روشی نوین جهت صرف واژه های فارسی

بهرنگ قاسمی زاده سعید رحیمی مرتضی سالاریان علی ترکمنی
حیدر سمیاری عباس کوچاری مجید نم نبات لقمان براری

دپارتمان بومی سازی، شرکت دیجیتال کلون
QasemiZadeh@DigitalClone.net

چکیده: در این مقاله رویکردی نوین برای صرف واژه در زبان فارسی ارائه شده است. صرف بخشی از دستور زبان است که ساخت واژه را مورد بررسی قرار می‌دهدازمهمترین مباحث صرف،تصریف و اشتقاق است که در زبان فارسی به کمک وندافزایی صورت می‌گیرد. در این مقاله روشی جدید با پیچیدگی محاسباتی مناسب جهت حل مشکل صرف تصریفی واژه‌ها در زبان فارسی ارائه شده است. روش پیشنهادی به سادگی قابل بسط برای دیگر زبان‌های هند و اروپایی است که در آن تصریف واژه از طریق وندافزایی صورت می‌گیرد. در روش پیشنهادی، مسئله تحلیل لغوی به مسئله جستجوی جزیی هدایت‌شده بر روی درخت جستجوی سه‌تایی تبدیل شده است. مشکل موجود در صرف تصریفی شامل استخراج قوانین، تعداد زیاد قوانین و نحوه اعمال آن بر روی واژه‌های زبان است. از جمله مشکلاتی که در تحلیل صرفی، مبتنی بر دانش و قوانین با آن مواجه هستیم می‌توان از تعداد زیاد قوانینی که می‌توانند همزمان اعمال شوند، عدم یکنواختی، و فضای جستجوی بزرگ نام برد. در این پژوهش سعی شده است روشی برای حل این مشکلات ارائه شود.

کلمات کلیدی: صرف واژه های فارسی، پردازش زبان طبیعی، تحلیل لغوی، واژگان، بازنمایی دانش

۱- مقدمه

می‌دانیم واحد زبان، جمله است و جمله خود از واحدهای کوچکتری که واژه[1] نام دارد تشکیل شده است. از کنار هم نشستن واژه‌ها با توجه به نحو زبان، جمله‌واره و جمله ساخته می‌شود. واژه از اجزای کوچکتری به نام تکواژ[2] تشکیل شده است. به بررسی ساختمان واژه، صرف[3] (ساخت واژه) گویند. تکواژ که کوچکترین واحد معنادارزبان است به دو دسته تقسیم می شود: ۱)تکواژ آزاد[4] مانند " درخت" که به‌طور مستقل می‌تواند در جمله ظاهر شود.۲) تکواژ مقید[5] مانند "-ی" در "شادی" که به تنهایی نمی‌تواند در جمله به‌کار رود. تکواژ مقید خود بر دو گونه است:۱) تکواژ صرفی[6] مانند شناسه‌های فعلی (ام، ای، ...) که افزودن آنها به تکواژ پایه به‌دلیل الزامات صرفی جهت ایفای نقش نحوی است و سبب تغییر مقوله زبانی واژه و نیز معنا نمی‌شود. در زبان فارسی این تکواژها بیشتر به‌صورت پسوند ظاهر می‌شوند. ۲) تکواژ اشتقاقی[7] مانند "-گاه" در "دانشگاه" که معمولاً حضور آن در واژه سبب تغییر مقوله واژه شده و تغییر معنایی را به‌همراه دارد. این تکواژها هم به‌صورت پیشوند و هم به‌صورت پسوند ظاهر می‌شوند.

از آنجا که تکواژهای مقید به پیش و یا پس یک تکواژ مقید و یا آزاد دیگر افزوده می شوند، به آنها وند[8] می‌گویند. به فرایند اضافه کردن وند به تکواژ پایه وندافزایی گویند. اگر این وند در پیش واژه قرار گیرد به آن پیشوند[9]، و اگر در پس واژه قرار گیردبه آن پسوند[10] گویند. بنابراین وند تکواژ مقیدی است که می‌توان آن را به دو نوع تصریفی و اشتقاقی تقسیم کرد.

اگر مقایسه‌ای ساده بین وند تصریفی و اشتقاقی صورت گیرد، خواهیم دید که وند تصریفی نقش نحوی دارد و واژه را برای ایفای نقش معینی در ساخت نحوی پردازش می‌کند و هرگز مقوله زبانی[11] تکواژ پایه را تغییر نمی‌دهد. کاربرد وند تصریفی قیاسی است به این مفهوم که با همه اعضای یک مقوله از واژه‌ها می‌تواند به‌کار رود و کمتر استثنا می‌پذیرد. اما وندهای اشتقاقی نقش واژه‌سازی داشته و واژه جدید می‌سازد. وند اشتقاقی اغلب مقوله زبانی تکواژ پایه را تغییر می‌دهد. کاربرد وندهای اشتقاقی سماعی است یعنی با همه اعضای

[1] Word
[2] Morpheme
[3] Morphology
[4] Free Morpheme
[5] Bound Morpheme
[6] Inflectional Morpheme
[7] Derivational Morpheme
[8] Affix
[9] Prefix
[10] Suffix
[11] Part of Speech (POS)

یک مقوله از واژه‌ها به‌کار نمی‌رود. تعداد وندهای تصریفی به مراتب کمتر از وندهای اشتقاقی است. برهمین اساس، بسامد کاربرد وند تصریفی از وند اشتقاقی بسیار بالاتر است. از نقطه نظر محل قرار گرفتن وندهای اشتقاقی در مقایسه با وند تصریفی، معمولاً وند اشتقاقی به پایه نزدیکتر است و در درونیترین لایه ساخت‌واژه قرار می‌گیرد. درحالی‌که وند تصریفی در برونترین لایه ساخت‌واژه قرار می‌گیرد. اما باید توجه نمود که در زبان فارسی بعضی از وندهای تصریفی قبل از وندهای اشتقاقی قرار می‌گیرند. وند تصریفی و اشتقاقی را می‌توان از لحاظ روابط معنایی از هم بازشناخت. در وند تصریفی رابطه میان معنی تکواژه پایه و معنی تکواژ جدید حاصل از افزایش وند به پایه کاملاً قابل پیش‌بینی است و از قوانین مشخصی تبعیت می نماید ؛ درحالی‌که در وند اشتقاقی رابطه میان تکواژ پایه و معنی تکواژ جدید بدست آمده از وندافزائی غیرقابل پیش‌بینی است[۱].

