تبدیل متن محاوره‌ای فارسی به رسمی به کمک N_gramها

نادیه آرمین [1]، مهرنوش شمس‌فرد [2]

[1] آزمایشگاه پردازش زبان طبیعی، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی
na.armin@gmail.com

[2] استادیار، آزمایشگاه پردازش زبان طبیعی، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی
m-shams@sbu.ac.ir

چکیده

با افزایش متون محاوره، تبدیل متن محاوره‌ای به رسمی یکی از چالش‌های موجود در پردازش زبان است. در این مقاله به ارائه و آزمون، راهکاری برای تبدیل متون محاوره‌ای به متون رسمی با استفاده از ترکیب روشهای مبتنی بر قاعده و مدل‌سازی آماری می-پردازیم. همچنین چگونگی ساخت پیکره، نحوه توکن‌بندی و نحوه یافتن ریشه کلمات، قوانین تبدیل کلمات محاوره‌ای به رسمی، الگوریتم پیشنهادی و نحوه بکارگیری N_gram بیان می‌شود.

کلمات کلیدی

پردازش زبان، متن محاوره، پیکره، N_gram

۱ –مقدمه

منظور از متن محاوره‌ای، نوشتن متن به شکلی است که معمولاً فارسی زبانان به آن شیوه، تکلّم می‌کنند و با فارسی رسمی معیار، متفاوت است. تبدیل متن محاوره‌ای به رسمی یکی از مسائل پیش رو برای پردازش زبان طبیعی است که کاربردهای فراوانی در تحلیل و پردازش گفتار گوینده، نظرات[1] بازدیدکنندگان سایت‌ها، بررسی وبلاگ‌ها، پیام-های کوتاه و نظایر این‌ها می‌تواند داشته باشد.

لازم به ذکر است که زبان محاوره با زبان عامیانه متفاوت است. برخی از واژه‌ها، نه محاوره‌اند و نه رسمی بلکه صرفا عامیانه‌اند. کلمات عامیانه از نظر معنایی مبهم و از نظر لحن جالبند و بسته به آوای کلام مردم از آنها استفاده می‌کنند. بسیاری از واژه‌های عامیانه صرفا از سوی برخی از افراد جامعه به کار می‌روند و تعداد قابل توجهی از مردم از آن بی‌اطلاع هستند. در واقع به تعداد گروه‌های اجتماعی جداگانه انواع واژه‌های عامیانه وجود دارد. به طور مثال واژه‌های عامیانه دانشجویان با دانش آموزان دبیرستان متفاوت است.

در این مقاله تمرکز بیشتر بر شناسایی و تبدیل متون محاوره ای است.

Popowich و همکاران در سال ۱۹۹۷ [2]، روی ترجمه زبان انگلیسی محاوره بررسی نظری و عملی انجام دادند. آنها با روشی لغوی، یک سیستم تمام خودکار برای ترجمه زبان محاوره انگلیسی به اسپانیایی ارائه کردند.

Megerdoomian در سال ۲۰۰۶[7]، سیستمی برای تجزیه و تحلیل ریخت شناسی[2] در زبان فارسی معرفی کرد. با توجه به این که در این کار روشی برای گسترش این سیستم روی متون وبلاگ های فارسی ارائه شده، برخی از قواعد ساختواژی برای متون محاوره ای معرفی و اعمال شده است.

اسدی در سال ۲۰۰۷[3]، تحقیقی بر روی آواهای حذف شده در محاورات فارسی ارائه کرد. پیکره استفاده شده در این تحقیق بیست دقیقه مکالمه سه فارسی زبان است. این تحقیق نشان می‌دهد حذف آواها در یک کلمه به نوع آن کلمه (کلمات دستوری که کاربرد بیشتری در کلام دارند، تغییر بیشتری نسبت به کلمات دیگر کرده-اند) و محل قرارگیری آن آوا در کنار حروف دیگر وابسته است.

بررسی منابع نشان می دهد اگر چه مطالعات زبانی در حوزه متون محاوره ای فارسی صورت گرفته ولی سابقه ای از تحقیق در باره تبدیل متون محاوره ای به رسمی جهت اعمال سایر پردازش های متداول دیده نمی شود. از آنجاکه بیشتر پردازش‌های زبانی که صورت گرفته است متمرکز بر متون رسمی بوده است، برای استفاده از این پردازش‌ها روی متون محاوره باید متن محاوره‌ای را به رسمی تبدیل کرد. در این مقاله به ارائه و آزمون راهکاری برای تبدیل متون محاوره‌ای به متون رسمی با استفاده از ترکیب روشهای مبتنی بر قاعده و مدل‌سازی آماری و پیکره بنیان می پردازیم.

