پنجمین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات
5th National Conference on Advanced Topics in Computer and Information Technology
ششم آبان ماه ۱۴۰۰

ارائه یک روش خلاصه‌سازی متن مبتنی بر تعبیه کلمه و الگوریتم کرم شب‌تاب

محبوبه آرمان فرد [۱] - مرجان عبدیزدان [۲]

(۱) گروه کامپیوتر - دانشگاه آزاد اسلامی واحد ماهشهر
mhbbarman@gmail.com
(۲) گروه کامپیوتر – دانشگاه آزاد اسلامی واحد ماهشهر
abdeyazdan87@yahoo.com

خلاصه: خلاصه‌سازی متن، جملات کلیدی را از متون ورودی، شناسایی و استخراج می‌کند تا خلاصه‌های متنی خودکار را از اسناد ورودی تولید کند. چندین خلاصه‌ساز خودکار در تحقیقات وجود دارند که قادرند خلاصه‌های با کیفیتی را تولید کنند، اما بر حفظ محتوا و معنای متن مورد نظر تمرکز نمی‌کنند. در روش پیشنهادی، معنای متن به‌عنوان یک ویژگی اصلی برای خلاصه کردن یک متن، حفظ و نگهداری می‌شود. یک خلاصه‌ساز خودکار با استفاده از تعبیه کلمه و الگوریتم کرم شب‌تاب ارائه می‌شود تا از طریق تعبیه کلمه، معنا برای تولید خلاصه‌های با کیفیت بالا، حفظ شود. برای شبیه‌سازی روش پیشنهادی از پایتون استفاده شده است. آزمایشاتی با دو مجموعه داده استاندارد DUC 2006 و DUC 2007 انجام شده‌اند. روش پیشنهادی، خلاصه‌هایی ایجاد می‌کند که بسیار نزدیک به مرکز خلاصه‌های مرجع برای هر مجموعه سند است. برای ارزیابی روش پیشنهادی، معیارهای Rouge-1، Rouge-2 و ROUGE-SU4 در آزمایشات انتخاب شده‌اند. روش پیشنهادی نسبت به روش Pv-dbow، توانسته است معیارهای Rouge-1 (۴,۶۱)، Rouge-2 (۷,۰۹) و Rouge su4 (۴,۸۵) را بهبود بخشد. همچنین نسبت به روش DocRebuild، معیارهای Rouge-1 (۳,۹۱)، Rouge-2 (۷,۸۲) و Rouge su4 (۵,۱۱) بهبود یافته‌اند.

کلمات کلیدی: خلاصه‌سازی متن، تعبیه کلمه، الگوریتم کرم شب‌تاب

۱ - مقدمه

با توسعه سریع فناوری اینترنت، حجم متون الکترونیکی در اینترنت با سرعت فوق‌العاده‌ای، افزایش یافته است. در عصر رشد سریع داده‌ها، انسان‌ها می‌توانند اطلاعات را از طریق منابع مختلف، فوراً به دست آورند و به اشتراک بگذارند. اینترنت در حال حاضر، دسترسی به میلیاردها متون را فراهم می‌کند. با افزایش هر ثانیه، رشد نمایی داده‌ها در مدت‌زمان کوتاهی مشاهده می‌شود [۱]. یکی از روش‌های پردازش زبان طبیعی (NLP[۱])، خلاصه‌سازی متن است که جملات مهم را از متون مرتبط، استخراج می‌کند. بسیاری از محققان در کنفرانس TAC[۲] و کنفرانس DUC[۳] به بررسی زبان‌های اروپایی و انگلیسی توجه کرده‌اند [۲]. علیرغم آخرین پیشرفتی که در خلاصه‌سازی متن مشاهده شده است، چالش‌های موجود هنوز به‌طور کامل برطرف نشده‌اند. هدف خلاصه‌سازی متن، یافتن روش‌هایی برای کشف مهم‌ترین اطلاعات در متن و متعاقباً متراکم کردن آن برای سهولت استفاده توسط خوانندگان است [۳]. علاوه بر این، به دست آوردن ویژگی‌های خاص با پردازش داده‌ها به‌منظور کوتاه کردن زمان صرف شده برای دستیابی به اطلاعات، ضروری است [۴]. این موضوعات باعث افزایش علاقه در زمینه سیستم‌های خلاصه‌سازی متن، شده است [۵]. با در نظر گرفتن یک مجموعه متن، یک مجموعه جملات کاندید $S = [s_1, s_2, \dots, s_n]$ ایجاد

می‌شود که شامل همه جملات موجود در همان مجموعه متن می‌باشد. بردارها برای تمامی جملات موجود در مجموعه جملات کاندید با استفاده از مدل تعبیه کلمه، محاسبه می‌شوند. به دلیل استفاده از تعبیه کلمه در روش پیشنهادی، تمامی لغات استفاده شده در یک زبان را می‌توان توسط مجموعه‌ای از اعداد اعشاری (در قالب یک بردار) بیان نمود. تعبیه کلمه، بردارهای n بعدی هستند که سعی دارند معنای لغات و محتوای آن‌ها را با مقادیر عددی خود، ثبت و حفظ کنند. بردارهای عددی نمایانگر کلمات یک لغت‌نامه هستند. تعبیه کلمه، روش پیشنهادی را قادر می‌سازد تا روابط و شباهت‌های بین کلمات را به‌صورت خودکار درک کند. برای پیشرفت بیشتر عملکرد مدل، از الگوریتم کرم شب‌تاب استفاده می‌شود. الگوریتم کرم شب‌تاب برای انتخاب جملات کلیدی برای غلبه بر مشکل افزونگی است. هدف در روش پیشنهادی، دستیابی به یک کد به ازای هر کلمه است و بر اساس این کدها (به آن‌ها تعبیه کلمه گفته می‌شود)، عملیات خلاصه‌سازی متن انجام می‌شود. شبیه‌سازی روش پیشنهادی، از طریق پایتون صورت می‌گیرد. در روش پیشنهادی، از مجموعه داده موجود در پژوهش [۶] استفاده می‌شود. آزمایش‌هایی با دو مجموعه داده خلاصه‌سازی استاندارد یعنی DUC 2006و DUC 2007 انجام می‌گیرد که برای ارزیابی توسط NIST ارائه شده‌اند. DUC 2006 و DUC 2007 به ترتیب دارای ۵۰ و ۴۵ مجموعه متن هستند و هر مجموعه متن شامل ۲۵ مقاله خبری و چهار نوشته انسانی خلاصه شده به‌عنوان شواهد تجربی[۴] است.