اگر وندها را با توجه به محل قرارگیریشان نسبت به پایه دسته‌بندی کنیم، سه دسته وند خواهیم داشت: پیشوند، پسوند و میانوند[1]. لازم به تذکر است که در فارسی میانوند وجود ندارد [۱][۲]. یکی از مشکلات مهم در زبان فارسی که باید به آن توجه نمود، هم‌نویسه بودن گروهی از وندهای تصریفی با وندهای اشتقاقی است؛ چراکه در تحلیل واژه‌های حاصل از افزایش چنین وندهایی، با ابهام روبرو هستیم مانند "-ی" نشانه نکره در مقایسه با "-ی" وند اشتقاقی . آنچه در این مقاله مورد بحث قرار می‌گیرد، بررسی ساخت‌واژه واژه‌های حاصل از افزایش وندهای مقید تصریفی است.

مسئله تحلیل لغوی یکی از مسائل قدیمی در پردازش زبان طبیعی می‌باشد. امروزه برای اکثر زبان‌های زنده دنیا، سیستم‌های پردازش لغوی وجود دارد. پوشش مناسب از واژه‌های زبان، بار محاسباتی کم، سرعت پردازش بالا و همچنین دقت کافی از مهمترین مسائل در طراحی پردازشگرهای لغوی است. روش‌های متعددی برای پردازش لغوی پیشنهاد شده است. بیشتر کارهای انجام شده، طراحی تحلیلگرهای لغوی مبتنی بر ماشین حالت متناهی [2] بوده است[۳][۴][۵] [۶][۷]. علاوه بر آن روش‌های آماری و یا مبتنی بر یادگیری ماشین نیز برای استفاده در طراحی سیستم‌های ریشه یاب[3] خودکار ارائه شده است[۸][۹][۱۰][۱۱].

در [۸] یک معماری مستقل از زبان مبتنی بر مدل مخفی مارکوف برای تحلیلگر لغوی ارائه شده است. این سیستم برای زبان‌های چینی، ژاپنی و انگلیسی آزمایش شده است. سیستم فوق پیش از استفاده نیاز است تا بر روی پیکره[4] زبانی که مقوله زبانی واژه های آن مشخص شده است آموزش داده شود. متاسفانه چنین پیکره هایی برای استفاده عمومی در زبان فارسی منتشر نشده است.

در [۹] یک سیستم تحلیلگر لغوی غیرقطعی[5] مبتنی بر آموزش حافظه مدار[6] ارائه شده است. در این روش یک نگاشت مستقیم از حروف در یک متن به لایه‌های متفاوتی که اطلاعات صرفی را در خود رمز نموده‌اند، تعریف شده است. این سیستم نیز پیش از استفاده نیاز است آموزش داده شود. در حقیقت این سیستم با تشخیص مرزبندی‌ها در یک واژه، آنها را طبقه بندی و برچسب دهی می‌نماید.

در [۱۲] یک روش مبتنی بر واژگان برای حل مشکل تحلیل لغوی ارائه شده است. در این مقاله از یک گراف غیرمدور برای ذخیره اطلاعات واژگان به همراه اطلاعات ساخت واژه استفاده شده است که پردازش فقط از طریق یکبار جستجو در این گراف صورت می‌گیرد.

برای زبان فارسی تا به حال چندین کار انجام شده است به‌عنوان مثال [۷] و [۱۳]. در [۱۳] از یک معماری مبتنی بر ماشین حالت متناهی جهت ارائه یک تحلیل گر لغوی استفاده شده است. آقای ریاضتی در [۷]، یک سیستم تحلیلگر لغوی دو سطحی برای تحلیل صرفی و اشتقاقی زبان فارسی ارائه نموده است. مدل Perslex با الهام از مدل Englex [۱۴] ارائه شده است. مشابه چنین کاری در[۱۵] و به کمک ابزارهای Xerox finite state برای ارائه یک سیستم تحلیلگر لغوی برای زبان فارسی استفاده شده است.

هدف تحلیلگر لغوی، تحلیل ویژگی‌های صرفی واژه و تشخیص تغییرات صورت گرفته در آن به واسطه نقش نحوی آن است. برای رسیدن به این هدف، ابتدا یک دسته‌بندی جدید از مقوله‌های زبانی[7] در زبان فارسی ارائه شده است. سپس با توجه به این دسته‌بندی و استفاده از روشی مناسب جهت بازنمایی واژگان، یک روش جدید برای تحلیل صرفی واژه ها ارائه شده است.

در بخش ۲، مقوله‌های زبانی در فارسی شرح داده می شود. در ادامه و در بخش ۳ به واژگان، ساختار و رابطه آن با تحلیل لغوی اشاره می شود.. در بخش ۴ روش پیشنهادی در طراحی تحلیلگر لغوی آورده شده است. ساختار سیستم در بخش ۵ شرح داده می شود. بخش ۶ بیانگر نتایج به‌دست آمده است و در انتها به جمع بندی و پیشنهاد برای کارهای آینده پرداخته می‌شود.

۲- مقوله های واژگانی در زبان فارسی

به‌طور کلی دسته‌بندی های انجام شده بر روی مقوله‌های زبانی در زبان فارسی به دو دسته تقسیم می‌شود. روش سنتی و روش نوین. روش سنتی بر پایه دیدگاه‌های قدیمی ساخت زبان استوار است. در

[1] Infix
[2] Finite State Automata
[3] Stemmer
[4] Corpus

[5] Non Deterministic
[6] Memory Based Learning
[7] Part of Speech (POS)

این دسته‌بندی، اجزای کلام فارسی به هفت گروه تقسیم بندی شده است[۱۶]. از آنجا که روش سنتی بر پایه دیدگاه‌های قدیمی استوار است، چندان به رابطه صرف و نحو تکیه نداشته و بیشتر به تحلیل واژه مستقل از روابط نحوی می‌پردازد[۱۷][۱۸] دسته‌های ارائه شده در روش سنتی عبارتند از[۱۶]:

- اسم: می تواند مستقیما و مستقلاً نهاد(مسند الیه)جمله باشد وآن برای نامیدن شخصی یا حیوانی یا چیزی و یا مفهومی به کار می رود.
- صفت: حالت و مقدار و شماره یا یکی دیگر از چگونگی های اسم را می رساند.
- فعل: در جمله جایگاه اسناد را اشغال می کند، یعنی یا خود به نهاد اسناد داده می شود یا کلمه ای را به نهاد اسناد می دهد و به تنهایی یا به کمک وابسته هایی اغلب به چهار مفهوم دلالت می کند: ۱.شخص ۲.شمار ۳.زمان ۴. حالت یا کار.
- قید: کلمه یا گروهی که مفهومی به مفهوم فعل ویا صفت یا مسند یا قید و مصدر دیگر می افزاید.
- ضمیر: کلمه ای که معمولاً به جای اسم می نشیند.
- عدد
- اصوات
- حروف: نقش دستوری واژه های دیگر را نشان می‌دهد.

