قطعه‌بندی عبارات متون فارسی با استفاده از شبکه‌های عصبی

محمدمهدی میردامادی، علی‌محمد زارع بیدکی و مهدی رضائیان

چکیده: قطعه‌بندی کلمات و عبارات متن، یکی از فعالیت‌های اصلی در حوزه پردازش زبان‌های طبیعی است. اکثر برنامه‌های پردازش زبان‌های طبیعی به یک پیش‌پردازش برای استخراج کلمات متن و تشخیص عبارات احتیاج دارند. هدف اصلی و نهایی قطعه‌بندی عبارات، به دست آوردن کلمات معنی‌دار همراه با پیشوندها و پسوندهایشان است و این فعالیت متناسب با زبان‌های طبیعی مختلف می‌تواند سخت یا آسان باشد. در زبان فارسی به علت وجود فاصله و نیم‌فاصله، عدم توجه کاربران به فاصله‌گذاری‌ها و نبود قواعد دقیق در نوشتن کلمات چندقسمتی، تشخیص و قطعه‌بندی کلمات چندقسمتی و مرکب با مشکلات و پیچیدگی‌های خاص خود روبه‌رو است.

در این مقاله برآنیم تا با استفاده از شبکه‌های عصبی، یک روش آماری برای قطعه‌بندی عبارات متون فارسی جهت استفاده در موتورهای جستجو ارائه کنیم. الگوریتم پیشنهادی شامل ٤ فاز است که با استفاده از احتمال رخداد تک‌کلمات و دوکلمه‌ای‌های موجود در پیکره و با دقت ۸۹٫٦٪ عمل قطعه‌بندی را انجام می‌دهد. نتایج آزمایشات نشان دادند این روش می‌تواند با قطعه‌بندی بهتر عبارات، بهبود نسبی در کارایی روش‌های معمول به وجود آورد.

کلید واژه: **پردازش زبان‌های طبیعی، شبکه‌های عصبی، قطعه‌بندی، موتور جستجو.**

۱- مقدمه

با گسترش روزافزون رسانه‌های ذخیره‌سازی الکترونیکی و رسانه‌های ارتباطی و همچنین پیشرفت سریع علم کامپیوتر و فراگیرشدن آن، امروزه با حجم عظیمی از متون نوشتاری دیجیتال و اسناد الکترونیکی مواجه هستیم [۱]. با گسترش این گونه اسناد، پردازش اسناد و متون مورد نظر از بین حجم عظیمی از اطلاعات متنی به صورت دستی کاری دشوار و در عمل غیر ممکن خواهد بود. از این رو پردازش خودکار متون نوشتاری مورد توجه قرار می‌گیرد که یکی از موضوعات پردازش زبان‌های طبیعی[1] است.

برای انجام پردازش خودکار متون نوشتاری به کوچک‌ترین واحد معنی‌دار متن یا کلمات بامفهوم نیاز داریم [۲]. کلمات بامفهوم، کلمات ساده، مرکب و یا جمعی هستند که یک مفهوم کلی را می‌رسانند، برای مثال "بین الملل" یک کلمه بامفهوم است. گرچه این کلمه در ظاهر دو کلمه املایی[2] (به دنباله‌ای از حروف اطلاق می‌شود که دارای معنی هستند) به نظر می‌رسد، اما در این مقاله آن را یک کلمه بامفهوم یا عبارت[3] در نظر می‌گیریم زیرا در کل به یک چیز اشاره می‌کند و در کنار هم، معنی کامل‌تری می‌دهد. قطعه‌بندی[4] به معنی تشخیص و استخراج این عبارات از متون نوشتاری یا گفتاری می‌باشد که یکی از مسایل اساسی در پردازش زبان‌های طبیعی است.

قطعه‌بندی و تشخیص صحیح مرز کلمات و عبارات در بسیاری از سیستم‌های پردازش زبان طبیعی مانند تشخیص گروه‌های نحوی و پردازش آنها در سیستم‌های ترجمه ماشینی[5]، استخراج اطلاعات، سیستم پرسش و پاسخ[6]، تشخیص نقش‌های موضوعی، موتورهای جستجو[7] و غیره نقش کلیدی ایفا می‌کند [۳]. با توجه به این کاربردها قطعه‌بندی صحیح کلمات می‌تواند موجب بهبود در بازدهی فعالیت‌های ذکرشده باشد.

شاید در ابتدای امر قطعه‌بندی عبارات امری ساده و آسان به نظر برسد، اما باید به این نکته توجه کرد که حتی در زبان‌هایی مانند فارسی و انگلیسی که از فاصله استفاده می‌کنند هم اگر تنها از فاصله به عنوان جداکننده برای قطعه‌بندی استفاده شود، نتیجه نهایی مطلوب نخواهد بود و باید تکنیکی استفاده شود که بتواند مرز کلمات بامفهوم کامل را به خوبی تشخیص دهد.

با توجه به ساختار زبان فارسی، در این زبان با مشکلات خاص خود مواجه هستیم که در ادامه به برخی از آنها اشاره می‌کنیم.

وجود رسم‌الخطهای مختلف و سبک‌های نگارش متفاوت در زبان فارسی باعث شده فاصله، معیار قطعی و دقیقی برای تشخیص مرز کلمه نباشد. به طور کل می‌توان گفت به تعداد افراد جامعه سبک‌های نگارش و رسم‌الخطهای مختلف وجود دارد. دو نوع فاصله درون کلمه و برون کلمه در متون نوشتاری زبان فارسی وجود دارد که در متون تایپی به ترتیب از نیم‌فاصله و فاصله استفاده می‌شود. وجود این دو نوع فاصله‌گذاری باعث شده کلمه‌ای مانند "می‌رفت" را بتوان به سه صورت "می رفت"، "می‌رفت" و میرفت" نوشت.

اکثر کاربران به فاصله‌گذاری‌ها توجه نمی‌کنند و همچنین قواعد دقیقی در نوشتن کلمات چندقسمتی وجود ندارد که باعث بروز مشکلات متعددی در قطعه‌بندی می‌شوند [۴]. همچنین زبان فارسی میان اشکال ابتدایی، میانی و انتهایی اغلب حروف بر حسب موقعیت آنها در یک کلمه تمایز قایل می‌شود. از این روست که وجود یک تا چهار مدل مختلف برای کاراکترها، موجب پیچیدگی در تشخیص برخی کلمات می‌شود. از دیگر مشکلات می‌توان به حروف و کلمات واردشده از زبان عربی به زبان فارسی اشاره کرد. برای مثال صداهای همزه و تنوین باعث می‌شوند کلمات "پائیز" و "پاییز" یا "حتماً" و "حتما" به دو شکل نوشته شوند. ابهام یونیکد برای حروف "ک" و "ی"، افعال و کلمات مرکب، تولید کلمات جدید و همچنین ورود واژه‌ها از زبان‌های دیگر که در به کارگیری

این مقاله در تاریخ ۶ مهر ماه ۱۳۹۱ دریافت و در تاریخ ۷ تیر ماه ۱۳۹۲ بازنگری شد.

محمدمهدی میردامادی، دانشکده برق و کامپیوتر، دانشگاه یزد، یزد، (email: mirdamadi@stu.yazduni.ac.ir).

علی‌محمد زارع بیدکی، دانشکده برق و کامپیوتر، دانشگاه یزد، یزد، (email: alizareh@yazduni.ac.ir).

مهدی رضائیان، دانشکده برق و کامپیوتر، دانشگاه یزد، یزد، (email: mrezaeian@yazduni.ac.ir).

1. Natural Language Processing
2. Orthographic Word
3. Phrase
4. Segmentation
5. Machine Translation Systems
6. Question Answering System
7. Search Engines

حروف برای نوشتن این کلمات ابهام ایجاد می‌کند، از دیگر مشکلاتی هستند که برای قطعه‌بندی با آنها روبه‌رو هستیم.

در این مقاله به دنبال ارائه روشی برای به دست آوردن عبارت‌های صحیح متون فارسی هستیم. ابتدا در بخش ۲ مروری بر کارهای گذشته که در این زمینه انجام شده است، داریم. در بخش ۳ به معرفی روش پیشنهادی برای قطعه‌بندی عبارات متون فارسی می‌پردازیم. بخش ۴ شامل نتایج پیاده‌سازی الگوریتم می‌باشد و نهایتاً در بخش ۵ نتیجه‌گیری و کارهای آینده آمده است.

