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

مجلـه‌ی علمی پژوهش در علوم رایانه (سال سوم)
شماره ۱۲ / زمستان ۱۳۹۷ / ص ۱۶-۱

خلاصه‌سازی تک سندی متون فارسی به کمک یادگیری عمیق ماشینی

علیرضا محسنی [۱]، ونوس مرضی [۲]، امیرحسین جدیدی نژاد [۳]

[۱] دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی، واحد تهران غرب، تهران، ایران.

[۲] استادیار، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد تهران غرب، تهران، ایران.

[۳] استادیار، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد قزوین، قزوین، ایران.

نام نویسنده مسئول:
علیرضا محسنی

چکیده

در سال‌های اخیر نرخِ رشدِ اطلاعات، بسیار فزاینده بوده و نیاز به سیستم‌های خلاصه‌ساز متن احساس می‌شود. خلاصه‌سازی خودکار سند به معنی تولید یک نسخه مختصرتر از سند اصلی با حفظ نکات کلیدی به وسیله کامپیوتر می‌باشد. روش‌های مختلفی جهت خلاصه‌سازی متون فارسی تاکنون به کار گرفته شده است که از روش‌های استخراجی استفاده می‌نمایند که در این روش جملات مهم براساس الگوریتم‌هایی استخراج و به یکدیگر متصل می‌گردند همچنین در آخرین روش‌ها برای بهبود نتایج از یادگیری ماشینی استفاده شده ، لذا تاکنون در زمینه‌ی خلاصه‌سازی به کمک یادگیری عمیق ماشینی کمتر کاری انجام شده است. یادگیری عمیق ماشینی از شبکه‌های عصبی عمیق مصنوعی استفاده می‌نماید که نتایج بسیار خوبی در مقایسه با روش‌های مرسوم از خود نشان داده است. ما جهت خلاصه‌سازی متون فارسی از شبکه‌های عمیق انکودر-دکودر استفاده نمودیم که این دسته از شبکه‌ها قابلیت بالایی در استخراج ویژگی دارا می‌باشند. ما جهت ایجاد زیرساخت‌های نرم افزاری لازم از آخرین فناوری شرکت گوگل یعنی تنسورفلو استفاده نمودیم همچنین یکی از دستاوردهای این تحقیق توسعه یک دیتاست فارسی سازگار با مدل نهایی ما بر پایه مجموعه همشهری می باشد و بر اساس آن شبکه خود را آموزش دادیم.

واژگان کلیدی: خلاصه‌سازی خودکار متن، تسورفلو، یادگیری عمیق، شبکه عصبی، انکودر-دکودر، شبکه عصبی بازگشتی.

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

مقدمه

در عصر حاضر با توجه به انفجار اطلاعات نیاز به سیستم‌هایی جهت خلاصه‌سازی اطلاعات و متون که از کارایی معقولی نیز برخوردار باشند احساس می گردد. در سالهای گذشته تلاشهای زیادی در جهت ایجاد سیستم‌های خلاصه‌ساز متن در زبان فارسی و سایر زبانها صورت پذیرفته که اکثر آنها از روش‌های استخراجی استفاده می‌نمودند در روش مذکور براساس الگوریتم‌هایی، به جملات امتیازاتی داده شده و جملات مهم استخراج و در نهایت به یکدیگر چسبانده تا جمله نهایی ایجاد گردد. روش دیگری در بحث خلاصه‌سازی وجود داشته که به روش چکیده معروف می‌باشدمی‌باشد این روش تلاش می نماید تا درک درستی از مفاهیم یک سند پیدا کرده سپس به بیان این مفاهیم به زبان طبیعی می‌پردازد. هدف از این خلاصه‌سازی استخراج مفهوم اصلی مـتن ورودی بوده بـی‌آنکه لزومـا از جملات متن اصلی استفاده نماید. اکثر روشهایی که در جهت خلاصه‌سازی متون فارسی بکار گرفته شده استخراجی می‌باشند؛ چندین تحقیق نیز در زمینه‌ی یادگیری ماشینی جهت خلاصه‌سازی انجام گرفته که اطلاعات دقیقی از نحوه اجرای دقیق آن موجود نیست[72]. با پیشرفتهایی که در زمینه یادگیری عمیق ماشینی صورت گرفته امکان طراحی عمیق‌تر شبکه‌ها به وجود آمده و در تعدادی از تحقیقات خارجی از این فناوری جهت خلاصه‌سازی استفاده شده است. بیشتر مطالعاتی که در زبان فارسی انجام شده از روش استخراجی استفاده می‌نمایند و تاکنون در زمینه‌ی خلاصه‌سازی به کمک یادگیری عمیق ماشینی کمتر تحقیقی به چشم می‌خورد. یادگیری عمیق انتخاب روشهایی برای ساخت مدل‌های یادگیری ماشین است که درشرایط پیچیده مفید می‌باشند[58]. یادگیری عمیق[1] اشاره به مجموعه‌ای از الگوریتم‌های یادگیری ماشین داردکه معمولا مبتنی بر شبکه‌های عصبی مصنوعی بوده و در تلاش‌اند تا انتزاعات سطح بالای موجود در داده‌ها رامدل نمایند. یادگیری عمیق به آموزش یک شبکه عصبی چند لایه با استفاده از تکنیک مبتنی بر شیب گویند[60]. ما در این تحقیق خلاصه‌سازی انتزاعی متن را با استفاده از شبکه‌های عصبی عمیق بازگشتی انکودر – دکودر مدل می‌نماییم. رمزگذار ما با استفاده از رمزگذار مبتنی بر مراقبت[48] مدلسازی شده است. این شبکه‌ها داده محور بوده و به‌صورت خودکار مهندسی ویژگی ها را انجام داده و انسان دخالتی در آن ندارد. در این تحقیق با بکارگیری شبکه‌های انکودر دکودر عمیق و استفاده از آخرین فناوری شرکت گوگل (تنسور فلو) به پیاده سازی یک شبکه عمیق جهت یادگیری و خلاصه‌سازی متون فارسی پرداخته و انکودر-دکودر را با کمک شبکه‌های عصبی بازگشتی دارای حافظه LSTM[2] ایجاد می‌نماییم انکودر دکودر یک شبکه عصبی مصنوعیبوده که جهت کدینگ از آن استفاده می‌گردد و دارای ۲ قسمت حیاتی با نامهای انکدر و دکودر یا رمزنگار و رمزگشا می‌باشد. از خود رمزگذارها برای فشرده‌سازی نمایش داده‌ها یا به عبارت دیگر تقلیل ابعاد استفاده می‌شود. انکودر-دکودر نوع خاصی از شبکه عصبی مصنوعی است که برای انکد کردن بهینه یادگیری مورد استفاده قرار می‌گیرد[65] به‌جای آموزش شبکه و پیش بینی مقدار هدف Y در ازای ورودی X یک انکودر-دکودرآموزش می‌بینید تا ورودی X خود را بازسازی کند. به‌صورت خلاصه می‌توان گفت این گونه شبکه‌ها قابلیت استخراج ویژگی‌ها را به صورت مطلوب دارا می‌باشند؛ همانطور که گفتیم انکودر-دکودر را با کمک شبکه‌های عصبی بازگشتی یا RNN ایجاد می‌نماییم. شبکه‌های مذکور مدل‌های یادگیری عمیقی هستند که قادر بوده جملات دارای طول متغیر را پردازش کرده و نمایش‌های سودمندی (حالت مخفی) را برای هر عبارت ایجاد نمایند. این شبکه‌ها هر عنصر دنباله (یعنی هر کلمه) را به شکل یک به یک پردازش کرده و برای هر ورودی جدید در دنباله، شبکه یک حالت مخفی جدیدی را به صورت تابعی از ورودی و حالت مخفی پیشین به عنوان خروجی می‌دهد از این نقطه نظر حالت مخفی محاسبه شده در هر کلمه، تابعی است از همهٔ کلمات خوانده شده تا به آن نقطه[45]، همچنین جهت آموزش شبکه از دیتاست اختصاصی خود که شامل متون فارسی و خلاصه‌ای از آنها می‌باشد استفاده می‌کنیم دیتاست مذکور براساس مجموعه همشهری و با توسعه برنامه اختصاصی تهیه گردیده و جهت مدل مورد نظر استفاده می‌شود. ساختار اصلی این مقاله به شرح ذیل می‌باشد، در بخش دوم این مقاله به مروری بر کارهای گذشته که در رابطه با خلاصه‌سازی متون در زبان فارسی و سایر زبان‌ها صورت گرفته است پرداخته و در بخش سوم به توصیف روش پیشنهادی خود جهت ایجاد سیستم خلاصه‌ساز خواهیم پرداخت و در ادامه مدل ریاضی خود را توصیف نموده و در بخش پنجم این تحقیق به تنظیم پارامترهای اساسی سیستم جهت خلاصه‌سازی همچنین اجرا و ارزیابی سیستم می‌پردازیم.

