دانشکده مهندسی کامپیوتر

خلاصه‌سازی گزینشی چندسندی متون فارسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش هوش مصنوعی

محسن مشکی

استاد راهنما:

دکتر مرتضی آنالویی

اردیبهشت‌ماه 1388

دانشکده مهندسی کامپیوتر

خلاصه‌سازی گزینشی چندسندی متون فارسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش هوش مصنوعی

محسن مشکی

استاد راهنما:

دکتر مرتضی آنالویی

اردیبهشت‌ماه 1388

به نام خداوند جان و خرد

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پایان‌نامه

نام دانشکده: مهندسی کامپیوتر

نام دانشجو: محسن مشکی

عنوان پایان‌نامه یا رساله: خلاصه‌سازی گزینشی چندسندی متون فارسی

تاریخ دفاع:

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
1	استاد راهنما	دکتر مرتضی آنالویی		دانشگاه علم و صنعت ایران	
2	استاد راهنما				
3	استاد مشاور				
4	استاد مشاور				
5	استاد مدعو خارجی	خانم دکتر مهرنوش شمس‌فرد		دانشگاه شهید بهشتی	
6	استاد مدعو خارجی				
7	استاد مدعو داخلی	دکتر محمدرضا کنگاوری		دانشگاه علم و صنعت ایران	
8	استاد مدعو داخلی				

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب محسن مشکی به شماره دانشجویی 85722135 دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی ارشد تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی:

امضا و تاریخ:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

- بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.
- بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.
- بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

نام استاد یا اساتید راهنما:

تاریخ:

امضا:

تقدیم به:

همه‌ی جویندگان دانش که فرهنگ اسلامی و ایرانی خود را پاس داشته و بـه زبـان شـیرین پارسـی ارج می‌نهند.

سپاسگزاری:

با سپاس از خانواده‌ی عزیزم که همواره پشتیبان من بوده‌اند، اساتیدی که پدرانه و دلسوزانه از عمر خود کاستند و بر دانش ما افزودند، دوستانی که همراهی با آنها مایه‌ی شادی و دلگرمی من است و همه‌ی کسانی که کار من ادامه‌ی تلاش‌های آنهاست.

چکیده

در این پایان‌نامه، یک روش مبتنی بر خوشه‌بندی برای خلاصه‌سازی چندسندی متون پیشنهاد شده است. یک سامانه‌ی خلاصه‌سازی گزینشی چندسندی، خلاصه‌سازی است که چند سند را به عنوان ورودی می‌گیرد و خلاصه‌ای تولید می‌کند که گزیده‌ای از جمله‌های سندهای اولیه است. اگر چه روش پیشنهادی محدود به حوزه نیست، اما ارزیابی آن روی یک مجموعه از خبرهای ورزشی فارسی صورت گرفته است.

یکی از بخش‌های اصلی روش پیشنهادی، خوشه‌بندی جمله‌ها است. در خوشه‌بندی جمله‌ها دو راهکار برای دسته‌بندی هر چه بهتر جمله‌ها بکار گرفته شده است، که عبارتند از:

- استفاده از خوشه‌بندی سلسله‌مراتبی منفرد محدود برای خوشه‌بندی جمله‌ها.
- تولید خودکار بردارهای همبستگی و بردارهای واژه-بافت و استفاده از آنها برای تعیین شباهت بین جمله‌ها.

خوشه‌بندی سلسله‌مراتبی محدود با در نظر گرفت یک کران بالا برای اندازه‌ی خوشه‌ها، از به وجود آمدن خوشه‌هایی با اندازه‌ی بیشتر جلوگیری می‌کند. استفاده از این روش خوشه‌بندی کمک شایانی به گزینش بهتر جمله‌ها می‌کند. همچنین، برای تعیین شباهت جمله‌ها که نقش مهمی در خوشه‌بندی دارد، دو روش پیشنهاد شده است. روش نخست، از همبستگی بین واژه‌ها بهره می‌گیرد که بر مبنای رخداد همزمان واژه‌ها در یک پنجره‌ی با اندازه ثابت بدست می‌آیند. در روش دوم، از شباهت بین بردارهای واژه-بافت واژه‌ها استفاده می‌شود که بیانگر شباهت آنها است. بنابر مطالعه مراجع مرتبط در زبان فارسی، به نظر می‌رسد منابع نامبرده برای نخستین بار در سطح کاربردی برای زبان فارسی تولید شده‌اند.

بیشتر راهکارهای در نظر گرفته شده که خاص زبان فارسی هستند، در بخش‌های پیش‌پردازش و تولید منابع زبانی صورت گرفته است. در بخش پیش‌پردازش، برای رفع مشکل وجود واژه‌های به هم

چسبیده، روشی برای شناسایی و جداسازی آنها پیشنهاد شد. همچنین برای بهبود سرعت محاسبه شباهت بین بردارهای واژه-بافت، بردارهای جدیدی به نام بردارهای هم‌بافت پیشنهاد شد. در بردار هم‌بافت یک واژه، تعدادی از واژه‌هایی که بیشترین شباهت (بین بردارهای واژه-بافت) را نسبت به واژه‌ی اصلی دارند، وجود دارد. هر واژه در این بردار دارای یک وزن است که بیانگر میزان شباهت آن با واژه‌ی اصلی است.

روش ارزیابی استفاده شده در این پایان‌نامه، یک روش ارزیابی مستقیم است. این روش شامل دو بخش است. در بخش نخست، خلاصه‌ی خودکار با تعدادی خلاصه‌ی مرجع که توسط افراد خبره تهیه شده است مقایسه می‌شود و با اهمیت بودن جمله‌های موجود در خلاصه مورد ارزیابی قرار می‌گیرد. در بخش دوم، میزان اطلاعات تکراری در جمله‌های گزینش شده ارزیابی می‌شود.