این مقاله بصورت ذیل سازماندهی شده است: در بخش ۲ به چگونگی ساخت پیکره می‌پردازیم. در بخش ۳ الگوریتم پیشنهادی و نحوه توکن‌بندی و نحوه یافتن ریشه کلمات بیان می‌شود. بخش ۴

مربوط به قوانین تبدیل کلمات محاوره‌ای به رسمی است. در بخش ۵ توضیحی در مورد N_gram و نحوه استفاده از آن در این مقاله بیان شده است و در آخر هم در بخش ۶ به بیان نتایج می‌پردازیم.

۲ -ساخت پیکره

ابتدا برای ساخت پیکره و متون تست، از کتاب‌های با متون محاوره و زیرنویس فارسی چندین فیلم استفاده شد. همچنین از ده‌ها وبلاگ مختلف با حجم بالای ۱۰۰۰ کلمه (همانطور که می‌دانید زبان فارسی چهارمین زبان از لحاظ حجم وبلاگ در اینترنت محسوب می‌شود) متون استخراج شدند. به این ترتیب پیکره‌ای با بیش از ۴۴۰۰۰ کلمه گرد آوری شد.

سپس به بررسی کلمات محاوره در پیکره موجود و یافتن قواعد احتمالی پرداخته شد. طی بررسی‌های انجام شده بیش از بیست قانون استخراج شد و برای کلماتی که قانون خاصی برای تبدیل آنها یافت نشد، پایگاه داده‌ای از کلمات محاوره و معادل رسمی آن با بیش از ۵۰ کلمه با تگ POS ساخته شد.

۳ -الگوریتم پیشنهادی

هدف در این الگوریتم یافتن لیست مرتبی از کلمات رسمی است که می‌تواند جایگزین یک کلمه محاوره ای در متن شود (مثال: خونه تبدیل می‌شود به خانه، خون است، خواند). انتظار می‌رود این لیست به ترتیب احتمال وقوع در بافتار متنی مورد نظر مرتب شده باشد. لذا برای پیشنهاد مناسب‌ترین کلمه در هر جمله از N_gram برای اولویت‌دهی به کلمات لیست استفاده می‌شود. الگوریتم پیشنهادی به صورت زیر است:

۱. توکن‌بندی متن محاوره‌ای
۲. جستجوی کلمه در واژگان[3] زبان رسمی
۳. یافتن ریشه کلمه به کمک ریشه یاب[4] و جستجوی ریشه در واژگان رسمی و واژگان محاوره
۴. استفاده از قوانین تبدیل محاوره به رسمی برای رسمی کردن ریشه کلمه و جستجوی ریشه در واژگان رسمی
۵. جستجو در پایگاه داده‌ی محاوره_ رسمی
۶. افزودن وندهای رسمی شده به ریشه
۷. رتبه‌بندی کلمات در لیست پیشنهادی با استفاده از N_gram کلمات متن.

در ادامه برای هریک از مراحل فوق توضیحاتی ارائه خواهد شد.

۳ ۱ - توکن بندی متون محاوره

به علت آنکه در متون محاوره‌ای ساختار کلمات دارای تفاوت‌هایی با شکل رسمی است، استفاده از قطعه‌بندهای متون رسمی به صورت مستقیم امکان‌پذیر نمی‌باشد و نیاز به ساخت قطعه‌بند جدید برای متون محاوره‌ای یا تغییر قطعه‌بندهای موجود داریم، مثلا (در متن رسمی ما پسوند "مان" را داریم در حالی که در متن محاوره‌ای پسوند "مون" جایگزین آن می‌شود (که در این صورت در متون رسمی "کوچه مان" را به عنوان یک کلمه در نظر می گیریم اما در متون محاوره "کوچه مون" به عنوان یک کلمه توکن بندی می شود)[4]. در این پروژه از شکل ساده‌ای از قطعه‌بند که تنها به به جداسازی کلمات توسط جداکننده‌هایی مانند: "فاصله "، "!"، ":"، "."، "؟"، ";" و امثال آنها بسنده شد.

