علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۳۹- ۴۶
مقاله عادی

یادگیری عمیق در خلاصه‌سازی چندسندی متون فارسی

شیما محرابی حمیدرضا احمدی‌فر سید ابوالقاسم میرروشندل

دانشکده فنی، دانشگاه گیلان، رشت، ایران

چکیده

با پیشرفت علوم و تکنولوژی و در نتیجه افزایش حجم اطلاعات متنی قابل دسترس از طریق اینترنت، وجود سامانه‌های خلاصه‌ساز که چکیده‌ای از اطلاعات موردنظر را در کوتاه‌ترین زمان ممکن در دسترس کاربر قرار دهند، ضروری به‌نظر می‌رسد. خلاصه‌سازی خودکار متون از دیرباز مورد توجه پژوهشگران حوزه‌ی پردازش زبان‌های طبیعی قرار گرفته است. امروزه با بهبود توان پردازشی سیستم‌های موجود و ظهور ابزارهای محاسباتی نوین، تلاش برای افزایش کارایی سیستم‌های خلاصه‌ساز ادامه دارد. در این مقاله به معرفی یک سامانه‌ی خلاصه‌ساز استخراجی چندسندی متون فارسی می‌پردازیم. این سامانه برای امتیازدهی به جملات از نظر میزان اهمیت آنها در سند، از روشی تحت عنوان یادگیری عمیق بهره می‌برد. یادگیری عمیق، روشی برای آموزش ماشین برپایه‌ی شبکه‌های عصبی مصنوعی است. پیش از این، یادگیری عمیق در زمینه‌های پردازش صوت و تصویر و همچنین پردازش زبان‌های طبیعی مورد استفاده قرار گرفته است. نتایج خوب بدست آمده از این روش در مقایسه با دیگر روش‌های مرسوم، انگیزه‌ای در بکارگیری این روش در خلاصه‌سازی خودکار چندسندی متون فارسی گشت. در خلاصه‌ساز پیشنهادی با استفاده از یک شبکه‌ی Autoencoder عمیق، عمل امتیازدهی به جملات انجام می‌گیرد و میزان دقت خلاصه‌ساز در ارزیابی جمله‌ای قابل‌قبول به‌نظر می‌رسد.

کلمات کلیدی: پردازش زبان فارسی، خلاصه‌سازی خودکار چندسندی، یادگیری عمیق، شبکه‌های عصبی مصنوعی.

۱- مقدمه

در دنیای امروزی، با پیشرفت علوم و تکنولوژی، شاهد افزایش روزافزون حجم اطلاعات متنی قابل دسترس هستیم. بنابراین دستیابی به اطلاعات مورد نظر در حجمی کمتر اما کاربردی و دارای جامعیت قابل‌قبول، بسیار مطلوب به‌نظر می‌رسد. خلاصه‌سازی متون[1] توسط انسان باوجود داشتن مزایایی از قبیل صحت و جامعیت مستلزم صرف وقت و هزینه‌ی بالایی است. همچنین خلاصه‌سازی اسناد بزرگ به‌صورت دستی برای انسان کاری دشوار است. این در حالی است که متون موجود برروی اینترنت فراوان است و اینترنت بیشترین اطلاعات مورد نیاز افراد را فراهم می‌کند. لذا وجود یک سیستم خلاصه‌ساز کامپیوتری باعث صرفه‌جویی در زمان و هزینه‌ی مصرفی در تولید متن خلاصه خواهد شد، هرچند که ممکن است از لحاظ صحت و جامعیت با خلاصه‌ی تولید شده توسط انسان برابری نکند.

بطور کلی هدف از خلاصه‌سازی خودکار متن، فشرده‌سازی و کوتاه نمودن متن اصلی با حفظ محتوا و معنای کلی آن است. خلاصه‌سازی متون منجر به استفاده از منابع بیشتر با سرعت بالاتر و در نتیجه دستیابی به اطلاعاتی غنی‌تر می‌شود. یک خلاصه خوب باید موضوعات گوناگون یک یا چند سند را بدون افزونگی دربرداشته باشد.

به شکل‌های مختلفی می‌توان خلاصه‌سازی را طبقه‌بندی کرد. یک روش طبقه‌بندی از نظر شکل و فرم سازماندهی متن خلاصه است. در این حالت متن خلاصه شده به یکی از روش‌های استخراجی[2] و یا چکیده ای[3] حاصل می‌گردد. در روش استخراجی، جملات مهم متن انتخاب شده و به همان شکل اصلی خود در متن خلاصه ظاهر می‌شوند. در خلاصه‌ی چکیده‌ای، خلاصه‌ی مفهومی از متن در خروجی خلاصه‌ساز تولید می‌شود. در واقع ممکن است، فرم جملات تغییر یابد. این روش مشابه روشی است که انسان برای خلاصه‌سازی متون بکار می‌گیرد [۱]. همچنین گونه‌ی دیگری از دسته‌بندی خلاصه، براساس تعداد اسناد شرکت داده شده در تولید خلاصه است. بر این اساس خلاصه‌سازی به دو دسته‌ی تک‌سندی و چندسندی تقسیم‌بندی می‌شود. در خلاصه‌سازی تک‌سندی، از یک سند برای ایجاد متن خلاصه استفاده می‌شود اما در خلاصه‌سازی چندسندی، ورودی سیستم چندین سند با موضوع کلی مشترک است که جنبه‌های مختلفی از آن موضوع را

تحت پوشش قرار می‌دهد

یکی از مهمترین چالش‌ها در امر خلاصه‌سازی متون، انتخاب بهترین جملات متن اصلی است به‌طوری که جنبه‌های مهم و کلیدی متن را شامل شده و در عین حال فاقد افزونگی باشد. در نتیجه لازم است که متن اصلی پیش‌پردازش شود و ویژگی‌هایی که در انتخاب بهترین جملات تاثیرگذارند، استخراج شوند. مرحله‌ی پیش‌پردازش و استخراج ویژگی‌ها نقش بسزایی در حصول نتیجه‌ی مطلوب، ایفا می‌کند. در خلاصه‌سازی استخراجی جملات در قالب بردارهایی قرار می‌گیرند که هر بردار شامل ویژگی‌هایی است که اهمیت یک جمله را از جهات مختلف بیان می‌دارد. یک بردار ویژگی برداری n عضوی است که هر عضو آن یک مقدار عددی دارد. تصمیم‌گیری در مورد میزان اهمیت یک جمله براساس این مقادیر صورت می‌گیرد.