نتایج حاصل از ارزیابی روش پیشنهادی نشان می‌دهند که استفاده از خوشه‌بندی سلسله‌مراتبی محدود می‌تواند به همراه استفاده از همبستگی لغوی جهت تعیین شباهت جمله‌ها، بهترین کیفیت را نسبت به روش‌های مشابه حاصل کند. با بکارگیری روش پیشنهادی، کارایی از 0,65 به 0,86 (نسبت به روش MEAD) بهبود یافت که این بهبود بدون بروز افزونگی (میزان افزونگی در دو روش یکسان است) بیشتر حاصل شد.

واژه‌های کلیدی: خلاصه‌سازی چند سندی، بردار واژه بافت، همبستگی لغوی، خوشه‌بندی، خوشه‌بندی سلسله مراتبی محدود.

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل 1:

مقدمه

1-1- مقدمه

با افزایش روز افزون منابع متنی در شبکه جهانی وب، هر روز بر گستره‌ی اطلاعات قابل دسترس برای کاربران افزوده می‌شود. اگر چه این رشد سریع مزایای فراوانی دارد، اما مشکلاتی نیز به همراه داشته است. نخستین چالش آن است که کاربر چگونه می‌تواند اطلاعات مورد نیاز خود را بیابد. امروزه، بکارگیری سامانه‌های بازیابی اطلاعات یکی از عمومی‌ترین روش‌های جستجو و کسب اطلاعات مورد نیاز است. خروجی یک سامانه‌های بازیابی اطلاعات معمول، به صورت سیاهه‌ای از عنوان‌ها است که هر کدام با توضیح کوتاهی همراه شده است. بدبختانه، در بیشتر موارد این سیاهه‌ها بسیار بلند هستند و بررسی همه‌ی سندهای بازیابی شده عملی نیست. به طور معمول، کاربران تنها چند ده سند اول را بررسی می‌کنند و بقیه را نادیده می‌گیرند. بیشتر کاربران گرایش به طرح پرس‌وجوهای کوتاه دارند [Franzen and Karlgren, 2000]، بنابر بسیاری از پرس‌وجوها نادقیق و مبهم هستند [Sanderson, 1994]. به این ترتیب، کاربران با چالش دیگری مواجه می‌شوند: چگونه اطلاعات مفید را گزینش کنند. سامانه‌ی خلاصه‌سازی خودکار متن، یک راه حل برای این مشکل است.

سه مزیت عمده‌ی تولید خودکار خلاصه به وسیله ماشین عبارتند از:

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

خلاصه‌سازی یکی از کاربردهای پردازش متن است. پردازش متن شامل چهار سطح است [شمس‌فرد، 1385] : پردازش لغوی، پردازش ساختواژی، پردازش نحوی و پردازش معنایی. هر یک از کاربردهای فراوان پردازش متن، از جمله بازیابی اطلاعات، خلاصه‌سازی، درک، تولید متن، ترجمه‌ی ماشینی، پرسش و پاسخ به زبان طبیعی، استخراج دانش از متون و موارد دیگر، با توجه به گستردگی و

پیچیدگی، در یک یا چند سطح فوق به انجام می‌رسد. خلاصه‌سازی، یکی از پیچیده‌ترین کاربردهای پردازش متن است و معمولا با چند سطح از پردازش متن همراه است.

در سال‌های اخیر فعالیت گسترده‌ای روی ساخت و توسعه‌ی سامانه‌های خودکار خلاصه‌سازی متن برای زبان‌های مختلف انجام شده است.

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

- خلاصه‌ی گزینشی با توجه به معیارهای آماری، شهودی[1] و یا ترکیبی از این دو تهیه می‌شود. از آنجا که در تولید این دسته از خلاصه‌ها، جملات متن تغییرات نحوی و معنایی ندارند، می‌توان آن را نوعی گزینش جملات قلمداد کرد.
- چکیده، تفسیری از متن اولیه است. در تولید چکیده، مفاهیم جملات متن اصلی به شکل کوتاه‌تر بازنویسی می‌شود. به عنوان مثال، جمله «او سیب، انگور و گیلاس‌ها را خورد» را می‌توان به صورت «او میوه‌ها را خورد» نوشت.

تقسیم‌بندی‌های دیگری نیز برای انواع خلاصه وجود دارد، که چند مورد از آنها عبارتند از:

- سامانه‌های خلاصه‌سازی با توجه به تعداد سندهای ورودیشان به دو نوع تک سندی و چند سندی تقسیم می‌شوند. در نوع تک سندی، ورودی تنها یک متن است، اما در نوع چند سندی خلاصه باید بر مبنای اطلاعات چند سند تهیه شود.
- بر حسب میزان عمومیت قابل پشتیبانی توسط سامانه‌ی خلاصه‌سازی، این سیستم‌ها به دو نوع عمومی و محدود تقسیم می‌شوند. در نوع عمومی، متن می‌تواند از هر حوزه‌ای به سامانه‌ی داده شود اما در نوع محدود، سامانه تنها قادر به خلاصه‌سازی موثر متون مربوط به یک حوزه‌ی معین است.
- دسته‌بندی دیگر مربوط به مبتنی بر پرس‌وجو بودن یا نبودن خلاصه‌سازی است. در نوع مبتنی بر پرس‌وجو، ابتدا یک پرس‌وجو از سوی کاربر مطرح می‌شود و خلاصه‌ساز خواسته‌ی کاربر را نیز در تهیه‌ی خلاصه در نظر می‌گیرد، اما در نوع دیگر چنین نیست.

