



معرفی یک الگوریتم ریشه‌یابی و لم‌یابی مبتنی بر قانون برای زبان فارسی

زینب رحیمی^۱، یاسر شکفته^۲

^۱ گروه پردازش صوت زبان طبیعی، پژوهشگاه توسعه فناوری های پیشرفته خواجه نصیرالدین طوسی، تهران،
rahimi-z@rcdat.ir

^۲ گروه پردازش صوت زبان طبیعی، پژوهشگاه توسعه فناوری های پیشرفته خواجه نصیرالدین طوسی، تهران،
shekofteh@rcdat.ir

چکیده

با توجه به ذات زایا و اشتقاق پذیر زبان فارسی و همپنین نیاز برنامه های کاربردی مختلف مرتبط با پردازش زبان طبیعی و بازیابی اطلاعات، ریشه‌یابی و لم‌یابی از مسائل مهم پیش پردازشی در پردازش زبان طبیعی فارسی به شمار می رود. در این راستا در این مقاله یک الگوریتم مناسب برای یافتن خودکار ریشه و لمای کلمات پیشنهاد شده است. این الگوریتم و ابزار پیاده سازی شده بر اساس آن، دارای چند حالت برای ریشه‌یابی و لم‌یابی انواع مختلف کلمات است که با روش مبتنی بر قانون و با استفاده از چندین منبع زبانی از جمله فهرستی از افعال زبان فارسی، جمع مکسر، واژگان زایای زبان فارسی و ... طراحی شده است. روال کلی انجام کار به این صورت است که ابتدا بررسی می‌شود که کلمه باید ریشه‌یابی شود یا خیر و در صورت لزوم الگوریتم اصلی اعمال می‌شود. برای لم‌یابی ابتدا برچسب اجزای کلام برای هر کلمه مشمول ریشه‌یابی تعیین می‌گردد و سپس اعمال قوانین صورت می‌گیرد. این امکان در ریشه‌یاب قرار داده شده که به تفکیک آرگومان، فعل‌ها، اسامی و صفت‌ها به تنهایی ریشه‌یابی شده و یا هر ۳ مورد در متن ریشه‌یابی شوند. همچنین با توجه به بار پردازشی برچسب زن اجزای کلام و زمان‌بری روال، یک مد ریشه‌یابی سبک نیز در برنامه لحاظ شده است که در آن فقط با توجه به شکل ظاهری کلمات، قوانین تعیین شده و ریشه‌یابی انجام می‌گیرد. نکته مورد توجه، جامعیت در قوانین و استثنائات مورد پوشش و استفاده از منابع متنی و پیش پردازشی دقیق در الگوریتم پیشنهادی است. نتایج ارزیابی نشان دهنده عملکرد مناسب سیستم پیشنهادی در هر دو حالت ریشه‌یابی و لم‌یابی است.

کلمات کلیدی

ریشه‌یابی، لم‌یابی، برچسب اجزای کلام، واژگان زایا، پردازش زبان طبیعی.

۱- مقدمه

ریشه‌یابی متفاوت است. در این زمینه اغلب با دو واژه ریشه‌یابی و لم‌یابی روبرو هستیم. مطابق تعریف [11] در ریشه‌یابی عموماً وندها جدا می‌شوند ولی در لم‌یابی با تحلیل مورفولوژی و بررسی کلمات زمینه، بن کلمه بازگردانده می‌شود. انتخاب از میان این دو وابسته به شرایط مورد نیاز در کاربردهای مختلف مثل عمق ریشه‌یابی مورد نیاز در برابر پیچیدگی زمانی الگوریتم است.

از یک نقطه نظر روش های ریشه‌یابی به ۳ دسته روش های ساختاری، روش های مبتنی بر جدول و روش های آماری تقسیم بندی می‌شوند. به صورت کلی روش های ساختاری وابسته به تحلیل ساخت واژی بوده، مبتنی بر قاعده عمل می‌کنند و با استفاده از مجموعه قانون های تعریف شده ریشه‌یابی را انجام می‌دهند. در اینگونه