یکی از معروف‌ترین سیستم‌های خلاصه‌سازی چندسندی، تحت عنوان MEAD شناخته می‌شود [۲]. Gistsumm نیز یک خلاصه‌ساز استخراجی است که از ۳ بخش تکه‌سازی متن، امتیازدهی به جملات و ایجاد خلاصه تشکیل شده است [۳]. قدیمی‌ترین کاری که در زبان فارسی برای یک سیستم خلاصه‌ساز صورت گرفته، یک خلاصه‌ساز استخراجی تک‌سندی به‌نام FarsiSum است [۴]. این خلاصه‌ساز یک برنامه‌ی کاربردی تحت وب است و شکل‌گیری آن بر پایه‌ی یک پروژه‌ی خلاصه‌سازی در زبان سوئدی است. در [۵] یک استخراج کننده متن فارسی تک‌سندی ارائه شده است که براساس زنجیره‌ی زبانی و روش‌های مبتنی بر گراف عمل می‌کند. در [۶] یک سیستم خلاصه‌ساز چندسندی- چندزبانی ایجاد و معرفی شده است که براساس SVR[۴] و دسته‌بندی چندسطحی عمل می‌کند. در [۷] یک خلاصه‌ساز براساس منطق فازی معرفی شده است. در این سیستم، ویژگی‌های یک متن نظیر طول جمله، میزان تشابه با عنوان و تشابه با کلمات کلیدی به‌عنوان ورودی سیستم فازی درنظر گرفته شده و عمل خلاصه‌سازی انجام می‌شود.

از سال ۲۰۰۶ حوزه‌ی جدیدی از تحقیقات یادگیری ماشین با نام یادگیری ساختاریافته عمیق و یا بطور مصطلح‌تر یادگیری عمیق[۵] پا به عرصه‌ی وجود نهاد [۸]. در طی چند سال اخیر تکنیک‌های حاصله از تحقیقات یادگیری عمیق، حجم وسیعی از پژوهش‌های مربوط به پردازش سیگنال و اطلاعات را تحت تأثیر خود قرار داده است. یادگیری عمیق عموماً از شبکه‌های عصبی مصنوعی بهره می‌برد. سطوح بالاتر در شبکه به واسطه‌ی اطلاعات حاصله از سطوح پایین‌تر تعریف می‌شوند. یادگیری عمیق را می‌توان تقاطعی میان حوزه‌های تحقیقاتی شامل هوش مصنوعی، شبکه‌های عصبی، تشخیص الگو و پردازش سیگنال دانست. ۳ عامل مهم در محبوبیت امروزی یادگیری عمیق، افزایش روزافزون توان پردازشی تراشه‌های محاسباتی، افزایش قابل‌توجه حجم اطلاعات مورد استفاده برای یادگیری و پیشرفت‌های اخیر در پژوهش‌های مرتبط با پردازش اطلاعات و سیگنال است [۹].

یکی از مقالات مهم در زمینه‌ی یادگیری عمیق در سال ۲۰۰۹ منتشر شد و در آن دلایل اهمیت این روش و مزایای آن برشمرده شد. امکان یادگیری چندسطحی و توانایی استخراج ویژگی‌ها بطور مستقیم از داده‌های ورودی از نقاط قوت این روش محسوب می‌شود [۱۰]. در [۱۱] شبکه‌های عصبی عمیق را در زمینه‌ی پردازش صوت و گفتار مورد آزمایش قرار داده‌اند. ارزیابی‌های صورت گرفته در مقاله‌ی مذکور، نشان داده است که استفاده از شبکه‌های عصبی عمیق در پردازش گفتار نسبت به روش‌های مرسوم نظیر مدل مارکوف پنهان، با افزایش کارایی همراه است.

نتایج بدست آمده توسط یادگیری عمیق در حوزه‌های پردازش صوت و تصویر، پژوهشگران پردازش زبان‌های طبیعی را برآن داشت که از این روش یادگیری در مسائل مربوط به پردازش زبان نیز بهره ببرند. در سال ۲۰۱۱ یک معماری واحد و یکپارچه معرفی شد که قابل‌اعمال بر روی مسائل متنوعی در پردازش زبان است. مسائل قابل حل توسط این معماری شامل تشخیص موجودیت‌های نامدار[۶]، برچسب‌گذاری اجزای کلام[۷]، برچسب‌گذاری نقش معنایی کلمات[۸] و تکه‌بندی گرامری جملات[۹] است [۱۲، ۱۳]. در [۱۴] از یادگیری عمیق در مدل‌سازی زبانی[۱۰] استفاده شده است. در این مقاله مشاهده شده است که میزان نرخ خطای کلمات و همچنین میزان حیرانی در انتخاب کلمه به نسبت روش‌های دیگر بهبود یافته است. در [۱۵] یک چهارچوب خلاصه‌سازی چندسندی با استفاده از یادگیری عمیق ارائه شده است که بردار ویژگی‌ها در آن شامل تعداد تکرار کلمات یک مجموعه لغات از پیش تعیین شده، در اسناد موردنظر برای خلاصه‌سازی است. بنابراین ورودی شبکه، برداری از ویژگی‌ها و خروجی آن مجموعه‌ای از جملات خواهد بود. برای ایجاد خلاصه ساز از یک RBM[۱۱] استفاده می‌شود. ابتدا شبکه سعی در حذف کلمات غیرضروری دارد، سپس کلمات مهم از بین کلمات باقی‌مانده تعیین می‌شوند. جملاتی که حاوی کلمات مهم هستند، استخراج شده و در نهایت با استفاده از برنامه‌نویسی پویا و با در نظر گرفتن محدودیت حجم از پیش تعریف شده متن خلاصه، از بین جملات مهم، جملات مناسب برای شکل‌گیری خلاصه انتخاب می‌گردد.

با توجه به دستاوردهای بدست آمده از یادگیری عمیق در پردازش زبان‌های طبیعی برآن شدیم که از توانایی‌های این روش یادگیری در خلاصه‌سازی خودکار متون فارسی بهره ببریم. در این پژوهش به معرفی یک سیستم خلاصه‌سازی استخراجی چندسندی برای متون فارسی می‌پردازیم که در آن عمل انتخاب جملات مهم براساس امتیازاتی است که به هر جمله توسط شبکه عمیق اختصاص داده شده است. شبکه‌ی عمیق پیشنهادی، یک شبکه‌ی ۹ لایه است که در لایه‌ی ورودی ویژگی‌های هر جمله شامل TF/IDF، میزان تشابه با عنوان سند، مکان رخداد جمله و امتیاز مربوط به برچسب اجزای کلام جمله به شبکه داده می‌شود. در نهایت طی آموزش لایه‌ای شبکه به هر جمله یک امتیاز اختصاص داده می‌شود، سپس جملات براساس امتیازشان مرتب شده و جملات با بالاترین امتیاز برای تشکیل متن خلاصه انتخاب می‌شوند. ساختار این مقاله بدین شرح است: در بخش دوم به معرفی یادگیری عمیق پرداخته می‌شود و در بخش سوم روش پیشنهادی برای خلاصه‌سازی چندسندی متون فارسی ارائه خواهد شد. بخش چهارم ارزیابی روش پیشنهادی را شامل می‌شود و در پایان نیز نتیجه‌گیری ذکر شده است.