تولید چکیده از یک متن، به مراتب سخت‌تر از ایجاد خلاصه‌ی گزینشی از آن است، زیرا فرآیند

[1] Heuristic

چکیده‌سازی نیازمند دانش فراوان از حوزه‌های مختلف و بکارگیری مناسب آنها است. خلاصه‌ی گزینشی ساده از یک سو، و چکیده‌ای که توسط یک فرد خبره تهیه شده از سوی دیگر را می‌توان دو کران خلاصه‌سازی دانست.

2-1- فرآیند خلاصه‌سازی

در [Lin & Hovy 1997]، سه مرحله اصلی برای خلاصه‌سازی در نظر گرفته شده است، که عبارتند از: 1) شناسایی عنوان، 2) تفسیر و 3) تولید خلاصه. البته این مراحل مربوط به پس از پردازش اولیه روی متن ورودی هستند؛ از آنجا که متون ورودی اغلب نیاز به پیش‌پردازش دارند، می‌توان این مرحله را نیز به سه مورد نامبرده، افزود.

در مرحله‌ی پیش‌پردازش، متن ورودی به ساختاری قابل پردازش برای خلاصه‌ساز تبدیل می‌شود. این ساختار بستگی به ویژگی‌هایی دارد که در بخش بعد از خلاصه‌سازی به کار گرفته می‌شوند. مانند بسیاری از سامانه‌های پردازش زبان دیگر مشخص کردن مرز جمله‌ها، تشخیص اسم‌ها و ریشه‌یابی، معمولا جزیی از این مرحله هستند.

هدف از شناسایی عنوان، تعیین بخش‌های مهم متن است. این کار از راه‌های مختلفی قابل انجام است که بعضی از آنها عبارتند از:

- در بعضی از انواع متنی، موقعیت واژه‌ها یا عبارت‌ها معنی خاصی دارد. در این حالت می‌توان بعضی از بخش‌های مهم متن مانند عنوان و جمله‌های اول هر بند را مورد توجه قرار داد.
- عبارات اشاره[1] مانند «به طور خلاصه»، «در پایان»، «مهم‌ترین»، «در این مقاله» و غیره می‌توانند مبین بخش‌های مهم متن به لحاظ محتوایی باشند.
- فراوانی واژه‌ها نیز می‌تواند میزان اهمیت آنها را نشان دهد (البته به شرطی که یک واژه عمومی[2] مانند حروف اضافه و ضمایر نباشد).

[1] Cue phrases
[2] Stop word

مرحله‌ی بعدی یعنی مرحله‌ی تفسیر، شامل اعمالی مانند ادغام مفاهیم و بخش‌های مرتبط، و حذف افزونگی و غیره می‌شود. میزان پردازش‌های صورت گرفته در این مرحله، در خلاصه‌سازی گزینشی ناچیز و در چکیده‌سازی زیاد است. به عنوان مثال، توالی مفاهیم «نشستن»، «خواندن صورت غذاها»، «سفارش دادن»، «خوردن غذا» و «خارج شدن» می‌تواند به صورت «رفتن به رستوران» تفسیر شود.
مرحله‌ی تولید، مرحله‌ی پایانی در سامانه‌های خلاصه‌سازی است. روش‌های گوناگونی برای تولید متن نهایی وجود دارد که بعضی از آنها عبارتند از:

- گزینش: عبارت‌ها و جمله‌های انتخاب شده در مرحله‌ی شناسایی عنوان، به صورت ساده در خروجی درج می‌شوند.
- لیستی از عنوان‌ها: لیستی از واژه‌ها مهم یا بازنمایی مفاهیم، بدون دست‌کاری در خروجی درج می‌شوند.
- چسباندن عبارت‌ها: دو یا چند عبارت بدون تغییر به هم چسبانده می‌شوند.
- ساختن جمله: یک جمله‌ساز برای تولید جمله‌های جدید مورد استفاده قرار می‌گیرد. ورودی جمله‌ساز، لیستی از مفاهیم و عنوان‌های به هم مرتبط است.

3-1- خلاصه‌سازی چندسندی

می‌توان روش‌های خلاصه‌سازی را بر مبنای تعداد سندهای ورودی به دو نوع تک‌سندی و چندسندی تقسیم کرد. فلسفه‌ی خلاصه‌سازی چند سندی آن است که اطلاعات متنی خاصیت توزیع‌شده و پراکنده دارند؛ برای گردآوری آنها نیاز به سامانه‌هایی است که آنها را یکپارچه کند و به صورت یک متن در آورد. اگر چه شباهت‌های زیادی بین روش‌های خلاصه‌سازی تک‌سندی و چندسندی وجود دارد، اما تفاوت‌های قابل توجهی نیز بین آنها وجود دارد. چند تفاوت عمده بین خلاصه‌سازی تک‌سندی و چندسندی عبارتند از:

- میزان افزونگی اطلاعات در یک دسته از متن‌هایی که به لحاظ موضوعی به هم مربوطند، بیشتر از میزان افزونگی موجود در یک متن است. در خلاصه‌سازی چندسندی، متن‌های

ورودی دارای زمینه‌ی مشترکی هستند و گاهی با هدف روشن ساختن موضوع مشابهی نگارش شده‌اند. این تفاوت بین خلاصه‌سازی تک‌سندی و چندسندی، ایجاب می‌کند که برای خلاصه‌سازی چندسندی از روش‌هایی استفاده شود که ضدافزونگی هستند.