مروری بر کارهای گذشته

تاریخچه سیستم‌های خلاصه‌ساز متن به سال ۱۹۵۰برمی‌گردد که استخراج متن‌خلاصه براساس‌تعداد تکرار کلمه و در حقیقت بر پایه‌ی روش آمـاری بـوده اسـت. به دلیل کمبود کامپیوترهای قدرتمند و مشکلات موجود برای پردازش زبان‌های طبیعی، کارهای اولیه بر روی مطالعه ظواهر متن مانند (موقعیت جمله و عبارات اشاره) متمرکز شده بود. سال ۱۹۷۰ تا ۱۹۸۰ از هوش مصنوعی استفاده گردید و ایده آن استخراج نمایش‌های دانش، مانند فریم ها یا الگوها، برای شناسایی موجودیت‌های مفهومی از متن و استخراج روابط بین

[1] Deep Learning
[2] Long Short Term Memory

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

موجودیت‌ها با مکانیزم‌های استنتاج بود. مشکل اصلی آن‌ها را می‌توان به این گونه عنوان نمود که فریم یا الگوهای تعریف شده محدودیت‌هایی دارند و ممکن است به تحلیل کامل موجودیت‌های مفهومی منجر نشود [27]. از اوایل ۱۹۹۰ تاکنون روش‌های بازیابی اطلاعات بکار گرفته شده است. روش[28] براساس مقادیر ویژگی‌های یک جمله احتمال حضور آن را در خلاصه نهایی تخمین می‌زند. در روش مذکور عمل خلاصه‌سازی به صورت یک مسئله دسته‌بندی درنظر گرفته شده و دسته‌بندی کننده‌های بیزین برای تعیین جملاتی که باید در خلاصه وارد شوند بکار گرفته می شوند. در مرجع [29]چندین الگوریتم مانند درخت تصمیم و دسته‌بندی کننده را برای استخراج قطعات جمله پیشنهاد دادند . این روش خلاصه‌سازی اسناد، در یک حوزه خاص عملکرد خوبی داشته گرچه برای یادگیری صحیح نیازمند مجموعه‌ای آموزشی بسیار بزرگ می‌باشد. مرجع[30] روشی برای تولید خلاصه با پیدا کردن زنجیره‌های لغوی معرفی کرد که به توزیع کلمه و اتصالات لغوی بین آنها برای تقریب زدن محتوا و ارائه یک نمایش از ساختار لغوی بهم پیوسته متن اتکا می‌کرد. کارهای انجام شده در زمینه‌ی پیاده‌سازی ابزارهای خلاصه‌سازی خودکار فارسی بسیار اندک می‌باشد. در ادامه آنها به طور مختصر مورد اشاره قرار خواهند گرفت. FarsiSum یک خلاصه‌ساز مبتنی بر وب است که SweSum را برای زبان فارسی مدل می‌نماید [32]. این سیستم قادر به خلاصه کردن متون روزنامه‌های فارسی با قالب HTML و متن کد شده با فرمت یونیکد می‌باشد و دارای خروجی از نوع گزینشی بوده و این سیستم جهت خلاصه‌سازی اخبار بکار می‌رود و همواره به اولین جملات و اولین پاراگراف‌ها وزن بیشتری اختصاص داده و تحت سیستم عامل ویندوز و به زبان پرل توسط[25] ارائه شده و معروف‌ترین سیستم خلاصه‌ساز فارسی می‌باشد. سیستم خلاصه‌ساز دیگری که توسط [۲]ارائه گردیده تک سندی و بر مبنای گزینش جمله‌ها عمل می‌نماید. این روش در گزینش جملات خروجی، ترکیبی از دو روش زنجیره‌ی لغوی و نظریه‌ی گراف می‌باشدبوده و خروجی نهایی می‌تواند کلی یا براساس پرس وجوی کاربر باشد. از جمله روش‌های خلاصه‌سازی چندسندی در زبان فارسی می‌توان به مرجع[۸] اشاره کرد که خروجی تولید شده توسط آن از نوع چکیده‌ای می‌باشد. در مرجع[۹] پس از مرور شیوه‌های خلاصه‌سازی موجود، گزارشی از پیاده‌سازی یک نمونه‌ی عملی برای خلاصه‌سازی متون فارسی ارائه گردیده است. درمرجع[۱۰] پس از بررسی چالش‌های مربوط به پردازش متون فارسی، انواع روش‌های موجود در خلاصه‌سازی مورد بررسی قرار گرفته است. در مرجع[۱۱] یک نمونه ابزار برای خلاصه‌سازی متون فارسی پیاده‌سازی شده است. این ابزار برای خلاصه‌سازی چندسندی متون فارسی مورد استفاده قرار گرفته، خروجی آن گزینشی بوده و از یک روش مبتنی بر خوشه‌بندی بهره می‌گیرد. درمرجع[33] یک سیستم خلاصه‌ساز چندسندی چند زبانی ارائه شد که براساس SVR[۳] عمل نموده و خلاصه‌ی حاصله از نوع گزینشی می‌باشد، روش مورد استفاده جزء دسته آماری محسوب می‌شود. مبنای این روش استفاده از روش سلسله مراتبی اتصال میانگین برای خوشه‌بندی جمله‌ها و بکارگیری روش تجزیه به مقادیر منفرد برای تعیین اهمیت جمله‌هاست. در این کار از دو منبع زبانی ساده یکی برای قطعه‌بندی متن به واژه‌ها و دیگری برای تعیین فراوانی واژه‌ها در سندها استفاده شده است. در مرجع[34] نیز با بهره‌گیری از شکل جدیدی از روش استخراج روابط معنایی موجود در متن (LSA[۴]) و تکنیک برچسب‌زنی معنایی نقش لغات (SRL[۵])، روشی جدید برای خلاصه‌سازی چندسندی ارائه گردید. در مقاله [41] به کمک یادگیری ماشینی شیوه‌ای برای خلاصه‌سازی متون فارسی پیشنهادشده است. الگوریتم طراحی شده درمرحله‌ی اول هر کلمه‌ای که در یک متن باید تفکیک شود را مشخص کرده و سپس پایگاه داده کلماتی که باید حذف شوند را به وجود می‌آورد؛ مانند چه‌کسی، بودن، سلام، بیشتر، اگر و غیره. این کلمات تکراری بدین دلیل می‌بایست حذف گردند که تنهایکبار در خلاصه قرار گیرند. مجموعه‌ای از کلمات باارزش به وجود می‌آید که در آن‌ها ارزش هر کلمه برابر با تعداد تکرار در یک متن می‌باشد. مرحله دوم تشخیص جملات متن است. دراین بخش کاراکترهای موجود دریک جمله می‌بایست بررسی شده و براساس آن‌ها تمامی کاراکترهای پیش از نقطه به‌عنوان جمله درنظرگرفته شده و آنالیز جمله در مراحل بعدی انجام می‌شود. ابعاد بردار برابر است با تعداد کلمات برای هر جمله و در نهایت وزنی به هر عنصر تخصیص داده می‌شود. بدین ترتیب ماتریسی براساس تعداد جملات در ردیف و تعداد کلمات کلیدی در ستون ایجاد می‌شود. در مرحله چهارم دسته‌بندی جمله بااستفاده از الگوریتم k-means انجام می‌شود .همچنین تحقیق دیگری که در سال ۲۰۱۴ در خصوص خلاصه‌سازی متون فارسی به کمک یادگیری ماشینی صورت گرفت [72] می‌باشدبدینگونه بود که در سیستم پیشنهادی ورودی سیستم براساس ویژگی‌هایی مانند شباهت با عنوان و مرکز متن، طول جمله همچنین موقعیت جمله استخراج و امتیاز دهی می‌شوند. معماری پیشنهادی به کمک شبکه‌های عصبی بوده و از دادگان همشهری جهت آموزش سیستم استفاده شده است. در راستای این تحقیق تاکنون کار قابل توجه‌ای به کمک یادگیری عمیق ماشینی صورت نگرفته است. تحقیقات خارجی زیادی در خصوص سیستم‌های خلاصه‌ساز صورت گرفته اما در ادامه تحقیقاتی را که به موضوع این تحقیق نزدیکتر می‌باشد معرفی می‌نماییم. باظهور یادگیری عمیق به عنوان یک راهکار مناسب برای بسیاری از وظایف NLP، محققان این چارچوب را بسیار باارزش یافته و آن را یک گزینه داده - محور برای

[3]Support Vector Regressio
[4]Latent semantic analysi
[5] Statistical relational learning

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

خلاصه‌سازی انتزاعی قلمداد کردند. این تحقیق[51] از یک مدل انکودر-دکودر استفاده کرده اما دکودر را با یک RNN جایگزین نموده است. درمقاله‌ی دیگری[52] یک مجموعه داده بزرگ برای خلاصه‌سازی متن کوتاه چینی به وجود آورده و نتایج امیدوارکننده‌ای را برای مجموعه داده چینی با استفاده از یک RNN انکودر-دکودر به دست آمد اما آزمایشات بر روی متون انگلیسی زبان صورت نگرفت. درمطالعه‌ی دیگری[51] از انکودر-دکودر مبتنی بر RNN برای خلاصه‌سازی استخراجی اسناد استفاده شد. این مدل دقیقاً مشابه مدل ما نیست چراکه چارچوب آن‌ها استخراجی و چارچوب کاری روش پیشنهادی ما انتزاعی می‌باشد. شبکه‌های اشاره‌گر [55] نیز پیش‌تر برای مسئله کلمات نادر درحوزه‌ی ترجمه‌ی ماشینی به‌کار گرفته می‌شدند.

