سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی

مهدی محسنی و بهروز مینایی بیدگلی
دانشگاه علم و صنعت ایران

چکیده

برچسب‌گذاری اجزای واژگانی کلام، موضوع تحقیقاتی مهمی در حوزه‌ی پردازش زبان طبیعی است و پایه‌ی بسیاری از دیگر مباحث مطرح در این حوزه است. تاکنون تحقیقات گسترده‌ای با رویکردهای متعدد در زبان‌های دیگر انجام و نتایج چشم‌گیری حاصل شده است. این موضوع سنگ‌بنای بسیاری از روش‌های مورد استفاده در حوزه‌های دیگر پردازش زبان طبیعی، هم‌چون ترجمه‌ی ماشینی، خطایاب، تبدیل متن به گفتار، تشخیص گفتار است، فعّالیت بر روی این موضوع تحقیقاتی می‌تواند راهگشای این مباحث در زبان فارسی باشد.

در این مقاله با بیان مسایل پیش رو در برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی، یک طرح کلی برای نیل به یک برچسب‌گذار خودکار با دقت بالا در زبان فارسی پیشنهاد می‌گردد. پس از آن تحلیل ساخت‌واژی و استفاده از آن را برای پوشش دادن تعداد زیادی از برچسب‌های پیکره با حفظ دقّت بالا در برچسب‌گذاری کلمات مورد بررسی دقیق‌تر قرار داده و تأثیر وجود یک تحلیل‌گر ساخت‌واژی در سطح تصریف را بر برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی بررسی می‌کنیم. نتایج به دست آمده نشان از کارآیی بسیار مناسب این روش پیشنهادی در برچسب‌گذاری دارد.

واژگان‌کلیدی: برچسب‌گذاری اجزای واژگانی کلام، ساخت‌واژه، برچسب‌گذارهای مارکوفی، برچسب‌گذار مبتنی بر حافظه

۱- مقدمه

برچسب‌گذاری اجزای واژگانی کلام[1] عمل انتساب برچسب‌های واژگانی به کلمات و نشانه‌های تشکیل‌دهنده‌ی یک متن است؛ به‌صورتی که این برچسب‌ها نشان‌دهنده‌ی نقش نحوی کلمات و نشانه‌ها در جمله باشند. درصد بالایی از کلمات از نقطه‌نظر برچسب واژگانی ابهام دارند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوت دارند. بنابراین برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجّه به بافت[2] مورد نظر است. برچسب‌گذاری عملی پایه‌ای برای بسیاری از روش‌های مورد استفاده در حوزه‌های دیگر پردازش زبان طبیعی از قبیل ترجمه‌ی ماشینی، خطایاب و تبدیل متن به گفتار است.

تاکنون مدل‌ها و روش‌های زیادی برای برچسب‌گذاری در زبان‌های مختلف استفاده شده است. این روش‌ها و مدل‌ها را می‌توان به دو دسته‌ی کلی تقسیم‌بندی کرد: دسته‌ی اوّل ره‌یافت‌های آماری است که از پیکره‌های[3] برچسب‌خورده[4] بهره می‌جویند؛ و دسته‌ی دیگر ره‌یافت‌های غیرآماری و مبتنی بر قاعده هستند که بر مبنای یادگیری ماشینی و دانش بشری استوارند. در این مجال تعدادی از روش‌های گزارش شده را ذکر می‌کنیم: مدل مخفی مارکوف[5] *(Charniak و همکاران ۱۹۹۳؛ Kupiec، ۱۹۹۲)*، سیستم‌های بیشینه‌ی آنتروپی[6] *(Ratenaparkhi، ۱۹۹۶)*، برچسب‌گذاری مبتنی بر تبدیل[7] *(Brill ۱۹۹۳؛ Brill، ۱۹۹۴)*، سیستم‌های مبتنی بر حافظه[8] *(Daelemans، ۱۹۹۶)*.

فعّالیت‌های بسیاری برای برچسب‌گذاری در زبان‌های دیگر انجام شده، ولی در زبان فارسی این‌گونه نبوده است.

[1] Part-Of-Speech tagging
[2] Context
[3] Corpus
[4] Tagged corpora
[5] Hidden Markov Model
[6] Maximum entropy systems
[7] Transformation-based tagger
[8] Memory-based systems

یکی از دلایل این امر، عدم دسترسی آسان به منابع زبانی هم‌چون پیکره‌های متنی مناسب بوده است. در این جا فعالیت‌های انجام شده برای برچسب‌گذاری در زبان فارسی را ذکر می‌کنیم.

اوّلین کار برای برچسب‌گذاری زبان فارسی توسط *(Assi و Haji Abdolhoseini، ۲۰۰۰)* انجام شده که بر مبنای روش *Schutze (۱۹۹۵)* است. ایده‌ی استفاده شده جمع‌آوری همسایه‌های یک کلمه در دو بردار به نام‌های بردار زمینه‌ی چپ[1] و بردار زمینه‌ی راست[2] است. بعد از آن، انواع کلمات بر طبق شباهت توزیعی طبقه‌بندی می‌شوند (شباهت آن‌ها به معنای اشتراک همسایه‌های یکسان می‌باشد)، و سپس هر طبقه برچسب‌گذاری می‌شود. این سیستم به‌عنوان بخشی از فرآیند برچسب‌گذاری یک پیکره‌ی فارسی به نام پایگاه داده‌ی زبان‌شناسی فارسی (FLDB) *(Assi، ۱۹۹۷)* طراحی شده است. مجموعه‌ی برچسب استفاده شده، متشکل از ۴۵ برچسب می‌باشد. دقّت ارائه شده به این شرح است: دقّت در اعداد، طبقات مختلف افعال و اسامی ٪۸۳-۶۹ بوده است و به‌طور کلی، دقت بخش خودکار سیستم ٪۵۷.۵ بوده است. سیستم ارائه شده قادر به ابهام‌زدایی از برچسب‌های کلمات نیست. هم‌چنین سیستم قادر به برچسب‌گذاری کلمات با فراوانی کم نیست. از طرف دیگر دقت سیستم برای صفت‌ها و قیدها پایین است.

تحقیق دیگر برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی توسط *(Megerdoomian، ۲۰۰۴)* انجام شده است. نگارنده‌ی مقاله‌ی مذکور تنها برخی از چالش‌هایی را که در توسعه‌ی یک برچسب‌گذار فارسی وجود دارد، بیان می‌کند. این تحقیق هیچ پیاده‌سازی عملی را شامل نمی‌شود.

در *(Raja و همکاران، ۲۰۰۵)* نتایج چند برچسب‌گذار بر روی پیکره‌ی *متنی زبان فارسی* ارائه شده است. در آن جا بخشی از پیکره‌ی متنی زبان فارسی مورد استفاده قرار گرفته است. مجموعه‌ی برچسب مورد استفاده دارای چهل برچسب است؛ یعنی برچسب کلمات از حدود ۵۶۰ برچسب به چهل برچسب تقلیل یافته است و نتایج برچسب‌گذاری بر روی این چهل برچسب گزارش شده است. نتایج ارائه شده، دقّت ٪۹۷-٪۹۴ را نشان داده است که نشان از کارآیی سیستم‌های برچسب‌گذاری آماری در زبان فارسی با این تعداد برچسب است.

کارهای انجام شده برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی که در بالا ذکر شد، همگی مسئله‌ی برچسب‌گذاری را محدود در نظر گرفته و بر اساس مجموعه برچسبی محدود (حداکثر ۴۵ برچسب) به برچسب‌گذاری می‌پردازند ولی در یک دیدگاه کلی‌تر، برچسب‌گذاری در زبان فارسی با توجه به خصوصیات ذاتی آن با مباحث دیگری پیوند خورده است. مسئله ساخت‌واژه[3] کلمات فارسی و هم‌نگاره‌ها[4] از این قبیل هستند. ما در اینجا همه‌ی این مسایل را به‌صورت یک فرآیند به هم پیوسته در نظر می‌گیریم و بر اساس آن یک طرح جامع که دربرگیرنده همه‌ی مسائل برای برچسب‌گذاری کلمات در زبان فارسی باشد، پیشنهاد می‌دهیم. با توجه به گستردگی مسائل ما تنها به یکی از این مسائل می‌پردازیم و آن بررسی تأثیر تحلیل ساخت‌واژی در سطح تصریف[5] در برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی می‌باشد.

ساختار مقاله در ادامه به‌صورت زیر است: ابتدا طرح جامع برچسب‌گذاری را ارائه داده و شرح مختصری بر بخش‌های مختلف آن بیان می‌کنیم. فعالیت ما بر روی پیکره‌ی *متنی زبان فارسی* صورت گرفته که در (بخش ۳) مشخصّات آن را بیان می‌کنیم. در (بخش ۴) به مبحث ساخت‌واژی و جایگاه مهم آن در برچسب‌گذاری کلمات زبان فارسی پرداخته می‌شود. در ادامه سه برچسب‌گذاری که ما برای نتایج تجربی استفاده می‌کنیم، شرح داده می‌شود. دو تا از این برچسب‌گذارها، برچسب‌گذاری‌های مارکوفی می‌باشد و دیگری برچسب‌گذار مبتنی بر حافظه. در (بخش ۶) نتایج تجربی به‌دست آمده بیان می‌شود و نتیجه‌گیری نهایی و کارهای آتی در (بخش ۷) ارائه می‌گردد.

