**شماره مستند: ۱۹۰/۲۵۳۷/۱/۴**

جمهوری اسلامی ایران

دبیرخانه شورای عالی اطلاع رسانی

# مطالعه و بررسی ابزارهای برچسب‌دهی خودکار به منظور به کارگیری در پیکره متنی زبان فارسی

**نسخه ۱.۰**

**دانشگاه علم و صنعت ایران**

**فروردین ۸۸**

# فهرست مطالب

**عنوان** **صفحه**

# مقدمه

برچسب‌گذاری اجزای واژگانی کلام[1] عمل انتساب برچسب‌های واژگانی به کلمات و نشانه‌های تشکیل‌دهنده یک متن است، به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جلمه باشند. درصد بالایی از کلمات از نقطه‌نظر برچسب واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوت دارند. بنابراین برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه[2] مورد نظر است. برچسب‌گذاری واژگانی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می‌باشد.

تاکنون مدل‌ها و روش‌های زیادی برای برچسب‌گذاری در زبان‌های مختلف استفاده شده است. این روش‌ها و مدل‌ها را می‌توان به دو دسته کلی تقسیم‌بندی کرد: دسته اول رهیافت‌های آماری است که از پیکره‌های برچسب‌خورده[3] بهره می‌جویند و دسته دیگر رهیافت‌های غیرآماری و مبتنی بر قانون[4] است که بر مبنای یادگیری ماشینی و دانش بشری استوارند. بعضی از روش‌های گزارش شده را ذکر می‌کنیم: مدل مخفی مارکوف[5] (Charniak et. al., 1993) (Kupiec, 1992) ، سیستم‌های ماکزیمم انتروپی[6] (Ratenaparkhi, 1996) ، برچسب‌گذاری مبتنی بر تبدیل[7] (Brill, 1993) ( Brill, 1994) ( Brill,1995)، ، سیستم‌های مبتنی بر حافظه[8] ( Daelemans et al.,1996).

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

هدف از این پایان‌نامه بررسی چالش‌ها و مشکلات برچسب‌گذاری خودکار اجزای واژگانی کلام در زبان فارسی و استفاده از پیکره متنی زبان فارسی (بی‌جن‌خان، ۱۳۸۱) در طراحی یک سیستم برچسب‌گذاری می‌باشد. برخی از سیستم‌های برچسب‌گذاری هنگام انتساب برچسب به کلمات یا یک برچسب به کلمه منتسب می‌کنند یا چند برچسب. اگر به کلمه چند برچسب منتسب شود برچسب کلمه اصطلاحا مبهم می‌باشد و نیاز به ابهام‌زدایی دارد. برخی سیستم‌های برچسب‌گذاری سعی می‌کنند به هر کلمه تنها یک برچسب منتسب کنند. خروجی این سیستم‌ها مبهم نمی‌باشد و نیاز به ابهام‌زدایی از برچسب‌های کلمات نیست. در طول این پایان‌نامه نتایج ارائه شده بر روی سیستم‌هایی است که به هر کلمه تنها یک برچسب منتسب می‌کنند، یعنی به عبارت دیگر می‌توان گفت عمل برچسب‌گذاری با عمل ابهام‌زدایی توأمان است.

برای ارائه مطالب پایان‌نامه، ابتدا در این فصل یک طرح کلی برای برچسب‌گذاری اجزا واژگانی کلام در زبان فارسی پیشنهاد می‌شود. این طرح با نگاهی کلان به مشکلات عدیده برچسب‌گذاری در زبان فارسی ارائه شده است و در آن سعی شده جوانب مختلف این مسئله مد نظر قرار گیرد؛ بر اساس مسائل مطرح در طرح یاد شده به کارهای انجام شده پرداخته می‌شود و شرح مختصری بر فعالیت‌های انجام شده ارائه می‌گردد.

# طرح کلی برچسب‌گذاری

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

۱- ساختواژه[9] فارسی و کلمات: اگر چند وند در یک کلمه ظاهر شوند، همه این وندها معمولا به کلمه می‌چسبند (Megerdoomian, 2000). نشانه‌های جمع، کسره اضافه، نکره، ضمایر ملکی و ...

[9] Morphology

می‌توانند به کلمه متصل شوند مانند "کتابهایم" که شامل کتاب + ها + ی + م می‌باشد. این ویژگی باعث اشکال متفاوتی از کلمات با ریشه 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 تاثیر می‌گذارد.

۲- ساختواژه افعال: از نقطه نظر ساختواژی، فعل شامل بن فعل و وندهای تصریفی می‌باشد. افعال در فارسی با توجه به شخص صرف می‌شوند و بنابراین اشکال متفاوتی از آن‌ها ایجاد می‌شود.

۳- ابهام در ساختواژه: شکل یکسان برخی از تکواژها ایجادکننده ابهام در متون فارسی است. برای مثال پسوند "ی" در کلمه "مردی" می‌تواند به عنوان نشانه نکره در نظر گرفته شود، همچنین می‌تواند به عنوان شناسه دوم شخص در یک فعل اسنادی لحاظ گردد. به علاوه در فارسی معمولا مصوت‌های کوتاه در متن ظاهر نمی‌شوند که این باعث ابهام در تحلیل می‌گردد مانند کلمه "مردم" که می‌تواند به صورت /mardam/ یا /mordam/ تلفظ شود. این ابهام، ابهام هم‌نگاره گفته می‌شود.

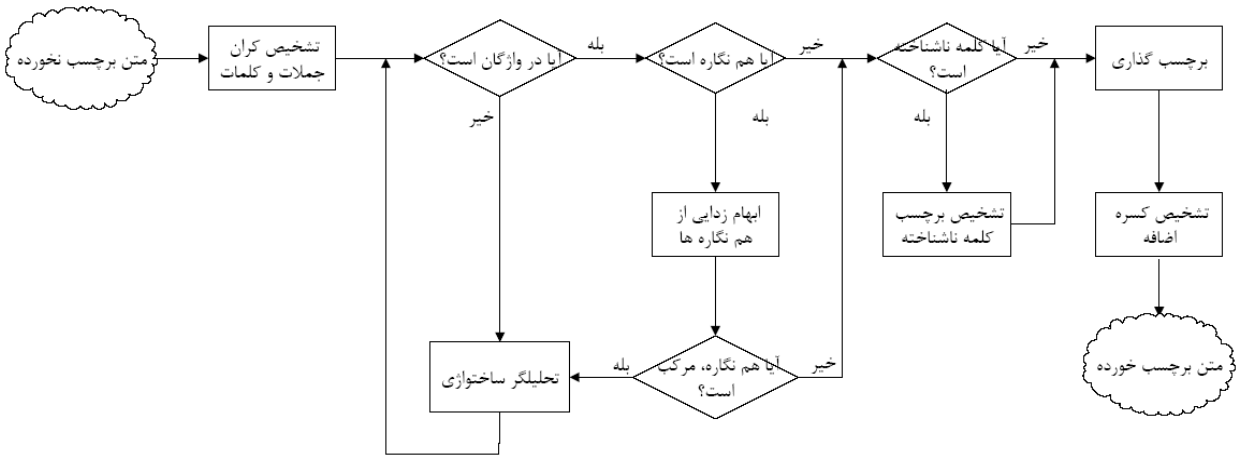
۴- تشخیص کران کلمات: فاصله عامل ابهام زیادی در متن فارسی است. برای مثال تکواژ جمع "ها" در اسامی می‌تواند به چند شکل ظاهر شود. به عنوان مثال در مورد کلمه "کتاب" سه شکل "کتابها"، "کتاب‌ها" و "کتاب ها" برای حالت جمع وجود دارد. در مورد افعال نیز می‌توان به نشانه زمان استمراری یعنی "می" اشاره کرد. مثلا فعل استمراری اول شخص حال از ستاک حال "رو" می‌تواند سه شکل "میروم"، "می‌روم"، "می روم" را داشته باشد. این مورد نیز عامل تاثیر گذار دیگری در برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی است، چون چند شکل نوشتاری متفاوت از یک کلمه عملا به عنوان کلمات متفاوت تفسیر می‌شود.

از بین مسائل مطرح شده فوق مسائلی که از ساختواژه زبان فارسی ناشی می‌شوند مهمتر است، زیرا ساختواژه علاوه بر شکل کلمه، برچسب کلمه را نیز تحت تاثیر قرار می‌دهد. ساختواژه زبان فارسی باعث می‌شود کلمات با اشکال متفاوت از یک بن‌واژه[10] یکسان ایجاد شوند که برچسب آن‌ها نیز در پیکره متفاوت خواهد بود و این امر باعث می‌شود تعداد برچسب‌های متمایز در پیکره بسیار زیاد شود (رک.: فصل ۲، پیکره متنی زبان فارسی).

شکل ۱-۱ طرح برچسب‌گذار پیشنهادی را نشان می‌دهد. شایان ذکر است که این طرح پس از

[10] Lemma

بررسی‌های فراوان راجع به زبان فارسی و مشکلات به وجود آمده هنگام طراحی سیستم‌های برچسب‌گذاری مختلف بر روی پیکره متنی زبان فارسی به دست آمده است و به همین دلیل امکان دارد این طرح از ساختار پیکره متنی زبان فارسی متاثر شده باشد. همچنین 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.




طرح کلی پیشنهادی برای برچسب‌گذاری در زبان فارسی

هنگامی که یک متن برچسب‌نخورده به سیستم داده می‌شود، اولین گام تشخیص کران جملات آن می‌باشد، زیرا ورودی برچسب‌گذارها جمله است و اکثر برچسب‌گذارها برچسب‌گذاری را در واحد جمله انجام می‌دهند. تشخیص کران کلمات نیز بسیار مهم است. باید معلوم شود که موردی مثل "می خورم" یک کلمه است ("می‌خورم") یا دو کلمه ("می" و "خورم") یا "وبا" یک کلمه است ("وبا") یا دو کلمه ("و" و "با"). برای این کار توجه به واژگان و ساختار آن مهم می‌باشد. به عنوان مثال اگر چند واحد[11] خاص در پیکره یک کلمه در نظر گرفته می‌شوند باید این چند واحد هنگام تشخیص کران کلمات یک کلمه در نظر گرفته شوند (مثل "از آن جایی که" که ممکن است در پیکره یک کلمه در نظر گرفته شود در متن برچسب‌نخورده نیز باید یک کلمه در نظر گرفته شود).

[11] Token

پس از تشخیص کران جملات و کلمات، کلمات هر جمله یکی یکی در واژگان[۱۲] جستجو می‌شود و با توجه به وجود یا عدم وجود کلمه در واژگان تصمیم‌گیری می‌شود. واژگان و ساختار آن بسیار مهم است. واژگان ذکر شده در شکل ۱-۱ می‌تواند یک مجموعه لغت ساده و بدون هیچ ساختار خاصی باشد همچنین می‌تواند نمایشی از ساختارهای پیچیده کلمات باشد. ساختار واژگان و اطلاعاتی که از هر کلمه ذخیره می‌کند تاثیر مستقیم بر تحلیل‌گر ساختواژی[۱۳] دارد. به عنوان مثال واژگان می‌تواند به طور ساده شامل مجموعه کلمات موجود در پیکره باشد همچنین می‌تواند برای کلمه‌ای مانند "مردی" با سه معنی متفاوت: مردی (صفت)، مرد+ی شناسه دوم شخص مفرد(=تو مردی)، مرد+ی نکره، سه برچسب مختلف آن را با سه تجزیه تصریفی مختلف ذخیره کند.

اگر کلمه در واژگان وجود نداشت این کلمه باید توسط تحلیل‌گر ساختواژی بررسی شود و در صورتی که کلمه با افزوده شدن تکواژهای غیر اشتقاقی به کلمه‌ای تبدیل شده که در واژگان وجود ندارد کلمه به اجزای تشکیل‌دهنده آن تجزیه می‌شود.

گام بعد بررسی هم‌نگاره بودن کلمه است. اگر کلمه هم‌نگاره[۱۴] باشد باید ابهام‌زدایی از آن انجام شود. بعد از ابهام‌زدایی اگر هم‌نگاره از نوع هم‌نگاره مرکب باشد (رک.: فصل ۷، هم‌نگاره‌ها در زبان فارسی) کلمه به تحلیل‌گر ساختواژی داده می‌شود تا تجزیه گردد.

یکی از مسائل مهم در سیستم‌های برچسب‌گذاری کلمات ناشناخته[۱۵] است. آزمون ناشناخته بودن کلمه گام بعدی می‌باشد. اگر کلمه ناشناخته باشد باید برچسب آن تشخیص داده شود. این کار به روش‌های گوناگونی می‌تواند انجام پذیرد. 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 معمولا برچسب‌گذاری کلمات ناشناخته و کلمات شناخته شده با هم انجام می‌شود.)

مرحله بعد 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.

---

[12] Lexicon
[13] Morphological analyzer
[14] Homograph
[15] Unknown words

در شکل ۱-۱ یک مرحله برای تشخیص وجود کسره اضافه نیز تعبیه شده است. دلیل این امر این است که در پیکره متنی زبان فارسی برای کسره اضافه یک برچسب لحاظ شده و تشخیص این مورد در کلمات را می‌توان جدای از برچسب‌های دیگر در نظر گرفت زیرا کسره اضافه طبیعتی متفاوت از برچسب‌های دیگر دارد.

بنابراین پس از طی مراحل فوق یک متن برچسب نخورده به متن برچسب خورده تبدیل می‌شود. انجام کامل هر کدام از مرحل فوق نیاز به تلاش بسیار زیاد و ایجاد پیش‌نیازهای بسیاری است. در این پایان‌نامه ما حد امکان موارد فوق را مورد بررسی قرار خواهیم داد.

## کارهای انجام شده

فعالیت‌های بسیاری برای برچسب‌گذاری در زبان‌های دیگر انجام شده است ولی در زبان فارسی فعالیت‌ها محدود بوده است. یکی از دلایل این امر عدم دسترسی آسان به پیکره‌های استاندارد می‌باشد. در این جا برخی از فعالیت‌های انجام شده در زبان فارسی را ذکر می‌کنیم.

اولین کار برای برچسب‌گذاری زبان فارسی توسط (Assi and Haji Abdolhoseini., 2000) انجام شده است و بر مبنای روش ( Schutze,1995) می‌باشد. ایده استفاده شده جمع‌آوری همسایه‌های یک کلمه در دو بردار به نام‌های بردار زمینه چپ و بردار زمینه راست است. بعد از آن، انواع کلمات بر طبق شباهت توزیعی طبقه‌بندی می‌شوند (شباهت آن‌ها به معنای اشتراک همسایه‌های یکسان می‌باشد)، و سپس هر طبقه را می‌توان برچسب‌گذاری کرد. این سیستم به عنوان بخشی از فرآیند برچسب‌گذاری یک پیکره فارسی به نام پایگاه داده زبان‌شناسی فارسی (FLDB) (Assi, 1997) طراحی شد. مجموعه برچسب استفاده شده متشکل از ۴۵ برچسب می‌باشد. دقت ارائه شده به این شرح است: دقت در اعداد، طبقات مختلف افعال و اسامی ٪۸۳-۶۹ بوده است و به طور کلی، دقت بخش اتوماتیک 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 می‌باشد.