- گاهی در خلاصه‌سازی چندسندی با متن‌هایی مواجه هستیم که دارای یک ترتیب زمانی هستند. به عنوان مثال در خلاصه‌سازی اخبار، هر خبر دارای یک زمان مشخص است و نادیده گرفتن آن در خلاصه‌سازی می‌تواند کیفیت خروجی را پایین بیاورد. در مواردی که متن‌ها دارای یک ترتیب زمانی هستند باید از روش‌هایی سود جست که وزن بیشتری به خبرهای جدید بدهند.
- در خلاصه‌سازی چندسندی، نرخ خلاصه‌سازی به طور چشمگیری کمتر از خلاصه‌سازی تک‌سندی است. این تفاوت از آنجا ناشی می‌شود که در خلاصه‌سازی چندسندی با حجم بیشتری از متن‌ها مواجه هستیم و همچنان انتظار می‌رود که خلاصه‌ساز یک خلاصه‌ی کوچک و مفید تولید کند.
- از آنجا که در خلاصه‌سازی چندسندی، جمله‌های گزینش شده از یک سند نیستند، استفاده از ترتیب اولیه‌ی جمله‌ها در سند اصلی مقدور نیست. در مجموع، مرتب‌سازی جمله‌های گزینش شده برای تولید خلاصه‌ی نهایی در نوع چندسندی بسیار مشکل‌تر از نوع تک‌سندی است.

یک سامانه‌ی نوعی خلاصه‌سازی گزینشی چندسندی با سه چالش عمده روبرو است. این چالش‌ها عبارتند از:

- گزینش اطلاعات: باید بتواند بخش‌های برجسته و مفید متن را شناسایی کند.
- حذف افزونگی: باید راهکاری برای جلوگیری از بروز افزونگی داشته باشد.
- مرتب‌سازی: باید روشی برای مرتب‌سازی اطلاعاتی داشته باشد که از متن‌های گوناگون انتخاب شده‌اند.

یک سامانه‌ی خلاصه‌سازی چندسندی خوب، سامانه‌ای است که از جنبه‌های زیر عملکرد مطلوبی داشته باشد:

- نرخ فشرده‌سازی: در کاربردهای واقعی، کاربران نیاز به تولید خلاصه‌هایی دارند که اندازه‌ی

مشخص دارد. به این ترتیب، یک سامانه‌ی خلاصه‌سازی مطلوب باید به ازای یک مجموعه سند ورودی بتواند خلاصه‌هایی با اندازه‌ی مختلف تولید کند.

- حفظ اطلاعات: سامانه‌ی خلاصه‌سازی باید بتواند تا جای که امکان دارد، اطلاعات معنایی سندهای اولیه را حفظ کند. برای این کار، سامانه‌ی خلاصه‌سازی نیازمند روش‌هایی برای تعیین زیر موضوع‌ها دارد.
- درستی ساختار: سامانه‌ی خلاصه‌سازی باید جمله‌هایی تولید کند که از نظر دستوری درست باشد. همچنین، چینش جمله‌ها در خلاصه باید به نحوی باشد که بین جمله‌های متوالی ارتباط منطقی بر قرار باشد.
- قابلیت توسعه: مطلوب است که یک سامانه‌ی خلاصه‌سازی با تغییرات اندک بتواند برای زبان‌ها و حوزه‌های دیگر توسعه یابد.

1-3-1- مسائل خاص خلاصه‌سازی چندسندی

روش پیشنهادی در این پایان‌نامه، یک روش چندسندی است. مانند بسیاری از روش‌های دیگر چندسندی، روش پیشنهادی را می‌توان در خلاصه‌سازی تک‌سندی نیز به کار گرفت. با این حال، چند ویژگی خاص در روش پیشنهادی وجود دارد، که موجب می‌شود آن را یک روش چندسندی بنامیم. نخست آنکه در آن، از روش‌هایی برای مرتب‌سازی جمله‌هایی که از چند سندهای مختلف به خلاصه می‌آیند استفاده شده است (بر مبنای اطلاعات زمانی متن‌ها و همچنین موقعیت جمله‌ها در متن اصلی، موقعیت آنها در خلاصه تعیین می‌شود). علاوه بر این، روش بکار رفته برای خوشه‌بندی جمله‌ها یعنی خوشه‌بندی سلسله مراتبی اتصال منفرد محدود زمانی ویژگی‌های خاص خود را نشان می‌دهد که با افزونگی زیادی مواجه باشد. اگر از این روش در خلاصه‌سازی تک‌سندی استفاده کنیم، به الگوریتم خوشه‌بندی سلسله مراتبی اتصال مفرد ساده کاهش می‌یابد (نمی‌تواند ویژگی‌های خاص خود را نشان دهد).

1-3-2- فرآیند خلاصه‌سازی چندسندی

خلاصه‌سازی چندسندی فرآیندی است که به ازای چند سند ورودی هم موضوع، یک متن کوتاه تولید می‌کند. این متن باید اطلاعات عمده‌ی موجود در سندهای اولیه را در بر داشته باشد. این فرآیند را می‌توان به صورت چهار مرحله‌ی پشت سر هم مانند زیر در نظر گرفت:

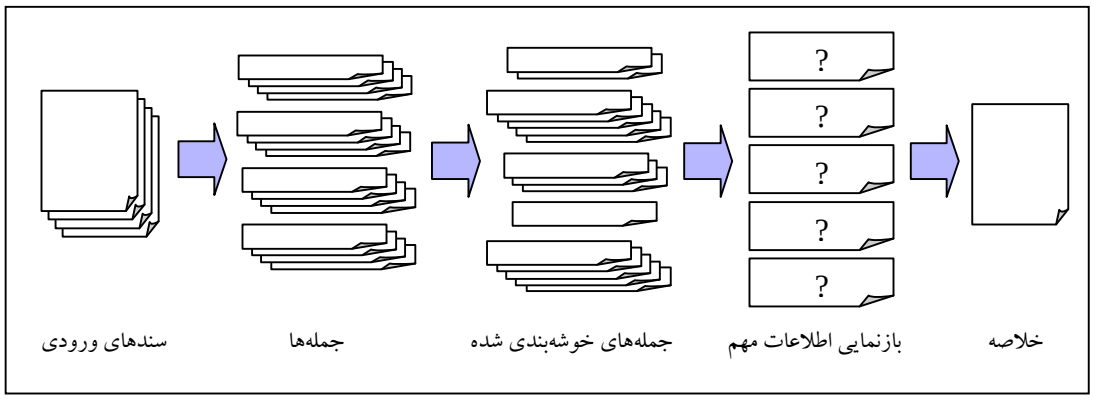
مرحله‌ی 1: سندهای ورودی به جمله‌های تجزیه می‌شوند.