۲- برچسب‌گذاری در زبان فارسی و ارائه‌ی یک طرح کلی برای آن

در این بخش یک طرح جامع برای برچسب‌گذاری در زبان فارسی ارائه می‌کنیم. آن‌چه ما را به این طرح رهنمون کرد برخورد تجربی با مشکلات و مسائلی بود که ناشی از ساختار زبان فارسی است. بنابراین ابتدا مشکلات برچسب‌گذاری در زبان فارسی را بیان می‌کنیم که عمدتاً مشکلات دیگر سیستم‌های پردازشی در زبان فارسی نیز هستند.

[1] Left Context Vector
[2] Right Context Vector

[3] Morphology
[4] Homograph
[5] Inflection

۲-۱- مشکلات برچسب‌گذاری در زبان فارسی

مشکلات زیادی در برچسب‌گذاری در زبان فارسی وجود دارد. برخی از مشکلات برچسب‌گذاری فارسی از دیدگاه تجربی به شرح زیر است:

۱- ساخت‌واژه[1] کلمات فارسی: اگر چند وند در یک کلمه ظاهر شوند، همه‌ی این وندها به‌طور معمول به کلمه می‌چسبند (*Megerdoomian*، ۲۰۰۰). وندهایی مانند نشانه‌های جمع، کسره اضافه، نکره و ضمایر ملکی می‌توانند به کلمه متصّل شوند مانند "کتاب‌هایم" که شامل کتاب + ها + ی (یای میانجی) + م است. در مورد افعال نیز فعل شامل بن فعل و وندهای تصریفی است. افعال در فارسی با توجه به شخص، صرف می‌شوند و بنابراین اشکال متفاوتی از آن‌ها ایجاد می‌شود. این ویژگی باعث اَشکال متفاوتی از کلمات با ریشه‌ی یکسان می‌شود و این کلمات در سیستم‌های محاسباتی، متفاوت از یکدیگر فرض می‌شوند. بنابراین فراوانی کلمات کاهش می‌یابد و این عامل بر دقت سیستم‌های مبتنی بر روش‌های آماری تأثیر می‌گذارد.

۲- ابهام در ساخت‌واژه: شکل یکسان برخی از تکواژها، ایجادکننده ابهام در متون فارسی است. برای مثال پسوند "ی" در کلمه "مردی" می‌تواند به‌عنوان نشانه‌ی نکره در نظر گرفته شود، همچنین می‌تواند به‌عنوان شناسه‌ی دوم شخص در یک فعل اسنادی لحاظ گردد. به‌علاوه در فارسی به‌طور معمول مصوت‌های کوتاه در متن ظاهر نمی‌شوند که این باعث ابهام در تحلیل می‌گردد مانند کلمه "مردم" که می‌تواند به صورت /mardam/ یا /mordam/ و یا /mardom/ تلفظ شود. این ابهام، یک نوع از انواع ابهام هم‌نگاره است که در ادامه بحث می‌شود.

۳- ابهام هم‌نگاره: هم‌نگاره‌ها کلمات با ساختار نوشتاری یکسان و معنی متفاوت، یکی از مهم‌ترین لایه‌های ابهام را در متن ایجاد می‌کنند.

۴- تشخیص مرز کلمات[2]: تشخیص مرز کلمات در زبان فارسی مشکل است. برای مثال تکواژ جمع "ها" در اسامی، می‌تواند به چند شکل ظاهر شود. به‌عنوان مثال در مورد کلمه‌ی "کتاب" سه شکل "کتابها"، "کتاب‌ها" و "کتاب ها" برای حالت جمع وجود دارد. در مورد افعال نیز می‌توان به نشانه‌ی زمان استمراری یعنی "می" اشاره کرد. به‌عنوان مثال فعل استمراری اوّل شخص حال از ستاک حال "رو" می‌تواند سه شکل "میروم"، "می‌روم"، "می روم" را داشته باشد.

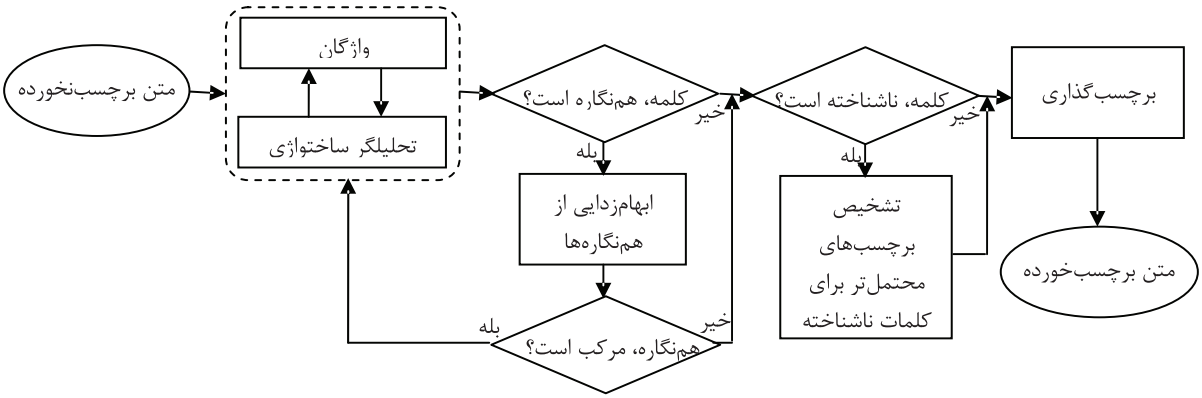
از بین مسائل مطرح شده‌ی فوق، مسائلی که از ساخت‌واژه زبان فارسی ناشی می‌شوند، مهم‌تر است؛ زیرا ساخت‌واژه علاوه‌بر شکل کلمه، برچسب کلمه را نیز تحت تأثیر قرار می‌دهد. ساخت‌واژه زبان فارسی باعث می‌شود کلمات با اشکال متفاوت از یک بن‌واژه[3] یکسان ایجاد شوند که برچسب آن‌ها نیز در پیکره متفاوت است و این امر باعث می‌شود تعداد برچسب‌های متمایز در پیکره بسیار زیاد شود.

۲-۲- ارائه‌ی طرحی جامع برای برچسب‌گذاری در زبان فارسی

(شکل ۱) طرح برچسب‌گذار پیشنهادی را نشان می‌دهد. گفتنی است که این طرح پس از بررسی‌های فراوان راجع‌به زبان فارسی و مشکلات به‌وجود آمده هنگام طراحی سیستم‌های برچسب‌گذاری مختلف، بر روی پیکره‌ی متنی زبان فارسی به دست آمده است. ساختار قسمت‌های مختلف وابسته به یکدیگرند و هر قسمت با توجه به قسمت‌های دیگر باید طراحی شود. در ادامه بخش‌های مختلف این طرح شرح داده می‌شود.

با این فرض که مرز کلمات مشخص شده است، در گام اوّل باید عمل تحلیل ساخت‌واژی تصریفی کلمات انجام شود. برای این‌کار ابتدا به یک واژگان[4] نیاز است. واژگان و ساختار آن بسیار مهم است. منظور از واژگان ذکر شده در (شکل ۱) به‌طوردقیق مترادف اصطلاح تخصصی واژگان در زبان‌شناسی نیست؛ بلکه واژگان می‌تواند یک مجموعه لغت ساده و بدون هیچ ساختار خاصّی باشد، هم‌چنین می‌تواند نمایشی از ساختارهای پیچیده‌ی کلمات باشد. ساختار واژگان و اطّلاعاتی که از هر کلمه ذخیره می‌کند تأثیر مستقیم بر تحلیل‌گر ساخت‌واژی[5] دارد. به‌عنوان مثال واژگان می‌تواند به‌طور ساده شامل مجموعه کلمات موجود در پیکره باشد؛ همچنین می‌تواند برای کلمه‌ای مانند "مردی" با امکان سه تفسیر متفاوت: مردی (صفت)، مرد+ی شناسه‌ی دوم شخص مفرد(=تو مردی)، مرد+ی نکره، سه برچسب مختلف آن را با سه تجزیه‌ی تصریفی مختلف ذخیره کند.