۲- معرفی یادگیری عمیق

مکانیزم پردازش اطلاعات توسط انسان نظیر بینایی و شنوایی، به نوعی بازگوکننده‌ی نیاز به معماری عمیق برای استخراج ساختارهای پیچیده‌ی داده‌های ورودی است. به‌عنوان مثال سیستم بینایی انسان ذاتاً از یک ساختار سلسله مراتبی برای درک تصاویر بهره می‌برد و به‌عنوان ورودی رنگ، اندازه، جهت و عواملی از این دست را مدنظر قرار می‌دهد. مفهوم یادگیری عمیق نشأت گرفته از پژوهش‌های حوزه‌ی شبکه‌های عصبی مصنوعی است. شبکه‌های پیش‌خور[۱۲] و یا پرسپترون چند لایه با تعداد لایه‌های مخفی زیاد، نمونه‌های خوبی از مدل‌هایی با معماری عمیق هستند. الگوریتم Back-Propagation که در دهه‌ی ۱۹۸۰ به محبوبیت رسید، یک الگوریتم معروف در یادگیری پارامترهای این شبکه محسوب می‌شود.

آموزش شبکه‌های عمیق دشوار است، روش‌هایی که بطور مؤثر و بهینه بر روی شبکه‌های کم‌عمق اعمال می‌شود، در شبکه‌های عمیق چندان کارآمد نیستند. این مشکل با معرفی روشی تحت عنوان پیش‌آموزش لایه‌ای بدون‌ناظر[۱۳]، در شبکه‌های عمیق مرتفع گشت. بطور دقیق‌تر، در یک ساختار یادگیری عمیق، هر لایه بطور مجزا در نظر گرفته می‌شود. در این روش به محض اینکه لایه‌های قبلی آموزش داده شدند، لایه‌ی بعدی با استفاده از داده‌های حاصل از لایه‌ی قبلی آموزش می‌بینند. سپس برروی کل شبکه یک مرحله تنظیم کلی یا fine-tuning انجام

می‌گیرد [۹].

از مدل‌های مرسوم در یادگیری عمیق می‌توان به ماشین بولتزمن محدود شده و Autoencoderها اشاره کرد. ماشین بولتزمن محدود شده و یا به اختصار RBM مدلی برای نمایش یک توزیع احتمال است. به‌منظور آموزش RBM و با فراهم آوردن مجموعه داده‌های آموزشی، این مدل سعی در تنظیم پارامترهای خود دارد، به قسمی که توزیع احتمال ارائه شده توسط RBM به بهترین شکل ممکن در برگیرنده‌ی داده‌های آموزشی باشد. با معرفی یک معماری چند لایه به‌نام شبکه‌های باور عمیق[۱۴] که متشکل از چندین RBM متصل به هم هستند، RBMها بیش از پیش مورد توجه قرار گرفتند. ایده‌ی این معماری استخراج ویژگی‌های مرتبط با نورون‌های ورودی توسط نورون‌های مخفی است. سپس این ویژگی‌ها به‌عنوان ورودی RBM بعدی عمل می‌نمایند. بدین شکل و با پشته‌سازیRBMها، شبکه قادر به یادگیری ویژگی‌های جدید از ویژگی‌های حاصله‌ی قبلی خواهد بود [۱۶].

نوع خاصی از شبکه‌های عصبی عمیق تحت عنوان Autoencoderها شناخته می‌شود که در آن‌ها بردار خروجی مشابه بردار ورودی است. این نوع شبکه اغلب برای یادگیری ویژگی‌ها با رمزگذاری مؤثر داده‌های ورودی بکار گرفته می‌شود. Autoencoder یک روش استخراج ویژگی به شکل غیرخطی و بدون استفاده از داده‌های برچسب‌دار است. یک Autoencoder دارای یک لایه‌ی ورودی است که نمایش‌دهنده‌ی داده‌های ورودی شبکه است (به‌عنوان مثال، پیکسل‌های یک تصویر). همچنین این مدل شامل یک یا چند لایه‌ی مخفی است که نشان‌دهنده‌ی ویژگی‌های تغییریافته هستند و دارای یک لایه خروجی مطابق لایه‌ی ورودی خود است. تعداد نورون‌های مخفی می‌توانند بیشتر و یا کمتر از تعداد نورون‌های ورودی باشد. یک Autoencoder اغلب توسط یکی از اشکال الگوریتم Back-Propagation آموزش می‌بینند و عموماً روش Stochastic gradient descent را مورد استفاده قرار می‌دهد [۱۷].

۳- روش پیشنهادی خلاصه‌سازی

یکی از مراحل پایه‌ای در امر خلاصه‌سازی متون، مرحله‌ی پیش‌پردازش متن ورودی است. اولین گام پیش‌پردازش متن، نرمال‌سازی آن است. نرمال‌سازی به عمل یکدست‌سازی واحدهای متنی به‌طوری که قابل پردازش توسط ماشین باشند، اطلاق می‌گردد. گاهی حروف به‌کار رفته در دو کلمه‌ی یکسان، با یکدیگر متفاوتند این در حالی است که از دید انسان آن دو کلمه یکسان محسوب می‌شوند. این مسأله باعث عدم‌شناسایی یکسان بودن دو کلمه توسط ماشین می‌شود در نتیجه لازم است متن نرمال‌سازی شود تا یک شکل واحد برای کلمات یکسان بدست آید. به‌عنوان مثال «ها» را می‌توان به ۳ شکل، چسبان، جدا بافاصله و جدا با نیم‌فاصله بکار برد. در نتیجه ۳ شکل متفاوت از یک کلمه‌ی جمع بسته شده با «ها» تولید می‌شود. به‌منظور نرمال‌سازی متن می‌بایست یکی از این حالات را به‌عنوان شکل قابل‌قبول درنظر گرفت و تمامی اشکال دیگر آن کلمه به شکل قابل‌قبول تعیین شده، تبدیل شوند.

