مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۶، زمستان ۱۳۹۸، ص ۵۴- ۳۸

ISSN: 2745-3673
http://www.Rcsj.ir

مجلـه‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۶ / زمستان ۱۳۹۸ / ص ۵۴- ۳۸

ارزیابی همزمان انسجام محلی و عمومی متن با بکارگیری ویژگی‌های آماری

محمد عبدالهی[۱]، الهه کفشی تقی آبادی[۲]

[۱] گروه کامپیوتر دانشگاه جامع علمی کاربردی، جهاد دانشگاهی مشهد.

[۲] گروه کامپیوتر دانشگاه جامع علمی کاربردی، جهاد دانشگاهی مشهد.

نام نویسنده مسئول:
محمد عبدالهی

چکیده

انسجام اجزای متن یکی از مهمترین ویژگی‌های آن بوده که آن را با مجموعه‌ای از جملات بدون هدف متمایز می‌کند. عدم انسجام یکی از مشکلات تمام سیستم‌های پردازش متن و نوشته‌های افراد با دانش نگارشی پایین بوده و همه آنان تلاش دارند با ارزیابی آن عملکرد سیستم خود را بهبود بخشند. بیشتر روش‌های قبل تکیه بر ارزیابی محلی انسجام داشته و انسجام عمومی متن را با دقت پایین‌تری سنجیده‌اند. در این مقاله با ارزیابی انسجام محلی در سطح درون پاراگرافی و انسجام عمومی در سطح ارتباط موضوعی پاراگراف‌های متوالی عمل سنجش انسجام عمومی و محلی را به طور همزمان انجام داده است. این روش با تبدیل جملات به ماتریس‌های عددی و ارزیابی فاصله آنان در یک پنجره وابستگی موضوعی آنان به هم را محاسبه کرده است. نتایج حاصل از روش پیشنهادی بر روی متن‌های داستانی و بلند برتری محسوسی را در مقایسه با رویکردهای مبتنی بر نهاد نشان می دهد.

واژگان کلیدی: پردازش زبان طبیعی، ارزیابی انسجام متن، فضای بردار واژه، ارزیابی انسجام محلی، ارزیابی انسجام عمومی.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۶، زمستان ۱۳۹۸، ص ۵۴- ۳۸

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

ارزیابی انسجام متن از حوزه‌های مهم در پردازش زبان طبیعی بوده که اخیرا به یکی از موضوعات مورد علاقه بسیاری از پژوهشگران تبدیل شده است. کاهش انسجام متن به دو صورت نگارش آن افراد با دانش پایین و یا عملیات پردازشی سیستم‌های ماشینی پردازش متن اتفاق می‌افتد. مهمترین متن‌های نگارشی غیر منسجم نوشته‌های دانش‌آموزان و دانشجویان، متن‌های ترجمه شده توسط مترجمان با دانش ترجمه پایین، متن‌های خلاصه شده و نوشته‌های ترکیبی و کپی شده از منابع مختلف است. بیشترین متن‌های غیر منسجم ماشینی خروجی سیستم‌های خلاصه سازی ماشینی [۱] [۲] [۳] [4]، تولید کننده متن [5]، ساده سازی متن، ترجمه ماشینی [6] [7] [8] [9]، آنالیز احساس [10] [11] [12] [13]، متن کاوی و شبکه‌های اجتماعی [14]، سیستم‌های پرسش و پاسخ، دسته بندی و خوشه بندی اسناد متنی [15] [16]، سیستم‌های جستجوگر متن [17] و سیستم های امتیازدهی خودکار مقالات [۱۸] [۱۹] [۲۰] [۲۱] [۲۲] بوده که اغلب در خروجی از انسجام موضوعی اولیه برخوردار نیستند. از این رو تمام سیستم‌های ماشینی اشاره شده تمایل داشته تا پس از اعمال رویکرد پردازشی خود میزان انسجام متن خروجی را سنجیده تا در صورت نامطلوب بودن آن، الگوریتم پردازشی خود را بهبود دهند. تولید متن منسجم، از ابتدای معرفی رویکردهای ماشینی خلاصه سازی متن توسط لون در سال ۱۹۵۸ مورد توجه قرار گرفت [23]. اما مهمترین ویژگی‌های یک متن منسجم و راه‌های بهبود آن در کتاب "انسجام در انگلیسی" توسط هالیدی و حسن [24] مطرح گردید. در تعریف نامبردگان انسجام یک مفهوم معنایی بوده که به روابط معنایی موجود در متن اشاره دارد. این روابط می‌توانند دستوری نباشند و اغلب بر اساس دانش مشترکی که بین نویسنده و خواننده موجود است شکل می‌گیرند. از این رو اغلب تعاریف در خصوص ارتباط موضوعی بین اجزا متن و انسجام برگرفته از ایده آنها است. انسجام یک متن در دو حوزه محلی و عمومی مورد ارزیابی و سنجش قرار می‌گیرد. انسجام محلی به وابستگی موضوعی جملات متوالی با فاصله کم پرداخته و انسجام عمومی وابستگی موضوعی کل متن را مورد ارزیابی قرار می دهد. با وجود اینکه سیستم‌های پردازش متن طراحی شده در سال‌های اخیر متن‌هایی با کیفیت بالا و بسیار نزدیک به نوشته‌های تولید شده توسط انسان ایجاد می‌کنند، اما اغلب آنها درگیر با مفاهیم معنایی واژه‌ها و الگوهای زبانشناسی شده و دارای چالش‌های بزرگی مانند محدودیت به یک حوزه ویژه، نداشتن قابلیت اعمال و گسترش به سایر زبان‌ها، الگوریتم‌های پیچیده و عدم دقت کافی هستند. اما بزرگ‌ترین مشکل رویکردهای ارائه شده ارزیابی وابستگی موضوعی جملات متوالی در محدوده‌ای کوچک و محلی و عدم دقت کافی در سنجش وابستگی موضوعی و انسجام عمومی کل متن است. این مشکل در متن‌های بزرگ‌تر و با جملات بیشتر بسیار مشهودتر است. مهمترین رویکردهای قبل مدل‌های مبتنی بر نهاد و گراف بوده که بسیار درگیر مفاهیم معنایی و زبانشناسی شده‌اند. این روش‌ها با محدود کردن خود به همرخدادی واژگان در جملات متوالی و در بخش کوچکی از متن از دقت کافی در ارزیابی انسجام عمومی برخوردار نبوده‌اند. ارزیابی انسجام عمومی در متن‌های بزرگ با توجه به تعداد زیاد جملات با تکیه بر مفاهیم معنایی موجود در جملات متوالی بسیار مشکل بوده و از دقت کافی برخوردار نخواهد بود.

با معرفی یادگیری عمیق و امکان تبدیل واژگان به بردارهای عددی نرمال راهکارهای جدیدی برای تبدیل متن به بردارهای عددی معرفی شدند. در این روش‌ها هر بردار واژه نشان دهنده معنای آن، موقعیت آن در جمله و میزان ارتباط آن با سایر واژگان است. با بکارگیری این بردارها جملات تبدیل به بردار یا ماتریس‌هایی شدند که امکان عملیات ریاضی بر روی آنان فراهم شده تا بتوان از آنان ویژگی‌هایی را استخراج کرده و با مقایسه این ویژگی‌ها بسیاری از عملیات پردازش متن را بهبود بخشید. با قرار گیری بردارهای واژگانی مربوط به هر جمله ماتریسی تشکیل شده که به آن ماتریس عددی جمله گفته می‌شود. اما یکی از بزرگ‌ترین چالش‌ها در تولید این ماتریس‌ها اختلاف در یکی از ابعاد آنان به دلیل برابر نبودن اندازه جملات در متن بوده که موجب اشکال در اعمال الگوریتم‌های پردازشی و مقایسه‌ای خواهد شد. تا به حال رویکردهای متفاوتی برای رفع این مشکل پیشنهاد شده که یکی از مهمترین آنان تبدیل ماتریس جمله به بردار جمله بوده که این عمل با میانگین گیری ستون‌های ماتریس انجام می‌شود[25] [26] روش میانگین گیری از الگوریتم ساده و سرعت بالایی برخوردار است. اما مشکلاتی مانند احتمال نزدیک بودن بردارهای جملات، کاهش شدید اطلاعات قابل استخراج از جمله، عدم اهمیت جایگاه واژه در جمله و دقت بسیار پایین در ارزیابی انسجام عمومی می‌شود. در رویکرد پیشنهادی دیگری پس از تبدیل جمله به ماتریس عددی با استفاده از مدل های زبانی ماتریس های جملات را هم اندازه و نرمال شده‌اند [27]. اما در رویکرد ارائه شده در این مقاله نیازی به هم سایز کردن ماتریس‌های تولیدی نبوده و الگوریتم پیشنهادی از روشی ساده‌تر و بدون نیاز به عمل نرمال سازی ماتریس‌ها مقایسه ماتریس‌ها را انجام می‌دهد. در این رویکرد با استفاده از معیارهایی آماری، بکارگیری دانش پنهان و وابستگی معنایی واژه‌های موجود به ارزیابی انسجام و وابستگی جملات در کل متن پرداخته است. مدل معرفی شده با ارتقای بررسی انسجام محلی از سطح دو جمله متوالی به سطح جملات با فاصله بیشتر ارزیابی دقیق‌تری را پیشنهاد داده است. مهمترین ویژگی مدل ارائه شده توانایی ارزیابی همزمان انسجام محلی و عمومی در متن‌های بزرگ و با تعداد جملات زیاد و با دقت قابل قبول است.