[1] Morphology
[2] Word Boundary Detection
[3] Lemma
[4] Lexicon
[5] Morphological analyzer



(شکل ۱): طرح جامع پیشنهادی برای برچسب‌گذاری در زبان فارسی

تحلیل‌گر ساخت‌واژی در تعامل با واژگان، کلمات را مورد بررسی قرار می‌دهد و درصورتی‌که کلمه با افزوده‌شدن وندهای غیر اشتقاقی به کلمه‌ای تبدیل شده که در واژگان وجود ندارد کلمه به اجزای تشکیل‌دهنده‌ی آن تجزیه می‌شود. در بخش ۴ به ساخت‌واژه‌ی تصریفی کلمات و تأثیر آن بر برچسب‌گذاری بیشتر می‌پردازیم. نتیجه‌ی تجربی نشان می‌دهد که وجود یک تحلیل‌گر ساخت‌واژی تأثیر به‌سزایی در برچسب‌گذاری می‌گذارد.

اگرچه درصد هم‌نگاره‌ها[1] نسبت به کلّ کلمات متن تا حدودی اندک است ولی در این طرح کلّی، آن‌ها را نیز در نظر می‌گیریم.

به‌طور کلّی هم‌نگارگی در زبان فارسی از بازنمایی واجی و صرفی و عناصر زبانی در خطّ فارسی ناشی می‌شود، به‌طوری‌که یک رابطه‌ی چند به چند و در بعضی موارد یک رابطه غیر نظام‌مند بین عناصر واجی و صرفی زبان فارسی و تظاهر نوشتاری آن‌ها در خطّ فارسی وجود دارد. درمجموع هم‌نگارگی در خطّ فارسی را می‌توان ناشی از عوامل زیر دانست *(بی‌جن‌خان و مرادزاده، ۱۳۸۳)*.

۱- عدم بازنمایی واکه‌های کوتاه در خط فارسی: مثل کلمه "مرد" با دو معنی و دو تلفظ /mard/ و /mord/.

۲- عدم تناظر یک به یک میان واج‌ها و حروف فارسی: مانند /u/ و /av/ مثل "جو".

۳- یکسانی تظاهر واجی و نوشتاری تکواژها: مانند یکسانی تکواژهای ضمیر متصل سوم شخص مفرد با تکواژ اسم‌ساز ‹‹ ش ››، مثل رویِش و رویَش.

۴- رابطه‌ی بین وزن کلمات عربی و بعضی پسوندهای فارسی: به‌عنوان مثال کلمه "منزلت" در باب مفعله یک معنی دارد و با تعبیر "منزل تو" دارای پی‌چسب ملکی "ت" معنی دیگری دارد.

یک طبقه‌بندی که برای هم‌نگاره‌ها ارائه شده است *(مرادزاده، ۱۳۸۳)* در اوّلین سطح، هم‌نگاره‌ها را به دو دسته‌ی تکیه‌ای و غیرتکیه‌ای تقسیم‌بندی می‌کند. هم‌نگاره‌های تکیه‌ای و غیرتکیه‌ای خود به دو دسته بسیط و مرکب تقسیم‌بندی می‌شوند. هم‌نگاره‌های بسیط ذاتاً هم‌نگاره هستند و از ساخت‌واژه زبان ناشی نمی‌شوند. هم‌نگاره‌های مرکب از ساخت‌واژه اشتقاقی و تصریفی زبان یا افزوده شدن واژه‌بست‌ها[2] به کلمات ناشی می‌شوند.

در برچسب‌گذاری فقط بخشی از هم‌نگاره‌ها یعنی هم‌نگاره‌هایی که دارای برچسب متفاوت باشند، ایجاد مشکل می‌کنند. اگر کلمه هم‌نگاره باشد باید ابهام‌زدایی از آن انجام شود. بعد از ابهام‌زدایی اگر هم‌نگاره از نوع هم‌نگاره مرکب تشخیص داده شود باید به تحلیل‌گر ساخت‌واژی داده شود تا تجزیه گردد.

یکی از مسائل مهم در سیستم‌های برچسب‌گذاری کلمات ناشناخته[3] است. آزمون ناشناخته بودن کلمه گام بعدی است. دقت برچسب‌گذاری کلمات ناشناخته کمتر از کلمات شناخته شده است. اگرچه برای بهتر بیان کردن روند عملیات، ما این مرحله را جداگانه در نظر گرفته‌ایم ولی به‌طور معمول برچسب‌گذاری کلمات ناشناخته و کلمات شناخته شده با هم انجام می‌شود. در این مرحله، برچسب‌های محتمل‌تر به برچسب‌گذار پیشنهاد شده و برچسب‌های نامربوط حذف می‌شود.

[1] Homograph

[2] Clitic
[3] Unknown words

درنهایت مرحله‌ی برچسب‌گذاری است که با استفاده از اطلاعات به دست آمده در مراحل قبل و با به کار بردن مدل‌ها و روش‌های مختلف برچسب‌گذاری، عمل برچسب‌گذاری انجام می‌شود و برچسب کلمات مشخص می‌شود.

پس از طی مراحل فوق یک متن برچسب‌نخورده به متن برچسب‌خورده تبدیل می‌شود. انجام کامل هر کدام از مرحل فوق نیاز به تلاش بسیار زیاد و ایجاد پیش‌نیازهای بسیاری است. در این مقاله هدف ما نشان دادن مسائل و مشکلات مختلف برچسب‌گذاری در زبان فارسی و بررسی تأثیر تحلیل ساخت‌واژی در سطح تصریف در برچسب‌گذاری کلمات فارسی است.

۳- پیکره

پیکره‌ی متنی زبان فارسی *(بی‌جن‌خان، ۱۳۸۱)* دارای دو بخش کلمات برچسب‌خورده و کلمات برچسب‌نخورده است که در این‌جا ما تنها مشخّصات بخش برچسب‌خورده را بیان می‌کنیم. این بخش از پیکره، شامل حدود ده میلیون کلمه است. متون تشکیل دهنده‌ی پیکره، ترکیبی از متون رسمی و متون غیر رسمی و محاوره‌ای است. این متون برگرفته شده از اینترنت، پایان‌نامه‌ها، روزنامه‌ها، مجلات و کتب مختلف هستند و شامل موضوعات گوناگونی مانند متون حقوقی، سیاسی، حسابداری، روانشناسی، آموزشی، ورزشی، دینی، اقتصادی، داستانی و ادبی می‌باشد. متونی که به‌صورت غیر رسمی و محاوره‌ای هستند، بیشتر متون داستانی، ادبی و نمایش‌نامه می‌باشند. در گفتار غیر رسمی و محاوره‌ای، نحو و تلفظ کلمات تغییر زیادی می‌کند و این خود مشکل دیگری پیش روی سیستم‌های پردازش زبان فارسی است.

ساختار در نظر گرفته شده برای برچسب کلمات در پیکره‌ی متنی زبان فارسی ساختار سلسله‌مراتبی است. در ساختار سلسله‌مراتبی، تمایز میان طبقات اصلی و زیربخش‌های آن‌ها نشان داده می‌شود. در پیکره‌ی متنی زبان فارسی هر مقوله یا برچسب با ویرگول از زیربخش‌های خود جدا می‌شود. به‌عنوان مثال برچسب «اسم، عام، مفرد، مکان» نشان می‌دهد که این برچسب به کلماتی منتسب می‌شود که مقوله‌ی آن‌ها اسم است؛ نوع آن‌ها عام و به‌لحاظ شمار، مفرد هستند و به‌لحاظ معنایی در دسته‌ی اسم‌های مکان قرار می‌گیرند.

در آخرین نسخه‌ی پیکره، تعداد برچسب‌های اصلی شانزده برچسب و برچسب‌های متمایز ۵۸۶ برچسب است و کلمات پیکره در این تعداد زیاد برچسب، دسته‌بندی می‌شوند.

۴- تحلیل‌گر ساخت‌واژی

سه نوع ساخت‌واژه می‌توان برای کلمات متصوّر شد و بر اساس آن سه تحلیل ساخت‌واژی می‌توان انجام داد. سه نوع مختلف ساخت‌واژه: اشتقاق، تصریف و ترکیب است. ایجاد تحلیل‌گر ساخت‌واژی خودکار با دقت بالا در این سه نوع مختلف کاری بسیار دشوار و پرچالش است.

با بررسی اجمالی برچسب‌های پیکره‌ی متنی زبان فارسی می‌توان متوجه شد که از سه نوع مختلف ساخت‌واژه، تنها ساخت‌واژه‌ی تصریفی[1] است که بر برچسب‌های کلمات تأثیر می‌گذارد و تعداد زیاد برچسب‌های متمایز پیکره (۵۸۶) ناشی از این مورد است. زیرا با افزودن وندهای تصریفی به بن‌واژه‌ها برچسبی نیز معادل وند تصریفی به برچسب بن‌واژه افزوده می‌شود و بنابراین یک کلمه با برچسبی متمایز به پیکره افزوده می‌شود. به‌عنوان مثال اگر به کلمه "کتاب" با برچسب "اسم، عام، ساده" ضمیر ملکی اوّل شخص "م" به آن افزوده شود حاصل کلمه "کتابم" با برچسب "اسم، عام، ساده، ۱" است، بنابراین علاوه بر برچسب "اسم، عام، ساده" برچسب "اسم، عام، ساده، ۱" به برچسب‌های متمایز در پیکره افزوده می‌شود . به همین ترتیب در مقوله‌ی اسم و مقولات دیگر تعداد برچسب‌های متمایز در پیکره افزوده می‌شود.