این دسته‌بندی دو مشکل اساسی دارد. اول این‌که دسته‌بندی سنتی اطلاعات کافی را جهت تحلیل نحوی زبان به کمک روش‌های نوین زبان‌شناسی فراهم نمی‌آورد. مشکل دوم این دسته‌بندی، عدم وجود دسته و گروهی خاص برای برخی از اجزای کلام در زبان فارسی است. این همان چیزی است که در زبان‌شناسی سنتی با نام ادات شناخته می‌شود. اما وجود چنین دسته‌بندی سبب فراهم آوردن اطلاعات افزوده‌ای نخواهد بود و کمکی در جهت تسهیل تحلیل رایانه‌ای زبان فراهم نمی‌آورد.

روشهای نوین زبان‌شناسی تعاریف جدیدی از مقوله های زبانی یا اجزای کلام در زبان فارسی ارائه داده‌اند [۱۹][۲۰]. با توجه به نواقص موجود در دسته‌بندی سنتی،[۲۰] دسته‌بندی جدیدی از مقوله‌های زبانی با توجه به زبان‌شناسی نوین ارائه کرده است. در این دسته بندی تکواژها بر پایه ویژگی‌های معنایی، صرفی و نحوی در چهار دسته جای می‌گیرند. در این دسته‌بندی تکواژ به‌عنوان کوچکترین واحد زبان بررسی می‌شود. دسته‌بندی مقوله‌های زبانی واژه های فارسی مطابق با این نظریه عبارتست از:

- تکواژ واژگانی یا قاموسی
  ۱- پایه فعل ۲- اسم ۳- صفت ۴- قید
  ۲- واژه‌های دستوری یا نقشی
  ۳- حرف اضافه ۲- حرف ربط ۳- ضمیر
- تکواژ اشتقاقی
- تکواژهای تصریفی
  ۱- نشانه‌های جمع
  ۲- نشانه‌های برتر برای واژه‌های صفت
  ۳- نشانه برتر برای واژه‌های قید
  ۴- نشانه‌های ماضی ساز
  ۵- نشانه‌های صفت مفعولی
  ۶- جزءِ پیشین فعلی
  ۷- شناسه‌های فعلی
  ۸- پی‌بست‌های ماضی نقلی
- واژه‌بست
  ۱- کسره اضافه و وصفی
  ۲- صورت‌های بودن
  ۳- ی نکره
  ۴- ضمایر متصل
  ۵- الفِ ندا : خدایا

این دسته‌بندی یک گروه‌بندی دقیق از مقوله‌های زبانی در فارسی ارائه می دهد؛ اما در بعضی موارد فاقد ارزش و اعتبار کافی برای تحلیل رایانه‌ای صورت نوشتاری واژه ها در زبان فارسی است. نگارندگان بر این عقیده‌اند که چگونگی دسته‌بندی ارائه شده برای مقوله‌های زبانی مستقیماً می‌تواند بر روی الگوریتم‌های طراحی شده تأثیرگذار باشد. به عبارتی دیگر می‌توان قسمتی از دانش زبانی مورد نیاز جهت تحلیل زبان را با کمک دسته‌بندی‌ها و روابط منطقی که میان این دسته‌ها برقرار است، به‌صورت ضمنی ارائه کرد. با توجه به آنچه که گفته شد و با توجه به مطالعات صورت‌گرفته و تجربیات کسب‌شده در طراحی و پیاده‌سازی نسخه‌های اولیه، دسته‌بندی جدیدی برای مقوله‌های زبانی تعریف شده است. دسته‌بندی ارائه‌شده، جز مواردی محدود، بسیار شبیه دسته‌بندی ارائه‌شده در روش گشتاری است.

دسته‌بندی پیشنهادی، مقوله‌های زبانی واژه‌های فارسی را در دو دسته مقوله‌های زبانی باز و مقوله‌های زبانی بسته قرار می دهد. مقوله‌های زبانی باز، فهرست بازی را تشکیل می‌دهند و در طول زمان ممکن است از آنها کاسته و یا به آنها اضافه شود. از طرف دیگر مقوله‌های زبانی بسته، فهرست بسته‌ای را تشکیل می‌دهد که تعداد آنها محدود و مشخص است و تنها برای نشان‌دادن نقش و یا رابطه دستوری خاصی به‌کار می‌رود، به عبارت دیگر استفاده از آنها جنبه نقشی دارد. در دسته‌بندی صورت‌گرفته، پاره‌ای از تعاریف سنتی در صرف فارسی، به دلیل ناکارآمدی در تحلیل رایانه‌ای، نادیده انگاشته شده است. این دسته‌بندی جدید در جدول های ۱ و ۲ نمایش داده شده است.

این دو گروه خود به دسته‌های کوچکتری تقسیم می‌شوند. این

دسته‌بندی‌ها با توجه به نحـو و نقـش متـداول واژه و جایگـاه آن در میان گروه‌ها و وابـسته‌هـای پیـشین و پـسین در گـروه‌هـای نحـوی مختلف صورت گرفته است. علاوه بر اینکه دسته‌بندی فوق به شکلی صورت گرفته است که روابط منطقی بین گروه‌ها وجود داشته باشـد. این روابط منطقی در قسمت بعد توضیح داده مـی‌شـود. ایـن روابـطِ میان-گروهی، اساس کار موتور تحلیل صرفی را تشکیل مـی‌دهـد. بـا مطالعه سطحی در دیگر زبان‌های هندی و اروپایی، دیده شـده اسـت که این روابـط بـرای اکثـر ایـن زبـان‌هـا وجـود دارد. تفـاوت دیگـر دسته‌بندی ارائه شده، قرارگیری مجموعه ونـدهای تـصریفی مختلـف در گروه های متفوت است. این کار به خصوص در مورد زبان فارسـی که درآن امکـان دارد ونـدهای مقیـد جـدا از پایـه مربـوط بـه آن در نوشته ظاهر شود، سبب سهولت در تحلیـل صـرفی و افـزایش دقـت خواهد شد.