۲- پیشینه تحقیق

در [۷]، روش عصبی- فازی به نام ANFIS برای افزایش توانایی سیستم خلاصه‌سازی متن پیشنهاد شده است که این رویکرد، توانسته است برخی از محدودیت‌ها در خلاصه‌سازی متن را کاهش دهد.

در [۶]، چارچوب بازسازی تراز متون بر اساس مدل PV-DBOW[۵] ارائه شده است. مدل ارائه شده نسبت به آخرین فناوری‌های بدون نظارت، بهتر عمل کرده است و پیشرفت‌های چشمگیری در Rouge2 و Rouge SU4 نشان داده است که مزایا و بهبود اصلی روش ارائه شده، هستند.

در [۸]، یک رویکرد از به‌کارگیری تغییرات ساده روش خلاصه‌سازی متن را نشان داده‌اند تا چندین سیستم کاندید ایجاد شوند. از سیستم‌های نامزد برای ارائه یک روش خلاصه جدید استخراجی، با تمرکز بیشتر بر روی پیدا کردن یک اندازه تشابه جمله استفاده شده است.

در [۹]، Gist را معرفی نموده‌اند که به‌طور خودکار می‌تواند حجم زیادی از متن را در جملات کلیدی خلاصه کند. با یادگیری بدون نظارت و تجزیه‌وتحلیل احساسات، Gist جملات را انتخاب کرده است که به بهترین شکل، مجموعه‌ای از بررسی‌ها را توصیف کرده‌اند.

در [۱۰]، خلاصه‌سازی متن را با استفاده از تعبیه کلمه انجام داده‌اند. یک روش خلاصه‌سازی خودکار را با استفاده از مدل معنایی توزیع پیشنهاد کرده‌اند تا معانی را برای تولید خلاصه‌هایی با کیفیت بالا، به دست آورند.

در [۱۱]، خلاصه‌سازی متن و طبقه‌بندی متن بدون نظارت را انجام داده‌اند. در روش ارائه شده، تلاش نموده‌اند نوع خلاصه‌سازی را با استفاده از RNN[۶] پیاده‌سازی کنند.

در [۱۲]، طبقه‌بندی کننده برای خلاصه‌سازی متن بررسی شده‌اند. روش‌های یادگیری ماشین برای مشکلات طبقه‌بندی در حوزه‌های مختلف اجرا شده‌اند. الگوریتم‌های k نزدیک‌ترین همسایه، جنگل تصادفی، ماشین بردار پشتیبان، پرسپترون چند لایه، درخت تصمیم و رگرسیون لجستیک در مجموعه داده‌های Newsroom اجرا شده‌اند.

در [۱۳]، یک رویکرد خلاصه‌سازی متن با استفاده از RNN ارائه شده است. در پژوهش ذکر شده، یک رویکرد "تولید متن" بررسی شده است. یک شبکه عصبی RNN برای این رویکرد تولید متن، استفاده شده است.

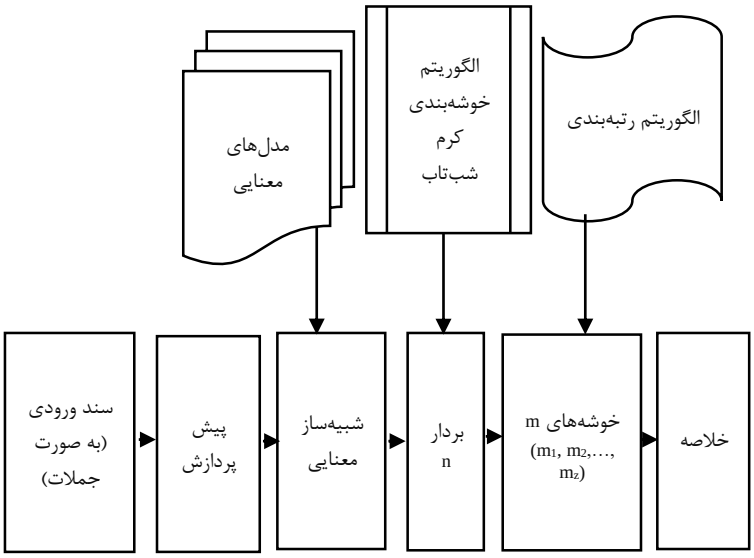
جدول (۱): مرور روش‌های پیشین

منبع	روش	عیب	مزیت
[۷]	روش عصبی- فازی به نام ANFIS برای افزایش توانایی سیستم خلاصه‌سازی متن	تنظیم قوانین از پیش تعریف شده	کاهش برخی از محدودیت‌ها در خلاصه‌سازی متن
[۶]	ارائه چارچوب بازسازی تراز متون بر اساس مدل PV-DBOW	عدم تعمیم آن به یک روش استخراج سطح عبارت مبتنی بر سیستم خلاصه‌سازی	عملکرد بهتر نسبت به آخرین فناوری‌های بدون نظارت
[۸]	نشان دادن به‌کارگیری تغییرات ساده روش‌های خلاصه‌سازی متن	ترکیب‌های ارائه شده در ابتدا، همیشه عملکرد بهتری ندارند.	تأثیر مراحل پیش‌پردازش و مراحل پردازش بر عملکرد
[۹]	خلاصه نمودن خودکار حجم زیادی از متن در جملات کلیدی	عدم استفاده از رویکرد فرامکاشفه‌ای جهت افزایش سرعت	کارآمد بودن برای تجزیه متن در مقیاس بزرگ و داده‌کاوی
[۱۰]	خلاصه‌سازی متن با استفاده از تعبیه کلمه	عدم بهبود نتایج از طریق روش‌های جدید بهینه‌سازی	ارائه نتایج بهتر و کاهش افزونگی‌های منبع
[۱۱]	خلاصه‌سازی متن و طبقه‌بندی متن بدون نظارت	عدم استفاده از بهینه‌سازی چندهدفه برای بهبود چند هدف	طول آن، لزوماً به طول متن منبع، وابسته نیست.
[۱۲]	بررسی طبقه‌بندی کننده‌ها برای خلاصه‌سازی متن	عدم بررسی رویکردهای فرامکاشفه‌ای در مقایسه با روش‌های یادگیری ماشین	آزمایش دسته‌بندهای یادگیری ماشین بر روی یک مجموعه داده
[۱۳]	خلاصه‌سازی متن با استفاده از شبکه عصبی	افزایش بار محاسبات به دلیل آموزش شبکه عصبی	پیش‌بینی بهتر توالی بعدی با استفاده از حافظه