۴-۱- تصریف و تفسیرهای متفاوت کلمات با بن‌واژه یکسان

تکواژهای تصریفی و واژه‌بست‌ها، باعث می‌شوند که کلمه‌ی حاصل، متفاوت از بن‌واژه‌ی آن تفسیر شود[2]. به عبارت دیگر افزودن تکواژهای تصریفی و واژه‌بست‌ها به یک کلمه باعث می‌شود کلمه‌ای جدید ایجاد شود. به‌عنوان مثال اگر کلمه‌ی "کتاب" با تکواژ "ها" جمع بسته شود کلمه‌ی حاصل، "کتاب‌ها" به‌عنوان یک کلمه جدید تفسیر می‌شود.

از طرف دیگر تکواژهای تصریفی و واژه‌بست‌هایی وجود دارند که می‌توانند بر روی کلمات موجود در مقولات متفاوت

[1] در این مقاله منظور ما از تصریف افزوده شدن تکواژهای تصریفی و واژه‌بست‌ها به کلمات است.

[2] منظور تفسیر یک شخص خبره (مثلا یک فارسی‌زبان یا زبان شناس) نیست بلکه منظور پردازش کلمه در یک سیستم رایانه‌ای است.

ظاهر شوند. به‌عنوان ضمایر متصل می‌توانند به اسم، صفت، قید و حرف اضافه افزوده شوند. علاوه‌بر این‌که کلمه‌ی جدید متفاوت از بن‌واژه تفسیر می‌شود، این باعث می‌شود که این تکواژهای با کاربرد یکسان در مقولات متفاوت، به‌صورت متفاوت تفسیر شوند. به‌عنوان نمونه ضمیر متصل "م" در کلمه "پسرم" و در عبارت "پسر خوبم" دو تفسیر متفاوت می‌شود زیرا "م" در کلمه‌ی اوّل (پسرم) بر روی اسم ظاهر شده و برچسب معادل آن به اسم اضافه شده (N,COM,SIM,1) و در کلمه دوم (خوبم) به یک صفت اضافه شده و برچسب معادل آن به صفت اضافه شده است (*ADJ,CMPR,SIM,1*).

بنابراین دو مشکل پیش می‌آید که در پیکره‌ی متنی زبان فارسی به‌طور کامل مشهود است. اولین مشکل این است که کلمات با بن‌واژه‌ی یکسان به اشکال متفاوتی با برچسب متفاوت ظاهر می‌شوند که این مورد بر فراوانی کلمات با بن‌واژه‌ی یکسان تأثیر می‌گذارد. دومین مشکل این است که تعداد برچسب‌های متمایز پیکره، بسیار زیاد خواهد شد و از طرفی نیز فراوانی آن‌ها کاهش می‌یابد. این دو مشکل تأثیر بسیار زیادی بر روی برچسب‌گذارهای آماری دارد.

۴-۲- حل مشکل تفسیرهای متفاوت کلمات با بن‌واژه یکسان

برای حل این دو مشکل، روشی به کار می‌بریم که، باعث کاهش تعداد برچسب‌ها و حذف تفسیرهای متفاوت کلمات با بن‌واژه‌ی یکسان در پیکره‌ی متنی زبان فارسی شود و هم در عین حال بتوانیم تعداد زیادی از برچسب‌های متمایز پیکره را برای عمل برچسب‌گذاری پوشش دهیم. در این روش سعی می‌کنیم که کلمات را از نقطه‌نظر تصریفی تجزیه کنیم. با این روش می‌توان کارآیی وجود یک تحلیل‌گر ساخت‌واژی در حوزه‌ی تصریف را در برچسب‌گذاری اجزای واژگانی کلام مشاهده نمود. مراحل این روش به صورت زیر است:

۱- چون هدف برچسب‌گذاری در زبان فارسی است، ابتدا برچسب‌هایی را که نشان‌گر مفاهیم معنایی هستند و یا برای عمل برچسب‌گذاری مناسب نیستند از مجموعه‌ی برچسب پیکره حذف می‌کنیم. ازجمله‌ی این برچسب‌ها می‌توان به DAY (روز)، LOC (مکان)، DIR (جهت)، SES (فصل) و MON (ماه)، SURN (لقب) و TIME (زمان) در مقوله‌ی اسم و LOC (مکان)، EXM (مثال)، ORD (ترتیبی)، REPT (تکراری) و NEGG (نفی) در مقوله‌ی قید اشاره کرد. تعداد کلمات منتسب به این برچسب‌ها کم است و امکان مدل‌کردن آماری آن‌ها وجود ندارد. البته بعضی از آن‌ها را می‌توان به‌طور مستقیم و بدون ایجاد مشکل بعد از برچسب‌گذاری به برچسب کلمه اضافه کرد.

۲- برای هر برچسب با این شرط که نشان‌گر تکواژ تصریفی و واژه‌بست باشد، باید همه‌ی تکواژهای منتسب به آن مشخص شود. برچسب‌های دیگر مانند برچسب مقولات اصلی مثل N (اسم)، ADJ (صفت) و ADV (قید)، برچسب‌های نشان‌گر نوع مانند COM (عام) و PR (خاص) و جز این‌ها که نشانه‌ی خاصی در کلمه ندارند در این مرحله در نظر گرفته نمی‌شوند. همان‌طور که در (جدول ۱) مشاهده می‌شود اگر برای اضافه شدن یک تکواژ به کلمه، نیاز به یک واکه میانجی بود، این واکه‌ی میانجی چون جزء کلمه نیست ما آن را جزء تکواژ در نظر می‌گیریم. به‌عنوان مثال در کلمه‌ی "کتاب‌هایم" هنگام اضافه شدن "م" به کلمه "کتاب‌ها" نیاز به واسط "ی" است یا در کلمه‌ی "خانه‌ام" نیاز به واسط "ا" است که این موارد را جز تکواژ در نظر می‌گیریم.

۳- حال که تکواژهای تصریفی و واژه‌بست‌هایی که بر روی کلمات به‌دست آمده است مشخص شد، اکنون می‌توان هر کلمه را تجزیه‌ی تصریفی کرد. درعمل برچسب‌گذاری هر جزء تشکیل‌دهنده‌ی کلمه به‌عنوان یک کلمه‌ی مجزا در نظر گرفته می‌شود. با این کار تعداد برچسب‌های متمایز کاهش می‌یابد و تفسیرهای متفاوت کلمات با بن‌واژه‌ی یکسان از بین خواهد رفت و درنتیجه فراوانی برچسب‌ها و کلمات بسیار افزایش می‌یابد که این تأثیر به‌سزایی در افزایش دقّت برچسب‌گذارهای آماری خواهد داشت.

۴- در این مرحله عمل برچسب‌گذاری بر روی این کلمات تجزیه‌شده با کمک روش‌های مختلف آماری می‌تواند انجام شود.

پس از تجزیه‌ی تصریفی کلمات پیکره در بخش نتایج تجربی با اعمال چند مدل برچسب‌گذاری تأثیر وجود یک تحلیل‌گر ساخت‌واژی در سطح تصریف نشان داده می‌شود.

۵- برچسب‌گذارهای مورد استفاده

در این بخش دو مدل برچسب‌گذار مارکوفی یعنی برچسب‌گذار مبتنی بر مدل مارکوف مرتبه‌ی دو یا

برچسب‌گذار bigram و برچسب‌گذار مبتنی بر مدل مارکوف مرتبه‌ی سه یا برچسب‌گذار trigram و مدل برچسب‌گذاری مبتنی بر حافظه را بیان می‌کنیم که در بخش نتایج تجربی از آن‌ها استفاده می‌کنیم.

۵-۱- برچسب‌گذارهای مارکوفی

۵-۱-۱- برچسب‌گذار bigram

ما در این‌جا از نشان‌گذاری مورد استفاده توسط *(Charniak و همکاران، ۱۹۹۳)* که مشارکت گسترده در استفاده از مدل مارکوف برای برچسب‌گذاری داشته‌اند، استفاده می‌کنیم.