ریشه‌یابی در لغت به معنای حذف پسوندها، پیشوندها و میانوندهای کلمه و به‌دست آوردن ریشه کلمه است. اگرچه در هر زبان، واژه‌ها با توجه به نقش معنایی و نحوی خود در جمله به شکل‌های متفاوتی ظاهر می‌شوند، اما با توجه به این که آن‌ها از یک ریشه مشتق شده‌اند، از نظر معنایی به هم نزدیک هستند. پس در واقع ریشه‌یابی واژه‌ها نه به معنای زبان شناسی آن بلکه به معنای دسته بندی کلمات در گروه های معنایی یکسان تعریف می‌شود. بنابراین ریشه‌یابی مسئله ای است که در بسیاری از زمینه های پردازش زبان طبیعی مورد نیاز می‌باشد و در بسیاری از کاربردها، نیاز داریم تا همه مشتقات یک واژه را به ریشه آن، که همان شکل ساده واژه می‌باشد، تبدیل نماییم و البته باید توجه داشت که نیاز کاربردهای مختلف برای نحوه

وندهای احتمالی کلمه حذف شده و مجدداً جستجو صورت می‌گیرد. ایراد اینگونه روش‌ها نیاز به فضای ذخیره سازی و به روز رسانی است.

[6] یک روش ریشه‌یابی متنی بر گراف و مدل آماری را معرفی می‌کند. هر کلمه در روش پیشنهادی به صورت دو بخش ریشه ووند و هر زیربرشته یک گره از گراف در نظر گرفته می‌شود. یال‌های گراف نشان‌دهنده نحوه ترکیب‌ها هستند. همچنین با استفاده از یک مجموعه داده و شمردن تعداد تکرار در ترکیب‌های مختلف یال‌ها وزندهی می‌شوند. هدف در این مقاله تخمین احتمال برای همه زیربرشته‌های ممکن و تحلیل آماری برای ریشه‌یابی است. ابزار ارائه شده در [7] با در نظر گرفتن صرف افعال و چند قانون مورفولوژیکی اشتقاق عمل می‌کند. همچنین در این سیستم از برچسب اجزای کلام (pos) کلمه و ریشه، FSA، دیتابیس‌هایی شامل کلمات و برچسب‌های pos آنها، ساختاری برای ذخیره سازی ریشه و برچسب‌های مورد نظر آن، ۲ لیست پیشوند و پسوند (که باید حذف شوند) و چند قانون برای تحلیل مورفولوژیک استفاده می‌شود.

در [8] یک جدول ریشه و لیستی از پسوندها (برای حذف) استفاده می‌شود. همچنین به سادگی از لیستی از کلمات جمع مکسر استفاده می‌گردد.

در [9] یک تحلیلگر مورفولوژیکی ۲ مرحله‌ای با استفاده از ابزار FS زیراکس معرفی شده است. به دلیل وجود مسائلی در مرزبندی مثل فاصله و نیم فاصله و صرف افعال پیچیده، در این مقاله چالش‌های مختلف سیستم ریشه‌یابی فارسی بیان شده است. این چالش‌ها به دو دسته مورفولوژی غیرفعلی مثل فاصله افتادن بین بخش‌های کلمات، توکن‌های پیچیده (مثل انکلیتیرینهایشان)، قوانین فوتنیکی و دسته مورفولوژی فعلی مانند مسائل مربوط به بن‌های ماضی و مضارع و وابستگی‌های طولانی تقسیم شده است.

روش پیشنهادی [10] مبتنی بر قانون و یک روش از پایین به بالاست. به این صورت که ابتدا تلاش می‌شود که زیربرشته‌های کلمه استخراج شوند (تشخیص همه زیربرشته‌های مجاز که هسته نامیده می‌شوند و تشخیص مرفه‌ها و خوشه بندی آنها). بعد سعی می‌کند که مطابق قوانین با چسباندن هسته به عناصر دیگر، کلمات مجاز ایجاد کند. هر هسته‌ای که حداقل یک تولید کلمه مجاز داشته باشد ریشه تشخیص داده می‌شود. در مرحله بعد حروفی که احیاناً اضافه هستند (مانند گ در ستارگان) بر طبق یکسری قانون تشخیص داده می‌شوند.

در ابزار معرفی شده در [11]، هم ریشه‌یابی و هم لم‌یابی انجام می‌شود. در الگوریتم ارائه شده برای بخش ریشه‌یابی از یک لیست حاوی کلمات، برچسب pos آنها و تعداد تکرار آنها و همچنین یک لیست کلمات جمع مکسر استفاده شده است. بدین صورت که ابتدا کلمات در لیست مکسرها جستجو می‌شود، اگر بود که به سادگی ریشه آن برگردانده می‌شود و در غیر اینصورت وندهای احتمالی آن حذف می‌شود. کلمه حاصل اگر در لیست واژگان زبانی فارسی [15] موجود بود بررسی می‌گردد که با توجه به برچسب اجزای کلام کلمه آیا وند حذف شده معتبر است یا خیر. در الگوریتم لم‌یابی نیز ابتدا همه ریشه‌های ممکن بدست آورده شده و بعد از بررسی برچسب اجزای کلام کلمه انتخاب مناسب صورت می‌گیرد.