۳- رو پیشنهادی

مدل پیشنهادی، معنای متن را به‌عنوان یک ویژگی در کنار ویژگی‌های آماری و ادبی استفاده می‌کند. شکل (۱)، معماری مدل پیشنهادی را نشان می‌دهد. به صورت دقیق‌تر، روش پیشنهادی شامل گام‌های زیر است:

- انجام پیش‌پردازش متن برای نرمال‌سازی متن و حذف ناسازگاری‌ها
- حفظ معنای متون با استفاده از مدل‌های معنایی توزیع‌شده
- استفاده از خوشه‌بندی برای گروه‌بندی معنایی جملات مشابه در خوشه‌های مشترک
- استفاده از الگوریتم رتبه‌بندی برای به دست آوردن امتیاز هر جمله از هر خوشه
- نرمال‌سازی امتیازها برای استخراج مؤثر جمله و به دست آوردن خلاصه



شکل (۱): عملکرد کلی سیستم

۳-۱- پیش‌پردازش

پیش‌پردازش، ناسازگاری‌های اطلاعات را حذف می‌کند. در نتیجه، یک داده نرمال شده را تولید می‌نماید. پیش‌پردازش، اولین مرحله در سیستم پیشنهادی است. در روش پیشنهادی، داده‌ها به خلاصه‌ساز وارد می‌شوند، سپس آن‌ها، پیش‌پردازش می‌گردند و به فرمت مناسبی برای الگوریتم خلاصه‌ساز و پردازش تبدیل می‌شوند. جزییات شامل مراحل زیر هستند: URL های موجود در متن ورودی توسط ماژول پیش‌پردازش حذف می‌گردند. تمام حروف متن ورودی توسط ماژول پیش‌پردازش به حروف کوچک تبدیل می‌شوند. کلمات توقف در عمل خلاصه‌سازی بی‌معنا هستند، بنابراین حذف می‌گردند. کلمات توقف با استفاده از بسته Stanford core NLP حذف می‌شوند. هر جمله در متن ورودی، پیش از پردازش بیشتر توسط کلمات، علامت‌گذاری می‌گردد. کلمات با استفاده از بسته Stanford core NLP کاهش می‌یابند.

۳-۲- حفظ معنا با استفاده از مدل‌های معنایی توزیعی

در روش پیشنهادی، برای حفظ معنای متن، از الگوریتم‌های معنایی توزیعی استفاده می‌شود، زیرا آن‌ها عمومی هستند و به آنالیز زبانی و لغوی نیاز ندارند. همچنین، این مدل‌ها از منابع خارجی برای به دست آوردن مبنای اطلاعات معنایی مستقل هستند. از مدل‌های معنایی توزیعی استفاده می‌شود تا سازگاری معنایی میان دو المان متنی حفظ شود. این مدل‌ها بر مبنای فرضیات توزیع‌شده هستند. فرضیات توزیع‌شده بیان می‌کنند که کلمات موجود در متن، اغلب معنای مشابهی دارند. بنابراین، می‌توان معنای آن‌ها را بر اساس کاربردشان استنباط نمود. علاوه بر این، مدل‌های توزیعی درج معنا با استفاده از محاسبات آماری محتوا برای وقوع کلمات تولید می‌شوند. در نتیجه، برای هر کلمه، بردارهای مقدار حقیقی با ابعاد زیاد محاسبه می‌گردند و به صورت تعبیه کلمه یا بردارهای کلمات محاسبه و ارائه می‌شوند. این نمایش در کنار ویژگی‌های هندسی در فضاهای برداری با ابعاد بیشتر از نظر گرامری و معنایی سودمند هستند تا سازگاری میان کاربردهای کلمات مختلف پیدا شود. بنابراین، اگر کلمات در فضاهای برداری با ابعاد زیاد به یکدیگر نزدیک باشند، از نظر معنایی و گرامری مشابه هستند. برای به دست آوردن شباهت توزیع، از Word2Vec استفاده می‌شود، که یک مدل معنایی توزیعی است که معنا را حفظ می‌کند. کاربرد Word2Vec برای حفظ شباهت معنایی متن، سودمند و مناسب است، زیرا در روش‌های مختلف استفاده شده است. Word2Vec یک شبکه عصبی هستند که متن را پردازش می‌کنند. مجموعه‌ای از بردارها را به‌عنوان خروجی برای متن ورودی تولید می‌نماید. بردارهای تولید شده توسط Word2Vec، بردارهای ویژگی کلمات هستند. الگوریتم Word2Vec به صورت یک فضای برداری از مؤلفه‌های استخراج شده از دو لایه شبکه عصبی آموزش داده می‌شوند. Word2Vec دو معماری دارد، Skip-gram و CBOW (بسته پیوسته‌ای از کلمات). این معماری‌ها نحوه ساخت کلمات درج شده توسط شبکه عصبی برای کلمات مختلف را توضیح می‌دهند. در حوزه CBOW، کلمات فعلی پیش‌بینی می‌شوند و Skip-gram، محتوای کلمات را پیش‌بینی می‌کند. در روش پیشنهادی از مدل Skip-gram در خلاصه‌ساز خودکار استفاده می‌شود. در روش پیشنهادی از مدل پیش آموزش Word2Vec که توسط پایگاه داده اخبار google آموزش داده شده است، بهره برده می‌شود. روش جدید اندازه‌گیری شباهت معنایی با استفاده از بردارهای BV[۷] برای خلاصه‌سازی متن در بخش بعدی معرفی می‌شود.

