Iranian Association of Information and Communication Technology
IKT 2013
Shiraz University
Regular Paper

بهبود کارآیی تشخیص خودکار مقولهٔ نحوی کلمات از طریق بکارگیری دانش تصریف و اشتقاق

بصیرا خاکی[۱]، مریم گیلی[۲]، بهرام وزیرنژاد[۳]، هادی عبدی قویدل[۴]

[۱] دانشگاه صنعتی شریف، basira@mehr.sharif.ir

[۲] دانشگاه صنعتی شریف، maryam_gili@mehr.sharif.ir

[۳] دانشگاه صنعتی شریف، bahram@sharif.ir

[۴] دانشگاه صنعتی شریف، hadi_abdighavidel@mehr.sharif.ir

چکیده

تشخیص خودکار مقولهٔ نحوی کلمات، یکی از پایه‌ای‌ترین امور در علم پردازش زبان طبیعی است. انجام این فرایند در زبان فارسی، تاکنون به واسطهٔ الگوریتم‌های مختلفی صورت گرفته است. در مقالهٔ حاضر با تمرکز بر روی یکی از این روش‌ها بنام تی‌ان‌تی سعی شده روش تشخیص را به اطلاعات تصریف و اشتقاق از حوزهٔ صرف مجهز سازیم. تی‌ان‌تی یا برچسب دهی بر اساس دنباله‌های سه-نگاشتی، یکی از معروف‌ترین روش آموزش پیکره‌های برچسب‌خورده است که برای اکثر زبان‌های جهان قابل اجرا می‌باشد. بدین ترتیب در مرحلهٔ نخست پژوهش حاضر، پیکرهٔ متنی برچسب‌خوردهٔ زبان فارسی توسط روش فوق‌الذکر آموزش داده شده است. سپس اطلاعات تصریف، وندهای دستوری چسبیده به آخر کلمات را به تفکیک مقولهٔ نحوی در اختیار مرحله آموزش قرار داده و از نقطه نظر اشتقاق اطلاعات نحوی شامل ترکیبات اسمی، صفتی و ... فراهم آورده شده است. در نهایت پس از ارزیابی، صحت۹۷،۶ درصد برای کلمات داخل واژگان و ۹۵،۶۹ درصد برای کلمات تصریفی و اشتقاقی خارج از واژگان به دست آمد که اطلاعات ترکیبی لایه‌های زبان (در مقاله حاضر صرف و نحو) تاثیر به‌سزایی در افزایش صحت این ابزار پایه‌ای در مقوله پردازش زبان طبیعی بر جای گذاشت.

واژه‌های کلیدی

تشخیص خودکار مقولهٔ نحوی، تصریف، اشتقاق

۱- مقدمه

یکی از ضروری‌ترین بخش‌هایی که در پردازش زبان طبیعی[1] به آن توجه می‌شود، تعیین مقوله نحوی کلمات می‌باشد. مقوله نحوی کلام (که معمولاً طبقه واژگان، طبقه لغوی یا مقوله واژگانی نیز تعریف می‌شود) در دستور زبان یک مقوله زبان‌شناختی از کلمات می‌باشد که عموماً تحت عنوان رفتار نحوی یا صرفی واژگان تعریف می‌شود. از جمله مقوله‌های واژگانی می‌توان اسم، فعل، صفت، حرف اضافه، قید و غیره را نام برد که رایج‌ترین این مقولات در بین تمام زبان‌ها، اسم و فعل می‌باشد.

طبقه‌بندی زبان‌ها به مقوله‌های واژگانی برای نخستین بار در دستور زبان سانسکریت صورت گرفت که چهار مقوله اسم، فعل، پیشوند و حروف تعریف را برای واژگان تعریف کردند. پس از آن در پایان قرن دوم، مقوله-های واژگانی به نه گروه اسم، فعل، صفت، حرف تعریف، حرف ندا، ضمایر، حروف اضافه، قیود و حروف عطف طبقه‌بندی شدند.