تحقیق دیگر برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی در ( Megerdoomian,

2004) انجام شده است. نگارنده تنها برخی از چالش‌هایی را که در توسعه یک برچسب‌گذار فارسی وجود دارد، بیان می‌کند. این تحقیق هیچ پیاده‌سازی عملی را شامل نمی‌شود.

در (Raja et al., 2007) نتایج چند برچسب‌گذار بر روی پیکره متنی زبان فارسی ارائه شده است. در آن جا بخشی از پیکره متنی زبان فارسی مورد استفاده قرار گرفته است. مجموعه برچسب مورد استفاده دارای ۴۰ برچسب بوده یعنی برچسب کلمات به ۴۰ برچسب تقلیل یافته است و نتایج برچسب‌گذاری بر روی این ۴۰ برچسب گزارش شده است. نتایج ارائه شده دقت ۹۷٪-۹۴٪ را نشان داده است که نشان از کارایی سیستم‌های برچسب‌گذاری آماری در زبان فارسی است.

در طرح پیشنهاد شده در بخش قبل از دیگر مسائل عمده‌ای که بیان شد مبحث هم‌نگاره‌هاست. ابهام‌زدایی هم‌نگاره‌ها عملی چالش بر انگیز در پردازش زبان طبیعی است. اگرچه تاکنون فعالیت‌های زیادی برای ابهام‌زدایی از هم‌نگاره‌ها در بسیاری از زبان‌ها با استفاده از روش‌های متفاوت انجام شده است ولی در زبان فارسی این حوزه چندان مورد توجه نبوده است. دو فعالیت انجام‌شده در رابطه با هم‌نگاره‌های فارسی یعنی (مرادزاده، ۱۳۸۳) و (بی‌جن‌خان و مرادزاده، ۱۳۸۳) فقط به طبقه‌بندی هم‌نگاره‌های فارسی می‌پردازند و وارد مبحث ابهام‌زدایی از آن ها نمی‌شوند. با وجود سختی کار، تاکنون روش‌های زیادی برای ابهام‌زدایی از هم‌نگاره‌ها در دیگر زبان‌ها توسعه یافته است که به برخی از آن‌ها در این جا اشاره می‌کنیم.

در (Hearst, 1991) با استفاده از یک الگوریتم به نام CatchWord مشخص می‌شود کدام یک از معانی از پیش تعیین شده باید به یک هم‌نگاره منتسب شود. این کار با بررسی زمینه اطراف هر کلمه و جمع‌آوری موارد از قبل مشاهده شده و انتخاب یک معنی که بیشترین احتمال را دارد، انجام می‌شود. حوزه این الگوریتم به اسامی محدود شده است و همچنین این الگوریتم در مواردی که زمینه مشابه وجود دارد مشکل دارد.

رده‌بندی‌کننده[16] Bayesian کلمات اطراف یک هم‌نگاره را به عنوان یک دسته[17] از کلمات مدل می‌کنند(Gale et al., 1992). تلفظ‌های متفاوت هم‌نگاره با توجه به ۱۰۰ کلمه نزدیک به هم‌نگاره از یکدیگر متمایز می‌شوند. این روش دو مشکل دارد. اولا فرض می‌کند که احتمال رخداد کلمات زمینه

[16] Classifier
[17] Bag

از یکدیگر مستقل هستند، ثانیا این روش نمی‌تواند بر اساس ساختار جمله عمل کند.

برچسب‌گذاری اجزای واژگانی کلام خود می‌تواند برای ابهام‌زدایی از دسته‌ای از هم‌نگاره‌ها به کار رود زیرا برخی از هم‌نگاره‌ها تلفظ‌های متفاوتی با برچسب واژگانی متفاوت دارند. به طور مثال هم‌نگاره "مرد" با دو تلفظ متفاوت /mard/ و /mord/ دارای برچسب‌های واژگانی متفاوت (یک اسم و دیگری فعل) می‌باشد. برچسب‌گذارهای N-gram (Merialdo, 1990) در سیستم تبدیل متن به گفتار AT&T (Sproat et al., 1992) برای برطرف سازی ابهام هم‌نگاره‌ها استفاده شده است. این برچسب‌گذار برچسب N کلمه قبل را مدل می‌کند. این مدل تنها می‌تواند بر اساس ساختار جمله حامل هم‌نگاره به طور مناسب عمل کند. برچسب‌گذار مبتنی بر تبدیل (Brill, 1995) همچنین برای ابهام‌زدایی از هم‌نگاره‌های انگلیسی استفاده شده است ( Wilks and Stevenson,1997).

لیست‌های تصمیم‌گیری (Yarowsky, 1994) یک روش ترکیبی حاصل از ترکیب رده‌بندی‌کننده Bayesian، برچسب‌گذار N-gram و درخت‌های تصمیم‌گیری است که نتایج خوبی در ابهام‌زدایی از هم‌نگاره‌های زبان انگلیسی از خود نشان داده است (رک.: فصل ۷، هم‌نگاره‌ها در زبان فارسی).

## رئوس مطالب فصول

بر اساس طرح پیشنهادی مسائل زیادی وجود دارد و این پایان‌نامه نیز بر اساس آن شامل مطالب زیادی می‌شود. در ادامه چون سیستم برچسب‌گذاری ما بر روی پیکره متنی زبان فارسی بنا شده است در فصل ۲ این پیکره را اجمالا شرح می‌دهیم. در فصل ۳ به برچسب‌گذاری با مدل‌های مارکوف می‌پردازیم. ما ابتدا برچسب‌گذاری با مدل مارکوف را مطرح می‌کنیم، چون استفاده از این مدل یک روش متداول برای برچسب‌گذاری بوده و نتایج خوبی نیز ارائه داده است. برای مقایسه نتایج برچسب‌گذاری بر اساس مدل مارکوف با یک مدل دیگر ما برچسب‌گذاری مبتنی بر حافظه را انتخاب کرده و آن را در فصل ۴ شرح می‌دهیم. فصل ۵ به تحلیل ساختواژی و تجزیه تصریفی کلمات فارسی و مباحثی راجع به ساختواژه کلمات فارسی می‌پردازد. برچسب‌گذاری کلمات ناشناخته بسیار مشکل می‌باشد. این مسئله در زبان فارسی به دلیل ساختار آن به مراتب مشکل‌تر است. یک روش برای مقابله با کلمات ناشناخته در فصل ۶ بررسی شده است. هم‌نگاره‌ها و ابهام‌زدایی از آن‌ها یک مسئله دیگر در

زبان فارسی است که در فصل ۷ به این مبحث می‌پردازیم. در فصل ۸ نتایج تجربی به دست آمده بیان می‌شود و نتیجه‌گیری نهایی و کارهای آتی در فصل ۹ ارائه می‌گردد.

شایان ذکر است که در طول این پایان‌نامه منظور از برچسب، برچسب اجزای واژگانی کلام، منظور از برچسب‌گذاری، برچسب‌گذاری اجزای واژگانی کلام و همچنین منظور از پیکره متنی زبان فارسی، بخش برچسب‌خورده این پیکره می‌باشد مگر آن‌که صریحا خلاف این ذکر شود.

# پیکره متنی زبان فارسی

## مقدمه

از آن جایی که برای فعالیت‌هایی که در این پایان‌نامه انجام گرفته از پیکره متنی زبان فارسی (بی‌جن‌خان، ۱۳۸۱) استفاده شده است لازم است که پیش از پرداختن به مطالب دیگر، قسمت برچسب‌خورده این پیکره، هرچند به صورت مختصر تشریح گردد. در (Atkins and Clear, 1992) تعریفی که برای پیکره ارائه شده است به این صورت است: حجم زیادی از داده‌های زبانی که براساس معیارهای مشخص برای هدف معینی جمع‌آوری و ذخیره شده‌اند به طوری که نمایندهٔ زبان یا گویش مورد مطالعه باشد. به طور کلی در طراحی و تهیه یک پیکره برچسب‌خورده یکی از مهمترین مسائلی که باید مورد توجه قرار گیرد مجموعه برچسب پیکره است که بر اساس هدف غایی پیکره و منظوری که از پیکره مد نظر است حاوی برچسب‌هایی خواهد شد که نیل به آن هدف و منظور را ممکن سازد. در طول انجام این پایان‌نامه پیکره متنی زبان فارسی دچار تغییراتی شد. مجموعه برچسب نسخه یک این پیکره بر اساس استاندارد ایگلز (Leech and Wilson, 1999) که یک استاندارد مشهور برای طراحی مجموعه برچسب پیکره می‌باشد به یک مجموعه برچسب جدید تبدیل شد و به تبع آن برچسب‌های کلمات نیز تغییر کرد. این تغییرات بر انجام این پایان‌نامه تاثیر زیادی گذاشت؛ از طرفی به علت گستردگی کارهای انجام شده بر روی نسخه یک پیکره، عدم ارائه نتایج حاصل از آن‌ها در پایان‌نامه ممکن نبود؛ و از طرفی نیز باید فعالیت‌ها بر اساس نسخه دو پیکره به روز می‌شد تا نتایج حاصل از پایان‌نامه قابلیت بیشتری برای مقایسه و تحلیل در کارهای دیگر را داشته باشد. به همین دلیل ما در این فصل هر دو نسخه پیکره متنی زبان فارسی را مورد بررسی قرار می‌دهیم تا امکان ارائه نتایج تجربی را بر روی دو نسخه پیکره فراهم کرده باشیم. منظور از نسخه یک پیکره، نسخه اولیه پیکره و منظور از نسخه دو پیکره، نسخه منطبق بر استاندارد ایگلز برای طراحی مجموعه برچسب می‌باشد.

ابتدا راجع به متون تشکیل‌دهنده پیکره مطلبی مختصر ارائه می‌دهیم و پس از آن مباحثی راجع

به مجموعه برچسب بیان می‌کنیم و سپس برچسب‌های پیکره متنی زبان فارسی و نحوه برچسب‌گذاری دستی این پیکره را شرح می‌دهیم. برای بررسی پیکره‌های زبانی توجه به آمارهایی راجع به تعداد کلمات، تعداد برچسب‌های متمایز کلمات و فراوانی برچسب‌های مختلف مهم می‌باشد که ما نیز برای ارائه یک دید کلی از پیکره متنی زبان فارسی این آمارها را در هر دو نسخه پیکره بیان می‌کنیم.

## متون تشکیل‌دهنده پیکره

پیکره متنی زبان فارسی شامل ۷.۵ میلیون کلمه برچسب‌خورده است. متون تشکیل دهنده پیکره یا متون رسمی است و یا متون غیر رسمی و محاوره‌ای. این متون برگرفته شده از اینترنت، پایان‌نامه‌ها، روزنامه‌ها، مجلات و کتب مختلف می‌باشد. همچنین این متون در موضوعات گوناگون می‌باشند مانند متون حقوقی، سیاسی، حسابداری، روانشناسی، آموزشی، ورزشی، دینی، اقتصادی، داستانی، ادبی. متونی که به صورت غیر رسمی و محاوره‌ای هستند بیشتر متون داستانی، ادبی و نمایش‌نامه‌ها 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.

## مجموعه برچسب

برچسب‌گذاری پیکره‌های زبانی به طور کلی می‌تواند در چهار سطح زبانی انجام شود که عبارتند از (بی‌جن‌خان، ۱۳۸۱):

۱- تعیین برچسب مقوله کلمه[1]

۲- برچسب‌گذاری نحوی: که شامل پردازش جملات و به دست آوردن درخت نحوی آن‌ها است.

۳- برچسب‌گذاری معنایی: که عبارت است از استخراج صورت منطقی جملات و به دست آوردن تعبیر معنایی آن‌ها. پیش‌نیاز این نوع برچسب‌دهی، انجام برچسب‌دهی مقوله کلمات است، به این تعبیر که پیش از آن‌که تعبیر معنایی کلمات و جملات یک متن به دست داده شود، تعیین مقوله کلمات آن‌ها ضروری است.

۴- برچسب‌گذاری کاربردشناختی[2]: که عبارت است از تعیین روابطی که میان دو کلمه در یک متن وجود دارد. به عنوان مثال مشخص کردن ضمایر و مرجع آن‌ها در متن در حوزه برچسب‌دهی کاربردشناختی قرار می‌گیرد.

بنا بر سطوح زبانی در برچسب‌گذاری پیکره‌های زبانی ابتدا باید یک مجموعه برچسب برای برچسب‌گذاری یک پیکره طراحی شود. تعیین برچسب مقولات اصلی اولین مرحله در تعیین برچسب‌های موجود در مجموعه برچسب است. بر این اساس در تهیه پیکره متنی زبان فارسی نیز سعی شده است مجموعه برچسب به گونه‌ای باشد که تمام مقوله‌های دستوری و انواع کلمات زبان فارسی را دربرگیرد و علاوه بر این مطابق با استانداردهای برچسب‌گذاری متون در زبان‌های دیگر مانند انگلیسی و زبان‌های اروپایی باشد.

برچسب‌هایی که برای یک پیکره در نظر گرفته می‌شود به طور کلی به سه دسته برچسب تقسیم می‌شوند که عبارتند از ( Cloern,1999):

۱- برچسب‌های نحوی-ساختواژی[3]: که اصلی‌ترین برچسب‌ها هستند. این برچسب‌ها شامل مقوله‌های نحوی اصلی از جمله فعل، اسم، صفت، قید و غیره هستند. اغلب کلماتی که در متون وجود دارند به یکی از این مقوله‌های اصلی تعلق دارند. اصلی‌ترین مقوله‌های نحوی که در اغلب پیکره‌های زبانی در نظر گرفته می‌شوند شامل مقوله اسم، صفت، حرف اضافه، حرف ربط، حرف تعریف، قید و عدد هستند.

[1] Wordclass tagging
[2] Pragmatic tagging
[3] Morphosyntactic

۲- برچسب‌های خاص[۴]: که شامل کلماتی هستند که در طبقه مقوله‌های اصلی قرار نمی‌گیرند، اما تعیین برچسب آن‌ها در استخراج اطلاعات زبانی از پیکره حائز اهمیت است. ادات شرط، حرف ندا، تکواژ صفت‌ساز از این دسته‌اند.

۳- برچسب‌های متفرقه[۵]: نیز شامل کلماتی است که در طبقه مقوله‌های اصلی قرار نمی‌گیرند اما در متن وجود دارند و به عنوان کلمات مجزا استخراج شده‌اند. کلمات خارجی، نشانه‌ها و علائم ریاضی و نیز علائم اختصاری در این طبقه قرار می‌گیرند.

مجموعه برچسب در نظر گرفته شده برای پیکره متنی زبان فارسی موارد فوق را شامل می‌شود و می‌توان برچسب‌های موجود در مجموعه برچسب را در این سه دسته قرار داد. در بخش بعد به این برچسب‌ها می‌پردازیم.

## برچسب‌های پیکره متنی

برچسب‌های پیکره متنی زبان فارسی در سه دسته تقسیم‌بندی می‌شوند(بی‌جن‌خان، ۱۳۸۱):

۱- برچسب‌های نحوی-ساختواژی: مانند اسم، فعل، صفت، قید، حرف ربط، حرف اضافه، حرف تعریف، ضمیر.