۳-۳- تولید بردار BV و تعبیه کلمه

بردار BV یک جمله، توسط زنجیره‌ای از کلمات مشابه به دست آمده به وسیله مدل پیش‌آموزش Word2Vec تولید می‌شود. فرض می‌شود $\varphi(\cdot)$ تابعی باشد که لیست زنجیره‌ای از m کلمه مشابه، به صورت $\omega' = \varphi(\omega) = \omega'_1 \oplus \omega'_2 \oplus \dots \oplus \omega'_m$ است. برای جمله‌ای به صورت $W = \{w_1, w_2, w_3, \dots, w_k\}$، یعنی یک جمله با k کلمه علامت‌گذاری شده، یک بردار BV توسط زنجیره‌ای از m کلمه مشابه

برای هر کلمه پُر می‌شود، یعنی $BV = \{\varphi_1(\omega_1)\oplus\varphi(\omega_2)\oplus \dots \oplus\varphi(\omega_k)\}$. بردارهای BV بیانگر اطلاعات معنایی غنی از یک جمله هستند و بر اساس فرضیه توزیع تشکیل می‌شوند. به‌طور خاص، از نظر معنایی مشابه نمایش بسته کلمات از یک جمله است که حاوی کلمات مشابه از نظر معنایی می‌باشد. بردارهای BV تمام جملات موجود در یک سند به دست می‌آیند. چون تعداد کلمات در یک جمله تغییر می‌کنند؛ بنابراین، بردارهای BV با اندازه متفاوتی تولید می‌گردند. برای غلبه بر این مسئله، اندازه بردارهای BV به اندازه n بُعد ثابت در نظر گرفته می‌شوند. در نتیجه، برای جملات کوتاه‌تر پُر شده و در جملات طولانی‌تر، طول به n محدود می‌گردد. تعبیه کلمه برای یک نمایش برداری پیوسته قادر است اطلاعات معنایی و گرامری یک کلمه را حفظ کند. چندین روش برای انجام تعبیه کلمه پیشنهاد شده است، که از فرضیه توزیع پیروی می‌کنند. در این تحقیق، از دو مدل استفاده می‌شود که شامل bag-of-word پیوسته و skip-gram هستند. این مدل‌ها با استفاده از یک مدل زبان شبکه عصبی یک نمایش برداری را برای هر کلمه آموزش می‌بینند و می‌توانند به صورت کارآمدی، میلیاردها کلمه را یاد بگیرند. Word2Vec، این امکان را فراهم می‌کند که با استفاده از عملگرهای برداری ساده مانند $vec(king) - vec(man) + vec(woman) \approx vec(queen)$ و $vec(Barrett) - vec(singer) + vec(guitarist) \approx vec(Gilmour)$، روابط معنایی پیچیده را یاد بگیرد. هر چند که این روش، عمومی است و سایر روش‌ها برای ساخت تعبیه کلمه می‌توانند استفاده شوند. برای ساخت یک بردار مرکزی با استفاده از تعبیه کلمه، ابتدا کلمات با معنی در متن انتخاب می‌شوند. برای سادگی و مقایسه مناسب با روش اصلی، کلماتی که از وزن $tf * idf$ بیشتری نسبت به آستانه عنوان برخوردارند، انتخاب می‌شوند.

۳-۴- خوشه‌بندی

هنگامی که نمایش جملات به فرم بردارهای BV انجام شد، از الگوریتم‌های خوشه‌بندی استفاده می‌شود تا جملاتی که از نظر معنایی، مشابه هستند، گروه‌بندی شوند. سپس، از الگوریتم رتبه‌بندی برای بازیابی و استخراج جملات مهم هر گروه استفاده می‌شود تا خلاصه استخراجی به دست آید. از الگوریتم بهینه‌سازی کرم شب‌تاب استفاده می‌شود تا خوشه‌هایی از بردارهای BV از نظر شباهت معنایی تشکیل شوند. این بردارهای BV از طریق TF-IDF و Token-weights جهت‌دهی می‌شوند و بردارهای BV جهت‌دار به‌عنوان ورودی به الگوریتم بهینه‌سازی کرم شب‌تاب وارد می‌شوند. سپس خوشه‌بندی انجام شده و خوشه‌های مختلف بر اساس شباهت معنایی تشکیل می‌گردند. الگوریتم کرم شب‌تاب از سه قانون پیروی می‌نماید:

- کرم‌های شب‌تاب از یک جنس قادر به جذب یکدیگر به وسیله شب‌تابی هستند.
- عامل جذابیت با توجه به روشنایی، همان‌طور که کرم به طرف کرم‌های شب‌تاب دیگر حرکت می‌کند، مد نظر قرار می‌گیرد.
- این روشنایی از طریق یک تابع هدف محاسبه می‌گردد.

