دانشگاه فردوسی مشهد
دانشکده مهندسی - گروه مهندسی کامپیوتر
آزمایشگاه فناوری وب

پایان‌نامه کارشناسی ارشد

خلاصه‌سازی چکیده‌ای مبتنی بر مشابهت جملات

نگارش

فاطمه پورغلامعلی

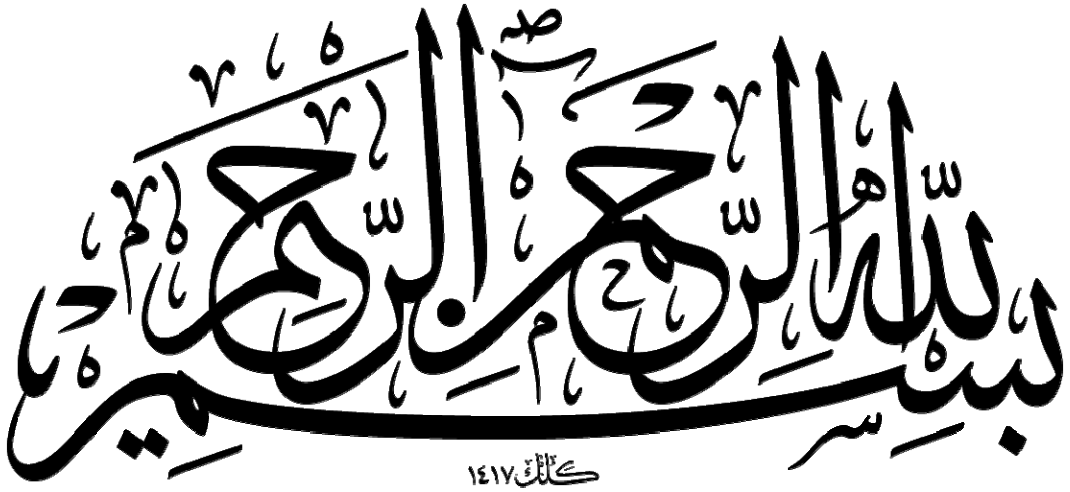
استاد راهنما

دکتر محسن کاهانی

استاد مشاور

دکتر نادر جهانگیری

بهمن ۹۰


بسم الله الرحمن الرحيم
كلك ١٤١٧

تقدیم

به گلهای زندگیم

پدر دلسوز و فداکارم که عشق ورزیدن را به من آموخت

مادر صبور و عطوفم که نعمت صمیمیت را به من هدیه داد

و همسر فداکار و مهربانم که همواره، صبورانه، دلسوزانه و عاشقانه یار یکرم بوده و هست

سپاس

پروردگار مهربانم را سپاس گذارم به خاطر نعمت استادی گرانقدر و معلمی عالم، جناب آقای دکتر محسن کاهانی، که در نگاهش جز شوق علم، در کلامش جز دلسوزی و محبت و در کردارش جز انسانیت و شرافت ندیدم. پندهایش همواره زمزمه زبانم و رفتارش همواره الگوی زندگی علمی‌ام خواهد بود.

از خداوند بزرگ تمنا دارم تا مجالی دهد، گوشه‌ای از زحمات ایشان را جبران نمایم.

همچنین از استاد گرانقدر جناب آقای دکتر نادر جهانگیری که بارها همیاری‌هایشان همچون پدری دلسوز یاریگرم بودند، صمیمانه سپاسگذارم.

علاوه بر این از تمامی دوستان و همکاران آزمایشگاه فناوری وب و دوستان گروه زبانشناسی مخصوصا خانم طباطبایی تشکر می‌نمایم.

چکیده

خلاصه‌سازی خودکار متون همزمان با رشد روز افزون اسناد و اطلاعات بیش از پیش مورد توجه علاقه‌مندان حوزه پردازش زبان طبیعی قرار گرفته است. از این میان خلاصه‌سازی چند سنده که در آن چندین سند به عنوان ورودی دریافت می‌گردد، مورد توجه زیادی قرار گرفته است. در بسیاری از روش‌ها تنها گزیده‌ای از جملات اولیه بدون تغییر به عنوان خلاصه برگردانده می‌شود. که به خلاصه-سازی گزینشی معروف است. در مقابل آن در زمینه خلاصه‌سازی چکیده‌ای -که در آن برگرفته‌ای از جملات اولیه را خواهیم داشت- کار چندانی انجام نگرفته‌است. در این پایان‌نامه به بیان روشی برای خلاصه‌سازی چکیده‌ای و چند سنده خواهیم پرداخت که بر مبنای نقش‌های معنایی، شباهت معنایی جملات را محاسبه نموده و بر مبنای آن جملات را فشرده‌سازی می‌نماید. پس از آن با الگوریتمی، به حذف و ادغام جملات می‌پردازد. نتایج ارزیابی نشان‌دهنده بهبود روش‌های پیشنهادی شباهت معنایی و فشرده‌سازی جملات نسبت به روش‌های مرتبط پیشین می‌باشند. همچنین ارزیابی سیستم پیشنهادی برروی داده‌های کنفرانس DUC و با استفاده از معیار ارزیابی ROUGE نشانگر بهبود نتایج نسبت به سایر روش‌های خلاصه‌سازی موجود می‌باشد.

کلمات کلیدی:

خلاصه‌سازی متن، خلاصه‌سازی چندسنده، خلاصه‌سازی چکیده‌ای، فشرده‌سازی جملات، شباهت معنایی، ادغام جملات.

فهرست عناوین

صفحه

۱ مقدمه ۱

۱.۱ مروری بر خلاصه سازی ۱

۲.۱ روش پیشنهادی ۳

۲ مرور ادبیات ۲

۱.۲ تعاریف پایه در پردازش زبان طبیعی ۲

۱.۱.۲ ریشه یابی ۲

۲.۱.۲ برچسب زنی بخش گفتاری ۳

۳.۱.۲ برچسب گذاری نقش معنایی ۴

۴.۱.۲ حذف ایست واژه ها ۵

۵.۱.۲ درخت تجزیه ۶

۲.۲ خلاصه سازی متن ۷

۱.۲.۲ خلاصه سازی تک سنده ۷

۲.۲.۲ خلاصه سازی چند سنده ۷

۳.۲.۲ خلاصه سازی گزینشی ۸

۴.۲.۲ خلاصه سازی چکیده ای ۹

۱.۴.۲.۲ فشرده سازی ۱۰

۲.۴.۲.۲ آمیختن جملات ۱۵

۳.۲ ارتباط معنایی کلمات ۱۶

۱.۳.۲ شبکه واژگان ۱۶

۲.۳.۲ روش های مبتنی بر طول مسیر ۱۸

۳.۳.۲ مقیاس بندی شبکه ۱۹

۴.۳.۲ روش های مبتنی بر اطلاعات ۲۱

۵.۳.۲ ارزیابی روش های محاسبه شباهت معنایی ۲۴

۴.۲ شباهت جملات ۲۶

۵.۲ مجموعه داده های استاندارد ۲۹

۱.۵.۲ داده های استاندارد DUC ۲۹

۶.۲ روش های ارزیابی خلاصه سازی ۳۰

۱.۶.۲ معیار ROUGE ۳۱

۳ روش پیشنهادی ۳۵

۱.۳ معماری روش پیشنهادی ۳۵

۲.۳ پیش پردازش ۳۶

۱.۲.۳ حذف ایست واژه ها ۳۶
۲.۲.۳ ریشه یابی ۳۷
۳.۲.۳ درخت تجزیه، برچسبگذاری بخشهای گفتاری و برچسبگذاری نقشهای معنایی ۳۸
۳.۳ اعمال الگوریتم خلاصه سازی گزینشی ۴۱
۴.۳ محاسبه شباهتهای معنایی ۴۳
۵.۳ فشرده سازی جملات ۴۶
۱.۵.۳ حذف کلمات اضافه ۴۶
۲.۵.۳ فشرده سازی مبتنی بر مشابهت جملات ۴۹
۶.۳ حذف یا ادغام جملات ۵۱
۱.۶.۳ مروری گذرا بر خوشهبندی طیفی ۵۲
۲.۶.۳ الگوریتم حذف و ادغام ۵۳
۴ پیاده سازی و ارزیابی ۵۶
۱.۴ ارزیابی روش مشابهت معنایی جملات ۵۶
۲.۴ ارزیابی روش فشرده سازی جملات ۵۸
۳.۴ ارزیابی تعداد دسته های مختلف ۶۰
۴.۴ ارزیابی نهایی ۶۱
۵ نتیجه‌گیری ۶۴
۱.۵ کارهای آینده ۶۴
۶ منابع و مراجع ۶۹

فهرست اشکال

صفحه

شکل ۱.۲ نمونه‌هایی از بخش های گفتار و برچسب های آنها [CCG1]..... ۴
شکل ۲.۲ نمونه‌هایی از نقشهای معنایی و برچسب‌های آنها [CCG2]..... ۵
شکل ۳.۲ قسمتی از سلسله مراتب شبکه واژگان [RES95] ۲۲
شکل ۴.۲ ارزیابی انسانی بر شباهت کلمات..... ۲۵
شکل ۵.۲ ارزیابی معیار هیرست بر شباهت کلمات [BUD06]..... ۲۵
شکل ۶.۲ ارزیابی معیار لین بر شباهت کلمات [BUD06]..... ۲۶
شکل ۲.۷ ارزیابی معیار لیکوک بر شباهت کلمات [BUD06]..... ۲۶
شکل ۸.۲ ارزیابی معیار ژیانگ بر شباهت کلمات [BUD06]..... ۲۶
شکل ۹.۲ ارزیابی معیار رسنیک بر شباهت کلمات [BUD06]..... ۲۶
شکل ۱.۳ معماری کلی روش پیشنهادی..... ۳۵
شکل ۲.۳ خروجی برچسب زن بخش های گفتاری..... ۳۹
شکل ۳.۳ خروجی برچسب زن نقش های معنایی..... ۴۰
شکل ۴.۳ خروجی تجزیه گر ۴۱
شکل ۵.۳ خروجی SVD برای ماتریس کلمه-سند..... ۴۲
شکل ۶.۳ ساختار درخت تجزیه برای نمونه بدل ۱ ..... ۴۷
شکل ۷.۳ساختار درخت تجزیه برای نمونه بدل ۲..... ۴۸
شکل ۸.۳ ساختار درخت تجزیه برای یک نمونه عبارت مخفف..... ۴۹
شکل ۱.۴ مقایسه نتایج حاصل از روشهای پیشین و روش پیشنهادی شباهت جملات ۵۷
شکل ۲.۴ مقایسه معیار ROUGE-2 جملات فشرده شده در ۵ حالت مختلف..... ۵۸
شکل ۳.۴ مقایسه معیار ROUGE-SU4 جملات فشرده شده در ۵ حالت مختلف ۵۹
شکل ۴.۴ مقایسه معیار F بر فشرده سازی ۵۹
شکل ۵.۴ مقایسه نرخ فشرده سازی..... ۶۰
شکل ۶.۴ مقایسه مقادیر مختلف تعداد دسته ها با معیار ROUGE-2..... ۶۱
شکل ۷.۴ مقایسه مقادیر مختلف تعداد دسته ها با معیار ROUGE-SU4..... ۶۱
شکل ۸.۴ مقایسه نتایج با معیار ROUGE-2..... ۶۲
شکل ۹.۴ مقایسه نتایج با معیار ROUGE-SU4..... ۶۲

فهرست جداول

صفحه

جدول ۱.۲ نمونه ای از ایست واژه ها ۶
جدول ۲.۲ مقایسه ضریب درستی معیارهای مختلف اندازه گیری شباهت بین کلمات مبتنی بر شبکه واژگان [BUD06] ۲۶
جدول ۳.۲ اطلاعات مربوط به مجموعه داده های DUC [پور۹۰] ۳۰
جدول ۱.۳ نمونه‌ای از ایست واژه‌های محذوف از لیست ایست واژه ها ۳۷
جدول ۲.۳ مقایسه ای موردی بین دو ریشه یاب ۳۸
جدول ۳.۳ نمونه‌ای از جملات فشرده شده ۵۱
جدول ۴.۳ نمونه‌ای از جملات ادغام شده ۱ ۵۴
جدول ۵.۳ نمونه‌ای از جملات ادغام شده ۲ ۵۴
جدول ۶.۳ نمونه‌ای از جملات ادغام شده ۳ ۵۴
جدول ۱.۴داده‌های حاصل از صحت روشهای پیشین و روش پیشنهادی برای شباهت جملات ۵۷

فصل اول

مقدمه

۱ مقدمه

در این فصل مروری کلی بر خلاصه‌سازی و انواع آن خواهیم داشت. سپس به طور خلاصه به روش پیشنهادی این پایان‌نامه اشاره شده است. در انتها، قالب پایان‌نامه با اشاره به بخش‌های موجود در آن ارائه گردیده است.

۱.۱ مروری بر خلاصه‌سازی

مبحث خلاصه‌سازی متن در سالهای اخیر توسط محققان حوزه پردازش زبان طبیعی مورد بحث و بررسی قرار گرفته است. برای اولین بار در اواخر دهه ۱۹۵۰ فعالیت‌هایی در زمینه خلاصه‌سازی انجام گرفت [LUH58] و با فعالیتهای خوبی نیز ادامه یافت [SKO70][EDM69] اما از ادامه دهه ۱۹۹۰ و اوایل قرن ۲۰ بود که به طور ویژه به توسعه سیستم‌های خلاصه‌سازی خودکار توجه شد. رقابت‌های سالانه DUC[1] و TAC[2] نمونه ای از این توجه و علاقه می‌باشد.

خلاصه‌سازی خودکار متن بر عمل ایجاد نسخه‌ای کوتاه از متن اولیه دلالت دارد که همچنان اطلاعات مفید متن را در بر داشته باشد. طبق تعریف [RAD02] یک خلاصه متنی است که از یک یا چند متن اولیه تولید شده است و حاوی قسمت ارزشمندی از اطلاعات متن یا متون اولیه می‌باشد و این متن خلاصه طولانی‌تر از نصف متن یا متون اولیه نیست. در تعریف دیگر [MAN99] خلاصه‌سازی متن به نوعی روند کسب مهمترین اطلاعات از یک یا چند منبع به منظور ایجاد یک نسخه مختصر برای کاربر یا کاربران خاص و به منظور انجام وظایف یا وظیفه خاص می‌باشد.

---

[1] Document Understanding Conference

[2] Text Analysis Conference

همانطور که از تعاریف نیز برمی‌آید، خلاصه‌سازی از دیدگاه‌های مختلف انواع گوناگونی خواهد داشت. اول اینکه خلاصه می‌تواند از یک متن و یا از مجموعه‌ای از متون حاصل شود. در همین راستا دو نوع خلاصه‌سازی تک سنده[۳] و چندسنده[۴] را برای این تقسیم‌بندی ذکر نموده‌اند. از دیدگاه دیگر خلاصه می‌تواند عمومی و یا خاص منظوره باشد. خلاصه‌های عمومی خود را مقید به عنوان یا دامنه خاصی نمی‌کنند. در مقابل خلاصه‌های خاص دامنه داریم که قادر به خلاصه‌سازی موثرتر متون در عنوان یا حوزه‌ای خاص[۵] می‌باشند.

این ویژه‌سازی علاوه بر دامنه کاربرد می‌تواند شامل کاربر نیز بشود. بدین معنا که خلاصه می‌تواند مبتنی بر پرس‌وجوی کاربر[۶] باشد. تقسیم‌بندی های دیگری نیز برای خلاصه‌سازی ذکر شده است از آن میان می‌توان به خلاصه‌سازی نشان‌دهنده[۷] در مقابل آموزنده[۸] اشاره نمود. خلاصه‌سازی نشان-دهنده تنها بیان می‌دارد که متن ورودی در چه موردی است و به بیان جزئیات آن نمی‌پردازد. خلاصه‌سازی آموزنده می‌تواند جایگزینی برای متن باشد. تقسیم‌بندی مهم دیگر به روند و نحوه خلاصه‌سازی مربوط می‌شود. در این تقسیم بندی، خلاصه‌ها می‌توانند گزینشی[۹] و یا چکیده‌ای[۱۰] باشند. در خلاصه‌سازی گزینشی گزیده‌ای از مجموعه قطعات متن اولیه بدون تغییر به عنوان خلاصه برگردانده می‌شود. اغلب جمله به عنوان واحد گزینش انتخاب می‌گردد. خروجی اغلب لیستی از جملات متن یا متون اولیه است که بر مبنای اهمیتشان مرتب شده‌اند. در خلاصه‌سازی چکیده‌ای

---

[3] Single-Document Summarization

[4] Multi-Document Summarization

[5] Topic Oriented Summarization

[6] Query Oriented Summarization

[7] Indicative Summarization

[8] Informative Summarization

[9] Extractive Summarization

[10] Abstractive Summarization

جملات متن اولیه تغییر کرده و در واقع برگرفته‌ای از متن اولیه را خواهیم داشت بدیهی است که تولید این نوع خلاصه بسیار پیچیده و دشوار می‌باشد.

۲.۱ روش پیشنهادی

در این پایان‌نامه به بیان روشی برای خلاصه‌سازی چند سنده و چکیده‌ای خواهیم پرداخت. روش مذکور شامل چندین فاز می‌باشد، پس از انجام پیش‌پردازش‌های لازم و مرتبط بهترین جملات با استفاده از یک روش خلاصه‌سازی گزینشی مناسب انتخاب می‌گردند، سپس روشی جدید برای مشابهت جملات معرفی می‌نماییم. و بعد بر مبنای این مشابهت و با تکیه بر نقش‌های معنایی جملات روشی غیرنظارتی برای فشرده‌سازی جملات ارائه خواهیم داد تا قسمت‌های غیر ضروری جملات حذف گردند. پس از آن جملات را با استفاده از معیار مشابهت مذکور دسته بندی نموده و روشی نیز برای یکی نمودن جملات موجود در دسته‌ها ارائه خواهیم نمود. نتایج حاصل از ارزیابی‌ها نشانگر بهبود روش‌های ارائه شده شباهت معنایی جملات، فشرده‌سازی آنها و روش خلاصه‌سازی ارائه شده در مقایسه با روش‌های مرتبط از جمله سیستم‌های قوی موجود در DUC2007 می‌باشند.

در ادامه این گزارش، در فصل دوم به مرور کارهای انجام شده در زمینه مرتبط با پایان‌نامه خواهیم پرداخت. در فصل سوم روش پیشنهادی و در فصل چهارم نتایج ارزیابی‌ها بیان گردیده‌اند. و در انتها جمع بندی و نتیجه‌گیری بر مطالب مذکور را خواهیم داشت.

فصل دوم

مرور ادبیات

۲ مرور ادبیات

این فصل به مرور کارهای انجام شده در ارتباط با موضوع این پایان‌نامه اختصاص دارد. ابتدا با تعاریف پایه در مبحث پردازش زبان طبیعی مطرح و معرفی می‌گردند. سپس مروری بر کارهای انجام شده در زمینه خلاصه‌سازی انجام می‌گیرد. پس از آن، از آنجا که مشابهت معنایی بین اجزاء مختلف متن، جایگاه ویژه‌ای در کاربردهای مختلف پردازش متن به ویژه خلاصه‌سازی دارا می‌باشد، به معرفی تلاش‌های انجام شده در زمینه مشابهت معنایی کلمات و شباهت جملات پرداخته شده است. در ادامه، روش‌های مورد استفاده پژوهشگران برای ارزیابی خلاصه‌های تولید شده توسط سیستم‌های رایانه‌ای مورد بررسی قرار می‌گیرد. در انتها نیز مجموعه داده‌های استاندارد موجود برای خلاصه‌سازی معرفی خواهند شد.

۱.۲ تعاریف پایه در پردازش زبان طبیعی

قبل از پرداختن به هر مطلبی در زمینه پردازش زبان طبیعی از جمله خلاصه‌سازی متن، بهتر است مفاهیم پایه و تعاریف اولیه آن که به منزله الفبای پردازش متن می‌باشند، معرفی گردند. اغلب فعالیت های مربوط به این مفاهیم نوعی پیش‌پردازش متن می‌باشند، به این معنی که انجام این پردازش‌ها بر روی متن در واقع آماده‌سازی آن به منظور اعمال فرایندها و فعالیت‌های بعدی می‌باشد.

۱.۱.۲ ریشه یابی

ریشه‌یابی به فرایند تبدیل کلمات به فرم ریشه‌ای و پایه‌ای آنها اشاره می‌نماید. این فرایند در پردازش متن اهمیت بسیاری دارد، چرا که به ماشین اجازه می‌دهد با دو کلمه هم‌خانواده اما ظاهرا

متفاوت، مانند دو کلمه‌ای که از لحاظ ریشه‌ای هیچ ارتباطی با هم ندارند برخورد ننماید. الگوریتم‌های مختلفی برای ریشه‌یابی لغات پیشنهاد شده است و مورد استفاده قرار می‌گیرد. الگوریتم پیشنهاد شده در [POR80] رایجترین الگوریتم در زبان انگلیسی می‌باشد. الگوریتم دیگر، ریشه‌یابی مبتنی بر شبکه واژگان[11] می‌باشد. تفاوتی که این دو الگوریتم در شکل خروجی خواهند داشت، کاربرد این دو ریشه-یاب را متفاوت خواهد کرد.

برای روشن‌تر شدن موضوع به این مثال توجه نمایید: در الگوریتم پورتر ریشه کلمات "country" و "countries" کلمه "countri" تشخیص داده خواهد شد، در حالیکه با ریشه‌یاب مبتنی بر شبکه واژگان کلمه "country" به عنوان ریشه دو کلمه مذکور برگردانده خواهد شد.

۲.۱.۲ برچسب زنی بخش گفتاری

در دستور زبان، بخش‌های گفتار (و یا کلاسهای لغوی) طبقه‌بندی‌هایی زبانی از کلمات هستند که رفتار نحوی یک قسمت از جمله را بیان می‌دارند. به طور عموم، تمام زبان‌ها دو بخش گفتاری فعل و اسم را دارند. بقیه بخش‌های گفتاری در زبانهای مختلف متفاوت می‌باشند. از جمله مهمترین بخش-های گفتاری در زبان انگلیسی اسم، ضمیر، صفت، قید، حرف اضافه و همراه را می‌توان نام برد.