در تحلیل نحوی، مقوله‌های زبانی ظاهرشده در زیرگـروه مقولـه‌هـای زبانی باز به‌همراه ویژگی‌های آنها به‌عنوان عناصر پایه‌ای دانش جهـت تحلیل نحوی واژه‌ها استفاده می‌شود. مجموعه ویژگی‌های مقوله‌هـای زبانی از چگونگی ترکیب مقوله‌های زبانی باز و بسته فراهم می شود و به شکل زوج خصیصه-ارزش[1] بازنمایی می‌شود.

جدول ۱- مقوله های زبانی باز

بن ماضی	بن مضارع	فعل
صفت	عدد	قید
اسم	ضمایر	شاخص
شبه جمله	حروف ربط	حروف اضافه

جدول ۲- مقوله‌های زبانی بسته

نشانه برتر برای قید	جز پیشین فعلی	نشانه‌های جمع
شناسه‌های فعلی	نشانه فعل دعایی	علامت نکره
الف ندا	نشانه‌های صفت مقایسه‌ای	ی واژه‌بست اضافی
نشانه‌های عدد ترتیبی	نشانه ماضی نقلی و بعید : ه	

۳- واژگان

در زبان‌شناسی لایه‌های متفاوتی برای بررسـی زبـان در نظـر گرفتـه شده است. دانشی که در پردازش زبان به کار می رود، با توجه به این لایه عبارتند از: دانش واجی(دانـشی دربـاره سـاخت هجـایی واژه در زبان) ، دانـش صـرف (سـاختواژی) (درمـورد سـاخت‌واژه در زبـان)، دانش نحوی (درباره ساخت جمله زبان بـر پایـه واژه‌هـای آن زبـان)، دانش معنایی[2](درباره مفهـوم و معنـی پـاره ای از گفتـار)، و دانـش کاربرد[3](درباره محل کاربرد واژه‌های زبان) و تحلیـل کـلام[4] (دربـاره مفاهیم و ارتباطات با توجه به بافت زبانی)، تقسیم‌بندی می‌شود. هر یک از این لایه‌هـا، از دانـش ایجادشـده در لایـه‌هـا بـالاتر اسـتفاده می‌کند.

واژگان[5] مجموعه اطلاعات دانش واژگانی است کـه در پـردازش زبـان مورد استفاده قرار می‌گیرد. این دانش به واژه‌های زبان وابسته اسـت. واژگان دانش اولیه و پایه در زبان‌شناسی رایانه‌ای را فـراهم مـی‌آورد. در انجام این پژوهش مجموعه‌ای از واژگان، با توجه بـه دسـته‌بنـدی ارائه شده در بخش ۲، تهیه شده است.

واژگان طراحی شده در پایین‌تـرین سـطح شـامل یـک پایگـاه داده[6] ساده است که هر یک از واژه‌های زبان در یـک رکـورد خـاص جـای گرفته اند. تراکنش‌های ایـن پایگـاه داده شـامل اضـافه نمـودن واژه جدید به واژگان، حـذف واژه و اصـلاح یـک واژه خـاص اسـت. بـرای بازنمایی واژه‌های زبان از یک ساختار خصیصه – ارزش استفاده شـده است. در این روش هر واژه با یک مقوله زبانی علامت خورده اسـت و هر مقوله زبانی ویژگی‌های خاص خود را دارد. واژه‌های هم‌نویسه[7] که به یک صورت نگارش می‌شوند دارای مقولـه‌هـای زبـانی و یـا معـانی متفاوت می‌باشند و هر یک بـصورت جداگانـه در پایگـاه داده ذخیـره شده اند. همان‌طور که گفته شد دسته‌بندی صورت گرفته و سـاختار مقوله‌های زبانی یا همان اجـزای کـلام سـبب فـراهم آمـدن دانـشی ضمنی بر روی واژگان شده است که از این دانش برای تحلیـل لغـوی استفاده می شود.

برای بازنمایی واژگان در سیستم پیشنهادی، از ساختمان داده درخت جستجوی سه‌تایی استفاده شده است. درخـت جـستجوی سـه‌تـایی اولـین بـار در [۲۱] معرفـی شـده اسـت. سـاختمان داده در درخـت جستجوی سه‌تایی به گونه ای است که رشته‌های اولیه مشترک میان داده‌های ذخیره شونده، تنها یک بـار ذخیـره مـی‌شـودکه نتیجـه آن فشرده‌سازی مناسب در بازنمایی داده های متنی است. در ایـن روش در هر گره از درخت، یک حرف ذخیره می‌شود. هر گره بـه سـه گـرهِ چپ، وسط و راست اشاره می‌کند. اشاره‌گر سمت چپ به حرف با کد کوچکتر و اشاره‌گر سمت راست به حرف با کد بزرگتر اشاره می‌کنـد. اشاره‌گر وسط بـه کـاراکتر بعـدی در رشـته ورودی اشـاره مـی‌کنـد. پیمایش درخت توسط گره های سمت چپ و راست، سبب پیمـایش در رشته ورودی نمی شود. در گره انتهـایی هـر واژه، مقولـه واژه آن کلمه به همراه لیست خصیصه – ارزش ذخیره شـده اسـت. بازنمـایی

[1] Attribute-Value pairs

[2] Semantic
[3] Pragmatic
[4] Discourse
[5] Lexicon
[6] Database
[7] Homograph

واژگان توسط این ساختار داده سبب فراهم‌آمدن ابزارهای قوی مانند جستجوی جزیی[1] و همچنین سرعت بالا در جستجوی رشته جهت پیمایش رشته می‌شود.