۲- مروری بر کارهای گذشته

روش‌های گوناگونی برای قطعه‌بندی کلمات و عبارات وجود دارند. پرکاربردترین آنها روش‌های مبتنی بر قواعد، روش‌های آماری، روش‌های مبتنی بر فرهنگ واژگان و روش‌های یادگیری هستند که [۳] به معرفی آنها پرداخته است. این روش‌ها ممکن است به تنهایی یا به صورت ترکیبی مورد استفاده قرار گیرند که در ادامه به اختصار توضیح داده شده‌اند.

۲-۱ روش‌های مبتنی بر قواعد

این روش‌ها به دانش زبان‌شناسی اعم از معنایی و نحوی احتیاج دارند. این قواعد می‌توانند توسط انسان به صورت دستی تعریف شوند و یا از منابع زبانی مانند پیکره[1]های برچسب خورده با استفاده از یک رویه یادگیری استخراج شوند.

۲-۲ روش‌های آماری

این روش‌ها به دانش زبان‌شناسی نیاز ندارند و میزان موفقیت آنها وابستگی زیادی به منابع آمارگیری و پیکره‌ها دارد. روش‌های آماری قابل حمل‌تر از روش‌های دیگر هستند و معمولاً مستقل از زبان می‌باشند. این روش‌ها، عبارات پررخداد زبان، فرکانس تکرار و احتمال وقوع عبارات مختلف را به عنوان اطلاعات آماری از منابع زبانی مانند پیکره‌های پردازش‌شده، اسناد وب، خروجی موتورهای جستجو و ... استخراج می‌کنند.

۲-۳ روش‌های مبتنی بر فرهنگ واژگان

این روش‌ها با تطبیق کلمات جمله با مدخل‌های یک فرهنگ واژگان، قطعه‌بندی کلمات و عبارات را انجام می‌دهند. دقت آنها به پوشش فرهنگ واژگان بستگی دارد و اگر با کلمه جدیدی روبه‌رو شوند، شکست می‌خورند. در این روش نیاز است از ابزارهای ریشه‌یابی و تحلیل ساخت‌واژی برای کاهش تعداد کلمات ناموجود در فرهنگ واژگان استفاده شود.

۲-٤ روش‌های یادگیری

در روش‌های یادگیری سیستم، اطلاعات مربوط به قطعه‌بندی را از منابع ورودی دریافت می‌کند. این اطلاعات می‌توانند مدل زبانی، قواعد نحوی و معنایی و یا اطلاعات آماری مورد نیاز سیستم باشند. در این روش‌ها منبع یادگیری عمدتاً پیکره‌ها و واژگان هستند. پیکره‌های برچسب خورده با مقوله نحوی که نشانه‌گذاری[2] شده‌اند یکی از مناسب‌ترین منابع زبانی برای یادگیری قواعد قطعه‌بندی می‌باشند که در بسیاری زبان‌ها از جمله فارسی در دسترس نمی‌باشند. لذا در این زبان‌ها، عدم وجود پیکره مناسب، استفاده از این روش را با مشکل روبه‌رو می‌کند.

در ادامه به معرفی چند سیستم قطعه‌بندی پرداخته‌ایم که در مقالات مختلف برای مقاصد گوناگون آورده شده‌اند و از روش‌های فوق برای قطعه‌بندی استفاده کرده‌اند.

شمس‌فرد در [۵] به معرفی سیستمی به نام STeP −۱ پرداخته است. در این سیستم برای قطعه‌بندی کلمات متون فارسی از ترکیب روش‌های مبتنی بر قواعد و مبتنی بر فرهنگ واژگان استفاده می‌کند.

در [۶] چانگ برای ترجمه ماشینی از زبان چینی به زبان انگلیسی و با هدف قطعه‌بندی کلمات چینی، سیستمی را ارائه کرده که از روش‌های آماری و مبتنی بر قواعد توأمان استفاده می‌کند. برای ترجمه ماشینی این سیستم، اطلاعات در دسترس از پیکره‌های موازی را جهت تشخیص کلمات ترکیب می‌کند و از احتمال شرطی و یادگیری بیزی برای به دست آوردن روشی دقیق‌تر استفاده کرده است.

فرونزا در [۷] یک روش یادگیری باناظر را برای انجام قطعه‌بندی پیشنهاد می‌کند. در روش او سعی شده عبارات مرکب به عنوان یک عبارت واحد تشخیص داده شود. روش به صورت خودکار، قوانین قطعه‌بندی را از یک پیکره قطعه‌بندی شده یاد می‌گیرد. این روش بر روی زبان‌های رومانی و انگلیسی، بهبود قابل توجه آماری دارد.

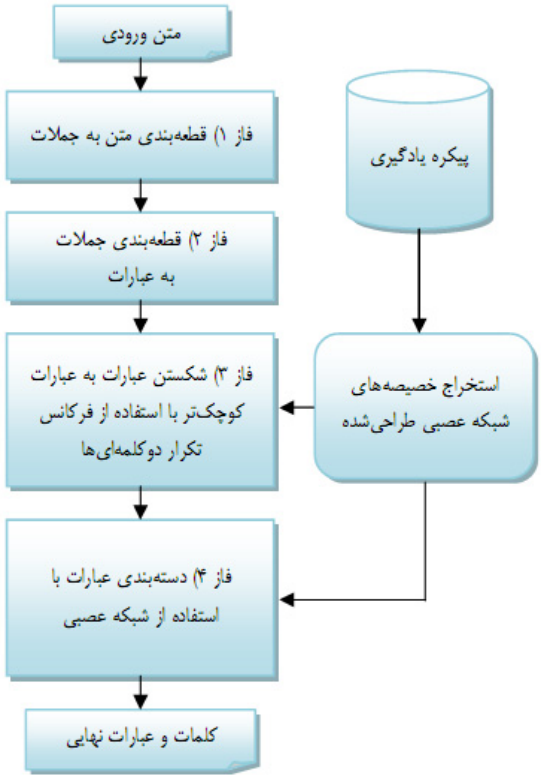
در [۳] یک سیستم قطعه‌بندی عبارات ارائه شده است. سیستم مذکور به برچسب‌گذاری با براکت و برچسب‌گذاری با فرمت IOB جهت تشخیص محدوده گروه‌ها روی آورده و با استفاده از شبکه عصبی که ورودی‌های آن ترکیبی از برچسب‌های POS[3] کلمات و خروجی آن برچسب IOB بود، به قطعه‌بندی عبارات پرداخته است. پس از ایجاد داده‌های ورودی و خروجی، یک شبکه عصبی MLP با سه لایه برای برچسب‌دهی خودکار IOB طراحی شده است. برای ارزیابی سیستم قطعه‌بندی عبارات از پیکره بیجن‌خان استفاده شده که میزان موفقیت آن حدود ۸۰٪ می‌باشد.

در [۸] یک برچسب‌زن[4] به گونه‌ای توسعه یافته که هم‌زمان با قطعه‌بندی کلمات، برچسب‌گذاری نیز می‌کند. برای این کار الگوریتم ویتربی[5] به گونه‌ای بر روی زبان اسپانیایی تعمیم داده شده که توانایی ارزیابی نشانه‌ها با طول‌های مختلف بر روی ساختار یکسان را داشته باشد. با استفاده از روش تعمیم‌یافته، قطعه‌بندی عبارات مبهم و پیچیدگی‌های آنها در یک فاز تحلیل می‌شود.

مرجع [۹] نمونه یک قطعه‌بندی با روش آماری برای زبان ویتنامی است که به دلیل کمبود پیکره و واژگان در این زبان، برای تعیین مستدل‌ترین قطعه‌بندی از اطلاعات آماری و الگوریتم ژنتیک استفاده می‌نماید. همچنین در [۱۰]، دو الگوریتم یادگیری C۴٫۵ و RIPPER برای قطعه‌بندی عبارات در زبان تایلندی با هم مقایسه شده‌اند.

مرجع [۱۱] بر روی قطعه‌بندی عبارات تمرکز کرده است. سیستم ارائه‌شده در [۱۱] یک روش ترکیبی برای قطعه‌بندی خودکار متون فارسی می‌باشد. سیستم مذکور در ابتدا از یک روش مبتنی بر قواعد برای ایجاد یک پیکره برچسب‌گذاری شده جهت آموزش شبکه عصبی بهره جسته است و سپس از یک شبکه عصبی چندلایه پرسپترون همراه با C - Means فازی برای قطعه‌بندی جملات و عبارات جدید استفاده می‌کند. نتایج آزمایشگاهی، نشان‌دهنده میانگین دقت ۸۵٫۷٪ برای این سیستم قطعه‌بندی می‌باشند.