در گام بعدی می‌بایست متن را به واحدهایی براساس جملات و کلمات تقسیم‌بندی کنیم. در واقع باید مرز جملات و کلمات شناسایی شود. علائمی نظیر نقطه (اگر محصور به عدد نباشد) و"\n" را می‌توان به‌عنوان پایان یک جمله قلمداد کرد. همچنین علائمی نظیر خط‌فاصله، فضای خالی و ویرگول را می‌توان به‌عنوان مرز کلمات در نظر گرفت. در مرحله‌ی بعد، کلمات ریشه‌یابی شده و کلمات ایست[۱۵] حذف می‌گردند. کلمات ایست، کلماتی پرتکرار و بی‌اهمیت از لحاظ بارمعنایی هستند که عدم حذف آن‌ها بواسطه‌ی تعداد تکرار بالایی که دارند، ممکن است سیستم را در شناسایی کلمات پراهمیت متن، دچار اشتباه کند. کلماتی نظیر «است»، «برای»، «شده» در مجموعه‌ی کلمات ایست قرار می‌گیرند.

۳-۱- تولید بردار ویژگی‌ها

در این پژوهش برای آموزش خلاصه‌ساز، ۴ ویژگی معرفی می‌شود. این ۴ ویژگی تشکیل یک بردار ویژگی را می‌دهند. هر جمله از متن دارای یک بردار ویژگی است. ویژگی‌ها شامل میزان تکرار کلمات، میزان تشابه جملات با عنوان سند، موقعیت قرارگیری جمله در سند و امتیاز مربوط به برچسب‌گذاری اجزای کلام جمله است.

۳-۱-۱- ویژگی میزان تکرار کلمات

برای اندازه‌گیری میزان تکرار کلمات از فراوانی وزنی TF/IDF بهره برده شده است. در این شیوه به کلمات یک وزن براساس فراوانی آنها در سند داده می‌شود. در واقع این سیستم وزن‌دهی نشان می‌دهد که به چه میزان یک کلمه در یک سند مهم است. میزان تکرار کلمه در یک سند با TF(f,d) نشان داده می‌شود و وزن نهایی با استفاده IDF بدست می‌آید. IDF به معنای عکس فراوانی سند است که نشان‌دهنده‌ی میزان فراوانی کلمه موردنظر در اسناد دیگر است، در واقع به این سوال پاسخ می‌دهد که آیا کلمه موردنظر، در تمامی اسناد متداول است یا خیر. عکس فراوانی سند با لگاریتم‌گیری از نتیجه‌ی تقسیم تعداد کل اسناد بر تعداد اسنادی که کلمه‌ی موردنظر در آن‌ها ظاهر شده، بدست می‌آید. در نهایت با ضرب TF در IDF میزان وزن اختصاص داده شده به کلمات براساس تکرار آنها بدست می‌آید. در رابطه‌ی (۱) روش محاسبه‌ی IDF و در رابطه‌ی (۲) نیز روش محاسبه‌ی معیار TF/IDF نشان داده شده است.

$$IDF(t, D) = \log(\frac{D}{d \in D : t \in d}) \quad (۱)$$

در رابطه‌ی (۱)، D به کلیه‌ی اسناد موجود و t به کلمه موردنظر اشاره دارد. $d \in D : t \in d$ تعداد اسنادی که کلمه t در آن موجود است را نشان می‌دهد.

$$TF/IDF\ (t.d.D): TF(t, d) \times IDF(t, D) \quad (۲)$$

در رابطه‌ی (۲)، منظور از TF(t,d)، تعداد دفعاتی است که کلمه‌ی t در سند d تکرار شده است. در روش پیشنهادی، میانگین مقادیر معیار TF/IDF کلمات یک جمله، بیانگر مقدار ویژگی میزان تکرار کلمات آن جمله خواهد بود، در واقع طول جملات بر حسب تعداد کلمات آن در محاسبه مقدار ویژگی تکرار کلمات یک جمله در نظر گرفته می‌شود. روش محاسبه ویژگی میزان تکرار کلمات یک جمله در رابطه (۳) نشان داده شده است.

$$Sentence\ TF/IDF\ Feature: \frac{\sum_{i=1}^{n} TF/IDF(w_i.d,D)}{n} \quad (۳)$$

در رابطه‌ی (۳)، w_i، کلمه iام در جمله S و n تعداد کل کلمات موجود در جمله‌ی S را نشان می‌دهد.

۳-۱-۲- ویژگی میزان تشابه جملات با عنوان سند

نسبت تعداد کلمات مشترک در یک جمله از سند به تعداد کل کلمات عنوان را امتیاز جمله از نظر میزان تشابه آن با عنوان، در نظر می‌گیریم. البته میزان تشابه با

عنوان بعد از انجام عملیات پیش‌پردازش جمله و عنوان، امتیازدهی می‌گردد. رابطه‌ی (۴) روش محاسبه‌ی ویژگی تشابه با عنوان را نشان می‌دهد. در این رابطه مقدار میزان اشتراک جمله مورد نظر بر تعداد کلمات عنوان نرمال می‌گردد، این عمل سبب می‌شود که درک مناسب‌تری از میزان تشابهات ایجاد گردد، چرا که اگر سند دارای یک عنوان طولانی باشد مطمئناً شمارش تعداد تشابهات به تنهایی محاسبه‌ی درستی از معیار نخواهد بود.

$$\text{TitleSimilarityFeature} : \frac{|S \cap T|}{|T|} \quad (۴)$$

در رابطه‌ی (۴)، $|S \cap T|$ نشان دهنده‌ی تعداد کلمات مشترک بین عنوان و جمله‌ی موردنظر است و $|T|$ تعداد کلمات عنوان را نشان می‌دهد.

۳-۱-۳- ویژگی موقعیت جمله

در متون خبری فارسی عموماً جملات ابتدایی و انتهایی متن حاوی بار معنایی بیشتری نسبت به جملات میانی هستند. از آنجایی که متون مورد استفاده در سامانه‌ی پیشنهادی، متون خبری است لذا از این قاعده به عنوان یکی از معیارهای امتیازدهی به جملات سند بهره می‌بریم. در واقع اگر جمله‌ی موردنظر در ابتدای متن و یا در انتهای متن ظاهر شود، میزان اهمیت آن جمله نسبت به مابقی جملات بیشتر خواهد بود لذا به ویژگی موقعیت جمله‌ی مربوط به جملات ابتدایی و انتهایی سند مقدار عددی ۱ اختصاص داده می‌شود و ویژگی موقعیت جمله در سایر جملات به صفر مقداردهی شده است. هدف از این ویژگی برجسته کردن میزان اهمیت جملات ابتدایی و انتهایی سند در فرآیند امتیازدهی کلی به یک جمله است.