مرحله‌ی 2: جمله‌ها در چند دسته خوشه‌بندی می‌شوند.

مرحله‌ی 3: هر دسته به صورت یک شکل میانی (که می‌تواند جمله باشد) بازنمایی می‌شود.

مرحله‌ی 4: با توجه به بازنمایی‌های مرحله‌ی 3، متن پایانی تولید می‌شود.

شکل 1-1، مراحل بالا را نشان می‌دهد. بیشتر سامانه‌های خلاصه‌سازی متن از این چهارچوب کلی پیروی می‌کنند.



شکل (1-1) معماری عمومی یک سامانه‌ی خلاصه‌سازی

1-3-3- رهیافت‌های عمده

در حوزه‌ی خلاصه‌سازی چندسندی، می‌توان به سه کار عمده اشاره نمود که از جنبه‌های مختلف پوشش‌دهنده‌ی فعالیت‌های صورت گرفته در این حوزه‌اند. این سه کار عبارتند از:

- MEAD [Radev et al., 2004]: MEAD یا خلاصه‌ساز چندسندی مبتنی بر مرکز[1]، یک روش خلاصه‌سازی گزینشی معرفی می‌کند، که در آن خلاصه زیر مجموعه‌ای از جمله‌های سندهای اولیه هستند.
- کاهش جمله [Knight and Marcu, 2002]: دو روش برای کاهش جمله‌های بلند به جمله‌های کوتاه معرفی کرده‌اند. با توجه به اینکه روش آنها جمله‌ها را تغییر می‌دهد، می‌توان آن را یک روش تولید چکیده دانست.
- آمیزش اطلاعات[2] [Barzilay et al., 1999]: روشی برای آمیختن اطلاعات پیشنهاد کرده‌اند. این روش هم برای تولید چکیده استفاده می‌شود.

در ادامه، هر یک از این سه روش به طور جداگانه بررسی خواهد شد.

♦ MEAD

MEAD یک سامانه‌ی خلاصه‌سازی گزینشی است که در آن، خلاصه مجموعه‌ای گزینش‌شده از جمله‌های اولیه هستند. مبنای MEAD، انتخاب جمله‌ها به شیوه‌ای است که دو معیار زیر بهینه شوند:

- سودمندی نسبی درون خوشه (CBRU[3]): بیانگر میزان ارتباط یک جمله‌ی معین با موضوع عمومی خوشه است (منظور از خوشه، یک مجموعه خبر مرتبط به هم است).
- اشتراک اطلاعات بین جمله‌ای (CSIS[4]): نشان می‌دهد که یک جمله تا چه اندازه اطلاعات عرضه شده در سایر جمله‌ها را تکرار می‌کند.

یک خلاصه‌ساز چندسندی در حالت ایده‌آل باید به طور همزمان CBRU را بیشینه و CSIS را کمینه کند. پیش از آنکه سندهای ورودی به MEAD داده شوند، اگر خوشه‌بندی نشده‌اند باید با بکارگیری یک روش مبتنی بر مرکز، خوشه‌بندی شوند.

در MEAD، هر خوشه از سندها مانند D به صورت برداری از واژه‌ها بازنمایی می‌شود که هر عنصر از این بردار متناظر با یک واژه‌ی موجود در سندها است. به هر واژه یک وزن نسبت داده می‌شوند که

---

[1] Centroid Based Multi-Document Summarizer

[2] Information fusion

[3] Cluster-based relative utility

[4] Cross-sentence information subsumption

این وزن با بکارگیری معیار TF.IDF محاسبه می‌شود.

رادف و همکارانش سه ویژگی را برای گزینش جمله‌ها معرفی می‌کنند: 1) میزان مرکزیت جمله 2) ارزش مکانی 3) هم‌پوشانی با جمله‌ی اول (برای آگاهی از چگونگی تعریف این ویژگی‌ها به [Radev et al., 2004] مراجعه کنید). هر سه ویژگی بین 0 و 1 نرمال‌اند. پس از محاسبه‌ی تک‌تک ویژگی‌ها، امتیاز خام یک جمله به صورت زیر مشخص می‌شود:

$$SCORE_O(S_{i,k}) = w_c C_{i,k} + w_p P_{i,k} + w_f F_{i,k} \qquad (1\text{-}1)$$

در این رابطه، $S_{i,k}$ جمله‌ی i ام از سند D_k در خوشه‌ی D است، $C_{i,k}$، $P_{i,k}$ و $F_{i,k}$ به ترتیب امتیازهای مرکزیت، ارزش مکانی و هم‌پوشانی با جمله‌ی نخست هستند و w_c، w_p و w_f وزن‌های سه ویژگی می‌باشند.