۴- روش پیشنهادی در طراحی تحلیلگر لغوی

هدف از تحلیل لغوی مشخص نمودن مقوله زبانی واژه و مجموعه ویژگی‌های صرفی آن است. دستیابی به این هدف نیازمند دانش وسیعی در محدوده صرف زبان می‌باشد. این دانش عموماً به کمک یکسری دانش رویه‌ای[2] که به شکل قوانین نمود پیدا می‌کند و دانش تشریحی[3] که در غالب واژگان ارائه می شود، فراهم می آید. با توجه به محدودیت‌های دانش رویه‌ای در ارائه دانش زبانی، روش‌های جدید همگی به سوی استفاده گسترده‌تر از دانش تشریحی که به شکل واژگان و پیکره‌های بسیار بزرگ نمود پیدا می‌کند، سوق داده شده‌اند.

تحلیل لغوی با صرف تصریفی در ارتباط است. از آنجا که صورت‌های صرفی هر واژه بسیار متنوع و زیاد است، در نتیجه امکان ذخیره آن در واژگان امکان‌پذیر نیست و کارایی لازم را به‌دنبال نخواهد داشت. خوشبختانه قوانین حاکم بر صرف تصریفی در زبان ها عمومی[4] بوده و کمتر استثناپذیر است. به‌طور کلی، برای تحلیل یک واژه صرف شده، از یکسری قوانین که به‌طور متوالی بر واژه‌های زبان اعمال می‌شود، استفاده می‌گردد. اشکال این روش آن است که ترتیب و نوع اعمال قوانین انتخاب شده برای تحلیل واژه‌ها اهمیت دارد. اعمال نادرست قوانین و یا حتی ترتیب اشتباه سبب در بهترین حالت سیی ناتوانی سیستم در تشخیص واژه و تحلیل آن می‌شود. علاوه بر این، جمع آوری قوانین و مشخص نمودن ترتیب مناسب جهت اعمال آنها، کاری پر زحمت و همراه با خطای فراوان است. در روش‌های سنتی زمانی تحلیل یک واژه به پایان می‌رسد که کلیه قوانین بر روی آن اعمال شده باشد. در بسیاری از موارد این کار سبب هدر رفتن زمان و منابع پردازشی می‌شود و حتی ممکن است در بعضی از موارد سبب ناپایداری سیستم به‌دلیل وجود قوانین زنجیره‌ای گردد. از سوی دیگر، درصورتی‌که شرط پایان یافتن تحلیل تنها یک جواب و یا به عبارت دیگر اولین جواب باشد، ممکن است قسمتی از اطلاعات از دست برود چراکه در بسیاری از موارد مانند واژه‌های هم‌نویسه یک واژه می‌تواند دارای چندین شکل صرفی و غیر صرفی باشد. به عبارت دیگر، روش‌های مبتی بر قانون روش‌هایی فرضیه محوراند که از روش‌های بالا به پایین جهت تحلیل استفاده می‌کنند.

در اینجا و در روش پیشنهاد شده برای تحلیل‌گر لغوی، از نحوه بازنمایی واژگان در درخت جستجوی سه‌تایی و دسته بندی ارائه شده برای مقوله های زبانی جهت انجام تحلیل لغوی استفاده می کند. درخت جستجوی سه‌تایی ابزار قدرتمندی را برای پیمایش رشته در اختیار کاربر قرار می دهد. روش پیشنهاد شده در این پژوهش از جستجوی جزیی و یک پایگاه دانش کوچک برای هدایت مسیر جستجو، جهت پردازش لغوی استفاده می شود. روش پیشنهاد شده، در مقابل روش‌های قبلی، روشی داده-محور است. به عبارت دیگر پردازش لغوی واژه به مسئله پیمایش هدایت شده بر روی درخت جستجوی سه تایی تبدیل شده است.

ورودی سیستم تحلیلگر لغوی یک رشته از حروف زبان، در اینجا فارسی، است. درصورتی‌که رشته حروف واردشده در سیستم، در زبان دارای معنی بوده و ریشه آن در واژگان ذخیره شده باشد با این تفاوت که شکل واژه ذخیره‌شده به دلیل تصریف تغییر یافته باشد، تحلیلگر لغوی ، رشته حروف ورودی را پردازش خواهد نمود و نوع واژه و ویژگی‌های صرفی مرتبط با آن را مشخص می‌نماید.

با بررسی صورت گرفته بر روی واژه‌های زبان فارسی، به این نتیجه رسیدیم که جایگاه هر یک از اجزا از مقوله های زبانی در یک ترکیب که تشکیل یک واژه صرف شده جدید را می دهند، ثابت است. می‌توان پس از یک واژه چندین جایگاه متفاوت در نظر گرفت. در هر یک از این جایگاه‌ها، تنها وندهای خاصی می‌توانند قرار گیرد. علاوه بر این‌که یکسری روابط درون گروهی بین مقوله‌های واژگانی که در درون یک ترکیب ظاهر می‌شوند وجود دارد.

ساختار درخت سه‌تایی امکان جستجوی جزیی را فراهم می‌آورد. به‌عنوان مثال اگر در واژگان واژه‌های 0,1,01,10 معرفی شده باشد آنگاه سیستم پس از مشاهده رشته 0110 متواند نتایج زیر را تولید کند: 0 , 1 , 1 , 0 و 10 , 1 , 0 و 0 , 1 , 01 و 01 , 10 . با زیادشدن تعداد واژه‌های زبان و افزایش طول رشته ورودی برای جستجوی جزیی، تعداد پیشنهادات سیستم و زمان پردازش به‌شدت افزایش می‌یابد. به عبارت دیگر، سیستم با یک فضای جستجوی بسیار بزرگ مواجه می‌شود. از میان پیشنهادات ارائه‌شده تنها یک یا چند مورد از آنها مورد قبول و مورد نظر است. افزایش فضای جستجو سبب افزایش زمان جستجو، ارائه جواب‌های اشتباه و در نهایت ناکارآمدی سیستم می‌شود. برای کاهش فضای جستجو از قوانین شهودی[5] استفاده می‌شود. این قوانین به‌عنوان یک راهنما جهت محدود کردن فضای جستجو استفاده می‌شود. در سیستم پیشنهادی این کار به‌وسیله قوانین صرف انجام می‌شود. برای این منظور قوانین پیشنهادی می‌بایستی جامع و مانع باشند. مشکل بزرگ این مرحله، جمع‌آوری قوانین و حصول اطمینان از جامع و مانع بودن آنهاست. نحوه ارائه قوانین و اعمال آن، بر روی کارآیی سیستم تاثیر به سزایی دارد. در سیستم پیشنهادی، این قوانین توسط دو ماتریس ارائه شده

[1] Partial Match
[2] Procedural Knowledge
[3] Declarative Knowledge
[4] General
[5] Heuristic

که قوانین صرفی زبان به سادگی در آنها رمز شده است. این دو ماتریس، ماتریس‌های *محدودیت دوطرفه*[1] و *محدودیت ترتیبی*[2] نام گرفته اند. جدول شماره ۳ و ۴ قسمتی از این دو ماتریس را نمایش می‌دهد.