روش پیشنهادی

اساس و پایه اصلی استراتژی ما در این تحقیق استفاده از یادگیری عمیق می‌باشد یادگیری عمیق انتخاب شیوه‌هایی برای ساخت مدل‌های یادگیری ماشینی بوده که در شرایط پیچیده مفید می‌باشند[58]. یادگیری عمیق در واقع همان یادگیری به وسیله شبکه‌های عصبی بوده که دارای لایه‌های پنهانی زیادی می‌باشند. به هر میزان در لایه‌های یک شبکه عصبی عمیق جلوتر می‌رویم، به مدل‌های پیچیده‌تر دست خواهیم یافت. با عمیق‌تر شدن شبکه‌ها و سلسله مراتبی که توسط لایه‌های مختلف ارائه می‌گردد می‌توان سطوح انتزاع مختلفی به‌دست آورد. این شبکه‌ها داده محور بوده و به‌صورت خودکار مهندسی ویژگی‌ها در آنها صورت گرفته و انسان دخالتی در آن ندارد. ما به کمک شبکه‌های بازگشتی انکودر-دکودر قصد ایجاد یک سیستم خلاصه‌ساز در زبان فارسی را داریم، این نوع شبکه‌ها قابلیت استخراج ویژگی‌ها را دارا می‌باشند. ویژگی‌هایی که در جهت خلاصه‌سازی متون فارسی مورد نیاز می‌باشد شامل مواردی همچون ارتباط بین کلمات، جملات و کلمات کلیدی‌تر در متن هستند. در این مطالعه برای ایجاد زیرساخت‌های لازم از تنسور فلو استفاده می‌نماییم. تنسورفلو یک کتابخانه‌ی نرم‌افزاری منبع باز قدرتمند برای محاسبات عددی بوده که به طور خاص برای یادگیری ماشین در مقیاس بزرگ ارائه و تنظیم شده است. همچنین مدل اولیه جهت خلاصه‌سازی برگرفته از آخرین تحقیقات صورت گرفته توسط آقایان زین پن و پیتر لیو از اعضای تیم تحقیقاتی مغز گوگل بوده که با توسعه سیستمی به کمک فناوری تنسورفلو همچنین با کمک گرفتن از زبان پایتون به خلاصه‌سازی متون در زبان انگلیسی پرداخته‌اند که از طریق آدرس اینترنتی ذیل در دسترس می‌باشد:

https://github.com/tensorflow/models/tree/master/research/textsum

همچنین از مکانیزم اصلی سیستم هیچگونه اطلاعاتی در دسترس نیست. سایر بسترهای نرم افزاری و سخت افزاری لازم جهت سیستم نهایی به شرح ذیل است:

جدول (۱) : نرم افزارهای مورد نیاز

نام	نسخه
سیستم عامل لینوکس	اوبونتو ورژن ۱۶
پایتون	نسخه ۲,۷ و ۳,۵
تنسور فلو	نسخه ۱
درایور های کارت گرافیک جهت استفاده از فناوری کودا و...	بسته به کارت گرافیک در این تحقیق از کارت گرافیک انویدیا تیتان اکس استفاده شد
آناکوندا	ورژن در دسترس
نصب NLTK	زبان‌های فارسی، هندی، شرق آسیا
نصب محیط مجازی	
نصب کتابخانه Numpy	

یکی از ارکان مهم در بحث یادگیری عمیق استفاده از توان محاسباتی کارت‌های گرافیکی مجهز به فناوری کودا می‌باشد. کودا دارای معماری بر پایه پردازش موازی است که این فناوری توسط شرکت انویدیا ابداع شد. در واقع کودا یک موتور قدرتمند محاسباتی(پردازش) بر مبنای کارت گرافیک است و همانطور که میدانیم اساس مبحث یادگیری عمیق، شبکه‌های عصبی عمیق بوده و اتفاقی هم که در این شبکه‌ها رخ میدهد چیزی جز محاسبات ریاضی و بطور خاص ماتریسی در مقیاس بزرگ نمی‌باشد. به همین دلیل اگر محاسبات مورد نیاز شبکه عصبی عمیق توسط پردازنده‌های معمولی صورت پذیرد نیاز به صرف زمان زیادی خواهد بود. در این مقاله محاسبات شبکه به کمک یک سیستم مجهز به کارت گرافیکی با مشخصات ذیل انجام گرفته است:

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

جدول (۲) : مشخصات کارت گرافیک

	مدل	مشخصات
کارت گرافیک	Geforce Titan x	CUDA ۳۰۷۲

مدل اساسی تحقیق به این شرح می‌باشد که مجموعه‌ای از داده‌های آموزشی با ساختار ذیل به عنوان ورودی شبکه تزریق تا ویژگی‌های موجود استخراج و شبکه آموزش ببیند . ویژگی‌های مورد نظر می‌تواند شامل ارتباط بین لغات و جملات، موضوع اصلی متن، همچنین موقعیت جمله باشد.

جدول (۳) : ساختار دیتاست

متن کامل	براساس آخرین تحلیل داده‌ها و نقشه‌های پیش‌بینی هواشناسی از بعدازظهر امروز پنج‌شنبه با شمالی شدن جریان هوا در شمال کشور و گذر موجتر از میانی جو، از سواحل غربی خزر ابرناکی و بارش باران شروع می‌شود. بارندگی روز جمعه.....
تگ	@Highlight
عنوان خبر	تشدید بارش‌ها از امشب در شمال و شمال‌شرق/ بارش برف در جاده‌های کوهستانی و کاهش دما

هر رکورد داده آموزشی از۲ قسمت اصلی تشکیل شده ، بخش ابتدایی که همان متن کامل بوده و بخش دوم که بعد از @highlightدرج شده به عنوان خلاصه همان متن در نظر گرفته می‌شود داده‌های مورد نیاز از مجموعه دادگان همشهری استخراج شده که مجموعه فوق پیکره‌ای است حاوی ۳۱۸ هزار سند مربوط به اخبار سال‌های ۱۳۷۵ تا ۱۳۸۶. موضوع قابل توجه این مهم می‌باشد که ورودی شبکه ما در اصل همان ۲ سطر ابتدایی هر متن کامل یا ۱۲۰ کلمه ابتدایی هر متن اصلی به همراه بخش بعد از @highlight یا همان تیتر یک متن می‌باشد. دلیل انتخاب ۲ سطر ابتدایی از متن اصلی به این دلیل بوده که معمولا ۲ سطر ابتدایی هر متن به خلاصه داستانی را که آن متن حول و حوش آن می‌چرخد اشاره می‌نماید از طرفی با این تکنیک سربار داده‌ای جهت ورودی شبکه کاهش یافته و به تعداد لایه‌ها و عناصر کمتر در شبکه کمتر مورد نیاز می‌باشد. لازم به ذکر است که دیتاست همشهری با مدل ما خوانایی نداشته لذا برنامه اختصاصی برای پردازش دیتاست مذکور و تبدیل آن به مدل مورد نظر ما توسعه داده شد. در نهایت شبکه با مقایسه‌ی ۲ سطر ابتدایی هر متن کامل و عنوان آن، به استخراج ویژگی‌ها و الگوها اقدام و عمل خلاصه‌سازی را آموزش می‌بینند. همچنین دیتاست به فایلی جهت استفاده در تنسور فلو تبدیل گشت. یکی از ارکان اصلی این تحقیق براساس دامنه لغات می‌باشد به طوری که در زمان آماده‌سازی داده‌ها یکی از وظایف مهمی که می‌بایست انجام شود استخراج تمامی لغات موجود در دیتاست به همراه میزان تکرار هر لغت در کل اسناد می‌باشد. به طوری که در زمان آموزش سیستم نسبت به بررسی درجه اهمیت هر کلمه با مراجعه به فایل مذکور اقدام می‌نماید. دامنه لغات با نام vocab شناخته شده و مقدار متغیر این پارامتر در این تحقیق ۲۰۰ هزار کلمه می‌باشد. جهت استخراج لغات از برنامه‌ای که به زبان پایتون توسعه داده شده است استفاده می‌گردد مهمترین بخش این برنامه از کتابخانه Numpy جهت استخراج لغات و تعیین مقدار تکرار های آن استفاده شده است.