۳-۱-۴- ویژگی برچسب اجزای کلام

برچسب‌زنی اجزای کلام به عمل دسته‌بندی کلمات یک متن براساس ماهیت آنها از نظر اجزای گرامری زبان است. بنابراین هر کلمه را می‌توان در یکی از دسته‌های فعل، اسم، صفت، قید و غیره قرار داد. اسامی، یکی از مهمترین کلمات جمله در انتقال بار معنایی آن محسوب می‌شوند. صفات نیز به‌عنوان یکی از اجزای کلام می‌تواند حاوی اطلاعات مفیدی از نظر معنای جمله باشد [۱۸، ۱۹]. در این پژوهش امتیاز مربوط به اجزای کلام، از مجموع تعداد اسامی و صفات بکاررفته در یک جمله، تقسیم بر تعداد کل کلمات آن جمله بدست می‌آید. رابطه‌ی (۵) نحوه محاسبه‌ی این ویژگی را نشان می‌دهد.

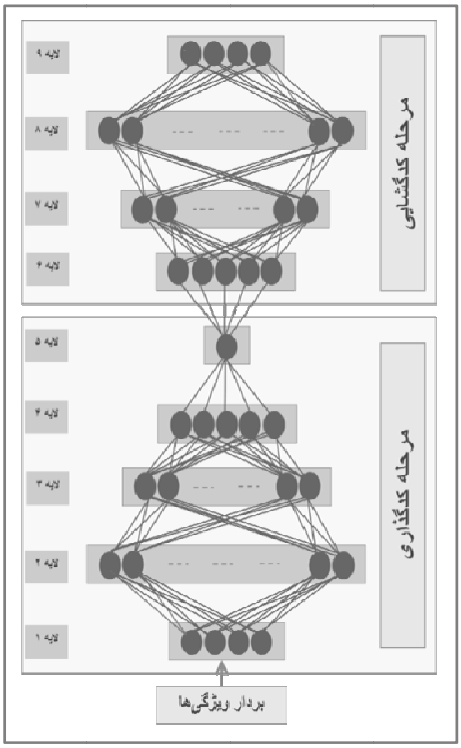
$$\text{POSScoreFeature} : (S_{|N|} + S_{|Adj|}) / |S| \quad (۵)$$

در رابطه‌ی (۵)، $S_{|N|}$ تعداد اسامی در جمله S، $S_{|Adj|}$ تعداد صفات در جمله‌ی S و $|S|$ تعداد کل کلمات موجود در جمله‌ی S را نشان می‌دهد.

۳-۲- استفاده از یادگیری عمیق در امتیازدهی به جملات و تولید متن خلاصه

پس از مرحله‌ی پیش‌پردازش متن و تولید بردار ویژگی‌ها می‌بایست به جملات امتیازی مبنی بر میزان اهمیت آنها اختصاص داده شود. روش پیشنهادی ما برای امتیازدهی به جملات، استفاده از یادگیری عمیق و به کمک یک شبکه‌ی عصبی است. در این پژوهش از یک شبکه‌ی Autoencoder که به انضمام لایه‌ی ورودی شامل ۹ لایه است، بهره برده‌ایم. در Autoencoderها لایه‌ی ورودی مشابه لایه‌ی خروجی است و هدف بازسازی مقادیر گره‌های ورودی در لایه‌ی خروجی است. در نتیجه داده‌های ما بدون برچسب هستند. عملکرد این نوع شبکه به دو بخش کدگذاری و کدگشایی تقسیم‌بندی می‌شود. در مرحله‌ی کدگذاری، شبکه سعی در کدگذاری داده‌ها و تولید ویژگی‌های جدید از ویژگی‌های ورودی دارد، پس از آن و در مرحله کدگشایی، شبکه به بازسازی داده‌های ورودی از روی ویژگی‌های جدید تولید شده در پایان مرحله‌ی کدگذاری، می‌پردازد. شبکه‌های Autoencoder در آموزش‌های بدون ناظر یعنی زمانی که داده‌های آموزشی ورودی فاقد برچسب هدف هستند، کاربرد دارند.

تصمیم‌گیری در مورد تعداد لایه‌ها و گره‌های مخفی در شبکه عملی تجربی بوده و با آزمایشات مکرر می‌بایست در مورد بهترین حالت تصمیم‌گیری نمود. در شبکه‌ی پیشنهادی ما لایه‌ی ورودی و یا همان لایه اول شامل ۴ نورون است که به‌عنوان ورودی، بردار ویژگی ۴ عضوی استخراج شده‌ی مربوط به ویژگی‌های هر جمله به آن تزریق می‌شود. در لایه‌های دوم، سوم و چهارم به‌ترتیب ۱۵، ۱۰ و ۵ نورون داریم و لایه‌ی پنجم، جایی که مرحله‌ی کدگذاری به پایان می‌رسد، شبکه دارای یک نورون است که مقدار آن امتیازی است که شبکه به جمله‌ی ورودی می‌دهد. بعد از آن، مرحله‌ی بازسازی ورودی و یا کدگشایی آغاز می‌شود. بنابراین لایه‌های ششم، هفتم، هشتم و نهم به ترتیب شامل ۵، ۱۰، ۱۵ و ۴ نورون هستند. شکل ۱ شمایی از شبکه‌ی پیشنهادی را نشان می‌دهد.



شکل ۱- شبکه عمیق پیشنهادی

برای آموزش و آزمایش سیستم خلاصه‌ساز پیشنهادی از پیکره‌ی پاسخ [۲۰] استفاده شده است. پیکره‌ی «پاسخ» اولین پیکره‌ی متنی برای ارزیابی خلاصه‌سازی تک‌سندی و چندسندی است که توسط آزمایشگاه فناوری وب دانشگاه فردوسی مشهد و با همکاری سازمان فناوری اطلاعات ایران تولید گردیده است. «پاسخ» یک پایگاه داده متشکل از حجم زیادی از اسناد خبری فارسی در موضوعات متنوع است. همچنین این پیکره دربردارنده‌ی خلاصه‌های تولید شده

توسط عوامل انسانی به اَشکال تک‌سندی، چندسندی، استخراجی و چکیده‌ای است. تیم تولیدکننده پیکره «پاسخ» برای ساخت متن خلاصه توسط عامل انسانی، از ۱۰ دانشجوی آموزش دیده، کمک گرفته است. به‌منظور جلوگیری از تاثیر سلایق و گرایشات شخصی در تولید خلاصه، عمل خلاصه‌سازی هر سند به ۵ نفر محول گشته است. پیکره «پاسخ» در بخش چندسندی دارای ۵۰ عنوان کلی خبری است که هر عنوان شامل ۲۰ سند است. در مجموع پیکره «پاسخ» در بخش چندسندی حاوی ۱۰۰۰ سند است که از این بین ۸۰۰ سند در مرحله آموزش و ۲۰۰ سند در مرحله آزمایش سامانه‌ی پیشنهادی مورد استفاده قرار گرفته است.