در زبان انگلیسی درصد بالایی از کلمات از لحاظ نحوی به طور مستقل دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف می‌توانند برچسب‌های نحوی متفاوتی داشته باشند. بنابراین، برچسب زنی نحوی بخش‌های گفتاری، عمل ابهام زدایی از کلمات با توجه به زمینه را انجام می‌دهد. برچسب-زنی بخش‌های گفتاری، عملی پایه‌ای برای بسیاری از حوزه‌های پردازش زبان طبیعی (NLP) از قبیل

[11] WordNet stemmer , http://projects.csail.mit.edu/jwi

ترجمه ماشینی و تبدیل متن به گفتار می‌باشد. در شکل (۱.۲) نمونه هایی از بخش‌های گفتار و برچسب‌های آنها معرفی شده است [CCG1].

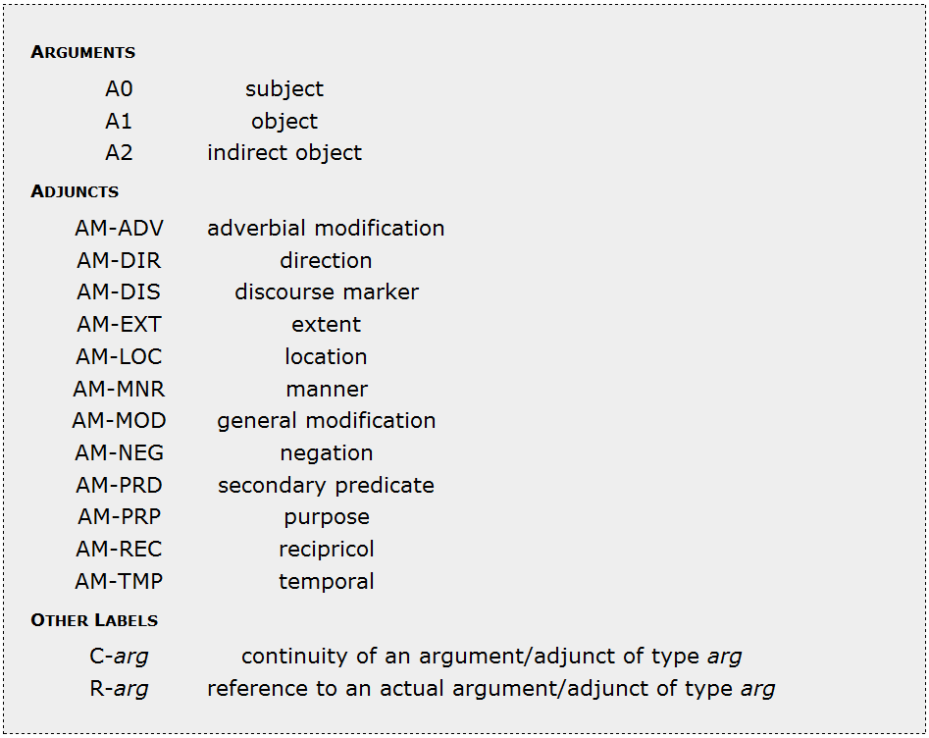
- # - Pound sign
- $ - Dollar sign
- '' - Close double quote
- `` - Open double quote
- ' - Close single quote
- ` - Open single quote
- , - Comma
- . - Final punctuation
- : - Colon, semi-colon
- -LRB- - Left bracket
- -RRB- - Right bracket
- CC - Coordinating conjunction
- CD - Cardinal number
- DT - Determiner
- EX - Existential there
- FW - Foreign word
- IN - Preposition
- JJ - Adjective
- JJR - Comparative adjective
- JJS - Superlative adjective
- LS - List Item Marker
- MD - Modal
- NN - Singular noun
- NNS - Plural noun
- NNP - Proper singular noun
- NNPS - Proper plural noun
- PDT - Predeterminer
- POS - Possesive ending
- PRP - Personal pronoun
- PP$ - Possesive pronoun
- RB - Adverb
- RBR - Comparative adverb
- RBS - Superlative Adverb
- RP - Particle
- SYM - Symbol
- TO - to
- UH - Interjection
- VB - Verb, base form
- VBD - Verb, past tense
- VBG - Verb, gerund/present participle
- VBN - Verb, past participle
- VBP - Verb, non 3rd ps. sing. present
- VBZ - Verb, 3rd ps. sing. present
- WDT - wh-determiner
- WP - wh-pronoun
- WP$ - Possesive wh-pronoun
- WRB - wh-adverb

شکل ۱.۲ نمونه هایی از بخش های گفتار و برچسب های آنها [CCG1]

۳.۱.۲ برچسب گذاری نقش معنایی

برچسب‌زنی معنایی کلمات[۱۲] مشابه برچسب گذاری بخش‌های گفتاری بوده با این تفاوت که وارد عمق بیشتری از معنا می‌شود. نقش‌های معنایی رابطه اجزاء مختلف جمله را در ارتباط با فعل متناظرشان معرفی می‌نمایند. برچسب‌زنی معنایی وظیفه استخراج نقش‌های معنایی جملات نظیر فاعل، مفعول مستقیم، مفعول غیر مستقیم، فعل و ... را بر عهده دارد. در شکل ۲.۲ یک مجموعه از نقش‌های معنایی نمایش داده شده است [CCG2].

[12] Semantic Role Labeling



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
OTHER LABELS	
C-*arg*	continuity of an argument/adjunct of type *arg*
R-*arg*	reference to an actual argument/adjunct of type *arg*

شکل ۲.۲ نمونه‌هایی از نقشهای معنایی و برچسب‌های آنها [CCG2]

۴.۱.۲ حذف ایست واژه ها

ایست واژه‌ها، لغات پرکاربرد و اغلب کم اهمیتی هستند که هنگام کار با متن به وفور با آنها برخورد می‌شود. در نگاه اولیه، کلمات ربط و تعریف، ایست‌واژه به نظر می‌آیند؛ در عین حال بسیاری از افعال، افعال کمکی، اسم‌ها، قیدها و صفات نیز ایست‌واژه شناخته شده‌اند. در اغلب کاربردهای متن حذف این کلمات، نتایج پردازش را بهبود می‌دهد. علاوه بر این، از آنجا که بیشتر کاربردهای پردازش متن با حجم عظیمی از داده ها رو به رو هستند، حذف این کلمات سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل، این کلمات عموما در فعالیت‌های مربوط به حوزه پردازش زبان طبیعی که با حجم انبوهی از داده‌ها روبه رو هستیم، در فاز پیش پردازش حذف می‌شوند. برای زبان انگلیسی چندین لیست از این کلمات منتشر شده است که بطور میانگین شامل ۵۰۰ کلمه می‌باشند. در جدول تعدادی از این لغات ذکر گردیده‌است.

جدول ۱.۲ نمونه ای از ایست واژه ها

صفت	قید	اسم	فعل کمکی	فعل	کلمات ربط
certain	currently	caption	do	appear	the
corresponding	probably	example	being	appreciate	and
dare	really	everybody	can	ask	A
immediate	truly	everything	could	believe	any

۵.۱.۲ درخت تجزیه

اجزای هر جمله را می‌توان در قالب گروه‌های اسمی، فعلی، حرف اضافه‌ای و ... تقسیم بندی نمود. گاه هر کدام از این گروه‌ها خود شامل زیرگروه دیگری می‌باشند و باز آنها نیز؛ علاوه بر این، این گروه-ها با خود نیز دارای روابطی می‌باشند، مثلا یک گروه اسمی متعلق به یک گروه فعلی می‌باشد. در نتیجه این روابط و تقسیم بندی‌های سلسله مراتبی، می‌توان یک ساختار درخت‌گونه از جمله داشت که درخت تجزیه نام دارد. به عبارت دیگر، درخت تجزیه درختی است که ساختار نحوی یک جمله را بر اساس برخی روابط گرامری موجود در آن به شکلی ساده قابل فهم برای کسانی که دانش عمیق زبانشناسانه‌ای ندارند، نمایان می‌سازد [MAR08]. ابزارهای مختلفی برای تجزیه جمله توسعه یافته‌اند که خروجی اغلب آنها به صورت رشته‌ای شامل پرانتزهای تو در تو به همراه برچسب‌ها و کلمات می-باشند. این مدل نمایش برای ورودی سیستم‌ها مناسب است، اما برای انسان خوانایی چندانی ندارد. در ابزار lfgParser[۱۳] شاهد نمایش گرافیکی و درخت‌گونه درخت تجزیه خواهیم بود.

[13] Http://lfg-demo.computing.edu.ie

۲.۲ خلاصه سازی متن

همان گونه که در مقدمه نیز ذکر شد، خلاصه سازی خودکار متن بر عمل ایجاد نسخه‌ای کوتاه از متن اولیه دلالت دارد که همچنان اطلاعات مفید را در بر داشته باشد و از دیدگاه‌های مختلف انواع مختلفی خواهد داشت. در ادامه به مهم‌ترین انواع آن و کارهای انجام شده اشاره شده است.

۱.۲.۲ خلاصه سازی تک سنده

همانگونه ذکر شد در خلاصه‌سازی تک سنده، ورودی سیستم خلاصه‌ساز، تنها یک سند می‌باشد. روش‌های متعددی برای این نوع خلاصه‌سازی پیشنهاد گردیده‌است. روش‌های مذکور در [MIH05] [SVO07] از جمله این روش‌ها می‌باشند. به طور کلی، پیچیدگی خلاصه‌سازی تک‌سنده بسیار کمتر از خلاصه‌سازی چندسنده می‌باشد؛ چرا که در خلاصه‌سازی تک سنده تنها با یک سند روبرو هستیم که با احتمال بالایی می‌توان ادعا نمود که متن آن سند در مورد یک موضوع و به صورت پیوسته بحث می‌نماید و فاقد زیر موضوعات ضد و نقیض خواهد بود. اما در خلاصه‌سازی چند سنده که با تعداد زیادی سند مواجهیم, پوشش مفاهیم عمده تمامی این اسناد کار دشوار و پیچیده‌ای می‌باشد.

۲.۲.۲ خلاصه سازی چند سنده

خلاصه‌سازی چندسنده، ارتباط تنگاتنگی با مباحث سیستم‌های پاسخگو[۱۴] و خلاصه‌سازی مبتنی بر پرس‌وجو دارد [HIR01]. در حقیقت خلاصه‌سازی چندسنده برروی اسنادی انجام می‌شود که در ارتباط با یک موضوع هستند ولی جهت دید آنها متفاوت از یکدیگر است. همانگونه که ذکر شد در خلاصه‌سازی چندسنده با پیچیدگی‌های بیشتری نسبت به خلاصه‌سازی تک‌سنده روبرو هستیم. در

[14] Question Answering

این مدل خلاصه‌سازی با دو چالش عمده مواجهیم [WAN07]: اول اینکه از آنجا که اطلاعاتی که در اسناد مختلف ذخیره شده‌اند در ارتباط با یک موضوع کلی می‌باشند، به ناچار با یکدیگر هم‌پوشانی خواهند داشت. از این رو نیازمند یک روش خلاصه‌سازی کارا جهت ادغام این اطلاعات و حذف افزونگی‌ها هستیم. دوم اینکه چون اسناد با دیدگاههای متفاوت به شرح یک موضوع می‌پردازند، احتمال موجود مضامین متناقض با یکدیگر وجود دارد. بنابراین تولید خلاصه‌ای با خوانایی بالا امری دشوار خواهد بود.

افزایش روزافزون اسناد الکترونیکی در موضوعات شبیه به یکدیگر و مشکلات کمبود زمان برای عموم کاربران و همچنین نزدیکی این بحث با موضوع سیستم‌های پاسخ گو و موتورهای جستجوگر، منجر به استقبال روز افزون دانش پژوهان جهت تحقیق در زمینه خلاصه‌سازی خودکار چندسندی شده است. با توجه به ماهیت این مدل از خلاصه‌سازی، در حال حاضر بیشتر توجهات برروی این مدل از خلاصه سازها می‌باشد [WAN07][BAR05][AKS09] . این پایان‌نامه هم بر مبنای این نوع خلاصه سازی قرار دارد.

۳.۲.۲ خلاصه سازی گزینشی

گرچه روشهای خلاصه سازی بسیار متنوع هستند اما اغلب آنها یک ویژگی مشترک دارند آن هم گزینشی بودن آنها است[STE05][GON01][CON01][CHA00][BAR99][KIA06]. همان طور که ذکر شد، خروجی این سیستم‌ها اغلب به صورت لیستی مرتب از جملات متن یا متون اولیه است و بر اساس میزان فشرده سازی (درصد فشرده سازی) که از قبل تعیین شده، تعدادی از جملات از اول لیست به عنوان متن خلاصه برگردانده می شود. اکثر این سیستم ها برای اینکه حداقل میزان افزونگی

اطلاعات را داشته باشند از ایجاد جملات مشابه به هم در متن خلاصه جلوگیری می‌نمایند [CAR98].

۴.۲.۲ خلاصه سازی چکیده ای

آنچه مسلم است این است که در خلاصه سازی خودکار، هدف اصلی رسیدن به خلاصه ای است که تا حد ممکن و از ابعاد مختلف به خلاصه انسانی شبیه‌تر و نزدیک‌تر باشد. فرایندها و روند استنتاج‌های مغز انسان در طی انجام عمل خلاصه‌سازی همانند سایر فعالیت‌های آن بسیار پیچیده و بعضاً غیر قابل شناخت و پیش‌بینی می‌باشند و بی‌شک پی بردن و فهم دقیق حتی جزء کوچکی از آن مستلزم انجام تحقیقات و آزمایش‌های روانشناسانه زمانبر و پرهزینه بسیاری می‌باشد [KIN78].

برای اینکه سیستم خلاصه سازی داشته باشیم که شبیه به انسان عمل کند بایستی این سیستم بتواند برداشتی از متن ورودی را به نحوی ایجاد و ذخیره نماید؛ به عبارتی متن ورودی را تفسیر نماید [SPA99] و سپس متن خلاصه را تولید نماید. متاسفانه تا کنون نیل به سطح قابل قبولی از این تفسیر مهیا نشده است به شکلی که به نظر می‌رسد با شرایط کنونی برای داشتن سیستمی کاربردی استفاده از خلاصه سازی گزینشی انتخاب مناسب تری باشد [MAN01]. اکثریت مطلق تلاشها برای ایجاد سیستم های خلاصه سازی بر روی توسعه سیستم‌های خلاصه سازی گزینشی متمرکز گشته است [SPA07].

با این وجود در سالهای اخیر، تلاش‌هایی در زمینه خلاصه‌سازی مبنی بر تغییر جملات اولیه و یا تولید جملات جدید به منظور نزدیک شدن به خلاصه‌سازی چکیده‌ای انجام پذیرفته است. به عنوان مثال، برای ایجاد فهرست مطالب یا تیترها از روشهای کم عمقی مانند برچسب گذاری بخشهای

گفتاری ، TFIDF[۱۵] و دوتایی[۱۶] ها استفاده شده است. مشکل روش های مذکور این است که آنچه که به عنوان تیتر برگردانده می‌شود بایستی یک عبارت یا چند کلمه کوتاه و گویا باشد اما آنچه که برگردانده می‌شود بیشتر شبیه به یک جمله است تا عبارت. Wan راهی پیشنهاد داده است که در آن چند کلمه اصلی و مهم از متن به عنوان ورودی داده می‌شود و خروجی، درخت وابستگی است که همه آن کلمات را به ترتیبی مناسب در بر می‌گیرد[WAN09][WAN08] . مشکل این روش آن است که جمله خروجی می‌تواند معنایی کاملا متفاوت با متن اولیه‌ای که کلمات اصلی را از آن استخراج کرده ایم داشته باشد.

به طور کلی تلاش‌های عمده در زمینه خلاصه‌سازی چکیده‌ای را در دو دسته می‌توان جای داد[WAN08]. دسته اول روش‌هایی هستند که اقدام به حذف قسمت‌هایی از جملات می‌نمایند که دارای بار اطلاعاتی زیادی نیستند و می‌توان بدون آنها نیز اصل مطلب را ابلاغ نمود. این روش‌ها فشرده‌سازی جملات نام دارند. روش‌های دسته دوم سعی بر آمیختن اطلاعات موجود در جملات مختلف دارند. اکثر تلاش‌های انجام گرفته در زمینه خلاصه‌سازی چکیده‌ای در دسته دوم قرار دارند و تا رسیدن به خلاصه چکیده‌ای مطلوب راه درازی در پیش روی محققین قرار دارد. در ادامه عمده‌ترین تلاش‌ها در زمینه خلاصه‌سازی چکیده‌ای بیان گردیده است.

۱.۴.۲.۲ فشرده سازی

فشرده سازی جملات به فرایند حذف قسمت‌هایی از جمله در حین حفظ اطلاعات مهم و اصلی آن اطلاق می‌گردد. از یک دیدگاه روش‌های پیشنهادی در این زمینه را می‌توان به دو دسته مبتنی بر درخت و مبتنی بر جمله تقسیم بندی نمود. روش‌های مبتنی بر درخت، با تبدیل جمله به درخت

---

[15] Term Frequency Inverse Document Frequency

[16] bigram

تجزیه و اعمال ویرایش و تصحیحاتی به آن، فرم فشرده‌ای از درخت تجزیه را تولید کرده و سپس درخت را به جمله تبدیل می‌نمایند[KNI02][GAL07][FIL08][GAL10]. روش‌های مبتنی بر جمله، تغییرات و تصحیحات را مستقیما به جمله اعمال کرده و فرم فشرده‌ای از جمله را تولید می‌نمایند [MCD06]. این روش‌ها این مزیت را دارند که پردازش کمتری نسبت به روش‌های مبتنی بر درخت دارند، اما از آنجا که درخت تجزیه فرم ساخت‌یافته تری نسبت جمله خام دارد، در روش‌های مبتنی بر درخت تجزیه فرایند تولید جمله فشرده به شکل ساختمندتری دنبال خواهد شد.

در تقسیم‌بندی دیگر روش‌های فشرده‌سازی جملات، روش‌های نظارتی[۱۷] و غیرنظارتی[۱۸] خواهیم داشت. در روش‌های نظارتی که شروع تلاش‌های مربوط به فشرده‌سازی جملات نیز با این روش‌ها انجام پذیرفت، ابتدا بایستی یک پیکره[۱۹] عظیم به منظور آموزش روند یادگیری[۲۰] ایجاد گردد[KNI02][MCD06][GAL11]. هرچه که پیکره آموزش غنی‌تر و با کیفیت تر باشد نتایج مطلوبتری کسب خواهد شد. بدیهی است روش‌های غیرنظارتی [CLA08][FIL08] که نیازمند تولید چنین پیکره‌ای نیستند از این لحاظ دارای مزیت هستند و در صورتیکه نتایج بهتر و یا حتی معادل روش‌های نظارتی کسب نمایند، برای بسیاری از محققین ترجیح خواهند داشت.

اکثر روش‌های فشرده‌سازی جملات برای بررسی صحت گرامر جملات تولید شده از یک مدل زبانی[۲۱] استفاده می‌نمایند [KNI02][GAL07][TUR05] و برخی با ایجاد یک سری قواعد و قوانین دست‌ساز [CLA08] و یا وضع یک سری قیود برای ایجاد درخت فشرده‌شده [FIL08] و اعمال این قوانین و قیود به این مقصود نایل می‌شوند. در صورتیکه بتوان روشی را در پیش گرفت که از همان

---

[17] Supervised Approaches

[18] Unsupervised Approaches

[19] Corpus

[20] Learning

[21] Language Model

ابتدا با قواعد گرامری موجود در جمله کار کند و جمله‌ای با گرامر قابل قبول تولید نماید یک مزیت نسبت به روش‌های پیشین و راه‌حلی برای این مشکل خواهیم داشت.

- **روش‌های نظارتی فشرده سازی جملات**

روش‌های نظارتی برای فشرده‌سازی جملات همانگونه که ذکر گردید می‌توانند جمله یا درخت تجزیه حاصل از آن را مبنای کار خود قرار دهند. از جمله روش هایی که مبتنی بر درخت تجریه می-باشند، روش‌های مورد استفاده در [KNI02] می‌باشد. نایت و مارکو دو روش برای کوتاه کردن و فشرده‌سازی جملات به منظور ارائه خلاصه‌ای بهتر ارائه نمودند. آنها برای شروع کار از پیکره Ziff-Davis که مجموعه‌ای از مقالاتی در مورد کالاهای کامپیوتری می‌باشد استفاده نمودند. هر یک از این مقالات شامل دو قسمت چکیده و متن اصلی می‌باشد. آنها تعداد ۱۰۶۷ جفت جمله بلند- کوتاه را از این پیکره استخراج نمودند. جمله اول جمله‌ای در متن اصلی و جمله دوم معادل همان جمله با طول کمتر در قسمت چکیده می‌باشد. پیکره تولید شده به عنوان پیکره آموزش مورد استفاده دو الگوریتم واقع می‌شود.

روش اول پیشنهادی آنها، از مدل کانال نویزی[۲۲] بهره می‌گیرد. این مدل در زمینه‌های مختلفی از جمله ترجمه ماشینی، تشخیص گفتار و غیره استفاده می‌گردد. به عنوان مثال، در ترجمه ماشینی در این مدل هنگامی که به یک جمله فرانسوی نگاه می‌کنیم، جمله فرانسوی را جمله‌ای انگلیسی می-انگاریم که مقداری نویز وارد آن شده‌است. حال هدف این است که با حذف نویزهای اضافه، به شکل اولیه‌ی جمله که همان جمله انگلیسی می‌باشد برسیم. در زمینه فشرده‌سازی جملات نیز وقتی به جمله بلند نگاه می‌کنیم، تصور می‌نماییم که این یک جمله کوتاه بوده که تعدادی کلمه دیگر به آن

[22] Noisy Channel

اضافه شده‌است. سعی بر آن است تا کلمات اضافه را حذف نماییم. روش کار به این ترتیب می‌باشد که فرض می‌کنیم s درخت تجزیه مربوط به جمله کوتاه و فشرده شده و l درخت تجزیه مربوط به جمله بلند اولیه باشد. روش از مدل زبانی در اینجا یعنی $P(s)$ و مدل کانالی در اینجا یعنی $P(l|s)$ استفاده می‌نماید. بهترین فرم فشرده شده، درختی است که مقدار $P(s)*P(l|s)$ را بیشینه نماید. برای تخمین مقدار $P(l|s)$ احتمال تمام عملیات بسطی که نیاز است انجام شود تا درخت تجزیه s به درخت تجزیه l تبدیل شود محاسبه می‌گردد.