۲- برچسب‌های خاص: مانند ادات شرط، کیفیت نما، کلمه پرسشی، جمله‌واره، حرف ندا، منادی، تکواژ صفت‌ساز و عربی.

۳- برچسب‌های متفرقه: جداکننده، علامت ریاضی.

جدای از این تقسیم‌بندی برچسب‌های مجموعه برچسب پیکره متنی زبان فارسی یا برچسب اصلی‌اند یا برچسب‌های هستند که به عنوان زیربخش‌های برچسب‌های اصلی استفاده می‌شوند. به عنوان مثال کلمه‌ای مانند "کتابهایم" در پیکره به صورت زیر نمایش داده شده است:

کتابهایم N, COM, PL,1 N

[4] Unique tags
[5] Residual /Miscellaneous tags

ستون اول از چپ (N) برچسب اصلی کلمه، ستون دوم (N, COM, PL,1) برچسب سلسله‌مراتبی[۶] و ستون آخر (کتابهایم) خود کلمه می‌باشد.

## ۲-۴-۱. برچسب‌های اصلی

در نسخه اول پیکره متنی زبان فارسی تعداد برچسب‌های اصلی ۲۵ برچسب است که در جدول ۲-۱ نشان داده شده است. بعد از استانداردسازی برچسب‌ها بر اساس استاندارد ایگلز تعداد برچسب‌های اصلی به ۱۶ برچسب تقلیل یافته است که در جدول ۲-۲ قابل مشاهده می‌باشد.

مجموعه برچسب پیکره (نسخه ۱)

| برچسب | توصیف برچسب | برچسب | توصیف برچسب |
|---|---|---|---|
| N | اسم | Alpha-per | حرف الفبای فارسی |
| NP | گروه اسمی | Alpha-eng | حرف الفبای انگلیسی |
| OH | حرف ندا | ADJ | صفت |
| OHH | منادی | ADV | قید |
| P | حرف اضافه | AR | کلمات عربی |
| PP | گروه حرف اضافه‌ای | CON | حرف ربط |
| PRO | ضمیر | DELM | جدا کننده |
| PS | جمله‌واره | DET | حرف تعریف |
| QUA | سور | IF | ادات شرط |
| RA | حرف اضافه معرفه‌ای (را) | INT | حرف صوت |
| SPEC | کیفیت نما | MORP | تکواژ |
| SUBJ | موضوع متن | MS | علامت ریاضی |
| V | فعل | | |

[6] Hierarchical tag

مجموعه برچسب پیکره (نسخه ۲)

| برچسب | توصیف برچسب | برچسب | توصیف برچسب |
|---|---|---|---|
| ADJ | صفت | NUM | عدد |
| ADV | قید | P | حرف اضافه |
| CON | حرف ربط | PRO | ضمیر |
| DELM | جدا کننده | SUBJ | موضوع |
| DET | حرف تعریف | V | فعل |
| INT | حرف صوت | IDEN | شاخص |
| MORP | تکواژ | RES | متفرقه |
| N | اسم | PSTP | حرف اضافه پسین |

## ۲-۴-۲. زیربخش‌های مقولات اصلی

هر برچسب نحوی-ساختواژی ممکن است بر اساس نوع[۷]، جنس[۸]، شمار[۹]، شخص[۱۰] و غیره به زیربخش‌هایی تقسیم شود به طوری که اطلاعات کامل‌تری از کلمه را بیان کند. به عنوان مثال یک اسم در فارسی از لحاظ نوع می‌تواند خاص و عام باشد، از لحاظ شمار، مفرد یا جمع باشد؛ یک فعل در فارسی اصلی است یا کمکی، بر اساس شخص (۱ تا ۶) صرف می‌شود. در جدول ۲-۳ نمونه‌هایی از زیربخش‌های چند برچسب آورده شده است. با توجه به این زیربخش‌ها که برخی کاربرد نحوی و برخی کاربرد معنایی دارند می‌توان گفت برچسب‌های پیکره متنی برچسب‌های نحوی-معنایی می‌باشد.

ساختار زبان فارسی و روش‌های ساخت کلمات باعث می‌شود که زیربخش‌هایی به برچسب کلمات اضافه شوند که ذاتا جز زیربخش‌های یک مقوله خاص نیستند. به عنوان نمونه می‌توان به افزوده شده واژه‌بست‌ها به کلمات اشاره کرد (برای توضیح بیشتر رک.: فصل ۵، تحلیل‌گر ساختواژی).

[7] Type
[8] Genre
[9] Number
[10] Person

زیربخش‌هایی چند برچسب اصلی

| برچسب اصلی | زیربخش‌ها | | برچسب اصلی | زیربخش‌ها | |
|---|---|---|---|---|---|
| صفت | ساده<br>تفضیلی<br>عالی | | اسم | عام<br>خاص<br>جمع | مفرد<br>مکان<br>زمان |
| قید | کلی<br>مقدار<br>ساده<br>تفضیلی | پرسشی<br>تکرار<br>ترتیبی<br>مکان | فعل | کمکی<br>اسنادی<br>امری<br>ماضی | مضارع<br>آینده<br>استمراری<br>غیرتصریفی |

## ۲-۴-۳. ساختار برچسب‌های پیکره متنی

ساختار در نظر گرفته شده برای برچسب کلمات در پیکره متنی ساختار سلسله‌مراتبی است. در ساختار سلسله‌مراتبی، تمایز میان طبقات اصلی و زیربخش‌های آن‌ها نشان داده می‌شود. به این ترتیب که اگر برچسب از سمت راست به چپ خوانده شود، اصلی‌ترین برچسب در سمت راست قرار دارد و از راست به چپ جزئیات و اطلاعات دقیق‌تر به صورت زیربخش‌هایی به برچسب اصلی افزوده می‌شود. هر یک از طبقات اصلی و زیربخش‌های آن‌ها نیز به وسیله یک کاما یا نقطه از یکدیگر مجزا می‌شوند (بی‌جن‌خان، ۱۳۸۱).

در پیکره متنی زبان فارسی هر مقوله یا برچسب به وسیله کاما از زیربخش‌های خود مجزا می‌شود. به عنوان مثال برچسب «اسم,عام,مفرد,مکان,کسره اضافه» نشان می‌دهد که این برچسب به کلماتی منتسب می‌شود که مقوله آن‌ها اسم است، نوع آن‌ها عام و به لحاظ شمار مفرد هستند و به لحاظ معنایی در دسته اسم‌های مکان قرار می‌گیرند و نیز یک نشانگر نحوی کسره اضافه دارند، چه این کسره اضافه به صورت آشکار در خط ظاهر شود و چه کسره اضافه در خط ظاهر نشود.

# آمارهایی از پیکره متنی زبان فارسی

در این بخش آماری‌هایی از تعداد برچسب‌های متمایز و فراوانی آن‌ها برای پیکره متنی زبان فارسی (نسخه ۱ و ۲) ارائه می‌دهیم.

در نسخه ۱ پیکره همان‌گونه که ذکر شد ، تعداد ۲۵ برچسب اصلی وجود دارد. این برچسب‌ها به همراه فراوانی آن‌ها در جدول ۲-۴ آمده است. در این پیکره تعداد برچسب‌های متمایز (برچسب‌های سلسله‌مراتبی) ۵۵۹ می‌باشد. به عبارت دیگر کلمات پیکره در ۵۵۹ برچسب دسته‌بندی می‌شوند. جدول ۲-۵ تعدادی از برچسب‌های با فراوانی بیشتر را نشان می‌دهد.

در نسخه ۲ پیکره تعداد برچسب‌های اصلی ۱۶ برچسب است. این برچسب‌ها به همراه فراوانی آن‌ها در جدول ۲-۶ آمده است. در این پیکره تعداد برچسب‌های متمایز ۵۸۶ برچسب است و کلمات پیکره در این تعداد برچسب دسته‌بندی می‌شوند. جدول ۲-۷ تعدادی از برچسب‌های پیکره که بالاترین فراوانی را در پیکره داشته‌اند نشان می‌دهد.

برچسب‌های اصلی به همراه فراوانی آن‌ها در پیکره (نسخه ۱)

| برچسب | توصیف برچسب | فراوانی | برچسب | توصیف برچسب | فراوانی |
|---|---|---|---|---|---|
| N | اسم | 4283088 | Alpha-per | حرف الفبای فارسی | 1780 |
| NP | گروه اسمی | 54 | Alpha-eng | حرف الفبای انگلیسی | 3307 |
| OH | حرف ندا | 837 | ADJ | صفت | 1090371 |
| OHH | منادی | 238 | ADV | قید | 164503 |
| P | حرف اضافه | 1096396 | AR | کلمات عربی | 18268 |
| PP | گروه حرف اضافه‌ای | 419 | CON | حرف ربط | 831200 |
| PRO | ضمیر | 239743 | DELM | جدا کننده | 1022593 |
| PS | جمله‌واره | 1898 | DET | حرف تعریف | 177987 |
| QUA | سور | 66451 | IF | ادات شرط | 12671 |
| RA | حرف اضافه معرفه‌ای (را) | 145966 | INT | حرف صوت | 542 |
| SPEC | کیفیت نما | 133352 | MORP | تکواژ | 7687 |
| SUBJ | موضوع متن | 14978 | MS | علامت ریاضی | 788 |
| V | فعل | 868573 | | | |

چند برچسب سلسله‌مراتبی پیکره (نسخه ۱) با بیشترین فراوانی

| ردیف | برچسب | توصیف برچسب | فراوانی |
|---|---|---|---|
| ۱ | N,SING,COM,GEN | اسم.مفرد.عام.کسره اضافه | 1190381 |
| ۲ | N,SING,COM | اسم.مفرد.عام | 1148934 |
| ۳ | DELM | جدا کننده | 1022593 |
| ۴ | P | حرف اضافه | 983307 |
| ۵ | CON | حرف ربط | 822060 |
| ۶ | ADJ,SIM | صفت.ساده | 636287 |
| ۷ | N,PL,COM,GEN | اسم.جمع.عام.کسره اضافه | 350664 |
| ۸ | N,SING,NUM | اسم.مفرد.عدد | 252719 |
| ۹ | ADJ,SIM,GEN | صفت.ساده.کسره اضافه | 225883 |
| ۱۰ | N,SING,LOC,PR | اسم.مفرد.مکان.خاص | 204758 |

برچسب‌های اصلی به همراه فراوانی آن‌ها در پیکره (نسخه ۲)

| برچسب | توصیف برچسب | فراوانی | برچسب | توصیف برچسب | فراوانی |
|---|---|---|---|---|---|
| ADJ | صفت | 927888 | NUM | عدد | 306353 |
| ADV | قید | 16594 | P | حرف اضافه | 1081619 |
| CON | حرف ربط | 831128 | PRO | ضمیر | 236566 |
| DELM | جدا کننده | 1005309 | SUBJ | موضوع | 1546 |
| DET | حرف تعریف | 245042 | V | فعل | 968019 |
| INT | حرف صوت | 788 | IDEN | شاخص | 20873 |
| MORP | تکواژ | 4067 | RES | متفرقه | 30276 |
| N | اسم | 3917090 | PSTP | حرف اضافه پسین | 143588 |

چند برچسب سلسله‌مراتبی پیکره (نسخه ۲) با بیشترین فراوانی

| ردیف | برچسب | توصیف برچسب | فراوانی |
|---|---|---|---|
| ۱ | N,COM,SING,GEN | اسم,عام,مفرد,کسره اضافه | 1204952 |
| ۲ | N,COM,SING | اسم,عام,مفرد | 1141081 |
| ۳ | DELM | جدا کننده | 1005309 |
| ۴ | P | حرف اضافه | 966708 |
| ۵ | CON | حرف ربط | 811728 |
| ۶ | ADJ,SIM | صفت,ساده | 622496 |
| ۷ | N,COM,PL,GEN | اسم,عام,جمع,کسره اضافه | 345081 |
| ۸ | NUM,CAR,NOMI,SING | عدد,اصلی,اسمی,مفرد | 242312 |
| ۹ | ADJ,SIM,GEN | صفت,ساده,کسره اضافه | 224767 |
| ۱۰ | N,PR,SING | اسم,خاص,مفرد | 205704 |

# نتیجه‌گیری

در این بخش توضیحاتی راجع به قسمت برچسب‌خورده پیکره متنی زبان فارسی ارائه شد و مشخصات کلی آن از نظر گذشت. ابتدا مطالبی در مورد مجموعه برچسب و خصوصیات برچسب‌های مختلف بیان شد و سپس نحوه برچسب‌گذاری پیکره متنی زبان فارسی تشریح شد. چون بخشی از نتایج ما بر روی نسخه اول پیکره و بخشی نیز بر روی نسخه دوم پیکره است سعی شد یک دید کلی نسبت به این دو نسخه پیکره متنی زبان فارسی ارائه شود. برای این کار هر دو نسخه پیکره مختصرا شرح داده شد و برچسب‌های اصلی آن‌ها به همراه گزارش آمارهایی در مورد فراوانی برچسب‌های موجود در دو نسخه ارائه گردید.

# برچسب‌گذاری با مدل‌های مارکوف

# مقدمه

مدل‌های مخفی مارکوف بدون شک یکی از پرکاربردترین روش‌های مورد استفاده برای برچسب‌گذاری اجزا واژگانی کلام بوده‌اند. این ادعا با رجوع به منابع موجود مرتبط با برچسب‌گذاری در زبان‌های مختلف کاملا مشهود است. اساس یک مدل مخفی مارکوف یک تابع احتمالی از یک زنجیر مارکوف[۱] می‌باشد. زنجیر مارکوف که به فرآیند ماکوف[۲] یا مدل مارکوف[۳] نیز شهرت دارد، ابتدا توسط Andrei A. Markov معرفی شد (Markov, 1913). ابتدا مدل مارکوف برای اهداف زبان‌شناسی مورد استفاده قرار گرفت ولی بعدها به عنوان یک ابزار آماری همگانی توسعه یافت.

چون اساس مدل‌های مخفی مارکوف زنجیر مارکوف می‌باشد در این فصل ابتدا زنجیر مارکوف بیان و مفروضات حاکم بر آن توصیف می‌گردد. مدل مخفی مارکوف یک مدل توسعه یافته از زنجیر مارکوف است که به همراه اجزا آن بررسی می‌شود.

برای استفاده از هر مدل آماری برای حل هر مسئله، ابتدا باید مسئله در قالب آن مدل آماری بیان شود و اجزا مسئله بر اجزا تشکیل دهنده مدل آماری انطباق یابد. لذا پس از آشنایی با مدل‌های مخفی مارکوف نحوه استفاده از آن‌ها را در برچسب‌گذاری بیان می‌کنیم. آموزش[۴] یک برچسب‌گذار مبتنی بر مدل مخفی مارکوف از روی پیکره برچسب‌خورده انجام می‌شود. هدف غایی از یک سیستم، استفاده از آن در راستای هدفی مشخص می‌باشد که باید پس از آموزش صورت گیرد. الگوریتم Viterbi یک الگوریتم رمزگشایی[۵] است که برای استفاده از سیستم‌های مبتنی بر مدل‌های مخفی مارکوف ارائه شده است. این الگوریتم نیز در ادامه مطلب بحث می‌گردد.

---

[1] Markov chain
[2] Markov process
[3] Markov model
[4] Training
[5] Decoding

یک مشکل اساسی که در سیستم‌های آماری به وجود می‌آید مشکل پراکندگی داده[۶] می‌باشد. این مشکل تاثیر زیادی بر کارایی سیستم‌ها دارد. برای حل این مشکل روش‌هایی به نام روش‌های هموارسازی[۷] استفاده می‌شود. ما چند روش هموارسازی را ذکر می‌کنیم و روش مورد استفاده خود و نحوه اعمال آن را در سیستم برچسب‌گذاری مبتنی بر مدل مخفی مارکوف بیان می‌داریم.

# مدل‌های مارکوف

## ۳-۲-۱. زنجیر مارکوف

فرض کنید $X=(X_1,\ldots,X_T)$ دنباله‌ای از متغیرهای تصادفی باشد که مقادیر خود را از مجموعه محدود $S=(s_1,\ldots,s_N)$ (فضای حالت) اختیار می‌کنند. خواص مارکوف عبارتند از:

۱- افق محدود[۸]:

$$P(X_{t+1}=s_k \mid X_1,\ldots,X_t)=P(X_{t+1}=s_k \mid X_t)$$

۲- مستقل از زمان بودن (ایستانی)[۹]:

$$P(X_{t+1}=s_k \mid X_t)=P(X_2=s_k \mid X_1)$$

[6] Data sparsity
[7] Smoothing methods
[8] Limited Horizon
[9] Time invariant (stationary)

آنگاه X را یک زنجیر مارکوف می‌نامند یا به عبارت دیگر x دارای خاصیت مارکوف است. زنجیر مارکوف را می‌توان با ماتریس انتقال A نشان داد که دارایه‌های آن عبارت است از:

$$a_{ij} = P\left(X_{t+1} = s_j \mid X_t = s_i\right)$$

که در آن $a_{ij} \geq 0, \forall i, j$ و $\sum_{j=1}^{N} a_{ij} = 1, \forall i$.

به علاوه احتمالات شروع حالات اولیه برای زنجیر مارکوف با $\Pi$ نشان داده می‌شود:

$$\pi_1 = P\left(X_1 = s_i\right)$$

که در آن $\sum_{i=1}^{N} \pi_i = 1$.

## ۳-۲-۲. مدل مخفی مارکوف

مدل مخفی مارکوف یک نوع خاص از زنجیر مارکوف می‌باشد. یک مدل مخفی مارکوف با پنج‌تایی $(S, K, \Pi, A, B)$ تعریف می‌شود که در آن، $S = \{s_1, \ldots, s_N\}$ مجموعه حالات، $K = \{k_1, \ldots, k_N\}$ مجموعه نشانه‌های خروجی، $\Pi = \{\pi_i\}, i \in S$ ماتریس احتمالات حالت اولیه، $A$ ماتریس احتمالات انتقال و $B$ ماتریس احتمالات خروجی می‌باشد.

یک مسیر در یک مدل مخفی مارکوف دنباله‌ای از انتقالات پی‌درپی است به طوری که حالت نهایی یک انتقال، حالت شروع انتقال بعدی در مسیر می‌باشد. احتمال یک مسیر حاصل‌ضرب احتمالات انتقالات می‌باشد. یک دنباله خروجی می‌تواند به وسیله چندین مسیر تولید شود ولی همیشه مسیری وجود دارد که محتمل‌ترین مسیر برای تولید این خروجی می‌باشد.

مدل‌های مخفی مارکوف بیشتر به دو روش استفاده می‌شوند. اول این‌که احتمال یک دنباله از خروجی‌ها برای چندین مدل محاسبه می‌شود تا مشخص شود کدام مدل احتمال بیشتری برای تولید

این دنباله از خروجی‌ها را دارد. یک مثال از این مورد تشخیص گفتار می‌باشد که در آن یک سیگنال صوتی با یک مدل مقایسه می‌شود تا مشخص شود که چه کلمه‌ای گفته شده است. در روش دوم، از مدل برای تعیین این‌که کدام مسیر برای تولید یک دنباله خروجی خاص طی شده است، استفاده می‌شود. این همان روشی است که از مدل مخفی مارکوف برای برچسب‌گذاری اجزا واژگانی کلام استفاده می‌شود که در بخش بعد به آن پرداخته شده است.

# مدل مخفی مارکوف برای برچسب‌گذاری

## ۳-۳-۱. مدل احتمالی

در برچسب‌گذاری با مدل مخفی مارکوف (Kupiec, 1992) (Charniak et al., 1993)، دنباله برچسب‌ها در یک متن به عنوان یک زنجیر مارکوف در نظر گرفته می‌شود. همان‌گونه که ذکر شد یک زنجیر مارکوف دو خاصیت افق محدود و مستقل از زمان بودن را داراست. تفسیر این دو خاصیت در برچسب‌گذاری با مدل مارکوف به این صورت است که ما فرض می‌کنیم برچسب یک کلمه تنها وابسته به برچسب کلمه قبلی است (افق محدود) و این وابستگی در طول زمان تغییر نمی‌کند (مستقل از زمان بودن). برای مثال، اگر یک صفت در اوایل جمله با احتمال ۰.۲ بعد از یک اسم ظاهر می‌شود، این احتمال در حین برچسب‌گذاری بقیه جمله یا یک جلمه دیگر تغییر نمی‌کند و ثابت فرض می‌شود. اگرچه پر واضح است که این دو خاصیت مارکوف چندان منطبق بر واقعیت نمی‌باشد زیرا به عنوان مثال خاصیت اول وابستگی‌های با فاصله زیاد را بین برچسب کلمات نادیده می‌گیرد. به طور مثال در یک جمله فارسی ارتباط ضمیر که اول جمله ظاهر می‌شود با فعلی که در آخر جمله می‌آید نادیده انگاشته می‌شود، و این فرضی ناصحیح است. ولی با وجود این مدل‌های مخفی مارکوف کارایی بسیار خوبی در کاربردهای حوزه پردازش زبان طبیعی از خود نشان داده‌اند.

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 $t_i...t_{i+m}$ برای کلمات $w_i...w_{i+m}$ |
| $w^l$ | کلمه l در واژگان |
| $t^j$ | jامین برچسب در مجموعه برچسب |
| $C(w^l)$ | تعداد رخداد $w^l$ در مجموعه آموزش[۱۰] |
| $C(t^j)$ | تعداد رخداد $t^j$ در مجموعه آموزش |
| $C(t^j,t^k)$ | تعداد رخداد $t^k$ بعد از $t^j$ |
| $C(t^i,t^j,t^k)$ | تعداد رخداد $t^k$ بعد از $t^j$ و $t^i$ |
| $C(w^l:t^j)$ | تعداد رخداد $w^l$ با برچسب $t^j$ |
| T | تعداد برچسب‌های در مجموعه برچسب |
| W | تعداد کلمات در واژگان |
| n | طول جمله |

جدول ۳-۱ نشان‌گذاری مورد استفاده را نمایش می‌دهد. فرض می‌کنیم که $\{w^1,w^2,\ldots,w^w\}$ یک مجموعه از کلمات در 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:

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

خاصیت افق محدود به صورت زیر بیان می‌شود:

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

رابطه فوق همان احتمالات انتقال می‌باشد. علاوه بر فرض افق محدود دو فرض دیگر راجع به

[10] Training set

کلمات در نظر می‌گیریم:

۱- کلمات از یکدیگر مستقل‌اند.

۲- یک کلمه مانند w تنها وابسته به برچسب خودش می‌باشد.

بنابراین، بر طبق مفروضات فوق می‌توان نوشت:

$$P(w_{1,n} \mid t_{1,n})P(t_{1,n}) = \prod_{i=1}^{n}[P(w_i \mid t_i) \times P(t_i \mid t_{i-1})]$$

فرض می‌کنیم هر جمله با کلمه فرضی SOS(Start Of Sentence) شروع می‌شود، بنابراین می‌توان تعریف کرد $P(t_1 \mid t_0) = 1.0$.

با توجه به تساوی‌های فوق، تساوی نهایی برای به دست آوردن برچسب‌های کلمات یک جمله به صورت زیر خواهد بود:

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

برای آموزش مدل فوق باید احتمالات انتقال و احتمالات خروجی از روی پیکره برچسب خورده به دست آید. احتمالات انتقال به سادگی به صورت زیر به دست می‌آید:

$$P(t^k \mid t^j) = \frac{C(t^j, t^k)}{C(t^j)}$$

و همچنین احتملات خروجی به صورت زیر محاسبه می‌شود:

$$P(w^l \mid t^j) = \frac{C(w^l, t^j)}{C(t^j)}$$

## ۳-۳-۲. برچسب‌گذار bigram و trigram

مدل مخفی مارکوف شرح داده شده در بخش قبل به برچسب‌گذار bigram [11] مشهور است، زیرا با اعمال خاصیت افق محدود فرض می‌شود هر برچسب تنها وابسته به برچسب قبل از خود است (یعنی مشابه مدل‌های n-gram با n=2). همان‌طور که بیان شد این فرض مبتنی بر واقعیت نمی‌باشد. می‌توان این وابستگی را گسترش داد و برچسب هر کلمه را وابسته به دو برچسب قبل فرض کرد و با این کار به دقت مدل و انطباق بیشتر آن با واقعیت افزود. این مدل جدید به برچسب‌گذار trigram [12] معرف می‌باشد. برای این کار رابطه ۳-۶ به صورت زیر تغییر می‌کند:

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

و به همین ترتیب رابطه ۳-۷ به صورت زیر تغییر می‌یابد:

$$P(w_{1,n} \mid t_{1,n}) P(t_{1,n}) = \prod_{i=2}^{n+1} [P(w_i \mid t_i) \times P(t_i \mid t_{i-2} t_{i-1})]$$

با توجه به دو رابطه فوق و رابطه ۳-۵، رابطه نهایی برچسب‌گذار trigram به صورت زیر به دست می‌آید:

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

برای آموزش برچسب‌گذار trigram تنها رابطه ۳-۹، احتمالات انتقال، باید تغییر کند:

[11] Bigram tagger
[12] Trigram tagger

$$P\left(t^k \mid t^i t^j\right) = \frac{C\left(t^i, t^j, t^k\right)}{C\left(t^i, t^j\right)}$$

## ۳-۳-۳. الگوریتم Viterbi

در رابطه ۳-۸ (یا در رابطه ۳-۱۳) می‌توان با محاسبه همه برچسب‌گذاری‌های $t_{1,n}$ ممکن برای یک جمله به طول n، $\hat{t}_{1,n}$ را به دست آورد. ولی با این روش پیچیدگی محاسباتی برچسب‌گذاری نسبت به طول جمله ورودی به صورت نمایی خواهد. یک الگوریتم کارا برای استفاده در مدل‌های مخفی مارکوف، الگوریتم Viterbi می‌باشد که بر اساس برنامه‌نویسی پویا طراحی شده است. این الگوریتم دارای سه مرحله می‌باشد: مقداردهی اولیه، استقرا، خاتمه و استخراج مسیر. تابع $\delta_i(j)$ احتمال حالت j (=برچسب j) را برای کلمه iام می‌دهد و تابع $\psi_{i+1}(j)$ با این فرض که در کلمه i+1 ام در حالت j هستیم محتمل‌ترین حالت (برچسب) را برای کلمه iام مشخص می‌کند به عبارت دیگر این تابع تعیین می‌کند کدام برچسب در کلمه قبل محتمل‌ترین انتقال را به برچسب j برای کلمه فعلی منجر می‌شود. شکل ۳-۱ الگوریتم Viterbi را برای برچسب‌گذار bigram نشان می‌دهد. در این شکل سه مرحله این الگوریتم مشخص شده و نحوه استفاده توابع مورد بحث نیز به روشنی ذکر شده است.

//Given a sentence of length n
//Initialization
$\delta_1(SOS) = 1.0$
$\delta_1(t) = 0.0$ for $t \neq SOS$
//Induction
for i=1 to n do
  for all tags $t^j$ do
    $\delta_{i+1}(t^j) = \max_{1 \le k \le T} \left[ \delta_i(t^k) \times P(t^j \mid t^k) \times P(w_{i+1} \mid t^j) \right]$
    $\psi_{i+1}(t^j) = \arg\max_{1 \le k \le T} \left[ \delta_i(t^k) \times P(t^j \mid t^k) \times P(w_{i+1} \mid t^j) \right]$
  end
end
//Termination and path extraction
$X_{n+1} = \arg\max_{1 \le j \le T} \delta_{n+1}(t^j)$
for j=n to 1 step -1 do
  $X_j = \psi_{j+1}(X_{j+1})$
end
$P(X_1, \ldots, X_n) = \max_{i \le j \le T} \delta_{n+1}(t^j)$

الگوریتم Viterbi برای برچسب‌گذار bigram

الگوریتم Viterbi برای برچسب‌گذار trigram مشابه الگوریتم Viterbi برای برچسب‌گذار bigram است و تفاوت در تعریف توابع $\delta_i(j)$ و $\psi_{i+1}(j)$ می‌باشد زیرا در برچسب‌گذار trigram هر برچسب وابسته به دو برچسب قبلی در نظر گرفته می‌شود بنابراین با تغییر اندکی در الگوریتم فوق الگوریتم Viterbi برای برچسب‌گذار trigram به دست می‌آید، لذا از بیان این الگوریتم به دلیل جلوگیری از اطناب خودداری می‌کنیم.

## ۳-۳-۴. پراکندگی داده و هموارسازی

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

در برچسب‌گذارهای مبتنی بر مدل مخفی مارکوف (bigram و trigram) تعداد پارامترهای مدل متناسب با تعداد کلمات در واژگان و تعداد برچسب‌های موجود در مجموعه برچسب می‌باشد. اگرچه به دلیل تعداد زیاد پارامترها در برچسب‌گذار bigram و trigram، پراکندگی داده از کارایی این دو برچسب‌گذار می‌کاهد ولی چون تعداد پارامترها در برچسب‌گذار trigram بیشتر از برچسب‌گذار bigram است (در برچسب‌گذار trigram آرایه احتمالات انتقال یعنی آرایه A سه بعدی است ولی در برچسب‌گذار bigram این آرایه دو بعدی می‌باشد) در برچسب‌گذار trigram این مشکل به مراتب نمود بیشتری از خود نشان می‌دهد. برای حل این مشکل از روش‌های هموارسازی استفاده می‌شود.

روش‌های زیادی برای هموارسازی ارائه شده است که هر کدام در مواردی کارایی بهتری از خود نشان می‌دهند. از جمله این روش‌ها روش افزایشی[۱۳] (Gale and Church, 1994)، روش Good-Turning (Good, 1953)، روش Jelinek-Mercer (Jelinek and Mercer, 1980) و روش Katz (Katz,1987) می‌باشد.