به‌منظور آموزش شبکه از ۸۰۰ سند متنی خبری در ۴۰ دامنه‌ی موضوعی متفاوت موجود در پیکره «پاسخ»، بهره گرفته شده است. مجموعاً تعداد ۱۳۷۹۵ جمله برای آموزش شبکه پیش‌پردازش شد و بردار ویژگی‌های مختص به هر جمله استخراج گردید.پس از آموزش شبکه و تنظیم پارامترهای مربوط به آن، در فاز ایجاد متن خلاصه، از شبکه‌ی آموزش دیده شده برای امتیازدهی به جملات استفاده می‌شود. سپس جملات براساس امتیاز اختصاص داده به آن‌ها مرتب شده و با توجه به نرخ فشرده‌سازی متن خلاصه که پیش فرض آن ۱۰ درصد است، جملاتی که بالاترین امتیاز را دارند برای تشکیل متن خلاصه انتخاب می‌شوند.

۴- ارزیابی روش پیشنهادی

برای ارزیابی روش پیشنهادی، با توجه به مطالعات و جستجوهای انجام شده، سیستم خلاصه‌ساز چندسندی قابل دسترسی دیگری که از پیکره‌ی پاسخ برای آموزش و ارزیابی استفاده کرده باشد، یافت نشد. یکی از شروط لازم برای مقایسه‌ی دو سامانه‌ی خلاصه‌سازی خودکار، اشتراک در داده‌های آزمایشی است. به دلیل عدم‌دسترسی به سیستم خلاصه‌ساز چندسندی فارسی دیگری که بتوان داده‌های آزمایشی برگرفته از پیکره «پاسخ» و مشابه با سامانه پیشنهادی را به آن اعمال نمود لذا ارزیابی‌های انجام شده در این مقاله محدود به مقایسه‌ی خروجی سیستم خلاصه‌ساز پیشنهادی با خلاصه‌های انسانی موجود در پیکره‌ی پاسخ، می‌شود.

سامانه‌ی پیشنهادی برروی ۱۰ مجموعه سند موجود در پیکره «پاسخ» که هر یک دارای ۲۰ سند خبری است، مورد ارزیابی و آزمایش قرار گرفت. تعداد ۲۴۹۳ جمله‌ی آزمایشی برای تولید بردار ویژگی پردازش گشت و شبکه‌ی پیشنهادی برای امتیازدهی به این داده‌های آزمایشی، به کار گماشته شد و خلاصه‌های متناظر با هر مجموعه سند تولید گشت.

یک روش ارزیابی استفاده شده در این مقاله، مقایسه‌ی جمله به جمله‌ی متن خروجی سامانه با خلاصه‌ی انسانی متناظر با آن است، جملاتی که عیناً مشابه یکدیگرند در امتیازدهی به عملکرد سامانه‌ی خلاصه‌ساز شرکت داده می‌شوند. معیارهای ارزیابی Precision، Recall و F-Score برای متن خلاصه محاسبه شده است. به‌عنوان مثال، برای یک مجموعه ۲۰ سندی که شامل ۱۸۰ جمله است، یک خلاصه‌ی ۱۸ جمله‌ای توسط خلاصه‌ساز تولید شد، خلاصه‌ی انسانی متناظر با این مجموعه سند شامل ۱۵جمله است. تعداد جملاتی که عیناً در هردو خلاصه‌ی تولید شده توسط سیستم و خلاصه‌ی تولید شده توسط انسان، ظاهر شده ۷ جمله است.

همچنین سیستم خلاصه‌ساز پیشنهادی را در ۵ حالت مختلف از نظر تعداد و نوع ویژگی‌های استخراج شده برای هر جمله، مورد ارزیابی قرار دادیم. بدین‌منظور از ۳ ویژگی دیگر برای جملات استفاده شده است که این ۳ ویژگی شامل تعداد کلمات ایست در جمله، وجود ضمیر در ابتدای جمله و طول جمله است [۲۱]. معمولاً جملاتی که حاوی تعداد کلمات ایست بالایی هستند، کلمات پراهمیت کمتری را شامل می‌شوند. در نتیجه نسبت تعداد کلمات ایست یک جمله به کل تعداد کلمات آن جمله به‌عنوان ویژگی تعداد کلمات ایست در جمله در نظر گرفته شده است. به‌طور کلی جملاتی که آغازگر آن‌ها یک ضمیر است به نوعی حاوی توضیحات در رابطه با جملات ماقبل خود هستند، در واقع این نوع جملات وابسته به جملات دیگر بوده و لحاظ کردن آن‌ها در متن خلاصه ممکن است میزان خوانایی آن متن را کاهش دهد. نسبتی از تاثیر مکان قرارگیری ضمیر در جمله، تعداد ضمایر در جمله و تعداد کلمات آن جمله بیانگر مقدار ویژگی تاثیر ضمیر در جمله است. ویژگی دیگر تاثیر طول جمله است، جملات طولانی و همچنین جملات بسیار کوتاه، جملات مناسبی برای انتخاب در متن خلاصه نیستند لذا حد آستانه‌ای برای طول جملات در نظر گرفته می‌شود و جملات نسبت به آن حد آستانه امتیازدهی می‌شوند. در هر یک از حالات سعی شده بهترین مجموعه ممکن از ویژگی‌ها انتخاب شود. جدول ۱ در بردارنده‌ی ویژگی‌های انتخابی در هر یک از این حالات است.

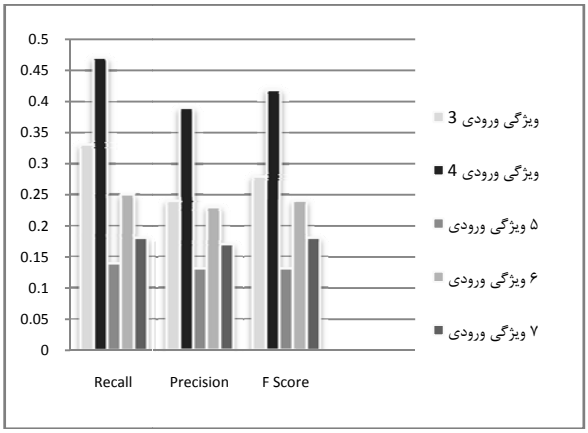
جدول ۱- حالات مختلف از نظر تعداد ویژگی‌های استخراج شده

تعدادویژگی‌ها / ویژگی‌ها	۳	۴	۵	۶	۷
میزان تشابه با عنوان	*	*	*	*	*
مکان رخداد جمله	*	*	*	*	*
TF/IDF	*	*	*	*	*
برچسب اجزای کلام		*	*	*	*
تعداد کلمات ایست در جمله			*	*	*
وجود ضمیر در ابتدای جمله				*	*
تاثیر طول جمله					*