روش دوم نیز بر مبنای مدل شرطی C4.5 [QUI93] می‌باشد. این روش سعی بر تبدیل مستقیم درخت تجزیه s به درخت تجزیه l دارد. این روش یاد می‌گیرد چه زمانی باید عملیات حذف و چه زمانی عملیات ترکیب زیر درخت‌ها را انجام دهد که در آن هر تصمیم (انتقال یا کاهش) بر مبنای آنچه در پشته باقی مانده و جمله نیمه تمام فشرده شده گرفته می‌شود.

مک‌دونالد [MCD06] نیز از همان پیکره به منظور یادگیری وزن‌ها و تشکیل بردارهای وزنی استفاده نمود. هر درخت کاندید برای فشرده‌سازی توسط ضرب داخلی بردارهای وزنی در بردارهای ویژگی که از برچسب‌های بخش گفتاری، چندتایی[۲۳]ها و وابستگی‌های موجود در درخت تجزیه به دست می‌آیند، امتیاز دهی می‌گردد. هر دنباله‌ای از کلمات که تابع هدف را بیشینه نماید بهترین فرم فشرده جمله را تشکیل خواهد داد. در [BER11] نیز یک روش یادگیری توامان برای گزینش و فشرده‌سازی جملات درون یک مدل یکپارچه ارائه شد.

تمام روش‌های نظارتی پیشنهاد شده برای فشرده‌سازی جملات نیازمند پیکره‌ای به منظور یادگیری هستند تا الگوریتم قادر باشد اقدام به حذف یا نگهداری اجزاء مختلف جمله نماید. همانگونه

---

[23] N-gram

که اشاره شد، این مطلب یکی از معایب اینگونه روش‌هاست. چرا که حصول یک پیکره غنی و قدرتمند با صرف هزینه و زمانی بسیار میسر است.

- **روش‌های غیر نظارتی فشرده‌سازی جملات**

گرچه اغلب روش‌های ارائه شده به منظور فشرده‌سازی جملات نظارتی هستند، روش‌هایی غیرنظارتی نیز بدین منظور ارائه گردیده‌است. کلارک [CLA08] روشی ارائه داد که بهترین فرم فشرده با استفاده از برنامه‌ریزی خطی صحیح[۲۴] یافت می‌گردد. تابع هدف با استفاده از یک مدل زبانی استفاده می‌نماید تا مشخص نماید کدام چندتایی بیشترین احتمال حذف را داراست. در نهایت برای بررسی صحت گرامری جملات تولیدی یک سری قوانین دست‌ساز به درختان تجزیه جملات اعمال می‌گردد. روش غیرنظارتی دیگر از مشابه روش قبل از برنامه ریزی خطی استفاده می‌نماید [FIL08]. این روش با تبدیل درخت تجزیه به درخت وابستگی اقدام به فشرده‌سازی درخت وابستگی به جای جمله اولیه می‌نماید. در تابع هزینه اهمیت کلمات جمله در متن و احتمال شرطی وابستگی‌های موجود در درخت لحاظ می‌گردند.

فرض کنید در درخت وابستگی یالی با برچسب l از گره h به گره w داشته باشیم، $x^{l}_{h,w}$ به شکل زیر تعریف می‌گردد:

$$x^{l}_{h,w} = \begin{cases} ۱ & \text{اگر وابستگی در درخت حضور داشته باشد} \\ ۰ & \text{در غیر این صورت} \end{cases} \qquad (۱.۲)$$

تابع هدف نیز به شکل زیر تعریف می‌گردد:

[24] Integer Linear Programming

$$f\ (X\ ) = \sum_{x} x^{l}_{h,w}\ * P\left(l \mid h\right) * I\left(w\right) \qquad (۲.۲)$$

که در آن اهمیت کلمات یا همان $I(w)$ از رابطه زیر به دست می‌آید:

$$I(w_i) = f_i\ * \log \frac{F_A}{F_i} \qquad (۳.۲)$$

که در آن f_i فراوانی کلمه w_i در سند، F_i فراوانی کلمه w_i در پیکره و F_A مجموع فراوانی همه کلمات اصلی در پیکره می‌باشد. $P\left(l \mid h\right)$ نیز احتمال حضور یالی با برچسب l از نود h می‌باشد. به عنوان مثال $P\left(with \mid work\right)$ یا $P\left(subj \mid work\right)$.

دو نوع قید به تابع هدف برای ساخت درخت اعمال می‌گردد. قیود ساختاری و قیود نحوی. قیود ساختاری از حضور وابستگی‌های اجباری در درخت اطمینان حاصل می‌نمایند و قیود نحوی بررسی می‌نمایند در صورتیکه یک گره در خروجی ظاهر نشود یالهای مربوط به آن نیز از خروجی حذف گردند. در نهایت درخت حاصل بایستی به جمله تبدیل گردد. اشکال این روش‌ها درگیری با درخت‌ها و محاسبات احتمالاتی فراوان می‌باشدکه امکان کاهش دقت را فراهم می‌آورند، به طوری که مشاهده کردیم برای اطمینان از صحت گرامری و ساختاری جملات نیاز به اعمال قیود و قوانین زیادی است که نوشتن آنها زمانبر می‌باشد. در فصل سه روشی روشی ارائه خواهد شد که بدون درگیری با درخت و قوانین اضافه و با تکیه بر نقش‌های معنایی جملات فشرده شده مناسبی تولید می‌نماید که از لحاظ گرامر نیز قابل قبول خواهند بود.

۲.۴.۲.۲ آمیختن جملات

آمیختن جملات عبارتست از یک فرایند تولید متن به متن، که تعدادی جمله مشابه را به عنوان ورودی می‌گیرد و جمله‌ای جدید ایجاد می‌نماید که حاوی اطلاعات مشترک بین جملات مشابه می-

باشد [BAR05]. در زمینه آمیختن جملات کار چندانی انجام نپذیرفته‌است. عمده‌ترین کار در این زمینه در [BAR99] و [BAR05] معرفی شده‌است. ابتدا جملات ورودی با استفاده از یک روش مشابهت جملات [HAT99] به دسته‌هایی تقسیم‌بندی می‌شوند. سپس در هر دسته یک جمله به عنوان جمله مرکزی انتخاب شده و درخت جمله مذکور با زیردرخت‌هایی از جملات دیگر موجود در دسته تجهیز شده و یک گراف پدید می‌آید. گراف مذکور می‌تواند به شکل‌های مختلف به درخت تبدیل شود، اما درخت حاصل بایستی شکل استاندار درخت تجزیه را دارا باشد. سپس درخت حاصل طی فرایند خطی سازی به جمله تبدیل خواهد شد. آنچه حاصل می‌شود اشتراکی از جملات موجود در دسته خواهد بود. برای بررسی صحت گرامر جملات هم از مدل زبانی استفاده می‌گردد.

۳.۲ ارتباط معنایی کلمات

در بسیاری از مواقع در کاربردهای مختلف پردازش متن نیازمند اندازه‌گیری ارتباط معنایی بین کلمات و متناظراً مفاهیم هستیم. این مبحث در کاربردهایی مثل رفع ابهام واژه‌ها، خلاصه سازی متن و تصحیح خودکار لغات، به شکل قابل توجهی تاثیر گذار است. روش‌هایی که برای اندازه‌گیری ارتباط معنایی کلمات از یک منبع لغوی استفاده می نمایند، آن منبع لغوی را به عنوان یک شبکه یا گراف می بینند و ارتباط معنایی را بر اساس خصوصیات مسیرها در این گراف محاسبه می‌نمایند.

۱.۳.۲ شبکه واژگان

در بین منابع موجود، شبکه واژگان[۲۵] مورد توجه بسیار قرار گرفته و روشهای متعددی برای محاسبه ارتباط بین کلمات بر اساس شبکه واژگان پیشنهاد گردیده است. شبکه واژگان شبکه ای از

[25] WordNet

واژگان انگلیسی با گستره‌ای وسیع می باشد [BUD06]. کلمات در کلاس‌های مختلف گفتاری یعنی اسم، فعل، قید و صفت هر کدام در زیرشبکه هایی از هم معنی[۲۶] ها به صورت سازماندهی شده قرار گرفته اند. هرکدام از مجموعه های هم معنی، نماینده یک مفهوم متناظر با آن مجموعه می‌باشند. مفاهیم توسط روابط مختلفی با هم در ارتباط هستند. شبکه مربوط به اسامی، اولین شبکه ای بود که به شکلی دقیق و کامل توسعه یافت و بسیاری از تلاش‌ها در زمینه اندازه‌گیری ارتباط معنایی کلمات خود را محدود به این شبکه کردند.

ستون فقرات شبکه اسمی، روابط سلسله مراتبی[۲۷] می‌باشد که حدود ۸۰ درصد حجم روابط را تشکیل می‌دهد. در بالاترین سطح از سلسله مراتب یازده مفهوم انتزاعی وجود دارد که آغازگرهای یکتا[۲۸] نامیده می‌شوند؛ مانند "entity" یا "psychological feature" . حداکثر عمق سلسله مراتب، یازده گره می‌باشد. علاوه بر رابطه هم معنی که ذکر شد، ۹ نوع رابطه دیگر نیز در زیرشبکه اسامی تعریف شده است. رابطه زیرمفهوم[۲۹] و معکوس آن ابر مفهوم، ۶ رابطه عضویت[۳۰] (شامل : جزئی از[۳۱]، عضوی از[۳۲] و نمونه‌ای از[۳۳]) و معکوس های آن ها و در نهایت رابطه متضاد[۳۴].

مستقل از هر فرمول و روش ارائه شده، ارتباط بین دو واژه از رابطه زیر به دست می‌آید [RES95]:

---

[26] synset

[27] Hyponymy/hypernymy relation

[28] unique beginners

[29] IS-A relation

[30] PART-OF relation

[31] COMPONENT-OF relation

[32] MEMBER-OF relation

[33] SUBSTANCE-OF relation

[34] COMPLEMENT-OF relation

$$rel(w_1,w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} \left[rel(c_1,c_2) \right] \qquad (۴.۲)$$

که در آن $s(w_i)$ مجموعه مفاهیمی در سلسله مراتب می باشد که متناظر معانی مختلف واژه w_i می‌باشند. به عبارتی دیگر، c_1 و c_2 مرتبط ترین مفاهیم در سلسله مراتب شبکه واژگان هستند که جزو مجموعه معانی w_1 و w_2 باشند. این بیان به مدل ذهنی انسان نیز بسیار نزدیک است؛ چرا که به عنوان مثال، هنگامی که چند واژه چند معنی به نحوی همزمان به انسان ارائه می‌شود ناخودآگاه ذهن به سمت معانی از آنها متوجه می‌شود که در نزدیکترین حالت به یکدیگر قرار داشته‌باشند. از این واقعیت به منظور محاسبه شباهت بین جملات در بخش ۳.۳ این پایان نامه استفاده شده است که در آنجا به تفصیل شرح داده خواهد شد.

در ادامه روش های مختلفی که برای بیان ارتباط بین مفاهیم با استفاده از شبکه واژگان ارائه شده‌اند معرفی می‌گردند.

۲.۳.۲ روش های مبتنی بر طول مسیر

یک راه برای محاسبه ارتباط بین مفاهیم درون شبکه واژگان آن است که آن را به صورت یک گراف تصور نموده و ارتباط بین مفاهیم را بر اساس طول مسیر بین آنها شکل دهیم. با این فرض که هر چه طول مسیر و یا تعداد یالهای بین آن دو مفهوم کمتر باشد ارتباط آنها با هم بیشتر خواهد بود و برعکس. بر همین مبنا هیرست [HIR95][HIR98] رابطه‌ای را برای بیان ارتباط بین دو مفهوم در شبکه واژگان بیان نمود که از این قرار می‌باشد:

$$rel_{HS}(c_1,c_2) = C - len(c_1,c_2) - k * turns(c_1,c_2) \qquad (۵.۲)$$

که در آن C و k دو مقدار ثابت می‌باشند که توسط پیشنهاد دهندگان فرمول متناظراً برای آنها مقادیر ۸ و ۱ در نظر گرفته شده است. تابع $len(c_1,c_2)$ تعداد یال‌ها در کوتاهترین مسیر بین دو مفهوم را برمی‌گرداند. $turns(c_1,c_2)$ نیز تعداد تغییر جهت در مسیر بین دو مفهوم می باشد. در طول یک مسیر سه جهت مختلف وجود خواهند داشت: بالا به پایین که در روابطی مثل زیرمفهوم و زیر عضو تحقق می‌یابد[۳۵]. پایین به بالا که در روابطی مثل ابر مفهوم[۳۶] تحقق می‌یابد و افقی که در رابطه متضاد تحقق می‌یابد. این الگوریتم قبلا توسط موریس [MOR91] بر روی پیکره Roget برای به دست آوردن فاصله معنایی معرفی شده بود.

۳.۳.۲ مقیاس بندی شبکه

روشهایی که تنها از طول مسیر برای بیان ارتباط مفاهیم بهره گرفته اند بر این مبنا بنا شده‌اند که تمام لینک‌های موجود در شبکه و یا یال‌ها بیانگر فاصله‌ای واحد بین مفاهیم می‌باشند. بدین معنا که به یال‌های مختلف از لحاظ میزان تاثیرگذاری در ارتباط مفاهیم، به دید واحدی نگریسته می‌شود. این فرض توسط برخی محققین نادرست تلقی شده و منجر به پیشنهاد روشهای بعدی گشته است.

در همین راستا در [SUS93] روشی بیان شده‌است بر مبنای این فرض که مفاهیم همزادی که در اعماق سلسله مراتب قرار دارند ارتباط نزدیک تری نسبت به مفاهیم همزادی دارند که در جایگاه-های کم عمق تر سلسله مراتب واقع شده اند. سپس وزنی برای هر یال با برچسب r که از مفهوم می‌آید تعیین می‌شود.

[35] generalization

[36] specialization

$$wt\left(c_1 \rightarrow_r\right) = \max_r - \frac{\max_r - \min_r}{edges_r\left(c_1\right)} \qquad (۶.۲)$$

که در آن $\max_r$ و $\min_r$ به ترتیب بیشترین و کمترین مقادیر ممکن برای وزن رابطه r می‌باشند. و $edges_r\left(c_1\right)$ بیان کننده تعداد یالهایی است که با برچسب رابطه r از مفهوم c_1 خارج می‌شوند. فاصله بین دو گره مجاور میانگین وزن هایی است که در جهات مختلف روی یال برآورد می‌شود که البته با عمق گره مقیاس‌بندی می‌شود.

$$dist_s\left(c_1,c_2\right) = \frac{wt\left(c_1 \rightarrow_r\right) + wt\left(c_2 \rightarrow_{r'}\right)}{2 * \max\left\{depth\left(c_1\right), depth\left(c_1\right)\right\}} \qquad (۷.۲)$$

که در آن r رابطه ای است که در شبکه واژگان بین دو مفهوم c_1 و c_2 برقرار است و r' نیز معکوس آن می‌باشد. در نهایت رابطه معنایی بین دو مفهوم دلخواه برابر خواهد بود با مجموع فواصل بین جفت مفاهیم همجواری که در طول کوتاه‌ترین مسیر بین آن دو قرار دارند.

در [WU94] نیز یک معیار رابطه بین دو مفهوم موجود در شبکه واژگان مطرح گردید که مشابهت مفهومی نام گرفت:

$$sim_{wp}\left(c_1,c_2\right) = \frac{2 * depth\left(lso\left(c_1,c_2\right)\right)}{len\left(c_1, lso\left(c_1,c_2\right)\right) + len\left(c_2, lso\left(c_1,c_2\right)\right) + 2 * depth\left(lso\left(c_1,c_2\right)\right)} \qquad (۸.۲)$$

$lso(c_1,c_2)$ مفهومی است که بر پایین ترین پدر مشترک[۳۷] دو مفهوم c_1 و c_2 در سلسله مراتب اشاره دارد. از آنجا که این مقدار، عددی است بین صفر و یک فاصله بین دو مفهوم به شکل روابط زیر بیان می‌گردد.

$$dist_{wp}(c_1,c_2) = 1 - sim_{wp}(c_1,c_2) \qquad (۹.۲)$$

و در نتیجه خواهیم داشت:

$$dist_{wp}(c_1,c_2) = \frac{len(c_1, lso(c_1,c_2)) + len(c_2, lso(c_1,c_2))}{len(c_1, lso(c_1,c_2)) + len(c_2, lso(c_1,c_2)) + 2*depth(lso(c_1,c_2))} \qquad (۱۰.۲)$$

در [LEA98] رابطه‌ای برای محاسبه مشابهت معنایی بین مفاهیم بیان گردیده است که از این قرار می‌باشد:

$$sim_{LC}(c_1,c_2) = -\log \frac{len(c_1,c_2)}{2*\max_{c \in WordNet} depth(c)} \qquad (۱۱.۲)$$

که در آن $len(c_1,c_2)$ به تعداد یالهای بین دو مفهوم در شبکه واژگان اشاره می‌نماید.

۴.۳.۲ روش های مبتنی بر اطلاعات[۳۸]

در این روش ها سعی بر آن شده تا دید را از نگاه صرف به شبکه واژگان فراتر برده و با استفاده از اطلاعات بیشتری، کیفیت روابط به دست آمده بین مفاهیم را ارتقا دهند. بدین منظور تابعی به نام

[37] Most-Specific Supreme

[38] Information-based approaches

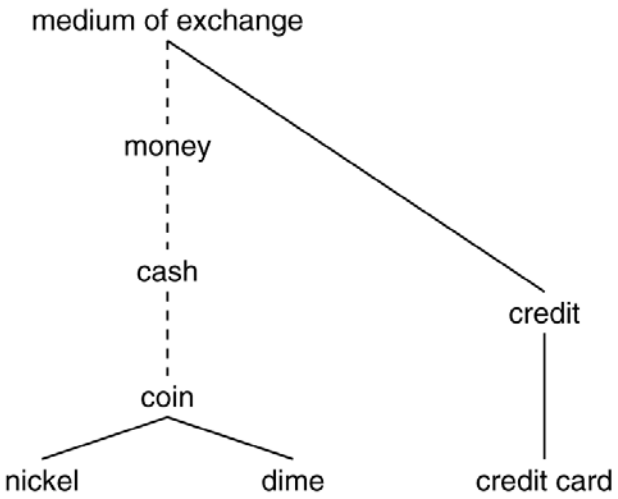
$p(c)$ معرفی می‌شود که معرف احتمال برخورد با **نمونه‌ای** از مفهوم c و یا خود مفهوم می‌باشد. با توجه به تعاریف استاندارد موجود در تئوری اطلاعات، محتوای اطلاعاتی[۳۹] مفهوم c یا $IC(c)$ برابر است با $-\log p(c)$. بر همین اساس مشابهت دو مفهوم از رابطه زیر بدست خواهد آمد[RES95]:

$$sim_R(c_1,c_2) = -\log p\left(lso(c_1,c_2)\right) \qquad (۱۲.۲)$$

قابل توجه است که در تابع p ، همچنان که بر روی سلسله مراتب بالا می‌رویم به شکل صعودی شاهد افزایش مقدار تابع خواهیم بود؛ به طوریکه قاعده زیر همواره برقرار خواهد بود:

If c_1 IS-A c_2 then $p(c_1) <= p(c_2)$

در نتیجه هر قدر که جایگاه پایین ترین پدر مشترک دو مفهوم در سلسله مراتب بالاتر باشد آن دو مفهوم مشابهت کمتری خواهند داشت.



شکل ۳.۲ قسمتی از سلسله مراتب شبکه واژگان [RES95]

[39] Information Content

شکل ۳.۲ قسمتی از سلسله مراتب شبکه واژگان را نمایش می‌دهد. پایین ترین پدر مشترک nickel و dime و همچنین nickel و credit card نشان داده شده است. خطوط پیوسته بیانگر روابط IS-A می‌باشند. در محل خطوط نقطه چین تعدادی گره حذف شده‌اند. رسنیک در کار خود برای محاسبه تابع p از فرکانس کلمات موجود در پیکره یک میلیون واژه ای Brown [FRA82] استفاده نمود. خصوصیت اصلی شمارش او برای تعیین فرکانس کلمات این بود که هنگام شمارش یک کلمه متناظر با یک مفهوم، کلمات متناظر با مفاهیم زیرین آن مفهوم در سلسله مراتب که در متن حضور داشته باشند نیز به حساب آورده می‌شوند. به عنوان مثال، هنگام شمارش کلمه coin، فرکانس کلمات dime و nickel نیز به فرکانس کلمه coin اضافه خواهند شد. نتیجه آنکه برای محاسبه تابع p از رابطه (۱۳.۲) استفاده می‌شود:

$$p(c) = \frac{\sum_{w \in W(c)} count(w)}{N} \qquad (۱۳.۲)$$

که در آن $W(c)$ مجموعه کلماتی هستند که مفهوم متناظر با آنها زیر مفهوم c باشند، و N تعداد کلمات مشترک بین شبکه واژگان و پیکره می‌باشد.