در این الگوریتم [12] از $regular\ expression$ ها برای جداسازی مرفه‌ها از ریشه استفاده می‌شود. ابتدا در جدول درهم سازی شده جستجو می‌شود، در صورت عدم یافته شدن با استفاده از عبارات منظم، جداسازی وندها از کلمه صورت می‌گیرد.

در این الگوریتم [13] از ۳۳ قانون، چند مکاشفه، بررسی استثنائات و حذف وندها استفاده می‌شود.

در [14] از یک دیکشنری و یک لیست وندها استفاده می‌شود. ابتدا با استفاده از شباهت ساختاری بین ریشه‌های کلمات، کلمات مشابه یافت می‌شوند (بر اساس تعداد و ترتیب حروف مثل کتابهایمان: تاب، کتاب و ایمان). سپس بررسی درستی وندها و لیست ریشه‌های طولانی ممکن صورت می‌گیرد و با حذف طولانی‌ترین وند ریشه برگردانده می‌شود.

۳- سیستم پیشنهادی

همانطور که پیش‌تر گفته شد، در سیستم پیشنهادی هم ریشه‌یابی انجام می‌شود و هم لم‌یابی. در هر دو این موارد، روال کلی انجام کار به این صورت است که ابتدا بررسی می‌شود که کلمه باید ریشه‌یابی شود یا خیر و در صورت لزوم الگوریتم ریشه‌یابی انجام می‌شود. یک کلمه ریشه‌یابی نمی‌شود اگر دارای کارکنه‌های غیرفارسی بوده، دارای عدد بوده، مخفف بوده و یا جزء اسامی خاص باشد. در حالت لم‌یابی ابتدا برچسب اجزای کلام برای هر کلمه مشمول ریشه‌یابی تعیین می‌گردد، چون برخی کلمات مانند قیود یا حروف اضافه نیاز به ریشه‌یابی ندارند. پس این امکان در لم‌یابی قرار داده شده که به تفکیک آرگومان، فعل‌ها، اسامی و صفتها به تنهایی ریشه‌یابی شده و یا هر ۳ مورد در متن ریشه‌یابی شوند. همچنین با توجه به بار پردازشی برچسب زن اجزای کلام و زمان بری روال، یک مد ریشه‌یابی سبک در برنامه لحاظ

روش‌ها با وجود قابلیت اعمال قانون به حجم وسیعی از کلمات فارسی، استثنائات بوجود آمده ممکن است ایجاد مشکل نمایند. در روش‌های مبتنی بر جدول، یک جدول برای کلمات شناخته شده و ریشه تعیین شده آنها در نظر گرفته می‌شود و برای هر کلمه با رجوع به جدول به آسانی ریشه برگردانده می‌شود. در صورت عدم وجود کلمه در جدول، ریشه آن با روش‌های معمول دیگر استخراج شده و به جدول افزوده می‌شود. ایراد این دسته روش‌ها حجم بالای دسترسی و ذخیره سازی و همچنین نیاز به روزرسانی است. در روش‌های آماری سعی می‌شود با استفاده از روش‌هایی از قبیل بررسی تعداد رخداد ترکیب ریشه‌های مختلف با وندهای مختلف، احتمال وقوع ریشه محاسبه شود و با استفاده از روش‌های تحلیل آماری ریشه کلمه استخراج گردد. در واقع ریشه‌یابی آماری بر اساس آمارهای یک پیکره متنی و قواعدی مرتبط با ساختار لغت انجام می‌شود.

در مقاله ارائه شده الگوریتمی معرفی شده که هر دو عملیات ریشه‌یابی و لم‌یابی را انجام می‌دهد و در دسته روش‌های ساختاری قرار می‌گیرد. در این روش سعی شده جامعیت قوانین و استثنائات مد نظر قرار گرفته، از منابع مختلف و ابزارهای پیش‌پردازشی مناسب و با دقت بالا استفاده شود و با اعمال تنظیماتی در حالت‌های مختلف ریشه‌یابی سبک و بن‌یابی برای مقوله‌های مختلف نحوی جداگانه یا باهم قابل استفاده باشد.