از آنجا که کلمات در زبان حاوی اطلاعات دستوری از قبیل شخص، شمار و جنسیت می‌باشند و هر کلمه ممکن است دارای مقوله‌های نحوی متفاوتی باشد، یکی از چالش‌های برچسب‌زن نحوی[2] در پردازش زبان طبیعی تعیین دقیق مقوله نحوی اجزای کلام است. به عمل انتساب برچسب واژگانی مانند اسم، فعل، صفت، قید و ... به کلمات و نشانه‌های تشکیل‌دهنده یک متن، برچسب‌دهی نحوی می‌گویند به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جمله باشند. اطلاعات ورودی در این سیستم یک متن و خروجی آن کلمات همراه با برچسب مورد نظر می‌باشد.

[1] Natural language processing(NLP)

[2] POS(Part Of Speech) tagger

رویکردهای مورد استفاده در طراحی برچسب‌زن‌ها عموماً به سه گروه تقسیم می‌شوند که عبارتند از: رویکرد آماری، رویکرد مبتنی بر قاعده و رویکرد گشتاری.

در رویکرد آماری، ابتدا برچسب‌زن توسط یک پیکره برچسب خورده آموزش داده می‌شود و احتمال هرکلمه یا برچسب یا توالی برچسب‌ها محاسبه شده و پس از آن وقتی که یک داده برای برچسب‌دهیبه آن داده می‌شود برچسب‌زن بر اساس بیشترین احتمال، برچسب کلمه مورد نظر را تعیین می‌کند. روش‌های به کار رفته در رویکرد آماری عبارتند از: مدل مخفی مارکوف، ماکزیمم آنتروپی وسیستم مبتنی بر حافظه.

روش‌های مبتنی بر قاعده که بر دانش انسانی استوارند، حاوی پایگاه داده بزرگی از قواعد دستوری می‌باشند. در این روش برچسب‌زن فرضیه‌ای می‌سازد و بر اساس قواعد پایگاه دادهٔ خود بهترین برچسب را انتخاب می‌کند. یکی از مزایای این روش این است که به پیکره برچسب خورده برای آموزش ماشین نیاز نمی‌باشد. علاوه بر آن، آموزش برچسب‌زن می‌تواند بصورت وابسته به حوزه خاص باشد. با این توضیح که به کارگیری یک برچسب‌زن برای متن یک حوزه دیگر معمولاً با دقت کمتری همراه است.

رویکرد سوم ترکیبی از دو رویکرد مبتنی بر قاعده و آماری می‌باشد که بهترین نتایج را در بر داشته است.

برچسب‌زن نحوی در حوزه‌های مختلف پردازش زبان طبیعی به کار می‌رود. برای مثال در سیستم‌های تبدیل متن به گفتار[3] برای مقاصد مختلفی از جمله تحلیل صرفی و رفع ابهام کلمات با املای یکسان و در زبان فارسی برای یافتن کسره اضافه که در املا نوشته نمی‌شود کاربرد دارد. در سیستم سنتز عروض[4] برای دیرش و آهنگ مدل و همچنین تخمین زیر و بمی استفاده می‌شود. در ابهام‌زدایی معنایی به منظور رفع ابهام از کلماتی که علاوه بر برچسب اصلی خود برچسب‌های دیگری نیز می‌گیرند می‌توان از برچسب‌زن استفاده نمود. علاوه بر این موارد، موارد دیگری از کاربردهای برچسب‌زن نحوی وجود دارد که عبارتند از: بازشناسی عبارت، شرح و تفسیر خودکار و آنالیز صرفی ترجمه ماشینی، آموزش زبان، بازیابی اطلاعات و بسیاری دیگر.

در مقالهٔ حاضر، سعی بر آن است نخست با مرور بر روش‌های پیشین تشخیص مقوله نحوی اجزای کلام، دادگان و روش کلی برچسب‌زنی و همچنین روش خاص مواجهه باکلمات تصریفی و اشتقاقی خارج از واژگان را به تفصیل شرح داده و در نهایت به ارزیابی نتایج بپردازیم.

۲- پیشینه

فعالیت‌های شکل گرفته در زمینهٔ برچسب‌زنی برای زبان فارسی در مقایسه با زبان‌های دیگر اندک بوده است که در این بخش به مرور بر برخی از فعالیت‌های انجام شده در زبان فارسی پرداخته می‌شود.