(جدول ۲): نشان‌گذاری

w_i	کلمه در موقعیت i
t_i	برچسب کلمه w_i
$w_{i,i+m}$	کلمات رخ داده در موقعیت i تا $i+m$
$t_{i,i+m}$	برچسب‌های $t_i \ldots t_{i+m}$ برای کلمات $w_i \ldots w_{i+m}$
n	طول جمله

فرض می‌کنیم که $\{w^1, w^2, \ldots, w^w\}$ یک مجموعه از کلمات در مجموعه واژگان و $\{t^1, t^2, \ldots, t^\tau\}$ یک مجموعه از برچسب‌های ممکن برای کلمات باشد. با فرض یک دنباله از کلمات از مجموعه کلمات، $w_{1,n}$، هدف یافتن محتمل‌ترین دنباله از برچسب‌ها از مجموعه برچسب‌ها، $t_{1,n}$ است. با به کار بردن قانون بیز می‌توان نوشت:

$$\arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) = \arg\max_{t_{1,n}} \frac{P(w_{1,n} \mid t_{1,n}) P(t_{1,n})}{P(w_{1,n})} = \arg\max_{t_{1,n}} P(w_{1,n} \mid t_{1,n}) P(t_{1,n}) \quad (۱)$$

امّا با توجّه به خاصیت افق محدود زنجیره‌ی مارکوف، می‌توان نوشت:

$$P(t_{i+1} \mid t_{1,i}) = P(t_{i+1} \mid t_i) \quad (۲)$$

یعنی هر برچسب تنها وابسته به برچسب قبل از خود فرض می‌شود. این مدل برچسب‌گذاری مارکوفی مدل برچسب‌گذاری bigram گفته می‌شود.

علاوه‌بر رابطه‌های فوق دو فرض دیگر را نیز در نظر می‌گیریم:

- کلمات از یکدیگر مستقل‌اند.
- ظهور هر کلمه تنها وابسته به برچسب خود است.

(جدول ۱): برچسب‌ها و تکواژهای مرتبط با آن‌ها

توصیف	برچسب	شکل رسمی	شکل محاوره‌ای
ضمایر متصل	1	م، ام، یم	_____
	2	ت، ات، یت	_____
	3	ش، اش، یش	_____
	4	مان، امان، یمان	مون، امون، یمون
	5	تان، اتان، یتان	تون، اتون، یتون
	6	شان، اشان، یشان	شون، اشون، یشون
نشانه‌ی (یاء نکره و موصولی)	YE	ی، ای، یی، ئی	_____
تکواژ جمع‌ساز	PL	ها، ان، یان، جات، گان، ات، یون، ون، ین	ا، ون
نشانه استمراری (فعل)	PRG	می	_____
نشانه حال (فعل)	PRES	می	_____
نشانه صفت مفعولی (فعل)	PAST-P	ه	_____
رابط (شناسه‌های فعلی)	1	م، ام، یم	_____
	2	ی، ای	_____
	3	ست	ه
	4	یم، ایم، ئیم	_____
	5	ید، اید، ئید	ین، این
	6	ند، اند، ینند	ن، ان
نشانه منفی (فعل)	NEG	ن، م	_____
نشانه التزامی (فعل)	SUB	ب	_____
نشانه امری (فعل)	IMP	ب	_____

ما فرض می‌کنیم هر جمله با کلمه‌ی فرضی SOS (شروع جمله) شروع می‌شود، بنابراین $P(t_1 \mid t_0)=1.0$.

با توجّه به مفروضات فوق رابطه‌ی (۱) در مدل bigram به صورت زیر تبدیل می‌شود:

$$\hat{t}_{1,n}=\arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n})= \arg\max_{t_{1,n}} \prod_{i=1}^{n}[P(w_i \mid t_i)\times P(t_i \mid t_{i-1})] \qquad (۳)$$

۵-۱-۲- برچسب‌گذار trigram

اگر در مدل شرح داده شده‌ی فوق، فرض وابستگی هر برچسب تنها به برچسب قبل را، که فرضی است که به‌طور دقیق با واقعیت منطبق نیست، به دو برچسب قبل گسترش دهیم، مدل دیگری که به trigram معروف است به دست می‌آید. برای این کار رابطه‌ی (۲) به رابطه‌ی زیر تغییر می‌کند:

$$P(t_{i+1} \mid t_{1,i})=P(t_{i+1} \mid t_{i-1}t_i) \qquad (۴)$$

و رابطه‌ی نهایی یعنی رابطه (۳) به‌صورت زیر تبدیل می‌شود:

$$\hat{t}_{1,n}=\arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n})= \arg\max_{t_{1,n}} \prod_{i=2}^{n+1}[P(w_i \mid t_i)\times P(t_i \mid t_{i-2}t_{i-1})] \qquad (۵)$$

۵-۱-۳-پراکندگی داده و هموارسازی

یک مشکل اساسی در استفاده از برچسب‌گذارهای مبتنی بر مدل مخفی مارکوف، پراکندگی داده است. هر چه تعداد پارامترهای مدل افزایش می‌یابد این مشکل تأثیر مخرب بیشتری خواهد داشت.

روش‌های زیادی برای هموارسازی، ارائه شده است که هر کدام در مواردی، کارآیی بهتری از خود نشان می‌دهند. ازجمله‌ی این روش‌ها روش افزایشی[1] (*Gale و Church ۱۹۹۴*)، روش Good-Turning (*Good ۱۹۵۳*)، روش *Jelinek و Mercer (۱۹۸۰)* و روش *Katz (۱۹۸۷)* می‌باشد.

روشی که ما مورد استفاده قرار می‌دادیم روشی است که از *(Thede و Harper ۱۹۹۹)* الهام گرفته‌ایم. ما به دو دلیل از این روش استفاده می‌کنیم: این روش درعمل هم کارآیی مناسبی از خود نشان می‌دهد و هم‌چنین محاسبات آن نسبت به روش‌های دیگر ساده است. در هنگام آموزش مدل، برای تعیین احتمالات انتقال در برچسب‌گذار bigram از رابطه‌ی زیر استفاده می‌کنیم:

$$P(t^k \mid t^j)=k_2\times\frac{C(t^j,t^k)}{C(t^j)}+(1-k_2)\times\frac{C(t^k)}{\sum_{\forall t^m}C(t^m)} \qquad (۶)$$

و در برچسب‌گذار trigram رابطه‌ی زیر را به کار می‌بریم:

$$P(t^k \mid t^i t^j)=k_3\times\frac{C(t^i,t^j,t^k)}{C(t^i,t^j)}+(1-k_3)\times k_2\times\frac{C(t^j,t^k)}{C(t^j)}+(1-k_3)\times(1-k_2)\times\frac{C(t^k)}{\sum_{\forall t^m}C(t^m)} \qquad (۷)$$

که در دو رابطه فوق:

$$k_2=\frac{\log(C(t^j,t^k)+1)+1}{\log(C(t^j,t^k)+1)+2} \qquad (۸)$$

و

$$k_3=\frac{\log(C(t^i,t^j,t^k)+1)+1}{\log(C(t^i,t^j,t^k)+1)+2} \qquad (۹)$$

۵-۱-۴- کلمات ناشناخته

کلمات ناشناخته نیز مسئله‌ی مهمی در برچسب‌گذاری است. کلمات ناشناخته بیش‌تر از مقولات باز[2] هستند. دو مسئله‌ی اساسی برای غلبه بر کلمات ناشناخته وجود دارد: یکی فقدان اطلاعات واژگانی راجع‌به کلمه ناشناخته و دیگری توزیع متفاوت کلمات ناشناخته با کلمات دیده شده در پیکره می‌باشد. به همین خاطر از پیش اطلاعات خاصّی راجع‌به کلمات ناشناخته وجود ندارد و نمی‌توان از اطلاعات کلمات حاضر در پیکره، برای مدل‌کردن دقیق رفتار توزیعی این کلمات استفاده کرد.

برای برخورد با کلمات ناشناخته در دو مدل bigram و trigram در دو مرحله عمل می‌کنیم: ابتدا در صورت امکان سعی می‌کنیم از قواعدی که با توجه به پیشوندها و پسوندهای کلمات استخراج شده‌اند، برچسب‌های محتمل را برای کلمه حدس بزنیم؛ درصورتی‌که این امکان وجود نداشت از توزیع احتمالی کلمات ناشناخته برای برچسب‌گذاری استفاده می‌کنیم.

برای استخراج قواعدی که از روی پیشوندها و پسوندهای کلمات به‌دست می‌آیند از یک روش خودکار که در *(Mikheev، ۱۹۹۷)* ارائه کرده است، استفاده می‌کنیم. ناگزیریم که از شرح این روش به دلیل جزییات زیاد آن خودداری کنیم.

مواردی نیز وجود دارد که قاعده یا قواعد استخراجی قابل اعمال بر کلمه ناشناخته نیستند. در این حالت از توزیع احتمالی کلمات ناشناخته استفاده می‌کنیم. جدول احتمالات استخراج شده، در بخش نتایج تجربی آورده شده است.