ISSN: 2745-3673
http://www.Rcsj.ir

در ادامه، این نوشته به شکل زیر سازماندهی شده است. در بخش دوم به مروری بر تاریخچه و کارهای انجام شده در زمینه ارزیابی انسجام متن پرداخته شده و بخش سوم پیش پردازش‌های انجام شده در این تحقیق معرفی گردیده است. بخش چهارم به توصیف مدل پیشنهادی در این مقاله پرداخته شده که درآن ابتدا به ارزیابی انسجام بین دو جمله، سپس ماتریس پنجره فاصله و در نهایت ارزیابی انسجام عمومی کل متن معرفی شده است. در بخش پنجم به معرفی پایگاه داده تولید شده و ارزیابی مدل پیشنهادی پرداخته شده و در بخش آخر نیز نتیجه گیری و پیشنهادات آتی آمده است.

مروری بر مبانی نظری و پیشینه تحقیق

تولید متن‌های طولانی، امکانات جستجو قوی در وب، ترکیب مجموعه بزرگ منابع متنی و پیشرفت سریع در تمامی حوزه‌های پردازش زبان طبیعی و متن، موجب شده است که انسجام و پیوستگی متون تولید شده از مهمترین دغدغه‌های پژوهشگران این حوزه‌ها باشد. نخستین مطالعات انجام شده بر روی ارزیابی انسجام متن به صورت کامپیوتری به کار انجام شده به رویکرد ارائه شده در سال ۱۹۹۸ باز می‌گردد [۲۸]. از نظر نامبردگان متن منسجم متنی است که ارتباطی معنایی بین دو جملات متوالی آن وجود داشته باشد. مدل مبتنی بر شبکه نهاد یکی از مهمترین رویکردهای پیشنهاد شده در حوزه ارزیابی انسجام متن است. در این مدل با استفاده از جایگاه گرامری اسم‌ها و عبارات اسمی وابستگی موضوعی آنان نسبت به هم ارزیابی می‌شود [۲۹] [۳۰]. کاستی بزرگ مدل‌های اولیه پیشنهادی این رویکرد استفاده از تکرار عینی اسم‌ها و عبارات اسمی برای مقایسه جملات متوالی بود. اما در روش پیشنهادی دیگری در سال ۲۰۱۵، راه حلی برای این کاستی معرفی شد [۹]. نامبردگان با استفاده از واژه‌های دارای ارتباط معنایی و هم خانواده دقت روش‌های قبلی را بیش از چهل درصد بهبود دادند. ترکیب مدل مبتنی بر شبکه نهاد با سایر تئوری‌ها و الگوریتم‌های یادگیری ماشین مانند شبکه‌های عصبی [۳۱]، گراف‌های دوقسمتی [۳۲]، موجب بهبود عملکرد روش‌های پیشین شده است. با توجه به ماهیت ویژه یک متن و ارتباط غیر خطی جملات با فاصله‌های متفاوت استفاده از نظریه گراف‌ها در تمامی حوزه‌های پردازش متن رایج بوده و همیشه به عنوان یکی از مهمترین و نخستین گزینه برای پژوهشگران حوزه پردازش متن مورد توجه قرار گرفته است [۱]. برخی از کاستی‌ها و مشکلات موجود در مدل مبتنی بر نهاد و ویژگی‌ها و امکانات خاص گراف‌ها عامل اصلی این توجه بوده و ترکیب این دو ایده موجب تولید رویکردهایی با کارایی بسیار بهتر شده است [۳۰]. در رویکردی دیگر با معرفی مدلی ترکیبی دامنه ارزیابی انسجام را از جملات متوالی فراتر رفته و ارتباط موضوعی جملات با فاصله بیشتر امکان پذیر شد [۳۳]. پترسون و سیمونسن نیز روشی ترکیبی از مدل مبتنی بر نهاد، تئوری گراف و آنتروپی برای ارزیابی ارتباط موضوعی جملات یک متن پیشنهاد کردند [۳۴]. زیرگراف‌های تکراری نیز در رویکردی دیگر برای استخراج الگوهای انسجامی یک متن استفاده شد [۳۵]. استفاده از گراف دوقسمتی و نگاشت آن به گراف‌های فشرده‌تر چالش‌هایی مانند کوچکتر شدن گراف و ماتریس تولید شده را به همراه داشته و موجب نادیده گرفته شدن برخی از اطلاعات و نشانه‌های انسجامی بین جملات که در گراف دوقسمتی اصلی می‌شود. کریستینا لیما و همکارانش با معرفی سه ویژگی روشی بدون نیاز به نگاشت گراف دوقسمتی به یک گراف یک قسمتی را پیشنهاد داده‌اند [۳۲]. در مدل دیگری، استفاده از تئوری گراف، بکارگیری آنتروپی و ترکیب آنان با رویکرد مبتنی بر نهاد موجب کاهش برخی از کاستی‌ها و مشکلات روش‌های اولیه مبتنی بر نهاد شده است [36].

زنجیره‌های واژگانی نیز در سال‌های اخیر در بسیاری از حوزه‌های مرتبط با پردازش متن بویژه در ارزیابی انسجام مورد توجه قرار گرفته است. در رویکردی پیشنهادی زنجیره‌های واژگانی برای ارزیابی انسجام متن خروجی در ترجمه آماری ماشینی استفاده شد [۳۷]. در مدل پیشنهاد شده دیگری نیز از زنجیره‌های واژگانی برای بررسی کیفیت مقالات استفاده شد [۳۸]. برای چیرگی بر کاستی‌های ویژگی‌های معنایی، رویکردهای نوین به سوی مدل‌هایی مبتنی بر شبکه‌های عصبی روی آوردند [۳۹]. این مدل‌ها راه حل‌های جدیدی جهت استخراج ویژگی‌ها پیشنهاد نموده و همچنین نمایشی ساده‌تر از مفهوم انسجام را به ارمغان آوردند. ترکیب مدل شبکه نهاد و شبکه‌های عصبی بازگشتی جهت ارزیابی انسجام محلی [۴۰] [۴۱]، شبکه‌های بزرگ با حافظه کوتاه مدت [۴۲]، مدل مبتنی بر چک لیست و شبکه‌های عصبی بازگشتی [۴۳]، مدل بدون ناظر و مبتنی بر مدل‌های زبانی شبکه‌های عصبی [۴۴] از جمله مهمترین رویکردهای پیشنهاد شده در این حوزه هستند. یادگیری عمیق نیز رویکردی نوین بوده که در سال‌های اخیر گام‌های بسیار موثری در ارزیابی انسجام محلی و عمومی متن برداشته است. ژانگ و همکارش نیز در مدل معرفی شده دیگری از شبکه‌های عصبی کانولوشنی برای طبقه‌بندی متن استفاده کرده‌اند [۴۵].

بیشترین پژوهش‌های انجام شده بر روی ارزیابی و ایجاد متن‌های منسجم تولید شده بر روی خروجی‌های سیستم‌های خلاصه‌سازی انجام شده است [۴۶] [۴۷]. یکی از مشکلات بزرگ در خلاصه سازی ماشینی متن، بویژه در رویکرد استخراجی، از بین رفتن انسجام و پیوستگی متن خلاصه شده است. مشکل دیگر خلاصه سازی استخراجی، انتخاب جملات بسیار شبیه به هم بوده که در عمل یک مفهوم را می‌رسانند [۴۸]. تولید خلاصه‌های منسجم از متون موجود در ویکی پدیا از این نمونه است [۴۹]. خلاصه سازی خودکار متن با استفاده از

ISSN: 2745-3673
http://www.Rcsj.ir

کلونی زنبورعسل [۵۰] و خلاصه سازی تک سندی متون به کمک یادگیری عمیق [۵۱] نیز از مهمترین گزینه‌های معرفی شده برای تولید خلاصه‌های منسجم هستند.

ارزیابی انسجام متون ترجمه شده توسط سیستم‌های ترجمه ماشینی نیز در سال‌های اخیر بسیار مورد توجه قرار گرفته است [۵۲]. عملیات ترجمه ماشینی یک عمل دو زبانه بوده و انسجام یک متن در دو زبان مورد بررسی قرار گرفته و تاثیر هر کدام بر روی دیگری کاملا مشخص نیست [۵۳]. بکارگیری الگوریتم‌های EM و IBM معرفی شده در روش‌های موجود در ترجمه آماری ماشینی [۷] [۳۷]، مدل مبتنی بر آموزش خطا [۵۴] و ترکیب سه روش شبکه نهاد، متریک‌های شباهت در شبکه‌های مبتنی بر گراف و مدل‌های مبتنی بر الگوهای نحوی از مهمترین رویکردهای پیشنهاد شده ارزیابی و تولید ترجمه منسجم در این حوزه هستند [۵۵]. ارزیابی انسجام خروجی‌های تولید شده در سیستم‌های تولید متن [۴۳] [۵۶]، بررسی کیفیت خروجی سیستم‌های ساده‌سازی نحوی متن [۵۷] و خروجی متن‌های ترکیبی [۱۹] [۵۸] نیز از مهمترین موضوعات مورد توجه در این حوزه هستند.