۳ ۲ - یافتن ریشه کلمه به کمک ریشه یاب

به علت تفاوت در ساختار کلمات، ریشه‌یاب رسمی در این پروژه قابل استفاده نبود. برای ساخت ریشه‌یاب ابتدا سعی شد با تغییردادن ریشه-یاب رسمی STeP-1 [5] برای کلمات محاوره ریشه‌یاب جدید ساخته شود. اما استفاده از ریشه‌یاب رسمی به علت پیچیده بودن، ریشه‌یابی کلمات در سطوح پایین و در نظر گرفتن جزئیات زیاد برای این پروژه کار را پیچیده‌تر و زمان اجرا را بیشتر می‌کرد. خصوصا اینکه برخی از اعمال برای یافتن ریشه عملا در این پروژه بدون کاربرد بودند.

بنابراین با توجه به اینکه ریشه‌یاب محاوره هم از لحاظ پایگاه ریشه‌ها و وندها و هم به جهت قوانین ساختواژی مورد استفاده با ریشه‌یاب رسمی متفاوت است بهتر دیدیم ریشه‌یاب تصریفی جدیدی خاص متون محاوره‌ای تهیه نماییم.این ریشه‌یاب با توجه به هدف این پژوهش که تبدیل متون محاوره‌ای به رسمی است ساده شده است.

لذا این ریشه‌یاب، در سطح ساختواژ تصریفی عمل می کند و تنها ادات جمع و "ی" نکره را در اسامی و شناسه‌ها، ضمایر مفعولی (متصل) و پیشوندهای فعلی مانند " نمی" ,"می" ,"ب" و " ن" را در افعال شناسایی و جدا می کند [6]. و ریشه‌یابی را تا جایی ادامه می‌دهد که ریشه یا خودش یا معادل رسمی آن در واژگان یافت شود.

برای یافتن معادل رسمی از قوانین معرفی شده استفاده می‌شود. و برای برخی که قانون خاصی پیدا نشده معادل رسمی آن از واژگان ساخته شده استخراج می‌شود.

۴ -استفاده از قوانین برای ایجاد معادل رسمی کلمه

یافتن قوانین برای تبدیل کلمات محاوره‌ای به رسمی کار مشکلی است. در زبان محاوره برخی از کلمات تنها به جهت آوایی متفاوت شده‌اند (مانند: دلت (Delat) که در محاوره به دلت (Delet) تبدیل می‌شود) و برخی از کلمات هم به لحاظ آوایی و هم به لحاظ نوشتاری دگرگون گشته‌اند (مانند: خیابان (khiaban) که در محاوره به خیابون (khiabun) تبدیل می‌شود). برخی از کلمات و اصطلاحات تنها در زبان محاوره کاربرد دارند و در گفتار و نوشتار رسمی به کار نمی‌روند (مانند "واسه " بجای "برای") و بلعکس برخی از کلمات و اصطلاحات تنها در سبک رسمی کاربرد دارند (مانند "به همین منظور"). تغییرات در زبان محاوره بیشتر به علت راحتی تلفظ و بیان رخ می‌دهد.

برای تبدیل زبان محاوره به رسمی قانون و قاعده خاصی وجود ندارد. و چنانچه قاعده‌ای هم بیابیم، تنها برای برخی از کلمات صدق می‌کند. افعال تغییرات بیشتری نسبت به کلمات دیگر دارند. به ویژه افعال مضارع که در زبان محاوره بیشتر به کار می‌روند.

برای یافتن معادل رسمی یک ریشه از تمام قوانین موجود استفاده می‌شود و لیستی از تبدیلات ممکن برای هر کلمه بدست می‌آید. برای انتخاب درست بین حالات محتمل نیاز به بررسی کلمه در متن هست که بوسیله مدل های زبانی مانند N_gram این کار صورت می‌گیرد. برخی از قوانین به کار رفته در این برنامه در ذیل آمده است.

۴ ۱ - قواعد تبدیل

- برخی از کلمات که دارای "ون" و "وم" هستند. "و" آنها را باید به "ا" تبدیل کرد.

خونه: خانه　　کدوم: کدام　　آسمون: آسمان
شونه: شانه　　آسون: آسان　　لونه: لانه

- برخی از کلمات به "ا" ختم می‌شوند باید به "ها" و یا "ان" جمع تبدیل کرد.

درختا: درختها　　درختا: درختان

- "و" ای که به آخر کلمات اضافه می‌شود و "رو" در بیشتر مواقع باید به "را" تبدیل شود.