نتایج حاصل از مقایسه‌ی این ۵ حالت در شکل ۲ نشان داده شده است. همانطور که قابل مشاهده است، زمانی که شبکه از ۴ ویژگی برای ارزیابی میزان اهمیت جمله استفاده نموده، نتایج بهتری را نسبت به ۴ حالت دیگر کسب کرده است. می‌توان یکی از دلایل رخداد این امر را مسأله‌ی تنک[16] بودن داده‌های آموزشی دانست. در واقع ممکن است افزایش و یا کاهش تعداد ویژگی‌ها و ترکیبات آن‌ها به‌دلیل عدم‌گستردگی حالات به وجود آمده در سطح مجموعه داده‌های در دسترس باعث گردد، اطلاعاتی که در رابطه با میزان اهمیت یک جمله بدست می‌آید، قابل تعمیم بر روی داده‌های آزمایشی نباشد. دلیل دیگر را می‌توان عدم‌هماهنگی معماری شبکه پیشنهادی با تغییر در تعداد ویژگی‌های ورودی دانست. در واقع با توجه به مساله تنک بودن داده‌ها و معماری شبکه، سامانه‌ی پیشنهادی در حالتی که از ۴ ویژگی شامل میزان تکرار کلمات، میزان تشابه جملات با عنوان سند، موقعیت قرارگیری جمله در سند و امتیاز مربوط به برچسب‌گذاری اجزای کلام استفاده می‌کند، بهترین عملکرد را از خود نشان داده است.

جدول ۲ مقادیر معیارهای ارزیابی مربوط به مقایسه‌ی جمله‌ای در حالتی که شبکه از ۴ ویژگی استفاده کرده است را نشان می‌دهد. با توجه به ماهیت چندسندی خلاصه‌ساز پیشنهادی، ممکن است در اسناد مختلف یک مجموعه، جملاتی وجود داشته باشند که در عین حالی‌که حامل بارمعنایی مشابه‌ای هستند، اما از کلمات و شیوه‌ی نگارش متفاوتی بهره برده‌اند. در ارزیابی‌های صورت گرفته بر روی خروجی خلاصه‌ساز، به جملاتی برخوردیم که از نظر مفهومی مشابه جملاتی در خلاصه‌ی تولید شده توسط انسان بودند، اما از آنجایی که معیار ارزیابی ما جملات کاملاً یکسان بود لذا جملات با مفهوم مشترک را در ارزیابی شرکت ندادیم هرچند که با احتساب اینگونه جملات می‌توان دقت خلاصه‌ساز را بالاتر از مقادیر موجود در جدول ۲ در نظر گرفت. با توجه به جدول ۲ می‌توان اینگونه

برداشت کرد که خروجی سامانه پیشنهادی، در قیاس جمله‌ای و در حالت میانگین نزدیک به ۵۰ درصد مشابه خلاصه‌ی انسانی موجود در پیکره است.



شکل ۲- ارزیابی خلاصه‌ساز با تعداد ویژگی‌های ورودی مختلف

جدول ۲- نتایج حاصل از ارزیابی جمله‌ای

معیارهای ارزیابی	Recall	Precision	F Score
مقادیر بدست آمده از مقایسه‌ی جمله‌ای	۰/۴۶۶۷	۰/۳۸۸۹	۰/۴۲۴۳

یک بسته‌ی نرم‌افزاری استاندارد برای ارزیابی سیستم‌های خلاصه‌سازی خودکار متون و ترجمه‌ی ماشینی تحت عنوان ROUGE موجود است. این نرم‌فزار به مقایسه‌ی خلاصه‌های تولید شده توسط سیستم‌های مختلف با یکدیگر و همچنین خلاصه‌های تولید شده توسط انسان می‌پردازد [۲۲]. در این پژوهش، توسط نرم‌افزار ROUGE، خلاصه‌ی تولید شده توسط سیستم خلاصه‌ساز پیشنهادی را با خلاصه‌های انسانی موجود در پیکره پاسخ و براساس unigramها و bigramهای مشابه مورد ارزیابی قرار دادیم. در واقع میزان کلمه‌ها و جفت کلمه‌های مشابه مابین خلاصه‌های تولید شده توسط سامانه و خلاصه‌های ایده‌آل مدنظر قرار می‌گیرند. لازم به‌ذکر است که این ارزیابی‌ها با حذف کلمات ایست و نادیده گرفتن آن‌ها انجام شده است. نتایج بدست آمده از این ارزیابی در جدول ۳ قابل مشاهده است.

جدول ۳- نتایج حاصل از ارزیابی توسط نرم‌افزار ROUGE

معیارهای ارزیابی	Recall	Precision	F Score
مقایسه براساس unigramهای مشابه	۰/۶۸۵۰	۰/۳۷۲۶	۰/۴۸۲۷
مقایسه براساس bigramهای مشابه	۰/۵۱۲۷	۰/۲۷۲۸	۰/۳۵۶۲

۵- نتیجه‌گیری

در این مقاله طراحی یک سیستم خلاصه‌ساز استخراجی چندسندی متون فارسی، با بهره‌گیری از پیکره‌ی پاسخ تشریح گشت. ۴ ویژگی تاثیرگذار در تصمیم‌گیری در رابطه با میزان اهمیت یک جمله از متن، برای تمامی جملات موجود در پیکره‌ی آموزشی استخراج شد. برای امتیازدهی به جملات براساس ویژگی‌های استخراج شده، از روشی تحت عنوان یادگیری عمیق استفاده کردیم. در این روش با استفاده از یک شبکه‌ی عصبی مصنوعی به‌نام Autoencoder و تزریق بردار ویژگی‌ها به‌عنوان ورودی به آن، به آموزش سیستم برای امتیازدهی به جملات پرداختیم. در فاز نهایی برای تولید متن خلاصه، جملات براساس امتیازهای بدست آمده، مرتب می‌شوند و جملاتی که بالاترین امتیاز را دارا هستند، تشکیل متن خلاصه را می‌دهند.

متن خلاصه‌ی تولید شده توسط سیستم پیشنهادی با خلاصه تولید شده توسط انسان مورد مقایسه و ارزیابی قرار گرفت. این ارزیابی به ۳ شکل، ارزیابی جمله‌ای، ارزیابی براساس unigramهای مشابه و ارزیابی براساس bigramهای مشابه صورت گرفت. نتایج حاصل از این ارزیابی‌ها نشان از عملکرد خوب سامانه‌ی پیشنهادی داشته و توانایی آن را در تولید یک خلاصه‌ی قابل‌قبول با بهره‌گیری از تعداد محدودی ویژگی ورودی را بازگو می‌نماید. در واقع یکی از مزایای استفاده از یادگیری عمیق در این سامانه، کاهش صرف وقت و هزینه در تولید و طراحی ویژگی‌ها و معیارهای اهمیت متن است، زیرا شبکه‌ی پیشنهادی به لطف استفاده از یادگیری عمیق قابلیت تولید ویژگی‌های جدید از ویژگی‌های محدود ورودی را دارد.