1. Corpus
2. Tokenization
3. Part of Speech
4. Tagger
5. Viterbi Algorithm



شکل ۱: شمای کلی سیستم قطعه‌بندی کلمات و عبارات.

۳- روش ارائه‌شده

توجه به این نکته ضروری به نظر می‌رسد که قطعه‌بندی برای موضوعات گوناگون، روند متفاوتی را می‌طلبد. فرض کنید یک سیستم قطعه‌بندی برای یک سیستم تشخیص گفتار طراحی شده و خروجی آن عباراتی باشد که برای سیستم مربوط مناسب باشد، اما ممکن است استفاده از همین سیستم قطعه‌بندی برای سیستم ترجمه ماشینی نتیجه مطلوبی به ما ندهد. از این روست که برای هر هدفی باید سیستم قطعه‌بندی متناسب با آن هدف طراحی کرد. هدف از ارائه این مقاله، پیشنهاد سیستمی برای قطعه‌بندی کلمات و عبارات متون و نوشته‌ها جهت استفاده در موتورهای جستجو می‌باشد. در موتورهای جستجو علاوه بر قطعه‌بندی کلمات سایت‌ها برای نمایه‌سازی، باید پرس و جوی کاربران نیز قطعه‌بندی شود.

برای نمونه، فرض کنید پرس و جوی کاربر "دانشگاه یزد" باشد. اگر برای قطعه‌بندی پرس و جو تکنیک خاصی به کار گرفته نشود و تنها از فاصله بین کلمات استفاده شود، ممکن است عبارت "دانشگاه آزاد اسلامی واحد یزد" به عنوان یکی از جواب‌های این پرس و جو برگردانده شود که جواب صحیح نمی‌باشد. بنابراین قطعه‌بندی کلمات برای موتورهای جستجو باید به گونه‌ای انجام شود که کلماتی مثل "دانشگاه یزد" را به عنوان یک عنصر در نظر بگیرد. برای انجام این کار روشی پیشنهاد کرده‌ایم که شامل ۴ فاز است. در هر فاز به هدفمان که قطعه‌بندی کلمات و عبارات است نزدیک‌تر می‌شویم و در نهایت در انتهای فاز آخر، عبارات مربوط به متن ورودی را در اختیار داریم. بدین منظور فرکانس تکرار تمام تک‌کلمات و دوکلمه‌ای‌های همسایه موجود در پیکره را استخراج می‌کنیم و بر اساس این فرکانس‌ها قطعه‌بندی را انجام می‌دهیم.

در این مقاله الگوریتمی را با هدف قطعه‌بندی کلمات و عبارات متن ورودی پیشنهاد می‌کنیم که در شکل ۱ شمای کلی آن نشان داده شده است. در ادامه به تشریح هر یک از فازهای الگوریتم می‌پردازیم.

فاز ۱) در فاز اول جملات متن ورودی را استخراج می‌کنیم. برای این کار از علامت‌هایی که برای پایان جمله استفاده می‌شود از قبیل "."، "!"، "؟" و غیره استفاده می‌کنیم.

فاز ۲) جمله‌های به دست آمده از فاز قبل را مجدداً به عبارات کوچک‌تر تقسیم می‌کنیم. برای این کار از ایست‌واژه[1]ها یا حروف اضافه‌ای که بین عبارات در جمله قرار دارند، استفاده می‌شود.

با توجه به این که حروف اضافه‌ای چون "و"، "در"، "از" و غیره خود یک کلمه بامفهوم هستند و از طرفی فرکانس تکرار آنها خیلی بیشتر از دیگر کلمات است، در جمله، هر یک از این حروف را به عنوان یک نشانه در نظر گرفته و عباراتی که بین آنها هستند به فاز بعد منتقل می‌شوند. منظور از نشانه، کلمه یا عبارتی با معنی و مفهوم کامل است که با توجه به سیستم مورد استفاده تعریف می‌شود. برای مثال در موتورهای جستجو عبارت "جمهوری اسلامی ایران" یک نشانه در نظر گرفته می‌شود زیرا به یک مفهوم خاص اشاره می‌کند.

در (۱) حرف اضافه ۱و۲ به عنوان نشانه در نظر گرفته می‌شوند و عبارات ۱ و ۲ و ۳ به فاز بعد منتقل خواهند شد

$$\text{جمله} = \text{عبارت ۱} + \text{حرف اضافه ۱} + \text{عبارت ۲} + \text{حرف اضافه ۲} + \text{عبارت ۳} + \text{نشان پایان جمله} \quad (۱)$$

فاز ۳) عبارات به دست آمده از فاز ۲ ممکن است حاوی یک کلمه یا بیشتر باشند. اگر هر یک از این عبارات شامل یک کلمه بودند، آن کلمه را یک نشانه در نظر می‌گیریم و در غیر این صورت فرکانس تکرار هر یک از دوکلمه‌ای‌های عبارت، در پیکره یادگیری را با یک حد آستانه مقایسه می‌کنیم. اگر این فرکانس تکرار از حد آستانه بیشتر بود، دوکلمه‌ای مورد نظر در یک عبارت قرار می‌گیرد، در غیر این صورت هر کدام از آن کلمات در عباراتی جدا قرار خواهند گرفت و به اصطلاح عبارت فاز ۲ از آن قسمت شکسته می‌شود.

برای مثال جمله "آئین نامه خرید کتاب" را در نظر بگیرید. فرض کنید در این جمله دوکلمه‌ای ۱ = "آئین نامه"، دوکلمه‌ای ۲ = "نامه خرید" و دوکلمه‌ای ۳ = "خرید کتاب" باشند.

اگر فرکانس تکرار دوکلمه‌ای ۱ و دوکلمه‌ای ۳ بیش از حدآستانه و فرکانس تکرار دوکلمه‌ای ۲ کمتر باشند، در این صورت عبارت بالا به دو عبارت "آئین نامه" و "خرید کتاب" شکسته شده و به فاز ۴ فرستاده می‌شوند. در این فاز آن دسته از کلماتی که در پیکره یادگیری کمتر در همسایگی هم بوده‌اند و به احتمال کمی تشکیل یک نشانه را می‌دهند، مشخص می‌شوند. قابل ذکر است که عبارات فرستاده‌شده به فاز ۴ می‌توانند بیش از ۲ کلمه داشته باشند، در این صورت هر دو کلمه همسایه در عبارت مذکور فرکانس تکراری بیش از حد آستانه مورد نظر دارند. حد آستانه انتخاب‌شده برای این فاز برابر ۲۵ می‌باشد که این مقدار با انجام آزمایش‌های متعدد و به صورت سعی و خطا به دست آمده است.

فاز ۴) اگر عبارت حاصل از فاز ۳ شامل یک کلمه بود، آن کلمه خود یک نشانه است. برای حالتی که عبارت بیش از یک کلمه داشته باشد، عبارت را به دو دسته "نشانه" و "غیرنشانه" دسته‌بندی[2] می‌کنیم. هدف از انجام این کار این است که اگر عبارت به دسته نشانه تعلق گرفت، کار در این مرحله تمام می‌شود و آن عبارت به عنوان یک نشانه در نظر گرفته

1. Stopword
2. Classification

شده و به قطعه‌بندی دیگر کلمات متن می‌پردازیم. اما در صورتی که عبارت به دسته غیر نشانه تعلق بگیرد، نمی‌تواند یک نشانه باشد. در این حالت عبارت کوچک‌تر شده و یکی از طول عبارت کم می‌شود، یعنی بدون کلمه ابتدایی یا انتهایی مورد بررسی قرار می‌گیرد. اگر در حالت جدید به دسته نشانه تعلق گرفت که کار تمام می‌شود، در غیر این صورت همین روال تکرار شده تا جایی که همه اجزای عبارت اولیه به دسته نشانه تعلق بگیرند و یا این که آن قدر تجزیه شوند که به تک‌کلمات برسیم. توجه به این نکته ضروری به نظر می‌رسد که تک‌کلمات به عنوان یک نشانه به حساب می‌آیند و وارد شبکه نمی‌شوند.