با توجه به رابطه رسنیک اگر دو جفت مفهوم دارای پایین ترین پدر مشترکی باشند دارای شباهت یکسانی خواهند بود؛ به عنوان مثال، شباهت money و credit با شباهت dime و credit card برابر است و این یکی از معایب این روش می‌باشد. ژیانگ و کنارث [JIA97] تصمیم به ترکیب روش های مبتنی بر یال و مبتنی بر گره گرفتند. به این ترتیب که در سلسله مراتب واژگان فاصله معنایی بیان شده توسط یالی که فرزند را به پدر متصل می‌نماید متناسب با احتمال شرطی $p(c \mid par(c))$ می‌باشد و از آنجا فاصله یک مفهوم و مفهوم بالای آن برابر است با تفاضل تابع IC آن دو. فاصله دو مفهوم دلخواه به صورت رابطه (۱۴.۲) خلاصه می‌گردد.

$$dist_{JC}(c_1,c_2) = IC(c_1) + IC(c_2) - 2*IC(lso(c_1,c_2)) \qquad (۱۴.۲)$$

لین [LIN98] روشی عمومی ارائه داد تا محاسبه شباهت بین دو شیء، منحصر به مفاهیم موجود در شبکه واژگان نباشد و برای هر دو شیء دلخواه نیز کاربرد داشته‌باشد. علاوه بر آن، او از یک سری فرضیات جانبی استفاده کرد تا جواب ها را از حالتی که صرفاً از طریق فرمول به دست می‌آیند بهبود دهد و در همین راستا فرضیه مشابهت خود را ارائه نمود:

> *"مشابهت بین دو شیء A و B با نسبت میزان اطلاعاتی که نیاز دارند تا مشترکات یکدیگر را بیان نمایند به اطلاعاتی که نیاز دارند تا دقیقاً خود را توصیف نمایند محاسبه می‌گردد."* [LIN98]

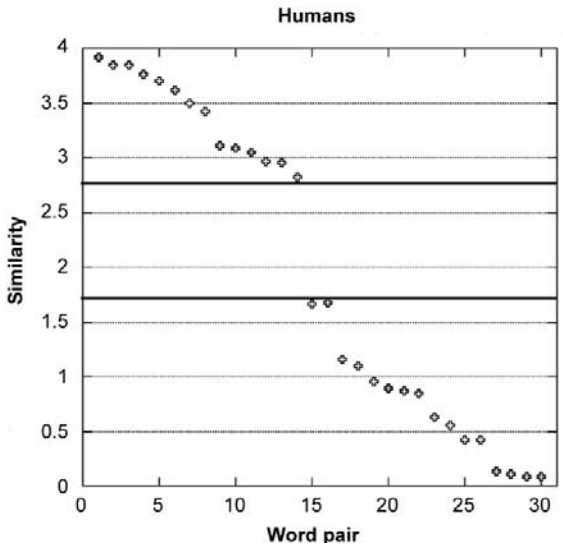
با به کار بردن این فرضیه برای دو مفهومی که در شبکه واژگان وجود دارند، فرمول مشابهت زیر به دست خواهد آمد:

$$sim_L(c_1,c_2) = \frac{2\log p(lso(c_1,c_2))}{\log p(c_1) + \log p(c_2)} \qquad (۱۵.۲)$$

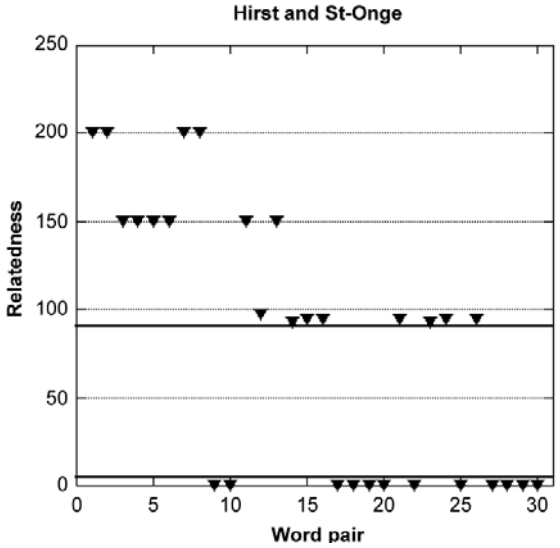
۵.۳.۲ ارزیابی روش های محاسبه شباهت معنایی

انتخاب روشی مناسب برای محاسبه میزان مشابهت کلمات از بین روش های مذکور تا حدودی بستگی به کاربرد و زمینه مورد استفاده دارد. علاوه بر این، روشهای متعددی می‌توان به کار برد تا کارایی این روش ها را سنجید. یکی از این روش ها مسلماً ارزیابی انسانی است. روش دیگر بررسی کارایی کاربردی است که از این روش ها برای رسیدن به مقصود خود استفاده می‌نمایند. رابن اشتاین [RUB65] نتایج قضاوت انسانی را بر روی ۶۵ جفت کلمه از ۵۱ موضوع مختلف ارائه نمود. جفت

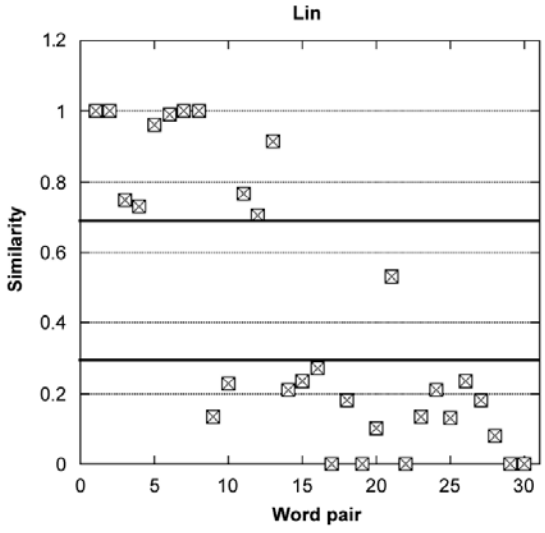
کلمات از "کاملاً هم معنی" تا "بی ارتباط" رتبه بندی شده و سپس به رنجی بین ۰ تا ۴ مقیاس بندی شدند. در شکل های زیر نتایج هر کدام از روش های مذکور به همراه نتایج قضاوت انسانی که به صورت عددی بین صفر تا چهار است بر روی داده های مذکور نشان داده شده است [BUD06]. ضریب درستی این روش ها نیز در جدول آمده است. همان طور که از شکل ها هم بر می‌آید، لین و لیکوک بهترین جواب ها را به خود اختصاص داده‌اند. از این رو در این پایان‌نامه نیز در مرحله‌ای که نیاز به محاسبه شباهت معنایی بین کلمات می‌باشد از معیار لین استفاده گردیده‌است.

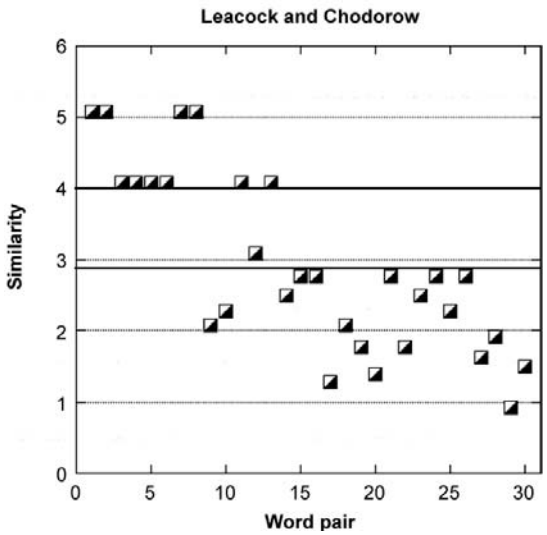


شکل ۴.۲ ارزیابی انسانی بر شباهت کلمات [BUD06]

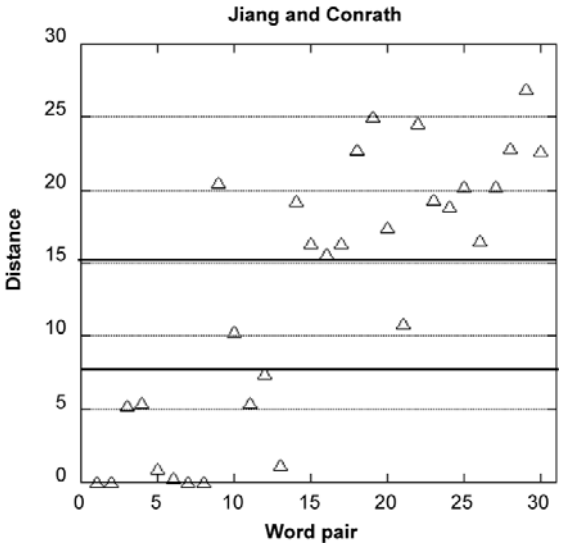


شکل ۵.۲ ارزیابی معیار هیرست بر شباهت کلمات [BUD06]

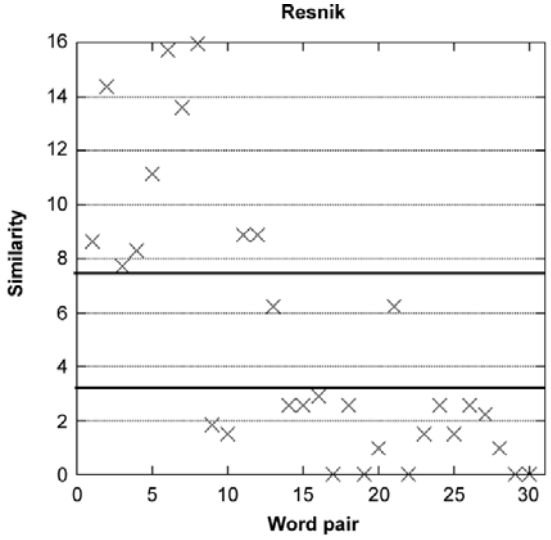




شکل ۶.۲ ارزیابی معیار لین بر شباهت کلمات [BUD06]



شکل ۷.۲ ارزیابی معیار لیکوک بر شباهت کلمات [BUD06]



شکل ۸.۲ ارزیابی معیار ژیانگ بر شباهت کلمات [BUD06]

شکل ۹.۲ ارزیابی معیار رسنیک بر شباهت کلمات [BUD06]

جدول ۲.۲ مقایسه ضریب درستی معیارهای مختلف اندازه گیری شباهت بین کلمات مبتنی بر شبکه واژگان [BUD06]

معیار	ضریب درستی
هیرست	۰.۷۶۸
ژیانگ	۰.۷۸۱
لیکوک	۰.۸۳۸
لین	۰.۸۱۹
رسنیک	۰.۷۷۹

۴.۲ شباهت جملات

شباهت جملات مبحثی است که در زمینه‌های مختلف پردازش زبان طبیعی بسیار تاثیرگذار می-باشد. سیستم‌های پرسش و پاسخ نیازمند تعیین شباهت بین جفت‌های سوال-پاسخ و یا سوال-سول می‌باشند[ACH08]. در زمینه خلاصه‌سازی مبتنی بر گراف برای وزن‌دهی به یالها به شباهت بین جملات نیاز است[ERK04]. کاربردهای دیگری چون دسته بندی متن [KO04]، و ترجمه ماشینی

[LIU04] از جمله زمینه‌های دیگری هستند که از شباهت جملات استفاده می‌نمایند. فرآیند محاسبه شباهت بین جملات فرایندی بسیار دشوار و پیچیده می‌باشد. علی غم اینکه بسیاری از کاربردها از معیارهای شباهت استفاده می‌کنند، اما بیشتر روش‌ها جملات را فقط بر مبنای سطح ظاهری مقایسه می‌کنند تا بر اساس معنا. علاوه بر این در اکثر روش‌هایی که پردازش‌های زبانی جمله را نیز در نظر می‌گیرند، تنها سطوح پایینی از آن مانند برچسب‌های بخشهای گفتاری را مد نظر قرار می‌دهند.

در این پایان نامه روشی برای محاسبه شباهت بین جملات ارائه شده است که در آن برای محاسبه شباهت بین کلمات علاوه بر شباهت ساده بین آن‌ها از شباهت معنایی کلمات بهره گرفته می‌شود. علاوه بر این سطوح بالایی از پردازش زبانی جمله مانند برچسب‌های نقش معنایی علاوه بر سطوح پایین آن در نظر گرفته می‌شود که این امر سبب افزاش دقت روش اندازه‌گیری می‌گردد. این روش در بخش ۴.۳ به تفصیل شرح داده خواهد شد.

روش‌های اندازه‌گیری شباهت جملات به سه دسته عمده تقسیم‌بندی می‌شوند [ACH08]، معیارهای هم‌پوشانی کلمات، معیارهای TFIDF و معیارهای زبانی. در ادامه شرحی از این روش‌ها آورده شده است.

در معیارهای هم‌پوشانی کلمات، شباهت دو جمله بر اساس فراوانی کلمات مشترک بین دو جمله محاسبه می‌گردد. در [MET05] دو معیار مبتنی بر اشتراک کلمات برای تعیین شباهت بین جملات معرفی شده است: اشترک ساده و اشتراک IDF. تابع اشترک ساده کلمات $sim_{overlap}$ به صورت نسبت تعداد کلمات مشترک دو جمله به طول جملات تعریف می شود. تابع اشتراک IDF هم به صورت تعداد کلمات مشترک دو جمله که توسط معکوس تکرار سندشان (IDF) وزن دهی شده باشند، تعریف می‌شود. یکی از مشکلات این روشها این است که تفاوتی بین عبارات تک کلمه ای و چند کلمه‌ای قائل نیستند. برای حل این مشکل [BAN03]، معیار همپوشانی را بر اساس رابطه بین

طول عبارات و تعداد رخدادشان در یک مجموعه متن تعریف می‌کند. همپوشانی برای m همپوشانیِ عبارتیِ n کلمه‌ای به صورت تابع زیر نشان داده می‌شود:

$$overlap_{phrase}(s_1, s_2) = \sum_{i=1}^{n} \sum_{m} i^2 \qquad (۱۶.۲)$$

که در آن m تعداد عبارت های i-کلمه‌ای می‌باشد که در هر دو جمله ظاهر شده‌اند.

معیارهای زبانی به شباهت معنایی که جلوتر اشاره شد و همچنین دانش زبانی مانند روابط بین کلمات و سایر اجزای جمله اشاره دارند. در [ISL08] یک روش برای اندازه‌گیری شباهت معنایی جملات ارائه شده است. این روش، بر مبنای شباهت معنایی کلمات و یک نسخه تغییر یافته و نرمال شده از LCS[40] برای نگاشت رشته ها می‌باشد. این روش، هم از اطلاعات معنایی و هم از اطلاعات نحوی موجود در جملات برای تعیین شباهت بین جملات استفاده می‌نماید. روش ارائه شده در [LEE11] دارای سه فاز اصلی می‌باشد. در این روش پس از پیش‌پردازش، برای محاسبه شباهت معنایی بین کلمات، از روش ارائه شده در [WU94] استفاده می‌نماید. این شباهتها در برچسب‌های مشخصی از گفتاری (اسم و فعل) محاسبه می‌شود و این شباهت ها به فضای برداری اسمی و فعلی نگاشت داده می‌شود. مقادیر هر فیلد در این بردارها برابر مقدار بیشینه مشابهت معنایی کلمه متناظر در بردار با سایر کلمات موجود در فضای برداری می‌باشد. شباهت معنایی بین هر جفت بردار توسط فاصله کسینوسی بین بردار متناظر (اسمی یا فعلی) محاسبه می‌گردد. دو مقدار به دست آمده با ضریبی با یکدیگر ترکیب شده و مشابهت نهایی را به دست می‌آورند. در [AKS09] فرکانس کلمات در نقش‌های معنایی محاسبه می‌شود و سپس از روش کسینوسی برای تعیین فاصله بین بردار جملات استفاده می‌گردد.

---

[40] Longest Common Substring

۵.۲ مجموعه داده های استاندارد

یکی از چالش‌های اساسی در زمینه خلاصه‌سازی خودکار متن، بحث ارزیابی سیستم‌های ارائه شده می‌باشد. یک ارزیابی مناسب و دقیق در گرو یک مجموعه داده مناسب و استاندارد می‌باشد. در مقالات از داده های مختلفی تاکنون استفاده شده است که از جمله آنها می توان به مجموعه داده‌های خبری BBC ، CNN ، TREC[۴۱]، CASTcorpus[۴۲]، DUCcorpus اشاره نمود [پور۹۰]. از آنجا که در زمینه خلاصه‌سازی مجموعه داده DUC جزء موارد پرکاربرد می‌باشد، در این پایان نامه نیز از آن استفاده شده است. در ادامه خصوصیات این مجموعه داده ها شرح داده شده است.

۱.۵.۲ داده های استاندارد DUC

سازمان NIST[۴۳] به عنوان یکی از نمایندگی‌های وزارت بازرگانی ایالات متحده، یکی از بخش‌های خود را به فعالیت‌های مربوط به بازیابی اطلاعات و پردازش زبان طبیعی اختصاص داده است. در بخش مذکور، گروه‌های مختلفی در حال انجام فعالیت می‌باشند که یکی از آنها به تولید مجموعه‌های تست به منظور بهبود فعالیت های مربوط به بازیابی اطلاعات از جمله خلاصه‌سازی می‌پردازد. یکی از پروژه-های این موسسه در زمینه خلاصه‌سازی، DUC می‌باشد. کنفرانس DUC تا کنون ۷ مجموعه از داده-ها را تحت عنوان DUC2001 تا DUC2007 ارائه نموده است. هر کدام از این مجموعه‌ها با اهداف خاصی انتشار یافته‌اند. با توجه به اینکه مجموعه DUC2007 آخرین مجموعه از این داده‌ها و کاملترین آنها می‌باشد، در این پایان‌نامه، از این مجموعه برای ارزیابی استفاده شده است.

---

[41] http://trec.nist.gov

[42] http://clg.wlv.ac.uk/projects/CAST/corpus/

[43] national institute of standards and technology

داده‌های DUC2007 که به منظور خلاصه‌سازی چند سنده فراهم آمده‌اند، در مجموع شامل ۴۵ عنوان مختلف می‌باشد. هر کدام از عناوین، شامل ۲۵ سند است. برای هر عنوان ۴ نفر به صورت تصادفی خلاصه‌های چکیده‌ای (انسانی) تولید نموده‌اند. ۳۲ سیستم خلاصه‌ساز در این پروژه، برای هر موضوع، خلاصه خودکار تولید کردند. این خلاصه ها با استفاده از ابزار ROUGE با خلاصه‌های انسانی مقایسه شده و سیستم ها بر اساس نتایج بدست آمده رتبه بندی می‌گردند. علاوه بر این، اعضای عضو تیم ارزیاب، سیستم ها را از دید قواعد زبان شناسی بررسی کرده و با در نظر گرفتن فاکتورهایی چون پیوستگی و خوانایی، دقت، ساختار گرامری و ... به آنها امتیاز داده‌اند. در جدول ۳.۲ اطلاعات کلی از این مجموعه داده‌ها آورده شده است.

جدول ۳.۲ اطلاعات مربوط به مجموعه داده های DUC [پور۹۰]

مجموعه داده	تعداد موضوعات	کاربرد	تعداد کل اسناد	اندازه خلاصه (تعداد کلمه)	تعداد سیستم های شرکت کننده
DUC 2001	۳۰	هردو	۳۰۳	۱۰۰ کلمه	۱۲
DUC 2002	۵۹	هردو	۶۰۰~	۱۰۰ کلمه	۱۲
DUC 2003	۶۰	هردو	۶۰۰	۷۵ بایت	۲۳
DUC 2004	۵۰	هردو	۵۰۰	۶۶۵ بایت	-
DUC 2005	۵۰	چند سنده	۱۳۰۰	۲۵۰ کلمه	۳۲
DUC 2006	۵۰	چند سنده	۱۲۲۵	۲۵۰ کلمه	۳۵
DUC 2007	۴۵	چند سنده	۱۱۲۵	۲۵۰ کلمه	۳۲

۶.۲ روش های ارزیابی خلاصه سازی

کار ارزیابی خودکار خلاصه نسبتاً کار دشواری است؛ چرا که یک خلاصه واحد و ایده‌آل برای یک یا چند سند وجود ندارد. از مقاله ها و مراجع مختلف اینگونه بر می‌آیدکه بین خلاصه‌های انسانی توافق کمی هم بر سر ارزیابی و هم بر سر ایجاد خلاصه وجود دارد [DAS07]. مساله دیگر در ارزیابی خلاصه‌ها مختلف بودن معیارهای ارزیابی مورد استفاده می‌باشد. در صورتیکه معیار استاندارد و

یکپارچه‌ای معرفی و مورد استفاده قرار گیرد، مساله ارزیابی و مقایسه سیستم‌های مختلف خلاصه-سازی به مراتب آسان‌تر خواهد بود. این مشکل در کاربردهای دیگر پردازش زبان طبیعی مثل ساخت درخت تجزیه و غیره ظهور و بروز پیدا نمی‌کند. از طرفی، ارزیابی دستی نیز بسیار هزینه بر می‌باشد. به شکلی که ارزیابی دستی خلاصه‌های موجود در کنفرانس DUC به بیش از ۳۰۰۰ ساعت تلاش انسانی نیاز دارد [LIN04].