تبدیل واژه‌ها به بردارهای عددی و اعمال الگوریتم‌های پردازشی بر روی این بردارها پنجره‌ای جدید را برای پژوهشگران حوزه پردازش متن باز کرد. استفاده از بردارهای عددی در ارزیابی انسجام متن حوزه‌ای جدید بوده که محققانی آن را بکار گرفته‌اند. مت کوسنر و همکارانش از روشی مبتنی بر بردارهای لغت برای تعیین فاصله بین دو بخش متن ارائه داده‌اند [۵۹]. در رویکرد پیشنهادی دیگری برای ارزیابی ارزش یک جمله از تبدیل آن به بردارهای واژگانی مبتنی بر یادگیری عمیق استفاده شد [۶۰]. رویکردهای دیگری نیز با استفاده از بردارهای واژگانی word2vec و تبدیل جملات به ماتریس‌هایی متشکل از این بردارها روشی کارا برای سیستم‌های پرسش و پاسخ ارائه دادند [۶۱] [۶۲].

در این مقاله با تبدیل واژه‌ها به بردارهای عددی و بکارگیری این بردارها در ایجاد ماتریس جملات مدلی برای ارزیابی انسجام جملات سرتاسر متن ارائه شده است. مدل معرفی شده از ویژگی‌های آماری برای ارزیابی فاصله و شباهت مفهومی جملات استفاده کرده و به هیچ عنوان درگیر مفاهیم معنایی واژه‌ها و جایگاه گرامری آنها نمی‌شود.

رویکرد پیشنهادی

در این بخش به معرفی رویکرد پیشنهادی در این مقاله پرداخته می‌شود. به همین جهت ابتدا به عملیات انجام شده برای پیش پردازش متن ورودی پرداخته شده و سپس الگوریتم ارزیابی کننده انسجام محلی و عمومی معرفی می‌شود.

پیش پردازش متن

یکی از مهم‌ترین عملیات در تمامی سیستم‌های پردازش متن آماده سازی متن ورودی برای اعمال عملیات پردازشی اصلی خواهد بود. در انتخاب نوع پیش پردازش‌های لازم عواملی مانند نوع متن، زبان آن، الگوریتم و نوع عملیات پردازشی بعدی تاثیر زیادی داشته برای هر حوزه پردازش متن مدلی خاص پیشنهاد می‌شود. مهمترین پیش پردازش‌های معرفی شده و قابل انجام بر روی یک متن عبارت از نشانه گذاری ، حذف ایست واژه‌ها ، ریشه یابی و برچسب زنی بخش‌های سخن هستند [۶۳] [۶۴]. انتخاب نوع پیش پردازش و درصد اعمال آن تاثیر بسیار بالایی بر دقت و سرعت عملیات اصلی پردازشی بعد خواهد داشت. با توجه به نیاز به تمامی اجزای جمله پیش پردازش‌های اعمال شده در این پژوهش تفاوت‌هایی با روش‌های قبل دارد. پیش پردازش‌های معرفی شده به دلایل زیر بر روی متن انجام نمی‌شوند:

حذف ایست واژه‌ها: در روش پیشنهادی هر واژه موجود در جمله تاثیری مستقیم بر روی سایر واژه‌ها داشته، برداری مخصوص به خود را تولید کرده و این بردار حتما باید در ماتریس نهایی جمله حضور داشته باشد. در نهایت با اعمال روش پیشنهادی ایست واژه‌های موجود در متن با توجه به درجه اهمیت در همان جمله و همان بخش از متن حذف شده در الگوریتم پردازشی شرکت داده نمی‌شوند. این ایست واژه‌ها با توجه به کم اهمیت بودن آنان در همان بخش از متن انتخاب شده و کاملا محلی هستند. با توجه به ایده موجود در این تحقیق درجه اهمیت ایست واژه‌ها در بخش‌های مختلف متن متفاوت بوده و نمی‌توان مانند سایر رویکردها همگی آنان را با استفاده از یک پایگاه داده و با یک روش و الگوریتم حذف کرد.

ریشه‌یابی: بین هر واژه و ریشه آن تفاوت‌هایی مفهومی و معنایی وجود دارد. در رویکرد پیشنهادی جایگاه واژه، شکل نگارشی آن و مفهوم واژه در ایجاد ماتریس ایجاد شده کاملا مهم بوده و تاثیر مستقیم دارد. لذا پیش پردازش ریشه‌یابی در مورد متن‌های ورودی در این رویکرد انجام نمی‌شود.

ISSN: 2745-3673
http://www.Rcsj.ir

برچسب زنی بخش‌های سخن: این پیش پردازش رفتار گرامری بخش‌های متفاوت جمله را بیان می‌کند. در رویکرد پیشنهادی رفتار گرامری واژه درون بردار ایجاد شده گنجانده شده است. این رویکرد کاملا آماری بوده و در آن نیازی به تشخیص جایگاه گرامری و برچسب‌زنی واژگان نیست.

عملیات پیش پردازش پیشنهادی در این رویکرد از قرار زیر خواهند بود:

ایجاد فاصله بین واژه و علامات نقطه‌گذاری بعد از آن: در متن‌های استاندارد هر علامت نقطه گذاری (نقطه، دو نقطه، ویرگول، علامت سوال، ...) با واژه قبلی خود هیچ فاصله‌ای نداشته و این مسئله موجب می‌شود که از دید سیستم‌های کامپیوتری این علامت جزو واژه محسوب شود. در عملیات پیش پردازش پیشنهادی علامت‌های نقطه گذاری به عنوان یک عضو جداگانه در نظر گرفته می‌شوند. همچنین در پایگاه داده بکار گرفته شده هر کدام از این علامت‌ها دارای برداری منحصر به فرد بوده که موجب افزایش بسیار زیاد دقت هر پردازش متنی در جمله خواهند شد.

یکسان سازی کاراکترهای لهجه: این کاراکترها معمولا در واژه‌ها و اسامی به قرض گرفته شده از سایر زبان‌ها که از سایر کاراکترهای نگارشی لاتین(È, Û, Ï, ...) استفاده می‌کنند وجود دارد [۶۵]. این عمل همه کاراکترهای متن را به کاراکترهای معتبر اسکی تبدیل می‌کند.

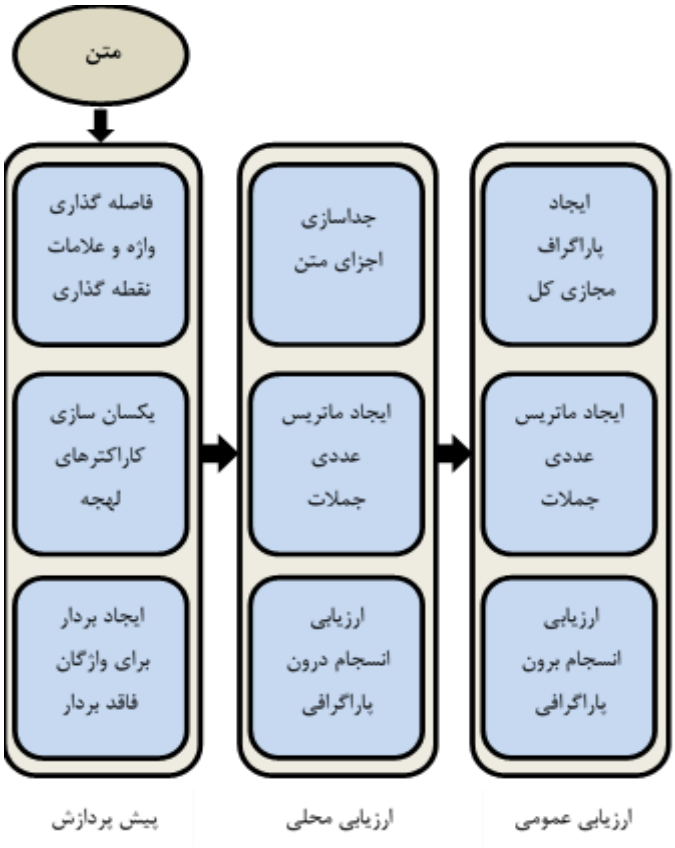
عدم تغییر در نوع حروف بزرگ یا کوچک واژگان موجود در متن: در پیش پردازش پیشنهادی در این پژوهش این عمل یکسان سازی انجام نشده و هر واژه به همان صورت که در متن اولیه ظاهر شده در الگوریتم اجرایی مورد پردازش قرار می‌گیرد. زیرا واژگان شروع شونده با حروف بزرگ دارای اندکی تفاوت در مفهوم با همان واژه در صورت شروع با حرف کوچک هستند. این تفاوت در موقعیت حضور واژه در جمله (ابتدا یا در بدنه اصلی)، در حالت گرامری (در جایگاه اسم خاص یا سایر حالات مانند صفت، قید، ...) و یا واژه‌ای به عاریت گرفته شده از زبان و فرهنگی خاص که شکل نگارش آن مهم بوده کاملا مشهود است.