موسوی میانگاه[۴] در مقالهٔ " ابهام‌زدایی اجزای کلام مبتنی بر پیکره در زبان فارسی" از احتمالات کلاس کلمه در پیکرهٔ آموزشی کوچک استفاده کرده که به طور خودکار متون فارسی را برچسب می‌زند. این آزمایش در دو سطح ابهام‌زدایی تک‌نگاشت و دونگاشت انجام گرفته است. مقایسهٔ نتایج به دست آمده از این دو روش حاکی از آن است که استفاده از بافتی که کلمه به آن بافت تعلق دارد، دقت سیستم را به مقدار قابل توجهی افزایش می‌دهد.

عاصی و حاج عبدالحسینی [۵] کار برچسب‌گذاری را بر پایهٔ روش شوتز[5] انجام داده‌اند. در این روش همسایه‌های یک کلمه در دو بردار زمینه چپ و زمینه راست جمع شده‌اند و سپس بر طبق شباهت توزیعی انواع کلمات دسته‌بندی می‌شوند و در نهایت این دسته‌ها برچسب می‌خورند. این سیستم عمل ابهام‌زدایی را از برچسب‌های کلمات انجام نمی‌دهد، به علاوه کلماتی که فراوانی کم دارند را برچسب‌گذاری نمی‌کند و از دقت کمی برای صفات و قیدها برخوردار است.

رجا و همکاران [۶] ازچند برچسب‌زن بر روی پیکرهٔ زبان فارسی استفاده کرده‌اند و نتایج به دست آمده را ارایه داده‌اند. در این مقاله از ۴۰ برچسب استفاده شده است و دقت به دست آمده ۹۷٪-۹۴٪ بوده است.

محسنی و همکاران [۷]سیستم برچسب‌زن اجزای کلام که مبتنی بر مدل مارکوف مرتبه اول است بر روی نسخهٔ قدیمی پیکرهٔ آرسی‌آی‌اس‌پی[6] به کار بستند. نتایج این سیستم در دسته‌های بزرگ کلمات فارسی گزارش شده است.

موسوی میانگاه و علی دلور خلفی [۸]روش بدون نظارت را برای ابهام‌زدایی اجزای کلام ارایه داده‌اند که برای زبان فارسی به کار گرفته شده است. این روش، مدل بازخورد بهبودیافتهٔ تکراری[7] نام دارد که مدلی شهودیاست. در خلال فرایند برچسب‌زنی این الگوریتم از میان چندین تکرارعبور می‌کند تا از هر کلمه که مطابق با سطوح تحلیل چند نگاشتی است، رفع ابهام کند.

۳- روش کار

۳-۱- دادگان و پیش‌پردازش

۳-۱-۱- پیکره متنی برچسب‌خورده زبان فارسی

در این مقاله از نسخه ۲ پیکرهٔ متنی برچسب‌خورده زبان فارسی [۹] استفاده شده است که دارای۱۶ عدد برچسب مقولات اصلی می‌باشد. این برچسب‌ها شامل اسم، فعل، صفت، قید و غیره می‌باشند. ساختار این برچسب‌ها به صورت سلسله مراتبی است که اصلی‌ترین برچسب در قسمت راست قرار گرفته و از راست به چپ جزئیات و اطلاعات دقیق‌تر افزوده می‌شود. این پیکره دارای ۷،۵میلیون کلمه برچسب‌خورده است. لازم به

[3] Test-to-speech system(TTS)

[4] Prosody synthesis system

[5] Schutze

[6] RCISP

[7] Iterative improved feedback

ذکر است که در مقاله حاضر به برچسب‌های اصلی بسنده کرده و از جزئیات آن تا حدی که اطلاعات برچسب اصلی را دچار ابهام نکند، صرف نظر می‌کنیم.

دادگان دیگر مورد استفاده در این مقاله، واژگان زایای زبان فارسی [۱] است. این مجموعه واژگان شامل بیش از شصت هزار کلمه همراه با ویژگی های زبان‌شناختی و صرفی آنهاست. کلمات نهفته در این واژگان شامل تقریباً تمامی تکواژهای رایج زبان فارسی اعم از تکواژهای آزاد و مقید می باشد.