شدت نور و جذابیت، دو متغیر اصلی در الگوریتم کرم شب‌تاب هستند. کرم شب‌تاب به طرف کرم شب‌تاب دیگری جذب می‌گردد که دارای تابش‌های نورانی‌تر نسبت به خودش است. جذابیت، بستگی به شدت نور دارد. شدت نور تابشی در شب، جذاب تعریف می‌گردد و دارای نسبت معکوسی با فاصله r از منبع نور است. این مسئله، کاهش جذابیت نسبت به فاصله را نشان می‌دهد و بالعکس که مطابق با رابطه (۱) است. I در رابطه (۱)، شدت نور تابشی را نشان می‌دهد. I_0، شدت نور تابشی اولیه است. $\tilde{a}$، نشانگر ضریب جذب نور می‌باشد و r، فاصله بین کرم شب‌تاب i و j است. شدت نور تابشی، متناسب با جذابیت می‌باشد و از طریق β مشخص می‌گردد که بر طبق رابطه (۲) است. β_0 ، نشان می‌دهد که جذابیت در r، صفر است. رابطه فاصله‌ای (۳)، به منظور محاسبه فاصله میان دو کرم شب‌تاب به کار گرفته می‌شود.

$$I(r) = I_0 e^{-\partial r^2} \quad (۱)$$

$$\beta = \beta_0 e^{-\partial r^2} \quad (۲)$$

$$r_0 = |x_i - x_j| = \sqrt{\sum_{k-1}^{d} \left(x_{i,k} - x_{j,k}\right)^2} \quad (۳)$$

حرکت کرم شب‌تاب i با جذابیت کمتر نسبت به کرم شب‌تاب j با جذابیت بیشتر، مطابق با رابطه (۴) مشخص می‌گردد. در رابطه (۴)، چنانچه ضریب جذب نوری به صفر نزدیک گردد، جذابیت تقریباً ثابت می‌باشد؛ در غیر این صورت چنانچه ضریب به بی‌نهایت نزدیک شود، جذابیت کاهش پیدا می‌کند. عبارت دوم در رابطه (۴)، به منظور موارد تصادفی مورد استفاده قرار می‌گیرد. در الگوریتمی که در بالا توصیف شده است، شدت نور توسط انرژی باقی‌مانده یک گره، جایگزین می‌شود، بنابراین جذابیت با انرژی، متناسب می‌باشد. $r_{i,j}$، فاصله میان دو گره x_i وx_j است.

$$\Delta x_i = \beta_0 e^{-\partial r^2}\left(x_j^t - x_i^t\right) + \beta\beta_i, x_i^{t+1} + \Delta x_i \quad (۴)$$

۳-۵- الگوریتم رتبه‌بندی

سپس از الگوریتم رتبه‌بندی برای رتبه‌بندی جملات هر خوشه استفاده می‌شود تا خلاصه استخراجی به دست آید. الگوریتم رتبه‌بندی، از ویژگی‌های آماری مختلف استفاده می‌کند تا n جمله برتر هر خوشه را استخراج نماید. ابتدا، الگوریتم رتبه‌بندی، امتیاز رتبه‌بندی هر جمله در یک خوشه را به دست می‌آورد، سپس این امتیازها، نرمال‌سازی شده و امتیاز نهایی برابر مجموع امتیاز نرمال‌سازی شده است. ویژگی‌های آماری به کار رفته توسط الگوریتم رتبه‌بندی عبارت‌اند از:

- طول جمله: اهمیت جمله به صورت مستقیم متناسب با طول جمله است. بنابراین، از طول جمله به‌عنوان یک ویژگی برای رتبه‌بندی جملات استفاده می‌شود. خلاصه‌ساز پیشنهادی از طول جمله به‌عنوان یک ویژگی آماری استفاده می‌کند تا اهمیت جمله را برای رتبه‌بندی محاسبه نماید. طول یک جمله، $[S_i]$ به صورت تعداد کلمات پس از پیش‌پردازش تعریف می‌شود.
- موقعیت جمله: جملات مهم‌تر در ابتدا و انتهای متن ورودی هستند. بنابراین، موقعیت جمله، یک ویژگی مهم برای رتبه‌بندی

جملات است. مهم‌ترین جملات، در آغاز متن ورودی هستند. امتیاز موقعیت جمله به صورت رابطه (۵) محاسبه می‌شود:

$$s_i^p = 1 - \frac{s_i - 1}{|S|} \quad (۵)$$

به‌طوری‌که، s_i^p امتیاز موقعیت جمله iام در متن ورودی S است و $|S|$، تعداد جملات در متن ورودی می‌باشد.

- فرکانس (TF-IDF): TF-IDF، یک ویژگی مهم و مشخص در هر سیستم خلاصه‌ساز متن است. TF-IDF، مهم‌ترین مؤلفه‌های هر متن را شناسایی می‌کند. الگوریتم رتبه‌بندی از این ویژگی استفاده می نماید تا مهم‌ترین کلمات در یک سند متنی و جملات شناسایی شوند. الگوریتم رتبه‌بندی، TF-IDF کلمات منحصربه‌فرد را محاسبه کرده و از کلمات منحصربه‌فرد برای محاسبه امتیاز TF-IDF جملات بهره می‌برد. به صورت دقیق، TF-IDF جمله برابر است با مجموع امتیاز TF-IDF کلمات در جمله. TF-IDF یک جمله s_i به صورت رابطه (۶) محاسبه می‌گردد:

$$s_i^{tf} = \sum_{w \in s_i} t_f(w) \quad (۶)$$

به‌طوری‌که $t_f(w)$، تابعی است که امتیاز TF-IDF کلمه w را به دست می‌آورد.

- عبارت اسمی و عبارت فعلی: یک جمله، حاوی عبارت اسمی و فعلی، مهم‌ترین جمله در متن ورودی است. جمله‌ای که حاوی یکی از دو عبارت باشد، یک جمله امری است و توسط الگوریتم رتبه‌بندی، رتبه بیشتری به دست می‌آورد. خلاصه‌ساز پیشنهادی، از Stanford POS tagger برای شناسایی عبارت‌های اسمی و فعلی در متن استفاده می‌کند و تگ‌های POS را به عبارات اسمی و فعلی اختصاص می‌دهد. پس از شناسایی این دو، از NVC (شمارنده اسم فعل) برای هر جمله استفاده می‌شود تا تعداد عبارت‌های اسم و فعل محاسبه شود. تعداد NVC بیشتر، موجب افزایش رتبه جمله می‌شود.
- نام مناسب: نام‌های مناسب، شامل مراجع مستقیم برای یک موضوع هستند، بنابراین حضور آن‌ها در جمله، آن را مهم‌تر می‌سازد. الگوریتم‌های رتبه‌بندی، رتبه‌های بیشتری را به جملات دارای نام‌های مناسب اختصاص می‌دهند. Stanford PoS tagger برای استخراج نام مناسب از متن ورودی به کار می‌رود.
- تجمیع شباهت کسینوس: شباهت کسینوس، برای محاسبه رابطه میان دو سند به کار می‌رود. از شباهت کسینوس به‌عنوان یک ویژگی در الگوریتم رتبه‌بندی استفاده می‌شود. کسینوس میان دو جمله محاسبه می‌گردد. به صورت دقیق‌تر، شباهت کسینوس بیشتر میان دو جمله، موجب دادن رتبه بیشتر به آن جمله‌ها می‌شود. میانگین شباهت کسینوس s_i^c در جمله iام به صورت رابطه (۷) است:

