معرفی یک الگوریتم ریشه یابی و لمیابی مبتنی بر قانون برای زبان فارسی

Conference Paper · January 2016

CITATIONS
0

READS
733

2 authors, including:

Yasser Shekofteh
Shahid Beheshti University
62 PUBLICATIONS **279** CITATIONS

SEE PROFILE

اولین همایش ملی جویشگر بومی
۲۷ و ۲۸ بهمن ماه سال ۱۳۹٤
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
وزارت ارتباطات و فناوری اطلاعات
ITRC

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 همچنین 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 می گیرد. 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 در لغت به معنای حذف پسوندها، پیشوندها و میانوندهای کلمه و به‌دست آوردن ریشهٔ کلمه است. اگرچه در هر زبان، واژه‌ها با توجه به نقش معنایی و نحوی خود در جلمه 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 می شود. بنابراین ریشه‌یابی مسئله ای است که در بسیاری از زمینه های پردازش زبان طبیعی مورد نیاز می باشد و در بسیاری از کاربردها، نیاز داریم تا همه مشتقات یک واژه را به ریشه‌ی آن، که همان شکل ساده واژه می‌باشد، تبدیل نماییم و البته باید توجه داشت که نیاز کاربردهای مختلف برای نحوه ریشه‌یابی متفاوت است. در این زمینه اغلب با دو واژه ریشه‌یابی و لم‌یابی روبرو هستیم. مطابق تعریف [11] در ریشه‌یابی عموما وندها جدا می شوند ولی در لم‌یابی با تحلیل مورفولوژی و بررسی کلمات زمینه، بن کلمه بازگردانده می شود. انتخاب از میان این دو وابسته به شرایط مورد نیاز در کاربردهای مختلف مثل عمق ریشه‌یابی مورد نیاز از 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 می شوند. به صورت کلی روش های ساختاری وابسته به تحلیل ساخت واژی بوده، مبتنی بر قاعده عمل می کنند و با استفاده از مجموعه قانون های تعریف شده ریشه‌یابی را انجام می دهند. در اینگونه

روش‌ها با وجود قابلیت اعمال قانون به حجم وسیعی از کلمات فارسی، استثنائات بوجود آمده ممکن است ایجاد مشکل نمایند. در روش های مبتنی بر جدول، یک جدول برای کلمات شناخته شده و ریشه تعیین شده آنها در نظر گرفته می شود و برای هر کلمه با رجوع به جدول به آسانی ریشه برگردانده می شود. در صورت عدم وجود کلمه در جدول، ریشه آن با روش های معمول دیگر استخراج شده و به جدول افزوده می شود. ایراد این دسته روش‌ها حجم بالای دسترسی و ذخیره سازی و همچنین نیاز به روزرسانی است. در روش های آماری سعی می شود با استفاده از روش هایی از قبیل بررسی تعداد رخداد ترکیب ریشه های مختلف با وندهای مختلف، احتمال وقوع ریشه محاسبه شود و با استفاده از روش های 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 می دهد و در دسته روش های ساختاری قرار می گیرد. در این روش سعی شده جامعیت قوانین و استثنائات مد نظر قرار گرفته، از منابع مختلف و ابزارهای پیش پردازشی مناسب و با دقت بالا استفاده شود و با اعمال تنظیماتی در حالت های مختلف ریشه‌یابی سبک و بن یابی برای مقوله های 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 نتیجه گیری مقاله آورده خواهد شد.

۲- مروری بر کارهای پیشین

گفتیم که در پردازش زبانهای طبیعی و سیستم های بازیابی اطلاعات، عموما مواردی وجود دارد که باید ریشه واژگان به صورت خودکار استخراج گردد. به همین منظور ریشه‌یاب های خودکار متنوعی برای زبانهای مختلف ساخته شده است. برای زبان انگلیسی روش های متعددی معرفی شده است که در میان روش های پایه مطرح شده برای زبان انگلیسی، می توان به الگوریتم های پورتر و کراوتز 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 نمی کند. همچنین الگوریتم کراوتز از روشهای ریخت شناسی و از یک فرهنگ لغت برای آزمودن ریشه های یافت شده استفاده می کند. این الگوریتم پسوند و پیشوند کلمات را بررسی می کند و در ماشینهای مترجم و برای زبانهایی که ساخت کلمات در آنها قانونمند است، کارائی خوبی را نشان داده است.