برای درک بهتر این مطلب مثالی را بیان می‌کنیم. فرض کنید عبارت رسیده به فاز ۴ دارای پنج کلمه به صورت "کلمه۱ کلمه۲ کلمه۳ کلمه۴ کلمه۵" باشد. همچنین فرض کنید اگر خروجی شبکه عصبی یک بود عبارت مربوط به دسته نشانه و اگر صفر بود به دسته غیر نشانه تعلق گیرد. در ابتدا خصیصه[1]های کل عبارت را به شبکه می‌دهیم، اگر خروجی شبکه یک بود، کل این پنج کلمه یک نشانه را تشکیل می‌دهند اما اگر خروجی صفر بود، این پنج کلمه یک نشانه نیستند و باید به عبارات کوچک‌تر تقسیم شوند. برای انجام این کار عبارت را به دو عبارت چهارکلمه‌ای و تک‌کلمه‌ای می‌شکنیم. البته باید توجه کرد که دو عبارت چهارکلمه‌ای و دو عبارت تک‌کلمه‌ای وجود دارد. یعنی یک بار عبارت را به صورت "کلمه۱" و "کلمه۲ کلمه۳ کلمه۴ کلمه۵" می‌شکنیم و در حالتی دیگر به "کلمه۱ کلمه۲ کلمه۳ کلمه۴" و "کلمه۵" تقسیم می‌کنیم.

برای هر دو حالت با استفاده از شبکه به دسته‌بندی زیرعباراتی که بیش از یک کلمه دارند، می‌پردازیم. اگر فقط برای یکی از حالت‌های بالا، زیرعباراتی که چهار کلمه دارد به دسته نشانه تعلق گرفت، آن گاه زیرعبارت‌های آن حالت هر کدام حکم یک نشانه را دارند. اگر هر دو چهارکلمه‌ای در هر دو حالت به دسته نشانه تعلق گرفتند آن گاه آن زیرعباراتی که درجه تعلق آن به دسته نشانه بیشتر است، یعنی خروجی شبکه عصبی به عدد یک نزدیک‌تر است، به عنوان نشانه انتخاب می‌شود و آن حالت، حالت برنده خواهد بود.

اگر هیچ کدام از چهارکلمه‌ای‌ها به دسته نشانه تعلق نداشتند عبارت اولیه را به سه حالت در می‌آوریم. حالت اول به صورت "کلمه۱ کلمه۲ کلمه۳" و "کلمه۴ کلمه۵"، حالت دوم به صورت "کلمه ۱"، "کلمه۲ کلمه۳ کلمه۴" و "کلمه۵" و حالت سوم نیز به صورت "کلمه۱ کلمه۲" و "کلمه۳ کلمه۴ کلمه۵" خواهد بود. در اینجا نیز مانند قبل اگر فقط در یکی از حالت‌ها همه زیرعبارات به دسته نشانه تعلق داشتند، آن حالت به عنوان حالت برنده در نظر گرفته می‌شود و همه زیرعبارات آن به عنوان نشانه‌های جداگانه قطعه‌بندی می‌شوند. اگر بیشتر از یک حالت داشتیم که زیرعبارات آنها به دسته نشانه تعلق داشتند، در این صورت حالتی برنده است که میانگین درجه تعلق زیرعبارات آن به دسته نشانه از دیگر حالت‌ها بیشتر باشد. در نهایت اگر هیچ کدام از حالت‌ها شرط ذکرشده را نداشتند باز از طول زیرعبارات کم شده و عبارت اولیه به زیرعبارات کوچک‌تری شکسته می‌شود و روال گفته‌شده ادامه می‌یابد تا جایی که به یکی از حالت‌های مورد قبول برسیم و یا تک‌کلماتی بمانند که هر کدام یک نشانه در نظر گرفته شوند. در انتهای فاز چهارم همه عباراتی که از فاز ۳ به فاز ۴ آمده‌اند قطعه‌بندی شده و در واقع همه کلمات متن ورودی قطعه‌بندی می‌شوند.

٤- استفاده از شبکه‌های عصبی برای دسته‌بندی عبارات فاز ٤

همان طور که گفته شد، در فاز ۴ به دسته‌بندی عبارات می‌پردازیم. برای این منظور از شبکه‌های عصبی که یکی از قوی‌ترین سامانه‌های پردازشی داده‌ها جهت دسته‌بندی می‌باشد، استفاده کرده‌ایم.

برای طراحی و ایجاد یک شبکه عصبی، باید در ابتدا خصیصه‌های مناسبی که می‌توان از مشاهدات استخراج کرد را مشخص کنیم. بعد از تعیین خصیصه‌ها، باید یک ساختار مناسب برای شبکه در نظر گرفت که در آن تعداد لایه‌ها و تعداد نورون‌های موجود در هر لایه مشخص باشد. همچنین باید داده‌های خوب و مناسبی به عنوان مشاهدات جمع‌آوری شود تا از آنها برای آموزش شبکه و به دست آوردن وزن یال[2]های شبکه استفاده گردد. بعد از این مرحله با آموزش‌دادن شبکه و به دست آوردن وزن یال‌ها، می‌توان از وزن‌های حاصل‌شده برای دسته‌بندی عبارات در فاز چهارم الگوریتم پیشنهادی استفاده کرد.

٤-١ تعیین خصیصه‌ها

سه پارامتر به عنوان خصیصه‌های اصلی مورد استفاده قرار گرفتند. تعداد کلمات عبارت، فراوانی تکرار تک‌کلمات عبارت و فراوانی تکرار دوکلمه‌ای‌های عبارت، سه پارامتر مورد نظر هستند.

باید توجه داشت در هر عبارت تعداد کلمات می‌تواند متفاوت باشد، بنابراین در هر عبارت به تعداد کلمات آن عبارت، فراوانی تکرار تک‌کلمات و فراوانی تکرار دوکلمه‌ای‌ها وجود دارد. یعنی ممکن است تعداد ورودی‌های شبکه تغییر کند، این در حالی است که تعداد ورودی‌های شبکه برای همه نمونه‌ها باید ثابت باشد. بنابراین نمی‌توان از فراوانی تکرار تک‌کلمات و دوکلمه‌ای‌ها به صورت مستقیم استفاده کرد و باید نماینده‌ای به جای همه این اعداد در نظر گرفت.

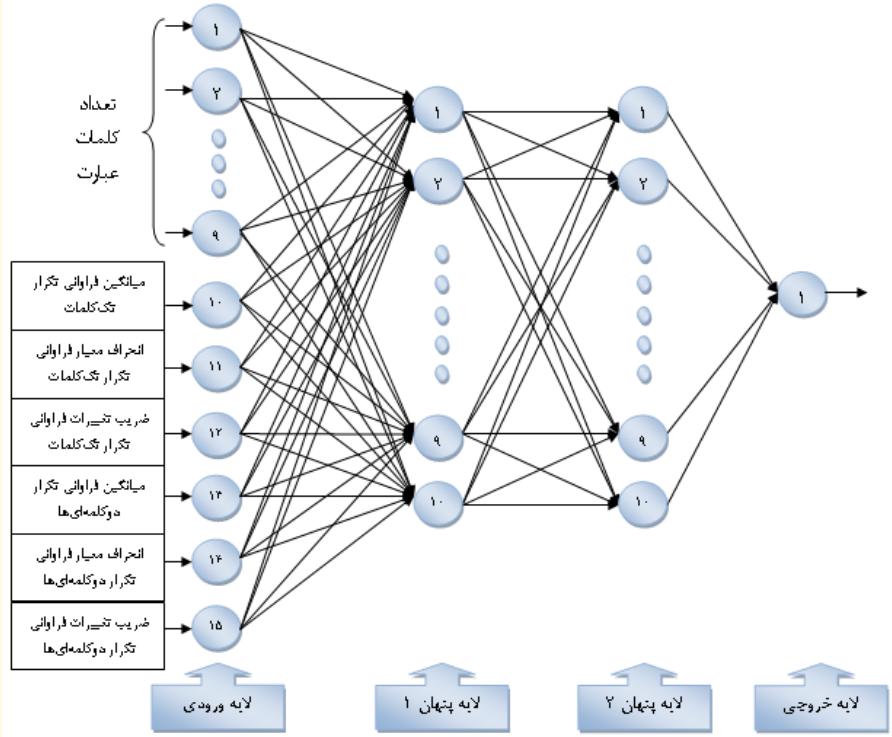
میانگین، انحراف معیار[3] و ضریب تغییرات[4] سه پارامتری هستند که به عنوان نماینده فراوانی تکرار تک‌کلمات و دوکلمه‌ای‌ها در نظر گرفتیم. این پارامترها می‌توانند به خوبی فاصله، پراکندگی و میانگین اعداد را در خود جای داده و به نمایندگی از فراوانی تکرار تک‌کلمات و دوکلمه‌ای‌ها به عنوان ورودی شبکه مورد استفاده قرار گیرند. تعریف ضریب تغییرات در پیوست آمده است.