$$s_i^c = \frac{\sum_{j=1, j \neq i}^{|S|} c(s_i, s_j)}{|S|} \quad (۷)$$

به‌طوری‌که، $c(s_i, s_j)$ شباهت کسینوس میان دو جمله را بیان می‌کند.

- عبارات اشاره: عبارات اشاره، مؤلفه‌های ارتباطی میان جملات هستند. اگر یک جمله در آغاز یک عبارت اشاره داشته باشد، بیانگر این است که آن به جمله قبلی وابسته بوده و این جمله باید در خلاصه باشد.

۴- نتایج شبیه سازی

در شبیه‌سازی روش پیشنهادی، برای آموزش، از ۷۵ درصد اسناد کل در داده، استفاده شده است. یک سرور با پردازنده TeslaV100 استفاده شده است. در شبیه‌سازی، تابع خطا MAE فرض می‌شود. برای اجرا باید مدل‌های پایه گرفته و ساخته شوند. برای شبیه‌سازی روش پیشنهادی از پایتون[۸] استفاده شده است. در این پژوهش، آزمایش‌هایی با دو مجموعه داده استاندارد معیار خلاصه DUC 2006 و DUC 2007 ارائه می‌شود. DUC 2006 و DUC 2007 به ترتیب شامل ۵۰ و ۴۵ مجموعه سند هستند. هر مجموعه سند شامل ۲۵ مقاله خبری و ۴ خلاصه نوشته شده توسط انسان است. طول خلاصه به ۲۵۰ کلمه محدود می‌شود. در جدول (۲)، پارامترهای کرم شب‌تاب برای ساخت مدل مشخص شده‌اند. جهت شبیه‌سازی روش پیشنهادی، رشته ورودی، ماتریس بردار کلمات به طول حداکثر ۱۰ می‌باشد. جمعیت، ۸۰ و تعداد تکرار، ۱۰۰، است. شعاع جذب، ۰,۷ می‌باشد. تابع هدف، کمینه‌سازی فاصله بین دو سری کلمه (جمله) است.

جدول (۲): پارامترهای کرم شب‌تاب برای ساخت مدل

پارامتر	مقدار
رشته ورودی	ماتریس بردار کلمات به طول حداکثر ۱۰
جمعیت	۸۰
تکرار	۱۰۰
شعاع جذب	۰,۷
تابع هدف	کمینه‌سازی فاصله بین دو سری کلمه (جمله)

۴-۱- معیار ارزیابی مورد استفاده

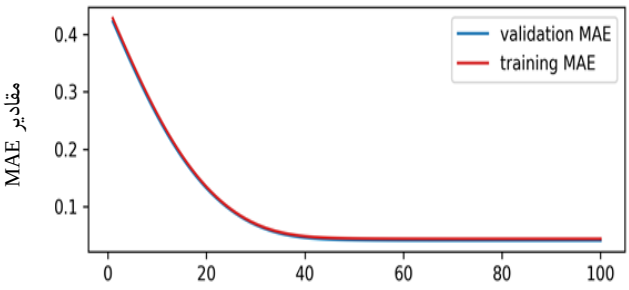
معیار ROUGE از جمله معروف‌ترین ابزارهای ارزیابی در خلاصه‌سازی خودکار می‌باشد. از آن در کاربردهای پردازش زبان طبیعی و بازیابی اطلاعات استفاده شده است. معیار ROUGE به "معنای ارزیابی مبتنی بر یادآوری برای خلاصه" است. ابزار ذکر شده، معیارهایی برای ارزیابی خلاصه‌سازی متن، مثل ترجمه ماشین را دارد. با مقایسه خلاصه استخراجی (خلاصه‌های تولیدشده به صورت خودکار) و خلاصه چکیده‌ای (خلاصه‌های تولیدشده توسط انسان) ارزیابی انجام می‌گردد. کلمات مشترک به تنهایی معیار ارزیابی نمی‌باشند؛ پس بهتر است از معیارهایی مثل $Precision$ و $Recall$ استفاده شود. معیار $Recall$ به صورت رابطه (۸) محاسبه می‌گردد. معیار $Precision$ به صورت رابطه (۹) محاسبه می‌گردد.

$$Recall = \frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary} \quad (۸)$$

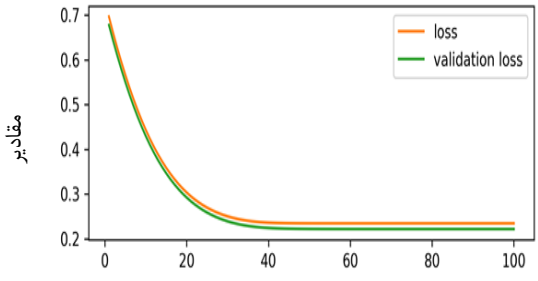
$$Precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary} \quad (۹)$$