در زمینه ریشه‌یابی کلمات فارسی نیز روش های متعددی معرفی شده است که در ادامه آورده شده اند که عموما ارزیابی جامعی برایشان انجام نشده بود. این ارزیابی‌ها غالبا 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 می شود و اگر خود کلمه ریشه تشخیص داده نشد با استفاده از قوانین، حذف وندها و بررسی مصادر و جمع های مکسر ریشه‌یابی صورت می گیرد. روش ارائه شده در [2] یکی از مشهورترین روش های مطرح شده برای زبان فارسی است. این الگوریتم شباهت زیادی به الگوریتم پورتر در انگلیسی دارد که بر مبنای ریخت شناسی است. هر دو ریشه‌یاب پسوندهای خاصی را جستجو می کنند و مراحل مختلفی را بر طبق لیست قوانین پسوندی پشته گذاری 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 زیررشته های نهایی (ترمینال) کلمه ورودی است که اگر در لیست وندها قرار داشت، حذف شود.

در [4] از دانش زبانی و الگوریتم های استاندارد قوانین ماشینی استفاده شده و همچنین پسوندهای جمع و استثنائات افعال مرکب در نظر گرفته شده اند (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 جمع های مکسر عمل می کند.

در [5] الگوریتم کراوتز برای زبان فارسی گسترش یافته است. این الگوریتم در دسته روش های ریشه‌یابی مبتنی بر جدول قرار گرفته و روال کلی آن به این صورت است که پس از ساخت جدول کلمات و ریشه های آنها، با ورود هر کلمه در جدول یک فرایند جستجو صورت می گیرد، اگر موجود بود که به سادگی ریشه کلمه برگردانده می شود و در غیر این صورت، وندهای احتمالی کلمه حذف شده و مجددا جستجو صورت می گیرد. 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 می کند. هر کلمه در روش پیشنهادی به صورت دو بخش ریشه و وند و هر زیررشته یک گره از گراف در نظر گرفته می شود. یال های گراف نشان دهنده نحوه ترکیب‌ها هستند. همچنین با استفاده از یک مجموعه داده و شمردن تعداد تکرار در ترکیب های مختلف یال‌ها وزندهی می شوند. هدف در این مقاله تخمین احتمال برای همه زیررشته های ممکن و تحلیل آماری برای ریشه‌یابی است.

ابزار ارائه شده در [7] با درنظر گرفتن صرف افعال و چند قانون مورفولوژیکی اشتقاق عمل می کند. همچنین در این سیستم از برچسب اجزای کلام (pos) کلمه و ریشه، FSA، دیتابیس هایی شامل کلمات و برچسب های pos آنها، ساختاری برای ذخیره سازی ریشه و برچسب های مورد نظر آن، ۲ لیست پیشوند و پسوند (که باید حذف شوند) و چند قانون برای تحلیل مورفولوژیک استفاده می شود.

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

در [9] یک تحلیلگر مورفولوژیکی ۲ مرحله ای 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 چالش های مختلف سیستم ریشه‌یابی فارسی بیان شده است. این چالش‌ها به دو دسته مورفولوژی غیرفعلی مثل فاصله افتادن بین بخش های کلمات، توکن های پیچیده (مثل انقلابیترینهایشان)، قوانین فونتیکی و دسته مورفولوژی فعلی مانند مسائل مربوط به بن های ماضی و مضارع و وابستگی های طولانی تقسیم شده است.

روش پیشنهادی [10] مبتنی بر قانون و یک روش از پایین به بالاست. به این صورت که ابتدا تلاش می شود که زیررشته های کلمه استخراج شوند (تشخیص همه زیررشته های مجاز که هسته نامیده می شوند و تشخیص مرفم‌ها و خوشه بندی آنها). بعد سعی می کند که مطابق قوانین با چسبانیدن هسته به عناصر دیگر ، کلمات مجاز ایجاد کند. هر هسته ای که حداقل یک تولید کلمه مجاز داشته باشد ریشه تشخیص داده می شود. در مرحله بعد حروفی که احیانا اضافه هستند (مانند گ در ستارگان) بر طبق یکسری قانون تشخیص داده می شوند.