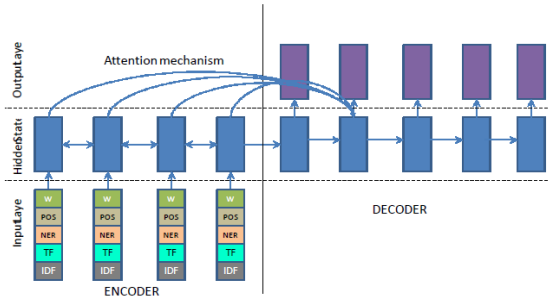
جدول (۴) : فرکانس تکرار لغات

ردیف	کلمه	میزان تکرار در اسناد
۱	و	۱۳۳۶۹۹۳
۲	در	۱۰۴۰۲۳۹
۳	به	۸۶۶۷۵۲
۴	از	۶۷۰۵۳۹
۵	که	۵۰۷۴۱۳
۶	این	۵۰۳۷۰۸
۷	را	۳۹۳۷۷۱
...	...	...
۱۹۷۹۸۸	دفاع و	۱

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

۱۹۷۹۸۹	پارلمان مرکزی	۱
۱۹۷۹۹۹	یانکرها	۱
۱۹۸۰۰۰	مفاد پرونده	۱

جهت ایجاد یک انکودر-دکودر بازگشتی از شبکه‌های عصبی بازگشتی RNN که دارای حافظه LSTM می‌باشد استفاده نمودیم. شبکه‌های عصبی بازگشتی RNN[6] مدل‌های یادگیری عمیقی هستند که می‌توانند جملات دارای طول متغیر را پردازش کرده و نمایش‌های سودمندی (حالت مخفی) را برای هرعبارت به دست بیاورند[45]. این شبکه‌ها در واقع برای پردازش سیگنال‌های دنباله دار به وجود آمدند. از این نظر حالت مخفی محاسبه شده در هر کلمه تابعی است از همهٔ کلمات خوانده شده تا به آن نقطه[45] و با کمینه‌سازی خطای نوسازی بهینه می‌شود. کد متناظر همان ویژگی فراگرفته شده است. فرآیند کلی این دسته از شبکه‌ها به این شرح می‌باشد که یک لایه‌ی تنها قادر به دریافت ویژگی‌های متمایز از داده خام نیست برای دریافت و استخراج ویژگی‌های خاص می‌توان از انکودر-دکودر عمیق استفاده نمود که در این نوع از شبکه‌ها کد یادگرفته شده از لایه قبلی را به لایه بعدی جهت به انجام رساندن کار خود ارسال می‌نمایند. اندازه‌ی لایه مخفی در این پژوهش ۴ لایه می‌باشد. در مدل فوق لایه مخفی از یک انکودر متشکل از یک GRU-RNN[7] دوجهتی تشکیل شده و همچنین دکودر متشکل از یک GRU-RNN تک جهتی می‌باشد.



شکل (۱) : معماری انکودر-دکودرجهت خلاصه‌سازی-[32]

دراین تحقیق، یکی از مهمترین قسمت‌ها بخش پیش پردازش متون ورودی می‌باشد که می‌بایست کلمات اضافی، تکراری و ... جدا و اهمیت کلمات نیز استخراج گردند. به دلیل عدم رعایت استانداردهای نگارشی باید پیش‌پردازش بر روی داده‌های ورودی صورت گیرد. درمرحله‌ی اول هر کلمه‌ای که در یک متن وجود دارد می‌بایست تفکیک شده و در ادامه باید حذف شود، مانند چه کسی، بودن، سلام، بیشتر، اگر و غیره. این کلمات تکراری با این هدف می‌بایست حذف شوند که تنها یکبار در خلاصه قرار بگیرند. به صورت خلاصه می‌توان گفت ریشه یابی، حذف کلمات توقف، برچسب‌زنی نقش کلمات و تعیین کلمات کلیدی مهمترین عملیاتی مربوط به خلاصه‌سازی متون هستند که این دسته عملیات را پردازش زبان طبیعی گویند؛ در حال حاضر یکی از ابزار موجود جهت زبان فارسی با نام هضم می‌باشد. هضم ابزاری برای اصلاح نویسه‌ها، تقطیع جمله‌ها و واژه‌ها، ریشه‌یابی کلمات، برچسب‌زنی اجزای سخن، تجزیهٔ وابستگی و واسطی برای پیکره‌های بیجن‌خان و همشهری است که سازگار با بستهٔ NLTK بوده و از پایتون پشتیبانی می‌نماید. در خلاصه‌سازی یکی از چالش‌های کلیدی دیگر شناسایی مفهوم کلی و ماهیت‌های کلیدی در سند است که مرکزیت موضوع حول آن می‌چرخد. به منظور دستیابی به این هدف، نیاز به تعیین ویژگی‌های زبانی و ذخیره‌ی آنها مانند تگ‌های اجزای سخن، آمارهایTF و IDF کلمات داریم درنتیجه ماتریس‌های قابل جستجو را برای کلمات هر نوع تگ ایجاد می‌کنیم.

روش TF-IDFدر این شیوه میزان تکرار یک کلمه در یک مستند را در مقابل تعداد تکرار آن در مجموعه کلیه مستندات در نظر می گیریم.

در روشTF-IDF وزن‌دهی کلمات تابعی از توزیع کلمات مختلف در مستندات است.برای پیاده‌سازی در ابتدا یک مجموعه اسناد را در نظر گرفته و به ازای تمام کلماتی که در پیکره وجود دارد ، بررسی می‌نماییم که هر کلمه در چه تعداد از سندها تکرار شده و آن را ذخیره می‌کنیم. سپس یک سند به عنوان ورودی دریافت می‌شود (هدف یافتن کلمات کلیدیِ سند دریافت شده است) برای این منظور ابتدا بررسی می‌شود که هر یک از کلمات سند ورودی چند بار در همان سند استفاده شده است و سپس به ازای تمام کلمات سند ورودی

[6]recurrent neural network
[7]Gated. Recurrent Unit *recurrent neural network*

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

بررسی می‌کنیم که هر کلمه در چه تعداد از اسناد پیکره اصلی وجوددارد، بعد از طی این مراحل به حساب کردن وزن کلمات می پردازیم، تعیین وزن کلمات با استفاده از دو معیار tf[8] و idf[9] انجام می‌شود که به شرح زیر محاسبه خواهند شد:

$$Tf(t,D)=0.5+\frac{0.5*f(t,d)}{max\{f(w,d):\omega \epsilon d\}}$$

که در آن:

f(t,d)f(t,d)تعداد تکرار کلمه t در سند d (سند هدف) است و max{f(w,d)}maxf(w,d) تعداد پر تکرارترین کلمه در سند d می‌باشد و

$$Idf(t,D)=log\frac{N}{|\{d \epsilon D:t \epsilon d\}|}$$

که در آن N تعداد کل اسناد موجود در پیکره است و

$|\{d \epsilon D:t \epsilon d\}|$

بیانگر تعداد اسنادی است که کلمه t در آنها وجود دارد.

در نهایت وزن هر کلمه به صورت زیر محاسبه خواهد شد:

$$Tfidf(t,d,D)=tf(t,d)*idf(t,D)$$

که همانطور که در شکل شماره ۱ مشخص گردیده تمام مقادیر به دست آمده به عنوان ورودی شبکه استفاده می‌گردد. در لایه مخفی می‌بایست ارتباطات بین کلمات همچنین استخراج ویژگی‌ها صورت پذیرد. با استفاده از انکدرها در لایه مخفی ویژگی‌های مهم استخراج می‌شود. ضمنا بایستی توجه نمود که اصطلاح رمزگذار نشاندهنده‌ی یک مدل زبان استاندارد می‌باشد و با ترکیب انکودر و آموزش و یکپارچه کردن آنها می‌توانیم متن ورودی را با تولید شده ترکیب نماییم. نکته قابل توجه این است که در این مرحله جملات به عنوان دنباله‌ای از کلمات تبدیل شده و انتهای هر جمله با نقطه مشخص می گردد. ارتباطات در دو سطح مشخص می شوند، در سطح جملات و در سطح کلمات و برای هر کلمه نیز یک ماتریس از ویژگی‌ها ایجاد، همچنین برای جملات نیز چنین رویکردی صورت می‌پذیرد. توزیع یکنواختی بر روی کلمات ورودی خواهیم داشت و برای خلاصه‌سازی، این مدل می‌تواند ارتباطات مهم بین کلمات و تشخیص محتوا از قبیل نوع کلمات و غیره را ذخیره نماید. برای درک کردن روابط مختلف بین کلمات از انکدر جهت ورودی‌ها استفاده می‌نماییم این معماری این امکان را می‌دهد که روابط بین کلمات و ارتباطات میان آنها به خوبی مشخص شوند. برای درک این موضوع جدول ذیل را در نظر بگیرید جمله کوتاه I enjoy flying. را در نظر بگیرید کلمات I و enjoy در کنار یکدیگر واقع شده‌اند بنابراین در سطر I و ستون enjoy عدد یک قرار داده شده است و این نشان‌دهنده‌ی همسایه بودن این ۲ کلمه می‌باشد و می‌توان بدین وسیله به‌سادگی روابط بین کلمات را استخراج نمود. در نمونه‌ی بعدی بین کلمات I و deep رابطه‌ای وجود ندارد بنابراین در سطر و ستونی که هردو همدیگر را قطع می‌نمایند عدد ۰ قرار می‌گیرد . همانطور که مشاهده می‌شود در سطر و ستون I و Like عدد ۲ درج گردیده و این موضوع نشان دهنده این مفهوم می‌باشد که در جملات ۲ بار این ۲ کلمه در کنار هم قرار گرفته‌اند. در حالت کلی به کمک این روش می‌توان از میان دیتاست‌ها ارتباط بین کلمات را به خوبی استخراج نمود؛ به عنوان نمونه در زبان فارسی بین کلمات آیت و الله رابط عمیقی وجود داشته وترتیب قرار گیری آنها به همین شیوه می‌باشد.