۴-۲- نتایج و مقایسه روش پیشنهادی

مدل اول، مربوط به مدل زبانی است. نمودار (۱)، نشان‌دهنده آموزش مدل زبانی است که از bert multilingual استفاده شده است. داده انگلیسی در آن حوزه‌ای که مجموعه داده‌ها هستند، آماده شده‌اند و فرمت آن‌ها به سیستم داده شده است و فاین تیون شده است تا سیستم از فرمت داده‌ای اطلاع داشته باشد. کار مدل به این شکل است که یک جمله مشخص می‌شود و برای هر کلمه با توجه به آن جمله، خروجی گرفته می‌شود. به این معنا که یک خروجی یا امبدینگ تولید می‌کند. این امبدینگ اندازه بردار ۳۰۰ تایی برای این کلمه‌ها می‌دهد. منظور از امبدینگ یا بردار جاسازی شده این است که به طور مثال در فضایی که کلمه "زیبا" مشاهده شود، برای آن، یک امبدینگ می‌دهد و برای کلمه "قشنگ" نیز یک امبدینگ می‌دهد. این دو امبدینگ عددهایشان نزدیک به هم می‌باشد؛ درحالی‌که ظاهر آن‌ها شبیه همدیگر نیست که این امر، یک مورد مفهومی گفته می‌شود. از طرف دیگر، اگر مجدداً کلمه "زیبا" داده شود، همان بردار را تکرار می‌کند و فاصله، صفر می‌شود. نمودار (۱) نشان می‌دهد که تابع هزینه یا loss درست شده است. سپس مطابق با نمودار (۲)، ساخت مدل خلاصه‌ساز به کمک روش پیشنهادی انجام شده است.

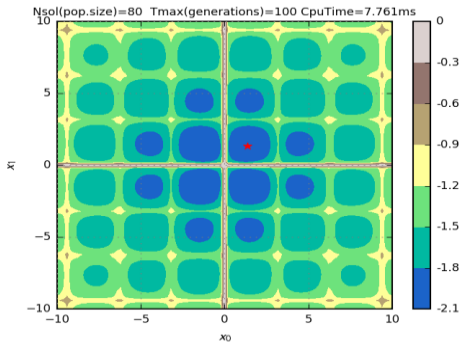


نمودار (۱): مرحله بردار تعبیه کلمات به کمک bert multilingual



نمودار (۲): نمودار ساخت مدل خلاصه‌ساز به کمک روش پیشنهادی

در شکل (۲)، نتایج ضربدری مشابهت میانگین جملات نشان داده شده‌اند. برای نشان دادن اثربخشی روش پیشنهادی، به طور تصادفی پنج مجموعه سند از مجموعه داده‌های DUC 2006 انتخاب شده‌اند و بردارهای خلاصه‌های مرجع و مدل تولید شده، محاسبه شده است. برای هر مجموعه سند، اسناد به همراه خلاصه تولید شده توسط سیستم و مرکز آن، ترسیم شده است. در شکل (۲)، هر رنگ مربوط به یک مجموعه اسناد است، خلاصه‌های تولید شده توسط سیستم با (x) و خلاصه‌های مرجع با (+) نشان داده می‌شوند. از شکل (۲) می‌توان دریافت که روش پیشنهادی، خلاصه‌هایی ایجاد می‌کند که بسیار نزدیک به مرکز خلاصه‌های مرجع برای هر مجموعه سند است. در شکل (۳)، نمونه نتایج روش پیشنهادی نشان داده شده است.



شکل (۲): نتایج ضربدری مشابهت میانگین جملات

ورودی

schizophrenia patients whose medication could n't stop the imaginary voices in their heads gained some relief after researchers repeatedly sent a magnetic field into a small area of their brains.

scientists trying to fathom the mystery of schizophrenia say they have found the strongest evidence to date that the disabling psychiatric disorder is caused by gene abnormalities, according to a researcher at two state universities.

a yale school of medicine study is expanding upon what scientists know about the link between schizophrenia and nicotine addiction.

exploring chaos in a search for order, scientists who study the reality-shattering mental disease schizophrenia are becoming fascinated by the chemical environment of areas of the brain where perception is regulated.

خروجی مورد انتظار

Magnetic treatment may ease or lessen occurrence of schizophrenic voices.

Evidence shows schizophrenia caused by gene abnormalities of Chromosome 1.

Researchers examining evidence of link between schizophrenia and nicotine addiction.

Scientists focusing on chemical environment of brain to understand schizophrenia.

Schizophrenia study shows disparity between what's known and what's provided to patients.

خروجی سیستم

Magnetic image aid of schizophreni.

Evidence of schizophrenia caused by gene abnormalities of Chromosome.

Researchers help of link between schizophrenia and addiction.

Scientists look on chemical of brain to understand schizophrenia.

Schizophrenia shows difference between what's known provided to patients.

شکل (۳): نمونه نتایج روش پیشنهادی

در این پژوهش مطابق با مقاله پایه، معیارهایROUGE[۹] اجرا شده‌اند. ROUGE کیفیت خلاصه را با شمارش واحدهای همپوشانی مانند توالی کلمات n گرم و جفت کلمات بین خلاصه تولید شده (تولید شده توسط الگوریتم‌ها) و مدل، اندازه‌گیری می‌شود. در روش پیشنهادی، معیارهای Rouge-1، Rouge-2 و ROUGE-SU4 در آزمایشات انتخاب شده‌اند. به طور رسمی، ROUGE-N یک یادآوری n گرم و ROUGE-SU4 یک UPSB[۱۰] با حداکثر فاصله پرش چهار بین یک خلاصه تولید شده از سیستم و یک مجموعه خلاصه مدل است. معیار ROUGE-N بر اساس مقایسه‌ی N تایی مشترک کلمات بین خلاصه استخراجی و خلاصه‌ی چکیده‌ای است. از ROUGE-1و ROUGE-2 به ترتیب برای مقایسه‌ی یک‌تایی و دوتایی مشترک کلمات بین خلاصه‌ی استخراجی و خلاصه‌ی چکیده‌ای استفاده می‌شود. نتایج برای همه آزمایشات انجام شده در مجموعه داده مورد نظر نشان داده شده است. همان‌طور که در جدول (۳) نشان داده شده است، Pv-dbow و DocRebuild عملکرد پایین‌تری نسبت به روش پیشنهادی داشته‌اند. آموزش مدل‌های مبتنی بر شبکه عصبی در مجموعه داده‌های کوچک، دشوار است. برای این منظور، ابتدا روش پیشنهادی بر روی مجموعه داده آموزش شده است و سپس بر روی مجموعه داده‌های DUC 2006 و DUC 2007 اجرا شده است. در روش پیشنهادی، مشابه مقاله پایه و آموزش مدل PV-DBOW، از کتابخانه gensim2 در پایتون استفاده شده است. مطابق با مطالب بیان شده، در جدول (۳)، مقایسه نتایج روش پیشنهادی با روش‌های پیشین انجام شده است.

