ضمایر و ارجاعات موجود در متن اصلی با مرجعشان جایگزین گردند تا متن از خوانایی مناسبی بر خوردار باشد. برای این مهم باید از ابزار های ارجاع ضمایر استفاده کرد. یکی از معروف ترین این ابزار ها، ابزار گیتار می باشد که با زبان پرل نوشته شده است. اگر در متن اصلی ضمایر با ارجاعاتشان جایگزین شوند خلاصه های تولید شده هم از خوانایی بیشتری بر خوردار خواهند بود.
- تاریخ رویداد: اگر جملات خلاصه حاوی فیلد تاریخ باشند می توان آنها را بر اساس کمترین تاریخ مرتب نمود.
- استفاده از جنبه Continue مربوط به نظریه مرکزیت : کامل شود.

فصل چهارم:

بررسی نتایج

۴– بررسی نتایج

در این قسمت نتایج ارزیابی خلاصه سازی چندسندی برروی مجموعه داده های استاندارد DUC2006 و DUC2007 ارائه شده است. برای ارزیابی هم از ابزار معروف ارزیابی ROUGE و از معیارهای ROUGE-2 و ROUGE-SU4 (که در گزارشات مربوط به این دو مجموعه داده به عنوان بهترین معیار تشخیص داده شده اند) استفاده شده است. در این ارزیابی ها علاوه بر مقایسه با سیستم های شرکت کننده در کنفرانس DUC سیستم های زیر هم پیاده سازی شده اند:

- سیستم ارائه شده توسط Gong [Gon 01] که مبتنی بر آنالیز کلمه- سند است.
- سیستم پیشنهادی در پایان نامه با اعمال فاز ۱و۲ به نام CAT1.
- سیستم پیشنهادی در پایان نامه با اعمال تمامی فازها به نام CAT2.
- میانگین سیستم های DUC2007 هم با نام AVG ارائه شده است.

۴–۱– نتایج آزمایش روی داده های DUC2007

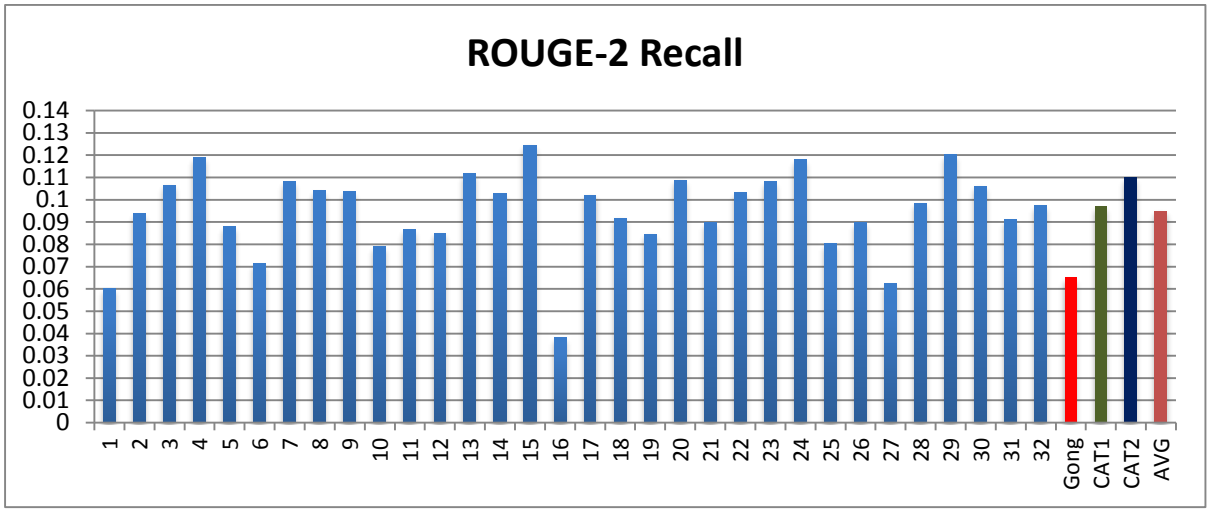
مدل مبتنی بر پیکره : پس از انجام مرحله پیش پردازش و ساخت ماتریس کلمه –سند برروی کل پیکره و اعمال سایر مراحل و مرتب سازی جملات خلاصه، ۲۵۰ کلمه از جملات مرتب شده را به ترتیب جدا کرده و برای پردازش به ابزار ROUGE می دهیم. ابزار ROUGE مبتنی بر سیستم عامل لیناکس بوده و با زبان پرل نوشته شده است.

پارامتر های مربوط به ابزار ROUGE مطابق با نتایج اولیه کنفرانس DUC برروی داده های DUC2007 به صورت زیر تنظیم شده است :

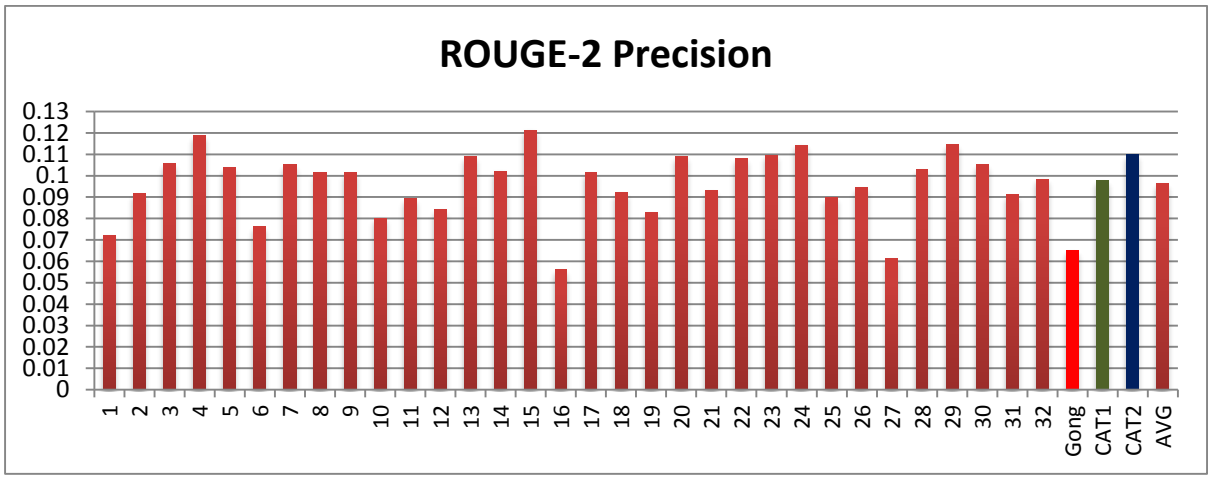
ROUGE run-time parameters:

ROUGE-1.5.5.pl -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d rougejk.in.