I Like deep learning
I Like NLP
I enjoy flying

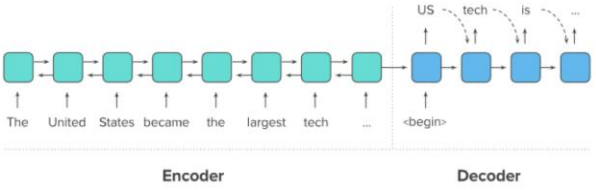
شکل (۲) : ارتباطمنطقی بین لغات در سطح جمله

Counts	i	like	enjoy	deep	learning	nlp	flying	.
i	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
nlp	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

[8]term-frequency
[9]Inverse Document Frequency

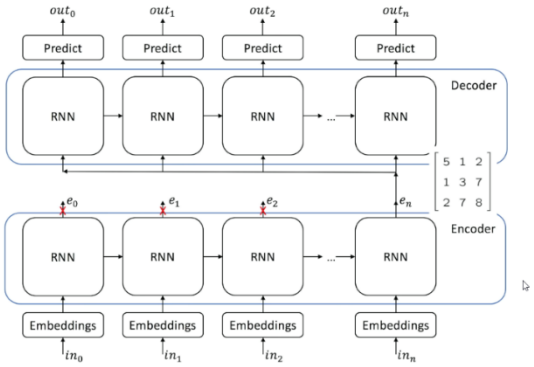
ISSN: 2745-3673
http://www.Rcsj.ir

روابط استخراج شده بین کلمات و جملات در لایه مخفی انکودر با سایر ویژگی های دیگری مانند IDF،TF ،وزن کلمات، معیارهای دوری و نزدیکی بین کلمات و جملات ترکیب شده و شبکه درک درستی از وضعیت جملات، ارتباط بین کلمات همچنین موضوع اصلی و هسته متن پیدا می نماید. در مدل ما RNNهای ورودی و خروجی با هم ترکیب می‌شوند و به یک مدل مشترک تبدیل می‌گردند که در آن حالت مخفی نهاییRNN ورودی به عنوان حالت مخفی اولیهRNN خروجی استفاده می‌شود. مدل مشترک که به این شکل همکاری می-نماید قادر می باشد هر متنی را خوانده و متن متفاوتی را از آن تولید نماید. این چارچوب با عنوان RNN انکودر-دکودر (یاseq2seq) شناخته می‌شود و مبنای مدل خلاصه‌سازی ما به شمار می‌رود همچنین در تحقیق فوق RNN انکودر سنتی را با یک انکودر دو طرفه جایگزین می‌کنیم که از دو RNN متفاوت برای خواندن دنباله ورودی استفاده می‌نماید بدین ترتیب که یکی از آن‌ها متن را از چپ به راست می‌خواند (شکل ۳) و دیگری بالعکس عمل می‌نماید. این راهکار می‌تواند به نمایش بهتر متن ورودی توسط مدل کمک نماید.



شکل (۳) : ارتباط معنایی بین انکدر - دکودر-[31]

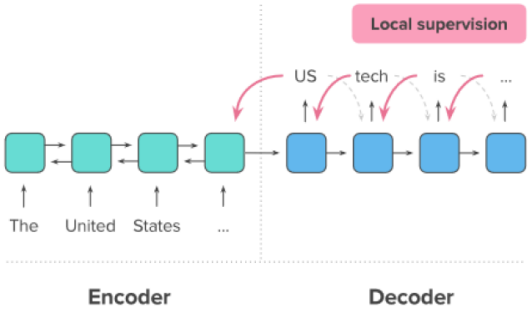
پس از استخراج ویژگی‌ها در بخش انکدر، همانطور که در شکل شماره ۴ مشاهده می‌شود ماتریس ویژگی‌ها به بخش دکودر ارسال می‌گردد.



شکل (۴) : ارسال ماتریس ویژگی ها

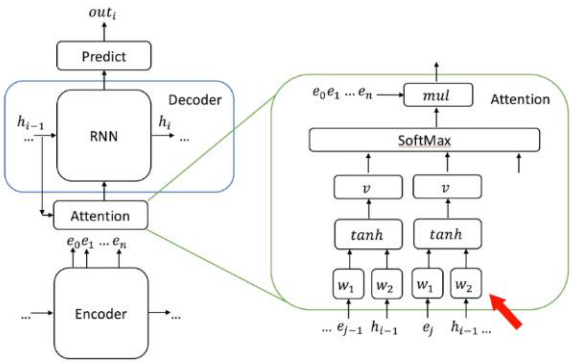
در قسمت دکودر براساس ورودی انکودرها نسبت به استخراج لغات هدف اقدام می‌گردد. یکی از مشکلاتی که در خلاصه سازها وجود دارد نادیده گرفتن لغاتی در خلاصه استخراج شده می باشد که در متن ورودی وجود داشته است.برای رفع مشکل مذکور دکودر در هنگام ایجاد یک کلمهٔ جدید از تکنیکی به نام توجه زمانی استفاده می نماید (به جای اتکا به حالت مخفی)، دکودر می‌تواند اطلاعات موضوعی را در مورد بخش‌های مختلف ورودی با تابع توجه به‌وجود آورده سپس این توجه می‌تواند مدوله شده تا اطمینان حاصل نماییم که مدل از بخش‌های مختلف ورودی در هنگام ایجاد متن خروجی بهره می‌گیرد و در نتیجه پوشش اطلاعاتی خلاصه را افزایش می‌دهد. به علاوه برای اطمینان از اینکه مدل خود را تکرار نمی‌نماید ما امکان بررسی حالت‌های مخفی پیشین از دکودر فراهم می‌کنیم. به شکل مشابه یک تابع توجه درون-دکودر را تعریف می‌کنیم که نگاهی به حالت‌های مخفی پیشینRNNهای دکودر دارد. در نهایت دکودر، ترکیبی از بردار موضوعی توجه زمانی با بردار توجه درون-دکودر برای ایجادکلمهٔ جدید در خلاصه خروجی را به‌وجود می‌آورد. شکل شماره ۵ نشان‌دهنده‌ی ترکیبی از این دو تابع توجه در مرحله رمزگشایی (دکودینگ) به دست می‌دهد.

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



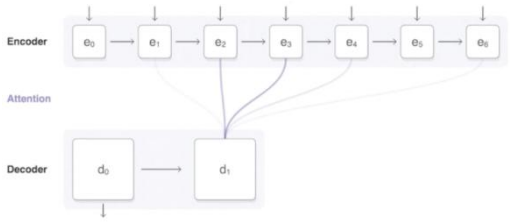
شکل (۵)- [31]

دکودر متشکل از یک GRU-RNN تک جهتی با اندازه حالت مخفی مشابه انکودر و یک مکانیسم مراقبتی برای حالت‌های منبع-مخفی و یک لایه soft-max بر روی لغات هدف برای ایجاد کلمات نهایی می‌باشد. در روش پیشنهادی رمز گشایی واژگان هر دسته کوچک محدود به کلمات سند منبع آن دسته می‌باشد. به‌علاوه رایج‌ترین کلمات در دیکشنری هدف تا زمانی اضافه می‌شوند که لغات به یک اندازه ثابت و تعریف شده در سیستم برسند. هدف از این تکنیک کاهش اندازه لایه soft-max دکودر است که مشکل محاسباتی اصلی می‌باشد.



شکل (۶) : تعیین لغات هدف

جهت استخراج لغات هدف، جمله خلاصه براساس اندازه ای که از پیش تعیین گردیده ایجاد می‌گردد. دکودر با توجه به ورودی انکدر و لغات قبلی در صف ایجاد شده و به کمک الگوریتم ویتربی و زنجیره‌های مخفی مارکوف نسبت به پیش‌بینی لغات هدف اقدام می‌نماید.