انتخاب نزدیک‌ترین بردار برای واژگانی که در بانک اطلاعاتی بردار وجود ندارند: بانک اطلاعاتی بردارهای واژگان مورد استفاده در این رساله حاوی ۱۳۲۴۳۰ بردار ۱۰۰ درایه‌ای برای ۱۳۲۴۳۰ کلمه است [۶۶]. اما هنوز هم ممکن است در متن‌های مورد پردازش واژگانی یافت شوند که در این پایگاه داده موجود نباشند. برای بدست آوردن شبیه‌ترین واژه به دنبال کلمه‌ای در پایگاه داده بوده که بزرگ‌ترین زیردنباله مشترک را با آن دارد.

روش پیشنهادی

در رویکرد پیشنهادی ابتدا انسجام درونی هر پاراگراف به عنوان انسجام محلی ارزیابی شده و سپس ارتباط موضوعی پاراگراف‌های متوالی به عنوان انسجام عمومی مورد ارزیابی قرار می‌گیرد. وقتی که خروجی سطح یک پاراگراف‌های منسجم باشند، انسجام بینامتنی می‌تواند ارتباط موضوعی هر پاراگراف را با پاراگراف‌های قبل و بعد و همچنین با موضوع اصلی متن (عنوان متن) ارزیابی کند. همخوانی و نزدیکی جملات موضوعی هر پاراگراف با عنوان متن، تشخیص تقدم و تاخر پاراگراف‌های تشکیل دهنده و ارتباط بینامتنی بین پاراگراف‌ها می‌تواند ابزار مناسبی برای ارزیابی انسجام عمومی باشد. رویکرد پیشنهادی یک پاراگراف مجازی متشکل از عنوان سند متنی به جای جمله موضوعی پاراگراف و جملات موضوعی هر پاراگراف به عنوان سایر جملات تشکیل دهنده آن ایجاد می‌کند. اعمال روش‌های تعریف شده ارزیابی محلی پاراگراف بر روی پاراگراف مجازی موجب تشخیص و تعیین اندازه انسجام عمومی کل متن خواهد شد. تصویر (۱) دیاگرام روش پیشنهادی را نشان می‌دهد.

ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۱: دیاگرام رویکرد پیشنهادی

بردارهای عددی واژگان

در اغلب روش‌هایی که متن را به بردارهایی عددی تبدیل می‌کنند از روش‌های مبتنی بر انرژی استفاده می‌شود. روش word2vec از خانواده این روش‌ها بوده و در سال ۲۰۱۳ توسط تیم گوگل معرفی شده است [۶۷]. این روش با الهام گیری از مدل‌های مبتنی بر شبکه عصبی قادر به حدس و تشخیص مفهوم یک واژه با دقت بسیار بالا بر پایه حضورهای قبلی آن در متن است. هدف اصلی این رویکرد گردآوری و کنار هم قرار دادن بردارهای واژگان شبیه به هم در یک فضای برداری است. عبارت (۱) فرمول کلی محاسبه این بردارها بوده که در آن nb(t) مجموعه‌ای از واژگان همسایه و $p(w_i \mid w_t)$ ماکزیمم احتمال گذر از واژه w_i به واژه w_t است.

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{j\in nb(t)}\log p\left(w_i \middle| w_t\right) \qquad (۱)$$

در روش پیشنهادی جمله به عنوان کوچکترین واحد منسجم در یک متن پذیرفته شده است. در نخستین مرحله جملات از هم تفکیک شده و برای ارزیابی انسجام یک پاراگراف پیوستگی و وابستگی موضوعی جملات متوالی آن ارزیابی می‌شوند. در این رویکرد انسجام محلی عبارت از ارتباط موضوعی جملات متوالی یک پاراگراف بوده و انسجام و پیوستگی موضوعی پاراگراف‌های متوالی به عنوان انسجام عمومی در نظر گرفته می‌شود. برای ارزیابی وابستگی موضوعی و ارتباط جملات، با بکارگیری بردارهای تولید شده توسط الگوریتم word2vec برای هر جمله یک ماتریس تولید می‌شود. در حالت کلی عمل پیش بینی جمله بعدی بستگی به (n-i) جمله قبلی آن دارد:

$$p(t) = p(s_1)p(s_2,\mid s_1)p(s_3 \mid s_1, s_2)\cdots p(s_n \mid s_1\cdots s_{n-1}) = \prod_{i=1}^{n} p(s_n \mid s_1\cdots s_{n-i}) \qquad (۲)$$

ISSN: 2745-3673
http://www.Rcsj.ir

در این معادله محاسبه عبارت $P(S_i \mid S_{i-1})$ با مقایسه ویژگی‌های استخراجی از ماتریس‌های جملات تولید شده توسط روش word2vec انجام می‌شود:

$$p(s_i \mid s_{i-1}) = p((a_{(i-1)} \cdots a_{i-n}) \mid a_{(i-1,1)}, a_{(i-1,2)}), \ldots, a_{(i-1,m)}) \quad (۳)$$

در این معادله $(a_{(i-1)}, a_{(i-2)}, \ldots a_{(i-n)})$ ویژگی‌های مربوط به جمله S_i ، S_{i-1} بوده و رویکرد بکار گرفته شده این بخش برگرفته از روش معرفی شده توسط روزنفلد بوده و از n-grams با فاصله استفاده می‌کند [۶۸]. در این مرحله با توجه به تئوری روزنفلد ابتدا انسجام جمله موضوعی و پنج جمله متوالی پس از آن سنجیده می‌شود. در این سنجش مقادیر انسجام هر یک از پنج جمله بعدی با جمله موضوعی، میانگین این مقادیر و اختلاف هر مقدار با مقدار قبلی بدست آمده و درون برداری n درایه‌ای قرار می‌گیرد. دلیل تعیین و بکارگیری اختلاف مقادیر بدست آمده اهمیت روند تغییرات و میزان کاهش مقادیر کسب شده بوده که از تفاضل دوبه‌دوی آنها بدست می‌آید. سپس جمله نخست پاراگراف حذف منطقی شده و عمل فوق در پاراگراف جدید ایجاد شده با یک جمله کمتر انجام می‌شود. این روند در یک پاراگراف k جمله‌ای k-5 بار تکرار شده که در نهایت یک ماتریس شامل k-5 بردار ایجاد می‌شود. با توجه به رویکرد روزنفلد ارتباط انسجامی جملات از فاصله صفر (دو جمله دنبال هم) تا فاصله پنج از کاهش ثابتی پیروی کرده و از فاصله پنج به بعد ثابت می‌ماند. به همین دلیل این مقایسه فقط بین جملات با فاصله صفر تا پنج انجام شده و فاصله‌های بیشتر مقایسه نمی‌شوند. با توجه به ذخیره سازی این فاصله‌ها در یک بردار و تشکیل ماتریس پاراگراف، هرچه مقادیر موجود در سطرهای ماتریس از فاصله کمتری برخوردار باشند انسجام محلی پاراگراف مورد نظر بیشتر است.

ارزیابی انسجام بین دو جمله

در این بخش به معرفی مدلی برای تشخیص و ارزیابی میزان ارتباط موضوعی (انسجام) اجزای یک سند متنی پرداخته می‌شود. مدل پیشنهادی دارای سه مرحله است:

- **مرحله نخست:** سند متنی به واحدهای منسجم آن (جملات) تقسیم شده و برای هر جمله ماتریسی متشکل از بردارهای word2vec تولید می‌شود.
- **مرحله دوم:** ماتریس‌های تولیدی نرمال می‌شوند [۵۹].
- **مرحله سوم:** ارتباط موضوعی جملات با توجه به ارتباط ماتریسی آنان ارزیابی می‌شود.

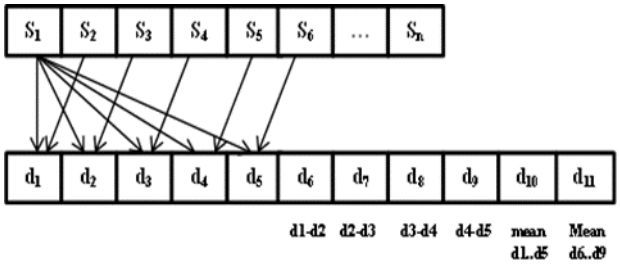
در روش پیشنهادی از دو ویژگی شباهت ماتریسی و معکوس فاصله برای ارزیابی میزان انسجام دو جمله استفاده می‌شود. شباهت بیشتر ماتریس‌های جملات نشان دهنده انسجام بیشتر موجود در متن است. کشف این شباهت‌ها و ارتباطات موضوعی در یک فضای کوچک مانند پاراگراف بسیار ساده است. برای ارزیابی شباهت دو جمله از شباهت کسینوسی (۴) ماتریس‌های آنان و برای ارزیابی معکوس فاصله دو جمله از معکوس فاصله مانهاتان (۵) دو ماتریس آنان استفاده می‌شود.. در این معادلات A_i و B_i درایه‌های ماتریس‌های A و B هستند.