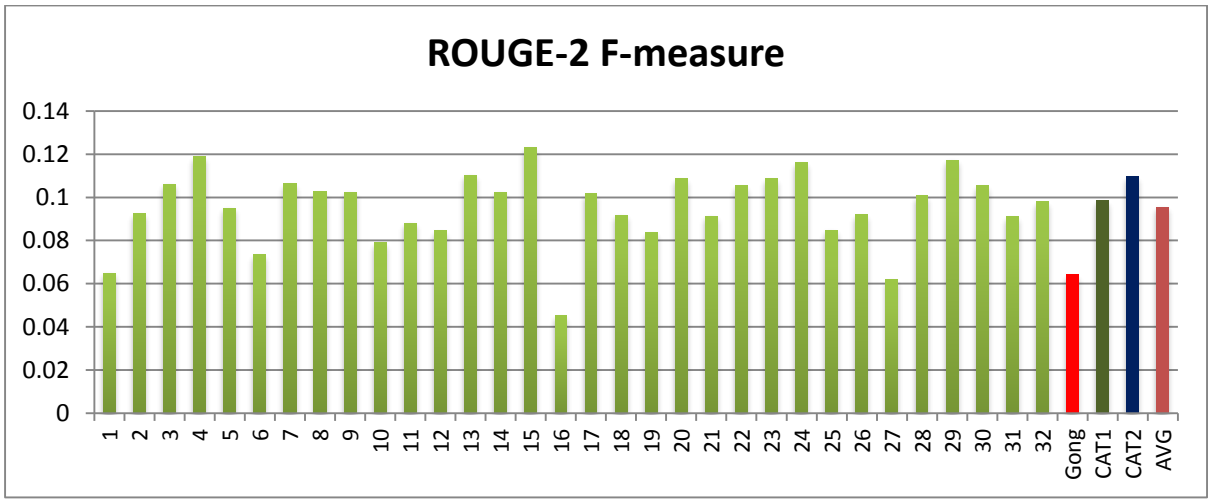
ابزار ROUGE میزان شباهت خلاصه های سیستمی را با ۴ تا از خلاصه های انسانی از پیش تهیه شده برای هر موضوع با استفاده از معیار های مختلف محاسبه می نماید. نتایج این ارزیابی ها در نمودارهای زیر آورده شده است.



شکل ۱۵- نمودار فراخوانی ROUGE-2 برای DUC2007



شکل ۱۶ - نمودار دقت ROUGE-2 برای DUC2007



شکل ۱۷- نمودار فاکتور F معیار ROUGE-2 برای DUC2007

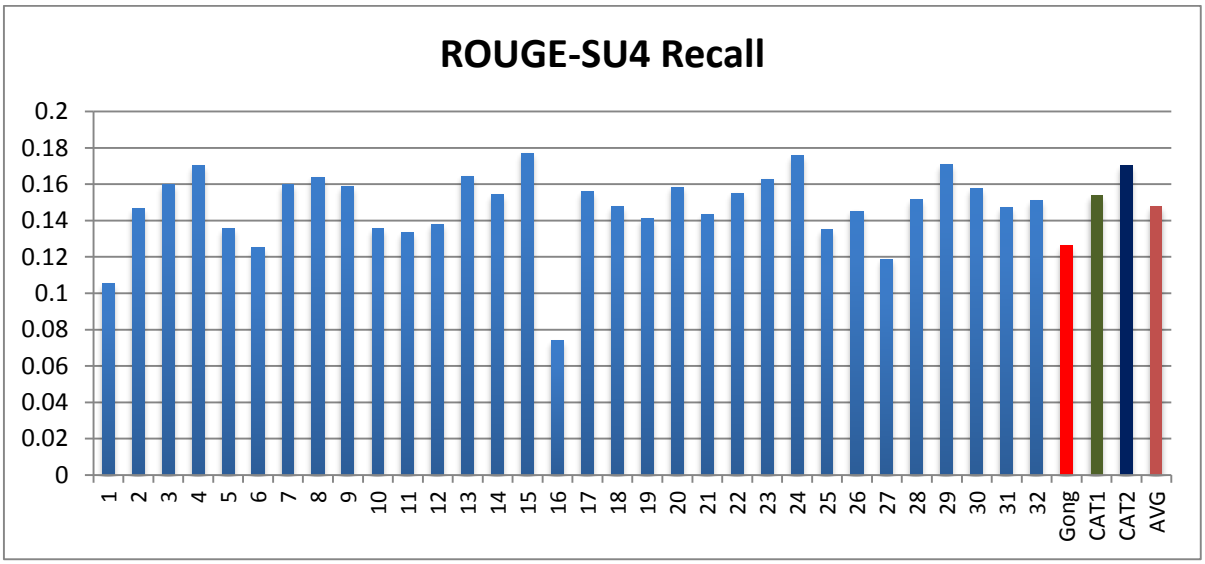
همانطور که پیش تر هم اشاره شد ه ارزیابی ROUGE-2 توسط رابطه زیر محاسبه می شود :

$$ROUGE-2 = \frac{\sum_{S\epsilon\{Reference\ Summaries\}} \sum_{gram_2} Count_{match}(gram_2)}{\sum_{S\epsilon\{Reference\ Summaries\}} \sum_{gram_2} Count(gram_2)}$$

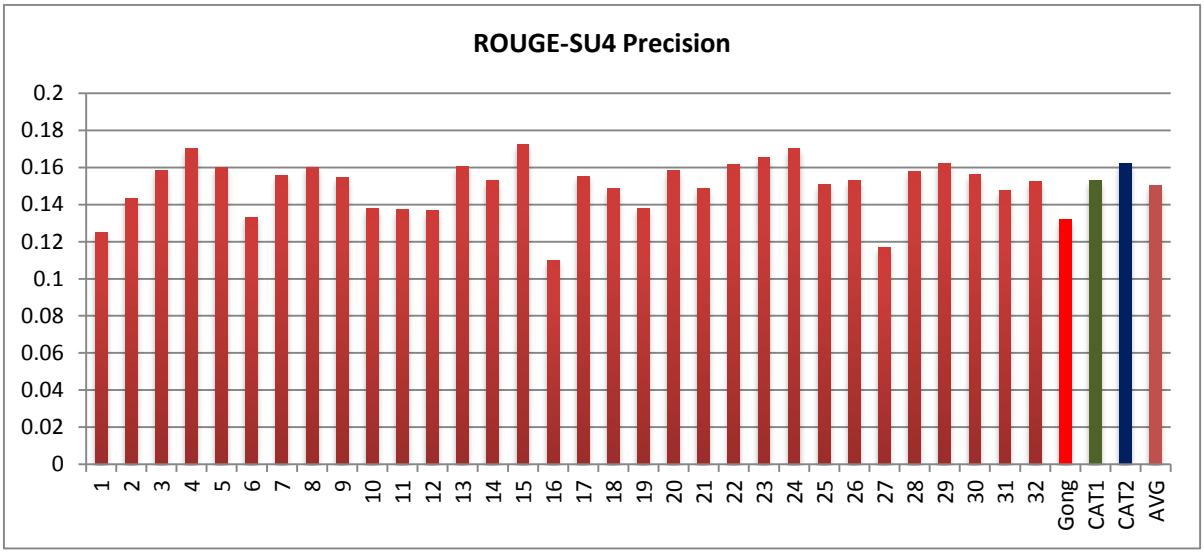
این رابطه میزان مشابهت بین خلاصه های سیستمی و خلاصه هـای مرجـع را براسـاس میـزان ۲ تـایی هـای مشترک بین آن دو محاسبه می نماید.