همچنین برای محاسبه‌ی معیار CSIS یا اشتراک اطلاعات بین جمله‌ای از رابطه زیر استفاده می‌شود:

$$R(S,S') = 2 \times \frac{length(S \cap S')}{length(S) + length(S')} \qquad (2\text{-}1)$$

الگوریتم MEAD به صورت بازگشتی امتیاز جمله‌ها را به صورت زیر محاسبه می‌کند:

$$SCORE_{r+1}(S_{i,k}) = SCORE_r(S_{i,k}) - SCORE_O(S') \cdot R(S_{i,k}, S') \qquad (3\text{-}1)$$

به طوریکه

$$S' = \underset{SCORE_r(\hat{S}) > SCORE_r(S_{i,k})}{\arg\max} R(S_{i,k}, \hat{S}) \qquad (4\text{-}1)$$

که در آن $S_{i,k}$ جمله‌ی i ام از سند D_k در خوشه‌ی D است. الگوریتم بازگشتی تا زمانی که تغییر امتیاز منجر به تغییر در ترتیب جمله‌ها نشود، ادامه پیدا می‌کند. پس از پایان فرآیند بازگشتی، الگوریتم جمله‌ها را بر حسب امتیاز آنها مرتب می‌کند و با توجه به نرخ فشرده‌سازی، تعدادی از آنها را برمی‌گزیند. در گام آخر، جمله‌ها بر مبنای زمان تولید سندی که از آن آمده‌اند، برچسب خورده و

مرتب می‌شوند. اگر دو یا چند جمله از یک سند آمده باشند، ترتیب مکانی آنها در سند اولیه مشخص کننده‌ی ترتیب آنها خواهد بود.

♦ کاهش جمله

نایت و مارکو [Knight and Marcu, 2002] دو الگوریتم برای فشرده‌سازی جمله‌ها معرفی کردند. ورودی هر الگوریتم یک جمله‌ی طولانی است و خروجی به طور معمول یک جمله‌ی کوتاه است که بیشتر اطلاعات معنایی جمله‌ی اولیه را در بر دارد. هر دو الگوریتم معرفی شده از تکنیک‌های یادگیری ماشین بهره می‌گیرند. این دو الگوریتم را می‌توان به صورت دو تبدیل کننده‌ی درخت به درخت دید که روی یک مجموعه داده‌ی آموزش می‌بینند.

برای سادگی، این دو الگوریتم را SC-1 و SC-2 می‌نامیم. الگوریتم SC-1 از یک مدل کانال نویزی[1] بهره می‌گیرد که شباهت فراوانی به مدل‌های کانال نویزی دارد که در ترجمه‌ی ماشینی و شناسایی گفتار بکار می‌روند. در SC-2 هم از یک مدل شرطی بهره گرفته شده است که در آن، هر تصمیم بر مبنای ورودی فعلی و سابقه‌ی پیشین گرفته می‌شود.

ورودی هر دو الگوریتم، یک درخت تجزیه است که به طور معمول توسط یک تجزیه‌گر خودکار تولید می‌شود. هر دو الگوریتم، تلاش می‌کنند این درخت تجزیه را به یک درخت تجزیه‌ی کوچک‌تر تبدیل کنند. فرآیند کاهش درخت تجزیه با اعمال دنباله‌ای از تبدیل‌ها پیاده می‌شود که هر تبدیل بخشی از درخت را حذف می‌کند.

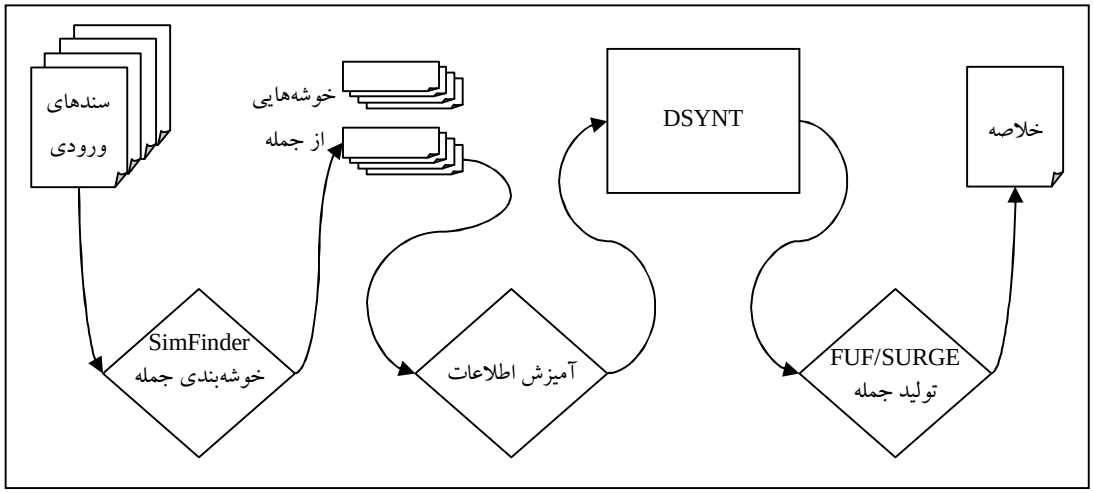
پیکره‌ی آموزشی بکار گرفته شده برای آموزش دو الگوریتم پیشنهاد شده توسط نایت و مارکو روی پیکره‌ی Ziff-Davis ساخته شده است. پیکره‌ی Ziff-Davis مجموعه‌ای از مقاله‌های روزنامه است که به اطلاع‌رسانی در مورد کالاهای کامپیوتری اختصاص دارند. پیکره‌ی آموزشی شامل 1067 جفت جمله به صورت (طولانی، کوتاه) است که جمله‌های طولانی از پیکره‌ی Ziff-Davis بدون دست‌کاری آمده‌اند و نسخه‌ی کوتاه هر جمله توسط یک انسان خبره تولید شده است.

در تولید جمله‌های کوتاه چند اصل رعایت شده است که عبارتند از: هر جمله کوتاه باید تنها با

---

[1] Noisy channel model

واژه‌های موجود در جمله‌ی طولانی اولیه ساخته شود و همچنین واژه‌ها باید با همان ترتیبی در جمله‌ی کوتاه ظاهر شوند که در جمله‌ی طولانی بین آنها برقرار بوده است. روش‌های MEAD و الگوریتم‌های کاهش جمله را می‌توان به صورت دو مولفه به طور همزمان در یک سامانه‌ی خلاصه‌سازی به کار گرفت و از ویژگی‌های مفید هر دو بهره گرفت.