در ابزار معرفی شده در [11]، هم ریشه‌یابی و هم لم‌یابی انجام می شود. 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 می شود، اگر بود که به سادگی ریشه آن برگردانده می شود و در غیر اینصورت وندهای احتمالی آن حذف می شود. کلمه حاصل اگر در لیست واژگان زایای زبان فارسی [۱۵] موجود بود بررسی می گردد که با توجه به برچسب اجزای کلام کلمه آیا وند حذف شده معتبر است یا خیر. در الگوریتم لم‌یابی نیز ابتدا همه ریشه های ممکن بدست آورده شده و بعد از بررسی برچسب اجزای کلام کلمه انتخاب مناسب صورت می گیرد.

در این الگوریتم [12] از regular expressionها برای جداسازی مرفم‌ها از ریشه استفاده می شود. ابتدا در جدول درهم سازی شده جستجو می شود، در صورت عدم یافته شدن با استفاده از عبارات منظم، جداسازی وندها از کلمه صورت می گیرد.

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

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

۳- سیستم پیشنهادی

همانطور که پیش تر گفته شد، در سیستم پیشنهادی هم ریشه‌یابی انجام می شود و هم لم‌یابی. در هر دو این موارد، روال کلی انجام کار به این صورت است که ابتدا بررسی می شود که کلمه باید ریشه‌یابی شود یا خیر و در صورت لزوم الگوریتم ریشه‌یابی انجام می شود. یک کلمه ریشه‌یابی نمی شود اگر دارای کارکترهای غیرفارسی بوده، دارای عدد بوده، مخفف بوده و یا جزء اسامی خاص باشد. در حالت لم‌یاب ابتدا برچسب اجزای کلام برای هر کلمه مشمول ریشه‌یابی تعیین می گردد، چون برخی کلمات مانند قیود یا حروف اضافه نیاز به ریشه‌یابی ندارند. پس این امکان در لم‌یاب قرار داده شده که به تفکیک آرگومان، فعل ها، 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

شده است که در آن فقط با توجه به ریخت شناسی کلمات قوانین تعیین شده و ریشه‌یابی انجام می گیرد.

الگوریتم لم‌یابی، برای هر ۳ نوع کلمات اسم، فعل یا صفت متفاوت است. اما در مد ریشه‌یابی سبک روال‌ها نسبتا 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 لم‌یابی 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 رخداده 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 ماژول متن باز ویراستیار اصلاح شده به همراه ماژول پیش پردازشی و پس پردازشی غلط یابی استفاده شده است.
- برای لم‌یابی، 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] 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.