همانطور که از نتایج مشخص است خروجی CAT1 نسبت به سیستم ارائه شده توسط Gong از دقـت بسـیار بالاتری برخوردار است که دلیل آنهم تاثیر تمامی اسناد با دید سراسری در تولیـد خلاصـه مـی باشـد. CAT1 هنوز نسبت به میانگین سیستم های DUC2007 از دقت بالایی برخوردار نیست که دلیـل آن وجـود جمـلات شبیه به هم می باشد. با برطرف شدن این مشکل در CAT2 شاهد افزایش دقت مناسب نسـبت بـه میـانگین سیستم های DUC هستیم. در مجموع با توجه به سایر مقالاتی که تاکنون در زمینه خلاصه سازی ارائه شـده اند، خروجی سیستم مطلوب ارزیابی می شود.

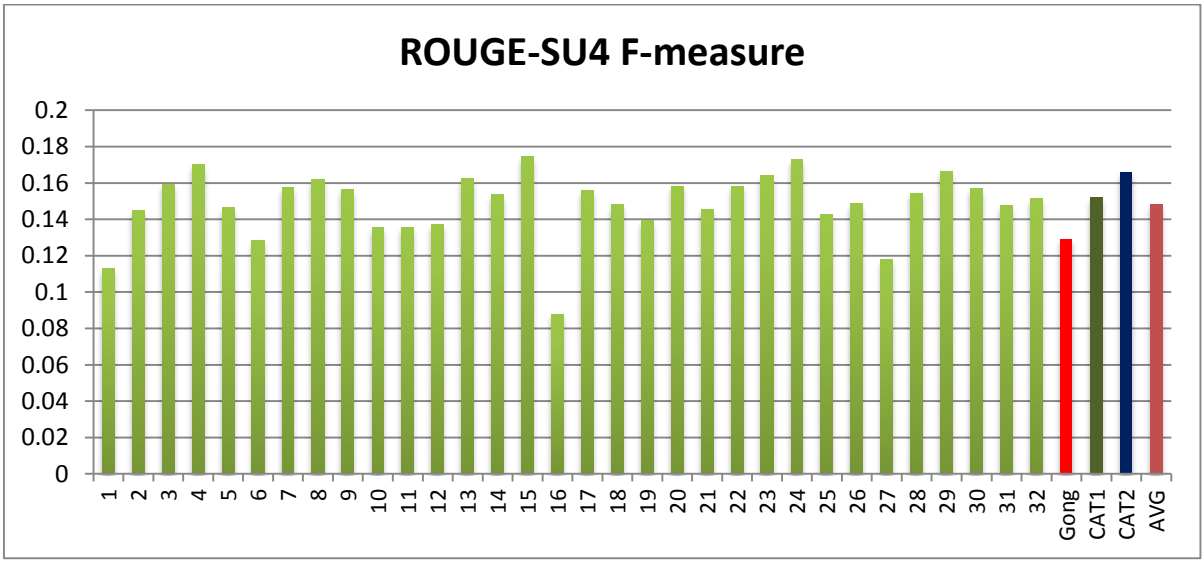
در نمودارهای زیر نتایج بررسی های ارزیابی با استفاده از معیار ROUGE-SU4 نمایش داده شده است.



شکل ۱۸- نمودار فراخوانی ROUGE-SU4 برای DUC2007



شکل ۱۹- نمودار دقت ROUGE-SU4 برای DUC2007

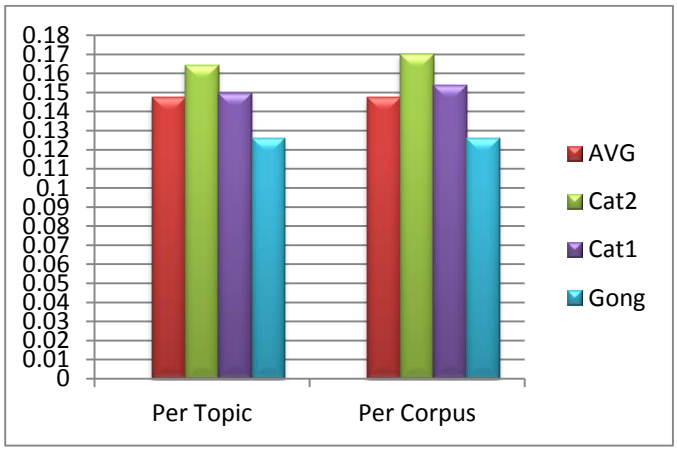


شکل ۲۰- نمودار فاکتور F معیار ROUGE-SU4 برای DUC2007

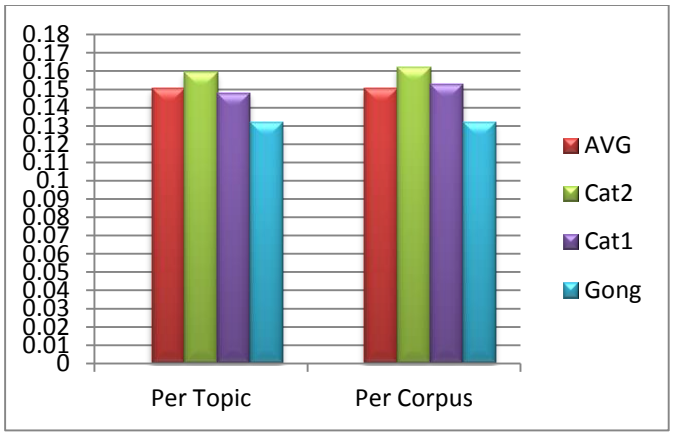
معیار ROUGE-SU4 آخرین معیاری است که توسط بسته ROUGE ارائه شده و در کنفرانس DUC هم مورد استفاده قرار گرفته است. این معیار بهترین نتایج ارزیابی را در بین معیارهای ROUGE برروی مجموعه داده های DUC2007 کسب کرده است. همانطور که در نمودار های فراخوانی، دقت و معیار –F هم نشان داده شده است سیستم ارائه شده نتایج بهتری نسبت به روش Gong و همچنین میانگین سیستم های DUC کسب کرده است.

مدل عمومی : در مدل عمومی پیاده سازی همانطور که پیش تر هم توضیح داده شد، ماتریس کلمه- سند

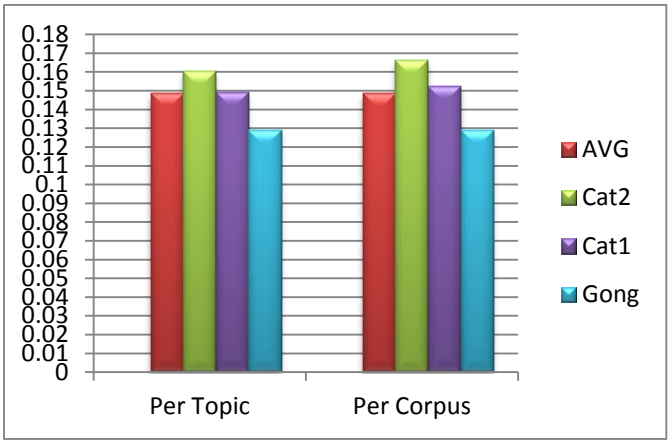
برروی هر موضوع (برخلاف روش مبتنی بر پیکره که برروی کل پیکره ساخته می شد) جداگانه ساخته شده و سایر مراحل دنبال می شود. در این مدل هم ارزیابی با معیارهای ROUGE-2 و ROUGE-SU4 انجام شده است که نتایج آن در نمودار های زیر آورده شده است. چون در این ارزیابی ها تفاوتی در نتایج سیستم های DUC ایجاد نشده بنابراین صرفا نتایج را با روش مبتنی بر پیکره و با میانگین سیستم های DUC مقایسه کرده ایم:



شکل ۲۱- نمودار فراخوانی ROUGE-SU4 برای DUC2007 (مدل عمومی)



شکل ۲۲- نمودار دقت ROUGE-SU4 برای DUC2007 (مدل عمومی)



شکل ۲۳- نمودار معیار-F ROUGE-SU4 برای DUC2007 (مدل عمومی)

با بررسی این نمودار ها مشخص می گردد که در حالت عمومی دقت ها کمی پایین تر از حالت مبتنی بر پیکره می باشد که البته این موضوع با خصوصیات مربوط به LSA هم مطابقت دارد چراکه معمولا در مقالات از LSA به عنوان روش مبتنی بر پیکره یاد می کنند[Dee 90].

با توجه به اینکه نتایج معیار ROUGE-2 هم شبیه به ROUGE-SU4 می باشد از نمایش نمودار مربوط به آن صرف نظر شده است.

۲-۴- نتایج آزمایش برروی داده های DUC2006

پارامتر های مربوط به ابزار ROUGE مطابق با نتایج اولیه کنفرانس DUC برروی داده های DUC2006 به صورت زیر تنظیم شده است :

ROUGE run-time parameters:

ROUGE-1.5.5.pl -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d rougejk.in

ابزار ROUGE میزان شباهت خلاصه های سیستمی را با ۴ تا از خلاصه های انسانی از پیش تهیه شده برای هر موضوع با استفاده از معیار های مختلف محاسبه می نماید. در مجموعه DUC2006، ۳۵ سیستم حضور دارند. نتایج مقایسه با میانگین این سیستم ها و همچنین بهترین و بدترین سیستم شرکت کننده در این کنفرانس در جداول زیر نشان داده شده است.

systems	ROUGE-2	ROUGE-SU
Gong	0.06524	0.10912

Cat1	0.07842	0.13221
Cat2	0.09207	0.15101
Avg	0.073609	0.1288
Best	0.09505	0.15464
Worst	0.02834	0.06394

جدول ۹- میزان فراخوانی برای DUC2006

systems	ROUGE-2	ROUGE-SU
Gong	0.07202	0.10932
Cat1	0.07719	0.13323
Cat2	0.09241	0.15012
Avg	0.076313	0.133371
Best	0.09524	0.15497
Worst	0.0429	0.09828

جدول ۱۰- نمودار دقت برای DUC2006

systems	ROUGE-2	ROUGE-SU
Gong	0.06503	0.11617
Cat1	0.07689	0.13218
Cat2	0.09203	0.15168
Avg	0.074779	0.130979
Best	0.09513	0.15478
Worst	0.03394	0.07704

جدول ۱۱- نمودار فاکتور F برای DUC2006

در هر سه نمودار سیستم پیشنهادی بهتر از میانگین سیستم های DUC و همچنین روش Gong می باشد.

۳-۴- نمایش تعدادی از جملات خلاصه شده توسط سیستم پیشنهادی

در شکل بخشی از جملات ابتدایی و انتهایی خلاصه تولید شده برای موضوع ۷۰۸ (" world-wide chronic potable water shortages") نشان داده شده است. همانطور که در تصویر هم مشخص است

جملات ابتدایی، جملات مهم و مرتبط با موضوع بوده و جملات انتهایی کاملا بی اهمیت و حتی بی ارتبـاط بـا موضوع می باشند.

The Bangladeshi capital of Dhaka is going alarmingly dry because of a serious shortage of water supply with no sign of immediate improvement.

Although Nepal is rich in water resources, water shortage is pervasive in the country because of its inability to tap the resources and the lack of well-managed supply system.

In 1996, a long spell of dry weather resulted in a low water level of Ruvu River and shortage of supplies to Dar es Salaam.

The Addis Ababa Regional Water and Sewerage Authority announced that the shortage of potable water in the capital city of Ethiopia will be solved in the last quarter of this year.

Iranian President Akbar Hashemi Rafsanjani today picked the ground for Mamloo reservoir dam near Tehran, a project designed to alleviate the water shortage problem in the capital.

A project will be launched this year to solve the water shortage problem for Yan'an, an old revolutionary base in northwest China's Shaanxi Province.

.
.
.

The police said the arrested will be release after each paying a fine of 500 pesos, or 60 U.S. dollars.

A shortfall usually rises during the summer season when load sheddings are more frequent than in the winter season.

The process will be implemented at paper mills across the country in the near future.

This year, the theme of the international disaster relief day is "preventing calamities with information."

تصویر – خروجی خلاصه های تولید شده سیستم پیشنهادی برای موضوع 708

شکل ۲۴ - نمونه ای از خلاصه های تولید شده برای موضوع ۷۰۸

۴-۴- پیاده سازی و تست سیستم برروی زبان فارسی

بعد از تست موفقیت آمیز سیستم برروی زبان انگلیسی، مشابه اینکار با همکاری اعضای آزمایشگاه وب معنایی[۱] و همچنین همکاری یکی از دانشجویان دانشگاه امیرکبیر پیاده سازی شد. البته با توجه به مشکلاتی که در زبان فارسی وجود دارد و عدم وجود تمامی ابزارهای مورد نیاز، بعضی از مراحل روش پیشنهادی به صورت دستی پیاده سازی شد. در ادامه مختصرا به کارهای انجام شده در این پیاده سازی اشاره شده است:

۱-۴-۴- تولید پیکره

متاسفانه در زبان فارسی، پیکره استانداردی برای خلاصه سازی متن وجود ندارد. بهمین دلیل پیکره کوچکی به صورت دستی و مطابق با استانداردهای استفاده شده در پیکره DUC و با همکاری گروه زبان شناسی دانشگاه فردوسی مشهد تولید شد. این پیکره شامل ۶ موضوع می باشد که از خبرگزاری های معروف تابناک،[۲] شبکه خبر،[۳] ایسنا،[۴] ایرنا[۵] و پرس تی وی[۶] جمع آوری شده اند. مشخصات کلی این پیکره در جدول زیر آورده شده است:

تعداد موضوعات	6
تعداد اسناد موجود در هر موضوع	30
تعداد جملات	3804
تعداد کلمات بدون کلمات ایست	17562
میزان خلاصه سازی	۴۵۰ کلمه

جدول ۱۲- مشخصات کلی پیکره تولید شده بروی زبان فارسی

پس از جمع آوری این اخبار، برای هر موضوع چهار خلاصه انسانی توسط چهار نفر تولید شده که در مرحله ارزیابی خلاصه سیستمی مورد استفاده قرار خواهند گرفت.

۱- wtlab.um.ac.ir

2 - www.tabnak.ir

3 - www.irirb.ir

4 - www.isna.ir

۵- www.irna.ir

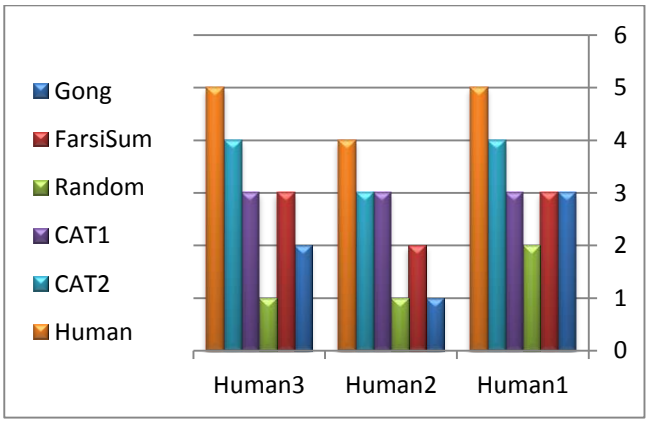
۶- www.presstv.ir

۲-۴-۴- پردازش های صورت گرفته

مراحل پیش پردازش شامل انتخاب جملات، کلمات، حذف کلمات ایست، ریشه یابی، تولید ماتریس و اعمال LSA و مرتب سازی جملات براساس شباهت با زمینه به صورت خودکار انجام شد. عملیات فاز سوم که شامل برچسب زنی و استفاده از شبکه واژگان می باشد به صورت دستی انجام شد.

۳-۴-۴- روش ارزیابی

- ارزیابی انسانی : متاسفانه برای ارزیابی متون در زبان فارسی هم ابزاری وجود ندارد و مقالات ارائه شده به صورت انسانی ارزیابی را انجام می دهند. برای ارزیابی انسانی در زبان فارسی از سه دانشجوی کارشناسی ارشد زبان شناسی و کامپیوتر خواسته شد تا به خلاصه های تولید شده توسط سیستم از ۱(خیلی بد) تا ۵(خیلی خوب) امتیاز دهند. نتایج این ارزیابی در تصویر زیر آورده شده است.



شکل ۲۵- نتایج ارزیابی انسانی خلاصه های فارسی تولید شده

- ارزیابی خودکار: خلاصه های تولید شده، توسط ابزار فارسی معادل ROUGE که توسط اعضای آزمایشگاه فن آوری وب دانشگاه فردوسی پیاده سازی شده است ارزیابی گشت. نتایج ارزیابی معیارهای ROUGE-2 و ROUGE-1 مربوط به این ارزیابی ها را در تصاویر زیر مشاهده می کنید:

موضوع : بحران نفتی خلیج مکزیک				
Docs	Sentences	Tokens	Stop Words	Top Words
30	514	2467	236	نفت، مکزیک، خلیج، شرکت، نشت، بریتیش

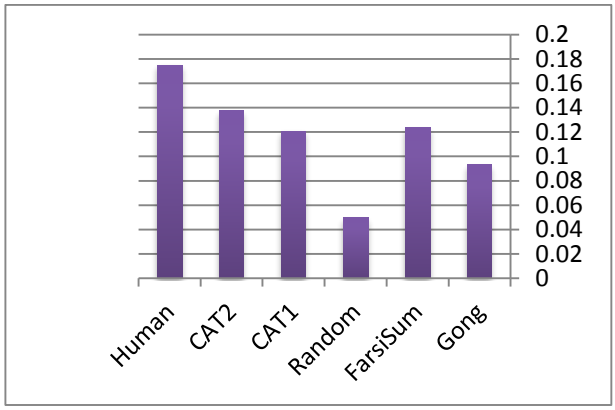
Summary_Size	Rouge1	Rouge2
500 کلمه	0.571	0.173
300 کلمه	0.413	0.149

موضوع : بازیهای آسیایی گوانگجو				
Docs	Sentences	Tokens	Stop Words	Top Words
30	602	2693	233	مدال، آسیایی، بازی، ایران ، تیم، گوانگجو
Summary_Size		Rouge1		Rouge2
500 کلمه		0.539		0.166
300 کلمه		0.40		0.134
موضوع : دستگیری عبدالمالک ریگی				
Docs	Sentences	Tokens	Stop Words	Top Words
31	610	3417	269	ریگی، دستگیری، ایران، عبدالمالک، کشور
Summary_Size		Rouge1		Rouge2
500 کلمه		0.501		0.153
300 کلمه		0.379		0.124
موضوع : تحولات مصر				
Docs	Sentences	Tokens	Stop Words	Top Words
29	663	3465	263	مصر، مبارک، انقلاب، کشور، اسلامی
Summary_Size		Rouge1		Rouge2
500 کلمه		0.594		0.164
300 کلمه		0.399		0.133
موضوع : زلزله هاییتی				
Docs	Sentences	Tokens	Stop Words	Top Words
29	690	3425	225	زلزله، هاییتی، تلفات، تخریب، امداد،
Summary_Size		Rouge1		Rouge2
500 کلمه		0.542		0.165
300 کلمه		0.417		0.141
موضوع : نقش آمریکا در تحولات خاورمیانه				
Docs	Sentences	Tokens	Stop Words	Top Words
29	705	3592	271	آمریکا، خاورمیانه، انقلاب، اسلام، مصر
Summary_Size		Rouge1		Rouge2