۳-۱-۲- استانداردسازی پیکره

قبل از اینکه بتوان از پیکرهٔ مذبور در ساخت برچسب‌زن نحوی تحقیق حاضر استفاده کرد، باید ابتدا پیش‌پردازشی روی آنها انجام گیرد تا صورت های غیراستاندارد[۸] موجود در پیکره به شکل استاندارد تبدیل گردند. به عنوان مثال اگر حروف، نشانه‌های نگارشی و کلمات فارسی به شکل یکسانی نوشته نشوند، متون مورد استفاده توسط سامانه‌های رایانه‌ای قابل تحلیل نخواهند بود. طی فرایند استانداردسازی، کلیه علایم نگارشی، حروف، فاصله‌های بین کلمات، اختصارات و غیره بدون ایجاد تغییرات معنایی در متن به شکل استاندارد تبدیل می‌گردند. در این مقاله از استانداردساز متنی تهیه شده در آزمایشگاه پردازش گفتار و زبان دانشگاه صنعتی شریف استفاده شده است.

۳-۲- الگوریتم تی‌ان‌تی

۳-۲-۱ شالودهٔ مدل

ابزار تی‌ان‌تی[۱۰] برای فرایند تشخیص خودکار مقولهٔ نحوی کلمات، از مدل مارکوف مرتبه دوم استفاده می‌کند. حالت‌ها در این مدل نمایانگر برچسب‌ها و خروجی آن‌ها نمایانگر کلمات می‌باشند. محاسبه احتمال گذر از حالتی به حالت دیگر، وابسته به حالت‌های موجود در هر دو طرف است. احتمال خروجی از سوی دیگر وابسته به جدیدترین مقوله است. برای واضح شدن این مطلب، فرمول ذیل را برای دنباله‌ای از کلمات به طول T محاسبه می‌کنیم:

$$\mathbf{argmax}_{t_1 \ldots t_T} \left[\prod_{i=1}^{T} P(t_i|t_{i-1}, t_{i-2}) P(w_i|t_i) \right] P(t_{T+1}|t_T) \quad (۱)$$

عناصر مجموعه برچسب‌ها را $t_1 \ldots t_T$ تشکیل می‌دهد. برچسب‌های اضافی نظیر t_{-1}، t_0 و t_{T+1} نشانگر دنباله‌های آغازی و پایانی هستند. استفاده از این برچسب‌های اضافی، نتایج برچسب‌زنی را بهبود می‌دهد.

احتمال گذر و احتمال خروجی بر اساس برچسب‌های موجود در پیکره برچسب‌خورده بدست می‌آید. در گام اول، احتمالات بیشینه P' که بسته به بسامد نسبی است، محاسبه می‌شود.

$$\mathrm{P}'(t_3) = \frac{f(t_3)}{N} \quad \text{(۲)تک نگاشتی}$$

$$\mathrm{P}'(t_3|t_2) = \frac{f(t_2,t_3)}{f(t_2)} \quad \text{(۳)دو نگاشتی}$$

$$\mathrm{P}'(t_3|t_1,t_2) = \frac{f(t_1,t_2,t_3)}{f(t_1,t_2)} \quad \text{(۴)سه نگاشتی}$$

$$\mathrm{P}'(w_3|t_3) = \frac{f(w_3,t_3)}{f(t_3)} \quad \text{(۵)واژگانی}$$

N تعداد کل نشانه‌ها در پیکره آموزشی می‌باشد.احتمال بیشینه زمانی صفر می‌شود که صورت یا مخرج آن فرمول صفر باشد. در گام بعد، بسامدهای بافتی هموارسازی شده و بسامدهای واژگانی با بررسی کلماتی که در واژه‌نامه نیستند، کنترل می‌شود.

۳-۲-۲ هموارسازی

احتمال‌های سه‌نگاشتی حاصل از پیکره عمدتا به خاطر تنک بودن قابل استفاده نیستند. این بدین معناست که نمونه‌های کافی برای هر سه‌نگاشت مهیا نیست تا بتواند احتمال قابل اطمینانی را تخمین بزند. علاوه بر آن، عدم رخداد سه‌نگاشتی در پیکره، تاثیر ناخواسته‌ای بر جای گذاشته و آن این است که احتمال یک دنبالهٔ کامل برای یک دنباله جدید را صفر می-کند.

الگوی هموارسازی برای تی‌ان‌تی، الحاق خطی تک‌نگاشتی، دو نگاشتی و سه‌نگاشتی است. بدین ترتیب، برای احتمال سه‌نگاشتی چنین محاسبه می‌کنیم.

$$\mathrm{P}'(t_3|t_1,t_2) = \lambda_1 \mathrm{P}'(t_3) + \lambda_2 \mathrm{P}'(t_3|t_2) + \lambda_3 \mathrm{P}'(t_3|t_1,t_2) \quad (۶)$$

P'تخمین احتمالات کلیهٔ احتمالات بوده و $\lambda_1 + \lambda_2 + \lambda_3 = 1$است. بنابراین Pتوزیع احتمالات را نشان می‌دهد.