شکل (۷)- [31]

مدل ریاضی

در این قسمت بحث را با تعریف وظیفه خلاصه‌ی جمله شروع می‌کنیم. باتوجه به یک جمله ورودی، هدف ارائه‌ی یک خلاصه‌ی فشرده از جمله ورودی می‌باشد. ورودی ما جملات و دنباله‌ای از کلمات می‌باشند که از یک فهرست واژگان ثابت که آن را V می‌نامیم آمده‌اندو اندازه آن برابر است با |V|=V که آن را به شکل زیر نشان می‌دهند[50]:

x1, . . . , xM

در نهایت ما قصد داریم هرکلمه را در یک بردار نشان دهیم:

xi∈ {0, 1} ----- X i ∈ {1, . . . , M}

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

جملات به صورت دنباله‌ای از شاخص‌ها می‌باشند و $\mathcal{Y}$مجموعه ای از ورودی‌های احتمالی می‌باشند علاوه بر این تعریف عبارتx[I,j,k] را برای نشان دادن ترتیب زیر ازعناصر k،j،i تعریف می‌کنیم. در جهت خلاصه‌سازی ما یک جمله با هر طول متغیر خواهیم داشت که آن را با x نشان داده و جمله خروجی ما که نتیجه سیستم خلاصه‌سازی می‌باشد را y می‌نامیم که می‌توان اینگونه نتیجه گرفت که طول خروجی از طول ورودی کوچکتر خواهد بود و آن را به این شکل نشان می‌دهیم N < M و همچنین فرض می‌کنیم که کلمات در خلاصه نیز از همان واژگان ثابت V آمده است وخروجی را تشکیل می‌دهد و خروجی توالی از y1, . . . , yNخواهد بود. نکته بسیار مهم در این مسئله این است که فرض می‌کنیم طول خروجی و خلاصه جملات ما ثابت می‌باشد و هسته سیستم خلاصه‌ساز از آن آگاهی دارد و براساس ثابت طول جمله نسبت به عمل خلاصه‌سازی اقدام می نماید. فرض بعدی در خصوص مسئله خلاصه‌سازی به شکل ذیل می‌باشد که Y مشخص کننده‌ی مجموعه از تمام احتمالات مربوط به جملاتی با هر طول N می‌باشد که به صورت فرموله شده زیر آن را نمایش می‌دهیم :

$$\gamma \subset (\{0,1\}^{v}, \dots, \{0,1\}^{v})$$

که در آن $y \in Y$ i and y و در آن y_i یک شاخص را نشان می‌دهد. همچنین به این نکته اشاره کنیم که یک سیستم انتزاعی می‌باشد در صورتی که تلاش نماید توالی مطلوب را از این مجموعهYپیدا نماید، تحت یک تابع امتیاز دهی

$$s: x \times y \mapsto R$$
$$\arg\max_{y \in Y} s(x, y)$$

در ذیل خلاصه جملات کامل استخراج شده است که کلمات را از ورودی منتقل می نماید.

$$\arg\max_{m \in \{1,\dots,M\}^N} s(x, x_{[m1,\dots mn]})$$

$$\arg\max_{m \in \{1,\dots,M\}^N, mi-1<mi} s(x, x_{[m1,\dots mn]})$$

در حالی که ایجاد سیستم خلاصه‌سازی براساس مدل انتزاعی یک چالش می‌باشد در صورتی که بتوان چنین سیستمی را ایجاد نمود می‌توان با طیف وسیعی از داده های آموزشی کار کرد. در این رابطه ما بر عملکرد تابع امتیازدهی با نام S، تمرکز می‌کنیم که این مهم با در نظر گرفتن یک پنجره ثابت ازکلمات قبلی به دست می‌آید[50]:

$$S(x,y) \approx \sum_{k=1}^{N-1} g(y_i + 1, x, y_c)$$

در جایی که ما یک پنجره با اندازه c جهت بررسی کلمات قبلی ایجاد می‌نماییم $y_c \triangleq y[i - C + 1, \dots, i]$، به صورت خاص در نظر می‌گیریم لوگاریتم احتمالاتی را از خلاصه‌ای از داده های ورودی که به شکل زیر آن را نمایش می دهیم :

$$S(x,y) = \log p(y|x;\theta)$$

که می‌توان آن را به شکل ذیل نوشت:

$$\log p(y|x;\theta) \approx \sum_{i=0}^{N-1} \log p(y_{i+1} + x, y_c; \theta)$$

ما بر اساس اندازه پنجره c که نشان دهنده تعداد لغات قبل از لغت فعلی می باشد سعی در استخراج لغت بعدی با کمک گرفتن از زنجیره های مارکوف می باشیم ما با استفاده از تابع امتیاز دهی سعی در استخراج بهترین لغت بر اساس بالاترین احتمالات می باشیم با این تابع امتیازدهی که در ذهن ما ایجاد شده ما قصد داریم بر روی مدل و شرایط محلی متمرکز شویم:

$$P(y_{i+1} + x, y_c; \theta)$$

مدل:

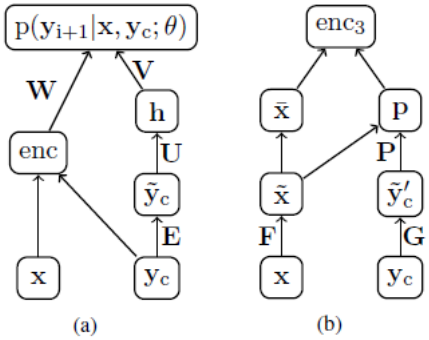
توزیع $P(y_{i+1} + x, y_c; \theta)$ یک مدل زبان شرطی براساس ورودی جمله X می‌باشد

$$arg_y \max \log p(y|x) = arg_y \max \log(p)\,|(x|y)$$

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

p(y) و p(x|y) به طور جداگانه تخمین زده می‌شوند. در اینجا ما مستقیما کار را بر روی شبکه‌های عصبی انجام می‌دهیم. شبکه شامل مدل زبان احتمالاتی و یک رمزنگار که به عنوان خلاصه شرطی عمل می‌کند، می‌باشد. هسته پارامتری، یک مدل زبانی برای تخمین احتمال کلمه بعدی بوده که با یک استاندارد شبکه‌های عصبی feed forward Language Model (NNLM) که مدل کامل آن به شکل ذیل می‌باشد سازگار است[50]:

$$P(y_{i+1}|y_c, x; \theta) \propto \exp(vh + Wenc(x, y_c),$$
$$y_c = [E_{yi-c+1}, \dots, E_{yi}],$$
$$h = \tanh(u^{y_c})$$



شکل (۸) : معماری انکودر-دکودر جهت خلاصه‌سازی- [50]

(a)یک نمودار شبکه برای رمزگذار NNLM با عنصر Encoder اضافی می‌باشد.

(b) یک نمودار شبکه بر پایه انکودر۳ می‌باشد.

پارامترها در زیر توضیح داده شده اند:

$$\theta = (E, U, V, W)$$

جایی $E \in R^{D \times V}$ که یک ماتریس تعبیه کلمه می‌باشد.

$U \in R^{(CD) \times H}, V \in R^{V \times H}, W \in R^{V \times H}$

سه پارامتر بالا وزن‌های ماتریس می‌باشند.

D اندازه کلمات تعبیه شده می‌باشد.

H یک لایه پنهان با انداره H می‌باشد.

تابع انکودر یک جعبه سیاه می‌باشد که یک بردار با اندازه H را برمی‌گرداند که نمایشی از محتویات و ورودی فعلی می‌باشد.

توجه داشته باشید که اصطلاح رمزگذار نشاندهنده‌ی یک مدل زبان استاندارد می‌باشد و با ترکیب انکودر و آموزش و یکپارچه کردن آنها می‌توانیم متن ورودی را با تولید شده ترکیب کنیم.

رمزگذاری کلمات:

مدل ما در ساده‌ترین شکل ممکن از کلمات از مخزنی از لغات ورودی استفاده به اندازه H استفاده نماید.