[1] Additive method

[2] Open category

۵-۲- برچسب گذاری مبتنی بر حافظه

برچسب‌گذاری مبتنی بر حافظه روشی برای برچسب‌گذاری است که بر اساس یادگیری مبتنی بر حافظه می‌باشد. یادگیری مبتنی بر حافظه، با به کارگیری روش نزدیک‌ترین k همسایه (KNN) برای رده‌بندی الگوهای آماری، در کاربردهای زیادی در حوزه‌ی پردازش زبان طبیعی موفق عمل کرده است. برچسب‌گذار مبتنی بر حافظه (MBT) *(Daelemans و همکاران، ۲۰۰۳)* از یک پایگاه نمونه برای کلمات شناخته‌شده و یک پایگاه نمونه برای کلمات ناشناخته استفاده می‌کند. بهترین الگوی ویژگی برای پایگاه‌های نمونه باید به‌صورت عملی و تجربی مشخص شود. نتایج به‌کارگیری MBT را در بخش بعد به همراه نتایج دو برچسب‌گذاری مارکوفی نشان می‌دهیم و تأثیر بسیار مفید وجود یک تحلیل‌گر ساخت‌واژی را در برچسب‌گذاری کلمات فارسی مشاهده می‌کنیم.

۶- نتایج تجربی

در این بخش نتایج حاصل از اعمال دو برچسب‌گذار مارکوفی bigram و trigram، و برچسب‌گذار مبتنی بر حافظه بر روی بخشی از پیکره ارائه می‌شود. برای نشان دادن بهتر تأثیر تحلیل‌گر ساخت‌واژی نتایج برچسب‌گذاری در مقولات اصلی را نیز بیان می‌کنیم.

ارزیابی برچسب‌گذارها با توجه به دقت[1] آن‌ها صورت می‌گیرد. برای ارزیابی از روش اعتبارسنجی متقابل پنج قسمتی[2] *(Tan و همکاران، ۲۰۰۵)* استفاده می‌کنیم تا ارزیابی هرچه دقیق‌تر انجام شود. در این روش ارزیابی، ابتدا داده‌ی آموزشی به پنج قسمت تقسیم می‌شود، سپس برچسب‌گذارهای شرح داده شده در فصول قبل، پنج مرتبه اجرا می‌شوند، به‌طوری که در هر مرتبه چهار قسمت برای آموزش[3] و یک قسمت برای آزمون[4] مورد استفاده قرار می‌گیرد و درنهایت دقّت برچسب‌گذاری با توجّه به نتایج پنج مرتبه اجرای متفاوت، به‌دست می‌آید. هم‌چنین دقت‌های جداگانه برای کلمات شناخته شده و ناشناخته و دقت کلی ارائه می‌شود.

به‌دلیل محدودیت‌های محاسباتی (پیچیدگی زمانی و مکانی) نمی‌توان همه‌ی پیکره را برای ارزیابی مورد استفاده قرار داد و ما تنها بخشی از پیکره را در نظر می‌گیریم و به‌عنوان ورودی به برچسب‌گذارها می‌دهیم. بخشی از پیکره را که برای برچسب‌گذاری و ارائه‌ی نتایج انتخاب کرده‌ایم برای همه‌ی آزمایش‌ها به‌طور یکسان به کار می‌بریم تا مقایسه‌ی نتایج بهتر انجام گیرد.

۶-۱- نتایج برچسب‌گذاری در مقولات اصلی

در برچسب‌گذاری مقولات اصلی توزیع احتمالی کلمات ناشناخته به‌صورت (جدول ۳) می‌باشد که اگر قواعد به‌دست آمده از روش ارائه شده توسط (Mikheev، ۱۹۹۷) (بخش ۵-۲-۴) قابل اعمال نباشد، استفاده می‌شود.

(جدول ۳): توزیع احتمالی کلمات ناشناخته در مقولات اصلی

برچسب	احتمال
اسم	٪۵۶
صفت	٪۲۰
قید	٪۱
فعل	٪۱۰
عدد	٪۲
متفرقه	٪۷

همان‌گونه که انتظار می‌رود کلمات ناشناخته بیشتر از مقولات باز مانند اسم، صفت و فعل هستند. دلیل این‌که برچسب متفرقه، درصد نسبتاً بالایی از کلمات ناشناخته را به خود اختصاص داده است این است که کلمات و عبارات عربی در متون فارسی که به‌طور معمول زیاد استفاده می‌شود در این مقوله قرار گرفته‌اند.

برای برچسب‌گذاری مبتنی بر حافظه از ابزار MBT *(Daelemans و همکاران، ۲۰۰۳)* استفاده می‌شود. برای تعیین این‌که چه اطلاعاتی مفیدتر است تا در پایگاه‌های نمونه ذخیره شود، باید به‌صورت تجربی عمل کرد. پس از آزمایش‌های متعدد به نظر می‌رسد که قراردادن این اطلاعات در پایگاه نمونه کلمات شناخته شده نتیجه‌ی بهتری ارائه می‌دهد: برچسب‌های دو کلمه‌ی قبل (برچسب‌های حاصل از برچسب‌گذاری دو کلمه قبل)، برچسب مبهم خود کلمه، برچسب‌های مبهم دو کلمه بعد (منظور از برچسب مبهم برچسب(های) ممکن برای کلمه است که در مرحله‌ی آموزش به‌دست می‌آید). هم‌چنین در پایگاه نمونه کلمات ناشناخته‌ی این اطلاعات ذخیره می‌شود: برچسب‌های دو کلمه‌ی قبل، برچسب مبهم کلمه‌ی بعد،

[1] Accuracy
[2] 5-fold cross validation
[3] Train
[4] Test

حرف اوّل کلمه، سه حرف آخر کلمه. برای همه آزمایش‌ها همین اطلاعات را در پایگاه‌های نمونه قرار می‌دهیم.

(جدول ۴) نتایج برچسب‌گذاری با برچسب‌گذارهای bigram و trigram و برچسب‌گذار مبتنی بر حافظه (MBT) را بر روی شانزده مقوله‌ی اصلی نشان می‌دهد.

(جدول ۴): نتایج سه برچسب‌گذار bigram، trigram و MBT در برچسب‌گذاری مقولات اصلی

	تعداد	Bigram	Trigram	MBT
کلمات شناخته‌شده	۱۱۶۵۵۴۴	٪۹۶.۲	٪۹۶.۴	٪۹۴.۸
کلمات ناشناخته	۲۳۸۰۹	٪۶۸.۷	٪۶۸.۷	٪۴۴.۸
کل کلمات	۱۱۸۹۳۵۳	٪۹۵.۷	٪۹۵.۸	٪۹۳.۸

۶-۲- نتایج برچسب‌گذاری با کمک تحلیل‌گر ساخت‌واژی

روش تجزیه تصریفی کلمات در (بخش ۴-۲) توضیح داده شد. طبق روش گفته شده، در مرحله یک این روش بعضی از برچسب‌های معنایی و برچسب‌های نامناسب از مجموعه برچسب‌ها حذف می‌شود. این کار با توجه به مقولات اصلی انجام می‌شود. پس از این مرحله تعداد برچسب‌های پیکره از ۵۸۶ برچسب متمایز به ۴۷۱ برچسب کاهش می‌یابد که این تعداد برچسب برای یک برچسب‌گذار بسیار زیاد می‌باشد و درعمل برچسب‌گذار را ناکارآمد می‌کند.

با توجّه به جدول تکواژها و واژه‌بست‌های استخراج شده (جدول ۱)، هر کلمه به بن‌واژه و تکواژهای تصریفی و واژه‌بست‌های تشکیل‌دهنده‌ی خود تجزیه می‌شود. این باعث می‌شود که تعداد برچسب‌های متمایز کلمات از ۴۷۱ برچسب به ۱۰۵ برچسب تقلیل یابد. حال می‌توانیم برچسب‌گذاری را با این تعداد برچسب انجام دهیم.

برای غلبه بر کلمات ناشناخته، مشابه بخش قبل عمل می‌شود. (جدول ۵) توزیع احتمالی کلمات ناشناخته را پس از تجزیه‌ی تصریفی کلمات نشان می‌دهد. همان‌طور که مشاهده می‌شود، بیشتر کلمات ناشناخته اسم عام، اسم خاص و یا صفت ساده هستند. همان‌طور که در بالا گفته شد کلمات در عبارات عربی تحت مقوله‌ی متفرقه قرار دارند.

نتایج برچسب‌گذاری سه برچسب‌گذار bigarm، trigram و مبتنی بر حافظه (MBT) بر روی این کلمات تجزیه شده در (جدول۶) قابل مشاهده می‌باشد. [(دلیل افزایش تعداد کلمات، تجزیه آن‌هاست وگرنه داده‌ی استفاده شده، همان داده‌ی استفاده شده در (جدول ۴) می‌باشد.)]

(جدول ۵): توزیع احتمالی کلمات ناشناخته پس از تجزیه تصریفی

برچسب	احتمال
اسم عام	٪۳۹
اسم خاص	٪۱۸
صفت ساده	٪۲۵
فعل‌ها	٪۲
متفرقه	٪۱۱
بقیه برچسب‌ها	٪۵

(جدول ۶): نتایج سه برچسب‌گذار bigram، trigram و MBT پس از تجزیه تصریفی کلمات (نسخه دو پیکره)

	تعداد	Bigram	Trigram	MBT
کلمات شناخته‌شده	۱۷۷۷۲۰۳	٪۹۶.۰	٪۹۶.۴	٪۹۸.۴
کلمات ناشناخته	۱۴۶۱۴	٪۴۳.۳	٪۵۸.۲	٪۶۶.۰
کل کلمات	۱۷۹۱۸۱۷	٪۹۵.۶	٪۹۶.۱	٪۹۸.۲

۶-۳- بحث و تحلیل

در این‌جا از جنبه‌های مختلف به تحلیل نتایج به دست آمده می‌پردازیم. ابتدا تحلیل نتایج را برای کلمات ناشناخته ارائه می‌کنیم.