کتاب رو: کتاب را　　خودشو: خودش را
کتابمو: کتابم را　　ماهانو: ماهان را

- حرف "ه" بعد از " ا" در آخر بیشتر کلمات حذف شده‌است.

نگا: نگاه　　سیا: سیاه　　کلا: کلاه

ضمایر

- تغییر ضمایر متصل:

م: م　　مون: مان
ت: ت　　تون: تان
ش: ش　　شون: شان

- تغییر ضمایر منفصل:

من: من　　ما: ما
تو: تو　　شما: شما
او، اون: او، آن　　ایشون، اونها، اونا: ایشان، آنها

حروف اضافه و ربط

- ضمایر متصل بعد از حروف اضافه به ضمایر منفصل تبدیل می‌شوند.

بهم: به من　　بهمون: به ما
بهت: به تو　　بهتون: به شما
بهش: به او، به آن　　بهشون: به آنها، به ایشان

ازم: از من　　ازمون: از ما
ازت: از تو　　ازتون: از شما
ازش: از او، از آن　　ازشون: از آنها، از ایشان

بعدم: بعد من　　بعدمون: بعد ما
بعدت: بعد تو　　بعدتون: بعد شما
بعدش: بعد او، بعد آن　　بعدشون: بعد ایشان، بعد آنها

- گاهی از اوقات برای اتصال ضمایر متصل بعضی از حروف کلمه حذف و بعضی اضافه می‌شوند.

برام: برای من، برایم　　برامون: برای ما، برایمان
برات: برای تو، برایت　　براتون: برای شما، برایتان
براش: برای او، برای آن، برایش　　براشون: برای آنها، برای ایشان، برایشان

باهام: با من　　باهامون: با ما
باهات: با تو　　باهاتون: با شما
باهاش: با او، با آن　　باهاشون: با آنها، با ایشان

"م" در آخر برخی از کلمات به "هم" تبدیل می‌شود.

بازم: باز هم　　منم: من هم

افعال

- برخی از افعال حذف می‌شوند.

خوبه: خوب است　　همینه: همین است
مبارکه: مبارک باشد　　خوبی: خوب هستی

افعال ماضی

- **ماضی ساده و استمراری و بعید و ملموس:** اگر تغییرات در بن فعل رخ نداده باشد، تغییرات در دوم شخص جمع و سوم شخص جمع رخ داده‌است (فقط اشکال تغییر یافته در زیر آمده).

رفتین: رفتید
رفتن: رفتند

می‌رفتین: می‌رفتید
می‌رفتن: می‌رفتند

رفته بودین: رفته بودید
رفته بودن: رفته بودند

داشتین می‌رفتین: داشتید می‌رفتید
داشتن می‌رفتن: داشتند می‌رفتند

تغییرات در بن فعل:

خوندن: خواندن　　موندن: ماندن

- **ماضی نقلی:** ماضی نقلی در محاوره معمولا به ماضی ساده تبدیل می‌شود، البته در سوم شخص مفرد "است" حذف می‌شود.

رفتم: رفته ام　　رفتیم: رفته ایم
رفتی: رفته ای　　رفتین: رفته اید
رفته: رفته است　　رفتن: رفته اند

- **ماضی التزامی:** تغییرات در سوم شخص مفرد، دوم شخص جمع و سوم شخص جمع رخ داده‌است.

رفته باشم　　رفته باشیم
رفته باشی　　رفته باشین: رفته باشید
رفته باشه: رفته باشد　　رفته باشن: رفته باشند

افعال مضارع

- **مضارع ساده و التزامی:** اگر تغییرات در بن فعل رخ نداده باشد، تغییرات تنها در شناسه‌های سوم شخص مفرد، دوم شخص جمع و سوم شخص جمع وجود دارد.

بدون تغییر در بن فعل:

می‌خورم　　می‌خوریم
می‌خوری　　می‌خورین: می‌خورید
می‌خوره: می‌خورد　　می‌خورن: می‌خورند

همراه تغییر در بن فعل:

می‌رم: می‌روم　　می‌ریم: می‌رویم
می‌ری: می‌روی　　می‌رین: می‌روید
می‌ره: می‌رود　　می‌رن: می‌روند

می‌گم: می‌گویم　　می‌گیم: می‌گوییم
می‌گی: می‌گویی　　می‌گین: می‌گویید
می‌گه: می‌گوید　　می‌گن: می‌گویند

می‌خونه: می‌خواند　　می‌خواد: می‌خواهد
می‌ذاره: می‌گذارد　　می‌آره: می‌آورد

- بن‌های مضارعی که به "و" ختم می‌شوند. "و" در محاوره حذف می‌شود.