شکل (1-2) معماری MultiGen

♦ آمیزش اطلاعات

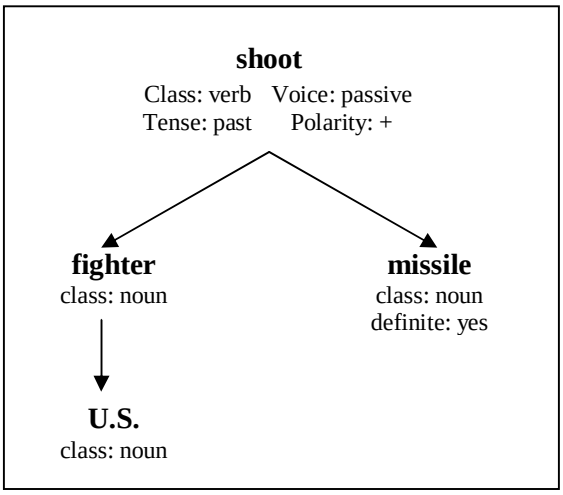
روش معرفی شده توسط بارزیلای و همکارانش [Barzilay et al., 1999]، مولفه‌ی مرکزی سامانه‌ی خلاصه‌سازی MultiGen است [McKeown et al., 1999] که توسط دانشگاه کلمبیا پیاده‌سازی شده است.

معماری MultiGen در شکل 1-2 به نمایش در آمده است. سامانه‌ی MultiGen از سه مولفه‌ی اصلی بهره می‌گیرد که دو مولفه‌ی SimFinder و FUF/SURGE در فرآیند آمیزش اطلاعات نقشی ندارند و الگوریتم‌های معمول را انجام می‌دهند.

مولفه‌ی SimFinder یک الگوریتم خوشه‌بندی جمله است [Hatzivassiloglou et al., 1999] که جمله‌های موجود در سندهای ورودی را به چند خوشه تقسیم می‌کند. مولفه‌ی FUF/SURGE هم یک مولد زبان است که بازنمایی کارکردی[1] یک جمله را می‌گیرد و یک جمله به زبان طبیعی باز

[1] Functional representation

می‌گرداند.



شکل (1-3) DSYNT مربوط به جمله‌ی "U.S. fighters was shot by missile"

مولفه‌ی اصلی MultiGen، مولفه‌ی آمیزش اطلاعات است. این مولفه، یک مجموعه جمله را از خوشه‌بند جمله‌ها یا همان SimFinder می‌گیرد و یک بازنمایی میانی تولید می‌کند که DSYNT نامیده می‌شود. نحوه‌ی کار به این صورت است که ابتدا یک DSYNT به ازای هر جمله در خوشه ساخته می‌شود و سپس به جستجو برای یافتن اشتراک بیشینه‌ای می‌گردد که همه‌ی جمله‌ها را در بر گیرد. یک DSYNT، یک بازنمایی وابستگی[1] است که با بکارگیری یک تجزیه‌گر آماری ساخته می‌شود. شکل 1-3 [Barzilay et al., 1999]، نمونه‌ای از یک DSYNT را نشان می‌دهد.

1-3-4- کارهای انجام شده در حوزه‌ی زبان فارسی

کارهای انجام شده در حوزه‌ی خلاصه‌سازی متون فارسی اندک و انگشت شمار هستند و بیشتر آنها به خلاصه‌سازی تک‌سندی پرداخته‌اند. در [Mazdak and Hassel, 2000]، یک سامانه‌ی خلاصه‌سازی به نام FarsiSum معرفی شده است. این سامانه، نسخه‌ی تغییر یافته‌ی یک سامانه خلاصه‌سازی متون سوئدی [Dalianis, 2000] به نام SweSum برای پوشش زبان فارسی است.

[1] Dependency representation

خلاصه‌ی خروجی SweSum از نوع گزینشی است و برای زبان‌های سوئدی، نروژی، دانمارکی، اسپانیایی، انگلیسی، فرانسوی و آلمانی پیاده‌سازی شده است. این سامانه خلاصه‌ساز، متن را در قالب متنی یا HTML دریافت می‌کند. متن ورودی می‌تواند از روزنامه انتخاب شود یا به صورت یک گزارش باشد.

در [کریمی و شمس‌فرد، 1385]، یک روش خلاصه‌سازی تک‌سندی دیگر پیشنهاد شده است. این روش مانند FarsiSum بر مبنای گزینش جمله‌ها کار می‌کند. همچنین، محتوی خلاصه می‌تواند کلی یا بر اساس پرس‌وجوی کاربر باشد. ایده‌ی بکار رفته در گزینش جمله‌ها در این خلاصه‌ساز، ترکیبی از دو روش زنجیره‌ی لغوی و نظریه‌ی گراف است. در [ورفریاری 1376]، علاوه بر مروری بر روش‌های خلاصه‌سازی موجود، گزارشی از پیاده‌سازی یک نمونه‌ی عملی برای خلاصه‌سازی فارسی ارایه شده است. در [مشکی، 1386]، پس از بررسی موضوع‌ها و چالش‌های مربوط به پردازش متون فارسی، مروری بر روش‌های خلاصه‌سازی موجود صورت گرفته است.