برچسب‌گذاری کلمات ناشناخته در زبان فارسی به دلیل ساختار آن بسیار مشکل است. به‌عنوان مثال بسیاری از صفات و اسامی در زبان فارسی نشانه خاصی برای تمایز از یکدیگر ندارند، لذا در برچسب‌گذاری کلمات ناشناخته در مقولات اصلی که مشکل عمده، تمایز بین اسم و صفت می‌باشد، دقت بالایی حاصل نشده است. البته پیش‌بینی می‌شود که وجود یک تحلیل‌گر ساخت‌واژی قوی بتواند نتایج را در بعد اشتقاق بهبود بخشد.

نتایج برچسب‌گذاری کلمات ناشناخته در برچسب‌گذار bigram و trigram پس از تجزیه‌ی تصریفی کلمات کاهش

یافته است. این مسئله دو علت دارد. یکی این‌که مشکل تمایز بین اسم و صفت همچنان باقی است و همان مواردی که در مقولات اصلی، برچسب اشتباه خورده‌اند هم‌چنان از موارد خطا می‌باشند. علاوه‌بر این مشکل، در این‌جا مقوله‌ی اسم در دو زیرمقوله‌ی جداگانه‌ی اسم عام و اسم خاص در نظر گرفته شده‌اند. یعنی کلماتی که در بخش قبل به‌طور صحیح در مقوله‌ی اسم تشخیص داده شده‌اند، ممکن است در تشخیص عام و یا خاص بودن آن‌ها سیستم برچسب‌گذار دچار خطا شود که این باعث می‌شود که برچسب تشخیص داده شده برای کلمه‌ی خطا لحاظ گردد. یعنی عدم وجود نشانه‌های خاص در زبان فارسی، برخلاف برخی از زبان‌ها مانند انگلیسی، برای تمایز اسامی عام از اسامی خاص باعث می‌شود که در تشخیص این دو مقوله خطای زیادی رخ دهد. نتایج برچسب‌گذاری مبتنی بر حافظه حتی با در نظر گرفتن یک حرف اول کلمه، سه حرف آخر کلمه، برچسب‌های ابهام‌زدایی شده از دو کلمه‌ی قبل و برچسب مبهم کلمه بعد بسیار پایین است و حتّی نتیجه معکوس می‌دهد؛ زیرا اگر تمام کلمات ناشناخته، اسم در نظر گرفته می‌شد، دقت بالاتری به‌دست می‌آمد چون با توجه به (جدول ۳) توزیع کلمات ناشناخته در مقوله اسم بیشتر است. یعنی روش غلبه بر کلمات ناشناخته در برچسب‌گذار مبتنی بر حافظه در بهترین حالت آن در زبان فارسی کاملاً ناکارآمد است. ولی در (جدول ۶) مشاهده می‌کنیم که این دقت بسیار بهبود یافته و حتی از برچسب‌گذاری‌های مارکوفی نیز نتیجه‌ی بهتری به‌دست آمده است. این مسئله دلیلی دارد که پس از تحلیل نتایج کلی بیان می‌شود.

مشکل دیگر، کلمات در عبارات عربی است که هم بر دقت برچسب‌گذاری کلمات ناشناخته تأثیر می‌گذارد و هم بر دقت برچسب‌گذاری کلمات شناخته شده. برای روشن‌تر شدن مطلب جمله‌ی زیر را که از بخشی از پیکره استخراج شده است در نظر می‌گیریم:

و CON پیام N,COM,SING,EZ الهی AJ,SIM فرمود DELM :V,PA,SIM,POS,3 الان RES,FW,AR و RES,FW,AR قد RES,FW,AR عصیت RES,FW,AR قبل RES,FW,AR و RES,FW,AR کنت RES,FW,AR من RES,FW,AR المفسدین RES,FW,AR. PUNC در پیکره‌ی متنی زبان فارسی به کلمات در عبارات و جملات عربی برچسب AR منتسب می‌شود. به‌طور مثال در جمله‌ی فوق همه‌ی کلمات در عبارت عربی برچسب RES,FW,AR دارند. در این حالت چون زمینه و بافت یکسان است، هیچ‌گاه برچسب‌گذار آماری نمی‌تواند اطلاعات مفیدی را استخراج کند و حتی این دنباله از برچسب‌های یکسان بر عملکرد برچسب‌گذار تأثیر منفی نیز دارد، زیرا به‌دلیل تفاوت کلمات عربی و ساختار آن‌ها با کلمات فارسی اطلاعات به‌دست آمده اثر منفی در نتایج برچسب‌گذاری بر کلمات ناشناخته فارسی می‌گذارد. کلمات "الان"، "قد"، "قبل" و "من" در عبارت عربی فوق چون شکل یکسانی با کلمات فارسی دارند و به‌طور معمول در فرآیند مجموعه‌ی آموزش حضور دارند در هنگام آموزش برچسب‌گذار به‌عنوان کلمات شناخته‌شده لحاظ می‌شوند. بنابراین می‌توان انتظار داشت که هنگام برچسب‌گذاری این کلمات برچسبی غیر از AR منتسب شود و خطا صورت گیرد. بنابراین از بحث حاضر این نکته استنباط می‌شود که در برچسب‌گذاری در زبان فارسی (و در دیگر کاربردهای پردازش زبان فارسی) باید سیستمی برای تشخیص عبارات عربی تعبیه شود.

حال به نتایج کلی می‌پردازیم. نتایج کلی برچسب‌گذارهای مارکوفی در برچسب‌گذاری مقولات اصلی بر روی نسخه‌ی یک و نسخه‌ی دو پیکره به‌طور تقریبی یکسان است و در هر دو مورد دقت این برچسب‌گذارها از برچسب‌گذار مبتنی بر حافظه اندکی بیشتر است. ولی پس از تجزیه‌ی تصریفی کلمات دقت برچسب‌گذار مبتنی بر حافظه از دو برچسب‌گذار مارکوفی بسیار بهتر بوده است. ما علت این پدیده را در این عامل جستجو می‌کنیم: اگر مقولات اصلی را در نظر داشته باشیم وابستگی، یک‌طرفه است و برچسب کلمات بیشتر به برچسب کلمات قبلی وابسته است تا برچسب کلمات بعدی (در برچسب‌گذارهای مبتنی بر مدل مارکوف این وابستگی در آرایه احتمال انتقالات کد می‌شود).

به همین علت برچسب‌گذارهای مارکوفی در مقولات اصلی، نتیجه‌ی خوبی ارائه می‌دهند و دقت آن‌ها از برچسب‌گذار مبتنی بر حافظه بهتر است هرچند که این برچسب‌گذار به برچسب دو کلمه‌ی بعد برای کلمات شناخته شده و برچسب یک کلمه‌ی بعد برای کلمات ناشناخته توجه کند (ساختار پایگاه‌های نمونه که در بخش پیش توضیح داده شد). ما برای این‌که تعداد زیادی از برچسب‌های کلمات پیکره را پوشش دهیم روش تجزیه تصریفی کلمات را به کار بردیم.

با اعمال این روش تعداد برچسب‌های متمایز پوشش داده شده (یعنی تعداد ۴۷۱ برچسب) در قالب ۱۰۵ برچسب

بیان می‌شوند. هنگامی که کلمات تجزیه و تعداد برچسب‌ها زیاد می‌شود وابستگی برچسب کلمات به برچسب کلمات بعدی افزایش می‌یابد یا به عبارت دقیق‌تر تشخیص برچسب کلمات با توجه به برچسب کلمات قبل و کلمات بعد بهتر انجام می‌شود. در ضمن با تجزیه کلمات اطلاعات زمینه افزایش می‌یابد.

حال برچسب‌گذار مبتنی بر حافظه که به‌دلیل استفاده از روش نزدیکترین k همسایه نسبت به زمینه بسیار حساس می‌باشد و در مورد کلمات شناخته شده علاوه‌بر برچسب‌های دو کلمه‌ی قبلی و خود کلمه، به برچسب دو کلمه‌ی بعدی توجه می‌کند؛ و در مورد کلمات ناشناخته علاوه بر برچسب‌های دو کلمه‌ی قبلی، برچسب کلمه‌ی بعدی را نیز در پایگاه نمونه کلمات ناشناخته در نظر می‌گیرد؛ می‌تواند به دقت بالاتری نسبت به دو برچسب‌گذار bigram و trigram دست یابد. این همان علتی است که دقت برچسب‌گذار مبتنی بر حافظه در برچسب‌گذاری کلمات ناشناخته در (جدول ۶) بسیار بهتر از دو برچسب‌گذار مارکوفی بوده است.