مدل را به شکل زیر نوشتیم:

$$enc_1(x, y_c) = p^T x,$$
$$P=[1/M,\dots 1/M]$$
$$X=[Fx1,\dots FxM]$$

جایی که ماتریس تعبیه ورودی $F \in R^{H \times V}$ تنها پارامتر جدید رمزگزار می‌باشد و $P \in [0,1]^M$ توزیع یکنواختی بر روی کلمات ورودی می‌باشد. برای خلاصه‌سازی این مدل می‌تواند ارتباطات مهم بین کلمات و تشخیص محتوا از قبیل نوع کلمات و غیره را ذخیره نماید. برای درک کردن روابط مختلف بین کلمات از انکودر عمیق جهت ورودی‌ها استفاده می‌نماییم. این معماری اجازه می‌دهد روابط بین کلمات و ارتباطات میان آنها به خوبی مشخص و درک شود. در بحث رمزگشایی الگوریتم رمزگشایی ویتربی را می‌توان جهت یافتن بهترین لغات

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

جهت ایجاد خلاصه نهایی در زمان $O(NV)^C$ استفاده نمود. در عمل هر مقدار که اندازه V (دامنه لغاتی که در خروجی سیستم می بایست از آن استفاده گردد) بزرگ باشد یافتن بهترین لغات با مشکل مواجه می- گردد. رویکرد حل مسئله را می توان در یک الگوریتم حریصانه مانند جستجوی پرتو جهت تابع arg max که نتیجه آن یک رمزگشای حریصانه می باشد یافت. در ذیل الگوریتم جستجوی پرتو که براساس مدل feed forward تغییر کرده است نشان داده شده:

Algorithm 1 Beam Search

Input: Parameters θ, beam size K, input $\mathbf{x}$
Output: Approx. K-best summaries

$\pi[0] \leftarrow \{\epsilon\}$
$\mathcal{S} = \mathcal{V}$ if abstractive else $\{\mathbf{x}_i \mid \forall i\}$
for $i = 0$ to $N-1$ **do**
  ▷ Generate Hypotheses
  $\mathcal{N} \leftarrow \{[\mathbf{y}, \mathbf{y}_{i+1}] \quad | \quad \mathbf{y} \in \pi[i], \mathbf{y}_{i+1} \in \mathcal{S}\}$
  ▷ Hypothesis Recombination
  $\mathcal{H} \leftarrow \left\{ \mathbf{y} \in \mathcal{N} \quad | \quad \begin{matrix} s(\mathbf{y},\mathbf{x}) > s(\mathbf{y}',\mathbf{x}) \\ \forall \mathbf{y}' \in \mathcal{N} \text{ s.t. } \mathbf{y}_c = \mathbf{y}'_c \end{matrix} \right\}$
  ▷ Filter K-Max
  $\pi[i+1] \leftarrow \underset{\mathbf{y} \in \mathcal{H}}{\text{K-arg max}}\, g(\mathbf{y}_{i+1}, \mathbf{y}_c, \mathbf{x}) + s(\mathbf{y}, \mathbf{x})$
end for
return $\pi[N]$

شکل (۹) : الگوریتم ویتربی–[50]

تنظیم پارامترهای اصلی و اجرا

سیستم خلاصه‌ساز از تعدادی فایل تشکیل گردیده که به زبان پایتون توسعه داده شده است که مرحله انکودر-دیکدر توسط فایل‌های ذیل صورت می‌گیرد.

جدول (۵) : ساختار کلی برنامه

ردیف	عنوان فایل
۱	seq2seq_attention.py
۲	seq2seq_attention_model.py
۳	seq2seq_attention_decode.py
۴	beam_search.py
۵	batch_reader.py

متغیرهای اصلی این تحقیق شامل موارد ذیل می‌باشد که تنظیماتی که جهت خلاصه‌سازی در زبان فارسی صورت گرفت به شرح ذیل می‌باشد:

جدول (۶) : تنظیمات پارامترها

عنوان پارامتر	مقدار
vocabulary size	200K
LSTM hidden units	256
word embedding size	128
Sampled softmax	4096
summary length	30
article length	first 2 sentences, total words within 120
bidirectional encoding layer	4
batch_size	64
Dataset	7k
min_input_len	6
learning rate	0.18

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

پس از طی مراحل ذکر شده در نهایت خروجی به صورت استخراج شده در فایل‌هایی با پسوند Dec قابل مشاهده می‌باشند

سرویس شهری حضور دست فروشان و توزیع مواد خوراکی آلوده و زیان آور از سوی آنان در مقابل مراکز تفریحی و فرهنگی که روزانه پذیرای صدها کودک و نوجوان هستند خطر بزرگی است که سلامتی و تندرستی جامعه به ویژه بچه‌ها را تهدید می‌کند.کارشناسان هشدار دادند این خوراکی ها به طریقه غیر بهداشتی و در محل کثیف و همچنین به د.....	۲ سطر ابتدایی هر متن که از متن اصلی جدا شده
هشدار کارشناسان بهداشتی درباره آلوده بودن موادغذایی دست فروشان	@highlight عنوان متن خلاصه
سرویس شهری هشدار کارشناسان بهداشتی برای توزیع مواد خوراکیآلوده در محل کثیف	خروجی سیستم

شکل (۱۰) : خروجی سیستم

به عنوان نمونه همانطور که در شکل شماره ۱۰ مشاهده می‌نمایید سیستم با جدا کردن ۲ سطر ابتدایی و استخراج ویژگی با کمک بخش عنوان متن خلاصه را تولید نموده است.

۱. ارزیابی

۲. در این تحقیق به کمک برنامه ای که به زبان پایتون توسعه داده شده است و به کمک معیار ارزیابی ROUGE نسبت به بررسی نتایج حاصل اقدام می‌نماییم . بسته‌ی ISI ROUGE که بعدها با نام ROUGE معروف شد، تلاشی برای ارزیابی خودکار خلاصه‌ها می‌باشد. این ابزار مبتنی بر فراخوانی n تایی‌های مشترک بین خلاصه‌های ماشینی و خلاصه‌های مرجع می‌باشد. نتایج حاصل از عمل خلاصه‌سازی در فایل‌هایی با پیشوند Dec و Ref ذخیره می‌گردند که فایلهای Ref شامل متن اصلی و فایل Dec حاوی متن خلاصه ، تولیدی توسط سیستم می‌باشد.همانطور که پیش تر ذکر گردید جهت آموزش شبکه از ۲ سطر ابتدایی هر متن ، همچنین چکیده آن که همان عنوان خبر می باشد استفاده گردید و انتخاب ۲ سطر به این جهت می باشد که موضوع اصلی و داستان هر متن در ۲ سطر ابتدایی یک متن گنجانده شده است همچنین این کار به علت کاهش سربار محاسباتی و بهبود مدل صورت می گیرد ما برای بررسی این موضوع و در آزمایش دیگری به جای استفاده از ۲ سطر از ۴ سطر جهت آموزش استفاده نمودیم که نتایج حاصل مشخص نموده بهترین پارامتر انتخاب ۲ سطر ابتدایی می باشد:

جدول (۷) : نتایج ارزیابی با ۲ سطر ورودی جهت آموزش شبکه

نتایج ارزیابی بر روی دیتاست اختصاصی	
Rouge_l_p	0.0971
Rouge_l_r	0.1412
Rouge_l_f	0.1189
Rouge_1_p	0.1114
Rouge_1_r	0.1932
Rouge_1_f	0.1083
Rouge_2_p	0.225
Rouge_2_r	0.0467
Rouge_2_f	0.0337

جدول (۸) : نتایج ارزیابی با ۴ سطر ورودی جهت آموزش شبکه

نتایج ارزیابی بر روی دیتاست اختصاصی	
Rouge_l_p	0.0670
Rouge_l_r	0.1167
Rouge_l_f	0.0669
Rouge_1_p	0.0719
Rouge_1_r	0.1460
Rouge_1_f	0.0897
Rouge_2_p	0.0125
Rouge_2_r	0.0167
Rouge_2_f	0.0137

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

همچنین نتایج مشابه در تیم تحقیقات گوگل که در زبان انگلیسی با مکانیزم مشابه بر روی دیتاست گیگاورد صورت پذیرفته به شرح ذیل می‌باشد:

جدول (۹) : نتایج ارزیابی در دیتاست گیگاورد

نتایج ارزیابی بر روی دیتاست Gigaword	
Rouge_1_p	0.5015
Rouge_1_r	0.3827
Rouge_1_f	0.4256
Rouge_2_p	0.2756
Rouge_2_r	0.2057
Rouge_2_f	0.2312

همانطور که مشاهده می‌نمایید نتایج ارزیابی بر روی دیتاست گیگاورد از امتیازات بیشتری برخوردار است و علت آن این است که همانطور که قبلا رکن اصلی و خوراک اصلی یادگیری عمیق حجم دیتاست آموزشی بوده به طوری که هرمقدار حجم که دیتاستی که در آموزش شبکه استفاده می‌گردید بیشتر می‌بود و از استاندارد های بالاتری برخوردار بود شبکه در استخراج ویژگی‌ها بهتر عمل می‌نمودیکی از مشکلات این تحقیق عدم وجود دیتاست استاندارد و با حجم بالا بود که ما مجبور به توسعه دیتاست مناسب‌تر بر پایه دادگان همشهری شدیم و در مدل اولیه این تحقیق آموزش شبکه بر پایه دیتاست گیگاورد صورت گرفت که شامل ۱۰ میلیون رکورد داده می‌شد در حالی که دیتاست همشهری شامل ۳۰۰ هزار رکورد بوده که بعد از تبدیلات لازم با توسعه برنامه اختصاصی ما به عدد نزدیک به ۱۰۰ هزار رکورد رسید که این مسئله بر نتایج خروجی اثر گذار بود.