در نخستین مرحله مقایسه جملات، الگوریتم یک بردار یازده درایه‌ای ایجاد می‌کند. پنج درایه نخست بردار تولید شده شامل مجموع مقدار معیار شباهت کسینوسی (CS) و معکوس فاصله مانهاتان (IMD) و جمله موضوعی (نخستین جمله) با پنج جمله متوالی بعدی است. چهار درایه بعدی شامل فاصله و تفاوت بین پنج درایه یک تا پنج است. درایه دهم میانگین پنج درایه نخست (۱..۵) و درایه یازدهم میانگین چهار درایه بعدی (۶..۹) است (شکل ۲). دلیل استفاده میزان کاهش متوالی تغییرات در جملات با فاصله یک تا پنج اهمیت میزان تغییرات و کاهش وابستگی آنان است. تغییرات انسجامی جملات متوالی در یک متن منسجم از یک روند کاهشی ثابتی پیروی می‌کنند [۶۸]. در مرحله بعد عمل مقایسه و ایجاد بردار از جمله دوم پاراگراف شروع شده و برای جملات بین دو تا شش بردار یازده درایه‌ای جدید تشکیل می‌شود (جمله نخست پاراگراف حذف منطقی شده و الگوریتم دوباره از ابتدای پاراگراف شروع می‌شود). برای هر پاراگراف این عمل (n-5) بار تکرار شده که n تعداد جملات پاراگراف است. در نهایت برای هر پاراگراف متن یک ماتریس (n*11) درایه‌ای تولید خواهد شد (شکل ۳). فاصله کم درایه‌های سطرهای ماتریس تولید شده نشانه انسجام بیشتر پاراگراف مورد ارزیابی است. برای سنجش شباهت این سطرها هم مجموع مقدار CS و IMD را محاسبه کرده‌ایم.

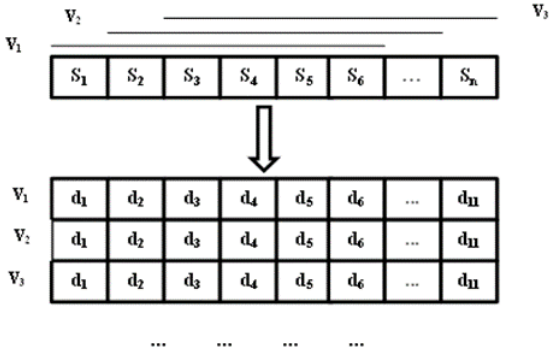
$$CS = \sum_{i-1}^{n} A_i * B_i \Big/ \left(\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}\right) \quad (۴)$$

ISSN: 2745-3673
http://www.Rcsj.ir

$$IMD = 1 - norm\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2 \Big/ (\partial_A^2 + \partial_B^2)} \qquad (۵)$$



شکل ۲: بردار یازده درایه‌ای مقایسه پنجره پنج جمله‌ای



شکل ۳: ماتریس انسجام پاراگراف

در نهایت جهت ارزیابی انسجام کل متن ضریب همبستگی بردارهای تولید شده هر پنجره (سطرهای ماتریس پاراگراف) نسبت به هم و به صورت متوالی اندازه گیری می‌شود (۶).

$$corr(V_1 \cdots V_m) = \left[\sum_{i=1}^{m} \frac{(V_i - \mu V_i)(V_{i+1} - \mu V_{i+1})}{\sigma V_i \sigma V_{i+1}} / m \right] \qquad (۶)$$

ضریب همبستگی شدت رابطه و نوع مستقیم یا معکوس بودن آنرا نسبت به هم در دو مجموعه داده مشخص کند. مقدار این ضریب بین ۱ و ۱- بوده که هرچه رابطه دو مقدار با هم بیشتر شود مقدار به یک نزدیک‌تر، مقدار صفر نشان دهنده عدم وجود رابطه بین دو متغیر و مقدار منفی نشان دهنده رابطه معکوس آنان است. ضریب همبستگی بالاتر در پاراگراف مورد ارزیابی نشان دهنده انسجام بالاتر کل متن است. برای ارزیابی انسجام عمومی (انسجام پاراگراف‌های متوالی) یک پاراگراف مجازی تولید می‌شود. برای ایجاد این پاراگراف ابتدا عنوان متن در جایگاه نخستین جمله آن قرار گرفته و سپس اولین جمله از هر پاراگراف (جمله موضوعی آن) به صورت متوالی در جایگاه جملات بعدی پاراگراف مجازی قرار می‌گیرد. در نهایت روش ارائه شده برای ارزیابی انسجام محلی پاراگرافی بر روی این پاراگراف مجازی نیز اعمال شده که نتیجه آن میزان انسجام عمومی متن است.

شبهه کد الگوریتم

۱- ابتدا مقدار معکوس فاصله انتقال واژه (IWMD) جمله اول پاراگراف (جمله i) و سایر پنج جمله بعدی محاسبه می‌شود

a.جمله نخست با جمله دوم

b. جمله نخست با جمله سوم

c. .....

ISSN: 2745-3673
http://www.Rcsj.ir

d. جمله نخست با جمله ششم

۲- مقدار IWMD جمله دوم (جمله i+1) با پنج جمله بعدی خود محاسبه می‌شود. (تشکیل بردار جمله)

۳- مرحله ۲ به تعداد n-5 بار تکرار می‌شود. (n تعداد جملات پاراگراف است. تشکیل ماتریس پاراگراف)

۴- پاراگرافی مجازی شامل عنوان متن و جملات موضوعی هر پاراگراف ایجاد می‌شود.

۵- مراحل ۱ تا ۳ بر روی پاراگراف مجازی اعمال می‌شود.

ارزیابی میزان انسجام و وابستگی موضوعی دو جمله

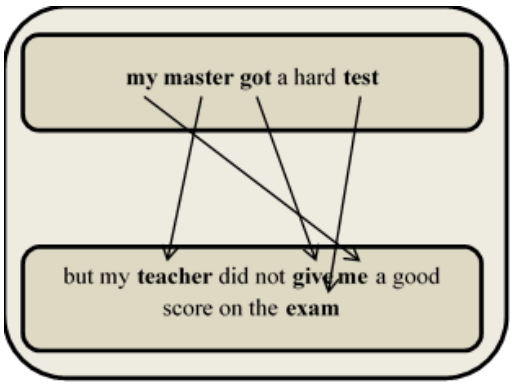
اغلب روش‌های پیشنهاد شده برای ارزیابی ارتباط دو جمله از کیسه واژگان (BOW) و یا از معیار فراوانی وزنی TF-IDF آنان استفاده کرده‌اند. بزرگ‌ترین اشکال دو معیار فوق وابستگی کامل ویژگی‌های استخراجی به شکل ظاهر شده و املای واژه‌های موجود در متن بوده و برای محاسبه فاصله یا شباهت دو واژه مشابه یا مترادف نیاز به معیارهایی قوی‌تر است. به دو جمله متوالی زیر اما با واژگان متفاوت دقت کنید:

my master got a hard test
but my teacher did not give me a good score on this exam

با توجه به اینکه این دو جمله هیچ واژه مشترکی ندارند اما یک نوع اطلاعات را منتشر کرده، و دارای نوعی انسجام و وابستگی موضوعی هستند. واژگان با مفاهیم نزدیک به هم مانند (my, me)، (mater, teacher)،(got, give) و (test, exam) در این دو جمله با هم تشابه معنایی داشته، موجب وابستگی و پیوستگی بین دو جمله شده و دنبال هم بودن آنان را نمایش می‌دهند. اما به هیچ عنوان امکان یافتن الگویی مشترک با معیارهای BOW و TF-IDF بین آنان وجود ندارد. رویکرد پیشنهادی لیانا ارماکووا و همکارانش برای تخمین و ارزیابی انسجام عمومی میانگین تمام شباهت‌های جملات متوالی را بدست آورده‌اند [۶۹].

$$coherence(d) = \frac{1}{|s|-1}\sum_{i=2}^{|s|} sc(S_{i-1}, S_i) \qquad (۷)$$

مت جی کوسنر برای ارزیابی انسجام مفهومی دو جمله معیار جدیدی با نام فاصله انتقال واژه را پیشنهاد و ارائه کرد [۵۰]. فاصله انتقال واژه عبارت از انتقال شباهت معنایی بین دو واژه از یک بخش به بخش دیگری از متن بوده و مشخص می‌کند که محتوای معنایی مجموعه واژگان موجود در یک جمله با چه فاصله‌ای به جمله بعدی انتقال پیدا کرده‌اند. هر چه این فاصله کمتر باشد ارتباط انسجامی دو جمله بیشتر است. فاصله معکوس انتقال واژه عکس فاصله انتقال واژه بوده که در این مقاله معرفی شده است. مقدار بیشتر این معیار نشان دهنده ارتباط انسجامی بیشتر دو جمله است. در مدل معرفی شده وابستگی انسجامی دو جمله با محاسبه مجموع دو مقدار کمترین فاصله و بیشترین شباهت بین دو جمله محاسبه می‌شود.