برای هر یک از مقادیر میانگین، انحراف معیار و ضریب تغییرات یک نود ورودی شبکه در نظر می‌گیریم. با توجه به این که این پارامترها هم برای فراوانی تکرار تک‌کلمات و هم برای فراوانی تکرار دوکلمه‌ای‌ها مورد استفاده قرار می‌گیرند، در مجموع ۶ نود ورودی برای دو خصیصه یادشده در نظر گرفته شد.

تعداد کلمات عبارت عددی صحیح است که می‌تواند بین دو تا ده کلمه باشد. با توجه به این که در شبکه‌های عصبی بهتر است پارامترهای ورودی مقادیر باینری به صورت صفر یا یک دریافت کنند، بنابراین برای خصیصه تعداد کلمات عبارت، نُه نود ورودی جدا در نظر گرفته شد که هر یک از آنها مشخص‌کننده یکی از اعداد بین دو تا ده بوده‌اند و مقدار باینری دریافت می‌کنند.

در کل ۱۵ نود ورودی داریم که همراه با خصیصه‌های تعریف‌شده در جدول ۱ آورده شده‌اند.

1. Feature
2. Link
3. Standard Deviation
4. Coefficient of Variation



شکل ۲: ساختار شبکه عصبی طراحی‌شده.

جدول ۱: خصیصه‌های شبکه عصبی طراحی‌شده.

خصیصه‌های شبکه	نحوه ورود خصیصه‌ها به شبکه	تعداد نودهای ورودی مربوط
تعداد کلمات عبارت	با استفاده از ۹ نود باینری	۹
فراوانی تکرار تک‌کلمات	با استفاده از میانگین، انحراف معیار و ضریب تغییرات	۳
فراوانی تکرار دوکلمه‌ای‌ها	با استفاده از میانگین، انحراف معیار و ضریب تغییرات	۳

در رابطه با تعداد نود استفاده‌شده برای خصیصه تعداد کلمات عبارت، می‌توان به جای استفاده از ۹ نود باینری، از ۴ نود برای ورودی‌های شبکه استفاده کرد اما باید به این نکته توجه شود که در صورت استفاده از ۴ نود، باید داده‌ها کُد شوند. در این صورت علاوه بر انجام یک فرایند اضافی برای کدکردن داده‌ها، ممکن است همبستگی متقابل[1] رخ دهد زیرا برای هر کدام از مقادیر ۱، ۳، ۵، ۷ و ۹ کم‌ارزش‌ترین بیت یک می‌شود. اما در صورت استفاده از ۹ نود باینری همبستگی متقابل رخ نمی‌دهد. همچنین با انتخاب ۹ نود ورودی به صورت باینری، در هر لحظه تنها یک نود از ۹ نود می‌تواند یک باشد و بقیه نودها باید صفر باشند، در این صورت پایداری شبکه و قابلیت اطمینان آن بیشتر شده و شبکه بهتر و سریع‌تر همگرا می‌شود.

٤-۲ تعیین ساختار شبکه

برای ساختار شبکه از یک شبکه MLP با دو لایه پنهان و یک لایه خروجی استفاده شد و لایه‌های اول و دوم شامل ۱۰ نورون پرسپترون بودند. با توجه به این که دو دسته نشانه و غیر نشانه داشتیم، یک نورون برای لایه خروجی در نظر گرفتیم. شبکه عصبی طراحی‌شده به صورت شکل ۲ می‌باشد.

٤-۳ تهیه داده‌های مناسب برای آموزش شبکه

بعد از مشخص‌کردن خصیصه‌ها و طراحی ساختار شبکه، باید داده‌های مناسبی را جهت آموزش شبکه عصبی و به دست آوردن وزن یال‌ها تهیه کنیم. باید سعی شود این مجموعه فاقد داده‌های دورافتاده[2] باشد. این داده‌ها سبب همگرا نشدن آموزش شبکه می‌شود و روند آموزش را از مسیر اصلی خود منحرف کرده و آن را کند می‌کند. با توجه به ساختار شبکه که در شکل ۲ مشخص است، تعداد یال‌های شبکه برابر با ۲۶۰ یال می‌باشد.

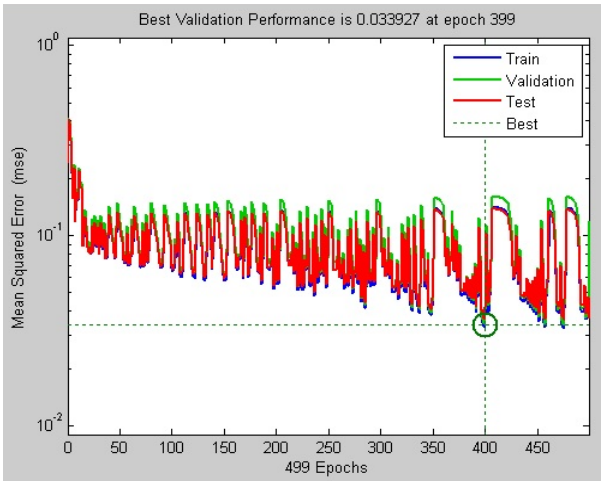
نزدیک به ۳۷۰۰ داده از پیکره بیجن‌خان به صورت دستی توسط نویسندگان مقاله انتخاب شد. داده‌های انتخابی شامل عبارات و کلماتی بودند که به یکی از دو دسته نشانه و غیر نشانه تعلق داشتند و تقریباً سهم هر دو دسته برابر بود. با توجه به این که ۷۰٪ داده‌ها برای آموزش در نظر گرفته می‌شود و ۳۰٪ مابقی برای مراحل آزمون و معتبرسازی[3] کنار گذاشته می‌شود، تقریباً ۱۰ برابر تعداد یال‌ها یعنی حدود ۲۶۰۰ داده برای آموزش شبکه و به دست آوردن وزن‌های یال‌ها در اختیار داشتیم. تقسیم داده‌ها برای مراحل آموزش و آزمون به صورت تصادفی انجام گرفته است.

٤-٤ آموزش شبکه طراحی‌شده

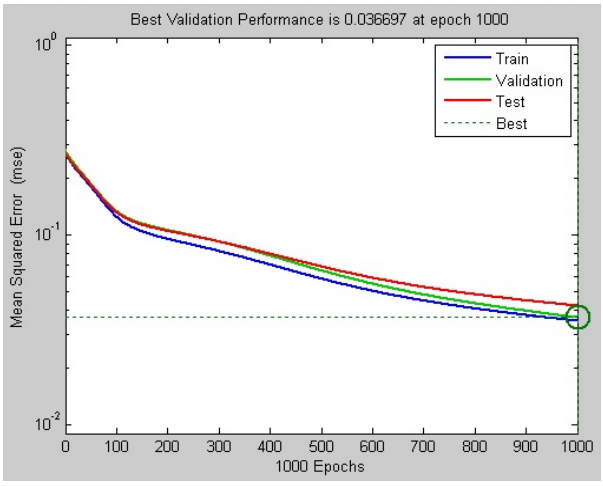
برای آموزش شبکه از روش پس‌انتشار[4] استفاده گردید. شبکه‌های پس‌انتشار یا BP، یک شبکه چندلایه با تابع انتقال غیر خطی می‌باشد و از بردار ورودی و هدف برای دسته‌بندی ورودی‌ها استفاده می‌کند. پس‌انتشار استاندارد یک الگوریتم با شیب نزولی[5] است که در آن وزن‌های شبکه در جهت خلاف تابع کارایی[6] حرکت می‌کنند.