در ادامه مقاله نخست در بخش ۲ مروری بر کارهای انجام شده در این زمینه خواهیم داشت، در بخش ۳ به معرفی روش پیشنهادی و جزئیات آن خواهیم پرداخت و در انتها ارزیابی و نتیجه‌گیری مقاله آورده خواهد شد.

۲- مروری بر کارهای پیشین

گفتیم که در پردازش زبانهای طبیعی و سیستم‌های بازیابی اطلاعات، عموماً مواردی وجود دارد که باید ریشه واژگان به صورت خودکار استخراج گردد. به همین منظور ریشه‌یاب‌های خودکار متنوعی برای زبانهای مختلف ساخته شده است. برای زبان انگلیسی روش‌های متعددی معرفی شده است که در میان روش‌های پایه مطرح شده برای زبان انگلیسی، می‌توان به الگوریتم‌های پورتر و کراوتر اشاره کرد. الگوریتم پورتر از ۵ مرحله تشکیل شده است و در طی این مراحل قواعدی بر روی کلمه اعمال می‌شود و بزرگترین پسوند موجود حذف می‌گردد. این الگوریتم سریع و ساده است، اما توجهی به پیشوندها نمی‌کند. همچنین الگوریتم کراوتر از روشهای ریخت‌شناسی و از یک فرهنگ لغت برای آزمون ریشه‌های یافت شده استفاده می‌کند. این الگوریتم پسوند و پیشوند کلمات را بررسی می‌کند و در ماشینهای مترجم و برای زبانهای که ساخت کلمات در آنها قانونمند است، کارائی خوبی را نشان داده است.

در زمینه ریشه‌یابی کلمات فارسی نیز روش‌های متعددی معرفی شده است که در ادامه آورده شده‌اند که عموماً ارزیابی جامعی برایشان انجام نشده بود. این ارزیابی‌ها غالباً بر روی تعداد بسیار محدود کلمات و یا در کاربرد خاص انجام شده بود.

اولین ریشه‌یاب فارسی [1] در سال ۲۰۰۲ ارائه شده است که در این روش ابتدا در یک دیکشنری جستجو انجام می‌شود و اگر خود کلمه ریشه تشخیص داده نشد با استفاده از قوانین، حذف وندها و بررسی مصادر و جمع‌های مکسر ریشه‌یابی صورت می‌گیرد. روش ارائه شده در [2] یکی از مشهورترین روش‌های مطرح شده برای زبان فارسی است. این الگوریتم شباهت زیادی به الگوریتم پورتر در انگلیسی دارد که بر مبنای ریخت‌شناسی است. هر دو ریشه‌یاب پسوندهای خاصی را جستجو می‌کنند و مراحل مختلفی را بر طبق لیست قوانین پسوندی پشته‌گذاری شده طی می‌کنند. با این حال تفاوت‌های مهمی بین این دو الگوریتم وجود دارد.

در [3] یک الگوریتم ریشه‌یابی بر پایه حذف یا افزودن وندها، استفاده از یک پایگاه داده برای ذخیره سازی استثنائات برای کاهش نرخ خطا بیان شده است و ادعا شده که نرخ خطا و سرعت ریشه‌یابی بهبود یافته است. در این الگوریتم اولین گام یافتن زیربرشته‌های نهایی (ترمیمال) کلمه ورودی است که اگر در لیست وندها قرار داشت، حذف شود.

در [4] از دانش زبانی و الگوریتم‌های استاندارد قوانین ماشینی استفاده شده و همچنین پسوندهای جمع و استثنائات افعال مرکب در نظر گرفته شده‌اند (۱۰ پسوند و ۲۰۰۰ استثنا اسامی مرکب). این سیستم با جداسازی کلمات مفرد و جمع، ایجاد تمایز بین پسوندها و حذف آنها و در نظر گرفتن استثنائات و جمع‌های مکسر عمل می‌کند.

در [5] الگوریتم کراوتر برای زبان فارسی گسترش یافته است. این الگوریتم در دسته روش‌های ریشه‌یابی مبتنی بر جدول قرار گرفته و روال کلی آن به این صورت است که پس از ساخت جدول کلمات و ریشه‌های آنها، با ورود هر کلمه در جدول یک فرایند جستجو صورت می‌گیرد، اگر موجود بود که به سادگی ریشه کلمه برگردانده می‌شود و در غیر این صورت،

لیست شناسه های ماضی و مضارع، لیست کلیتیک ها، لیست اسامی خاص، لیست افعال و بن های ماضی و مضارع آنها، لیست جمع های مکسر و لیست استثنائات (شامل مواردی مانند ستارگان) استفاده شده است.