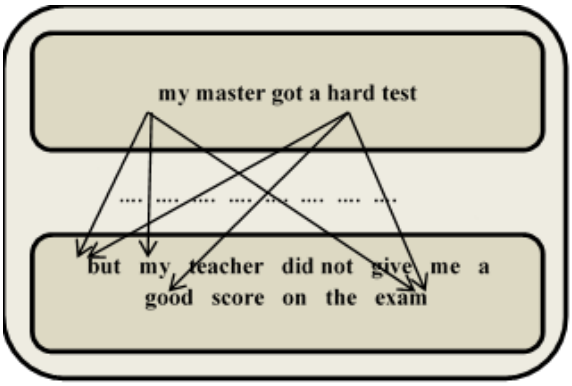


شکل ۴: نمایش فاصله انتقال واژه (WMD) در چند واژه [۵۹]

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۶، زمستان ۱۳۹۸، ص ۵۴- ۳۸

ISSN: 2745-3673
http://www.Rcsj.ir

فاصله انتقال واژه عبارت از انتقال شباهت معنایی بین دو واژه از یک بخش به بخش دیگر متن مانند master و teacher است. در این پژوهش برای تعیین فاصله بین این واژه‌ها از معیارهای شباهت کسینوسی و معکوس فاصله مانهاتان و در فضای برداری بردارهای word2vec استفاده شده است. این روش هزینه انتقال بین هر واژه از جمله مبدا را با تمام واژگان جمله مقصد محاسبه خواهد کرد. ابتدا هر واژه i در جمله d به تمام واژه‌های جمله d’ منتقل می‌شود. سپس برای بدست آوردن نرخ انتقال از شباهت کسینوسی بین هر واژه از جمله d با جمله d’ استفاده می‌شود. این عمل با محاسبه شباهت بین بردارهای word2vec آنان انجام می‌شود (شکل ۵).



شکل۵: نرخ انتقال بین هر واژه از جمله d با جمله d’

برای دو جمله فوق اگر $T1 \in R^{n \times m}$ یک ماتریس نشان دهنده این انتقال باشد به عنوان مثال جدول (۱) مشخص کننده تمام این انتقالات با توجه به شباهت کسینوسی واژه‌های دو جمله خواهد بود. تمامی مقادیر جدول ها به درصد هستند. در مرحله بعد به همین ترتیب معکوس فاصله مانهاتان انتقال هر واژه i در جمله d به تمام واژه‌های جمله d’ محاسبه می‌شود. چون این مقادیر در بیشتر حالات مقدار عددی بزرگی شده آنان را نرمال کرده و نتیجه $T2 \in R^{n \times m}$ در جدول دیگری قرار می‌گیرد. این جدول به دلیل محدودیت فضای مقاله نمایش داده نشده است. برای هر انتقال در جدول شباهت‌های کسینوسی(CS) سه مقدار بیشینه انتخاب و بقیه حذف می‌شوند. این سه انتخاب مشخص کننده سه واژه با نزدیک‌ترین فاصله انتقال مفهوم (واژه مترادف یا دارای نزدیک‌ترین مفهوم) در جمله مقصد با واژه مورد نظر در جمله مبدا هستند. برای هر انتقال در جدول فاصله معکوس مانهاتان(IMD) نیز سه مقدار برگزیده و بقیه حذف می‌شوند. اما معیار انتخاب این سه مقدار موقعیت مکانی آنان در جدول بوده و درایه‌های انتخابی مشابه جدول شباهت‌های کسینوسی برگزیده می‌شوند. مقدار (IWMD) جمع جبری مقادیر مشابه CS و IMD هستند (جدول ۲).

با توجه به جدول مشاهده شده که روش پیشنهادی به صورت خودکار و آماری ایست واژه‌ها را حذف کرده است. البته این عمل به صورت صد در صد نبوده و ممکن است برخی از ایست واژه‌ها را حذف نشده و برخی واژه‌های دیگر را به جای ایست واژه حذف شوند. اما این روش دارای دو ویژگی عدم درگیری عملیات حذف ایست واژه‌ها در بخش پیش پردازش و محلی بودن ایست واژه‌های کشف شده است. در نهایت جمع جبری مقادیر موجود در ماتریس اسپارس حاصل (جدول ۲) به عنوان نخستین مقدار بردار V_i در نظر گرفته خواهد شد. این عمل برای جملات با فاصله یک تا پنج انجام شده و با توجه به الگوریتم توصیف شده یازده درایه بردار V مقدار می‌گیرند. در نهایت عمل مربوطه بر روی تمامی پنج جمله‌های متوالی هر پاراگراف تکرار شده که منجر به تولید ماتریس عددی پاراگراف می‌شود. در نهایت جهت ارزیابی انسجام پاراگراف ضریب همبستگی بردارهای تولید شده هر پنجره (سطرهای ماتریس عددی پاراگراف) نسبت به هم و به صورت متوالی اندازه گیری می‌شود (۶).

یافته‌های پژوهش

در این بخش به ارزیابی سیستم پیشنهادی پرداخته می‌شود. جهت ارزیابی روش ارائه شده بیست داستان کوتاه از سری داستان‌های هانس کریستین آندرسن انتخاب و برای هر داستان ده نمونه غیر منسجم ایجاد شده است. این ده نمونه عبارت از پنج نمونه متن غیر منسجم با جابجایی اتفاقی جملات آن به میزان ۱۰ درصد، ۲۰ درصد، ۳۰ درصد، ۴۰ درصد و ۵۰ درصد و پنج نمونه متن غیر منسجم با حذف اتفاقی جملات (ایجاد متن کوتاه‌تر) هستند. پنج نمونه خلاصه ایجاد شده شامل دو متن با حذف ۱۰ درصد جملات، دو متن با حذف ۲۰ درصد جملات و یک متن با حذف ۳۰ درصد جملات هستند. متن‌های خلاصه بدون اعمال هیچگونه الگوی خلاصه سازی بوده و کاملا اتفاقی ایجاد شده که انسجام آنان کاهش پیدا کرده است. در نتیجه در پایگاه داده مورد مطالعه ۲۲۰ متن شامل ۲۰ متن دارای انسجام کامل و ۲۰۰ متن با انسجام کاهش یافته با درصدهای ذکر شده به دو صورت حذف اتفاقی بخش‌های مختلف و جابجایی اتفاقی جملات

ISSN: 2745-3673
http://www.Rcsj.ir

است. در نتیجه متن‌های ایجاد شده با جابجایی یا کاهش جملات دارای کاهش انسجامی متناسب با میزان جابجایی یا کاهش جملات هستند. سپس روش پیشنهادی بر روی نمونه اصلی و نمونه های کاهش یافته انسجام اعمال شده و دقت سیستم اندازه گیری شده است. با توجه به اینکه کاهش انسجام در نمونه‌های تولید شده از قاعده خاصی (کاهش ده درصدی) پیروی می‌کند انسجام ارزیابی شده نیز از همین قاعده پیروی کرده و دقت سیستم معرفی شده را نشان می‌دهد.

جهت ارزیابی اولیه مدل ده متن با طول متفاوت از متن‌های موجود در پایگاه داده ایجاد شده به همراه تمامی نمونه‌های غیر منسجم تولیدی از هر کدام انتخاب شده است. برای ارزیابی اولیه ابتدا یک متن با طول بلند (شامل ۱۹۲ جمله و ۴۵۰۶ واژه) به همراه ده نمونه متن کاهش انسجام یافته آن را انتخاب و مدل پیشنهادی را بر روی آن اعمال شده و نتایج حاصل مورد ارزیاب قرار گرفته است (جدول ۳).

جدول ۱: شباهت کسینوسی بین واژه‌های دو جمله

	my	Master	...	test
but	0.276	-0.026	...	0.277
my	5.170	1.023	...	0.067
teacher	0.500	1.179	...	0.744
...	...	...	...	...
on	3.296	0.083	...	0.714
this	0.434	-0.1651	...	0.445
exam	0.719	1.896	...	1.973

جدول ۲: جدول تجمعی سه مقدار بیشینه مقدار تشابه بین واژه‌های دو جمله

	my	master	got	a	hard	test
My	6.170		1.121	1.327	1.074	
teacher		1.351				0.870
Me	0.972		2.866			
A				2.954		
Good	1.188				1.856	
Score		1.382				2.161
This				1.369		
Exam		2.086	1.231		0.802	2.171

جدول ۳: مقایسه دقت مدل پیشنهادی و انسجام تئوری

کاهش انسجام	انسجام تئوری	مدل پیشنهادی	مقایسه دقت مقدار واقعی با تئوری
جابجایی جملات	۹۰	۸۲	۹۱,۱۱
	۸۰	۶۹	۸۶,۲۵
	۷۰	۶۰	۸۵,۷۱
	۶۰	۵۸	۹۶,۶۷
	۵۰	۳۹	۷۸
کاهش اتفاقی تعداد جملات	۹۰ (نمونه ۱)	۷۷	۸۵,۵۶
	۹۰ (نمونه ۲)	۸۱	۹۰
	۸۰ (نمونه ۱)	۶۶	۸۲,۵
	۸۰ (نمونه ۲)	۶۸	۸۵
	۷۰	۵۲	۷۴,۲۹
دقت مدل			**۸۵,۵۱**

ISSN: 2745-3673
http://www.Rcsj.ir

لیوما و تریسان با ارائه پارامترهایی قابل استخراج از خود گراف دوقسمتی و عدم نیاز به تصویر کردن آن به گراف معمولی غیر جهت دار روشی کارا را برای ارزیابی انسجام ارائه کرده‌اند [۳۲]. این روش به دلیل ارزیابی انسجام عمومی در سطحی وسیع‌تر نسبت به رویکردهای قبلی، گزینه مناسبی جهت مقایسه با روش پیشنهادی ما و ارزیابی آن است. جهت ارزیابی مدل پیشنهادی این مقاله، مدل مربوطه را با روش ارائه شده (BGSEG) توسط لیوما و تریسان مقایسه شده است. در ارزیابی انجام شده ابتدا متن ۱۹۲ جمله‌ای و سپس ده متن با اندازه‌های متفاوت به همراه ده نمونه کاهش یافته انسجام هر کدام انجام شده است (جدول ۴).