در این بخش، الحاق خطی غیر وابسته به بافت استفاده می‌شود. به عبارت دیگر، تعیین مقادیر λ به سه‌نگاشت خاصی وابسته نیست. این مقادیر بوسیلهٔ الحاق‌های حذف‌شده تعیین می‌گردد. این روش، به طور متوالی هر سه‌نگاشت را از پیکره آموزشی حذف کرده و بهترین مقدار از تمامی چند نگاشتی‌ها را نتیجه می‌دهد.

۳-۳- غلبه بر کلمات تصریفی و اشتقاقی خارج از واژگان

مدل محاسبات آماری محض هرگز نمی‌تواند به عنوان مدل برتر در پردازش زبان باشد. همواره در کنار این محاسبات، باید یافته‌های زبانی نیز حضور داشته باشند تا الحاق بخش‌های زیرین زبان به سامانه‌های پردازش زبان

[8] NSWS

طبیعی به طور کاملا ناآگاهانه و یا به صورت تک‌حوزه‌ای انجام نشود. نحو به تنهایی پاسخگوی شناخت مقولهٔ نحوی کلمات نیست و همواره وابسته به لایه‌های پیشین و پسین خود در زبان است. بدین ترتیب، در این بخش از تحقیق حاضر، با الحاق نحو به صرف بر روی صورت و ساختار کلمات در زبان، فرایند تشخیص رادقیق‌تر کرده و اطلاعاتی از آن که به این فرایند کمک کند را به این مدل اضافه می‌کنیم.

مدل تی‌ان‌تی دارای بخشی مجزا برای مواجهه با کلمات ناآشنا است. ما به جای وابسته نمودن این بخش به پیکره، با استفاده از وندهای شناخته‌شده زبان فارسی و دسته‌بندی آن‌ها برای هر مقولهٔ نحوی و همچنین افزودن استثناهای موجود از طریق شناسایی آن در فرهنگ واژگان زایا مقوله نحوی کلمات تصریفی و اشتقاقی خارج از واژگان در فرایند تشخیص را حدس می‌زنیم.

همانطور که می‌دانیم، مطالعهٔ صورت و ساختار واژگان در زبان در حوزهٔ صرف می‌گنجد و مطالعات صرف به لحاظ تکواژهای وابسته، در دو بخش تصریف و اشتقاق صورت می‌گیرد. لازم به ذکر است که محور مطالعات تصریف و اشتقاق در تحقیق حاضر بر روی پسوندهاست.

۳-۳-۱- تصریف

تصریف واژه‌ها را برای قرار گرفتن در ساختار نحوی آماده می‌کند. در این فرایند، کلمه جدید ساخته نشده و مقولهٔ نحوی دچار تغییر نمی‌شود. از ویژگی‌های مهم وندهای تصریف‌ساز فارسی می‌توان به این موارد اشاره کرد: قاعده‌مندی، محدود بودن تعداد آنها، مختص زبان خاص بودن، پسوند بودن آنها در اکثر موارد و قرار گرفتن بعد از وندهای اشتقاقی از نظر جایگاه.کلیهٔ وندهای تصریفی موجود در زبان فارسی از کتاب شقاقی[۲] برگرفته‌شده و در جدول ۱ آمده است.

جدول ۱: پسوندهای تصریفی

ها ، ان، ات، ون	چشمها، گیاهان، انقلابیون، آزمایشات
تر، ترین	بهتر، بهترین
م، ی، د، یم، ید، ند	خواندم
م، ت، ش، مان، تان، شان	کتابم

۳-۳-۲- اشتقاق

فرایندی که با گذاردن وندها در اول، میان ویاپایان تکواژهای آزاد یا واژه‌ها منجر به افزودن بار معنایی آنها و یا تغییر مقولهٔ نحوی آن‌ها می‌گردد، اشتقاق نام دارد.کشانی[۳] این وندها را دسته‌بندی کرده که در جدول ۲و۳با تمرکز بر پسوندها آورده شده است.