۳-۳- مد لم یایی

در این بخش الگوریتم ریشه یابی کلمات با توجه به برچسب اجزای کلام آنها یا همان لم یایی معرفی می گردد. همانطور که ذکر شد، از برچسب زن ۱۰۰ برچسبی استفاده شده است که اطلاعات مفیدی برای استخراج ریشه ارائه می کند. لم یایی برای ۴ برچسب فعل، اسم، صفت و مصدر طراحی شده است که می توان هریک را جداگانه و یا همه را باهم در یک متن ریشه یابی نمود. برای افعال، برچسب اجزای کلام اطلاعاتی مانند منفی بودن، اسنادی بودن، کمکی بودن، زمان و داشتن کلیتیک را ارائه می کند. پس برای افعال پس از حذف علامت نفی و کلیتیک ها، با توجه به برچسب POS نوع فعل تعیین شده و با توجه به برچسب، پیشوند و پسوند های آن نوع فعل جدا شده و در نهایت بن مضارع آن برگردانده می شود.

برچسب اجزای کلام برای اسامی مواردی چون جمع یا مفرد بودن، داشتن کلیتیک، اسم خاص بودن را ارائه می کند. در مورد اسامی و صفات نیز به تفکیک مجموعه ای از شناسه ها (مان، تان و ...) و وندها (ات، ها، هایی، ان، ... برای اسم و با، بی، تر، مند، .. برای صفات) در نظر گرفته شده و با بررسی حضور کلمه پس از حذف آن وند در واژه نامه برای حذف یا عدم حذف وند تصمیم گیری می شود. همچنین مصدرها و استثنائاتی مانند جمع های مکسر و حروف میانجی (مانند گان، یشان و ...) نیز بررسی می شود.

۳-۴- مد ریشه یابی سبک

در این مد الگوریتم ریشه یابی ساده ای بر روی کلماتی که باید روی آنها ریشه یابی صورت گیرد (همه به جز موارد استثنا ذکر شده در بخش ۳) اعمال می گردد. به صورت کلی در کلماتی که در دیکشنری نبوده و به کلیتیک یا وند ختم می شوند، وند یا کلیتیک حذف شده و بررسی می گردد که آیا کلمه جدید در دیکشنری هست یا نه. اگر باشد کلمه جدید به عنوان ریشه برگردانده می شود. برای افعال (که معمولاً در قاعده بالا نمی گنجد) پس از حذف کلیتیک و شناسه و پیشوند، در لیست بن ها جستجو می شود و در صورت وجود، ریشه برگردانده می شود. شکل (۱) فلوجارت روش لم یایی را نشان می دهد.

شده است که در آن فقط با توجه به ریخت شناسی کلمات قوانین تعیین شده و ریشه یابی انجام می گیرد. الگوریتم لم یایی، برای هر ۳ نوع کلمات اسم، فعل یا صفت متفاوت است. اما در مد ریشه یابی سبک روالها نسبتاً مشابه است.