روشی که ما مورد استفاده قرار دادیم روشی است که در (Thede and Harper, 1999) بحث شده است. به دو دلیل ما از این روش استفاده می‌کنیم: این روش در عمل هم کارایی مناسبی از خود نشان می‌دهد و همچنین محاسبات آن نسبت به روش‌هایی دیگر ساده می‌باشد. در هنگام آموزش مدل، برای تعیین احتمالات انتقال در برچسب‌گذار bigram به جای رابطه ۳-۹ از رابطه زیر استفاده می‌کنیم:

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

و به جای رابطه ۳-۱۴ در برچسب‌گذار trigram رابطه زیر را به کار می‌بریم:

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

که در دو رابطه فوق:

[13] Additive method

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

و

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

# نتیجه‌گیری

در این بخش ابتدا به طور مختصر به مبحث زنجیر مارکوف پرداختیم و مبنای تئوریکی آن را بیان کردیم. همان‌طور که ذکر شد زنجیر مارکوف اساس مدل‌های مخفی مارکوف است. سپس مدل‌های مخفی مارکوف مورد بررسی قرار گرفت و پس از آن چگونگی بیان مسئله برچسب‌گذاری در قالب مدل‌های مخفی مارکوف طرح و نحوه استفاده از این مدل‌ها برای برچسب‌گذاری شرح داده شد. دو مدل رایج برچسب‌گذارهای مبتنی بر مدل مخفی مارکوف برچسب‌گذارهای bigram و trigram می‌باشد که مفهوم آن‌ها بیان گردید. نحوه آموزش یک برچسب‌گذار مبتنی بر مدل‌های مارکوف و همچنین الگوریتم رمزگشایی برای مدل‌های مخفی مارکوف یعنی الگوریتم Viterbi مورد توجه قرار گرفت. از جمله مشکلات توام با مدل‌های آماری مانند مدل‌های مارکوف مشکل پراکندگی داده می‌باشد که برای حل این مشکل از روش‌های هموارسازی استفاده می‌شود. ما در پایان این فصل به این روش‌ها پرداختیم و روش مورد استفاده خود را در نتایج تجربی بیان داشتیم.

# برچسب‌گذار مبتنی بر حافظه

# مقدمه

در این فصل به برچسب‌گذاری مبتنی بر حافظه می‌پردازیم. روش یادگیری مبتنی بر حافظه یکی از روش‌های یادگیری ماشین است که به صورت باناظر[1] عمل می‌کند. این روش یادگیری سعی می‌کند با یادگیری اطلاعاتی از روی نمونه‌های قبلی، راجع به نمونه‌های جدید تصمیم‌گیری کند. لذا برای برچسب‌گذاری مبتنی بر حافظه نیاز به پیکره آموزشی است تا از آن اطلاعاتی استخراج شود که برای برچسب‌گذاری متون جدید استفاده شود. این برچسب‌گذار مزایایی از قبیل عدم نیاز به پیکره‌های خیلی بزرگ برای یادگیری، آموزش و برچسب‌گذاری سریع، طبیعت غیر پارامتریک آن و کارایی مناسب در برابر دیگر برچسب‌گذارها دارد. برچسب‌گذار مبتنی بر حافظه از جهتی مشابه برچسب‌گذارهای مبتنی بر قانون است و از جهتی هم مشابه رده‌بندهای آماری. هر نمونه ذخیره شده در حافظه را می‌توان یک قانون خاص و هر استدلال مبتنی بر تشابه را می‌توان به عنوان شکلی از مکانیزم انتخاب قانون در نظر گرفت و از این جهت این برچسب‌گذار مشابه برچسب‌گذارهای مبتنی بر قانون است. روش یادگیری مبتنی بر حافظه برای یافتن کلاس یک نمونه جدید از روش نزدیکترین $k$ همسایه[2] که یک روش معروف در شناسایی آماری الگو می‌باشد، استفاده می‌کند و از این جهت برچسب‌گذار مبتنی بر حافظه را می‌توان مشابه رده‌بندهای آماری در نظر گرفت.

[1] Supervised
[2] k-nearest neighbors (knn)

## یادگیری مبتنی بر حافظه

یادگیری مبتنی بر حافظه شکلی از یادگیری قیاسی[۳] باناظر است. نمونه‌هایی که یادگیری از روی آن‌ها انجام می‌شود با برداری از ویژگی‌ها نمایش داده می‌شوند و هر نمونه یک برچسب[۴] خاص دارد که نشانگر کلاس نمونه می‌باشد. در طی عمل آموزش، مجموعه‌ای از نمونه‌ها (مجموعه آموزش) به رده‌بند داده شده و به حافظه افزوده می‌گردد. برای رده‌بندی یک نمونه جدید، فاصله آن نمونه با نمونه‌های موجود در حافظه محاسبه می‌شود و برچسب نمونه (نمونه‌های) با حداقل فاصله برای پیش‌بینی برچسب نمونه جدید استفاده می‌شود.

بنابراین در برچسب‌گذاری مبتنی بر حافظه، عمل برچسب‌گذاری در اصل به عمل رده‌بندی نمونه‌های جدید بر اساس نمونه‌های قبلی تبدیل می‌شود که نمونه‌ها در آن، کلمات می‌باشند.

## برچسب‌گذار مبتنی بر حافظه

مسئله مهم در برچسب‌گذار مبتنی بر حافظه معیار فاصله یا به عبارت دیگر معیار شباهت[۵] می‌باشد که ابتدا به آن می‌پردازیم و سپس ساختار برچسب‌گذار شرح داده می‌شود.

### ۴-۳-۱. معیار شباهت

کارایی یک برچسب‌گذار مبتنی بر حافظه به معیار فاصله (شباهت) وابسته می‌باشد. معیار فاصله را می‌توان مانند زیر بدون وزن در نظر گرفت و به همه ویژگی‌ها وزن یکسانی داد:

---

[3] Inductive learning
[4] Lable
[5] Similarity measure

$$\Delta(X,Y)=\sum_{i=1}^{n}\delta(x_i,y_i)$$

که X و Y دو نمونه‌ای هستند که باید فاصله بین آن‌ها محاسبه شود و $\delta(x_i,y_i)$ فاصله بین مقادیر ویژگی iام در نمونه‌های با n ویژگی است. فاصله بین دو ویژگی نیز به صورت زیر محاسبه می‌شود:

$$\delta(x_i,y_i)=\begin{cases}0 & if\ x_i=y_i\\ 1 & otherwise\end{cases}$$

که برای تعیین فاصله بین مقادیر ویژگی‌های غیر عددی استفاده می‌شود (در برچسب‌گذاری نیز هیچ ویژگی با مقادیر عددی وجود ندارد). این روش به IB1 (Aha et al., 1991) شناخته می‌شود. این معیار فاصله برای برچسب‌گذاری مناسب نیست زیرا ویژگی‌های مختلف در برچسب‌گذاری وزن یکسانی ندارند و برخی ویژگی‌ها از ارزش بیشتری برخوردارند. به عنوان مثال خود کلمه از کلمات دیگر در زمینه مهم‌تر است. بنابراین هر ویژگی با بهره اطلاعاتی[6] آن ویژگی وزن‌دهی می‌شود. بهره اطلاعاتی عددی است که مقدار میانگین کاهش اینروپی اطلاعات مجموعه آموزشی را هنگامی که مقدار ویژگی معلوم باشد، نشان می‌دهد (Daelemans et al., 1996). این معیار فاصله، IB-IG نامیده می‌شود و در برچسب‌گذاری مبتنی بر حافظه از آن استفاده می‌شود و به صورت زیر تعریف می‌گردد:

$$\Delta(X,Y)=\sum_{i=1}^{n}G(f_i)\delta(x_i,y_i)$$

[6] Information Gain

## ۴-۳-۲. ساختار برچسب‌گذار مبتنی بر حافظه

یک برچسب‌گذار مبتنی بر حافظه دارای سه بخش اصلی است: واژگان، یک پایگاه نمونه[۷] برای کلمات شناخته شده (کلمات حاضر در پیکره) و یک پایگاه نمونه برای کلمات ناشناخته. این سه بخش از روی پیکره آموزشی ساخته می‌شود.

### ۴-۳-۲-۱. ساخت واژگان

واژگان از کلمات حاضر در پیکره آموزشی و با محاسبه فراوانی رخداد هر کلمه با هر برچسب به دست می‌آید. به عبارت دیگر برای هر کلمه برچسب‌هایی که به آن‌ها منتسب شده و تعداد تکرار آن‌ها باید از پیکره آموزشی استخراج شود.

### ۴-۳-۲-۲. کلمات شناخته شده

همان‌گونه که قبلا ذکر شد عمل برچسب‌گذاری مبتنی بر حافظه به صورت یک مسئله رده‌بندی می‌باشد. برای این کار از روش پنجره[۸] (Sejnowski and Rosenberg, 1987) استفاده می‌شود. یک نمونه می‌تواند شامل اطلاعاتی راجع به خود کلمه، زمینه راست و زمینه چپ کلمه و برچسب‌های کلمه در پیکره باشد.

بنابراین اطلاعات زیادی وجود دارد که می‌تواند در پایگاه نمونه ذخیره شود. در مورد کلمات شناخته‌شده معمولا از اطلاعاتی مانند خود کلمه، برچسب‌های منتسب شده به آن در پیکره، کلمات در زمینه سمت راست و برچسب‌های ابهام‌زدایی شده در زمینه سمت راست (که از نتیجه

[7] Case base
[8] Windowing approach

برچسب‌گذاری کلمات قبل حاصل شده است) کلمات در زمینه سمت چپ و برچسب‌های مبهم آن‌ها استفاده می‌شود. اندازه پنجره در الگوریتم می‌تواند بسته به اطلاعات موجود در زمینه حتی به صورت پویا نیز تغییر کند.

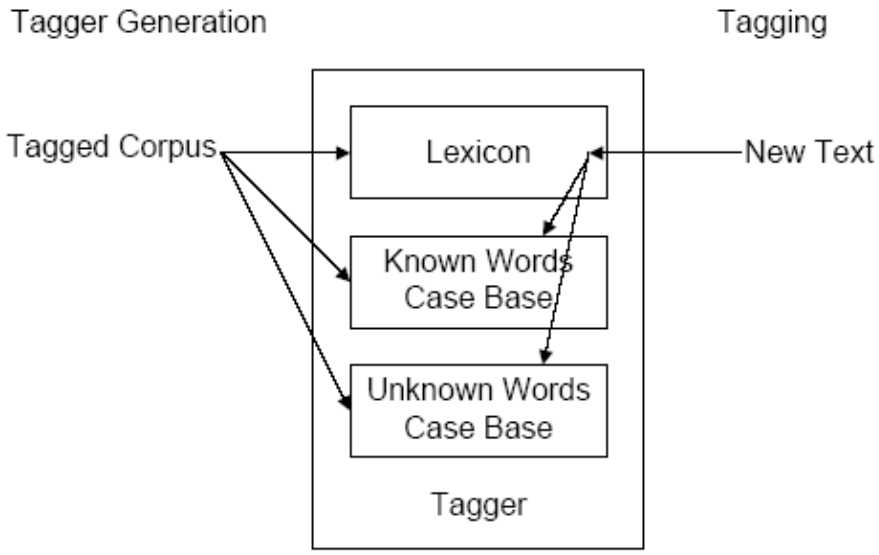
### ۴-۳-۲-۳. کلمات ناشناخته

اگر یک کلمه در واژگان حضور نداشته باشد نمی‌توان برچسب‌هایی که کلمه به آن‌ها منتسب می‌شود را بازیابی کرد. در این صورت برچسب کلمه را تنها می‌توان با توجه به شکل کلمه یا زمینه آن حدس زد. در این حالت از ترکیب اطلاعات استخراجی از شکل کلمه و زمینه‌ای که در آن رخ داده استفاده می‌شود تا برچسب‌هایی که کلمه می‌تواند به آن‌ها منتسب شود تعیین شود. پس از تعیین این برچسب‌ها، با کلمه ناشناخته مانند یک کلمه شناخته شده رفتار می‌شود.

در این‌جا می‌توان از چند حرف ابتدایی و یا انتهایی کلمه، کلمات در زمینه سمت راست و برچسب‌های ابهام‌زدایی شده در زمینه سمت راست، کلمات در زمینه سمت چپ و برچسب‌های مبهم آن‌ها برای ابهام‌زدایی از کلمات ناشناخته بهره برد.

### ۴-۳-۲-۴. عملکرد

شکل ۴-۱ (Daelemans et al., 1996) ساختار برچسب‌گذار مبتنی بر حافظه را نشان می‌دهد. برچسب‌گذار با استخراج واژگان و دو پایگاه نمونه از پیکره برچسب‌خورده ساخته می‌شود. در طی عمل برچسب‌گذاری به صورت زیر عمل می‌شود: کلمات در واژگان جستجو و به دو دسته کلمات شناخته شده و ناشناخته تقسیم می‌شوند. سپس نمونه‌ها از پایگاه نمونه کلمات شناخته شده و پایگاه نمونه کلمات ناشناخته بازیابی می‌شوند و بر اساس معیار شباهت برچسب کلمه تعیین می‌شود.




ساختار برچسب‌گذار مبتنی بر حافظه (Daelemans et al. 1996)

# نتیجه‌گیری

در این فصل ابتدا روش یادگیری مبتنی بر حافظه به طور مختصر تشریح شد. سپس به قسمت‌های برچسب‌گذار مبتنی بر حافظه پرداخته شد و ساختار آن مورد بررسی قرار گرفت. تعیین معیار شباهت یکی از اساسی‌ترین مواردی است که باید در نظر گرفته شود. واژگان و پایگاه نمونه کلمات ناشناخته و پایگاه کلمات شناخته شده ساختار برچسب‌گذار مبتنی بر حافظه را تشکیل می‌دهند. در ساخت پایگاه‌های نمونه، نوع اطلاعاتی که باید در آن‌ها قرار گیرد بسیار مهم است و برای تعیین این اطلاعات باید به صورت تجربی عمل کرد تا مشخص شود ذخیره کدام اطلاعات برای کلمات شناخته شده و کلمات ناشناخته نتیجه بهتری ارائه می‌دهد. نتایج این برچسب‌گذار در مقایسه با نتایج برچسب‌گذاری با روش‌های مبتنی بر مدل مارکوف در فصل نتایج تجربی ارائه می‌گردد.

# تحلیل‌گر ساختواژی

## مقدمه

در زبان‌هایی که ساختواژه تصریفی و اشتقاقی[1] آن‌ها پیچیده می‌باشد ظرفیت بالایی برای ساخت کلمات با اشکال جدید وجود دارد. زیرا در این زبان‌ها تکواژها به روش‌های گوناگون به یکدیگر متصل می‌شوند و کلمات جدیدی تولید می‌کنند که امکان دارد این کلمات حتی در پیکره‌های بزرگ وجود نداشته باشند و یا فراوانی آن‌ها بسیار اندک باشد. تکواژهای تصریفی[2] معمولا برای ساخت کلماتی استفاده می‌شوند که حامل مفاهیم دستوری در جمله باشند و تکواژهای اشتقاقی[3] کلمات جدید را می‌سازند که بار مفهومی خاصی را منتقل می‌کنند. گویشوران این زبان‌ها چون آشنایی کامل با قواعد زبانی دارند معمولا هنگام مواجه با کلمات این‌چنینی حتی اگر قبلا با این کلمات برخورد نکرده باشند، دچار مشکل نمی‌شوند. ولی این مطلب راجع به سیستم‌های پردازش رایانه‌ای صحیح نمی‌باشد. این سیستم‌ها به خصوص سیستم‌های برچسب‌گذاری به دلیل عدم اشراف به قواعد زبانی در برخورد با کلمات دارای پیچیدگی‌های ساختواژی توان محدودی دارند.