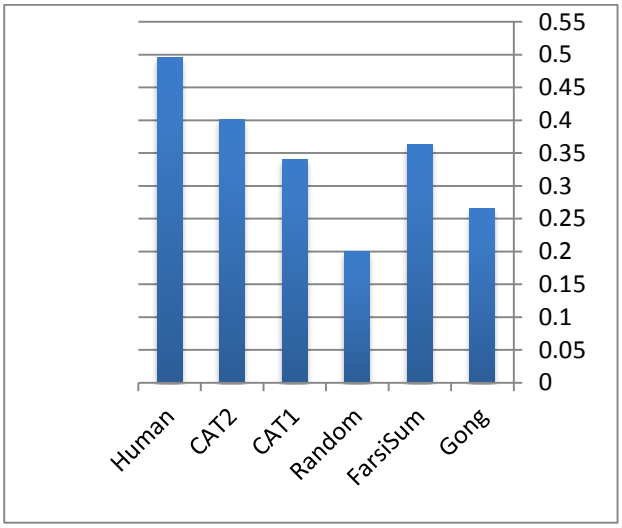
500 کلمه	0.581	0.17
300 کلمه	0.403	0.145

جدول ۱۳- مشخصات موضوعات پیکره

نتایج مقایسه فراخوانی روش پیشنهادی برای زبان فارسی با روش های دیگر در نمودار زیر نشان داده شده است:



شکل ۲۶- نمودار فرخوانی ROUGE-2 با فشرده سازی ۳۰۰ کلمه ای



شکل ۲۷- نمودار فرخوانی ROUGE-1 با فشرده سازی ۳۰۰ کلمه ای

نتایج ارزیابی سیستم پیشنهادی با سیستم های FarsiSum و روش Gong با معیارهای ROUGE-2 و ROUGE-1 نشان گر دقت مناسب سیستم پیشنهادی برای متون چند سندی فارسی می باشد.

۵-۴- خلاصه

در این بخش نتایج بدست آمده از روش پیشنهادی در فصل قبل ارائه شد. در ابتدا خلاصه سازی برروی دادهای استاندارد DUC2007 پیاده سازی شد. خلاصه های تولید شده با استفاده از معیارهای ارزیابی ROUGE-2 و ROUGE-SU4 ارزیابی شد. نتایج ارزیابی ها نشان داد که سیستم پیشنهادی بهتر از بسیاری از سیستم های DUC 2007 بوده و همچنین از روش پیاده سازی شده توسط Gong هم بهتر می باشد. این پیاده‌سازی‌ها در دو مدل مبتنی بر پیکره و مدل عمومی انجام شد. در مدل مبتنی بر پیکره نتایج بهتری حاصل شد که این موضوع با ویژگی LSA که روش مبتنی بر پیکره می باشد تطابق دارد. همچنین روش پیشنهادی برروی داده های DUC2006 هم پیاده سازی شد که نتایج این ارزیابی ها هم مناسب بود.

در ادامه روش پیشنهادی برروی زبان فارسی بررسی شد. برای تست برروی زبان فارسی در ابتدا یک پیکره کوچک با ۶ موضوع تولید شد و سپس برای هر موضوع خلاصه های انسانی تولید گشت. سیستم پیشنهادی برروی این پیکره پیاده سازی شد و به دو روش ارزیابی شد. در روش اول از تعدادی انسان خواسته شد که ارزیابی را به صورت دستی انجام دهند. در روش دوم معادل ابزار ROUGE برای زبان فارسی پیاده سازی و با استفاده از معیارهای ROUGE-1 و ROUGE-2 ارزیابی انجام شد. نتایج این ارزیابی ها با سیستم خلاصه سازی فارسی FarsiSum و همچنین روش Gong (که برای زبان فارسی پیاده‌سازی شد) مقایسه شد که نتایج حاکی از دقت مناسب روش پیشنهادی بود.

فصل پنجم:

نتیجه گیری و

کارهای آتی

۵– نتیجه‌گیری و کارهای آتی

۵–۱– نتیجه‌گیری

کامل شود.

۵–۲– کارهای آتی

کامل شود (بررسی سایر روش ها برای استخراج زمینه)

منابع

۶- منابع

[All 86] Allison, L. and Dix, T. 1986. A bit-string longest-common-subsequence algorithm.Inf. Proc. Lett. 23, 305-310.

[Aks 09]Gem Aksoy, Ahmet Bugdayci, Tunay Gur, ibra Uysal, Faz Can, Semantic argument frequency-based multi-document summarization.,ISCIS, 2009@IEEE.

[Aon 97] Aone, C., Okurowski, M. E., Gorlinsky, J., & Larsen, B. (1997).A scalable summarization system using robust NLP.In Proceedings ofthe ACL'97/EACL'97 workshop on intelligent scalable text summarization (pp. 10–17), Madrid, Spain.

[Kia 06] B A Kiani, T M R Akbarzadeh, Automatic text summarization using: Hybrid Fuzzy GA-GP (2006).Proceedings of the IEEE International Conference on Fuzzy Systems, July 16-21

[Azz 99]Azzam, S., Humphreys, K., & Gaizauskas, R. (1999). Using coreference chains for text summarization. In Proceedings of the ACL'99workshop on coreference andits applications (pp. 77–84), College Park, MD, USA.

[Bar 97] Regina Barzilay and MichaelElhadad.Using Lexical Chains for Text Summarization.In Proceedings of the ACL Workshop on Intelligent ScalableText Summarization, pages 10–17, Madrid, Spain,August 1997. Association for Computational Linguistics.

[Bar 99] Barzilay R., McKeown K. R., and Elhadad M. (1999), Information fusion in the context of multi-document summarization, in Proceedings of the 37th Association for Computational Linguistics, 1999, Maryland

[Bos05] Wauter Bosma (2005). Query-Based Summarization using Rhetorical StructureTheory. In: Ton van der Wouden, Michaela Poss, H. Reckman and C. Cremers, ed.,15th Meeting of CLIN, 2005

[Cha 00] Chuang and Yang: "Extracting Sentence Segments for Text Summarization: A Machine Learning Approach", in: Proceedings of the ACM SIGIR 2000.