جدول ۳- قسمتی از ماتریس محدودیت دوطرفه

ی نکره	شناسه فعلی	جز پیشین فعلی	علامت جمع	
1	0	0	1	**اسم**
1	0	0	1	**صفت**
1	0	0	1	**اعداد اصلی**
1	0	0	1	**ممیزها**
1	0	0	1	**شاخص‌ها**
0	0	0	1	**ضمیر**

همانطور که گفته شد، وجود یک مقوله زبانی در ساختمان یک کلمه سبب می شود تا یک محدودیت برای حضور مقوله های زبانی دیگر بوجود بیاید. به‌عنوان مثال درصورتی‌که در ترکیبی شناسه فعلی وجود داشته باشد آن‌گاه امکان حضور "ی" نکره در آن ترکیب نخواهد بود و بالعکس. مثالی دیگر برای زبان انگلیسی، حضور نشانه گذشته "ed-" در انتهای یک فعل است که مانع از حضور نشانه استمرار "ing-" در آن ترکیب می‌شود. ماتریس *محدودیت دوطرفه* همین واقعیت را نمایش می‌دهد. سطرها و ستون‌های این ماتریس با مقوله‌های زبانی نام‌گذاری شده‌اند. سطر و ستون متناظر با هر زوج مقوله زبانی نشان می‌دهد که آیا این دو مقوله می‌توانند با یکدیگر در یک ترکیب وجود داشته باشند یا خیر.

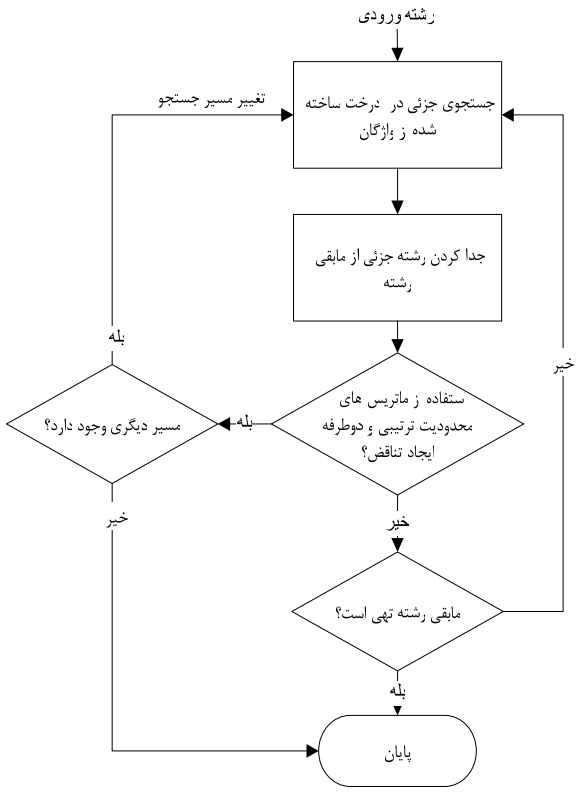
جدول ۴- قسمتی از ماتریس محدودیت دو طرفه

ی نکره	نشانه دعایی	شناسه فعلی	علامت جمع	
1	0	0	0	**علامت جمع**
1	0	0	1	**اعداد ترتیبی**
0	1	0	0	**شناسه فعلی**
0	0	0	1	**نکره**
0	0	0	0	**الف ندا**
0	1	1	0	**بن مضارع**
0	1	1	0	**بن ماضی**

ترتیب ظهور هر یک از اجزای ترکیب در واژه سازی مهم است.

[1] Mutual restriction
[2] Order Restriction

به‌عنوان مثال هیچگاه نمی‌توان شناسه را قبل از فعل آورد. این همان چیزی است که در سیستم‌های مبتنی بر قانون، به کمک ترتیب اعمال قوانین مورد توجه قرار می‌گرفت. در سیستم پیشنهادی چنین قوانینی در ماتریس *محدودیت ترتیبی* مدل‌سازی شده است. در اینجا نیز سطرها و ستون‌های ماتریس با مقوله‌های زبانی نام‌گذاری شده است. سطر و ستون متناظر با هر زوج مقوله زبانی نشان می‌دهد که آیا مقوله متناظر با یک سطر می‌تواند پس از مقوله زبانی مربوط به یک ستون بیاید یا خیر.



شکل ۱- بلوک دیاگرام الگوریتم پیشنهادی

روش کار بسیار ساده است. هنگامیکه یک رشته وارد سیستم می شود، رشته ورودی بر روی درخت جستجوی سه‌تایی ساخته شده از واژگان به شکل جزئی پیمایش می‌شود. هرگاه در این پیمایش جزئی، یک قسمت از واژه با یکی از واژه‌های واژگان سیستم نظیر شود، از مابقی رشته ورودی حذف می شود و به بقیه رشته های جزئی پیدا شده اضافه می شوند. در هر قدم پیش از اضافه شدن رشته جزئی به بقیه رشته های جزئی، از ماتریسهای محدودیت دوطرفه و محدودیت ترتیبی برای اطمینان از عدم تناقض استفاده می شود. در صورت عدم وجود تناقض برای مابقی رشته در واژگان جستجو می‌شود تا هنگامی که به انتهای رشته ورودی برسد و یا با یک رشته پوچ مواجه شود. در صورت وجود تناقض، امکان پیمایش درخت توسط

مسیرهای دیگر چک می شود. در صورت عدم وجود مسیر دیگر جهت پیمایش، کار پایان می پذیرد. بدیهی است که هریک از رشته های جزیی با یک مقوله زبانی مشخص شده اند. همانطور که پیشتر گفته شد، از این مقوله ها و ماتریسهای محدودیت دو طرفه و محدودیت ترتیبی برای اطمینان از عدم وجود تناقض استفاده می شود.به عبارت دیگر، مسئله تحلیل صرفی به مسئله پیمایش درخت سه تایی تبدیل شده است. بلوک دیاگرام الگوریتم در شکل ۱ دیده می‌شود.