زبان فارسی نیز از این منظر قابل تامل و بررسی است. زیرا در زبان فارسی تکواژهای تصریفی و اشتقاقی زیادی وجود دارد که با اتصال به کلمات باعث تغییر برچسب کلمات در پیکره‌های زبانی و همچنین کاهش فراوانی کلمات می‌گردد و این موارد باعث کاهش کارایی سیستم‌های برچسب‌گذاری آماری در زبان فارسی می‌شود. برای حل این مشکل می‌توان سیستم برچسب‌گذاری را به یک تحلیل‌گر ساختواژی مجهز کرد. چون اشتقاق[4] کلمه‌ای کاملا جدید تولید می‌کند (زیرا بن‌واژه کلمه تغییر می‌یابد) تحلیل‌گر ساختواژی را نیز می‌توان محدود به تصریف[5] کرد و اشتقاق را در نظر نگرفت.

---

[1] Inflectional and Derivational Morphology
[2] Inflectional Morpheme
[3] Derivational Morpheme
[4] Derivation
[5] Inflection

لذا این تحلیل‌گر ساختواژی در واقع یک تحلیل‌گر تصریفی خواهد بود.

بنابراین در این فصل به مبحث تکواژها در زبان فارسی و انواع آن‌ها، انواع ساختواژه و ویژگی‌های ساختواژی مقولات مهم در زبان فارسی، تجزیه تصریفی کلمات پیکره متنی زبان فارسی و استفاده از تجزیه تصریفی کلمات برای برچسب‌گذاری آماری کلمات می‌پردازیم.

# تکواژ

در هر زبان، واژه‌هایی بسیاری وجود دارد که از یک یا چند واحد کوچکتر معنادار ساخته می‌شوند. بر اساس همین واحدهای معنی‌دار کوچکتر است که می‌توان ویژگی‌های ساختواژی هر واژه را مشخص کرد. به کوچکترین واحد معنی‌دار که در ساخت واژه مشخص می‌گردد، تکواژ گفته می‌شود (مشکوةالدینی، ۱۳۸۴). بر پایه ویژگی‌های معنایی و کاربردی، تکواژها به پنج دسته تقسیم می‌شوند که هر کدام را به طور خلاصه شرح می‌دهیم.

## ۵-۲-۱. تکواژ مادی، قاموسی یا پایه‌واژه

به تکواژهایی که معنی مادی یا قاموسی دارند و ممکن است به تنهایی به صورت واژه ساده به کار روند و یا تکواژ اصلی فعل یا غیر فعل را تشکیل دهند، تکواژ مادی یا قاموسی یا پایه‌واژه گفته می‌شود. واژه‌های مقولات اصلی: اسم، فعل، صفت، قید و حرف اضافه تکواژهای مادی یا قاموسی را تشکیل می‌دهند.

## ۵-۲-۲. تکواژ یا واژه نقشی یا دستوری

تکواژها یا واژه‌های نقشی در واقع کلمات موجود در مقولات بسته[۶] می‌باشند و تعداد آن‌ها محدود است. این واژه‌ها محتوای معنایی مادی ندارند، بلکه مفهوم نقش یا رابطه معنایی خاصی را نشان می‌دهند. ضمایر، حروف ربط، حرف نشانه "را" و برخی دیگر واژه دستوری یا نقشی‌اند.

## ۵-۲-۳. وندها یا تکواژهای اشتقاقی

به تکواژهایی که تنها به طور محدود به عنوان پیشوند یا پسوند به همراه پایه‌واژه‌های خاصی به کار می‌روند وند یا تکواژ اشتقاقی گفته می‌شود. تکواژهای اشتقاقی دو ویژگی عمده دارند:

۱- مقوله دستوری واژه حاصل از اشتقاق ساختواژی را با مقوله دستوری پایه‌واژه متفاوت می‌سازند. مانند کلمه "دانش" (دان + ش) که اسم است ولی پایه‌واژه آن پایه فعل است.

۲- با همه پایه‌واژه‌های یک مقوله به کار نمی‌روند. به عنوان مثال کلمه "دانش" کاربرد دارد ولی کلمه "آورش" کاربرد ندارد.

## ۵-۲-۴. تکواژ صرفی یا وند صرفی

به تکواژهایی که به مفهوم دستوری خاص (نکره، معرفه، برتر، برترین، زمان، مطابقه، استمراری و جز این‌ها) اشاره می‌کنند و به عنوان پسوند و یا پیشوند صرفی به همراه واژه‌های اسم، صفت و قید و یا در ساختواژه فعل به کار می‌روند، تکواژ یا وند صرفی گفته می‌شود. تکواژهای صرفی روابط دستوری خاصی را در جلمه نشان می‌دهند. تکواژ صرفی دو ویژگی عمده دارند:

---

[6] Closed category/wordclass

۱- به همراه همه واژه‌های یک دسته یا مقوله به طور یکسان و تقریبا بدون استثنا به کار می‌روند.
۲- مقوله واژه همراه خود را تغییر نمی‌دهند.

### ۵-۲-۵. واژه‌بست

به دسته خاصی از تکواژهای دستوری که ویژگی‌هایی به شرح زیر دارند، واژه‌بست[۷] گفته می‌شود:
۱- به دنبال واژه می‌چسبند، اما بخشی از ساختواژه اشتقاقی و یا صرفی محسوب نمی‌شود.
۲- از لحاظ معنی، به رابطه ساختی یا دستوری خاص از جمله، کسره اضافه (اضافه اسمی و اضافه صفتی)، رابط، ضمیر متصل با رابطه دستوری اضافی اسمی، متمم حرف اضافه یا مفعول صریح و ... اشاره می‌کنند.
۳- از لحاظ آوایی تکیه اصلی واژه بر آن‌ها ظاهر نمی‌شود.

## ساختواژه کلمات فارسی

### ۵-۳-۱. ساختواژه

مطالعه ساختار کلمات و چگونگی ساخت کلمات از ترکیب واحدهای کوچکتر زبانی یعنی تکواژها، ساختواژه نامیده می‌شود. تکواژها بسته به چگونگی رخدادشان در زبان به دو دسته تکواژهای آزاد[۸] و تکواژهای مقید[۹] تقسیم‌بندی می‌شوند. تکواژهای آزاد به تنهایی می‌توانند به عنوان یک کلمه ظاهر

[7] Clitic
[8] Free morpheme
[9] Bound morpheme

شوند ولی تکواژهای مقید به تنهایی مورد استفاده قرار نمی‌گیرند و باید با تکواژهای آزاد ترکیب شوند (Sproat, 1992). طبق این تعریف تکواژ مادی، قاموسی و یا پایه‌واژه و تکواژ یا واژه نقشی یا دستوری تکواژ آزاد می‌باشند[۱۰] زیرا به تنهایی می‌توانند مورد استفاده قرار گیرند و همچنین وند یا تکواژ اشتقاقی و وند یا تکواژ صرفی و واژه‌بست تکواژ مقید می‌باشند. روشی که تکواژها با یکدیگر ترکیب می‌شوند و اطلاعاتی که در هنگام ترکیب تکواژها، هر تکواژ منتقل می‌کند از زبانی به زبان دیگر متفاوت است. با این حال سه نوع ساختواژه را مستقل از هر زبان خاصی نام می‌برند:

۱- اشتقاق: در ساختواژه اشتقاقی یک کلمه جدید از تکواژهای اشتقاقی و یک واژه به وجود می‌آید و معمولا کلمه جدید مقوله واژگانی متفاوتی از کلمه اولیه دارد (همان ویژگی ذکر شده در بخش ۵-۲-۳).

۲- تصریف: در ساختواژه تصریفی تکواژهایی که به کلمه اضافه می‌شوند معمولا معرف اطلاعات دستوری خاصی می‌باشند. این فرآیند معمولا مقوله واژگانی کلمه را تغییر نمی‌دهد (همان ویژگی ذکر شده در بخش ۵-۲-۴). به نظر می‌رسد در زبان فارسی علاوه بر وندها و تکواژهای صرفی افزودن اکثر واژه‌بست‌ها به کلمات را نیز بتوان جز تصریف و ساختواژه تصریفی در نظر گرفت.

۳- ترکیب[۱۱]: در ترکیب دو یا چند تکواژ مادی، قاموسی یا پایه‌واژه به هم متصل می‌شوند و یک واژه جدید را تشکیل می‌دهند.

نکته قابل توجه این است که از سه نوع ساختواژه، معمولا تصریف است که بر روی مجموعه برچسب پیکره و به تبع آن بر روی برچسب کلمات تاثیر می‌گذارد زیرا تکواژهای افزوده شده به کلمه در تصریف جز کلمه لحاظ نمی‌شوند[۱۲].

در ادامه بحث به ویژگی‌های ساختواژی چند مقوله اصلی مانند اسم، صفت، قید و فعل در زبان فارسی می‌پردازیم و برای این کار از (مشکوةالدینی، ۱۳۸۴) کمک می‌گیریم. چون بحث ما معطوف به

---

[۱۰] اگرچه ذکر این نکته حائز اهمیت است که در (مشکوه‌الدینی، ۱۳۸۴) تنها از تکواژ مادی، قاموسی و یا پایه‌واژه تحت عنوان تکواژ آزاد یاد شده است.

[11] Compounding

[۱۲] حداقل این مطلب در مورد پیکره متنی زبان فارسی که مقصود ماست در اکثر موارد صادق می‌باشد، هرچند موارد نادری مثل کلمه "یکبار" با برچسب "عدد,اصلی,اسم,عام,مفرد" وجود دارد که نمی‌توان آن را جز تصریف لحاظ کرد.

برچسب‌گذاری است سعی می‌کنیم خود را محدود به بیان ویژگی‌های تصریفی کنیم که بر برچسب واژگانی کلمه تاثیر می‌گذارند (زیرا همان‌طور که ذکر شد برچسب‌های کلمات از تصریف تاثیر می‌پذیرند) و از ذکر ویژگی‌های اشتقاق و ترکیب خودداری می‌کنیم.

## ۵-۳-۲. اسم

اسم به طور کلی به پدیده مادی یا غیر مادی مانند گیاه، جانور، انسان، روح و جز اینها اشاره می‌کند. اسامی با دو تکواژ تصریفی یعنی نشانه جمع و نشانه نکره ظاهر می‌شوند. نشانه معرفه و مفرد تهی می‌باشد. علاوه بر این دو تکواژ تصریفی واژه‌بست‌ها نیز می‌توانند بر روی کلمات ظاهر شوند. کسره اضافه (اضافه اسمی و یا اضافه صفتی) و ضمایر متصل از این جمله‌اند. حاصل افزودن رابط‌هایی مانند "م"، "ی" و ... به اسامی افعال اسنادی را سبب می‌شود مانند "خانه‌ام" (= در خانه هستم) و "مردی" (= تو مردی).

## ۵-۳-۳. صفت و قید

صفت و قید مشابهت زیادی به یکدیگر دارند. به گروهی از کلمات که به لحاظ معنی به ویژگی یا حالتی اشاره می‌کنند و به لحاظ پیوند ساختی به عنوان وابسته اسم و یا فعل به کار می‌روند به ترتیب، صفت و قید گفته می‌شود.

صفت‌ها یا ثابتند یا درجه‌ای. صفت ثابت صفتی است که صورت‌های برتر و برترین ندارد مانند مونث، مذکر، زنده، مرده. صفت‌های درجه‌ای صفت‌هایی هستند که علاوه بر شکل ساده با تکواژهای "تر" و "ترین" به صورت صفت برتر و برترین استفاده می‌شوند. قیدها نیز مانند صفت‌ها دو نوع ثابت و درجه‌ای دارند. قید ثابت صورت برتر ندارد مانند ناگهان، هرگز. قید درجه‌ای دو صورت ساده و قید برتر دارد. قید برتر با افزودن تکواژ "تر" به قید حاصل می‌شود.

علاوه بر تکواژهای فوق صفت و قید مانند اسم، بسیاری از واژه‌بست‌ها مانند ضمایر متصل، رابط‌ها و

کسره اضافه را نیز می‌پذیرند.

## ۵-۳-۴. فعل

از لحاظ ساختواژه، واژه فعل پایه فعل و وندهای صرفی خاصی را شامل می‌شود. به علاوه پایه فعل به یکی از مفاهیم «کنش»، «حالت»، «دگرگونی» و یا ترکیبی از آن‌ها اشاره می‌کند. تعداد زمان‌های[۱۳] فعلی در زبان فارسی نسبتا زیاد است اگرچه کمتر از زبان‌هایی مانند انگلیسی است. این زمان‌ها عبارتند از: حال ساده اخباری، حال استمراری، حال التزامی، امر، آینده، گذشته استمراری، گذشته نقلی، گذشته نقلی استمراری، گذشته بعید و گذشته التزامی. برخی از زمان‌ها تکواژهایی مخصوص به خود دارند مانند استمراری یا نقلی و ...، و هر فعل در زبان فارسی برای شش شخص صرف می‌گردد و تکواژهای صرفی برای هر شخص در هر زمان مشخص است. به دلیل آشنایی هر فارسی زبان با این موارد از ذکر جزئیات مربوط به این تکواژها خودداری می‌کنیم.

تقسیم‌بندی دیگری که در مورد افعال می‌توان قائل شد افعال اسنادی و افعال غیراسنادی است. افعال اسنادی که از پایه فعل (بن‌های ماضی و مضارع) و شناسه‌ها به وجود می‌آید افعال غیراسنادی هستند. فعل "است" به همراه افعالی که از یک اسم، صفت، قید، حرف اضافه و یا ضمیر به علاوه رابط‌ها (همان شناسه‌های فعلی) ایجاد می‌شوند فعل اسنادی نامند. نمونه‌هایی از این دسته را می‌توان نام برد: حاضرم، خوبی (= تو خوب هستی)، چطورست، بی‌بهره‌ایم، شمایید.

## ۵-۳-۵. دیگر مقولات

در این‌جا به طور وسیع به مقولات دیگر نمی‌پردازیم زیرا مقولات دیگر جزئیات ساختواژی کمتری

[13] Tense

دارند و بیشتر نیز در متون محاوره‌ای مانند نمایش‌نامه‌ها و کتب داستانی با زبان عامیانه مشاهده می‌شوند. به عنوان مثال ضمایر و حروف اضافه امکان دارد که با برخی واژه‌بست‌ها به کار روند. یا در گفتار عامیانه بسیار اتفاق می‌افتد که ضمایر با تکواژ جمع ساز "ها" جمع بسته می‌شوند مانند: ماها، شماها. گاهی مشاهده می‌شود که حرف ربط "هم" به صورت "م" و نشانه "را" به صورت "و" در انتهای کلمات ظاهر می‌شود مانند: منم آمدم، حسینو دیدم. از این موارد به وفور می‌توان بیان نمود اگرچه معمولا در متون رسمی کاربرد کمی دارند یا فراوانی تعداد نمونه‌ها آن‌ها بسیار کم است. برای مطالعه راجع با این موارد خواننده را به کتب مرتبط با دستور زبان فارسی و یا بررسی کلمات و برچسب‌های موجود در پیکره‌های برچسب‌خورده مانند پیکره متنی زبان فارسی ارجاع می‌دهیم.