۱.۶.۲ معیار ROUGE[۴۴]

لین در سال ۲۰۰۴ یک سری معیار برای ارزیابی خلاصه‌های اتوماتیک به نام ROUGE معرفی نمود [LIN04]. این معیارها به استانداردهایی در زمینه ارزیابی خلاصه‌ها بدل گشته‌اند. البته از این معیارها در دیگر کاربردهای بازیابی اطلاعات و پردازش زبان طبیعی نیز استفاده می‌شود. این معیارها با نام‌های *ROUGE-N*، *ROUGE-L*، *ROUGE-W* و *ROUGE-S* و برخی دیگر شناخته می‌شوند. هر کدام از این معیارها با توجه به فرمولی که برای آن معرفی شده است به مقایسه برخی جنبه‌ها بین خلاصه تولید شده توسط سیستم و خلاصه انسانی (مرجع) که از قبل موجود است می‌پردازد. در مجموعه داده DUC2007 از معیارهای ROUGE-2 و ROUGE-SU4 برای ارزیابی خلاصه‌ها استفاده می‌شود. فرض کنید $R = \{r_1, r_2, ..., r_m\}$ مجموعه‌ی خلاصه‌های مرجع و s خلاصه تولید شده توسط یک سیستم باشد. $\phi_n(d)$ برداری دودویی باشد که n-تایی های موجود در سند d را نشان می‌دهد. iامین عنصر در $\phi_n{}^i(d)$ یک است در صورتیکه i امین n-تایی در سند d وجود داشته باشد. در غیر این صورت i امین عنصر در $\phi_n{}^i(d)$ صفر است. معیار $ROUGE-N$ یک معیار آماری مبتنی بر فراخوانی n-تایی‌ها به صورت زیر تعریف می‌شود:

---

[44] Recall-Oriented Understudy for Gisting Evaluation

$$ROUGE-N\left(s\right)=\frac{\sum_{r\in R}\left\langle \phi_n\left(r\right),\phi_n\left(s\right)\right\rangle}{\sum_{r\in R}\left\langle \phi_n\left(r\right),\phi_n\left(r\right)\right\rangle} \qquad (۱۷.۲)$$

که در آن $\langle .,. \rangle$ به ضرب داخلی معمول بردارها اشاره دارد. ویژگی این معیار این است که برای خلاصه‌های مرجع چندتایی نیز می‌تواند مورد استفاده قرار گیرد، که در عمل بسیار مناسب‌تر از مقایسه تنها با یک مرجع عمل خواهد کرد. فرمول دیگری که برای $ROUGE-N$ معرفی شده است، خلاصه سیستمی را با شبیه‌ترین خلاصه مرجع مقایسه می‌نماید:

$$ROUGE-N_{multi}\left(s\right)=\max_{r\in R}\frac{\left\langle \phi_n\left(r\right),\phi_n\left(s\right)\right\rangle}{\left\langle \phi_n\left(r\right),\phi_n\left(r\right)\right\rangle} \qquad (۱۸.۲)$$

معیار دیگر معرفی شده توسط لین، $ROUGE-S$ می‌باشد. این معیار در واقع نسخه توسعه‌یافته $ROUGE-N$ با n = 2 که اصطلاحا دو تایی جهشی[۴۵] نامیده می‌شود، می‌باشد. فرض کنید $\psi_2\left(d\right)$ برداری دودویی اندیس گذاری شده توسط جفت‌های مرتب از کلمات باشد. iامین عنصر در $\psi_2^i\left(d\right)$ یک است در صورتیکه i امین جفت (دو-تایی) زیررشته‌ای از سند d باشد. در غیر این صورت i امین عنصر در $\psi_2^i\left(d\right)$ صفر است. معیار $ROUGE-S$ توسط فرمول زیر معرفی می‌گردد.

$$ROUGE-S\left(s\right)=\frac{\left(1+\beta^2\right)R_sP_s}{R_s+\beta^2P_s} \qquad (۱۹.۲)$$

که در آن

$$R_s\left(s\right)=\frac{\sum_{i=1}^{u}\left\langle \psi_2\left(r_i\right),\psi_2\left(s\right)\right\rangle}{\sum_{i=1}^{u}\left\langle \psi_2\left(r_i\right),\psi_2\left(r_i\right)\right\rangle} \qquad (۲۰.۲)$$

و

[45] Skip bigram

$$P_S(s) = \frac{\sum_{i=1}^{u} \langle \psi_2(r_i), \psi_2(s) \rangle}{\langle \psi_2(s), \psi_2(s) \rangle} \qquad (۲۱.۲)$$

نسخه‌های گوناگون ROUGE با محاسبه ضریب همبستگی بین امتیازات ROUGE و امتیازات قضاوت انسانی ارزیابی می‌شوند. از میان معیارهای *ROUGE-N*، *ROUGE-2* بهترین معیار شناخته شده است. *ROUGE-L*، *ROUGE-W* و *ROUGE-S* همگی بر روی مجموعه داده DUC2001 و DUC2002 که بیشتر به منظور خلاصه‌سازی تک سنده فراهم آمده‌اند بسیار خوب عمل کرده‌اند. اما همبستگی‌های به دست آمده برای خلاصه‌سازی چند سنده به خوبی خلاصه‌سازی تک سنده نبوده و در این زمینه هنوز جای کار بسیاری وجود دارد. با این حال در مجموعه داده DUC2007 که به منظور خلاصه‌سازی چندسنده فراهم آمده، سیستم‌ها بر اساس معیار *ROUGE-2* و *ROUGE-SU4* (نوعی از *ROUGE-S* ، بر مبنای دو-تایی های جدا شده توسط حداکثر چهار کلمه) ارزیابی می‌گردند. همانگونه که در بخش قبل نیز ذکر گردید، نتایج خلاصه‌سازی چندین سیستم قوی خلاصه‌سازی به عنوان مرجع و معیار سیستم‌های خلاصه سازی در این مجموعه موجود می‌باشد. در این پایان‌نامه نیز نتایج با این سیستم ها مورد ارزیابی قرار خواهد گرفت.

فصل سوم

روش پیشنهادی

۳ روش پیشنهادی

در این فصل مراحل مختلف روش پیشنهادی برای خلاصه‌سازی چکیده‌ای شرح داده شده است.

۱.۳ معماری روش پیشنهادی

شکل ۱.۳ معماری کلی روش پیشنهادی

معماری روش پیشنهادی در شکل نشان داده شده‌است. ابتدا بر روی جملات ورودی عملیات پیش‌پردازشی از جمله حذف ایست‌واژه‌ها، ریشه‌یابی، برچسب زنی بخش‌های گفتاری و برچسب گذاری نقش معنایی انجام می‌پذیرد. پس از آن جملات با استفاده از یک الگوریتم خلاصه‌سازی گزینشی مناسب [پور۹۰] جملات اسناد ورودی مرتب و امتیازدهی خواهند شد. مجموعه‌ای از بهترین این جملات انتخاب شده و فازهای بعد روی آنها انجام می‌پذیرد. پس از آن شباهت معنایی بین کلمات، نقش‌های معنایی، سطوح معنایی و جملات طبق روشی که در ادامه شرح داده خواهد شد، محاسبه می‌گردد. با استفاده از این معیار شباهت جملات دسته بندی شده و الگوریتم‌های ادغام جملات و فشرده سازی آن‌ها با انتخاب سطح معنایی مناسب انجام می‌پذیرد.

۲.۳ پیش پردازش

در روند هرگونه پردازش روی متن‌های زبان طبیعی انجام یک سری پیش‌پردازش امری اجتناب نشدنی است. علاوه بر آن دقت این پیش‌پردازش ها تاثیر بسزایی در فازهای بعدی نتایج اعمال الگوریتم‌ها دارد. هرچقدر که دقت پیش‌پردازش بیشتر باشد الگوریتم‌ها به نتایج واقعی خود نزدیک تر خواهند شد. در ادامه به پیش‌پردازش‌های مورد استفاده در این پایان‌نامه و ابزارهای به کار رفته اشاره می‌گردد.

۱.۲.۳ حذف ایست واژه ها

در این فاز کلمات کم اهمیت تر و یا ایست‌واژه ها که در بخش های قبل معرفی شدند حذف می-گردند. همانطور که ذکر شد بسیاری از ایست‌واژه ها فعل و اسم می‌باشند. هنگامی که با حجم انبوهی داده سروکار داریم و واحد کاری پردازش متن، اسناد بزرگی باشد، حذف این ایست‌واژه ها مطلوب می‌نماید. چنان که در ادامه توضیح داده خواهد شد، در این پایان نامه از فاز ۳ به بعد واحد کاری مورد

استفاده، جملات خواهند بود. به عنوان مثال، در یک مرحله نیازمند محاسبه شباهت معنـایی بـین دو جمله هستیم. در بسیاری مواقع امکان دارد فعل اصلی و هسته‌ای جمله یکی از همـین ایسـت‌واژه هـا باشد. به جمله زیر توجه کنید:

Original sentence: *preparations for the design of the Euro note have already begun.*
After stopword removing: *preparations design Euro note.*

همان گونه که مشاهده می‌شود، حذف تمام ایست‌واژه ها سبب از بین رفتن قسمت‌های مهمی از جمله می‌شود که بدون آنها معنی جمله متفاوت و ناقص خواهد بود. از این رو اقدام به تغییر لیست ایست‌واژه‌های منتشر شده گردید تا از ایجاد مشکل مذکور جلوگیری به عمل آید. در جدول ۱.۳ تعدادی از ایست‌واژه هایی که در فازهای سوم به بعد پروژه از لیست اولیه ایست‌واژه ها حذف گردیدند تا ایست‌واژه محسوب نشوند مشاهده می‌گردد.

جدول ۱.۳ نمونه ای از ایست واژه های محذوف از لیست ایست واژه ها

new	can	novel
concern	near	name
cause	except	consider
outside	believe	begin

۲.۲.۳ ریشه‌یابی

در این مرحله به منظور اولا یکسان سازی اشکال مختلف یک کلمه و یکپارچه سازی و ثانیا اعمـال پردازش های بعدی بایستی کلمات به ریشه خود بدل گردند. الگوریتم رایج مورد استفاده برای ریشـه-یابی الگوریتم پورتر [POR80] می‌باشد. اما از آنجا که خروجی ریشه یاب در فازهای بعد مورد استفاده های گوناگون از جمله اندازه‌گیری شباهت معنایی بر مبنای شبکه واژگان قرار می‌گیـرد، بایـد بررسـی

شود تا خروجی ریشه‌یاب، ورودی مناسبی برای آن فازها باشد. بدین منظور ریشه‌یاب پورتر را با ریشه‌یاب مبتنی بر شبکه واژگان مقایسه می‌نماییم. برای این مقایسه در جدول ۲.۳ چند مثال آورده شده است:

جدول ۲.۳ مقایسه ای موردی بین دو ریشه یاب

ریشه یاب	کلمه ۱	کلمه ۲	ریشه کلمه ۱	ریشه کلمه ۲	مشابهت معنایی با استفاده از روش لین
پورتر	countries	Iran	countri	Iran	0.0
ریشه یاب شبکه واژگان	countries	Iran	country	Iran	0.646

همان طور که مشاهده می‌شود نتیجه مشابهت معنایی با استفاده از ریشه‌یاب شبکه واژگان منطقی تر از نتیجه مشابهت معنایی با استفاده از ریشه‌یاب پورتر می‌باشد. دلیل این امر این است که کلمه "countri" که خروجی ریشه‌یاب پورتر می‌باشد، برای ماژول مشابهت معنایی ناشناخته می‌باشد، از این رو این ماژول، مشابهت را صفر برمی‌گرداند. این امر برای اکثر اسمها و فعلها در ریشه‌یاب پورتر اتفاق می‌افتد. از این رو ریشه‌یاب شبکه واژگان برای ادامه کار انتخاب می‌گردد.

۳.۲.۳ درخت تجزیه، برچسب‌گذاری بخشهای گفتاری و برچسب‌گذاری نقشهای معنایی

پیش‌پردازش‌های دیگر لازم در این پایان‌نامه توسط ابزارهای مربوطه، توسعه یافته توسط دانشگاه ایلینویز[۴۶] انجام گرفته است. برای جمله‌ای که در ادامه آمده است نمونه هایی از هرکدام در ادامه آورده شده است.

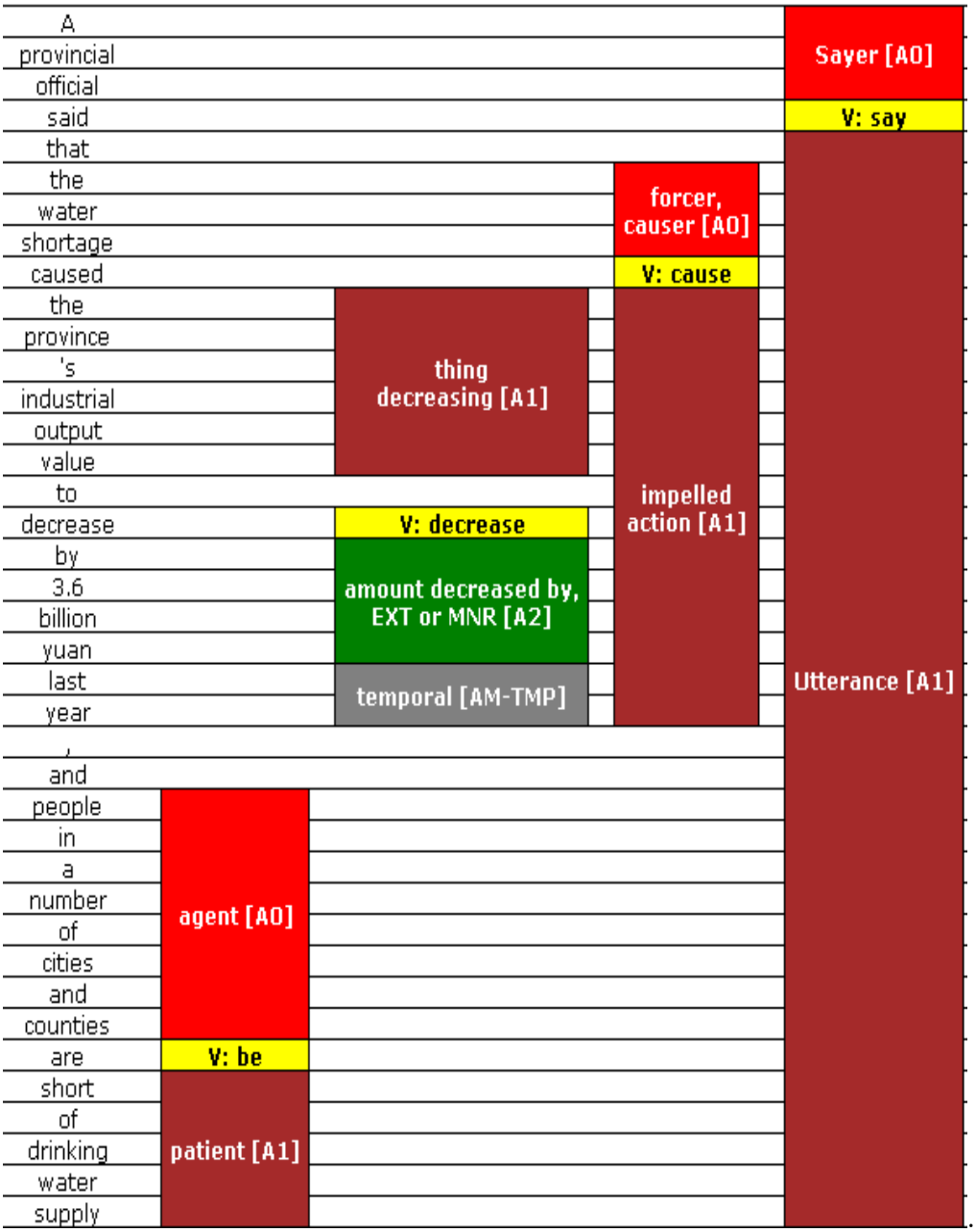
[46] http://cogcomp.cs.illinois.edu/

A provincial official said that the water shortage caused the province's industrial output value to decrease by 3.6 billion yuan last year, and people in a number of cities and counties are short of drinking water supply.

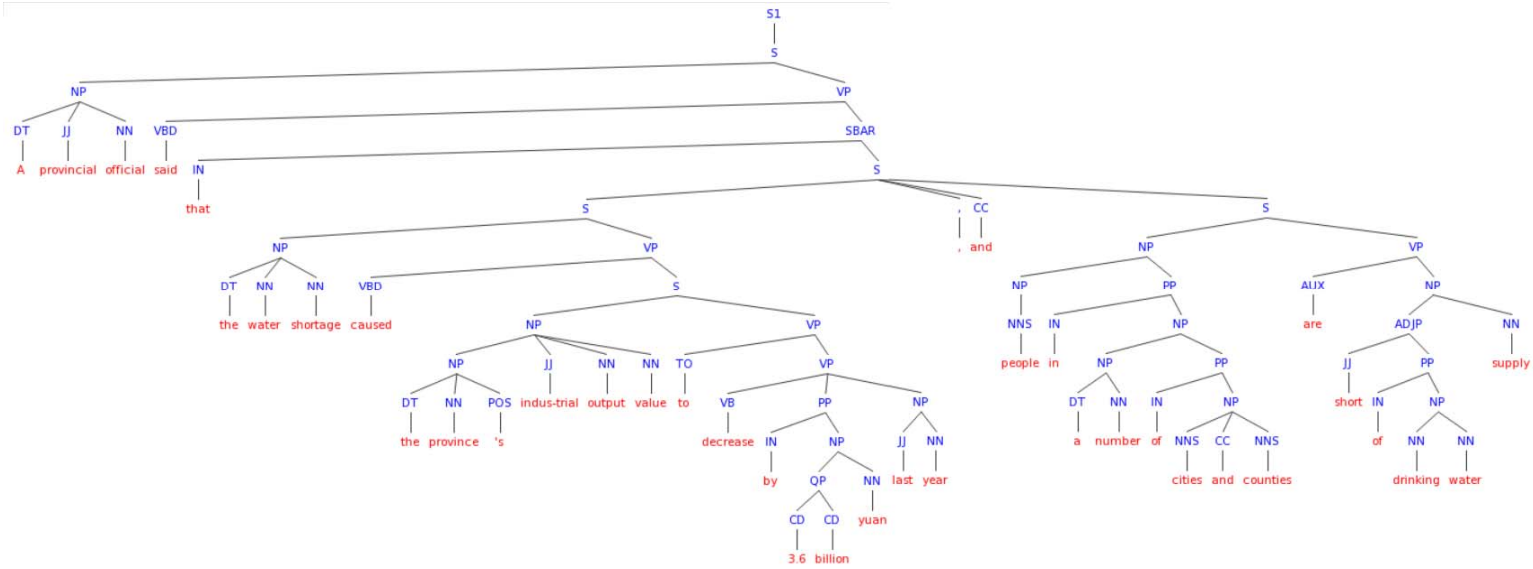
در ادامه خروجی هر کدام از ابزارهای ذکر شده نمایش داده شده است.

```
DT/ A JJ/ provincial NN/ official VBD/ said IN/ that DT/ the NN/ water NN/ shortage VBD/
caused DT/ the NN/ province POS/ 's JJ/ indus-trial NN/ output NN/ value TO/ to VB/
decrease IN/ by CD/ 3.6 CD/ billion NN/ yuan JJ/ last NN/ year ,/ , CC/ and NNS/ people
IN/ in DT/ a NN/ number IN/ of NNS/ cities CC/ and NNS/ counties VBP/ are JJ/ short IN/
of NN/ drinking NN/ water NN/ supply ./ .
```

شکل ۲.۳ خروجی برچسب زن بخش های گفتاری



شکل ۳.۳ خروجی برچسب زن نقش های معنایی



شکل ۴.۳ خروجی تجزیه گر

۳.۳ اعمال الگوریتم خلاصه‌سازی گزینشی

روش پیشنهادی در واقع ترکیبی از روش‌های خلاصه‌سازی گزینشی و چکیده‌ای می‌باشد. به این ترتیب که ابتدا جملات مهم‌تر توسط یک روش خلاصه‌سازی گزینشی [پور۹۰] انتخاب شده و سپس مراحل ایجاد خلاصه چکیده‌ای بر روی جملات منتخب انجام می‌گیرد. در الگوریتم خلاصه‌سازی گزینشی مورد استفاده پس از پردازش‌های اولیه‌ای که برروی داده‌ها انجام می‌گیرد، ماتریس کلمه-سند ساخته می‌شود. سپس از فرم تغییر یافته‌ای از روش LSA[۴۷] روی ماتریس کلمه-سند برای استخراج زمینه استفاده می‌شود. به این شکل که به جای استفاده از معیار وزن دهی در واحد جمله، از معیار وزن دهی TFIDF[۴۸] برای کلمات در واحد سند استفاده می‌شود تا اثر تمامی اسناد لحاظ شود. در گام بعدی SVD[۴۹] بر روی این ماتریس اعمال شده و ماتریس بردارهای یکه استخراج می‌شود

[47] Latent Semantic Analysis

[48] Term Frequency Inverse Document Frequency

[49] Singular Value Decomposition

(شکل ۵.۳). در مرحله بعد نیز بجای ماتریس V از ماتریس ستونی U استفاده می‌شود. 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 می‌باشد.

$$\underset{term-document}{\begin{bmatrix} w_{1,1} & \dots & w_{1,n} \\ w_{2,1} & \dots & w_{2,n} \\ & & \\ w_{m,1} & \dots & w_{m,n} \end{bmatrix}} = \underset{term-concept\,(u)}{\begin{bmatrix} u_{1,1} & \dots & u_{1,m} \\ u_{2,1} & \dots & u_{2,m} \\ & & \\ u_{m,1} & \dots & u_{m,m} \end{bmatrix}} \times \begin{bmatrix} \sigma_{1,1}, & 0 & \dots, & 0 \\ 0 & ,\sigma_{2,n}, & 0,\dots, & 0 \\ & & & \\ 0 & , & \dots & , \quad 0 \\ 0 & \dots, & 0 & ,\sigma_{n,n} \end{bmatrix} \times \underset{concept-document\,(v)}{\begin{bmatrix} v_{1,1} & \dots, & v_{1,n} \\ v_{2,1} & \dots, & v_{2,n} \\ & & \\ v_{n,1} & \dots, & v_{n,n} \end{bmatrix}}$$

شکل ۵.۳ خروجی SVD برای ماتریس کلمه-سند

در مرحله بعد باید مفاهیم به موضوعات مربوطه منتسب شوند. 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 رابطه (۱.۳) محاسبه می‌شود.

$$\cos(C_i, D_j) = \frac{\sum_k c_{ik} d_{jk}}{\sqrt{\sum_k c_{ik}^2} \times \sqrt{\sum_k d_{jk}^2}} \qquad (۱.۳)$$

بر همین اساس، نزدیکترین بردار به بردار کلمات هر موضوع به عنوان نماینده آن موضوع انتخاب می‌گردد.

در مرحله بعد جملات بر اساس شباهتشان با زمینه مرتب سازی می‌گردند. ابتدا بردار کلمات هر جمله ساخته می‌شود. سپس مجددا از فاصله کسینوسی بین بردار کلمات جملات، $S_j = [s_{j1}, s_{j2}, \dots, s_{jm}]$ و بردار کلمات مفاهیم، $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ برای مرتب کردن جملات استفاده می‌گردد.

$$\cos\left(C_i, S_j\right) = \frac{\sum_k c_{ik} s_{jk}}{\sqrt{\sum_k c_{ik}^2} \times \sqrt{\sum_k s_{jk}^2}} \qquad (۲.۳)$$

پس از محاسبه فاصله کسینوسی بردارهای تمامی جملات با بردار زمینه اصلی، جملات براساس مقدار شباهت، امتیازدهی و به صورت نزولی مرتب می‌شوند. جهت اطلاعات بیشتر از نحوه عملکرد این روش به [پور۹۰] رجوع شود.

۴.۳ محاسبه شباهتهای معنایی

در این مرحله شباهت معنایی بین سطوح معنایی موجود در جملات محاسبه می‌گردد. ابتدا نیازمند معیاری برای اندازه گیری شباهت بین کلمات هستیم. روش‌های متعددی از شبکه واژگان به این منظور استفاده نموده‌اند که در بخش‌های ۲.۳.۲ تا ۴.۳.۲ به تفصیل بحث شد. همان گونه که در آنجا شرح داده شد لین [LIN04] با ارائه فرضیه مشابهت خود و به کار بردن این فرضیه برای دو مفهومی که در شبکه واژگان وجود دارند، فرمول مشابهتی را برای اندازه‌گیری مشابهت بین کلمات ارائه نمود. طبق ارزیابی‌های انجام شده، این معیار نتایج خوبی را در مقایسه با سایر معیارها به دست آورده است [BUD06]. از این رو در این پایان نامه از آن به منظور محاسبه شباهت معنایی بین کلمات استفاده شده است.