جدول ۴: مقایسه مدل پیشنهادی و مدل BGSEG

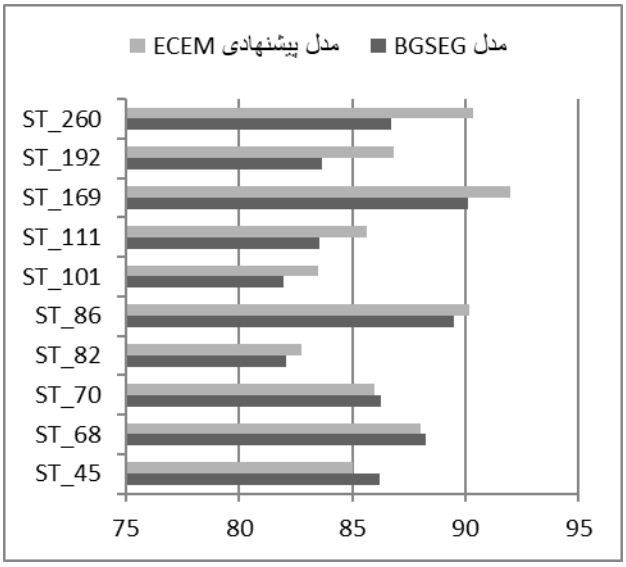
نوع کاهش انسجام	A	B	C	D	E
متن اصلی	۱۰۰	۸۸	۸۸	۸۶	۸۶
جابجایی جملات	۹۰	۸۲	۹۱,۱۱	۸۱	۹۰
	۸۰	۶۹	۸۶,۲۵	۶۶	۸۲,۵
	۷۰	۶۰	۸۵,۷۱	۵۷	۸۱,۴۳
	۶۰	۵۸	۹۶,۶۷	۵۷	۹۵
	۵۰	۳۹	۷۸	۳۷	۷۴
کاهش اتفاقی تعداد جملات	۹۰ (نمونه ۱)	۷۷	۸۵,۵۶	۷۴	۸۲,۲۲
	۹۰ (نمونه ۲)	۸۱	۹۰	۷۹	۸۷,۷۸
	۸۰ (نمونه ۱)	۶۶	۸۲,۵	۶۵	۸۱,۲۵
	۸۰ (نمونه ۲)	۶۸	۸۵	۶۶	۸۲,۵
	۷۰	۵۲	۷۴,۲۹	۵۰	۷۱,۴۳
دقت هر مدل			**۸۵,۷۳**		**۸۳,۱**

در این جدول (۴) A انسجام تئوری، B دقت مدل ECEM، C مقایسه مدل ECEM با انسجام واقعی، D دقت مدل BGSEG و E مقایسه مدل BGSEG با انسجام واقعی است. نتایج اعمال روش، ارزیابی و مقایسه نتایج مشخص می‌کند که روش پیشنهادی در متن‌های بزرگ‌تر عملکردی بهتر داشته و نتیجه‌ای نزدیک‌تر به مقدار واقعی با ۱,۱۹ درصد بهینه سازی را بدست می‌آورد (جدول ۵).

جدول ۵: مقایسه نتایج مدل پیشنهادی و مدل (BGSEG) بر روی ده متن با اندازه‌های متفاوت

متن‌های انتخابی	تعداد جملات	مدل BGSEG	مدل پیشنهادی ECEM
ST_45	۴۵	۸۶,۲۲	۸۵
ST_68	۶۸	۸۸,۲۵	۸۸
ST_70	۷۰	۸۶,۲۵	۸۶
ST_82	۸۲	۸۲,۱۱	۸۲,۷۵
ST_86	۸۶	۸۹,۵	۹۰,۲
ST_101	۱۰۱	۸۲	۸۳,۵
ST_111	۱۱۱	۸۳,۵۵	۸۵,۶۵
ST_169	۱۶۹	۹۰,۱۲	۹۲
ST_192	۱۹۲	۸۳,۶۷	۸۶,۸۵
ST_260	۲۶۰	۸۶,۷	۹۰,۳۵
دقت مدل		**۸۵,۸۴**	**۸۷,۰۳**

ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۶: نمودار مقایسه نتایج مدل پیشنهادی و مدل (BGSEG) بر روی ده متن با اندازه‌های متفاوت

نتیجه‌گیری

در این مقاله با بکارگیری بردارهای عددی word2vec، تبدیل جملات به ماتریس‌های هم اندازه و نرمال رویکردی ساده و کارا برای نمایش و ارزیابی انسجام متن پیشنهاد شده است. در روش پیشنهادی شباهت کسینوسی و معکوس فاصله مانهاتان نرمال شده بین واژه‌های دو جمله دو معیار مهم جهت ارزیابی وابستگی موضوعی دو جمله است. مدل ارائه شده فاصله گذر جملات موجود در یک پاراگراف را درون یک بردار قرار داده که بردار تولید شده بردار انسجام پاراگراف نام دارد. برخلاف رویکردهای مبتنی بر کیسه واژگان و زنجیره‌های واژگانی این روش اصلا متکی به شکل نگارش واژه و مفهوم ظاهری آن نبوده، به راحتی ارتباط مفهومی جملات حاوی آنها را کشف می‌کند. نتایج اعمال شده بر روی متن های داستانی عملکرد این روش را بر روی متن های داستانی با تعداد جملات بالا ۱,۱۹ درصد بهبود را نشان می‌دهد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Blanco, R., & Lioma, C. (2012). Graph-based term weighting for information retrieval. Information retrieval, Vol 15, No. 1, pp. 54-92.

[2] Parveen, D., & Strube, M. (2015). Integrating importance, non-redundancy and coherence in graph-based extractive summarization. in proc the twenty-fourth international joint conference on artificial intelligence (IJCAI). pp. 1298–1304.

[3] Zhang, R. (2011). Sentence ordering driven by local and global coherence for summary generation. In proc the ACL-HLT 2011 student session 6–11.

[4] Celikyilmaz, A., & Hakkani-Tur, D. (2011), Discovery of topically coherent sentences for extractive summarization. In proc the 49th annual meeting of the association for computational linguistics, 491–499.

[5] Ferreira, T. C., Krahmer, E., & Wubben, S. (2016). Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In proc of the 54th annual meeting of the association for computational linguistics.1, 568-577.

[6] Fox, H. J. (2002). Phrasal cohesion and statistical machine translation. In proc the conference on empirical methods in natural language processing (EMNLP), 304-311.

[7] Lin, Z., Liu, C., Ng., H. T., & Kan, M., (2012). Combining coherence models and machine translation evaluation metrics for summarization evaluation. In proc the 50th annual meeting of the association for computational linguistics, 1, 1006–1014.

[8] Xiong, D., Ding, Y., Zhang, M., & Tan, C. L. (2013). Lexical chain-based cohesion models for document-level statistical machine translation. In proc 2013 conference on empirical methods in natural language processing, 1563–1573.

[9] Zhang, M., Feng. V. W., Qin, B., Hirst, G., Liu, T., & Huang, J. (2015). Encoding world knowledge in the evaluation of local coherence. in proc of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies. 1087-1096.

[۱۰] پورمعصومی، آ. صدوقی یزدی، ه. قائمی، ه و دلخسته. ز (۱۳۹۵). آنالیز حس اسناد فارسی با طراحی حوزه تبدیل بهینه. نشریه مهندسی برق و مهندسی کامپیوتر ایران. (۱۴)۲.

[۱۱] گلپر رابوکی، ع. ضرغامی فر، س، و رضایی نور، ج (۱۳۹۵). استخراج ویژگی ها و بسط لغت نامه در اندیشه کاوی مورد استفاده در متون فارسی. نشریه مهندسی برق و مهندسی کامپیوتر ایران، (۱۴)۳.

[۱۲] نوری هفت چشمه، ک. خدادادی، ر. اکبری، ی . رضوی، س. م، و احمدی ترشیزی. ح (۱۳۹۵). تشخیص جنسیت نویسنده مستقل از متن و زبان نوشتاری با استفاده از پالایش پویای نمادین مبتنی بر تبدیل رادان. نشریه مهندسی برق و مهندسی کامپیوتر ایران، (۱۴)۴.

[۱۳] عباسی، ف. سهرابی، ب. مائیان، ا. و خدیور، آ (۱۳۹۶). ارائه مدلی جهت دسته بند احساسات خریداران کتاب با استفاده از رویکرد ترکیبی. فصلنامه مطالعات مدیریت کسب و کار هوشمند، (۶)۲۱، ۶۵- ۹۲.

[۱۴] کشاورزیان، س. و براردخت، ح (۱۳۹۶). جایگاه کتاب و کتابخوانی در سایت تبیان با رویکرد متن کاوی و تحلیل شبکه‌های اجتماعی. فصلنامه مطالعات مدیریت کسب و کار هوشمند، (۶)۲۱ ۱۶۹- ۱۸۸.