بروم: برم　　بشوم: بشم　　بگویم: بگم

افعال مستقبل

- ترکیب تغییرات در مضارع اخباری" خواه" و مضارع الترامی افعال است.

می‌خوام برم: می‌خواهم بروم
می‌خوایم بریم: می‌خواهیم برویم
می‌خوای بری: می‌خواهی بروی
می‌خواین برین: می‌خواهید بروید
می‌خواد بره: می‌خواهد برود
می‌خوان برن: می‌خواهند بروند

بدون قاعده‌ها

- "ه" آخر برخی از کلمات به "ر" تبدیل می‌شود.

اگه: اگر　　مگه: مگر　　دیگه: دیگر

- در بعضی از کلمات تعدادی از حروف اضافه می‌شوند.

چار: چهار　　چل: چهل　　مثه: مثله

- در برخی از کلمات "ی" ای که برای تلفظ راحتتر اضافه شده‌است، حذف می‌شود.

شیش: شش　　هیفده: هفده
هیجده: هجده　　کوچیک: کوچک

- "چی " برخی اوقات به "چه" و برخی اوقات به "چیز" تبدیل می‌شود.

چی کار: چه کار　　همه چی: همه چیز

۵ اولویت دهی به کمک N_grams

برای ساخت متن رسمی لازم است بین لیست کلمات رسمی محتمل برای یک کلمه محاوره انتخاب صحیح صورت گیرد. برای این‌کار استفاده از N_gramها برای یافتن احتمال رویداد کلمه در کنار کلمات دیگر و اولویت دادن به یک کلمه پیشنهاد می‌شود.

در ابتدا تعریفی کوتاه از N_gram ارائه می‌شود. یک N_gram یک زیر رشته به طول n از یک رشته از کلمات می‌باشد. از مفهوم N_gram در حیطه‌های وسیعی از جمله پردازش زبانهای طبیعی، استفاده می‌شود. می‌توان نشان داد که یک N_gram، یک مدل مارکوف از مرتبه‌ی n-1 می‌باشد.[۱]

فرکانس تکرار یک N_gram در یک متن تا حدود زیادی می‌تواند ارتباط آن متن با عبارت مورد نظر را نشان دهد. در این میان حجم سند مورد نظر نیز از اهمیت زیادی برخوردار است. به این معنی که تعداد N_gramها در واحد حجم سند پارامتر مناسبی محسوب می‌شود و تعداد آنها به تنهایی ملاک مطلوبی نیست.[۱]

در این الگوریتم ابتدا از Bi_gram استفاده می‌شود سپس برای تمایز بین دو احتمال مشابه از Uni_gram استفاده می‌گردد. برای این کار روی پیکره بیجن خان احتمال وقوع هر کلمه وابسته به کلمه قبلی آن(Bi_gram) را محاسبه کرده، سپس احتمال وقوع هر کلمه (Uni_gram) محاسبه شد. توسط احتمال محاسبه شده، لیست کلمات رسمی محتمل برای هر کلمه محاوره‌ای، مرتب می‌شود. نتایج استفاده از آن در بیشتر مواقع قابل توجه بوده است. و اولویت دهی خوبی بین کلمات ایجاد می‌کند.

مثلا کلمه خونه دارای لیست احتمالی (خانه، خون است، خواند) است. برای پیدا کردن کلمه جایگزین مناسب آن ابتدا از Bi_gram

استفاده می‌شود. مثلا اگر کلمه قبلی "می"، باشد احتمال جایگزینی کلمه "خواند" به جای "خونه" بالا است، اما احتمال وقوع کلمه "خانه" بعد از "می" پایین است.

نمونه‌هایی از ورودی و خروجی الگوریتم در جدول(۱) آمده است. همانطور که مشاهده می شود، لیست احتمالی هر کلمه بر اساس N_gram مرتب شده است.