۳-۳-۲- نحوهٔ برچسب‌دهی کلمات تصریفی و اشتقاقی خارج از واژگان

کلمات ناآشنا به کلماتی اطلاق می‌شود که در پیکره آموزشی موجود نبوده و تنها دردادگان بخش آزمون دیده می‌شوند. در مقاله حاضر، به جای وابستگی به اطلاعات آماری پسوندهای برچسب‌های نحوی اصلی در پیکره،از یافته‌های زبان‌شناسی استفاده می‌کنیم تا مقوله نحوی کلمات تصریفی و اشتقاقی خارج از واژگان معلوم گردند. بدین ترتیب، جدول ۱،۲ و ۳ فرایند تشخیص را کامل می‌کنند.

جدول ۲: پسوندهای اشتقاقی صفت و قیدساز

آ	فریبا	فام	سرخ فام
سا	پریسا	گون	زردگون
آسا	معجزه‌آسا	گین	شرمگین
مند	ثروتمند	آگین	زهرآگین
وار	بزرگوار	و	ریشو
وش	ماهوش	باره	شکمباره
آنه	مردانه		
آنی	نورانی		
وی	دنیوی		

۳-۴ جستجوی بیم[9]

جستجوی بیم، زمان پردازش الگوریتم ویتربی را به طور قابل توجهی کاهش می‌دهد. هر حالت با دریافت مقدار δ کمتر از δبزرگتر که بر θ تقسیم شده باشد، از پردازش‌های اضافی آن جلوگیری می‌شود. با اینکه اساس کار الگوریتم ویتربی پیدا کردن آن دسته از دنباله حالت‌هاست که دارای بیشترین احتمال باشد، این امر در مورد جستجوی بیم صادق نیست. اما برای انتخاب صحیح θ تفاوتی میان الگوریتم دارای بیم و بدون بیم ایجاد نمی‌کند. در عمل، θ–1000 معقول است زیرا سرعت فرایند را بدون اینکه بر صحت آن تاثیر بگذارد افزایش می‌دهد.

٤- آزمایش و نتیجه

فرایند ارزیابی تشخیص مقولهٔ نحوی کلمات بر روی ۹۰٪ از کل پیکره به عنوان داده آموزشی و ۱۰٪ به عنوان بخش آزمون می‌باشد. سپس صحت[10] برچسب‌زن با توجه به اینکه چقدر درست عمل‌کرده و چقدر دارای خطا بوده محاسبه می‌شود.

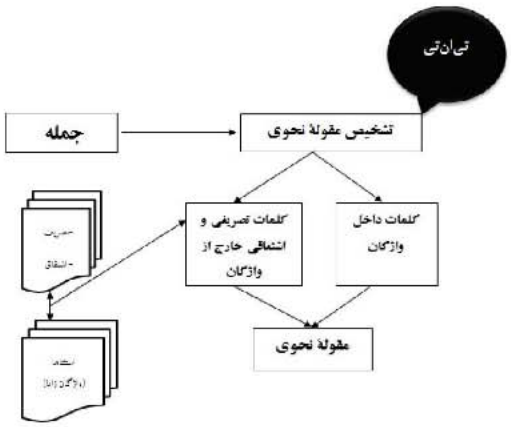
$$\text{صحت} = \frac{\text{تعداد برچسب‌های درست}}{\text{تعداد کل کلمات}} \quad (۷)$$

شمای کلی فرایند تشخیص مقولهٔ نحوی کلمات در شکل ۱ آمده است.

[9] Beam Search
[10] Accuracy

جدول ۳: پسوندهای اشتقاقی اسم‌ساز

	پسوند	مثال
اسم مصدر	آر	گفتار
	ش	آموزش
	ـار	ـنگـار
	گری	وحشی‌گری
	مان	زایمان
	ی	خرابی
	بازی	لاتبازی
اسم فاعل	آر	پرستار
	گار	آموزگار
	کار	رنگکار
	گر	آهنگر
	باز	شطرنج‌باز
	بان	دربان
	نده	فروشنده
	چی	تلفنچی
	ی	شیری
اسم شی	دان	شمعدان
	مان	ریسمان
	ه	دسته
اسم معنی	آ	درازا
	ش	دانش
	یت	محدودیت
	بازی	گدابازی
اسم مکان	نا	تنگنا
	لاخ	سنگلاخ
	بار	جویبار
	زار	گندمزار
	ـار	کوهـار
	دان	کاهدان
	کده	دانشکده
	ستان	کوهستان



شکل۱: شمای کلی فرایند تشخیص مقولهٔ نحوی کلمات

صحت برای کلمات داخل واژگان و کلمات تصریفی و اشتقاقی خارج از واژگان به صورت جدا محاسبه شده و در نهایت در جدول زیر آورده شده است.