این الگوریتم به صورت بازگشتی پیاده سازی می شود و در حقیقت چک کردن قوانین صرفی در هنگام فراخوانی مجدد و با کمک ماتریسهای تعبیه شده صورت می گیرد. شکل زیر نمونه‌ای از خروجی سیستم برای دو واژه "کودکانشان" و "می‌زدندشان" را نمایش می‌دهد.

1	1	1	کودک	اسم
1	1	2	ان	علامت جمع ان
1	1	3	شان	ضمير ملكي

1	1	1	می	جزپیشین فعلي
1	1	2	زد	بن ماضي
1	1	3	ند	شناسه فعلي
1	1	4	شان	ضمير مفعولي

شکل ۳- نمونه خروجی سیستم برای دو واژه "کودکشان" و "می‌زدندشان"

۶- ارزیابی عملی

سیستم پیشنهادشده بر روی روزنامه شرق آنلاین در تاریخ ۸۴/۱/۲۹ آزمایش شد. پیکره مورد آزمایش مشتمل بر ۹۴۴۱۵ واژه کاربردی[1] است. برای انجام این آزمایش از یک واژگان مشتمل بر ۲۲۰۰۰ واژه استفاده شد. در ۸۵٪ موارد واژه‌های پیکره عیناً در واژگان وجود داشت. ۱۲٪ واژه‌ها در پردازش لغوی تشخیص داده شد و تنها در ۳٪ موارد سیستم قادر به تشخیص واژه ورودی نبود. جدول شماره ۵ به‌طور خلاصه نتیجه آزمایش را با در نظر گرفتن واژه های تکراری نمایش می‌دهد. از سوی دیگر، واژه‌های استفاده شده در این پیکره مشتمل بر ۹۰۵۸ واژه نماینده[2] است. ۵۹۵۰ واژه از واژه‌های نماینده عیناً در واژگان وجود داشت. تحلیل گر لغوی ۲۳۸۹ واژه از مابقی تعداد واژه‌های نماینده را تشخیص داد و ۷۱۹ واژه باقیمانده دیگر را تشخیص نداد. به عبارت دیگر، در ۷۱٪ موارد واژه عیناً در واژگان موجود بوده است. ۱۹٪ واژه‌ها توسط تحلیلگر لغوی تشخیص داده شده و در ۸٪ موارد، واژه وارد شده تشخیص داده نشده است. جدول شماره۶ خلاصه این نتایج را نمایش می دهد.

جدول ۵- نتایج حاصل با احتساب واژه‌های تکراری

		تعداد	درصد
واژه تشخیص داده‌شده	عیناً موجود در واژگان	80253	85
	صرف شده	11330	12
کلمات تشخیص داده نشده	با شکل غیرصرفی	2002	0.022
	صرف‌شده	830	0.008
تعداد کل واژه‌ها (با احتساب واژه‌های تکراری)		94415	

جدول ۶- نتایج حاصل بدون احتساب واژه‌های تکراری

		تعداد	درصد
واژه تشخیص داده‌شده	عیناً موجود در واژگان	6502	0.73
	صرف‌شده	1756	0.19
واژه تشخیص داده‌نشده	با شکل غیرصرفی	596	0.07
	صرف شده	204	0.03
تعداد کل واژه‌ها (بدون احتساب واژه‌های تکراری)		8854	

میزان پوشش سیستم در تشخیص واژه‌های واردشده با احتساب واژه‌های تکراری برابر ۹۷٪، و بدون احتساب واژه‌های تکراری در حدود ۹۳٪ بوده است.

از واژه‌های تشخیص داده‌نشده، با احتساب واژه‌های تکراری، ۸۳۰ مورد از واژه ها از ۲۸۳۲ مورد واژه تشخیص داده نشده شکل صرفی داشته اند. بدون احتساب واژه‌های تکراری، ۲۰۴ مورد از ۸۰۰ واژه تشخیص داده‌نشده، شکل صرفی داشته اند. فرمول شماره۱ یک رابطه را برای ارزیابی دقت سیستم نمایش می دهد.

دقت سیستم = تعداد واژه های صرف شده در پیکره/ تعداد واژه های صرف شده تشخیص داده شده توسط سیستم — فرمول(۱)

مطابق این رابطه دقت سیستم با احتساب واژه‌های تکراری برابر ۹۳٪، و در صورت عدم احتساب واژه‌های تکراری برابر ۸۹٪ ارزیابی می‌شود. افزایش دقت به راحتی با اضافه نمودن ریشه واژه‌های تشخیص داده‌نشده امکان‌پذیر است.

۷- جمع‌بندی و کارهای آینده

در این مقاله روشی نوین جهت تحلیل لغوی واژه‌های زبان فارسی به کمک هماهنگی میان روش بازنمایی واژگان و الگوریتم پیمایش رشته با استفاده از ساختمان داده درختی ارائه شده است. در اینجا با استفاده از یک پایگاه دانش کوچک ارائه‌شده در غالب دو ماتریس که از قوانین صرف زبان فارسی و چگونگی همنشینی مقوله‌های زبانی به‌دست آمده است، جهت هدایت پیمایش بر روی یالهای درخت

[1] Token
[2] Type

استفاده شده است. به عبارت دیگر، مسئله پردازش لغوی به مسئله پیمایش هدایت‌شده بر روی درخت جستجوی سه‌تایی تبدیل شده است. نتایج حاصل نمایانگر کارآمدی این روش برای تحلیل لغوی واژه‌های زبان فارسی است.