1. Cross Correlation

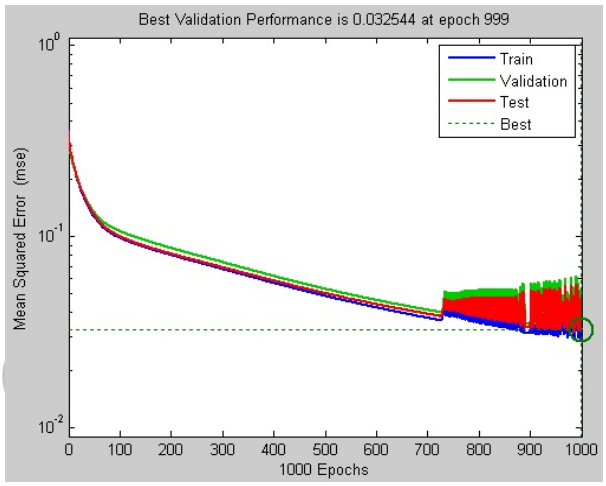
2. Outlier
3. Validation
4. Backpropagation
5. Gradient Descent
6. Performance Function



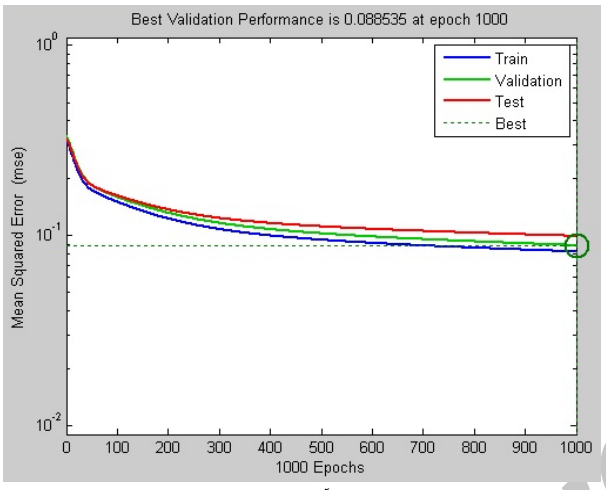
شکل ۳: میانگین مربعات خطای شبکه آموزش دیده با نرخ یادگیری ۰/۹.



شکل ۵: میانگین مربعات خطای شبکه آموزش دیده با نرخ یادگیری ۰/۰۷.



شکل ۴: میانگین مربعات خطای شبکه آموزش دیده با نرخ یادگیری ۰/۱.



شکل ۶: میانگین مربعات خطای شبکه آموزش دیده با نرخ یادگیری ۰/۰۱.

شبکه طراحی‌شده دارای ۱۰ نورون در لایه پنهان ۱ با تابع انتقال tan sigmoid و ۱۰ نورون در لایه پنهان ۲ که آن نیز از تابع انتقال tan sigmoid استفاده می‌کند، است. همچنین این شبکه یک نورون در لایه خروجی با تابع انتقال log sigmoid دارد.

شبکه را با تابع یادگیری TRAINGD[1] در نرم‌افزار MATLAB آموزش دادیم که یک روش آموزش دسته‌ای کاهش شیب[2] می‌باشد. در این روش وزن‌ها و بایاس‌ها در جهت عکس تابع کارایی به روز می‌شوند. در ابتدا وزن‌ها به صورت تصادفی مقداردهی اولیه می‌شوند و سپس با توجه به تابع یادگیری که در اینجا کاهش شیب می‌باشد به اصلاح وزن‌های شبکه می‌پردازیم. در هر تکرار[3]، نرخ یادگیری که در ابتدای امر مقداردهی شده است را کاهش می‌دهیم و با این عمل در ابتدا از اطلاعات به دست آمده در تکرار جاری بهای بیشتری نسبت به تجربیات گذشته داده می‌شود، اما هرچه به تکرارهای آخر نزدیک می‌شویم این قضیه عکس خواهد شد.

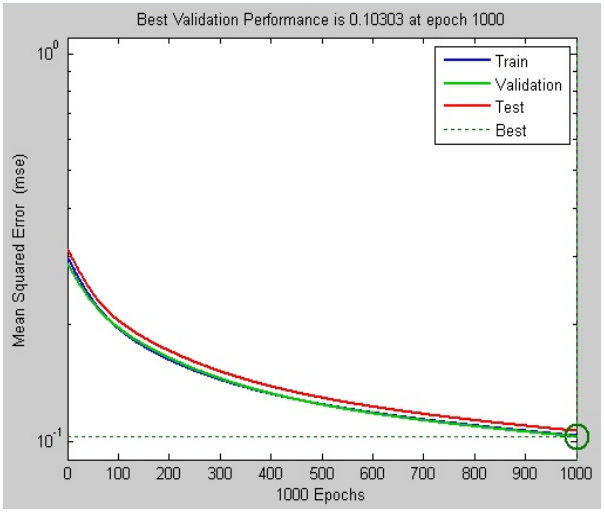
حاصل آموزش شبکه طراحی‌شده با استفاده از تابع TRAINGD شبکه‌ای شد که میانگین مربعات خطا[4]ی آن برای نرخ‌های یادگیری متفاوت در شکل‌های ۳ تا ۶ قابل مشاهده است. در هر یک از این نمودارها، میانگین مربعات خطای مربوط به داده‌های آموزش، معتبرسازی و آزمون آورده شده‌اند.

همان طور که در شکل ۳ ملاحظه می‌شود، نرخ یادگیری نامناسب سبب شده شبکه به خوبی آموزش نبیند. می‌توان علت این رخداد را به این صورت بیان کرد که نرخ یادگیری اولیه برای آموزش این شبکه زیاد است و با توجه به فراوانی داده‌ها در آموزش نوساناتی به وجود می‌آید. هرچه نرخ یادگیری اولیه کمتر می‌شود این نوسانات نیز کمتر می‌شود و آموزش باثبات‌تر می‌گردد. تعداد بالای مشاهدات و کم‌بودن دسته‌ها می‌تواند باعث این امر در حین دسته‌بندی مشاهدات شود. مقدار بیشینه شکست[5] برای نمودارهای بالا ۱۰۰ در نظر گرفته شده بود و همان طور که مشاهده می‌شود، آموزش قبل از رسیدن به تکرار هزارم متوقف می‌گردد. یعنی با استفاده از داده‌های معتبرسازی، انحراف از مسیر آموزش تشخیص داده شده و روند آموزش متوقف گردیده است.

با توجه به مطالب گفته‌شده نرخ یادگیری اولیه را کاهش می‌دهیم تا فرایند آموزش با دقت بیشتری طی شود. بدین ترتیب شیب نمودار با سرعت کمتری تغییر می‌یابد.

در شکل ۴ شبکه در ابتدا خوب آموزش دیده است اما در ادامه برای تکرارهای بالاتر از ۷۰۰ بار، شروع به نوسان می‌کند. علت این پدیده را می‌توان بیش‌برازش[6] دانست که با حفظ‌کردن داده‌های آموزش

1. Training Function
2. Batch Gradient Descent
3. Iteration
4. Mean Square Error
5. Max Fail
6. Overfitting



شکل ۷: میانگین مربعات خطای شبکه آموزش دیده با نرخ یادگیری ۰/۰۷ در حالتی که از ضریب تغییرات به عنوان ورودی‌های شبکه استفاده نشود.

توسط شبکه رخ می‌دهد. به همین جهت است که میانگین مربعات خطای داده‌های آموزش کم می‌شود اما برای داده‌های آزمون و معتبرسازی نوسان وجود دارد. مقدار بیشینه شکست نیز در اینجا برابر ۱۰۰ در نظر گرفته شده که به همین دلیل شبکه تا تکرار هزارم پیشروی می‌کند.

با توجه به نمودارهای رسم‌شده در شکل‌های ۳ تا ۶ می‌توان نتیجه گرفت که بهترین آموزش مربوط به شکل ۵ می‌باشد که با استفاده از نرخ یادگیری ۰/۰۷ به آموزش شبکه پرداخته است. بهترین نتیجه برای میانگین مربعات خطا در شکل مذکور برابر ۰/۰۳۶۷ در تکرار هزارم می‌باشد. مجذور این عدد تقریباً برابر با ۰/۱۹ است، یعنی شبکه با نرخ یادگیری ۰/۰۷ دقتی در حدود ۸۱٪ دارا می‌باشد. با استفاده از وزن‌های به دست آمده از شبکه آموزش‌دیده، می‌توان به انجام قطعه‌بندی در فاز ۴ پرداخت.