جدول۴: درصد فرایند تشخیص

	صحت در تشخیص مقوله کلمات تصریفی و اشتقاقی خارج از واژگان	صحت در تشخیص مقوله کلمات داخل واژگان
پیکره برچسب‌خورده زبان فارسی	٪۹۵،۶۹	٪۹۷،۶

۵- نتیجه‌گیری و بحث

در مقالهٔ حاضر با تمرکز بر روی روش تی‌ان‌تی سعی شد، روش تشخیص را به دو رویکرد تصریف و اشتقاق از حوزهٔ صرف مجهز سازیم. تصریف با اطلاعات وندهای دستوری چسبیده به آخر کلمات و اشتقاق با اطلاعات نحوی ترکیبات اسمی، صفتی و ... در بخش کلمات تصریفی و اشتقاقی کمک کرد تا صحت کلی فرایند تشخیص ۹۷.۶ درصد برای کلمات داخل واژگان و ۹۵.۶۹ درصد برای کلمات تصریفی و اشتقاقی خارج از واژگان رقم بخورد.

در این مقاله، خروج از ذهن آماری محض و ایجاد پلی بین دو لایهٔ زبانی صرف و نحو به عنوان نقش بسیار موثر عمل کرده است.

برای به ۱۰۰٪ رساندن این فرایند، می‌بایست راه حلی برای برخورد با کلمات هم‌نگاره یافت. همچنین، واژه‌بست‌هایی نظیر "م" در کلمه خوبم که برای کلمه خوب دو مقولهٔ نحوی متفاوت نتیجه می‌دهد.خوبم (خوب{صفت} هستم) و (خوب{اسم} من).

امید که نظریه‌پردازان حوزهٔ زبان‌شناسی با تدبیری متفکرانه در صدد حل این مشکلات برآیند.

٦- سپاسگزاری

بر خود لازم می‌دانیم از زحمات بی‌دریغ وراهنمایی‌های ارزنده جناب آقای دکتر محمد بحرانی در به سرانجام رساندن این تحقیق کمال سپاسگزاری را نماییم.

مراجع

[۱] اسلامی، محرم و شریفی، مسعود و علیزاده، صدیقه و زندی، طاهره. "واژگان زایای زبان فارسی"، *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*، صص ۶-۱۱، ۱۳۸۳.

[۲]شقاقی، ویدا.**مبانی صرف**، سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه ها (سمت)، مرکز تحقیق و توسعهٔ علوم انسانی، تهران، ۱۳۸۹.

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

[4] M. Mosavi. "Corpus-based part of speech disambiguation of Persian", ACEEE Int. J. on Information Technology, Mar 2011

[5] S. M. Assi, and M. Haji Abdolhoseini. "Grammatical Tagging of a Persian Corpus."International Journal of Corpus Linguistics, Vol. 5, Number 1, pp. 69-81(13). 2000.

[6] F. Raja, H. Amiri, S. Tasharofi, M. Sarmadi, H. Hojjat and F. Oroumchian. "Evaluation of Part of Speech Tagging on Persian Text." InProceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages Linguistic Institute, Stanford, California, USA, pp. 21-22. 2007.

[7] M. Mohseni, H. Motallebi, B. Minaei-bidgoli and M. Shokrollahi-far. "A Farsi Part-Of-Speech Tagger Based on Markov Model."23rd ACM Symposium on Applied Computing, Brazil. 2008.

[8] M. Mosavia and A. DelvarKhalafi. "Un supervised part of speech tagging for Persian", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[9] M. Bijankhan, J. Seikhzadeghan & M. Bahrani & M. Ghayoomi, "Lessons from Creation of a Persian Written Corpus: Peykare". Language Resources and Evaluation Journal.Vol. 45, No. 2.143-164, 2011.

[10] T. Brants."TNT – a Statistical Part-of-Speech Tagger" In the Proceedings of 6th conference on applied natural language processing (ANLP), USA (2000).