نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

خلاصه سازی چکیده‌ای مبتنی بر مشابهت جملات

فاطمه پورغلامعلی[۱]، محسن کاهانی[۲]، آصف پورمعصومی[۲]

fatemehjfg@gmail.com, kahani@um.ac.ir, as.poormasoomi@stu-mail.um.ac.ir

[۱] گروه مهندسی کامپیوتر دانشگاه ولیعصر (عج) رفسنجان
[۲] آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد

چکیده

خلاصه‌سازی خودکار متن مبحثی مورد علاقه در زمینه‌های مختلف بازیابی اطلاعات می‌باشد. در یک تقسیم بندی کلی روش‌های خلاصه‌سازی خودکار متن به دو دسته تک سنده و چندسنده تقسیم بندی می‌شوند. روش ارائه شده در این مقاله در دسته دوم قرار می‌گیرد. این روش ترکیبی از روشهای گزینشی و چکیده‌ای می‌باشد. پس از پیش‌پردازش‌های لازم بر روی جملات اسناد و گزینش بهترین آنها از دیدگاه خلاصه‌سازی، یک معیار برای شباهت معنایی بین جملات ارائه می‌گردد. بر اساس این شباهت و بر مبنای نقش‌های معنایی جملات یک الگوریتم فشرده سازی به منظور حذف قسمت‌های غیرضروری جملات اعمال می‌گردد. الگوریتم فشرده سازی پیشنهادی غیر نظارتی بوده و دارای نتایجی بهتر نسبت به روشهای غیرنظارتی فشرده سازی جملات میباشد. پس از آن جملات به گروه‌هایی تقسیم شده و جملات موجود در هر گروه حذف و یا با یکدیگر ادغام می‌گردند. نتایج حاصل بر روی مجموعه داده DUC2007 نشانگر بهبود خلاصه‌سازی نسبت به بسیاری از روش‌های مذکور می‌باشند.

کلمات کلیدی :خلاصه‌سازی خودکار متن، خلاصه‌سازی گزینشی، خلاصه‌سازی چکیده‌ای، شباهت معنایی

۱. مقدمه

خلاصه سازی خودکار متن به فرایند ایجاد نسخه‌ای کوتاه از متن اشاره دارد که اطلاعات مفید را برای کاربر فراهم آورد.خلاصه-سازی مبحثی است که در سال‌های اخیر توجه ویژه‌ای به آن شده است و در زمینه بازیابی اطلاعات نقشی مهم دارد. با رشد روزافزون اسناد و اطلاعات، در صورتیکه بتوان صورتیاز اسناد مرتبط با هر موضوع را به شکل خلاصه و فشرده ارائه داد، برای پیدا کردن اطلاعات و اسناد، کمک بزرگی به کاربران نموده-ایم[1]. خلاصه سازی چندسنده عبارتست از تولید یک خلاصه از محتوای اطلاعاتی یک مجموعه سند درباره یک موضوع اصلی [2]. به طور کلی خلاصه‌سازی متن به دو دسته گزینشی و چکیده‌ای تقسیم‌بندی می‌شود. اکثر سیستم‌های خلاصه‌سازی گزینشی هستند. در خلاصه‌سازی گزینشی، از متون اولیه چندین بخش را که اغلب جمله واحد کاری معمول است انتخاب کرده و بر اساس یک معیار اولویت آن‌ها را مرتب می‌نمایند. مشکل روش‌های خلاصه‌سازی گزینشی این است که در بسیاری موارد ما شاهد یک میزان اشتراک اطلاعاتی بین جملات هستیم. در این موارد در صورتیکه برخی جملات حذف شوند، میزانی از اطلاعات را از دست داده‌ایم و در صورتیکه همه جملات آورده شوند، دچار افزونگی اطلاعات خواهیم شد و از هدف خلاصه-سازی دورخواهیم شد. در فرایند خلاصه سازی چکیده‌ای به گزینش ساده جملات اکتفا نکرده و در واقع برگرفته‌ای از متن را خواهیم داشت. فرایند خلاصه‌سازی چکیده‌ای فرایندی بسیار پیچیده و دشوار است، چرا که نیازمند نمایشی مفهومی از متن می‌باشد و رسیدن به این نمایش بسیار مشکل خواهد بود. علاوه بر این برای ساخت جمله‌ای جدید نیاز به اطلاعات زبان‌شناسی بسیار قوی می‌باشد. به همین سبب در این زمینه کارهای کمی به انجام رسیده‌است. اکثر کارهای انجام شده بر مبنای درخت تجزیه حاصل از جملات عمل کرده و اقدام به اعمال تصحیحاتی بر روی آن می‌پردازند [3]، [4]، [5]، [6]، [7]. نتیجه اعمال تغییر در درخت تجزیه، گاه درخت‌هایی ناقص و گاه گراف‌هایی پدید می‌آورد که برای تبدیل به فرم مناسب درخت تجزیه بایستی مراحل دیگری را طی نماید. علاوه بر این درخت حاصل بایستی به شکلی مناسب به جمله‌ای درست و گویا بدل شود. این نیز خود بایستی پردازشی دقیق به همراه داشته باشد. برای بررسی صحت گرامری جملات تولید شده، اغلب روشها از یک مدل زبانی استفاده می‌کنند [3][4][8] و برخی نیز یک سری قوانین دستی بر روی جملات اعمال می‌نمایند [9]. دو گام اساسی مطرح شده در زمینه خلاصه‌سازی چکیده‌ای عبارتند از فشرده سازی جملات و ادغام جملات. فشرده سازی جملات به فرایند حذف اجزایی از جمله در حین حفظ اطلاعات اصلی آن اتلاق می‌شود. در این زمینه اغلب روش‌های ارائه شده، از روش-