مساله دیگری که در شباهت کلمات در نظر گرفته شده است، مساله جایگاه کلمه در عبارت فعلی می‌باشد. با بررسی جملات مختلف این نتیجه حاصل شد که غالبا هرچه کلمه در موقعیت های پایانی عبارت اسمی قرار گرفته باشد از اهمیت بیشتری در عبارت اسمی برخوردار می‌باشد. برای روشن‌تر شدن موضوع چند مثال ذکر گردیده‌اند.

European Central Bank President Wim Duisenberg *said Thursday that …*

The European single currency euro *will go ahead on schedule on January 1, 1999.*

Because of ***a serious shortage of water supply*** *with no sign of immediate improvement …*

همان طور که در مثال‌ها دیده می‌شود غالب کلماتی که در موقعیتهای پایانی گروه‌های اسمی قرار گرفته‌اند دارای اهمیت بیشتری می‌باشند. برای لحاظ نمودن این میزان اهمیت، برای هر کلمه با توجه به موقعیتی که در جمله دارد، وزنی اختصاص می‌دهیم.

$$weight(w) = \frac{position_w}{N} \qquad (۳.۳)$$

که در آن $position_w$ موقعیت کلمه w در کوچکترین گروه اسمی خود از سمت چپ و N تعداد کل کلمات موجود در گروه اسمی مذکور می‌باشد، و از آنجا خواهیم داشت:

$$WordSim(w_1, w_2) = Sim_L(w_1, w_2) * \left(\frac{weight(w_1) + weight(w_2)}{2} \right) \qquad (۴.۳)$$

لازم به ذکر است که برای جداسازی گروه‌های اسمی استفاده از ابزار تجزیه‌گر گزینه بسیار مناسبی می‌باشد. چرا که این ابزار تمامی گروه‌های اسمی و فعلی را به شکل درخت‌گونه جداسازی می‌نماید. تجزیه‌گر چارنیاک نیز که تجزیه‌گر مورد استفاده می‌باشد دارای دقت بسیار بالایی است.

پس از آن شباهت بین نقش‌های معنایی محاسبه می‌گردد. فرض می‌کنیم $CommonRoles\left(L_{\boldsymbol{i}}{}^{\boldsymbol{A}}, L_{\boldsymbol{j}}{}^{\boldsymbol{B}}\right) = \left\{r_1, r_2, \ldots, r_{\boldsymbol{n_{cm}}}\right\}$ اشتراک نقش‌های معنایی بین دو جمله A و B در سطوح iام و jام باشد. و $words_r\left(L_i^A\right) = \left\{w_1, w_2, \ldots, w_{m_A}\right\}$ مجموعه کلمات جمله A (غیر از

ایست‌واژه‌ها) در سطح معنایی iام ، در نقش معنایی r و پس از ریشه‌یابی می‌باشند. شباهت دو نقش معنایی با برچسب‌های یکسان در رابطه (۵.۳) نشان داده شده است:

$$RoleSim_r\left(words_r\left(L_i^A\right),words_r\left(L_j^B\right)\right)= \frac{\sum_{k=1}^{m_A}\left(\max_{w_B\in words_r\left(L_j^B\right),w_k\in words_r\left(L_i^A\right)} wordSim\left(w_k,w_B\right)\right)}{m_A+m_B}+ \frac{\sum_{k=1}^{m_B}\left(\max_{w_A\in words_r\left(L_i^A\right),w_k\in words_r\left(L_j^B\right)} wordSim\left(w_A,w_k\right)\right)}{m_A+m_B} \qquad (۵.۳)$$

با در نظرگرفتن قاعده کلی مذکور در بخش ۱.۳.۲ که هنگام محاسبه شباهت معنایی بین دو شیء ذهن به سمت معانی از آنها متوجه می‌شود که در نزدیکترین حالت به یکدیگر قرار داشته‌باشند؛ در این مرحله نیز برای هر کلمه‌ی موجود در مجموعه لغات A بیشترین شباهت کسب شده در مقایسه با تمامی کلمات مجموعه لغات B محاسبه می‌گردد. این مقدار برای تمامی کلمات موجود در مجموعه لغات A و همچنین کلمات موجود در مجموعه لغات B محاسبه شده و سپس میانگین این مقادیر محاسبه می‌گردد. پس از آن شباهت سطوح معنایی باید محاسبه گردد. بدین منظور از رابطه (۶.۳) استفاده می‌شود:

$$LevelSim\left(L_i^A,L_j^B\right)= \frac{\sum_{r=1}^{n_{cm}} RoleSim_r\left(words_r\left(L_i^A\right),words_r\left(L_j^B\right)\right)}{n_{cm}} \qquad (۶.۳)$$

در واقع بین هر جفت نقش معنایی یکسان در سطوح i و j مقدار قبل محاسبه شده و در نهایت بین آنها میانگین را محاسبه می‌نماییم. نهایتاً شباهت دو جمله با توجه به همان قاعده کلی عبارت است از شباهت شبیه‌ترین جفت سطح معنایی موجود در دو جمله. رابطه (۷.۳) بیانگر این مطلب است.

$$SenSim(A,B) = \max_{i \in levelsOfA,\, j \in levelsOfB} LevelSim\left(L_i^A, L_j^B\right) \qquad (۷.۳)$$

حال با داشتن شباهت بین جملات می‌توان ماتریس شباهت جملات را تشکیل داد و با اعمال یک الگوریتم خوشه‌بندی جملات مشابه را در یک دسته قرار داد. اما پیش از آن از شباهت های به دست آمده برای سطوح معنایی الگوریتم فشرده‌سازی جملات را به کار می‌بریم. این الگوریتم در بخش بعد توضیح داده شده‌است.

۵.۳ فشرده سازی جملات

فشرده‌سازی در روش پیشنهادی در دو فاز انجام می‌پذیرد. ابتدا با بررسی های انجام شده تعدادی کلمات را قابل حذف تشخیص داده و با پیدا کردن الگوی آنها اقدام به حذف آنها می‌نماییم. سپس الگوریتمی برای فشرده‌سازی مبتنی بر مشابهت جملات ارائه می‌نماییم.

۱.۵.۳ حذف کلمات اضافه

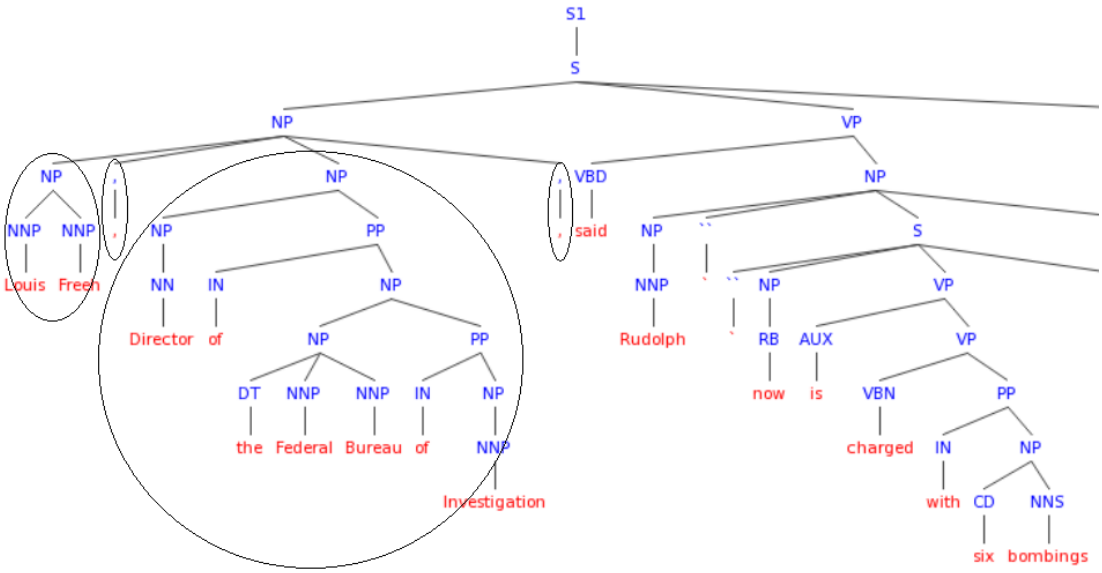
در این مرحله با بررسی های انجام شده روی داده‌های مختلف چند الگوی قابل حذف و دارای فراوانی قابل توجه در اسناد به دست آمد. دو الگوی اصلی تشخیص داده شده عبارتند از بدل‌ها و عبارات مخفف، که در ادامه به توضیح آنها پرداخته شده‌است.

- **بدل‌ها**

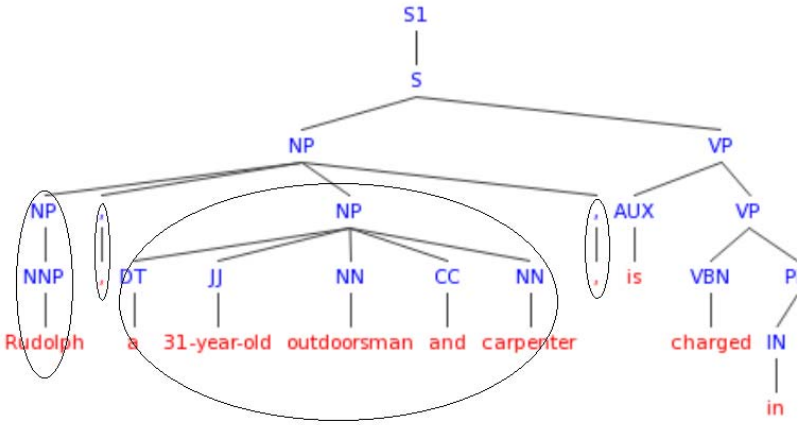
بدل‌ها و یا عبارات تمیز، عباراتی هستند که در ادامه‌ی اغلب یک اسم و یا گروه اسمی آمده و توضیح و مکمل برای آن گروه اسمی ایفا می‌نمایند. در ادامه چند مثال از وجود بدل در جمله آمده-است. بدل‌ها با قسمت‌های رنگی مشخص شده اند.

- *Louis Freeh, Director of the Federal Bureau of Investigation, said Rudolph "now is charged with six bombings".*
- *Rudolph, a 31-year-old outdoorsman and carpenter, is charged in …*
- *The mass printing of the banknotes of the single European currency, the euro, would be started at the beginning of 1999.*
- *Water shortage in Pokhara, a famous city for tourism in Nepal, is that surgical operations on patients …*

عبارات توضیحی مذکور به منظور رسیدن به خلاصه‌ای هر چه کوتاهتر و در عین حال حاوی اطلاعات مفید باید به نحوی مدیریت شوند. پس از بررسی های انجام شده، ابزار مناسب برای تشخیص عبارات توضیحی یا بدل‌ها، درخت تجزیه تشخیص داده شد؛ چرا که ساختار درخت تجزیه برای اینگونه عبارات، ساختار مشخصی است. در شکلهای (۶.۳) و (۷.۳) این ساختار نمایش داده شده است.



شکل ۶.۳ ساختار درخت تجزیه برای نمونه بدل ۱

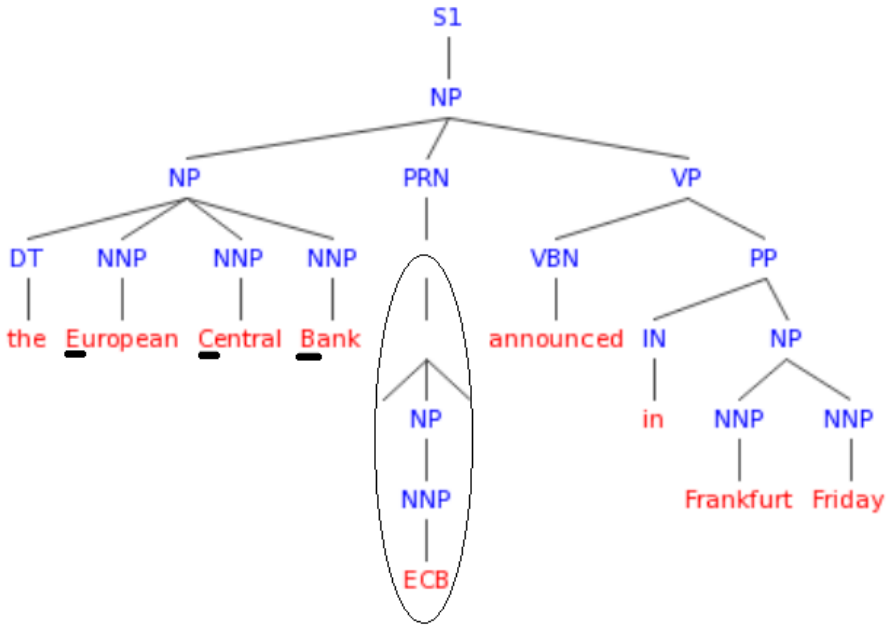


شکل ۷.۳ساختار درخت تجزیه برای نمونه بدل ۲

با توجه به اینکه در خلاصه‌سازی کوتاه بودن در حین حاوی اطلاعات بودن یک مزیت محسوب می‌شود و از آنجا که دو عبارت استخراج شده به وسیله بدل‌ها به نوعی معادل هم هستند، عبارت کوتاهتر را نگه داشته و عبارت طولانی‌تر حذف می‌گردد.

- **عبارات مخفف**

عمل دیگری که برای کوتاه سازی جملات می‌توان انجام داد، یافتن عبارات مخفف و نگه داشتن سرنام و حذف کلمات مربوطه می‌باشد. الگوی تشخیصی این عبارات نیز با پیمایش درخت تجزیه قابل تشخیص می‌باشد. در شکل (۸.۳) نمونه‌ای از جملات حاوی این عبارات و ساختار درخت مربوطه نمایش داده شده است.



شکل ۸.۳ ساختار درخت تجزیه برای یک نمونه عبارت مخفف

۲.۵.۳ فشرده سازی مبتنی بر مشابهت جملات

در این بخش به توضیح الگوریتمی می‌پردازیم که جملات را بر مبنای شباهتی که با سایر جملات مجاور دارند فشرده می‌نماید. به این ترتیب که برای جمله A سطحی از آن انتخاب می‌گردد که بیشترین شباهت را با سطوح موجود در جمله B کسب نماید. البته با توجه به خصوصیت خلاصه‌سازی چندسنده و با مطالعات انجام شده، در انتخاب سطح معنایی یک قانون ساده هم اعمال می‌شود و آن اینکه فعل اصلی سطح انتخاب شده نباید فعل نقل قولی باشد.

$$\begin{aligned} &Compressed_k\left(A\right)=GenCompressed\left(A,l\right) \\ &\text{where } LevelSim\left(L_l^A,L_j^{B_k}\right)\geq all\ LevelSim\left(L_i^A,L_j^{B_k}\right) \end{aligned} \qquad (۸.۳)$$

در رابطه (۸.۳) در $GenCompressed$ سطح انتخاب شده l از جمله A به همراه فعل اصلی موجود در آن سطح و شناسه‌های اصلی آن فعل، جمله فشرده شده نهایی را تولید خواهند کرد. در نتیجه به

ازای هر جمله‌ای که در مقابل جمله A برای فشرده‌سازی قرار می‌گیرد، یک فرم فشرده شده برای آن خواهیم داشت که بنا به نیاز و مقتضیات هر کدام از آنها را می‌توان انتخاب نمود.

ایده این روش از مطالعات و بررسی های انجام شده در زمینه خلاصه‌سازی چندسنده به دست آمده است. از آنجا که در خلاصه‌سازی چندسنده تعداد زیادی سند در مورد یک عنوان وجود دارند و هدف به دست آوردن مهم‌ترین و مرتبط ترین جملات با عنوان مورد بحث می‌باشد، استفاده از مشابهت جملات با سایر جملات مفید به نظر می‌رسد. علاوه بر این، روش مذکور یک روش غیرنظارتی برای فشرده‌سازی جملات محسوب می‌گردد و این یک مزیت نسبت به روشهای نظارتی فشرده‌سازی جملات -که غالب روش‌های فشرده‌سازی نیز از این دست می‌باشند- می‌باشد. چرا که روشهای نظارتی نیازمند صرف زمان و هزینه بسیار برای تولید پیکره‌ای مناسب و تا حد نیاز عظیم برای آموزش می‌باشند.

مزیت دیگری که این روش دارد تکیه به سطوح معنایی می‌باشد، چرا که هر سطح معنایی در جملات یک فعل به عنوان هسته در خود دارد و این فعل به همراه شناسه‌هایش، خود اغلب بیانگر جمله‌ای با گرامر درست و معنی کاملی می‌باشد. از این رو بدون نیاز به اعمال قوانین و قیود اضافه به جمله‌ای با گرامری قابل قبول دست یافته‌ایم. نتایج ارزیابی نشان دهنده بهبود نتایج این روش نسبت به روش‌های پیشین می‌باشد. در جدول ۳.۳ نمونه‌ای از جملات فشرده شده آورده شده‌است.

جدول ۳.۳ **نمونه‌ای از جملات فشرده شده**

Despite skepticism about the actual realization of a single European currency as scheduled on January 1, 1999, preparations for the design of the Euro note have already begun.	جمله ۱
preparations for the design of the Euro note begun.	فشرده شده جمله ۱
Thailand is considering using the European single currency, the euro, in the country's foreign reserves, the Nation reported Tuesday.	جمله ۲
Thailand considering using the euro.	فشرده شده جمله ۲
The European Commission (EC) spokesman today formally denied newspaper reports that the EC is "secretly" drawing up plans to delay the launch of the single currency, the Euro.	جمله ۳
the EC delay the launch of the Euro.	فشرده شده جمله ۳
Di Bisceglie said the antiviral drug interferon had reduced symptoms in up to 20 percent of patients with hepatitis C who were treated for six months.	جمله ۴
interferon reduced symptoms.	فشرده شده جمله ۴

۶.۳ حذف یا ادغام جملات

حال نوبت به حذف و یا ادغام جملات مشابه می‌رسد. همان طور که اشاره شد، با اعمال الگوریتم خوشه بندی جملات مشابه در یک دسته قرار می‌گیرند. با توجه به شرایط مساله و ابزار مورد استفاده برای ارزیابی (در نهایت ۲۵۰ کلمه به عنوان خلاصه به ارزیاب داده خواهد شد) تعداد ۳۰ جمله از جملات اولیه انتخاب می‌گردند و ماتریس شباهت با استفاده از الگوریتم شباهت ارائه شده برای این جملات ساخته می‌شود. با شرایط مذکور در این پایان‌نامه و بررسی‌های انجام شده، الگوریتم خوشه-بندی طیفی[۵۰] مورد استفاده قرار گرفت.

[50] Spectral Clustering

۱.۶.۳ مروری گذرا بر خوشه‌بندی طیفی

از آنجا که موضوع خوشه‌بندی ارتباط مستقیمی با موضوع این پایان‌نامه ندارد، در این بخش به شکل گذرا مروری بر الگوریتم خوشه‌بندی مورد استفاده خواهیم داشت. جهت اطلاعات بیشتر به مراجع ذکر شده رجوع شود.

خوشه‌بندی طیفی با مدل کردن میزان شباهت داده‌ها و استفاده از بردارهای ویژه انواع ماتریس-های لاپلاسین، سعی می‌نماید، از دانشی که در مجموعه داده وجود دارد برای افراز آن استفاده کند. الگوریتم‌های زیادی برای خوشه‌بندی طیفی ارائه گردیده‌اند. SLH [SCO90]، PF [PER98]، برش نرمال [SHI00] و NJW [NG02] از جمله این روش‌ها هستند. در این پایان نامه از الگوریتم NJW استفاده شده‌است. ایده این الگوریتم استفاده از اولین K بردار ویژه متناظر با بزرگترین K مقدار ویژه ماتریس لاپلاسین است. مراحل این الگوریتم به این صورت می‌باشد که پس از ساخت ماتریس شباهت W، ماتریس درجه D و ماتریس لاپلاسین L توسط رابطه (۹.۳) محاسبه می‌گردند. در این رابطه، D ماتریس درجه گراف است و به صورت یک ماتریس قطری تعریف می‌شود که عنصر $d_i = D_{ii} = \sum_{j=1}^{N} w_{ij}$ در آن درجه نود iام در گراف را نشان می‌دهد.

$$L_N = D^{-1/2} W D^{-1/2} \qquad (۹.۳)$$

سپس اولین K بردار ویژه $\mathbf{v}^1, \mathbf{v}^2, ..., \mathbf{v}^k$ متناظر با اولین K بزرگ‌ترین مقدار ویژه ماتریس L_N به دست می‌آید و ماتریس ستونی $\mathbf{V}=[\mathbf{v}^1,\mathbf{v}^2,...,\mathbf{v}^k] \in \mathbf{R}^{n \times k}$ تشکیل می‌گردد. پس از آن مجددا V نرمال سازی شده و Y به طوریکه همه سطرهای آن طول واحد داشته باشند از طریق رابطه (۱۰.۳) تشکیل می-شود. در نهایت مجموعه داده بازنمایی شده Y با استفاده از K-Means [SCH96] خوشه‌بندی می‌گردد [اشک ۹۰].

$$\mathbf{Y}_{ij} = \mathbf{V}_{ij} / \sqrt{\sum_{j} \mathbf{V}_{ij}^{2}} \qquad (۱۰.۳)$$

۲.۶.۳ الگوریتم حذف و ادغام

به منظور انجام فاز آخر، بین دو جمله‌ای که درون یک خوشه قرار دارند، شبیه‌ترین سطح معنایی انتخاب و بررسی می‌گردد. در صورتیکه فعل و تمام شناسه‌های اصلی آن به جز یکی از شناسه‌ها شباهت زیادی داشته باشند، نتیجه می‌شود که می‌توان دو جمله را در نقش معنایی ناهمسان با یکدیگر ترکیب نمود. بدین منظور یکی از جمله‌ها را به عنوان پایه در نظر گرفته و ترکیب دو نقش ناهمسان را با کلمه "and" به همراه سایر نقش‌های اصلی جمله به عنوان جمله جدید به خروجی می-بریم. در صورتیکه شرط مذکور برقرار نبود یکی از جملات حذف می‌گردد. برای انتخاب جمله، از میزان امتیازی که جمله طی فرایند خلاصه‌سازی گزینشی به دست آورده استفاده می‌گردد و جمله‌ای حذف می‌گردد که امتیاز کمتری کسب کرده‌باشد. این الگوریتم برای جفت جمله‌های موجود در هر دسته از بالا به پایین اجرا شده تا در نهایت به یک جمله برسیم. در ادامه روند کار به شکل سازماندهی شده بیان گردیده‌است. جمله‌های درون دسته را بر اساس اهمیتشان مرتب می‌کنیم و تا زمانیکه خوشه تک عضوی نشده این عملیات را ادامه می‌دهیم:

- دو جمله اول خوشه را بردار
- شبیه ترین جفت سطح معنایی آن دو را بررسی کن
- در صورتیکه تمام شناسه های اصلی فعل به جز یکی به هم شبیه بودند، جمله اول را همراه با شناسه های اصلی سطح انتخاب شده به همراه شناسه متفاوت از جمله اول در قالب یک جمله بیان کن و جمله حاصل را به جای این دو جمله جایگزین کن

- در غیر این صورت شناسه های اصلی سطح انتخاب شده از جمله اول را در قالب یک جمله بیان کن و به جای دو جمله اول جایگزین نما

در جدول ۴.۳ نمونه ای از جملات ادغام شده آورده شده‌است:

جدول ۴.۳ نمونه‌ای از جملات ادغام شده ۱

جمله ۱	Italy and France have adopted the euro, as the European Union’s new single currency is known.
جمله ۲	A number of countries are already planning to hold the euro as part of their foreign currency reserves, the magazine quoted European Central Bank chief Wim Duisenberg as saying.
جمله ادغام شده	Italy and France and A number of countries adopted the euro

جدول ۵.۳ نمونه‌ای از جملات ادغام شده ۲

جمله ۱	The Pentagon conducted two tests Thursday of important elements of the proposed national missile defense system in preparation for another attempt to shoot down a target in space.
جمله ۲	It was designed to test elements of the national missile defense system -- such as an "in-flight interceptor communication system" used to send information from the ground radar to the interceptor missile -- that will be used in the next attempt to shoot down a mock warhead in space .
جمله ادغام شده	shoot a target in space and a mock warhead in space.