[Con 01] Conroy, J.M. and D.P. O'leary, 2001.Textsummarization via hidden markov models.Proceedings of the 24th Annual International ACMSIGIR Conference on Research and Developmentin Information Retrieval, Sept. 9-12, New Orleans,Louisiana, United States, pp: 406-407.

[Con 05] John M. Conroy, Judith D. Schlesinger, Jade Goldstein Stewart (2005). CLASSYQuery-Based Multi-Document Summarization.In DUC 05 Conference Proceedings,Boston, USA.

[Dau 00] H.Daume, A.Echihabi, D.Marcu, D.Stefan, R.Soricut.GLEANS : A Generator of Logical Extracts and Abstracts forNice Summaries. 2000

[Dee 90] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman.Indexing by latentsemantic analysis. Journal of the American Society for Information Science and Technology(JASIS),41(6):391-470, 1990.

[Dej 79] DeJong, G. F. (1979). Skimming stories in real time: an experiment in integrated understanding.Doctoral Dissertation.ComputerScience Department, Yale University.

[Dup 07] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using clickthrough data and a user model. In WWW '07: Proceedings of International Conference on World Wide Web, Banff, Canada, 2007

[Eri 99] E r i c, C., T. G. K o l d a. New Term Weighting Formulas for the Vector Space Method in Information Retrieval. − Technical Report, ORNL/TM-13756, Oak Ridge National Laboratory, 1999.

[Fox 05] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve websearch. ACM Transactions on Information Systems,23(2):147–168, 2005.

[Gao 09] J.Gao, W.Yuan, X.Li, K.Deng, J.Nie.Smoothing Clickthrough Data for Web Search Ranking.SIGIR, 2009, Boston, Massachusetts, USA, ACM.

[Gol 99] Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999).Summarizing text documents: sentence selection and evaluationmetrics.In Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval(SIGIR'99) (pp. 121–128), Berkeley, CA, USA.

[Gon 01] Gong, Y., & Liu, X. (2001).Generic text summarization using relevance measure and latent semantic analysis.In Proceedings of the24th annual international ACM SIGIR conference on research andd evelopment in information retrieval (SIGIR'01) (pp. 19–25), NewOrleans, LA, USA.

[Gra 81] Graesser, A. C. (1981). Prose Comprehension beyondthe Word. NY: Springer-Verlag.

[Hal 07] W. S. A. Halabi, M. Kubat, and M. Tapia. Time spent on a web page is sufficient to infer a user's interest. In IMSA '07: Proceedings of IASTED European Conference, pages 41–46, Anaheim, CA, USA, 2007. ACTA Press.

[Hir 01] T. Hirao, Y. Sasaki, and H. Isozaki. An extrinsic evaluation for question-biased text summarization on qa tasks. In Proceedings of NAACL workshop on Automatic summarization, 2001.

[Hir 03] T.Hirao, K.Takeuchi, H.Isozaki, Y.Sasaki, E.Maeda. NTT/NAIST's Text Summarization Systems.2003.

[Hov 97] Hovy, E., & Lin, C. Y. (1997).Automatic text summarization in SUMMARIST.In Proceedings of the ACL'97/EACL'97 workshop onintelligent scalable text summarization (pp. 18–24), Madrid, Spain.

[Lin 02] C.Lin, E.Hovy. Automated Multi-document Summarization in NeATS. 2002

[Luh 58] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research andDevelopment, 2(2), 159–165.

[Lv 06] Y. Lv, L. Sun, J. Zhang, J.-Y.Nie, W. Chen, and W. Zhang.An iterative implicit feedback approach to personalized search. In ACL '06: Proceedings of International Conference on Computational Linguistics, pages 585–592, Morristown, NJ, USA, 2006. Association for Computational Linguistics

[Jag 07] Jagadeesh J, Prasad Pingali, Vasudeva Varma (2007). Capturing Sentence Prior forQuery-Based Multi-Document Summarization. In RIAO, http://dblp.uni-trier.de.

[Jia 97] Jiang, J. AND Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy.In Proceedings of the International Conference on Research in Computational Linguistics

[Joa 02] T. Joachims. Optimizing search engines using clickthrough data. In KDD '02: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 133–142, New York, NY, USA, 2002. ACM

[Kel 01] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 408–409, New York, NY, USA, 2001. ACM.

[Kel 04] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 377–384, New York, NY, USA, 2004. ACM.

[Kia 04] Khosrow Kaikhah, "Automatic Text Summarization withNeural Networks", in Proceedings of second internationalConference on intelligent systems, IEEE, 40-44, Texas,USA, June 2004.

[Kni 02] Knight K. and Marcu D. (2002), Summarization beyond sentence extraction: A probabilistic approach to sentence compression, Artificial Intelligence, 139(1), 2002.

[Kon 05] Kondark, G. 2005. N-gram similarity and distance.In Proceedings of the 12h International Conferenceon String Processing and Information Retrieval (Buenos Aires, Argentina).115–126.

[Koy 09] Koychev I., Nikolov, R. and Dicheva D.: SmartBook: The New Generation e-Book, Proc. of BooksOnline'09 Workshop, in conjunction with ECDL 2009, Corfu,October 2, 2009.

[Kup 95] Kupiec, J., Pedersen, J., Chen, & F. (1995).A trainable document summarizer. In Proceedings of the 18th annual international ACMSIGIR conference on research andd evelopment in information retrieval (SIGIR'95) (pp. 68–73), Seattle, WA, USA.

[Kyo 08] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami, Pooya Khosravyan Dehkordy, Asghar Tajoddin. Optimizing Text Summarization Based on Fuzzy Logic. Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008).

[Lac 06] F.Lacatusu, A.Hickl, K.Roberts, Y.Shi, J.Bensley,B.Rink, P.Wang, L.Taylor. LCC's GISTexter at DUC

2006: Multi-Strategy Multi-Document Summarization.2006.

[Lin 03] C. –Y. Lin and Hovy.Automatic evaluation of summaries using n-gram co-occurrence statistics.In Proccedings of NLT-NAACL, 2003.

[Liu 02] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In CIKM '02: Proceedings of the 11th ACM International Conference on Information and Knowledge Management, pages 558–565, New York, NY, USA, 2002. ACM.

[Man 99] Mani, I.& Maybury, M. T. (Eds.). (1999). Advances in automated text summarization. Cambridge, MA: The MIT Press.

[Mar 97] D.Marcu. From Discourse Structures to Text Summaries.Proceedings of the 14th National Conference on Artificial Intelligence AAAI-97.