در [اخوان و عرفانی، 1386]، یک روش خلاصه‌سازی گزینشی پیشنهاد شده است که قابلیت بکارگیری در هر دو حالت تک‌سندی و چندسندی را دارد. در کار آنها، از معیارهایی مانند وجود واژه‌های مهم، وجود واژه‌ها و عبارت‌های اشاره، وجود واژه‌های عنوان، وجود نقل‌وقول و ... برای امتیازدهی به جمله‌ها بهره‌گیری شده است. همچنین، برای جلوگیری از افزونگی چنانچه دو جمله شباهتی بیش از یک مقدار آستانه داشته باشند، جمله‌ای که امتیاز کمتر دارد را نادیده می‌گیرد. از دیگر ویژگی‌های کار آنها می‌توان به قابلیت دریافت درخواست از کاربر و غیر وابسته بودن به قلمرو (در مورد سندهای ورودی) اشاره نمود. در [Honarpisheh et al., 2008]، یک روش برای خلاصه‌سازی چندسندی پیشنهاد شده است. مبنای این روش، استفاده از روش سلسله مراتبی اتصال میانگین برای خوشه‌بندی جمله‌ها و بکارگیری روش تجزیه به مقادیر منفرد یا SVD[1] برای تعیین اهمیت جمله‌ها است. در این کار، از دو منبع زبانی ساده برای قطعه‌بندی متن به واژه‌ها و دیگری برای تعیین فراوانی واژه‌ها در سندها استفاده شده است.

در [شهابی، 1381]، یک روش خلاصه‌سازی چندسندی معرفی شده است که نوع خروجی تولید شده

[1] Singular value decomposition

توسط آن نیز گزینشی است. در [مشکی و آنالویی 1388-الف]، یک روشی برای خلاصه‌سازی چندسندی متون فارسی پیشنهاد شده است که در آن از الگوریتم خوشه‌بندی Kmeans برای خوشه‌بندی جمله‌ها استفاده شده است. همچنین، در [مشکی و آنالویی 1388-ب؛ مشکی و آنالویی 1388-ج]، نوعی از خلاصه‌ی خبری به نام خلاصه‌های پیشینه-خبر معرفی شده است. در این نوع از خلاصه، به همراه هر خبر خلاصه‌ای از خبرهای پیشین مرتبط با آن در اختیار کاربر قرار می‌گیرد. در [مشکی و آنالویی 1388-ب] روشی برای بازیابی خبرهای پیشین و در [مشکی و آنالویی 1388-ج] روشی برای خلاصه‌سازی خبرهای پیشین پیشنهاد شده است.

1-3-5- کارهای دیگر

علاوه بر کارهایی که در بخش رهیافت‌های عمده و کارهای انجام شده در زبان فارسی به آنها اشاره شد، در این بخش به بررسی روش‌های دیگری که به موضوع پایان‌نامه مربوط هستند، می‌پردازیم.

بخش عمده‌ای از کارهای انجام شده به معرفی روش‌های نوینی برای خوشه‌بندی جمله‌ها پرداخته‌اند. در [Hatzivassiloglou et al., 2001] روشی پیشنهاد شده که در آن خوشه‌بندی بر مبنای یک تابع هدف صورت می‌گیرد. این تابع، تابعی است که بر مبنای عدم شباهت بین جمله‌ها در خوشه تعریف شده است و باید کمینه شود. برای کمینه کردن این تابع، ابتدا خوشه‌ها با استفاده از روش Kmeans خوشه‌بندی می‌شوند. به دنبال آن، تا جایی که امکان جابجایی جمله‌ها بین خوشه‌ها وجود داشته باشد به نحوی که جابجایی منجر به کاهش تابع هدف گردد، فرآیند جابجایی ادامه می‌یابد.

همچنین از میان کارهایی که به ایجاد منابع خودکار پرداخته‌اند، می‌توان به [Santos, 2006] اشاره نمود. در این کار، روشی برای ایجاد خودکار زنجیره‌های لغوی از یک پیکره‌ی خام پیشنهاد شده است. در این کار، همچنین برای سنجش سودمندی زنجیره‌های حاصل، از آن در خلاصه‌سازی متن استفاده شده است.

1-4- نگاهی کوتاه بر پایان‌نامه

این پایان‌نامه، به معرفی یک روش جدید برای خلاصه‌سازی چندسندی متون فارسی اختصاص دارد. روش پیشنهادی یک روش گزینشی است که خلاصه‌ی تولید شده در آن، گزیده‌ای از جمله‌های سندهای اولیه است. از میان سه رهیافت عمده‌ای که در 1-3-2 مورد بررسی قرار گرفتند، روش MEAD شباهت بیشتری با روش پیشنهادی دارد. مهم‌ترین شباهت این دو روش، گزینشی بودن آنهاست. با وجود شباهت‌های قابل توجه بین دو روش، تفاوت‌هایی نیز بین آنها وجود دارد. بیشتر این تفاوت‌ها، مربوط به نحوه‌ی گزینش جمله‌ها در آنهاست. در روش پیشنهادی، علاوه بر استفاده از یک روش خوشه‌بندی متفاوت نسبت به روش MEAD، از معیارهای شباهت دیگری برای تعیین شباهت بین جمله‌ها استفاده شده است.

در ادامه و در فصل دوم از پایان‌نامه، نگاهی کوتاه به دو حوزه مرتبط یعنی خوشه‌بندی و بازیابی اطلاعات شده است. فصل سوم، به بررسی فرآیند پیش‌پردازش متون فارسی و همچنین نحوه‌ی تولید منابع زبانی مورد نیاز روش پیشنهادی اختصاص دارد. در فصل چهارم، علاوه بر معرفی فرآیند کلی روش پیشنهادی، روش جدیدی برای خوشه‌بندی جمله‌ها پیشنهاد شده است. در این فصل، همچنین دو معیار برای سنجش شباهت بین جمله‌ها معرفی شده است. فصل پنجم پایان‌نامه، حاوی نتایج پیاده‌سازی و ارزیابی روش پیشنهادی و همچنین مقایسه‌ی آن با سه روش دیگر از جمله MEAD است. فصل ششم و پایانی پایان‌نامه نیز به جمع‌بندی مطالب و همچنین ارایه چند پیشنهاد برای بهبود روش پیشنهادی اختصاص یافته است.