[۱۵] امیری، م. ختن لو، ح. (۱۳۹۲). خوشه بندی اسناد مبتنی بر آنتولوژی و رویکرد فازی. فصلنامه علمی پژوهشی فنآوری اطلاعات و ارتباطات ایران. (۵)۱۸،۱۷

[۱۶] میردامادی، م. م. زارع بیدکی، ع. م،و رضاییان، م (۱۳۹۳). قطعه بندی عبارات متون فارسی با استفاده از شبکه های عصبی. نشریه مهندسی برق و مهندسی کامپیوتر ایران، (۱۱)۲.

ISSN: 2745-3673
http://www.Rcsj.ir

[۱۷] برنجیان شاپوررضا، دستغیب محمد باقر، جستجوگر واژه‌های مصوب فرهنگستان زبان و ادب فارسی و واژه‌های معادل و رایج آنها در سیستم‌های بازیابی اطلاعات، مجله علمی پژوهش در علوم رایانه، شماره ۱۲، زمستان ۱۳۹۷، ص۲۵-۳۸.

[18] Higgins, D., Burstin, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In proc. NAACL-HLT. 185- 192.

[19] Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In proc NAACL-HLT. 681-684.

[20] Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In proc the seventh workshop on building educational applications using NLP, 33-43.

[21] Huang, G., Tan, M., Huang, S., Mo, R., & Zhou, Y. (2017). A discourse coherence model for analyzing Chinese students' essay. In progress in informatics and computing (PIC).430-434: IEEE.

[22] Arunsirot, S. (2013). An analysis of textual metafunction in Thai EFL students' writing. Novitas-ROYAL (Research on Youth and Language), (7)2, 160-174.

[23] Luhn, H. P. (1958). A business intelligence system. IBM journal of research and development, Vol. 2, No. 4, pp. 314-319.

[24] Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English, ed: London: longman.

[25] Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daume, H. (2015). Deep unordered composition rivals' syntactic methods for text classification. In proc of the 53rd annual meeting of the association for computational linguistics.1, 1681-1691.

[26] Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). Embedding words and sentences via character n-grams. in proc of the 2016 conference on empirical methods in natural language processing, Austin. 1504–1515

[27] Abdolahi, M., & Zahedi, M. (2017). Sentence matrix normalization using most likely n-grams vector. presented at the 2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI), Tehran, Iran.

[28] Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. Discourse processes, Vol. 25, No. 2, pp. 285-307.

[29] Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach, Computational linguistics, Vol. 34, No. 1, pp. 1-34.

[30] Putra, G. V. G., & Tokunaga, T. (2017). Evaluating text coherence based on semantic similarity graph. In proc of TextGraphs-11: the workshop on graph-based methods for natural language processing. 76-85.

[31] Xu, F., Du, S., Li, M., & Wang, M. (2017). An entity-driven recursive neural network model for chinese discourse coherence modeling. arXiv preprint arXiv:1704.04336.

[32] Lioma, C., Tarissan, F., Simonsen, J. G., Petersen, G., & Larsen, B. (2016). Exploiting the bipartite structure of entity grids for document coherence and retrieval. In proc of the 2016 ACM international conference on the theory of information retrieval. 11-20.

[33] Guinaudeau, C., & Strube, M. (2013), Graph-based local coherence modeling. In proc of the 51st annual meeting of the association for computational linguistics, 1, 93-103.

[34] Petersen, C., Lioma, C., Simonsen, J. G., & Larsen. B. (2015). Entropy and graph-based modeling of document coherence using discourse entities: An application to

ISSN: 2745-3673
http://www.Rcsj.ir

IR. In proc of the 2015 international conference on the theory of information retrieval.191-200: ACM.

[35] Mesgar, M., & Strube, M. (2015). Graph-based coherence modeling for assessing readability. In proc of the fourth joint conference on lexical and computational semantics. 309- 318.

[۳۶] عبدالهی، م، و زاهدی، م (۱۳۹۶). بهبود روش‌های ارزیابی انسجام متن با ترکیب مزایای سه رویکرد مبتنی بر موجودیت، گراف و آنتروپی. سومین کنفرانس بین المللی بازشناسی الگو و تحلیل تصویر ایران.

[37] Xiong, D., Zhang, M., & Wang, X. (2015). Topic-based coherence modeling for statistical machine translation. IEEE/ACM transactions on audio, speech and language processing (TASLP), Vol. 23, No. 3, pp. 483-493.

[38] Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. in proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers. 950- 961.

[39] Nguyen, D. T., & Joty, S. (2017). A neural local coherence model. In proc of the 55th annual meeting of the association for computational linguistics. 1, 1320-1330.

[40] Logeswaran, L., Lee, H., & Radev, D. (2016). Sentence ordering using recurrent neural networks, arXiv preprint arXiv:1611.02654.

[41] Logeswaran, L., Lee, H., & Radev, D. (2018). Sentence Ordering and Coherence Modeling using Recurrent Neural Networks. arXiv:1611.02654 [cs.CL].5285-5292.

[42] Li, J., & Jurafsky, D. (2016). Neural net models for open-domain discourse coherence. arXiv preprint arXiv:1606.01545.

[43] Kiddon, C., Zettlemoyer, L., & Choi, Y. (2016). Globally coherent text generation with neural checklist models. In proc of the 2016 conference on empirical methods in natural language processing. 329- 339.

[44] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[45] Zhang, X., & LeCun, Y. (2015). Text understanding from scratch, arXiv preprint arXiv:1502.01710.

[46] Christensen, J., Soderland, S., & Etzioni, O. (2013). Towards coherent multi-document summarization. In proc of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies. 1163- 1173.

[47] Zhang, R., Li, W., Liu, N., & Gao, D. (2016). Coherent narrative summarization with a cognitive model. Computer speech & language, Vol. 35, pp. 134-160.

[48] Gambhir, M., &Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial intelligence review, Vol. 47, No. 1, pp. 1-66.

[49] Liu, P. J. (2018). Generating Wikipedia by summarizing long sequences, arXiv preprint arXiv:1801.10198.

[۵۰] کیومرثی فرشاد، بختیاری شقایق، هادی پور مریم، خلاصه سازی خودکار متن با استفاده از کلونی زنبورعسل، مجله علمی پژوهش در علوم رایانه، شماره ۶، زمستان ۱۳۹۶، ص۴۶-۵۹.

[۵۱] محسنی علیرضا، ونوس مرضی، جدیدی نژاد امیر حسین، خلاصه سازی تک سندی متون فارسی به کمک یادگیری عمیق ماشینی، مجله علمی پژوهش در علوم رایانه، شماره ۱۲، زمستان ۱۳۹۲، ص۱-۱۶.

[52] Han, A. F., & Wong, D. F. (2016). Machine translation evaluation: A survey, arXiv preprint arXiv:1605.04515.

ISSN: 2745-3673
http://www.Rcsj.ir

[53] Sim Smith, K. (2018). Coherence in Machine Translation. Doctor of Philosophy thesis. University of Sheffield.

[54] Smith, K.S., Aziz, W., & Specia, L. (2015). A proposal for a coherence corpus in machine translation. in proc of the second workshop on discourse in machine translation. 52-58.

[55] Smith, K. S., Aziz, W., & Specia. L. (2016). The trouble with machine translation coherence. in proc of the 19th annual conference of the European association for machine translation. 178- 189.

[56] Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., & Carin, L. (2017). Adversarial feature matching for text generation. arXiv preprint arXiv:1706.03850.

[57] Siddharthan, A. (2014). A survey of research on text simplification. ITL-International journal of applied linguistics, Vol. 165, No. 2, pp. 259-298.

[58] Song, W., Fu, R., Liu, L., & Liu, T. (2015). Discourse element identification in student essays based on global and local cohesion. In proc of the 2015 conference on empirical methods in natural language processing. 2255-2261.

[59] Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. in international conference on machine learning. 957- 966.

[60] Lee, G. H., & Lee, K. G. (2017). Automatic text summarization using reinforcement learning with embedding features. In proc of the eighth international joint conference on natural language processing. 2, 193-197.

[61] Severyn, A., & Moschitti, A. (2015). Learning to rank short text pairs with convolutional deep neural networks. in proc of the 38th international ACM SIGIR conference on research and development in information retrieva: ACM.

[62] Severyn, A., & Moschitti, A.(2016). Modeling relational information in question-answer pairs with convolutional neural networks, arXiv preprint arXiv:1604.01178.

[63] Vijayarani, S., Ilamathi, M. G., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International journal of computer science & communication networks, Vol. 5, No. 1, pp. 7- 16.

[64] Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Political analysis, Vol. 26, No. 2, pp. 168- 189.

[65] Simard, M. (1998). Automatic insertion of accents in French text. in proc of the third conference on empirical methods for natural language processing. 27-35.

[66] https://developer.syn.co.in/tutorial/bot/oscova/pretrained-vectors.html

[67] Mikolov, T., & Sutskever, I. (2013). Distributed representations of words and phrases and their compositionality. in proc. NIPS 2013. 3111–3119.

[68] Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. Computer speech & language, Vol. 10, No. 3, pp. 187-228.

[69] Ermakova, L., Mothe, J., & Firsov, A. (2017). A Metric for sentence ordering assessment based on topic-comment structure. in proc of the 40th international ACM SIGIR conference on research and development in information retrieval. 1061-1064: ACM.