## تجزیه تصریفی کلمات فارسی

از بحثی که تاکنون راجع به تکواژها و ساختواژه کلمات فارسی مطرح شد می‌توان نتیجه گرفت که هر کلمه می‌تواند به واحدهای معنی‌دار تشکیل دهنده آن یعنی تکواژها تجزیه شود. بنابر سه نوع ساختواژه (اشتقاق، تصریف و ترکیب) مطرح شده سه نوع تجزیه را نیز می‌توان برای کلمات تصور کرد. همان‌طور که ذکر شد در بین سه نوع ساختواژه، تکواژهای تصریفی افزوده شده به کلمه در فرآیند تصریف، برچسب کلمه را تغییر می‌دهند و همین طور بر تعداد برچسب‌های پیکره تاثیر می‌گذارند. این مطلب با بررسی اجمالی پیکره متنی زبان فارسی و برچسب‌های کلمات آن کاملا مشهود است.

حال اگر کلمات بر اساس تصریف آن‌ها تجزیه شوند یعنی هر کلمه به بن‌واژه و تکواژهای صرفی افزوده شده به آن، تجزیه شود به تبع این تجزیه برچسب کلمه نیز ناگزیر تجزیه می‌شود. به عنوان مثال کلمه "کتاب‌ها" دارای برچسب "اسم،عام،جمع" (N,COM,PL) است. تجزیه این کلمه به صورت "کتاب" + "ها" می‌باشد. برچسب این کلمه نیز بر اساس این تجزیه به دو قسمت تجزیه می‌شود: N,COM برای جز "کتاب" و PL برای جز "ها". بنابراین مشخص شد که تصریف و تجزیه تصریفی هر دو، هم بر کلمه و هم بر برچسب کلمه تاثیر می‌گذارند. هدف ما از طرح این مقدمه استفاده از تجزیه تصریفی کلمات در برچسب‌گذاری می‌باشد که در بخش بعد به آن می‌پردازیم.

# استفاده از تحلیل‌گر ساختواژی با امکان تجزیه تصریفی برای برچسب‌گذاری

## ۵-۵-۱. تصریف و تفسیرهای متفاوت کلمات با بن‌واژه یکسان

تکواژهای تصریفی و واژه‌بست‌ها به هر روشی که بر روی کلمات ظاهر شوند (چه بدون فاصله، چه با فاصله و چه با نیم‌فاصله) باعث می‌شوند که کلمه حاصل در هر کدام از اشکالش متفاوت از بن‌واژه آن تفسیر شود[۱۴]. به عبارت دیگر افزودن تکواژهای تصریفی و واژه‌بست‌ها به یک کلمه باعث می‌شود کلمه‌ای جدید ایجاد شود. به عنوان مثال اگر کلمه "کتاب" با تکواژ "ها" جمع بسته شود چه کلمه به شکل "کتاب ها" ظاهر شود چه به شکل "کتاب‌ها" و چه به شکل "کتابها"، در هر صورت به عنوان یک کلمه جدید تفسیر می‌شود.

از طرف دیگر همان‌طور که در بیان ویژگی‌های ساختواژی مقولات مختلف ذکر شد، تکواژهای تصریفی و واژه‌بست‌هایی وجود دارند که می‌توانند بر روی کلمات موجود در مقولات متفاوت ظاهر شوند. مثلا ضمایر متصل می‌توانند به اسم، صفت، قید و حرف اضافه افزوده شوند. علاوه بر این‌که کلمه جدید متفاوت از بن‌واژه تفسیر می‌شود این باعث می‌شود که این تکواژهای با کابرد یکسان در مقولات متفاوت، به صورت متفاوت تفسیر شوند. به عنوان نمونه ضمیر متصل "م" در کلمه "پسرم" و در عبارت "پسر خوبم" دو تفسیر متفاوت می‌شود زیرا "م" در کلمه اول (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 پیکره متنی زبان فارسی کاملا مشهود است. اولین مشکل این است که تعداد برچسب‌های متمایز پیکره بسیار زیاد خواهد شد و از طرفی فراوانی آن‌ها کاهش می‌یابد. دومین مشکل این است که کلمات با بن‌واژه یکسان به اشکال متفاوتی با برچسب متفاوت

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

ظاهر می‌شوند که این مورد نیز بر فراوانی کلمات با بن‌واژه یکسان تاثیر می‌گذارد. این دو مشکل تاثیر بسیار زیادی بر روی برچسب‌گذارهای آماری دارد.

## ۵-۵-۲. حل مشکل تفسیرهای متفاوت کلمات با بن‌واژه یکسان

برای حل این دو مشکل ما روشی به کار می‌بریم که باعث کاهش تعداد برچسب‌ها و حذف تفسیرهای متفاوت کلمات با بن‌واژه یکسان در پیکره متنی زبان فارسی شود 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 نقطه نظر تصریفی تجزیه کنیم. 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 نمی‌باشند از مجموعه برچسب پیکره حذف می‌کنیم. از جمله این برچسب‌ها می‌توان به DAY (روز)، LOC (مکان)، DIR (جهت)، SES (فصل) و MON (ماه)، SURN (لقب) و TIME (زمان) در مقوله اسم و LOC (مکان)، EXM (مثال)، ORD (ترتیبی)، REPT (تکراری) و NEGG (منفی) در مقوله قید اشاره کرد. تعداد این برچسب‌ها کم و بعضی آن‌ها را می‌توان بدون ایجاد مشکل بعد از برچسب‌گذاری به برچسب کلمه اضافه کرد.

۲- برای هر برچسب با این شرط که نشانگر تکواژهای تصریفی و واژبست‌ها باشد همه تکواژهایی که منتسب به این برچسب می‌شوند باید 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 نشانه خاصی در کلمه ندارند در این مرحله در نظر گرفته نمی‌شوند. جدول ۵-۱ این موارد را نشان می‌دهد. همان طور که مشاهده می‌شود اگر برای اضافه شدن یک تکواژ به کلمه نیاز به یک واسط بود این واسط چون جز کلمه نیست ما آن را جز تکواژ در نظر می‌گیریم. به عنوان مثال در کلمه "کتاب‌هایم" هنگام اضافه شدن "م" به کلمه "کتاب‌ها" نیاز به واسط "ی" است یا در کلمه "خانه‌ام" نیاز به واسط "ا" است که این موارد را جز تکواژ در نظر می‌گیریم.

چند نمونه‌ای از 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 |
| نشانه ی (یاء نکره و مصولی | 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، یید، ئید | ین، این |
| | 6 | ند، اند، یند | ن، ان |
| نشانه منفی (فعل) | NEG | ن، م | _____ |
| نشانه التزامی (فعل) | SUB | ب | _____ |
| نشانه امری (فعل) | IMP | ب | _____ |

۳- حال که تکواژهای تصریفی و واژه‌بست‌هایی که بر روی کلمات به دست آمده است مشخص شد، اکنون می‌توان بر اساس بحث مطرح شده در بخش ۵-۴ هر کلمه را تجزیه تصریفی کرد. در عمل برچسب‌گذاری هر جز تشکیل‌دهنده کلمه به عنوان یک کلمه مجزا در نظر گرفته خواهد شد. با این کار تعداد برچسب‌های متمایز کاهش می‌یابد و تفسیرهای متفاوت کلمات با بن‌واژه یکسان از بین خواهد رفت و در نتیجه فراوانی برچسب‌ها و کلمات بسیار افزایش می‌یابد که این تاثیر به‌سزایی در افزایش دقت برچسب‌گذارهای آماری خواهد داشت.

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

در بخش نتایج تجربی با اعمال این روش کارایی آن را خواهیم دید.

# نتیجه‌گیری

در این بخش به بررسی تکواژها و ساختواژه کلمات فارسی پرداختیم. انواع تکواژهای فارسی که در پنج دسته تقسیم‌بندی می‌شوند، بیان و برخی از ویژگی‌های آن‌ها مطرح شد. انواع فرآیندهای ساختواژی و شرح هر کدام ذکر شد. پس از آن به بررسی ویژگی‌های ساختواژی مقولات اصلی در زبان فارسی پرداخته شد و خواص ساختواژه تصریفی کلمات فارسی بیان گردید. سپس به تجزیه تصریفی کلمات فارسی پرداختیم که یکی از انواع تجزیه ساختواژی است. افزودن تکواژهای تصریفی و واژه‌بست‌ها به کلمات منجر به مشکلاتی در قبال سیستم‌های برچسب‌گذاری آماری می‌شد که این موارد بیان شد و برای حل آن روشی بر اساس تجزیه تصریفی کلمات ارائه گردید. این روش باعث جلوگیری از تفسیرهای متفاوت کلمات با بن‌واژه یکسان و افزایش فراوانی برچسب‌ها و کلمات و نهایتا افزایش کارایی سیستم‌های برچسب‌گذاری آماری خواهد شد.

# کلمات ناشناخته

## مقدمه

در فصول پیش صحبت از برچسب‌گذاری کلمات بر اساس روش‌های مختلف و با کمک پیکره‌های آموزشی به میان آمد. در این روش‌ها برچسب کلمات با توجه به اطلاعات قبلی که از پیکره آموزشی استخراج می‌شود تعیین می‌گردد. حال اگر کلمه‌ای قبلا در پیکره آموزشی ظاهر نشده باشد نمی‌توان از پیکره آموزشی اطلاعات دقیقی راجع به آن کلمه به دست آورد و حتی از توزیع کلمات در پیکره نیز نمی‌توان استفاده کرد زیرا توزیع کلمات ناشناخته کاملا متفاوت می‌باشد. یک شاهد برای این مطلب این است که کلمات ناشناخته به ندرت از مقولات بسته انتخاب می‌شوند زیرا این مقولات تعداد کلمات محدودی دارند و آن کلمات معمولا در پیکره ظاهر می‌شوند. پس مشخص می‌شود که غلبه بر کلمات ناشناخته مشکل می‌باشد و به سادگی انجام نمی‌شود.

دو مسئله اساسی برای غلبه بر کلمات ناشناخته وجود دارد: یکی فقدان اطلاعات واژگانی راجع به کلمه ناشناخته و دیگری توزیع متفاوت کلمات ناشناخته با کلمات دیده شده در پیکره. به همین خاطر از پیش اطلاعات خاصی راجع به کلمات ناشناخته وجود ندارد و نمی‌توان از اطلاعات کلمات حاضر در پیکره برای مدل کردن دقیف رفتار توزیعی این کلمات استفاده کرد.

در این فصل ابتدا مختصری راجع به رفتار توزیعی کلمات ناشناخته و نحوه به دست آوردن توزیع این کلمات هرچند به صورت تقریبی صحبت می‌کنیم و سپس چند روش غلبه بر کلمات ناشناخته را مورد بررسی قرار می‌دهیم.

## رفتار توزیعی کلمات ناشناخته

رفتار و طبیعت کلمات ناشناخته کاملا با رفتار کلمات در زبان (= پیکره) متفاوت است و با مشاهده

همه کلمات در پیکره آموزشی نمی‌توان قضاوت درستی راجع به آن‌ها داشت. اگر پیکره به اندازه کافی بزرگ باشد احتمال این‌که کلمات ناشناخته از مقولات بسته باشند بسیار کم است. بنابراین انتخاب برچسب برای کلمات شناخته شده از بین تعداد بیشتری برچسب انجام می‌شود در حالی که این تعداد برای کلمات ناشناخته کمتر است؛ ولی با این حال دقت برچسب‌گذاری کلمات ناشناخته از کلمات شناخته شده به مراتب کمتر است.

برای بررسی و مطالعه کلمات ناشناخته باید توزیع آماری این کلمات را به همراه یک مجموعه نمونه از این کلمات از یک پیکره استخراج شود. اولین روشی که به ذهن می‌رسد متقابل قرار دادن یک مجموعه آزمون[1] و یک مجموعه آموزش می‌باشد، به طوری که کلماتی از مجموعه آزمون که در مجموعه آموزش حضور ندارند به عنوان کلمات ناشناخته در نظر گرفته شده و توزیع آماری بر اساس برچسب این کلمات به دست می‌آید. نقص این روش این است که باید حجم پیکره بسیار بزرگ باشد تا بتوان تخمین مناسبی از توزیع کلمات ناشناخته به دست آورد.

روش دیگر که ساده ولی بسیار کاراست در (Dermates and Kokkinakis, 1995) ارائه شد. ایده روش به این صورت است که احتمال این‌که یک کلمه ناشناخته دارای یک برچسب واژگانی خاص باشد می‌تواند از توزیع احتمالی کلماتی که تنها یک مرتبه در پیکره رخ داده‌اند، تخمین زده شود. این کلمات را اصطلاحا hapax words می‌گویند. این روش بیان می‌کند که توزیع کلمات ناشناخته باید بسیار مشابه کلماتی باشد که تنها یک مرتبه در پیکره ظاهر شده‌اند.

## غلبه بر کلمات ناشناخته

برای غلبه بر کلمات ناشناخته تاکنون روش‌هایی مختلفی استفاده شده است. ساده‌ترین روش برای انتساب برچسب به کلمات ناشناخته، انتساب همه برچسب‌های ممکن به آن کلمه و یا انتساب محتمل‌ترین برچسب در پیکره به آن کلمه می‌باشد. به عنوان مثال در زبان انگلیسی به کلمات

[1] Test set

ناشناخته‌ای که با حروف بزرگ آغاز می‌شوند اسم مفرد خاص و به دیگر کلمات ناشناخته اسم مفرد عام منتسب می‌شود که البته زبان فارسی به دلیل عدم تمایز ظاهری بین اسامی خاص و عام باید به همه کلمات ناشناخته برچسب اسم مفرد عام منتسب کرد زیرا این برچسب از دیگر برچسب‌ها محتمل‌تر است. روش دیگر این است که به جای انتساب همه برچسب‌ها به کلمه، فقط برچسب‌های مقولات باز[۲] برای کلمه ناشناخته منظور شود که البته تعداد این برچسب‌ها نیز زیاد می‌باشد.

اگر به کلمات ناشناخته تعداد زیادی برچسب منتسب شود ابهام‌زدایی از برچسب‌ها و انتخاب برچسب صحیح مشکل می‌شود. اگر برچسب کلمات ناشناخته یک برچسب خاص مانند اسم مفرد عام در نظر گرفته شود نیازی به ابهام‌زدایی برای تشخیص برچسب کلمه نیست ولی دقت این روش پایین است. در هر صورت روشن است که این روش‌ها مناسب کار ما نیستند چون اولا تاکید ما بر برچسب‌گذاری بدون ابهام است یعنی هر کلمه در نهایت باید یک برچسب داشته باشد که عملا دو روش از روش‌های گفته حذف می‌گردند و ثانیا ناکارامدی روش دیگر یعنی در نظر گرفتن همه کلمات ناشناخته به عنوان اسم مفرد عام کاملا مشخص است.

ما دو روش را برای غلبه بر کلمات ناشناخته استفاده می‌کنیم که در زیر به آن‌ها می‌پردازیم.