جدول (۳): مقایسه نتایج روش پیشنهادی با روش‌های پیشین بر اساس F-measure

روش	Rouge-1	Rouge-2	Rouge su4
Pv-dbow [۶]	42.7	11.23	16.5
DocRebuild [۱۴]	43.4	10.5	16.24
روش پیشنهادی	47.31	18.32	21.35

مطابق با نتایج ارائه شده در جدول (۳)، DocRebuild، با استفاده از چارچوب بازسازی سطح سند، عملکرد مناسبی دارد و PV-DBOW نیز به عملکرد مناسبی دست یافته است؛ اما مدل اصلی روش پیشنهادی، از روش‌های Pv-dbow و DocRebuild، بهتر عمل کرده است. دیده می‌شود که بهبود امتیازات Rouge-2 و Rouge-SU4 در مقایسه با امتیازات Rouge-1 قابل‌توجه‌تر است. امتیازات بالاتر Rouge-2 و Rouge-SU4، نشان می‌دهد که روش پیشنهادی که مبتنی بر تعبیه سازی کلمه و کرم شب‌تاب است، توانایی بیشتری در مدیریت n گرم نسبت به کلمات دارد.

۵- نتیجه گیری

در این پژوهش، یک خلاصه‌ساز خودکار با استفاده از تعبیه کلمه و الگوریتم کرم شب‌تاب ارائه شده است. برای شبیه‌سازی روش پیشنهادی از پایتون استفاده شده است. آزمایش‌هایی با دو مجموعه داده استاندارد DUC 2006 و DUC 2007 انجام شده‌اند. روش پیشنهادی، خلاصه‌هایی ایجاد می‌کند که بسیار نزدیک به مرکز خلاصه‌های مرجع برای هر مجموعه سند است. برای ارزیابی روش پیشنهادی، معیارهای Rouge-1، Rouge-2 و ROUGE-SU4 در آزمایشات انتخاب شده‌اند. روش پیشنهادی، از روش‌های Pv-dbow و DocRebuild، بهتر عمل کرده است. برای پژوهش‌های آتی می‌توان از ایده‌های زیر، استفاده نمود:

- استفاده از روش‌های یادگیری عمیق بجای رویکرد تعبیه کلمه جهت بهبود فرایند آموزش.
- ترکیب الگوریتم‌های فرا مکاشفه‌ای برای بهبود نتایج بهینه‌سازی.

مراجع

[1] Uçkan, T., & Karcı, A. (2020). Extractive multi-document text summarization based on graph independent sets. Egyptian Informatics Journal.

[2] Mallick, C., Das, A. K., Dutta, M., Das, A. K., & Sarkar, A. (2019). Graph-based text summarization using modified TextRank. In Soft Computing in Data Analytics (pp. 137-146). Springer, Singapore.

[3] Ermakova, L., Cossu, J. V., & Mothe, J. (2019). A survey on evaluation of summarization methods. Information Processing & Management, 56(5), 1794-1814.

[4] Hark, C., Myo, A., Seyyarer, A., Myo, G., Uçkan, T., Myo, B., & Karci, A. (2017, September). Doğal dil İşleme yaklaşimlari ile yapisal olmayan dökümanlarin benzerliği. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-6). IEEE.

[5] Yao, J. G., Wan, X., & Xiao, J. (2017). Recent advances in document summarization. Knowledge and Information Systems, 53(2), 297-336.

[6] Mani, K., Verma, I., Meisheri, H., & Dey, L. (2018, December). Multi-document summarization using distributed bag-of-words model. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 672-675). IEEE.

[7] Azhari, M., & Jaya Kumar, Y. (2017). Improving text summarization using neuro-fuzzy approach. Journal of Information and Telecommunication, 1(4), 367-379.

[8] Mehta, P., & Majumder, P. (2018). Effective aggregation of various summarization techniques. Information Processing & Management, 54(2), 145-158.

[9] Lovinger, J., Valova, I., & Clough, C. (2019). Gist: general integrated summarization of text and reviews. Soft Computing, 23(5), 1589-1601.

[10] Mohd, M., Jan, R., & Shah, M. (2020). Text document summarization using word embedding. Expert Systems with Applications, 143, 112958.

[11] Shrivastava, A., & Bilgaiyan, S. (2020). Abstractive Text Summarization and Unsupervised Text Classifier. In Machine Learning and Information Processing (pp. 355-365). Springer, Singapore.

[12] Pattanaik, A., Mishra, S. S., & Das, M. (2020). A Comparative Study of Classifiers for Extractive Text

Summarization. In Machine Learning and Information Processing (pp. 173-181). Springer, Singapore.

[13] Abujar, S., Masum, A. K. M., Islam, M. S., Faisal, F., & Hossain, S. A. (2020). A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN. In Innovations in Computer Science and Engineering (pp. 509-518). Springer, Singapore.

[14] Ma, S., Deng, Z. H., & Yang, Y. (2016, December). An unsupervised multi-document summarization framework based on neural document model. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers . pp. 1514-1523.

پانویس

[1] Natural Language Processing (Nlp)
[2] Text Analysis Conference (Tac)
[3] Document Understanding Conference (Duc)
[4] Ground Truth
[5] Present A Document Level Reconstruction Framework Based On Distributed Bag Of Words Model
[6] Recurrent Neural Network
[7] Big-Vector (Bv)
[8] Python
[9] Understudy-Recall-Oriented For Gisting Evaluation (Rouge)
[10] Unigram Plus Skip-Bigram (Upsb)