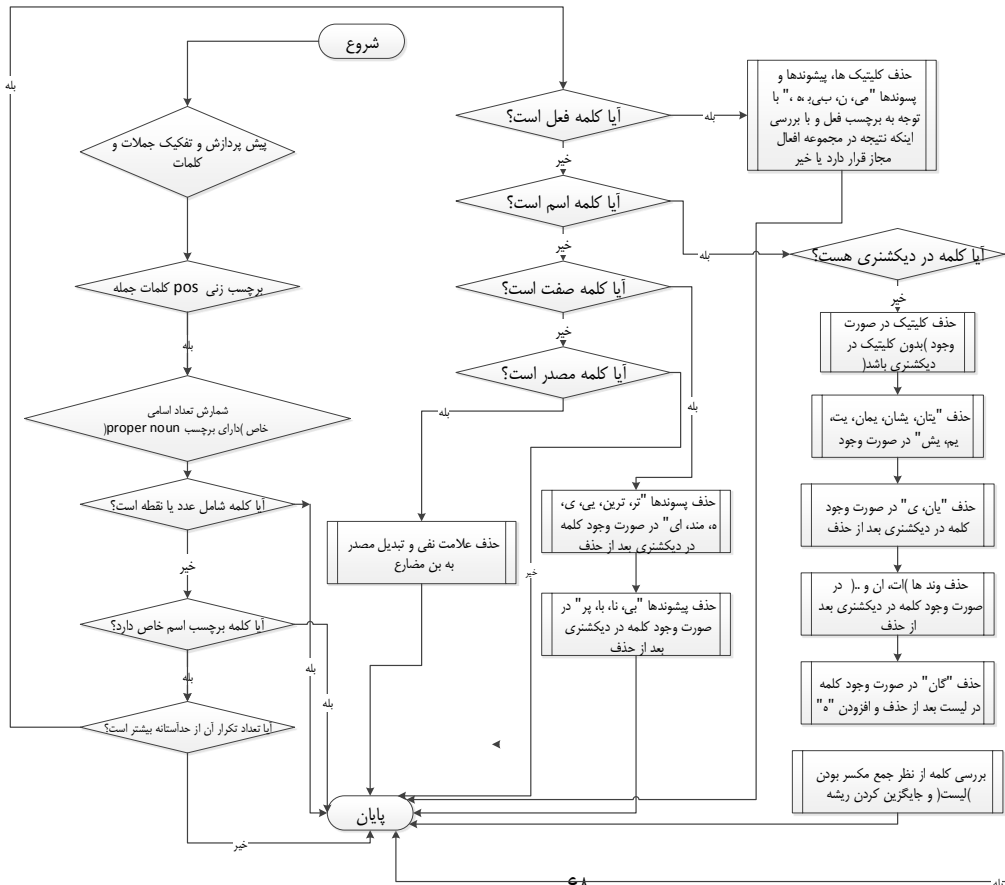
۳-۱- پیش پردازش

در کلیه پردازش های مربوط به زبان طبیعی نیاز است تا پیش از فرآیند اصلی پیش پردازش هایی بر روی متون اعمال گردد. در الگوریتم ارائه شده بخش پیش پردازش شامل مراحل نرمال سازی، جداسازی مرز جملات و کلمات و در صورت نیاز غلط یابی است. همچنین برای لم یایی نیاز به برچسب زنی اجزای کلام نیز وجود دارد.

- در بخش نرمال سازی یکسان سازی نویسه ها، حذف کارکترها و فواصل اضافی و اصلاح نشانه گذاری انجام می گیرد.
- بخش جداسازی مرز جملات شامل الگوریتم دقیقی است که هنگام جداسازی جملات به استثنائات توجه می کند تا حتی الامکان در شکستن متن به جملات اشتباه رخ ندهد. در این راستا در کنار جداسازی جملات با علائم جدا کننده ای مانند خط جدید، علامت تعجب و سوال، برای نقطه، هنگام رخداد نقطه بررسی مکان نقطه، کلمات مجاور آن، طول کلمه حاوی نقطه و تعداد نقاط رخ داده در هر توکن بررسی شده و در صورت رخداد حروف انگلیسی یا فارسی یا صورت تلفظی آنها، توکن های مختص آدرس های اینترنتی و یا اسامی خاص جداسازی صورت نمی گیرد.
- بخش جداسازی کلمات با استفاده از جداکننده های رایجی مانند فاصله، علائم و .. جداسازی کلمات را انجام می دهد.
- برای غلط یابی از مازول متن باز ویراستیار اصلاح شده به همراه مازول پیش پردازشی و پس پردازشی غلط یابی استفاده شده است.
- برای لم یایی، برچسب زن اجزای کلام با ۱۰۰ برچسب آموزش دیده بر روی بخش با برچسب پیکره متنی شامل ۸ میلیون کلمه مورد استفاده قرار می گیرد که بیش از ۹۴ درصد دقت دارد.

۳-۲- منابع مورد استفاده

در این الگوریتم از منابعی مانند لیست کلمات، لیست وندها (استخراج شده از flexicon) [۱۵]



شکل (۱): فلوجارت روش پیشنهادی برای یافتن لمای کلمه

۴- ارزیابی

روش مبتنی بر قانون و با استفاده از چندین منبع زبانی از جمله فهرستی از افعال زبان فارسی، جمع مکسر، واژگان زبانی زبان فارسی و ... تولید شده است.