یکی از نقاط قوت روش ارائه شده، سادگی استفاده از روش پیشنهادی برای دیگر زبان‌های هندی و اروپایی است که در آنها تصریف واژه‌ها از طریق وندافزایی صورت می‌گیرد. مزیت دیگر این روش، سادگی جمع‌آوری دانش و ارائه آن می‌باشد بطور کلی ساختار پیشنهادی یک چارچوب کلی برای جمع آوری دانش صرفی زبان ارائه می نماید. سرعت بالا در پردازش، ارائه کلیه حالت‌های صرفی واژه به‌خصوص در مورد واژه‌های هم‌نویسه از دیگر مزیت های روش پیشنهاد شده است. در روش‌های قبلی برای تحلیل لغوی واژه های فارسی ، عموماً قوانین کاربردی برای هر واژه براساس مقوله صرفی آن واژه مشخص می شود و بدین ترتیب تحلیل صرفی برای واژه ها به دو دسته تحلیل واژه های فعلی و غیر فعلی تقسیم بندی شده است. انتخاب قوانین براساس نوع وند ظاهرشده در واژه ها صورت می‌گرد و درصورتی‌که بخواهیم همگی صورت‌های ممکن برای تحلیل لغوی یک واژه را داشته باشیم، نتیجه عمل تحلیل در به‌کار بردن همگی قوانین و ترتیب مختلف آنها، سبب افزایش زمان تحلیل و افزایش حجم پردازش است. بر خلاف روشهای معمول، سیستم پیشنهادی نیاز به فرضیات اولیه جهت شروع به کار ندارد.

پژوهش صورت گرفته تنها در محدوده صرف تصریفی برای واژه های فارسی انجام شده است اما به راحتی امکان استفاده از چارچوب ارائه شده برای صرف اشتقاقی نیز وجود دارد. در این حالت کافی است تا سطر و ستون مربوط به وندهای اشتقاقی در زبان فارسی، به ماتریس های محدودیت ترتیبی و محدودیت دوطرفه اضافه شود.

یکی از مشکلات مهم که در طول کار با آن مواجه شدیم، عدم وجود استاندارد برای متون الکترونیکی فارسی است. این مشکل چه از نظر کد کارکترهای استفاده‌شده برای نمایش حروف و چه در نحوه تایپ و رعایت فاصله در مرز برونی واژه‌ها و نیم‌فاصله‌ها در مرز درونی واژه به چشم می‌خورد. شاید یک قدم بسیار مهم در این جهت، ارائه چارچوبی هماهنگ برای چگونگی بازنمایی متون الکترونیکی فارسی و ارائه استانداردی فراگیر باشد.

برای ادامه کار، استفاده از روش‌های یادگیری ماشین جهت تنظیم ماتریس‌ها و امکان رتبه بندی نتایج ممکن در نظر گرفته شده است. امید است که با این کار بتوان از واژه‌های هم‌نویسه در متون خاص، همزمان با تحلیل لغوی رفع ابهام صورت گیرد. همچنین، گسترش حجم واژگان جهت پوشش بیشتر از زبان مد نظر است. علاوه بر این‌که بسط و گسترش سیستم جهت تحلیل اشتقاقی واژه‌های فارسی، یکی از کارهای مد نظر در آینده می باشد.

مراجع:

[۱] ایران کلباسی، **ساخت اشتقاقی واژه در فارسی امروز**، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ۱۳۸۰.

[۲] کشانی، خسرو، **اشتقاق پسوندی در زبان فارسی امروز**، مرکز نشر دانشگاهی، تهران، ۱۳۷۱.

[3] Kuhn, J. *Compounding and derivational morphology in a finite-state setting*, ACL, 2003.

[4] Attia, M. A. *Developing a robust Arabic morphological transducer using finite state technology*, 8th Annual CLUK Research Colloquium, Manchester, 2005.

[5] Vaillette, N. *Logical Specification of Finite-State Transductions for Natural Language Processing* PhD Thesis, Ohio University, 2004.

[6] Radek, S. Pavel, S. *Automatic processing of Czech inflectional and derivative morphology*, FIMU-RS-2001-03, 2001.

[7] Riazati, D. *Computational Analysis of Persian Morphology*, MS Thesis, Department of Computer Science, RMIT, 1997.

[8] Yamashita, T. Matsumoto, Y. *Language independent morphological analysis*, In Proceedings of 6th Applied Natural Language Processing Conference, pp.232-238, 2000.

[9] Bosch, A. Daelemans, W. , *Memory-Based Morphological Analysis*, ACL , 1999.

[10] Snover M. G. Brent, M. R. *A probabilistic model for learning concatenative morphology*, NIPS, 2002.

[11] Goldsmith J. A. *Unsupervised learning of the morphology of a natural languag*e, Computational Linguistics 27:2 pp. 153-198, 2000.

[12] Sgarbas, K. N. Fakotakis, N. D. Kokkinakis, K. G. *A straightforward approach to morphological analysis and synthesis*, Proc. COMLEX , Greece, 2000.

[13] Megerdoomian, K. *Unification-Based Persian Morphology*, Proceedings of the CICling, Mexico, 2000.

[14] Evan A. L. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*, Occasional Publications in Academic Computing No. 16. Dallas, TX: Summer Institute of Linguistics, 1990.

[15] Megerdoomian, K. *Finite-State Morphological Analysis of Persian*, COLING, 2004.

[۱۶] انوری، احمدی گیوی، **دستورزبان فارسی ۲** (ویرایش دوم)، انتشارات فاطمی، تهران، ۱۳۸۲.

[۱۷] طباطبائی، علاالدین، **اسم و صفت مرکب در زبان فارسی**، مرکز نشر دانشگاهی، تهران، ۱۳۸۲.

[۱۸] ماهوتیان، شهرزاد، **دستور زبان فارسی از دیدگاه رده شناسی**، ترجمه سمائی، مهدی، نشر مرکز، تهران، ۱۳۸۲.

[۱۹] دبیرمقدم، محمد، **زبان‌شناسی نظری پیدایش و زایش دستور زایشی (ویراست دوم)**، سازمان مطالعه و تدوین کتب علوم انسانی دانشگاهها(سمت)، تهران، ۱۳۸۳.

[۲۰] مشکاۀالدینی، مهدی، **دستور زبان فارسی بر پایه نظریه گشتاری**، انتشارات دانشگاه فردوسی مشهد، ۱۳۸۲.

[21] Bentley J. L. Sedgewick, R. *Fast algorithms for sorting and searching strings*, Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, 1997.