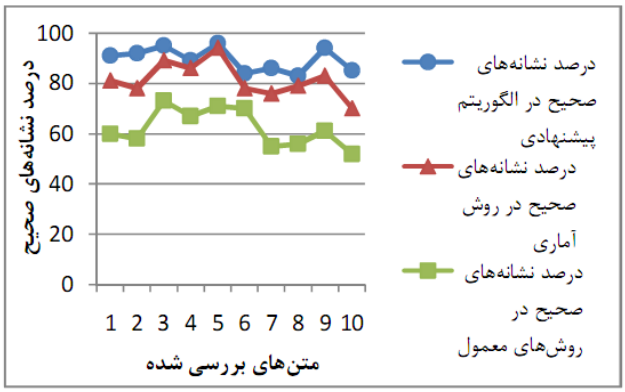
توجه به این نکته ضروری به نظر می‌رسد که بعد از آزمایش‌های گوناگون و بررسی‌های فراوان و با توجه به تعاریف مربوط به شبکه عصبی و این که چه نوع خصیصه‌هایی برای ارائه به شبکه مناسب هستند، خصیصه‌های ذکرشده انتخاب گردیدند. برای مثال هنگامی که از ضریب تغییرات به عنوان ورودی‌های شبکه استفاده نگردد، میانگین مربعات خطای شبکه آموزش‌دیده به صورت شکل ۷ می‌شود.

همان طور که ملاحظه می‌شود دقت این حالت خیلی کمتر از حالتی است که از ضریب تغییرات در ورودی‌های شبکه استفاده شده و نرخ یادگیری ۰/۰۷ می‌باشد (حالتی که در شکل ۵ به نمایش درآمده است).

۵- نتایج پیاده‌سازی

برای پیاده‌سازی از پیکره بیجن‌خان استفاده گردید. این پیکره در آزمایشگاه زبان‌شناسی دانشگاه تهران نگهداری می‌شود که از برخی اخبار روزنامه‌ها و متون معمولی جمع‌آوری شده است و مناسب برای تحقیقات پردازش زبان طبیعی در زبان فارسی می‌باشد. یکی از ویژگی‌های این پیکره این است که اسناد تحت عناوین سیاسی، فرهنگی، اقتصادی و غیره دسته‌بندی شده‌اند. این پیکره شامل حدود ۲/۶ میلیون واژه و ۵۵۰ برچسب می‌باشد که به طور دستی برچسب زده شده است [۱۲].

برای ارزیابی از ده متن مختلف موجود در پیکره بیجن‌خان استفاده شد. این متن‌ها از قسمت‌های مختلفی در پیکره بیجن‌خان استخراج شده‌اند و سعی شده از لحاظ محتوا متنوع باشند، برای مثال گزیده‌ای از اخبار سیاسی، فرهنگی، اجتماعی و ورزشی در این متون آمده‌اند تا با استفاده از



شکل ۸: مقایسه الگوریتم پیشنهادی با روش‌های آماری و معمول.

آنها بتوان به درصد دقیق‌تری از کارایی الگوریتم پی برد. هر یک از این متون مشتمل بر حدوداً ۲۰۰ تا ۳۰۰ کلمه هستند. برای هر یک از متن‌ها با استفاده از سه روش معمول، آماری و روش پیشنهادی درصد نشانه‌های درست به دست آمد. منظور از روش‌های معمول، روش‌هایی است که تنها از فاصله موجود بین کلمات برای قطعه‌بندی استفاده کرده و از الگوریتم خاصی پیروی نمی‌کنند.

روش آماری آورده‌شده نیز قطعه‌بندی را بدون استفاده از شبکه‌های عصبی انجام می‌دهد. این روش نیز شامل چهار فاز است که سه فاز اول آن دقیقاً مشابه سه فاز اول الگوریتم پیشنهادی می‌باشد و در فاز چهارم متفاوت است. به عبارت دیگر بعد از قطعه‌بندی متن به جملات و سپس عبارات و بعد از شکستن عبارات به عبارات کوچک‌تر، در فاز چهارم قطعه‌بندی عبارات با توجه به فاصله فراوانی تکرار دوکلمه‌ای‌های همسایه صورت می‌پذیرد. برای این منظور اگر عبارت حاصل از فاز ۳ شامل یک یا دو کلمه بود، کل عبارت یک نشانه است. برای حالتی که عبارت بیش از دو کلمه داشته باشد، در ابتدا فرض می‌کنیم عبارت از سه کلمه مانند (۲) تشکیل شده است

$$\text{عبارت} = \text{کلمه ۱}\ \ \text{کلمه ۲}\ \ \text{کلمه ۳} \qquad (۲)$$

فراوانی تکرار دوکلمه‌ای‌های "کلمه ۱ کلمه ۲" و "کلمه ۲ کلمه ۳" از حد آستانه بیشتر است و "کلمه ۲" وجه اشتراک این دوکلمه‌ای‌ها می‌باشد. در صورتی که فراوانی تکرار این دوکلمه‌ای‌ها نزدیک به هم باشند، به احتمال زیاد هر سه کلمه یک نشانه‌اند. زیرا این سه کلمه در همسایگی هم به تعداد زیاد و به نسبت‌های نزدیک به هم در پیکره یادگیری ظاهر شده‌اند.

برای به دست آوردن نسبت تکرار می‌توان تعداد تکرارشان را بر هم تقسیم کرد. به صورت (۳)

$$\begin{aligned} &if\ [\mathrm{Freq}(word۱, word۲) > \mathrm{Freq}(word۲, word۳)] \rightarrow \\ &\quad [\mathrm{Freq}(word۱, word۲)/\mathrm{Freq}(word۲, word۳)] < \\ &Threshold \\ &else \rightarrow \\ &\quad [\mathrm{Freq}(word۲, word۳)/\mathrm{Freq}(word۱, word۲)] < \\ &Threshold \end{aligned} \qquad (۳)$$

حاصل این تقسیم را با حد آستانه دیگری مقایسه کرده و در صورتی که از حد آستانه کوچک‌تر باشد، این سه کلمه، یک نشانه در نظر گرفته می‌شوند. برای عبارات بیش از سه کلمه، ابتدا برای سه کلمه اول همین روال صورت می‌پذیرد، سپس برای کلمات دوم تا چهارم، بعد کلمات سوم تا پنجم و به همین ترتیب به صورت زنجیره‌وار این فرایند تکرار می‌شود. در هر زنجیره تا جایی که (۳) برقرار باشد، کلمات یک نشانه را تشکیل

می‌دهند. در مجموع روش آماری با استفاده از فرکانس تکرار دوکلمه‌ای‌ها به قطعه‌بندی پرداخته است و از الگوریتم‌های یادگیری استفاده نمی‌کند.

شکل ۸ مقایسه بین درصد نشانه‌هایی که در ده متن متفاوت با استفاده از الگوریتم پیشنهادی، روش آماری و روش‌های معمول، به درستی تشخیص داده شده‌اند را نشان می‌دهد.

قطعه‌بندی با استفاده از روش‌های معمول می‌تواند به طور میانگین ۶۲٫۳٪ از کلمات و عبارات ده متن انتخابی را به درستی قطعه‌بندی کند که با اعمال روش آماری، میانگین دقت[1] سیستم به ۸۱٫۴٪ افزایش یافت. در این مبحث دقت عبارت است از، نسبت نشانه‌هایی که به درستی تشخیص داده شده‌اند به کل نشانه‌هایی که توسط سیستم قطعه‌بندی برگردانده می‌شوند، که به صورت (۴) نشان داده می‌شود

$$\text{Precision} = \frac{\text{Correct Tokens}}{\text{Correct Tokens} + \text{Incorrect Tokens}} \qquad (۴)$$

همان طور که در شکل ۸ مشاهده می‌کنید، استفاده از الگوریتم پیشنهادی در این مقاله موجب بهبود دقت قطعه‌بندی شده و از دو روش دیگر بهتر عمل می‌کند. این روش به طور میانگین ۸۹٫۶٪ از کلمات ده متن انتخابی را به درستی قطعه‌بندی می‌کند، یعنی با استفاده از این روش به طور میانگین از بین ۱۰۰ کلمه موجود در متن ۸۹٫۶ از کلمات در نشانه‌های صحیح از نظر کارایی در موتور جستجو قرار گرفته‌اند. می‌توان دلیل این افزایش دقت را اضافه‌کردن پارامترهای بیشتر برای انجام قطعه‌بندی و استفاده از یک روش یادگیری نظیر شبکه عصبی و همچنین استفاده از یک پیکره مناسب که به دقت نشانه‌گذاری شده بود، بیان کرد.