جدول (۱) : نمونه خروجی

متن محاوره	تبدیلات رسمی
کتاباشو خونه برده.	کتابهایش: اسم. را: حرف اضافه پسین/ رو: بن مضارع. خواند: فعل مضارع/ خانه: اسم/ خون است. برده است / برده: اسم.
مثه اینکه نمیاد.	مثل: اسم. اینکه: قید. نمی‌آید: فعل مضارع.
میگه: اومد بشینه.	می‌گوید: فعل مضارع/ میگ است. آمد: فعل ماضی. بنشیند: فعل مضارع.

۶ -نتیجه

در این مقاله به نحوه تبدیل متن محاوره‌ای به رسمی پرداخته شد. ابتدا با استفاده از برخی قوانین کلمه محاوره‌ای به لیستی از کلمات رسمی مبدل می‌شود، سپس بوسیله N_gram لیست احتمالی برای هر کلمه مرتب خواهد شد.

خطای پروژه در مورد قوانین محدود و نسبتا قابل قبول است. به طور متوسط از هر ۱۰۰ کلمه محاوره ای داده شده به برنامه ۹۳ کلمه درست تشخیص داده شده است. یعنی برنامه دارای دقت متوسط ۹۳٪ روی پیکره آزمایشی است. برخی از خطاها که برنامه دارد به علت کامل نبودن قطعه‌بند است. و با کامل کردن آن (استفاده از پیکره برای قطعه‌بندی) جواب‌ها بهبود خواهند یافت.

به نظر می‌رسد استفاده از یک الگوریتم یادگیر در کنار قوانینی از این نوع الزامی است. زیرا بیان تمام تغییرات در قالب قانون کاری ناممکن به نظر می‌رسد. و استفاده از یک الگوریتم یادگیر کار را به صورت نیمه خودکار پیش خواهد برد. و به علت اینکه زبان محاوره در طول زمان در حال تغییر است الگوریتم یادگیر می‌تواند قوانین را به مرور تصحیح کند. اما الگوریتم های یادگیر نیاز به وجود پیکره های موازی محاوره-رسمی و یا پیکره های محاوره ای برچسب خورده با برچسب های رسمی دارند که درحال حاضر برای زبان فارسی موجود نمی‌باشند.

بررسی بیشتر کلمات محاوره بخصوص به جهت آوایی می‌تواند برای یافتن قوانین بهتر و ساخت مبدل‌هایی از این دست راه‌گشا باشد. زیرا بیشتر تغییرات در کلمات محاوره برای بیان راحتتر کلمات بوجود آمده‌اند. پس بررسی نوع تغییر در آوای کلمات قدم بعدی در این رابطه خواهد بود.

به علت جستجوهای پی‌درپی در واژگان سرعت الگوریتم پایین می‌آید. بعلت آنکه بیشتر جستجوها بر اساس نوع انجام می‌شود. جدا کردن کلمات واژگان به لحاظ نوع در جداول متفاوت می تواند به بهبود سرعت الگوریتم بیانجامد.

سپاسگزاری

نویسندگان این مقاله از همفکری تمام اعضای آزمایشگاه پردازش زبان طبیعی و خانم سکینه دهقان برای همیاری در بررسی پیکره و مرتب کردن قوانین کمال سپاسگذاری را دارند.

مراجع

[۱] کاوسی، کاوه، مشیری، بهزاد، **طراحی معماری یک عامل هوشمند تطبیقی برای جستجوی اطلاعات تجاری با استفاده از تئوری ترکیب اطلاعات**، نشریه جهانی رسانه، دوره: ۴، ۲۰۰۷.

[2] Popowich, F., Turcato, D., Laurens, O., and McFetridge, P., ***A Lexicalist Approach to the Translation of Colloquial Text***, the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, 1997.

[3] Assadi, Sh.***, Sound Deletion in colloquial Persian***, 2007.

[4] Davarpanah, M.R., Sanji M., Aramideh M., ***Farsi lexical analysis and stop word list***, Library Hi, 27 (3), 435-449, 2009.

[5] Shamsfard M., Jafari, HS., Ilbeygi M., ***STeP-1: A Set of Fundamental Tools for Persian Text Processing***, LREC 2010, Malta, 2010.

[6] Sharifloo, A. Shamsfard, M., ***A Bottom up Approach to Persian Stemming***, Proceedings of the Third International Joint Conference on Natural Language Processing, 2007.

[7] Megerdoomian, K., ***,Extending a Persian Morphological Analyzer to Blogs,*** In: Proceedings of the Second Workshop on Persian Language and Computers, 2006.

زیرنویس‌ها

[1] Comment
[2] Morphological
[3] Lexicon
[4] Stemmer