جدول ۶.۳ نمونه‌ای از جملات ادغام شده ۳

جمله ۱	The European single currency euro will go ahead on schedule on January 1, 1999 with a broad membership, according to a survey of some prominent British economists.
جمله ۲	as national currency boundaries vanish most European stocks and bonds are denominated in euros
جمله ادغام شده	The European single currency euro and national currency boundaries go ahead on schedule on January 1 , 1999

فصل چهارم

پیاده سازی و ارزیابی

۴ پیاده سازی و ارزیابی

در این فصل به شرح نتایج ازریابی های مختلف انجام شده در این پایان‌نامه می‌پردازیم.

۱.۴ ارزیابی روش مشابهت معنایی جملات

به منظور ارزیابی روش ارائه شده مشابهت معنایی جملات از مجموعه داده مایکروسافت[۵۱] استفاده نمودیم. این مجموعه داده حاوی ۴۰۷۷ جفت جمله که جهت آموزش و ۱۷۲۶ جفت جمله که جهت تست آماده شده‌اند می‌باشد. هر کدام از این جفت جمله‌ها دارای یک فیلد به نام کیفیت (quality) می‌باشد که شباهت جفت جمله مذکور را مشخص می‌نماید. در صورتیکه این فیلد ۱ باشد به معنی شباهت جفت جمله و در صورتیکه ۰ باشد به معنی عدم شباهت جفت جمله خواهد بود. ما به شکل تصادفی تعداد ۵۰۰ جمله از این مجموعه داده را انتخاب نمودیم. همانطور که در شکل (۳) نشان داده شده است، روش پیشنهادی پیشرفتی قابل توجه نسبت به روش‌های ارائه شده لی [LEE11] و ایزلام [ISL08] نشان می‌دهد.

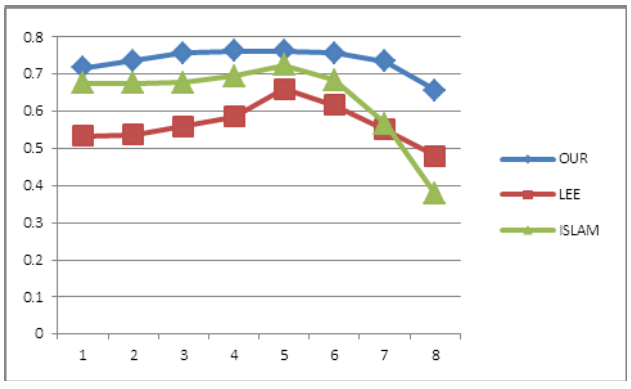
از آنجا که روش‌های پیشنهادی از جمله روش پیشنهادی این مقاله برای اندازه‌گیری شباهت جملات اغلب یک عدد بین صفر تا یک به عنوان میزان شباهت برمی‌گردانند، بایستی یک میزان آستانه در نظر گرفت تا اگر شباهت برگردانده شده کمتر از آن بود عدم شباهت نتیجه شود و اگر این شباهت بیشتر از آن بود شباهت جملات نتیجه گردد. مقدار این حد آستانه بستگی به مقیاس کاری الگوریتم و روش مورد استفاده برای اندازه‌گیری شباهت معنایی بین کلمات دارد. از این رو مقدار کمّی آن مهم نیست و مقدار نسبی آن اهمیت دارد. حد آستانه‌ای که بیشترین دقت را توسط الگوریتم

[51] http://research.microsoft.com

پیشنهادی ما به دست آورده‌است ۰.۰۹ می‌باشد. در دو روش دیگر این حد آستانه ۰.۶ می‌باشد. میزان صحت الگوریتم‌های مذکور و روش پیشنهادی بر اساس مقدار آستانه‌ای که بیشترین دقت را به دست آورده به همراه چند مقدار آستانه در همسایگی آن در جدول (۱.۴) و شکل (۱.۴) آورده شده‌است. همان طور که از داده ها و شکل مشهود است، روش پیشنهادی بهبودی قابل توجه در نتایج حاصل نموده است.

جدول ۱.۴داده‌های حاصل از صحت روش‌های پیشین و روش پیشنهادی برای شباهت جملات

thresholds	Our method accuracy	Lee's method accuracy	Islam's method accuracy
thre.1	0.7166	0.5330	0.6754
thre.2	0.7365	0.5374	0.6759
thre.3	0.7566	0.5594	0.6774
thre.4	0.7619	0.5859	0.6953
thre.5(max)	**0.7619**	**0.6608**	**0.7242**
thre.6	0.7566	0.6167	0.6845
thre.7	0.7355	0.5506	0.5667
thre.8	0.6560	0.4802	0.3778



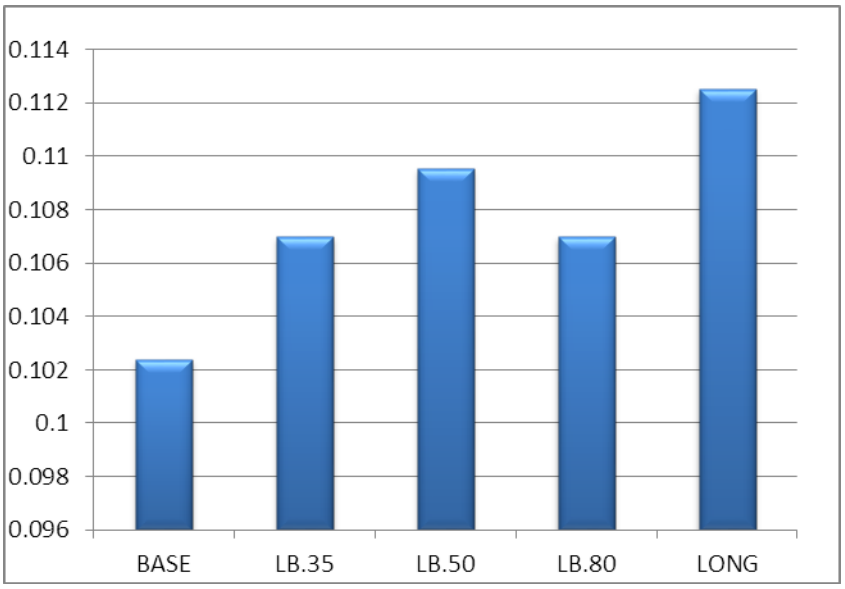
شکل ۱.۴ مقایسه نتایج حاصل از روش‌های پیشین و روش پیشنهادی شباهت جملات

۲.۴ ارزیابی روش فشرده سازی جملات

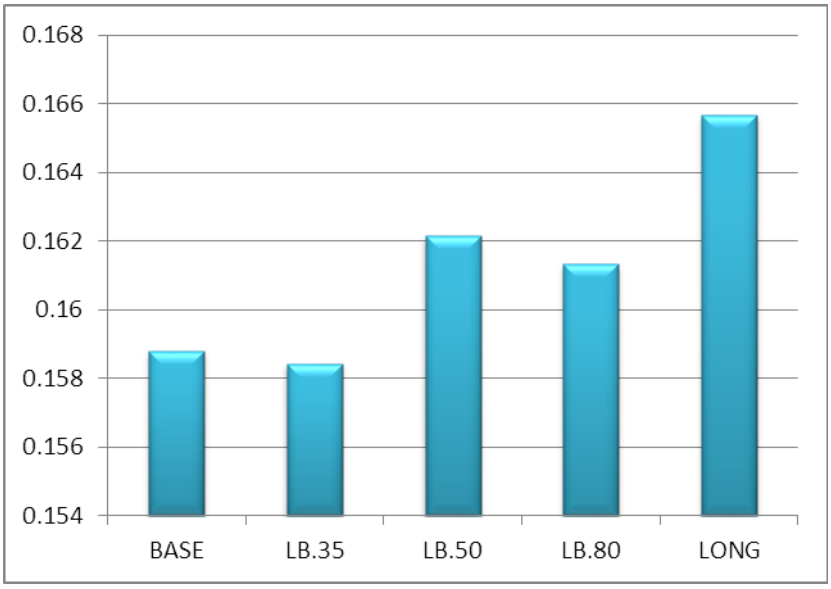
همان‌گونه که در فصل ۳ توضیح داده شد، ما در مرحله فشرده سازی جملاتی با طول‌های مختلف به دست آوردیم. برای انتخاب بهترین حالت، ۵ حالت مختلف را برای طول جمله انتخابی در نظر گرفتیم:

- انتخاب جمله اولیه (بدون فشرده سازی) (base)
- انتخاب کوتاهترین جمله با حد پایین ۳۵ کلمه به عنوان جمله فشرده شده (LB.35)
- انتخاب کوتاهترین جمله با حد پایین ۵۰ کلمه به عنوان جمله فشرده شده (LB.50)
- انتخاب کوتاهترین جمله با حد پایین ۸۰ کلمه به عنوان جمله فشرده شده (LB.80)
- انتخاب بلندترین جمله به عنوان جمله فشرده شده (long)

در شکل (۲.۴) و (۳.۴) نتایج حاصل از انتخاب هر کدام از این ۵ حالت را بر روی معیار ROUGE-2 و ROUGE-SU4 نشان داده شده است:

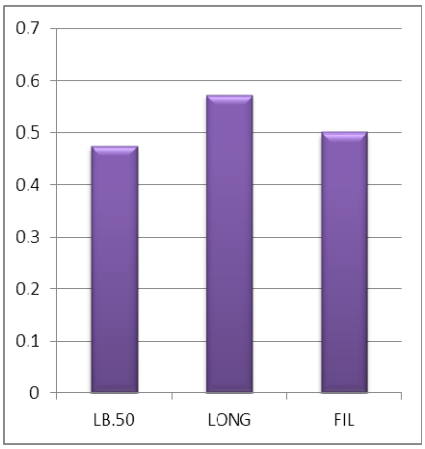


شکل ۲.۴ مقایسه معیار ROUGE-2 جملات فشرده شده در ۵ حالت مختلف



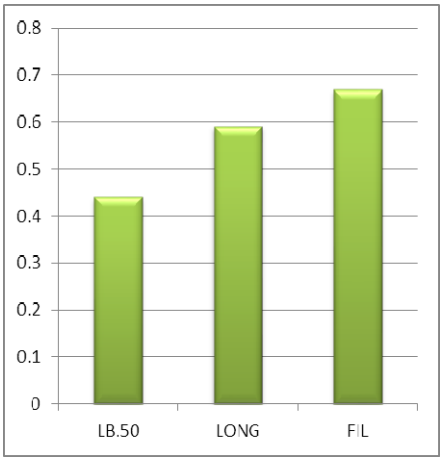
شکل ۳.۴ مقایسه معیار ROUGE-SU4 جملات فشرده شده در ۵ حالت مختلف

همانطور که مشاهده می‌شود در هر دو معیار حالت پنجم و بعد از آن حالت سوم بهترین نتایج را کسب نموده اند. در نتیجه حالت پنجم گزینه مناسبی به نظر می‌رسد. در شکل (۴.۴) و (۵.۴) معیار F و نرخ فشرده سازی دو حالت سوم و پنجم با روش ارائه شده در [FIL08] مقایسه شده است. همانطور که از شکل‌ها هم مشخص است، هنگامی که بلندترین جمله به عنوان جمله فشرده شده انتخاب می-گردد (حالت پنجم)، علاوه بر اینکه نرخ فشرده سازی کمتر می‌شود (جملات کوتاه‌تری تولید می‌شود) معیار F نیز بهبود می‌یابد.



شکل ۴.۴ مقایسه معیار F بر فشرده سازی

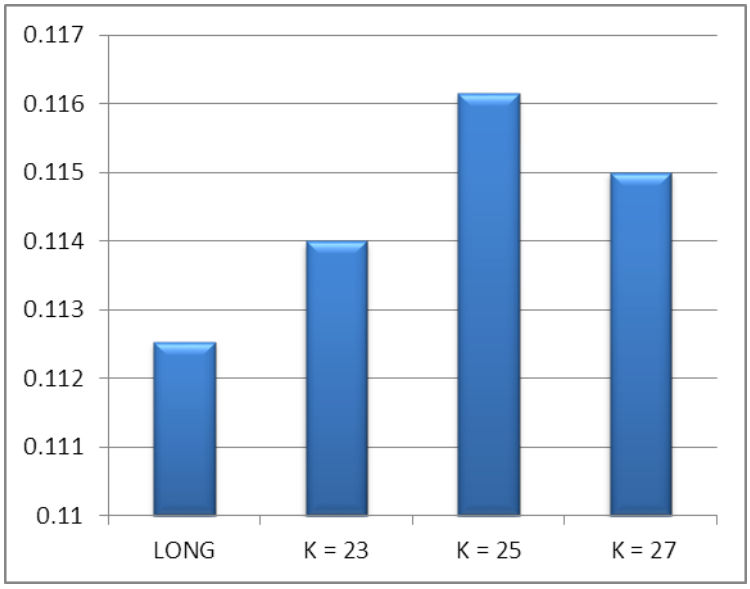
در اینجا دقت و فراخوانی عبارتست از نسبت روابط گرامری مشترک بین جمله خروجی و جمله فشرده شده انسانی به ترتیب به تعداد روابط گرامری جمله خروجی و جمله انسانی. و نرخ فشرده سازی عبارتست از طول جمله فشرده شده به طول اولیه جمله.



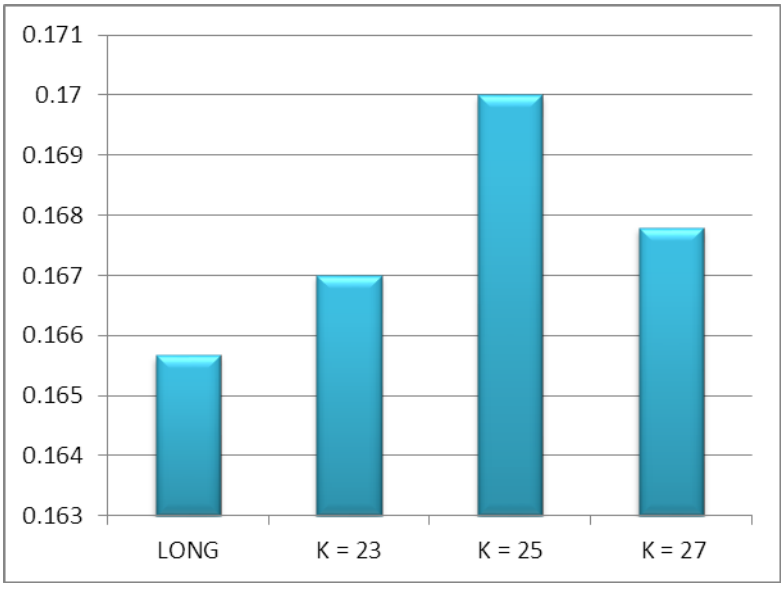
شکل ۵.۴ مقایسه نرخ فشرده سازی

۳.۴ ارزیابی تعداد دسته های مختلف

همان گونه که ذکر شد در فاز آخر جملات ابتدا دسته‌بندی می‌گردند. انتخاب تعداد دسته مناسب برای جملات نیز اهمیت دارد. از این رو تعداد دسته‌های مختلف را نیز آزمودیم. نتایج این آزمون در شکل‌های (۶.۴) و (۷.۴) نمایش داده شده است. با توجه به نتایج، تعداد ۲۵ برای دسته‌ها انتخاب می-گردد.



شکل ۶.۴ مقایسه مقادیر مختلف تعداد دسته ها با معیار ROUGE-2

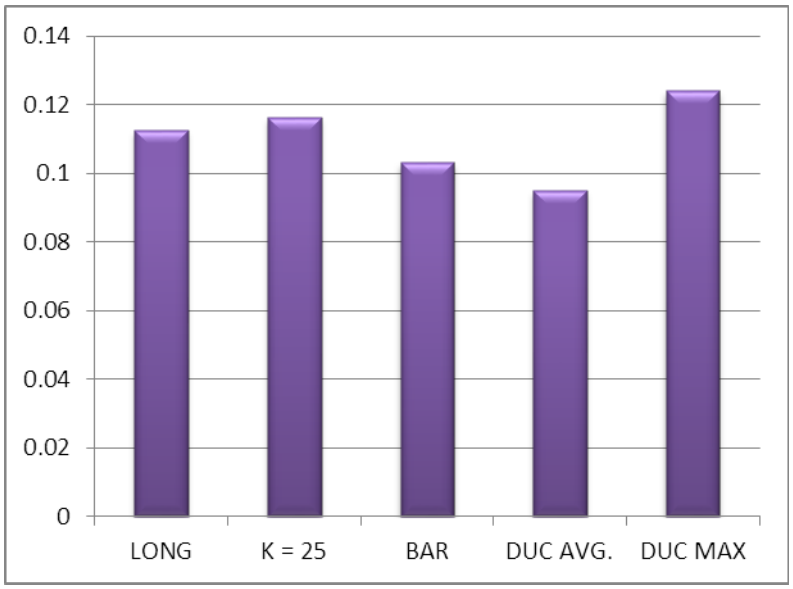


شکل ۷.۴ مقایسه مقادیر مختلف تعداد دسته ها با معیار ROUGE-SU4

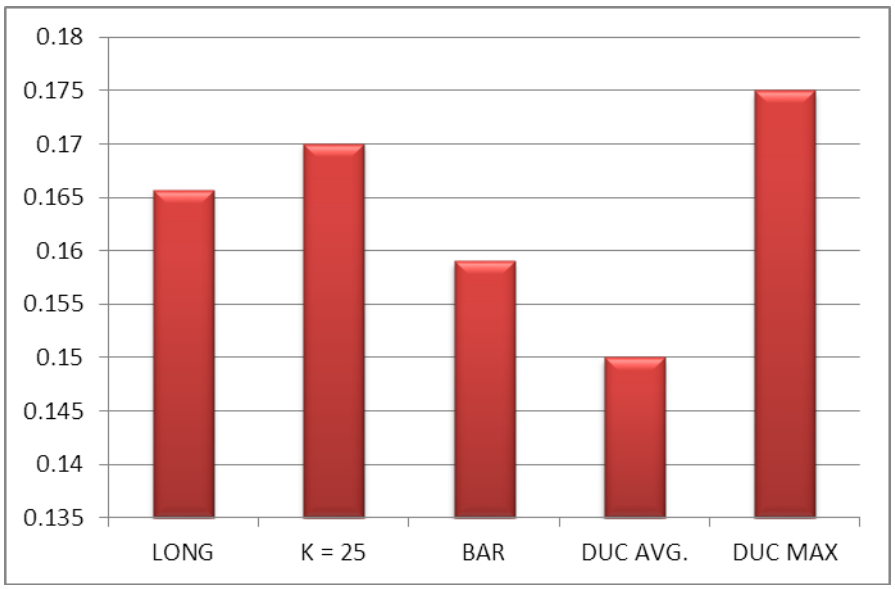
۴.۴ ارزیابی نهایی

نتایج نشان داده شده در شکل (۸.۴) و (۹.۴) نیز حاکی از بهبود قابل ملاحظه در مقایسه با میانگین سیستم‌های موجود در DUC2007 می‌باشد. به شکلی که روش پیشنهادی نتایجی تقریبا

نزدیک به بیشینه سیستمهای موجود در DUC2007 کسب نموده‌است. لازم به ذکر است سیستم‌های موجود در مجموعه داده مذکور سیستم‌های قوی می‌باشند؛ به شکلی که بسیاری مقالات خود را با میانگین آنها مقایسه می‌نمایند. علاوه بر این، روش پیشنهادی در مقایسه با روش پیشنهادی در [BAR05] که عمده‌ترین کار در زمینه خلاصه‌سازی چکیده‌ای است، دارای نتایجی بهتر می‌باشد.