الگوریتم اصلی مورد استفاده در این ابزار به گونه‌ای است که ابتدا بررسی می‌شود که کلمه باید ریشه‌یابی شود یا خیر و در صورت لزوم الگوریتم ریشه‌یابی انجام می‌شود. در الگوریتم ریشه‌یابی ابتدا برچسب اجزای کلام برای هر کلمه تعیین می‌گردد. این امکان در ریشه‌یاب قرار داده شده که یا فقط از کلماتی که یکی از سه برچسب، فعل‌ها، اسامی و صفت‌ها را دارند به تنهایی ریشه‌یابی شوند (فقط فعل‌ها یا فقط اسم‌ها و یا فقط صفت‌ها ریشه‌یابی شوند) و یا همه موارد در متن ریشه‌یابی شوند. همچنین با توجه به بار پردازشی برچسب زن اجزای کلام و زمان‌بری روال، یک مد سبک در برنامه لحاظ شده است که در آن فقط با توجه به شکل ظاهری کلمات قوانین تعیین شده و ریشه‌یابی انجام می‌گیرد. با توجه به تنوع اقسام ریشه‌یاب و بن‌یاب و عدم وجود پیکره تست استاندارد، ارزیابی مقایسه‌ای برای این ابزار انجام نشده است ولی نتیجه ارزیابی این ابزار هم با ارزیابی انسانی و همچنین با ارزیابی روی مجموعه داده محدود دقتی قابل قبولی را به دست آورده است. این نتایج مناسب با توجه به جامعیت ابزار و در نظر گرفتن تمامی نکات مثبت که تقریباً در تمامی روش‌های ریشه‌یابی فارسی گزارش شده است و همچنین استفاده مناسب از اطلاعات برچسب زنی اجزای کلام و الگوریتم‌های مناسب استفاده از الگوریتم پیشنهادی را قابل توجه می‌نماید.

مراجع

- [1] M. Tashakori, M. Meybodi, and F. Oroumchian, "Bon: the Persian stemmer," in EurAsia-ICT 2002: Information and Communication Technology. Heidelberg, Germany: Springer, 2002, pp. 487-494.
- [2] K. Taghva, R. Beckley, and M. Sadeh, "A stemming algorithm for the Farsi language," in Proceedings of 2005 International Conference on Information Technology: Coding and Computing (ITCC), Las Vegas, NV, 2005, pp. 158-162.
- [3] S. Estahbanati and R. Javidan, "A new stemmer for Farsi language," in Proceedings of 2011 CSI International Symposium on Computer Science and Software Engineering (CSSE), Tehran, Iran, 2011, pp. 25-29.
- [4] J. Mehrad and S. R. Berenjian, "Providing a Persian language singular-stemmer system (RiCeST Stemmer)," International Journal of Information Science and Management, vol. 9, no. 2, pp. 13-22, 2011.
- [5] R. Hesamifard and G. Ghassem-Sani, "A stemming algorithm for the Persian words," in Proceedings of the 11th Annual International CSI Computer Conference (CSICC2006), Tehran, Iran, 2006, pp. 515-519.
- [6] M. M. Nasiri, K. S. Esmaili, and H. Abolhassani, "A statistical stemmer for Persian language," in Proceedings of 11th International CSI Computer Conference (CSICC2006), Tehran, Iran, 2006.
- [7] Shamsfard, M., Jafari, H.S., Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. LREC 2010 - 8th Language Resources and Evaluation Conference, 19-21 May, Malta.
- [8] A. Rahimi, Hybrid stemming for Persian, CoRR,(2015).
- [9] K. megerdoomian, Finite-state morphological analysis of Persian. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Coling 2004.
- [10] A. A. Sharifloo and M. Shamsfard, "A bottom up approach to Persian stemming," in Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP), Hyderabad, India, 2008, pp. 583-588.
- [11] Sarabi, Zahra, Hooman Mahyar, and Mojgan Farhoodi. "ParsiPardaz: Persian Language Processing Toolkit." In Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on, pp. 73-79. IEEE, 2013.
- [12] A. H. Jadidinejad, F. Mahmoudi, and J. Dehdari, "Evaluation of PerStem: a simple and efficient stemming algorithm for Persian," in Multilingual Information Access Evaluation I: Text Retrieval Experiments. Heidelberg, Germany: Springer, 2010, pp. 98-101.
- [13] E. Rahimtoroghi, H. Faili, and A. Shakery, "A structural rule-based stemmer for Persian," in Proceedings of 2010 5th International Symposium on Telecommunications (IST), Tehran, Iran, 2010, pp. 574-578.
- [14] M. H. Dianati, M. H. Sadreddini, A. H. Rasekh, S. M. Fakhrahmad, and H. Taghi-Zadeh, "Words stemming based on structural and semantic similarity," Computer Engineering and Applications Journal, vol. 3, No. 2, pp.89-99, 2014.

[۱۵] محرم اسلامی، مسعود شریفی آتشگاه، صدیقه علیزاده لمجری، و طاهره زندی، ۱۳۸۳، واژگان زبانی زبان فارسی. مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران.