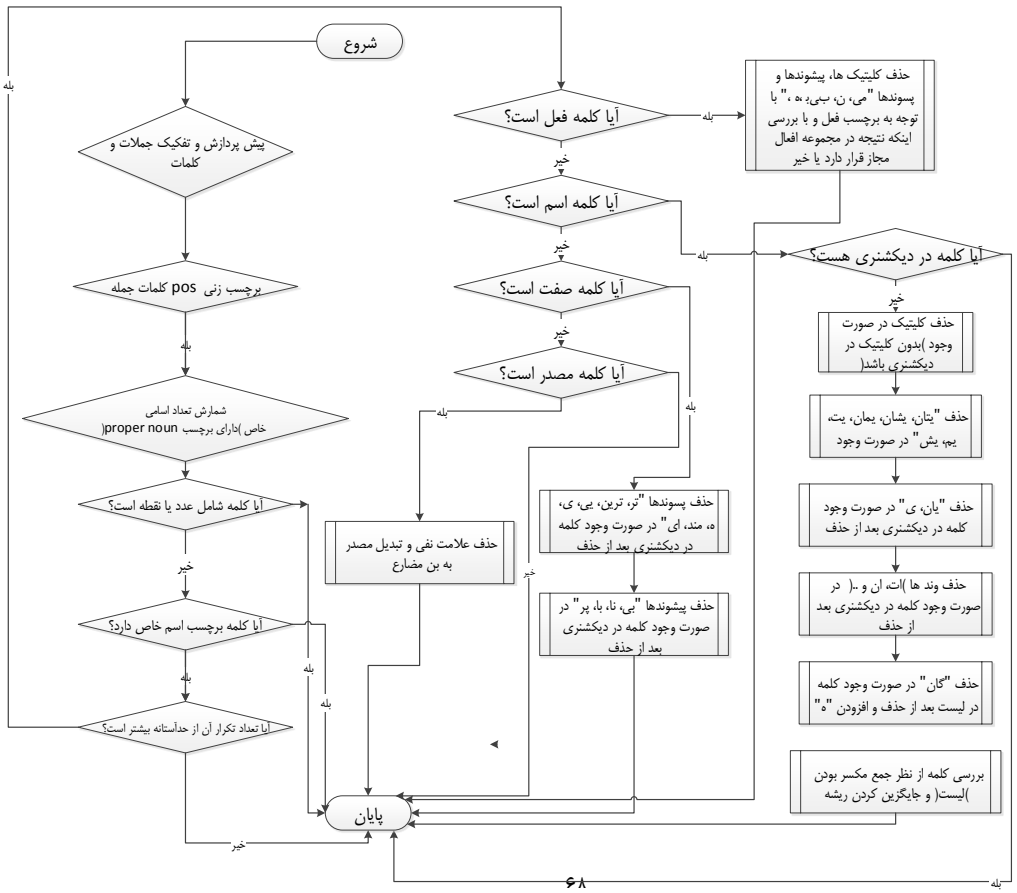
۳-۳- مد لم‌یابی

در این بخش الگوریتم ریشه‌یابی کلمات با توجه به برچسب اجزای کلام آنها یا همان لم‌یابی معرفی می گردد. همانطور که ذکر شد، از برچسب زن ۱۰۰ برچسبی استفاده شده است که اطلاعات مفیدی برای استخراج ریشه ارائه می کند. لم‌یابی 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 ریشه‌یابی 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 پسوندهای 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 شناسه‌ها (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 ریشه‌یابی سبک

در این مد الگوریتم ریشه‌یابی ساده ای بر روی کلماتی که باید روی آنها ریشه‌یابی 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 معمولا در قاعده بالا نمی گنجند) پس از حذف کلیتیک و شناسه و پیشوند، در لیست بن‌ها جستجو می شود و در صورت وجود، ریشه برگردانده می شود. شکل (۱) فلوچارت روش لم‌یابی را نشان می دهد.



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

۴- ارزیابی

برای ارزیابی عملکرد ریشه‌یاب پیشنهادی دو روش در نظر گرفته شده است. روش اول اینکه دقت ریشه‌یاب پیشنهادی مستقلا 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 مقایسه ای با الگوریتم های 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 می توان عملکرد مناسب الگوریتم پیشنهادی را مشاهده نمود.

۴-۱- ارزیابی مستقل ریشه‌یاب

برای ارزیابی مستقل ماژول ریشه‌یاب و لم یاب، 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 سایت های 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 زبان شناسی رایانشی تایید شده است.

جدول (۱): ارزیابی سیستم ریشه‌یاب پیشنهادی

تعداد واژه‌ها	۱۰۵۳
درصد ریشه‌های صحیح بن یاب	۹۲,۵
درصد ریشه‌های صحیح ریشه‌یاب	۹۰,۴

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 روش های خوشه بندی سلسله مراتبی است که بر روی ۵۰ فایل متنی خبری یک‌بار بدون اعمال ریشه‌یابی و یک‌بار با اعمال ریشه یابی (موثر در بخش سنجش شباهت بین جملات متن) تست شده است و با فایل های 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 انتخاب های سیستم)، بازخوانی (نسبت جملات انتخاب شده صحیح به آنچه باید انتخاب می شده) است، انجام شده است.

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

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

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 مقاله های منتشر شده برای مسئله ریشه‌یابی و بن یابی 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 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 ... 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 مقایسه ای 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 تقریبا در تمامی روش های ریشه‌یابی فارسی گزارش شده است و همچنین استفاده مناسب از اطلاعات برچسب زنی اجزای کلام و الگوریتم های مناسب استفاده از الگوریتم پیشنهادی را قابل توجیه می نماید.

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. Esmaeili, 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 لمجیری، و طاهره زندی، ۱۳۸۳، واژگان زایای زبان فارسی. مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران.