خلاصه سازی انتزاعی متن مبتنی بر برچسب گذاری نقش معنایی

سینا دامی[۱]، حسین داوطلب محمودی[۲]

[۱]استادیار، دانشگاه آزاد اسلامی واحد تهران غرب، گروه کامپیوتر، تهران، ایران dami@wtiau.ac.ir
[۲]دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد تهران غرب، گروه کامپیوتر، تهران، ایران hd.mahmoodi@yahoo.com

چکیده

خلاصه سازی متن یک پروسه استخراج اطلاعات برجسته از متن منبع و تولید خلاصه مطلوب برای ارائه به کاربر می‌باشد. تولید خلاصه دستی خصوصا دراسناد حجیم، کاری دشوار و زمان بر است. با این وجود خلاصه‌های اتوماتیک با تحلیل عمقی‌تر از متن روند کار را بهبود بخشیده‌اند. از آنجا که هدف ما بکارگیری یک شیوه مبتنی بر برچسب‌گذاری نقش معنایی برای تولید جملات جدید و استفاده از آن‌ها در خلاصه است، شیوه‌ای را پیشنهاد می‌دهیم که در آن جملات متن به درون مجموعه گراف‌های انتزاعی تجزیه می‌شوند، تا پس از حذف اطلاعات زائد طبق معیار برداشت، به یک گراف خلاصه منتقل شوند. نتایج تجربی حاکی از بهبود عملکرد روش پیشنهادی در مقایسه با روش‌های پایه می‌باشد.

واژه های کلیدی

خلاصه سازی انتزاعی، تحلیل معنایی، خلاصه، افزونگی داده، برچسب زنی نقش معنایی

۱- مقدمه

در عصر حاضر با توجه به افزایش تقاضا برای خلاصه سازی حجم وسیعی از اطلاعات، روش‌های متفاوتی از قبیل استخراجی و انتزاعی عرضه شده‌اند. اما در این میان خلاصه‌هایی قابل قبول هستند که در حین اخباری بودن اطلاعات ارزشمندی را برای کاربر فراهم می‌سازند [1]. خلاصه سازی استخراجی شامل استخراج جملات از منبع و اضافه کردن آن‌ها به خلاصه است. پیاده سازی این شیوه ساده بوده و مبتنی بر ویژگی‌های آماری است نه ارتباطات معنایی. این پروسه به مرحله پیش پردازش و پردازش تقسیم می‌شود. فاز پیش پردازش که در [2] آمده به این شکل پیش می‌رود که ابتدا مرز جملات با تشخیص خاتمه دهنده‌ها آغاز می‌شود. سپس کلمات توقف و اطلاعات غیرضروری حذف می‌شوند. و در نهایت برای هر کلمه یک دنباله معنادار ساخته می‌شود. در فاز پردازش یک برآوردی از جملات مرتبط انجام می‌شود و وزن‌ها را با شیوه یادگیری وزنی[۱] (آزمون و خطا) به جملات اختصاص می‌دهد. در نهایت جملات با امتیاز بالا به متن خلاصه اضافه می‌شوند. مشکلات این نوع خلاصه سازی که در مقالات جیمی لین و جکی سی کی [3,4] آورده شده است شامل موارد زیر است:

جملات خلاصه شده طولانی بوده و نیازمند فضای ذخیره سازی بالایی است.

جملات بدست آمده از لحاظ معنایی به هم مرتبط نیستند.

اطلاعات آن چنان دقیقی در خلاصه ارائه نمی‌شود.

در مقابل خلاصه سازی انتزاعی نیازمند فهمی از متن اصلی و تولید خلاصه با توجه به ارتباطات معنایی است در این وضعیت نتایج کوتاه‌تری حاصل می‌شوند که تنها با مشکل نمایش نهایی مواجه هستند.

۲- کارهای مرتبط

توجه و علاقه به خلاصه‌سازی خودکار متن، اولین بار حدود دهه پنجاه به وجود آمد. اولین فعالیت‌ها در این زمان توسط فردی به نام لوهن شروع شد [5]. در ابتدا اساس کار او، یافتن کلمات با بیش‌ترین تکرار بود. از نظر وی کلمات با فرکانس تکرار بیشتر در یک متن مهم تر از سایر کلمات بوده و جملاتی که تعداد بیش تری از این کلمات را دارند مهمترین بخش‌های متن بوده و باید در متن خلاصه شده قرار بگیرند. البته روش اولیه ارائه شده توسط وی خطای زیادی داشت که بعد ها توسط خود او اصلاحاتی بر روی آن انجام شد. به عنوان مثال برخی افعال و حروف اضافه دارای فرکانس بالایی در تمامی متون بوده اما حاوی اطلاعات مهمی نبوده که در ایده قبل او دارای اهمیت محسوب می‌شدند. در اصلاحات بعدی حروف اضافه و برخی از کلمات پرتکرار حذف شده و در الگوریتم استخراجی شرکت داده نمی‌شدند. گرچه که روش پیشنهادی وی دارای دقت خوبی نبود اما به عنوان پایه‌گذار اصلی خلاصه سازی بسیار مورد توجه قرار گرفت. از آن به بعد خلاصه سازی متن یکی از حوزه‌های مهم تحقیقاتی در پردازش زبان طبیعی در نظر گرفته شده و تحقیقات زیادی بر روی آن انجام گرفته است.روش‌های زیادی به همراه ابزارهایی قدرتمند به کار گرفته شده اند تا بتوانند پردازش متن را مانند آنچه که در مغز انسان انجام می‌شود، شبیه‌سازی کنند. ادمنسن [6] از جمله کسانی بود که بعد از وی از سایر ویژگی‌های موجود در متن برای ایجاد خلاصه‌های بهینه استفاده کرد.وی برای

[1] Weight Learning