مراجع

[1] D. Das, and A. Martins, "A Survey on Automatic Text Summarization," Literature Survey for the Language and Statistics II Course at Carnegie Mellon University, 2007.

[2] D. Radev, and T. Allison, "MEAD – A platform for multidocument multilingual text summarization," Proceedings ofthe 4th International Conference on Language Resources and Evaluation (LREC), pp. 699-702, 2004.

[3] A. Thiago, and P. Salgueiro, "GistSumm: A Summarization Tool Based on a New Extractive Method," Proceedings of 6th International Conference on Computational, pp. 210-218, 2012.

[4] M. Hassel, and N. Mazdak, "FarsiSum - A Persian text summarizer," Proceedings of the Workshop On Computational Approaches to Arabic Script-Based Languages, pp. 82-84, 2004.

[۵] ز. کریمی، و م. شمس‌فرد، "خلاصه‌سازی متون فارسی،" در مجموعه مقالات یازدهمین کنفرانس سالانه انجمن کامپیوتر ایران، ص ۱۲۹۴-۱۲۸۶، ۱۳۸۴.

[6] M. A. Honarpisheh, Gh. Ghassem-Sani, and G. Mirroshandel, "A Multi-Document Multi-Lingual Automatic Summarization System," Proceedings of the Joint Conference on Natural Language Processing, pp. 733-738, 2008.

[7] F. Kiyoumarsi, and F. Rhimi Esfahani, "Optimizing Persian Text Summarization Based on Fuzzy Logic Approach," Proceedings of theInternational Conference on Intelligent Building and Management, pp. 264-269, 2011.

[8] G. Hinton, O. Simon, and T. Yee-Whye, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, vol. 18, pp. 1527-1554, 2006.

[9] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-Moisy, "An Introduction to Deep Learning," Proceedings of theEuropean Symposium on Artificial Neural Networks-Computational Intelligence and Machine Learning, pp. 477-488, 2011.

[10] Y. Bengio, "Learning Deep Architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, and A. Mohamed, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[12] R. Collobert, and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," Proceedings of theInternational Conference on Machine Learning, pp. 160-167, 2008.

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, M. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (almost) from Scratch," Journal of the Machine Learning Research, vol. 12, pp. 2493-2537, 2011.

[14] E. Arısoy, T. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep Neural Network Language Models," Proceedings of the NAACL-HLT, pp. 20-28, 2012.

[15] Y. Liu, S. Zhong, and W. Li, "Query-Oriented Multi-Document Summarization via Unsupervised Deep Learning," Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp. 1699-1705, 2012.

[16] A. Fischer, and Ch. Igel, "An Introduction to Restricted Boltzmann Machines," Proceedings of the 17th Iberoamerican Congress on Pattern Recognition, pp. 14-36, 2012.

[17] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked DenoisingAutoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of the Machine Learning Research, vol. 11, no.11, pp. 3371-3408, 2010.

[18] N. Prabhakar, and N.Chandra, "Automatic Text Summarizaion Based on Pragmatic Analysis," International Journal of the Scientific and Research Publications, vol. 2, Issue 5, pp. 1-4, 2012.

[19] R. Mihalcea, and P. Tarau, "TextRank: Bringing Order into Texts,"Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 404-411, 2004.

[20] B. Behmadi Moghaddas, M. Kahani, S. A. Toosi, A. Pourmasoumi, and A. Estiri, "Pasokh: A standard corpus for the evaluation of Persian text summarizers," Proceedingsof the Computer and Knowledge Engineering (ICCKE), pp. 471-475, 2013.

[۲۱] الف. پورمعصومی، م. کاهانی، الف. طوسی، الف. استیری، و ه. قائمی "ایجاز: یک سامانه‌ی عملیاتی برای خلاصه‌سازی تک‌سندی متون خبری فارسی،" دوفصلنامه‌ی پردازش علائم و داده‌ها، شماره ۱، پیاپی ۲۱، ص۳۳-۴۸، ۱۳۹۳.

[22] Ch. Lin, "Rouge: A package for automatic evaluation of summaries," Proceedings of the ACL workshop on Text Summarization Branches Out, pp. 74-81, 2004.

شیما محرابی فارغ‌التحصیل ازموسسه آموزش عالی طبرستان در مقطع کارشناسی و دانشگاه گیلان در مقطع کارشناسی‌ارشد در رشته مهندسی کامپیوتر گرایش نرم‌افزار. زمینه‌های تحقیقاتی مورد علاقه یادگیری ماشین، یادگیری عمیق، پردازش زبان‌های طبیعی، خلاصه‌سازی خودکار متون هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shima.mehrabi85@gmail.com

حمیدرضا احمدی‌فر فارغ‌التحصیل از دانشگاه شهید بهشتی در مقاطع کارشناسی و دکتری و دانشگاه صنعتی امیرکبیر در مقطع کارشناسی‌ارشد در رشته مهندسی کامپیوتر با گرایش معماری سیستم‌های کامپیوتری. از سال ۱۳۸۲ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه حساب کامپیوتری، سیستم‌های توزیع شده، رایانش ابری و بکارگیری شبکه‌های عصبی در حل مسائل مختلف هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

ahmadifar@guilan.ac.ir

سید ابوالقاسم میرروشندل فارغ‌التحصیل از دانشکده فنی دانشگاه تهران در مقطع کارشناسی در رشته مهندسی کامپیوتر با گرایش نرم‌افزار و دانشگاه صنعتی شریف در مقاطع کارشناسی‌ارشد و دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی. از سال ۱۳۹۱ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه ایشان پردازش زبان‌های طبیعی، داده‌کاوی، یادگیری ماشینی و پردازش تصویر هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

mirroshandel@guilan.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۰۷/۱۱

تاریخ اصلاح: ۱۳۹۵/۱۰/۰۲

تاریخ قبول شدن: ۱۳۹۵/۱۰/۲۵

نویسنده مرتبط: دکتر حمیدرضا احمدی‌فر، دانشکده فنی، دانشگاه گیلان، رشت، ایران.

[1] Text Summarization
[2] Extractive Summarization
[3] Abstractive Summarization
[4] Support Vector Regression
[5] Deep Learning
[6] Name Entity Recognition
[7] Part Of Speech Tagging
[8] Semantic Role Labeling
[9] Chunking

[10]Language Modeling
[11]Restricted Boltzmann Machine
[12]Feed-Forward Neural Networks
[13]Unsupervised Layer-Wise Pre-Training
[14]Deep Belief Networks
[15]Stop Words
[16]Data Sparseness