[Mck 95] McKeown, K., & Radev, D. R. (1995).Generating summaries of multiple news articles. In Proceedings of the 18th annual internationalACM SIGIR conference on research andd evelopment in information retrieval (SIGIR'95) (pp. 74–82), Seattle, WA, USA.

[Mih 06] Mihalcea, R., Corley, C., and Strapparava, C. 2006. Corpus-based and knowledge-based measuresof text semantic similarity.In Proceedings of the American Association for Artificial Intelligence.(Boston, MA).

[Moh 06] Ahmed A. Mohamed, Sanguthevar Rajasekaran (2006). Query-Based SummarizationBased on Document Graphs. In Proceedings of IEEE International Symposiumon Signal Processing and Information Technology, pp.408-410, Vancouver,Canada, 2006.

[Nie 03] - Nielsen netratings search engine ratings report. http://searchenginewatch.com/reports/article.php/2156461,2003.

[Pap 00] C. H. Papadimitrious, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing:A probabilistic analysis. J. Comput. Syst. Sci., 61(2):217-235, 2000.

[Par 01] T.A.S.Pardo, L.H.M.Rino, M.d.G.V.Nunes.GistSumm: A Summarization Tool Based on a New Extractive Method.2001.

[Por 80] Porter, M. "An algorithm for suffix stripping. Program".14(3), pp.130-137, 1980.

[Qaz 08]Vahed Qazvinian, Leila Sharif Hassanabadi, Ramin Halavati, Summarising text with a genetic algorithm-basedsentence extraction.Int. J. Knowledge Management Studies, Vol. 2, No. 4, 2008.

[Rad 05] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In KDD '05: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pages 239–248, New York, NY, USA, 2005. ACM.

[Rat 61] GJ Rath, A Resnick, TR Savage. The formation of abstracts by the selection of sentences. American Documentation, 12(2): 139143, April 1961.

[Sag 02] H.Saggion, and G.Lapalme."Generating Informative and Indicative Summaries with SumUM", Computational Linguistics, Special Issue on Automatic Summarization. 2002.

[Sal 97] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997).Automatic text structuring and summarization. Information Processing &Management, 33(2), 193–207.

[Sch 77] Schank, R., & Abelson, R. (1977).Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Lawrence Erlbaum Associates.

[Sch 08] Frank Schilder, Ravikumar Kondadadi (2008). FastSum: Fast and accurate querybasedmulti-document summarization. In Proceedings of the 46th meeting of theAssociation for Computational Linguistics, Columbus, Ohio.

[Sha 09] Mehrnoush Shamsfard, Tara Akhavan and Mona Erfani Joorabchi, Persian Document Summarization by Parsumist, World Applied Sciences Journal 7 (Special Issue of Computer & IT): 199-205, 2009.

[Sil 04] Silla, J., Nascimento, C., Pappa, G.L., Freitas, A.A. and Kaestner, C.A.A. 'Automatictext summarization with genetic algorithm-based attribute selection', Lecture Notes in ArtificialIntelligence, 2004.

[Spe 05] M. Speretta and S. Gauch. Personalized search based on user search histories. In WI '05: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, pages 622–628, Washington, DC, USA,2005.

IEEE Computer Society.

[Ste 05] Steinberger, J., & Kabadjov, M.A. & Poesio, M., & Sanchez-Graillet,O . Improving LSA-basedsummarization with anaphora resolution.In Proceedings of the conference on Human Language Technologyand Empirical Methods in Natural Language Processing. 2005.

[Sua 09] Ladda Suanmali, Mohammed Salem, Binwahlan andNaomie Salim, "Sentence Features Fusion for Textsummarization using Fuzzy Logic, IEEE, 142-145, 2009.

[Svo 07] Svore, K., Vanderwende, L., and Burges, C. (2007). Enhancing single-documentsummarization by combining RankNet and third-party sources. In Proceedings ofthe EMNLP-CoNLL, pages 448{457.

[Swe 00] H. Dalianis, "SweSum A Text Summarizer for Swedish",Technical report", TRITANA-P0015, IPLab-174, 2000.

[Teu 97] Teufel, S. H., & Moens, M. (1997). Sentence extraction as a classification task.In Proceedings of the ACL'97/EACL'97 workshop onintelligent scalable text summarization (pp. 58–65), Madrid, Spain.

[Wan 07] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi document summarization. In Proceedings of IJCAI, 2007.

[Whi 01] M.White, T.Korelsky, C.Cardie, V.Ng , D.Pierce, K.Wagstaff. Multidocument Summarization via Information Extraction.2001.

[Win 02] Winkel, A. and D. Radev. MEADeval: Anevaluation framework for extractive summarization.2002, http://perun.si.umich.edu/clair/meadeval/.

[Xu 09] S.Xu, H.Jiang, F.C.M. Lau.User-Oriented Document Summarization through Vision-Based Eye-Tracking., Sanibel Island, Florida, 2009, USA.

[Yeh 02] Yeh, J. Y., Ke, H. R., & Yang, W. P. (2002).Chinese text summarization using a trainable summarizer and latent semantic analysis.InLecture Notes in Computer Science (Vol. 2555).In Proceedings of the 5th international conference on Asian digital libraries(ICADL'02) (pp. 76–87), Singapore. Berlin: Springer-Verlag.

[Yeh 05] Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. Text summarization using a trainable summarizer and latent semantic analysis. Information Processing and Management, 41, 75-95, 2005.

[Yu 09] Yu, H. News summarization based on semantic similarity measure. Ninth International Conference on Hybrid Intelligent Systems, vol. 1, pp.180-183, 2009.

[You 85] Young, S. R., & Hayes, P. J. (1985).Automatic classification and summarization of banking telexes.In Proceedings of the 2ndConference on Artificial Intelligence Applications (pp. 402–408).

[Zam 08] Azadeh Zamanifar, Behrouz minaei-Bidgoli and Mohsen Sharifi," A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of Text ", In Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, 635-639, Iran, 2008.

[Zec 00]K.Zechner, A.Waible.DiaSumm : Flexible Summarization of Spontaneous Dialogues in UnrestrictedDomains. 2000.

[کری ۸۵]- زهره کریمی ، مهرنوش شمس فرد ، ۱۳۸۵ ، " سیستم خلاصه سازی خودکار متون فارسی" ، دوازدهمین کنفرانس بین المللی انجمن کامپیوتر ، تهران، ایران

ضمایم

چند نمونه از جملات خلاصه انتخاب شده انگلیسی و فارسی آورده شود :

شبه کد :

Abstract:

To DO

Keywords:To Do

Department of Computer Engineering

Ferdowsi University of Mashhad

Master's Thesis

Concept-based Automatic Multi-document Summarization

By:

AsefPourmasoumi

Supervisor:

Dr. Mohsen Kahani

Advisor:

Dr. NaderJahangiri

September 2011