شکل ۸.۴ مقایسه نتایج با معیار ROUGE-2



شکل ۹.۴ مقایسه نتایج با معیار ROUGE-SU4

فصل پنجم

نتیجه گیری و کارهای آتی

۵ نتیجه‌گیری

در این پایان‌نامه به ارائه روشی جدید برای خلاصه‌سازی چند سنده و چکیده‌ای پرداختیم. روش مذکور شامل چندین فاز می‌باشد، ابتدا بهترین جملات با استفاده از یک روش خلاصه‌سازی گزینشی مناسب انتخاب می‌گردند، سپس به معرفی روشی جدید برای مشابهت جملات پرداختیم. که روش مذکور در مقایسه با روش‌های جدید مشابهت جملات پیشرفت زیادی کسب نمود. سپس بر مبنای این مشابهت و با تکیه بر نقش‌های معنایی جملات روشی غیرنظارتی برای فشرده‌سازی جملات ارائه نمودیم تا قسمت‌های غیر ضروری جملات حذف گردند. بر این مبنا ۵ حالت مختلف فشرده سازی ارائه گردید که از آن میان حالت پنجم نتایجی بهتر هم در مقایسه با سایر حالات با معیار ROUGE و هم در مقایسه با سایر روش‌های فشرده‌سازی غیرنظارتی کسب نمود. پس از آن جملات را با استفاده از معیار مشابهت مذکور دسته بندی نموده و روشی نیز برای یکی نمودن جملات موجود در دسته‌ها ارائه نمودیم. روش خلاصه‌سازی مذکور نتایجی مطلوب تر نسبت به روش‌های خلاصه‌سازی مرتبط و بسیار بهتر از میانگین سیستم‌های قوی موجود در DUC2007 و نزدیک به بیشینه آنها کسب نمود.

۱.۵ کارهای آینده

به عنوان کارهای آینده این تحقیق، تمرکز بیشتر بر روی جملات پیچیده به منظور کسب نتایج بهتر در فاز آخر این روش می‌باشد. چرا که با اینکه جملات ساده و کوتاه نتایج خوب و قابل قبولی داشتند، اما در مورد جملات پیچیده شاهد تولید جملاتی گاها نامفهوم بودیم. از این رو استفاده از روش‌های ساده‌سازی جمله[۵۲] می‌تواند گزینه مناسبی برای حل این مشکل باشد. علاوه بر این، از آنجا

[52] Sentence simplification

که این پایان‌نامه دارای فازهای متعدد می‌باشد، تلاش برای بهبود هر یک از این فازها، نتایج نهایی را بهبود خواهد بخشید. بهبود روش‌های گزینش جملات در فاز اول، مشابهت جملات، فشرده‌سازی و حتی خوشه‌بندی جملات در فاز آخر در بهبود نتایج اثربخش خواهد بود؛ به عنوان مثال، در مرحله فشرده سازی جملات، مجموعه‌ای از عبارات حاصل شدند که معادل هم تلقی می‌گردند. با ذخیره-سازی این عبارات معادل و بهره‌گیری از آنها، می‌توان روش شباهت جملات را ارتقا داد.

علاوه بر این، ابزارهای متعددی نیز در این پایان‌نامه مورد استفاده قرار گرفت، که گرچه سعی بر آن بود که از بهترین ابزارهای موجود استفاده گردد، اما هیچ یک دقت صد در صدی ندارند و استفاده از ابزارهای دقیق‌تر ، به طور قطع دقت کار را بالا خواهد برد.

به طور کلی خلاصه‌سازی چکیده‌ای چون کودکی نوپا در حال تلاش برای برداشتن اولین قدم-های خود است و تا رسیدن به سطحی مطلوب و قابل قبول راهی دراز در پیش است.

۶ منابع و مراجع

[اشک ۹۰] اشکذری، س. *"انتخاب بردارهای ویژه و تبدیل فضا در خوشه‌بندی طیفی"* پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد، گروه مهندسی کامپیوتر، ۱۳۹۰

[پور ۹۰] پورمعصومی، آ. *"خلاصه‌سازی خودکار چندسندی مبتنی بر استخراج مفاهیم"* ، پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد، گروه مهندسی کامپیوتر، ۱۳۹۰

[ACH08] Achananuparp, P., Hu, X., Xiajiong, S. "*The Evaluation of Sentence Similarity Measures*". Proceeding DaWaK '08 Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery Springer-Verlag Berlin, Heidelberg, 2008.

[AKS09] Aksoy, G., Bugdayci, A., Gur, T., Uysal, I., Can, F. "*Semantic argument frequency-based multi-document summarization*". ISCIS, 2009.

[BAN03] Banerjee, S. and Pedersen, T. "*Extended gloss overlap as a measure of semantic relatedness*". In Proceedings of IJCAI'03, Acapulco, Mexico, 805-810, 2003.

[BAR99] Barzilay, R. and Elhadad, M. "*Using Lexical Chains for Text Summarization"*. In Proceedings of the ACL Workshop on Intelligent ScalableText Summarization, pages 10–17, Madrid, Spain,August 1999.

[BAR05] Barzilay, R. and K. R. McKeown. "*Sentence fusion for multidocument news summarization"*. Computational Linguistics, 31(3):297–327, 2005.

[BER11] Berg-Kirkpatrick, T., Gillick, D., Klein, D. "*Jointly learning to extract and compress"*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, Stroudsburg, PA, USA. Association for Computational Linguistics, pages 481–490.2011.

[BUD06] Budanitsky, A. and Hirst, G. "*Evaluating wordnet-based measures of lexical semantic relatedness"*. Computational Linguistics, 32(1):13–47, 2006.

[CAR98] Carbonell, J. G. , J. Goldstein . "*The use of MMR, diversity-based reranking for reordering documents and producing summaries"*. In Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998, pp. 335–336, 1998.

[CCG1] Cognitive Computation Group, University of Illinois at Urbana Champaign, Part Of Speech Tagger Demo, http://cogcomp.cs.illinois.edu/demo/pos

[CCG2] Cognitive Computation Group, University of Illinois at Urbana Champaign, Semantic Role Labeling Demo, http://cogcomp.cs.illinois.edu/demo/srl

[CHA00] Chuang and Yang: "*Extracting Sentence Segments for Text Summarization: A Machine Learning Approach*", in: Proceedings of the ACM SIGIR, 2000. [CLA08]

Clarke and Lapata, M. Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 31(1):399–429, 2008.

[CLA08] Clarke and Lapata, M. "*Global inference for sentence compression: An integer linear programming approach"*. Journal of Artificial Intelligence Research, (2008) 31(1):399–429.

[CON01] Conroy, J.M. and O'leary, D.P. "*Text summarization 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, 2001.

[DAS07] Das D. and Martins A. "*A Survey on Automatic Text Summarization.* 2007.

[EDM96] Edmundson, H. "*New methods in automatic extracting"*. Journal of the Association for Computing Machinery, 16(2):264–285. Reprinted in Advances in Automatic Text Summarization, Mani, I. and Maybury,M.T. (Eds.), Cambridge,Mass.: MIT Press, 1999, pp.21-42, 1969.

[ERK04] Erkan, G., Radev, D. R. *"LexRank: Graph-based Lexical Centrality as Salience in Text Summarization*". Journal of Articial Intelligence Research 22, 457-479 , 2004.

[FIL08] Filippova, K., Strube, M. *"Dependency tree based sentence compression"*. In ProceedingOf INLG-08, 25–32, 2008.

[FRA82] Francis,W. N. and Henry K. *"Frequency Analysis of English Usage:Lexicon and Grammar"*. Houghton Mifflin, Boston, 1982.

[GAL07] Galley, M., McKeown, K. "*Lexicalized markov grammars for sentence compression"*. In Proceedings of NAACL/HLT, 180–187, 2007.

[GAL10] Galanis, D., Androutsopoulos, I. *"An extractive supervised two-stage method for sentence compression"*. In Proceedings of NAACL, 2010.

[GAL11] Galanis, D., Androutsopoulos, I.A "New Sentence Compression Dataset and Its Use in an Abstractive Generate-and-Rank Sentence Compressor". In Proceedings of the Language Generation and Evaluation Workshop (UCNLG+Eval), at the Conference on Empirical Methods on Natural Language Processing, Edinburgh, UK, 2011.

[GON01] Gong, Y., Liu, X. *"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, 2001.

[HAT99] Hatzivassiloglou, V, Judith, K, Eleazar, E. "*Detecting text similarity over short passage: Exploring linguistic feature combinations via machine learning"*. In priceeding of thr Jiont SIGDAT Conference on Emprical method in Natural Language Processing and Very Large Corpora. 1999

[HIR98] Hirst, G. , St-Onge, D. "*Lexical chains as representations of context for the detection and correction of malapropisms"*. In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, pages 305–332, 1998.

[HIR01] Hirao, T., Sasaki, Y., Isozaki, H. "*An extrinsic evaluation for question-biased text summarization on qa tasks*". In Proceedings of NAACL workshop on Automatic summarization, 2001

[ISL08] Islam, A. , Inkpen, D. "*Semantic text Similarity using corpus-based word similarity and string similarity"*. ACM transaction on Knowledge Discovery from Data, 2(2), 1-25. 2008.

[JIA97] Jiang, Jay J. , David W. Conrath. "*Semantic similarity based on corpus statistics and lexical taxonomy"*. In Proceedings of International Conference on Research in Computational Linguistics (ROCLING X), Taiwan, pages 19–33,1997.

[KIA06] Kiani, B.A., Akbarzadeh, T.M.R. "*Automatic text summarization using: Hybrid Fuzzy GA-GP* ".Proceedings of the IEEE International Conference on Fuzzy Systems, July 16-21, 2006.

[KIN78] Kintsch, W. , T. van Dijk . "*Toward a model of text comprehension and production*." Psychological Review, 85(5):363–394, 1998.

[KNI02] Knight, K., Marcu, D. "*Summarization beyond sentence extraction: A probabilistic approach to sentence compression"*. Artificial Intelligence, 139(1):91–107, 2002.

[KO04] Ko, Y., Park, J., and Seo, J., "*Improving Text Categorization using the importance of sentences"*. Inf. Proc. Manage. 40. 65-49, 2004.

[LEA98] Leacock, C. , Martin C. "*Combining local context and WordNet similarity for word sense identification"*. In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, pages 265–283. 1998.

[LEE11] Lee, M. C. "*A novel sentence similarity measure for semantic-based expert systems"*. Journal of Expert System With Application, 38 p.6392-6399, (2011)

[LIN98] Lin, D. "*An information-theoretic definition of similarity"*. In Proceedings of the 15th International Conference on Machine Learning, Madison,WI, pages 296–304, July, 1998.

[LIN04] Lin, C.Y. "*Rouge: A package for automatic evaluation of summaries*". In Proceedings of the ACL-04Workshop: Text Summarization Branches Out, Barcelona, Spain, 74–81, 2004.

[LIU04] Liu, Y., Zong, C. "*Example-Based chinese-english MT"*. In Proceeding of IEEE International Conference on Systems. Man, and Cybernetics. Vol.1-7, 2004.

[LUH58] Luhn, H. "*The automatic creation of literature abstracts"*. IBM Journal of Research and Development, 2:159–165, 1958.

[MAN99] Mani, I. , M. T. Maybury (Eds.). "*Advances in Automatic Text Summarization*." Cambridge, Mass.: MIT Press, 1999.

[MAN01] Mani, I. "*Automatic Summarization"*. Amsterdam, Philadelphia: John Benjamins, 2001.

[MAR08] Marie-Catherine de Marneffe , Christopher D. Manning. *"Stanford typed dependencies manual"*2008.

[MCD06] McDonald, R. "*Discriminative sentence compression with soft syntactic evidence"*. In Proceedings of EACL, 297–304, 2006.

[MET05] Metzler, D., Bernstein, Y., Croft, W., Moffat, A., and Zobel, J. "*Similarity measures for tracking information flow"*. Proceedings of CIKM, 517–524, 2005.

[MIH05] Mihalcea, R. , Tarau, P. "*An Algorithm for Language Independent Single and Multiple Document Summarization*". In Proceedings of the International Joint Conference on Natural Language Processing, Korea, 2005

[MOR91] Morris, J. , Hirst, G. "*Lexical cohesion computed by thesaural relations as an indicator of the structure of text"*. Computational Linguistics, 17(1):21–48. 1991.

[MOR05] Morris, J. , Graeme H. "*The subjectivity of lexical cohesion in text"*. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, Computing Attitude and Affect in Text. Springer, New York, pages 41–48, 2005.

[NG02] A. Y. Ng, M. Jordan, *et al., "On spectral clustering: analysis and an algorithm,"* Advances in Neural Information Processing Systems, vol. 14, pp. 849- 856, 2002.

[PER98] P. Perona , W. T. Freeman, "*A factorization approach to grouping*," in ECCV, 1998, pp. 655-670, 1998.

[PON07] Ponzetto, S.P. , Strube, M. "*Knowledge Derived From Wikipedia for Computing Semantic Relatedness*", Journal of Artificial Intelligence Research, 30, 181-212, 2007.

[POR80] Porter, M. "*An algorithm for suffix stripping. Program*".14(3), pp.130-137, 1980.

[QUI93] Quinlan, J. R. "*C4.5: Programs for Machine Learning"*. Morgan Kaufmann Publishers, 1993.

[RAD02] Radev, D. R., Hovy, E., , McKeown, K. "*Introduction to the special issue on summarization"*. Computational Linguistics., 28(4):399-408. [1, 2], 2002.

[RES95] Resnik, P. "*Using information content to evaluate semantic similarity"*. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal, Canada, August, 1995.

[RUB65] Rubenstein, Herbert , John B. Goodenough. *"Contextual correlates of synonymy"*. Communications of the ACM, 8(10):627–633, 1965.

[SCH96] Schölkopf B. ,Smola A., et al., "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," Neural computation, vol. 10, pp. 1299-1319, 1996.

[SCO90] G. L. Scott , H. C. Longuet-Higgins, "*Feature grouping by relocalisation of eigenvectors of the proxmity matrix*," presented at the British Machine Vision Conference, 1990.

[SHI00] J. Shi , J. Malik, "*Normalized cuts and image segmentation*," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, pp. 888-905, 2000.

[SKO72] Skorochod'ko, E. *"Adaptive method of automatic abstracting and indexing"*. *In*, In Information Processing 71: Proceedings of the IFIP Congress 71, pp. 1179–1182, 1972.

[SPA99] Sp¨arck Jones, K. *"Automatic summarizing: Factors and directions"*. In I. Mani & M. T. Maybury (Eds.), Advances in Automatic Text Summarization, pp. 1–12. Cambridge, Mass.: MIT Press, 1999.

[SPA07] Sp¨arck-Jones, K. *"Automatic summarising: A review and discussion of the state ofthe art"*. Technical Report UCAM-CL-TR-679, Cambridge, U.K.: University of Cambridge,Computer Laboratory, 2007.

[STE05] Steinberger, J., Kabadjov, M.A. , Poesio, M., Sanchez-Graillet,O . "*Improving LSA-based summarization with anaphora resolution*". In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005.

[SUS93] Sussna, M. *"Word sense disambiguation for free-text indexing using a massive semantic network"*. In Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93), pages 67–74, Arlington, VA.1993.

[SVO07] Svore, K., Vanderwende, L., and Burges, C. "*Enhancing single-document summarization by combining Rank Net and third-party sources*". In Proceedings of the EMNLP-CoNLL, 2007.

[TUR05] Turner, J., Charniak, E. *"Supervised and unsupervised learning for sentence compression"*. In Proceedings of ACL, 2005.

[WAN08] Wan, S., R. Dale, M. Dras , C. Paris. *Seed and grow: "Augmenting statistically generated summary sentences using schematic word patterns"*. In Proceedings of the 2008 Conference on EmpiricalMethods in Natural Language Processing, Honolulu, Hawaii, 25– 27 October 2008, pp. 543–552.2008.

[WAN09] Wan, S., M. Dras, R. Dale , C. Paris. "*Improving grammaticality in statistical sentence generation: Introducing a dependency spanning tree algorithm with an argument satisfaction model"*. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March – 3 April, pp. 852–860, 2009.

[WAN07] Wan, X., Yang, J. , Xiao, J. "*Manifold-ranking based topic-focused multi document summarization*". In Proceedings of IJCAI, 2007.

[WU94] Wu, Z. , Palmer, M. *"Verb semantics and lexical selection"*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pages 133–138, Las Cruces, NM, June,1994.

اثبات خاصیت تشابه برای روش پیشنهادی شباهت جملات

همانگونه که نتایج ارزیابی در فصل پنجم بیان می‌نمایند، روش پیشنهادی ذکر شده در این پایان‌نامه برای محاسبه شباهت بین جملات روشی مناسب با نتایجی مطلوب و منطقی می‌باشد. در این بخش برای اطمینان بیشتر از درستی نحوه عملکرد الگوریتم مذکور، به بررسی خاصیت عمومی تشابه (خاصیت تقارنی تشابه) که برای هر تابع تشابه (مستقل از اینکه تشابه بر چه شیءای اعمال گردد) صادق خواهد بود، خواهیم پرداخت. به منظور یاداوری، فرمول‌های مربوطه را ذکر می‌نماییم. فرض می‌کنیم $CommonRoles\left(L_{\boldsymbol{i}}{}^{\boldsymbol{A}}, L_{\boldsymbol{j}}{}^{\boldsymbol{B}}\right)=\left\{r_1, r_2, \ldots, r_{\boldsymbol{n_{cm}}}\right\}$ اشتراک نقش‌های معنایی بین دو جمله A و B در سطوح iام و jام باشد و $words_r\left(L_i^A\right)=\left\{w_1, w_2, \ldots, w_{m_A}\right\}$ مجموعه کلمات جمله A (غیر از ایست‌واژه‌ها) در سطح معنایی iام ، در نقش معنایی r و پس از ریشه‌یابی می‌باشند. شباهت بین دو جمله از فرمولهای پ.۱ تا پ.۶ محاسبه می‌گردد. توضیحات این فرمولها در بخش ۳ مفصلا آورده شده است.

$$SenSim\left(A,B\right)=\max_{i\in levelsOfA,\, j\in levelsOfB} LevelSim\left(L_i^A, L_j^B\right) \qquad (پ.۱)$$

$$LevelSim\left(L_i^A, L_j^B\right)= \frac{\sum_{r=1}^{\boldsymbol{n_{cm}}} RoleSim_{\boldsymbol{r}}\left(words_r\left(L_i^A\right), words_r\left(L_j^B\right)\right)}{\boldsymbol{n_{cm}}} \qquad (پ.۲)$$

$$RoleSim_r\left(words_r\left(L_i^A\right),words_r\left(L_j^B\right)\right)=$$
$$\frac{\sum_{k=1}^{m_A}\left(\max_{w_B\in words_r\left(L_j^B\right),w_k\in words_r\left(L_i^A\right)} wordSim\left(w_k,w_B\right)\right)}{m_A+m_B}+ \qquad (پ.۳)$$
$$\frac{\sum_{k=1}^{m_B}\left(\max_{w_A\in words_r\left(L_i^A\right),w_k\in words_r\left(L_j^B\right)} wordSim\left(w_A,w_k\right)\right)}{m_A+m_B}$$

$$WordSim\left(w_1,w_2\right)=Sim_L\left(w_1,w_2\right)*\left(\frac{weight\left(w_1\right)+weight\left(w_2\right)}{2}\right) \qquad (پ.۴)$$

$$weight\left(w\right)=\frac{position_w}{N} \qquad (پ.۵)$$

$$sim_L\left(c_1,c_2\right)=\frac{2\log p\left(lso\left(c_1,c_2\right)\right)}{\log p\left(c_1\right)+\log p\left(c_1\right)} \qquad (پ.۶)$$

برای اثبات خاصیت جابه‌جایی، باید این خاصیت در فرمول پ.۱ بررسی شود. این امر را از پایین به بالا مورد بررسی قرار می‌دهیم. در فرمول پ.۶ $lso\left(c_1,c_2\right)$ به پایین ترین پدر مشترک دو مفهوم در شبکه واژگان اشاره دارد. همانطور که واضح است والد مشترک بین دو نود در یک درخت به ترتیب قرار گرفتن آن دو نود بستگی ندارد. مخرج کسر نیز خاصیت جابه‌جایی را داراست در نتیجه فرمول پ.۶ خاصیت جابه‌جایی را دارد. با در نظر گرفتن خاصیت جابه‌جایی تابع sim_L به راحتی وجود این خاصیت در رابطه پ.۴ اثبات می‌گردد. در نتیجه خواهیم داشت:

$$RoleSim_r\left(words_r\left(L_j^B\right),words_r\left(L_i^A\right)\right)=$$
$$\frac{\sum_{k=1}^{m_B}\left(\max_{w_A\in words_r\left(L_i^A\right),w_k\in words_r\left(L_j^B\right)} wordSim\left(w_k,w_A\right)\right)}{m_A+m_B}+$$
$$\frac{\sum_{k=1}^{m_A}\left(\max_{w_B\in words_r\left(L_j^B\right),w_k\in words_r\left(L_i^A\right)} wordSim\left(w_B,w_k\right)\right)}{m_A+m_B}=$$
$$\frac{\sum_{k=1}^{m_B}\left(\max_{w_A\in words_r\left(L_i^A\right),w_k\in words_r\left(L_j^B\right)} wordSim\left(w_A,w_k\right)\right)}{m_A+m_B}+$$
$$\frac{\sum_{k=1}^{m_A}\left(\max_{w_B\in words_r\left(L_j^B\right),w_k\in words_r\left(L_i^A\right)} wordSim\left(w_k,w_B\right)\right)}{m_A+m_B}=$$
$$RoleSim_r\left(words_r\left(L_i^A\right),words_r\left(L_j^B\right)\right) \qquad (پ.۷)$$

یعنی خاصیت جابه‌جایی برای رابطه پ.۳ نیز برقرار می‌باشد. با در نظر گرفتن این نتیجه برقراری خاصیت جابه‌جایی در روابط پ.۲ و پ.۱ به وضوح حاصل می‌گردد.

Abstract

In recent years, with the increasing growth of information and documents, automated text summarization has been the subjects of many researches. Especially, multi-document summarization in which a summary of several documents are extracted, has received a lot of attention. However, most of these researches have focused on extractive summarization. In contrast, little work is done on abstractive summarization, in which an abstract of the concepts rather than a selection of sentences, is returned.

In this thesis, a new approach is described for multi-document abstractive summarization task, which is based on semantic roles, semantic similarity between sentences, and sentence compression, removal, and merging. The evaluation results show that the proposed method improves the semantic similarity of sentences and compression compared to the current methods. Furthermore, evaluating the proposed method on the DUC dataset, using ROUGE measure, shows some improvement over the other summarization systems.

Keywords: text summarization, multi-document summarization, abstractive summarization, sentence compression, semantic similarity, sentence merger

Ferdowsi University of Mashhad

Faculty of Engineering

Computer Engineering Department

Web Technology Lab

Master of Science thesis

Abstractive summarization based on sentences' similarities

By

Fatemeh Pourgholamali

Supervisor

Dr. Mohsen Kahani

Advisor

Dr. Nader Jahangiri

February 2012