برای ارزیابی عملکرد ریشه‌یاب پیشنهادی دو روش در نظر گرفته شده است. روش اول اینکه دقت ریشه‌یاب پیشنهادی مستقلاً با استفاده از معیار دقت (تعداد عملیات ریشه‌یابی صحیح به کل واژگان) و همچنین نظر کارشناس انسانی اندازه گرفته شود و دیگر اینکه این ماژول در یک برنامه کاربردی قرار داده شود و میزان تاثیر آن در کارایی برنامه اندازه گیری شود.

با توجه به تنوع اقسام ریشه‌یاب و بن‌یاب و عدم وجود پیکره تست استاندارد در این حوزه برای زبان فارسی، ارزیابی یکنواخت مقایسه‌ای با الگوریتم‌های دیگر برای این ابزار انجام نشده است ولی برای این ابزار با دو روش عنوان شده بر روی حجم مناسبی از داده ارزیابی انجام شده است. با توجه به نتایج ارائه شده در ادامه می‌توان عملکرد مناسب الگوریتم پیشنهادی را مشاهده نمود.

۴-۱- ارزیابی مستقل ریشه‌یاب

برای ارزیابی مستقل ماژول ریشه‌یاب و لم‌یاب، میزان دقت ریشه‌های خروجی حاصل از ریشه‌یاب، جهت ارزیابی کارایی آن مورد بررسی قرار گرفت. با مناسب به نظر رسیدن این روش، به کمک ده سند خبری که از سایت‌های مختلف خبری انتخاب شده‌اند، میزان دقت ریشه‌های به دست آمده از ریشه‌یاب پیشنهادی بررسی گردید. نتایج در جدول (۱) آمده است. از سوی دیگر نتایج عنوان شده در جدول زیر با نظر دو کارشناس زبانی شناسی رایانشی تایید شده است.

جدول (۱): ارزیابی سیستم ریشه‌یاب پیشنهادی

۱۰۵۳	تعداد واژه‌ها
۹۲،۵	درصد ریشه‌های صحیح بن‌یاب
۹۰،۴	درصد ریشه‌های صحیح ریشه‌یاب

۴-۲- ارزیابی ماژول در کاربرد

در ادامه روال ارزیابی، از دادگان لم‌یابی شده توسط الگوریتم پیشنهادی در یک سیستم خلاصه ساز استخراجی متن استفاده شد. نتایج بررسی دقت عملکرد سیستم خلاصه ساز، با و بدون استفاده از ماژول لم‌یاب پیشنهادی در جدول ۲ آورده شده است.

خلاصه ساز استفاده شده مبتنی بر روش‌های خوشه‌بندی سلسله‌مراتبی است که بر روی ۵۰ فایل متنی خبری یک‌بار بدون اعمال ریشه‌یابی و یک‌بار با اعمال ریشه‌یابی (موثر در بخش سنجش شباهت بین جملات متن) تست شده است و با فایل‌های مرجع تولید شده انسانی مقایسه شده است. این مقایسه بر اساس معیار f که میانگین هارمونیک دو معیار دقت (نسبت جملات صحیح انتخاب شده به انتخاب‌های سیستم)، بازخوانی (نسبت جملات انتخاب شده صحیح به آنچه باید انتخاب می‌شده) است، انجام شده است.

جدول (۲): نتایج اعمال ریشه‌یاب پیشنهادی روی خلاصه ساز

معیار f	دقت بازخوانی	بدون اعمال ریشه‌یاب
۳۷،۲	۵۵	۲۸،۲
۳۸،۴	۵۶،۷	۲۹،۱

۵- نتیجه

وظیفه ابزار ریشه‌یاب و لم‌یاب، یافتن ریشه و بن کلمات است. با توجه به نیاز حوزه پردازش زبان طبیعی زبان فارسی به ابزار مناسب ریشه‌یاب و بن‌یاب، در این مقاله سعی شده است که علاوه بر بررسی جامع بر روی گزارشها و مقاله‌های منتشر شده برای مسئله ریشه‌یابی و بن‌یابی در حوزه زبان فارسی، نقاط قوت و ضعف آنها به طور مناسبی تحلیل شود و بر اساس نیازهای این حوزه یک الگوریتم کارا و جامع برای این مسئله پیشنهاد شود. ابزار تولید شده بر اساس الگوریتم پیشنهادی در زیرگروه پردازش متن پژوهشگاه خواجه نصیرالدین طوسی تهیه شده و دارای چند حالت مختلف برای ریشه‌یابی و بن‌یابی انواع مختلف کلمات است. این ابزار با