با توجّه به نتایج به‌دست آمده می‌توان این نتیجه را گرفت که وجود یک تحلیل‌گر ساخت‌واژی در سطح تصریف، کارآیی بسیار زیادی در برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی دارد. با این روش می‌توان در حالی که دقت بالا را حفظ کرده‌ایم تعداد بسیار زیادی از برچسب‌ها را پوشش دهیم.

۷- نتیجه‌گیری و کارهای آتی

در این مقاله به بررسی تأثیر تحیل‌گر ساخت‌واژی در برچسب‌گذاری خودکار اجزای واژگانی کلام در پیکره‌ی متنی زبان فارسی پرداختیم. با توجه به مشکلات پیش رو در برچسب‌گذاری در زبان فارسی، ابتدا یک طرح کلی برای برچسب‌گذاری ارائه کردیم که در آن مسائل مختلف در برچسب‌گذاری پیش‌بینی شده بود. پس از آن با توجّه به ساخت‌واژه‌ی تصریفی کلمات فارسی و مسائلی ناشی شده از آن بر برچسب‌گذاری، نحوه‌ی تجزیه‌ی تصریفی کلمات پیکره‌ی متنی زبان فارسی تشریح شد. با این تجزیه تعداد برچسب‌های متمایز پیکره کاهش قابل توجهی می‌یافت و به این ترتیب امکان پوشش تعداد زیادی از برچسب‌ها در برچسب‌گذارهای آماری که قبلاً در عمل ممکن نبود، ممکن گشت. تعداد برچسب‌هایی که برای برچسب‌گذاری انتخاب شد ۴۷۱ برچسب بود که برچسب‌گذارهای آماری در برخورد با این تعداد برچسب ناکارآمد هستند. پس از تجزیه‌ی کلمات این تعداد برچسب به ۱۰۵ برچسب تقلیل یافت. پس از این مرحله دو برچسب‌گذار مبتنی بر مدل مارکوف (bigram و trigram) و برچسب‌گذار مبتنی بر حافظه را شرح دادیم. نحوه‌ی عمل‌کرد این برچسب‌گذارها در بخش نتایج تجربی نشان داده شد. ابتدا نتایج برچسب‌گذارها در مقولات اصلی (۱۶ مقوله) و پس از آن نتایج برچسب‌گذاری با پوشش ۴۷۱ برچسب که پس از تجزیه تصریفی به ۱۰۵ برچسب تقلیل یافته بود ارائه شد. نتایج نشان داد که با این روش می‌توان بخش اعظم برچسب‌های کلمات در پیکره را پوشش داد، درحالی‌که دقت برچسب‌گذاری در سطح بسیار مناسبی حفظ شود.

با توجّه به موارد بیان شده در طرح کلی ارائه شده مشکلات زیادی در برچسب‌گذاری وجود دارد که برای دست‌یابی به یک برچسب‌گذار کامل در زبان فارسی باید بر آن‌ها غلبه کرد. موارد زیر از این قبیل‌اند: تشخیص کلمات، عبارات و جملات غیر رسمی و عامیانه و بررسی رفتار آن‌ها؛ تشخیص عبارات و جملات عربی در بین متون فارسی؛ تشخیص مرز جملات و کلمات در متون فارسی؛ طراحی یک تحلیل‌گر ساخت‌واژی خودکار در حوزه‌ی اشتقاق، تصریف و ترکیب، با کارایی بالا؛ تشخیص هم‌نگاره بودن یک کلمه به‌خصوص در هم‌نگاره‌های مرکب و ارائه‌ی روش‌های ابهام‌زدایی از هم‌نگاره‌های فارسی؛ بررسی دقیق کلمات ناشناخته در زبان فارسی و ارائه‌ی روش‌هایی برای غلبه بر آن‌ها.

۸- منابع

بی‌جن‌خان، محمود. ۱۳۸۱. طرح مدل‌سازی زبان فارسی، مرحله دوم. آزمایشگاه گروه زبان‌شناسی، دانشکده ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۸۱.

بی‌جن‌خان، محمود و مرادزاده، شهروز. ۱۳۸۳. هم‌نگاره‌های خط فارسی. مجموعه سخنرانی‌ها و گزارش‌ها و چکیده طرح‌ها، اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشکده ادبیات و علوم انسانی دانشگاه تهران. ص ۵۳-۶۳.

مرادزاده، شهروز. ۱۳۸۳. طبقه‌بندی هم‌نگاره‌های خط فارسی. دبیرخانه شورای عالی اطلاع‌رسانی (کارگروه خط و زبان فارسی).

Megerdoomian, K. 2000. Unification-Based Persian Morphology. In Proceedings of CICLing 2000, Mexico.

Megerdoomian, K. 2004. Developing a Persian part-of-speech tagger. In Proceedings of First Workshop on Persian Language and Computers. Tehran University, Iran.

Mikheev, A. 1997. Automatic Rule Induction for Unknown-Word Guessing. Computational Linguistics 23(3), pp. 405-423.

Ratenaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. Proceeding of the Conference on Empirical Methods in Natural Language Processing, pp. 133-142.

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

Schutze, H. 1995. Distributional Part-of-Speech Tagging From Texts to Tags: Issues in Multilingual Language Analysis. Online Proceedings of the ACL SIDGAT Workshop, On the Internet at http://xxx.lanl.gov/find/cmp-lg.

Tan, P., Steinbach, M., Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley. Chapter 3: Classification: Basic Concepts, Decision Trees, and Model Evaluation, pp. 97.

Thede, S. and Harper, M. 1999. A Second-Order Hidden Markov Model for Part-Of-Speech Tagging. Proceedings of the 37th Conference on Association for Computational Linguistics, pp. 1475-1482.

مهدی محسنی کارشناسی ارشد خود را در رشته‌ی هوش مصنوعی و رباتیک از دانشگاه علم و صنعت ایران در سال ۱۳۸۷ اخذ نمود. حوزه‌ی مورد علاقه‌ی او پردازش زبان طبیعی و تحلیل‌های زبانی است و پایان‌نامه‌ی او در زمینه‌ی برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی است. او هم‌اکنون در دبیرخانه‌ی شورای عالی اطلاع‌رسانی در

Assi, S. M. 1997. Farsi Linguistic Database (FLDB). International Journal of Lexicography, Vol. 10, No. 3, EURALEX Newsletter p. 5.

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

Brill, E. 1993. A corpus-based approach to language learning. University of Pennsylvania: Ph.D. Dissertation, Department of Computer and Information Science.

Brill, E. 1994. A report of recent progress in transformation-based error-driven learning. Proceeding of the Twelfth National Conference on Artificial Intelligence, pp. 722-727.

Charniak, E., Hendrickson, C., Jacobson N. and Perkowitz, M. 1993. Equation for part-of-speech tagging. Proceedings of the Eleventh National Conference on Artificial Intelligence, pp. 784-789.

Daelemans, W., Zavral J., Berck P. and Gillis, S. 1996. MBT: A memory based part of speech tagger-generator. proceeding of the Fourth Workshop on Very Large Corpora, pp 14-27.

Daelemans, W., Zavrel J., van den Bosch, A. and Sloot, K. 2003. MBT: Memory-Based Tagger, version 2.0, Reference Guide, ILK Technical Report – ILK 03-13.

Gale, W. and Church, K. 1994. What's wrong with adding one? In Corpus-Based Research into Language. Rodolpi, Amsterdam.

Good, J. 1953. The population frequencies of species and the estimation of population parameters. Biometriks, 40, pp. 237-264.

Jelinek, F. and Mercer, R. 1980. Interpolated estimation of markov source parameters from sparse data. Proceeding of the Workshop on pattern Recognition in Practice.

Katz, S. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Proceeding, 35(3), pp.400-401.

Kupiec, J. 1992. Robust part-of-speech tagging using a hidden Markov model, Computer Speech and Language, 6(3), pp 225-242.

کارگروه خط و زبان فارسی در محیط رایانه‌ای مشغول فعالیت است.

نشانی رایانامک ایشان عبارت است از:

mohseni@comp.iust.ac.ir

بهروز مینایی بیدگلی دکترای خود را در رشته‌ی علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفت. تخصص او هوش مصنوعی و داده‌کاوی است. و هم اکنون به‌عنوان عضو هیأت علمی دانشکده‌ی مهندسی کامپیوتر دانشگاه علم و صنعت به تدریس دروس هوش مصنوعی و نرم‌افزار مشغول می‌باشد. ایشان سرپرستی گروه متن‌کاوی برای متون عربی و فارسی را در پژوهشکده‌ی داده‌کاوی نور نیز به عهده دارد. از سال ۱۳۸۶ ریاست بنیاد ملّی بازی‌های رایانه‌ای بر عهده‌ی ایشان است.

نشانی رایانامک ایشان عبارت است از:

b_minaei@iust.ac.ir