کارهای آینده

یکی از مشکلات در این تحقیق عدم وجود دیتاست مناسب با فرمت لازم جهت آموزش شبکه بوده است. یکی از مفاهیم اصلی که به دفعات متعدد در مبحث یادگیری ماشینی مورد استفاده قرار می‌گیرد اصطلاح دیتاست یا مجموعه داده‌ها می‌باشد. دیتاست‌ها در اصل ورودی اصلی و الگوهای اصلی در سیستم‌های هوش مصنوعی می‌باشند و به طور ساده می‌توان نوع کاربری آنها را اینگونه تعریف نمود که به مدل اصلی سیستم تزریق می‌شوند و شبکه عصبی از روی مجموعه داده‌ها، الگوها را استخراج و آموزش می‌بینند. در مبحث شبکه‌های عصبی عمیق به هر مقدار که حجم دیتاست بالاتر باشد سیستم بهتر آموزش دیده و خروجی‌های مناسب‌تری حاصل می‌گردد. به عنوان نمونه در زبان انگلیسی دیتاست‌های مناسبی جهت انجام تحقیقات مرتبط در حوزه خلاصه‌سازی متون وجود دارد که از آن جمله به دیتاست گیگاورد می‌توان اشاره نمود در مقابل آن در زبان فارسی دیتاست استانداردی با حجم استاندارد وجود ندارد و این مهم مشکلات جدی را در راه تحقیقات مرتبط در مبحث خلاصه‌سازی متون فارسی را ایجاد نموده است. در آینده سعی خواهد شد با توسعه دیتاست استاندارد و حجیم و تغییرات بنیادی‌تر در پارامترها و طراحی بهتر شبکه نتایج این تحقیقات را هرچه بیشتر بهبود بخشید.

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

نتیجه گیری

در دهه حاضر نرخِ رشدِ اطلاعات بسیار فزاینده بوده و نیاز به سیستم‌های خلاصه‌ساز متون به شدت احساس می‌شود. اکثر کارهایی که در خصوص خلاصه‌سازی متون فارسی صورت گرفته از روش استخراجی استفاده نموده‌اند و تعداد کمی نیز از یادگیری ماشینی وظیفه خلاصه‌سازی را انجام داده که از نحوه پیاده‌سازی آنها اطلاعات دقیقی در دسترس نیست. در سال‌های اخیر با ایجاد زیر ساخت‌های مناسب در حوزه‌های نرم‌افزار و سخت‌افزار امکان ایجاد شبکه‌های عمیق عصبی فراهم شده است ، این نوع شبکه‌ها به قدرت پردازشی بالایی نیاز داشته و امکان استخراج ویژگی‌های بهتری را از داده‌ها فراهم می‌نمایند. در این تحقیق با استفاده از مدل ارائه شده توسط تیم مغز گوگل و با کمک گرفتن از ابزار تنسورفلو همچنین ایجاد تغییراتی در پارامترهای کدهای توسعه داده شده نسبت به انجام خلاصه سازی متون فارسی اقدام نمودیم همچنین یکی از موانع موجود در راه رسیدن به هدف مذکور عدم وجود دیتاست مناسب و منطبق بر مدل تحقیقی ما بود ، بدین منظور نرم افزاری جهت انجام تحلیل متنی و تولید دیتاست مورد نظر به زبان سی شارپ توسعه داده شد که سیستم تولیدی از دادگان همشهری جهت ایجاد دیتاست مناسب استفاده نمود. دادگان همشهری پیکره‌ای حاوی ۳۱۸ هزار سند مربوط به اخبار سال‌های ۱۳۷۵ تا ۱۳۸۶می باشد مدل ما با کمک گرفتن از انکودر-دکودر عمیق بازگشتی به نتایج قابل توجهی در خلاصه‌سازی متون فارسی دست یافت گرچه نتایج به دست آمده با تحقیقاتی که در زبان انگلیسی و به کمک گرفتن از دیتاست گیگاورد که حاوی ۱۰ میلیون رکورد استاندارد می‌باشد فاصله زیادی دارد و مشخص گردید یکی از وظایف مهمی که در تکمیل این مسیر و در جهت تحقیقات آینده می بایست صورت پذیرد ایجاد مجموعه آموزشی استاندارد با حجم بالا می باشد. ضمنا پس از ارزیابی های صورت گرفته به کمک معیار Rouge مشخص گردید استفاده از ۲ جمله ابتدایی هر متن به همراه چکیده آن جهت ورودی شبکه نتایج مناسبتری را به همراه خواهد داشت. با توجه به این که تاکنون در زمینه‌ی خلاصه‌سازی به کمک یادگیری عمیق ماشینی کمتر کاری انجام شده و اطلاعات بسیار کمی جهت بهبود معماری شبکه عمیق منطبق با زبان فارسی در دسترس می باشد همچنین داده های آموزشی مناسبی بدین منظور در دسترس نمی باشد نتایج دلگرم کننده ای حاصل گردید.

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

منابع و مراجع

[۱] حسینی خواه،طیبه؛احمدی،عباس؛محبی،آزاده(۱۳۹۴) .بهبود خلاصه سازی خودکار متون فارسی با استفاده از روش های پردازش زبان طبیعی و گراف شباهت.

[۲] شمس فرد، مهرنوش؛ شفیعی، فاطمه (۱۳۹۳) . سیستم خودکار خلاصه ساز متون فارسی

[۳] برشبان،یاسمن؛ یوسفی نسب،حامد؛میر روشندل، ابوالقاسم،(۱۳۹۳) توسعه یک پیکره متنی فارسی پرسش و پاسخ

[۴] اخوان،تارا؛شمس فرد،مهرنوش؛ عرفانی جورابچی ،مونا (۱۳۸۷) : پارس سامیت خلاصه ساز تک سندی و چند سندی متون فارسی

[۵] حسین خانی،جواد؛ یزدانی،امید(۱۳۹۵) سیستم خلاصه ساز چند سندی فارسی ، جهت بهبود ، خوانایی و پیوستگی از طریق ایجاد گراف متقاطع اسناد

[۶] پورمعصومی، آصف؛کاهانی محسن؛طوسی،سید احمد؛استیری،احمد؛قائمی هادی(۱۳۹۳) ایجاز:یک سامانه عملیاتی برای خلاصه سازی تک سندی متون خبری فارسی

[۷] دانشگاه علم و صنعت ایران(۱۳۸۸) پروژه طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی

[۸] امیر شهاب شهابی، "چکیده‌سازی متون زبان فارسی"، دومین کنفرانس علوم شناختی، صفحه ۵۶ ، تهران ۱۳۸۱.

[۹] و. ورفریاری، "بررسی جامع سیستم‌های خلاصه‌ساز فارسی و تولید یک نمونه عملی"، پایان‌نامه کارشناسی ارشد، دانشکده‌ی مهندسی کامپیوتر، دانشگاه علم وصنعت ایران ۱۳۷۶.

[۱۰]محسن مشکی، "بررسی روش‌های خلاصه‌سازی متون غیرساخت یافته‌ی فارسی"، سمینار کارشناسی ارشد، دانشکده‌ی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران ۱۳۸۶.

[11] M. Honarpisheh, Gh. Ghassem-Sani, Gh. Mirroshandel, “A Multi-Document Multi-Lingual Automatic Summarization System”, Proceedings of the 3rd International Joint Conference on natural language processing (IJCNLP), pp. 733-738, 2007.

[12] A. Poormasoomi, M. Kahani, S. Varasteh and H. Kamyar, “Context-based Persian multi-document summarization (global view)”, IALP, Malaysia, 2011.

[13] - H.Saggion, and G.Lapalme.”Generating Informative and Indicative Summaries with SumUM”, Computational Linguistics, Special Issue on Automatic Summarization. 2002

[14] M.White, T.Korelsky, C.Cardie, V.Ng , D.Pierce, K.Wagstaff. Multidocument Summarization via Information Extraction.2001.

[15] T.Hirao,K.Takeuchi, H.Isozaki, Y.Sasaki, E.Maeda. NTT/NAIST’s Text Summarization Systems.2003

[16] T.A.S.Pardo, L.H.M.Rino, M.d.G.V.Nunes.GistSumm: A Summarization Tool Based on a New Extractive Method.2001

[17] F.Lacatusu, A.Hickl, K.Roberts, Y.Shi, J.Bensley,B.Rink, P.Wang, L.Taylor. LCC’s GISTexter at DUC 2006: Multi-Strategy Multi-Document Summarization.2006

[18] K.Zechner, A.Waible., “DiaSumm: Flexible Summarization of Spontaneous Dialogues in UnrestrictedDomains.”, 2000

[19] M.BAZRFKAN,M. RADMANESH. USING MACHINE LEARNING METHODS TO SUMMARIZE PERSIAN TEXTS,2014

[20] R. Nallapati,B. Zhou,c. dos Santos, Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond,2016

[21] P.Priya, G. and K. Duraiswamy, AN APPROACH FOR TEXT SUMMARIZATION USING DEEP LEARNING ALGORITHM,2014

[22] I,Sutskever,O.Vinyals,Q. V. Le, Sequence to Sequence Learning with Neural Networks,2014

[23] R. Paulus,C. Xiong,R. Socher, Your tl;dr by an ai: a deep reinforced model for abstractive summarization,2017

[24] M. Yousefi-Azar,L. Hamey, Text summarization using unsupervised deep learning,2017