طبق نتایج پیاده‌سازی، شبکه عصبی برای دسته‌بندی عبارات به دو دسته نشانه و غیر نشانه به خوبی عمل می‌کند و در تشخیص کلماتی که به هم مرتبط هستند، مانند صفت و موصوف و مضاف و مضاف‌الیه عملکرد بهتری نسبت به دو روش دیگر دارد. این امر موجب بهبود کارایی موتور جستجو نیز می‌شود.

درصدهای ارائه‌شده در شکل ۸ با توجه به نشانه‌های مطلوب در موتور جستجو به دست آمده‌اند. هدف از انجام این پروژه، طراحی یک سیستم مؤثر برای قطعه‌بندی کلمات و عبارات به جهت استفاده در موتورهای جستجو بود. در موتورهای جستجو مطلوب آن است کلماتی که به یک مفهوم اشاره می‌کنند، یک نشانه در نظر گرفته شوند. برای مثال عبارت "دانشگاه یزد" یک منظور را می‌رساند که در سیستم ارائه‌شده این عبارت به عنوان یک نشانه شناخته می‌شود.

٦- نتیجه‌گیری و کارهای آینده

هدف از ارائه این مقاله، پیشنهاد سیستمی برای قطعه‌بندی کلمات و عبارات متون و نوشته‌ها جهت استفاده در موتورهای جستجو بود. برای این منظور با استفاده از روش‌های آماری و شبکه‌های عصبی، سعی بر طراحی سیستمی شد که بتواند اهداف ذکرشده را تحقق بخشد. روش ارائه‌شده با استفاده از تعداد کلمات عبارت، احتمال رخداد تک‌کلمات و دوکلمه‌ای‌های موجود در پیکره در چهار فاز انجام گرفت.

در هر فاز تعدادی از نشانه‌ها تشخیص داده شد و نسبت به فازهای قبلی عبارات کوچک‌تری به دست آمد. در پایان فاز آخر متن ورودی، به کلمات و عبارات بامفهوم قطعه‌بندی شد. نتایج پیاده‌سازی نشان داد که روش مذکور، کارایی روش‌های معمول را از ۶۲٫۳٪ به ۸۹٫۶٪ افزایش داد.

از دیگر مزایای این روش مستقل از زبان بودن آن است که می‌تواند برای زبان‌های دیگر از جمله انگلیسی به کار رود. در صورت وجود پیکره متنی مناسب در زبان‌های دیگر می‌توان به عنوان کارهای آینده از این روش برای قطعه‌بندی کلمات و عبارات در آن زبان‌ها نیز استفاده کرد.

از دیگر پیشنهادها می‌توان به استفاده از وب به عنوان پیکره به جای پیکره بیجن‌خان برای مراحل آموزش و آزمون اشاره کرد. همچنین می‌توان از روش ارائه‌شده در موتور جستجوی پارسی‌جو برای مراحل نمایه‌سازی و قطعه‌بندی کلمات پرس و جو استفاده کرد [۱۳]. این موتور جستجو در دانشگاه یزد طراحی و پیاده‌سازی شده است.

پیوست

ضریب تغییرات: در نظریه آمار و احتمال ضریب تغییرات یک معیار است که برای اندازه‌گیری توزیع داده‌های آماری به کار می‌رود و از تقسیم انحراف معیار بر میانگین طبق (پ-۱) قابل محاسبه است

$$C_v = \frac{\sigma}{\mu} \qquad (پ-۱)$$

مراجع

[۱] م. محمدی جنقرا و م. آنالویی، "استخراج کلمات کلیدی اسناد فارسی،" *سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران*، جزیره کیش، اسفند ۱۳۸۶.

[2] B. Habert, et al., "Towards tokenization evaluation," in *Proc. 1st Int.l Conf. on Language Resources and Evaluation, LREC*, vol. 1, pp. 427-431, Spain, May 1998.

[۳] س. کیانی و م. شمس‌فرد، "تعیین مرز کلمات و عبارات در متون نوشتاری فارسی،" *چهاردهمین کنفرانس سالانه انجمن کامپیوتر ایران*، تهران، اسفند ۱۳۸۷.

[۴] س. م. غفوری، س. راحتی، م. ر. پهلوان‌نژاد و ع. عظیمی‌زاده، "نرمال‌ساز متون فارسی،" *پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران*، تهران، اسفند ۱۳۸۸.

[5] M. Shamsfard, S. Kiani, and Y. Shahedi, "Step - 1: standard text preparation for Persian Language," in Proc. of the 3rd Workshop on ComputationalApproaches to Arabic Script-based Languages MTSummit XII, Ottawa, Canada, 2009.

[6] T. Chung and D. Gildea, "Unsupervised tokenization for machine translation," in *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, vol. 2, pp. 718-726, Singapore, Aug. 2009.

[7] O. Frunza, "A trainable tokenizer, solution for multilingual texts and compound expression tokenization," in *Proc. of the 6th Int. Conf. on Language Resources and Evaluation, LREC'08*, Marrakech, May 2008.

[8] J. Grana, M. A. Alonso, and M. Vilares, "A common solution for tokenization and part - of - speech tagging," in *Proc. of the 5th Int. Conf. on Text, Speech, and Dialogue*, *TSD'02*, vol. 1, pp. 3-11, London, Sep. 2002.

[9] T. V. Nguyen, H. K. Tran, T. T. Nguyen, and H. Nguyen, "Word segmentation for vietnamese text categorization: an online corpus approach," in *Proc. 4th IEEE Int. Conf. in Computer Science, Research, Innovation and Vision of the Future, RIVF'06,* Hochiminh, Vietnam, Feb. 2006.

[10] V. Tesprasit, P. Charenpornsawat, and V. Sornlertlamvanich, "Learning phrase break detection in thai text - to - speech," in *Proc. of 8th European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003.

[11] S. Kiani, T. Akhavan, and M. Shamsfard, "Developing a persian chunker using a hybrid approach," in *Proc. of IEEE Int. Multiconf. on. Computer Science and Information Technology, IMCSIT'09,* vol. 1, pp. 227-234, Oct. 2009.

[12] BijanKhan Corpus, http://ece.ut.ac.ir/dbrg/Bijankhan/, 2012.

[13] Parsijoo Search Engine, http://www.parsijoo.ir, 2012.

1. Precision

محمدمهدی میردامادی در سال ۱۳۸۹ مدرک کارشناسی مهندسی کامپیوتر گرایش نرم‌افزار خود را از دانشگاه آزاد اسلامی واحد اصفهان و در سال ۱۳۹۱ مدرک کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی خود را از دانشگاه یزد دریافت نمود. موضوع پایان‌نامه‌ی کارشناسی ارشد نامبرده "نشانه‌گذاری آماری متون فارسی مبتنی بر محتوای وب" بوده و ایشان در حال حاضر از پژوهشگران مرکز تحقیقات مخابرات می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: بازیابی هوشمند اطلاعات، پردازش زبان‌های طبیعی و یادگیری ماشین.

علی محمد زارع بیدکی در سال ۱۳۷۸ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه صنعتی اصفهان و مقاطع کارشناسی ارشد و دکتری مهندسی کامپیوتر خود را به‌ترتیب در سال‌های ۱۳۸۱ و ۱۳۸۸ از دانشگاه تهران به پایان رسانده است. دکتر زارع بیدکی از سال ۱۳۸۸ در دانشکده مهندسی برق و کامپیوتر دانشگاه یزد مشغول به فعالیت گردید و اینک نیز عضو هیأت علمی این دانشگاه می‌باشد. زمینه‌های علمی مورد علاقه نام‌برده شامل موضوعاتی مانند موتورهای جستجو، رتبه-بندی صفحات وب، خزش و پردازش داده‌های با حجم بالا می‌باشد.

مهدی رضائیان در سال ۱۳۷۲ مدرک کارشناسی مهندسی برق – گرایش الکترونیک خود را از دانشگاه صنعتی خواجه نصیرالدین طوسی و در سال ۱۳۷۶ مدرک کارشناسی ارشد مهندسی برق گرایش بیوالکتریک خود را از دانشگاه تهران دریافت نمود. از سال‌های ۱۳۷۸ الی ۱۳۸۲ به عنوان عضو هیات علمی دانشکده فنی دانشگاه تهران به فعالیت مشغول بود و پس از آن برای ادامه تحصیل در دوره دکتری به دانشگاه ایالتی زوریخ سوییس وارد گردید. هم‌اکنون استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه یزد می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: ماشین بینایی و فتوگرامتری، محاسبات نرم و کاربردهای آن.