## ۶-۳-۱. توزیع احتمالی کلمات ناشناخته

همان‌گونه که ذکر شد، توزیع احتمالی کلمات ناشناخته متفاوت از توزیع کل کلمات در زبان است و می‌توان توزیع کلمات ناشناخته را از روی یک مجموعه نمونه به دست آورد. این توزیع از دو جهت مفید است: یکی این‌که می‌تواند به عنوان یک منبع نظری توسط کاربر برای بررسی و توجیه رفتار کلمات ناشناخته در قبال کلمات شناخته شده مورد استفاده قرار گیرد؛ دیگر این‌که مقادیر احتمالی توزیع می‌تواند در برچسب‌گذارهای آماری به طور مستقیم استفاده شود.

[2] Open category/wordclass

## ۶-۳-۲. توجه به وندها

روش‌های پیشرفته‌تر برای تشخیص برچسب کلمات ناشناخته در نظر گرفتن ویژگی‌هایی از قبیل پیشوند و پسوند کلمات است[۳]. با استفاده از این روش‌ها دقت‌های بالاتری برای کلمات ناشناخته در زبان انگلیسی گزارش شده است (Brill, 1994) (Weischedel, 1993). به عنوان نمونه برچسب‌گذار Xerox (Cutting, 1992) از قوانینی بر اساس بخش‌های پایانی کلمات یعنی حروف انتهای کلمات برای انتساب یک مجموعه برچسب به کلمات ناشناخته استفاده می‌کند.

یک روش خودکار برای کشف قوانینی که می‌توانند برای پیش‌بینی برچسب کلمات ناشناخته استفاده شوند توسط (Mikheev, 1997) ارائه شده است. توجه این روش هم به پیشوندهاست و هم به پسوندها. در ادامه به جزئیات این روش با تکیه بر زبان فارسی می‌پردازیم.

### ۶-۳-۲-۱. الگوی قوانین

در این جا دو نوع قانون در نظر گرفته می‌شود: قوانین ساختواژی و قوانین غیر ساختواژی. یک قانون ساختواژی با فرض این‌که با حذف یک وند از کلمه ناشناخته کلمه‌ای حاصل می‌شود که شناخته شده است، راجع به برچسب کلمه با توجه به برچسب کلمه حاصل و وند حذف شده قضاوت می‌کند. در حالی که قوانین غیر ساختواژی تنها بر اساس پیشوندها و پسوندهای کلمه ناشناخته در مورد برچسب کلمه قضاوت را انجام می‌دهند. قوانین غیر ساختواژی برای یافتن برچسب کلمات ناشناخته‌ای که حاصل تصریف و یا اشتقاق نیستند مفید می‌باشد.

برای تعریف این قوانین الگویی در نظر گرفته می‌شود که به صورت زیر می‌باشد:

[۳] در این‌جا منظور از پیشوند و پسوند و به طور کلی وند، لزوما آن اصطلاحات زبان‌شناسی معمول نیست و می‌تواند یک یا چند حرف اول یا آخر کلمات نیز باشد.

$$G=_{x\in\{b,e\}}\left[-S+M\,?\,I-Class\rightarrow R-Class\right]$$

در این الگو:

۳- x: تعیین می‌کند که قانون به ابتدای کلمه یا انتهای کلمه اعمال می‌شود. b به معنای ابتدا و e به معنای انتها می‌باشد.

۴- S: وندی است که باید از کلمه جدا شود. این وند از ابتدا یا انتهای کلمه ناشناخته (بسته به مقدار x) حذف می‌شود.

۵- M: بخشی است که دچار تغییر شده است و باید به رشته حاصل بعد از جداسازی وند اضافه شود.

۶- I-Class: برچسب کلمه جدید است که با حذف S و افزودن M به دست آمده است. کلمه جدید باید در واژگان جستجو شود تا مشخص شود که دارای این برچسب خاص می‌باشد یا خیر. اگر در یک قانون I-Class تهی بود این جستجو نیاز نیست.

۷- R-Class: برچسب جدیدی است که اگر عملیات فوق موفق بوده باشد به کلمه ناشناخته منتسب می‌شود.

ذکر این نکته ضروری است که در واقع I-Class و R-Class مجموعه‌هایی هستند که دارای یک یا چند برچسب می‌باشند. برای شرح موارد فوق یک مثال می‌آوریم. قانون زیر را در نظر بگیرید:

$$G_0=_e\left[-\text{گان}\quad+\text{ه}\;?(N,COM,SING)\rightarrow(N,COM,PL)\right]$$

این قانون بیان می‌کند که اگر یک کلمه ناشناخته داریم که با پسوند "گان" پایان می‌یابد، ما باید این قسمت را از پایان کلمه حذف و "ه" را به آن اضافه کنیم. حال اگر کلمه جدید در واژگان دارای برچسب N,COM,SING (اسم مفرد عام) بود ما به کلمه ناشناخته برچسب N,COM,PL (اسم جمع عام) را نسبت می‌دهیم. مثلا اگر کلمه "پرندگان" در واژگان ناشناخته باشد این قانون ابتدا این کلمه را به دو بخش تقسیم می‌کند (پرند + گان = پرندگان). سپس "ه" را با آخر کلمه اضافه می‌کند (پرند + ه = پرنده)، اگر کلمه حاصل (پرنده) در واژگان برچسب N,COM,SING را داشت برچسب کلمه "پرندگان"، N,COM,SING در نظر گرفته می‌شود. (البته ظاهرا قوانین با این شکل در فارسی بسیار

نادر است که در حین افزوده شدن وندی به کلمه بخشی از کلمه حذف و یا دچار تغییر شود، بنابراین مقدار M در قوانین معمولا تهی خواهد بود.)

## ۶-۳-۲-۲. استخراج قوانین

برای استخراج قوانین ساختواژی عملگر $\nabla_n$ را تعریف می‌کنیم که n طول بخشی که دچار تغییر می‌شود را مشخص می‌کند. اگر n صفر در نظر گرفته شود نتیجه عملگر $\nabla_0$ یک قانون ساختواژی است که هیچ بخشی از آن دچار تغییر نمی‌شود یعنی $M = \phi$. عملگر $\nabla_1$ قوانینی را استخراج می‌کند که طول بخش تغییرکننده ۱ کاراکتر می‌باشد. هنگامی عملگر به یک جفت از مدخل‌های واژگان( $[Word\ Tag(s)]_i$ و $[Word\ Tag(s)]_j$) اعمال می‌شود، ابتدا n کاراکتر بسته به مقدار x از ابتدا یا انتهای کلمه با طول کوتاه‌تر جدا می‌شود و در M قرار می‌گیرد. سپس این کلمه به دست آمده از کلمه با طول بزرگتر تفریق می‌شود (در صورت امکان). اگر حاصل تفریق یک رشته غیرتهی بود، سیستم یک قانون ساختواژی می‌سازد به طوری که I-Class برچسب کلمه کوتاه‌تر و R-Class برچسب کلمه طولانی‌تر خواهد بود و وند به دست آمده از عمل تفریق، در S قرار می‌گیرد. برای مثال:

$$[\text{پرندگان} \quad (N,COM,PL)]\nabla_1[\text{پرنده} \quad (N,COM,SING)] \rightarrow$$
$${}_e[-\text{گان} \quad +\text{ه}\ ?(N,COM,SING) \rightarrow (N,COM,PL)]$$

عملگر $\nabla_n$ به همه جفت مدخل‌های ممکن در واژگان اعمال می‌شود. اگر قانونی تولید شد که قبلا استخراج شده بود فراوانی آن افزایش می‌یابد و اگر یک قانون جدید بود به قوانین استخراجی افزوده می‌شود و مقدار فراوانی آن ۱ لحاظ می‌گردد. بنابراین قوانین ساختواژی پیشوندی و پسوندی به همراه فراوانی آن‌ها استخراج می‌شود. حال می‌توان با در نظر گرفتن یک حد آستانه[4] مثلا بین ۲ تا ۴ قوانین با فراوانی کم را حذف کرد.

برای استخراج قوانین غیرساختواژی حداکثر طول بخش جداشونده از ابتدا یا انتهای کلمه را مثلا ۵

[4] Threshold

در نظر می‌گیریم. بنابراین همه بخش‌های ابتدایی و انتهایی با طول ۱، ۲، ۳، ۴ و ۵ از کلمات در واژگان به همراه برچسب آن‌ها استخراج می‌شود. برای استخراج قوانین غیر ساختواژی عملگر $\Delta$ را تعریف می‌کنیم. این عملگر از یک کلمه در واژگان ($[Word\ Tag(s)]_i$) مجموعه‌ای از قوانین را استخراج می‌کند. برای مثال عملگر $\Delta$ از یک کلمه مانند $[(ADJ)$ *تاثیرپنیر*$]$ قوانین زیر که دلالت بر پایان کلمه می‌کنند را استخراج می‌کند:

$$\Delta[\text{تاثیرپنیر}\quad (ADJ)] = \begin{cases} {}_e[-\text{ر}\ ?- \rightarrow (ADJ)] \\ {}_e[-\text{یر}\ \ ?- \rightarrow (ADJ)] \\ {}_e[-\text{نیر}\quad ?- \rightarrow (ADJ)] \\ {}_e[-\text{پنیر}\quad\ ?- \rightarrow (ADJ)] \\ {}_e[-\text{ریپنیر}\quad\ ?- \rightarrow (ADJ)] \end{cases}$$

عملگر $\Delta$ به همه مدخل‌های واژگان اعمال می‌شود. برای هر قانونی که به دست می‌آید اگر قبلا استخراج شده بود فراوانی آن افزایش می‌یابد وگرنه به عنوان یک قانون جدید در نظر گرفته می‌شود. پس از استخراج قوانین غیرساختواژی می‌توان با در نظر گرفتن یک حد آستانه مثلا بین ۲۰ تا ۵۰ قوانین با فراوانی کم را حذف کرد. چون ماهیت قوانین ساختواژی و غیرساختواژی متفاوت است لذا حد آستانه نیز برای آن‌ها متفاوت در نظر گرفته می‌شود.

## ۶-۳-۲-۳. امتیازدهی به قوانین

تمام قوانین به دست آمده در پیش‌بینی برچسب کلمات ناشناخته یکسان عمل نمی‌کنند و برخی از برخی دیگر دقیق‌ترند. برای ارزیابی کارایی قوانین استخراج شده به صورت زیر عمل می‌شود: قوانین یک به یک از مجموعه قوانین انتخاب می‌شوند و در صورت سازگار بودن (قابل اعمال بودن) به کلمات در واژگان اعمال و برچسب کلمه حدس زده می‌شود. بر اساس تعداد موارد صحیح حدس زده شده و تعداد موارد سازگار دقت پیش‌بینی هر قانون به دست می‌آید:

$$\hat{p}_i = \frac{x_i}{n_i}$$

که در آن x تعداد موارد صحیح و n تعداد موارد سازگار است. مقدار $\hat{p}$ برای بیان دقت قوانین

مناسب است اما استفاده از این معیار با مشکل خطای تخمین روبروست که به دلیل عدم وجود داده آموزشی کافی رخ می‌دهد. برای مثال اگر یک قانون تنها یک مرتبه قابل اعمال باشد و در این یک مورد نیز صحیح عمل کند مقدار $\hat{p}$ برابر ۱ می‌شود که این تخمین صحیح نمی‌باشد زیرا ممکن است با افزایش داده آموزشی، مقدار $\hat{p}$ تغییر زیادی پیدا کند. این مشکل را می‌توان با محاسبه حد پایین اطمینان[5] ($\pi_l$) و استفاده از آن به جای $\hat{p}$ برای ارزیابی قوانین حل کرد. با اطمینان[6] $\alpha$ می‌توان فرض کرد که اگر داده آموزشی بیشتری استفاده شود مقدار $\hat{p}$ از $\pi_l$ بدتر نخواهد شد. ابتدا برای حذف صفرها از مخرج به جای $\hat{p}$ از $\hat{p}^*$ استفاده می‌شود:

$$\hat{p}_i^* = \frac{x_i + 0.5}{n_i + 1}$$

حد پایین اطمینان ($\pi_l$) به صورت زیر محاسبه می‌شود (Hayslett,1981):

$$\pi_{l_i} = \hat{p}_i^* - t_{(1-\alpha)/2}^{(n-1)} * \sqrt{\frac{\hat{p}_i^* * \left(1 - \hat{p}_i^*\right)}{n}}$$

$t_{(1-\alpha)/2}^{d_f}$ یک اطمینان از توزیع t است. $\alpha$ سطح اطمینان و $d_f$ درجه آزادی است که یکی از تعداد نمونه‌ها یعنی n کمتر است ($d_f = n - 1$). ما مقدار $\alpha$ را ۹۰٪ در نظر می‌گیریم یعنی $t_{(1-0.90)/2}^{d_f} = t_{0.05}^{(n-1)}$. مقدار $t_{0.05}^{(n-1)}$ بر اساس تعداد نمونه‌ها در جدول ۶-۱ آمده است. این جدول در کتب آماری در مبحث مربوط به توزیع t برای مقادیر متفاوت $\alpha$ قابل دسترسی است.

با این نحوه ارزیابی جدید اگر یک قانون مقدار $\hat{p}$ بالایی روی یک تعداد نمونه کوچک داشته باشد ولی قانون دیگر مقدار $\hat{p}$ کمتری روی یک تعداد نمونه بزرگتر داشته باشد، مقدار $\pi_l$ قانون دوم ممکن است بیشتر شود.

[5] Lower Confidence Limit
[6] Confidence

جدول توزیع t برای اطمینان ۹۰٪

| ۱ | ۲ | ۳ | ۴ | ۵ | ... | ۴۰ | ۵۰ | ۶۰ | ۸۰ | ۱۰۰ | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ۶.۳۱۴ | ۲.۹۲ | ۲.۳۵۳ | ۲.۱۳۲ | ۲.۰۱۵ | ... | ۱.۶۸۴ | ۱.۶۷۶ | ۱.۶۷۱ | ۱.۶۶۴ | ۱.۶۶ | ۱.۶۴۵ |

مورد مهم دیگر که باید در نظر گرفته شود توجه به طول پیشوند یا پسوند (S) است. زیرا هر چه طول پیشوند یا پسوند در یک قانون بیشتر باشد، آن قانون می‌تواند کارایی بالاتری داشته باشد ولی در عین حال طول بیشتر پیشوند یا پسوند از تعداد کلمات سازگار (قابل اعمال) می‌کاهد و در ارزیابی تاثیر می‌گذارد. لذا طول S (پیشوند یا پسوند) را نیز به صورت زیر در ارزیابی تاثیر می‌دهیم:

$$\pi_{l_i} = \hat{p}_i^* - t_{(1-\alpha)/2}^{(n-1)} * \sqrt{\frac{\hat{p}_i^* * \left(1 - \hat{p}_i^*\right)}{n}} / \left(1 + \log\left(|S_i|\right)\right)$$

هنگامی که طول S یک باشد مقدار امتیاز قانون تغییر نمی‌کند ولی برای طول بیشتر مقدار امتیاز افزایش می‌یابد (یا به عبارت دیگر در عمل تفریق کمتر کاهش می‌یابد).

## ۶-۳-۲-۴. حذف قوانین غیر مفید

پس از این‌که قوانین امتیاز دهی شد باید قوانین غیر مفید حذف و زیرمجموعه‌ای از قوانین که کارایی بهتری دارند، انتخاب شود. برای این کار باید به صورت تجربی عمل کرد. قوانین بر اساس امتیازشان ابتدا مرتب می‌شوند. با انتخاب یک حد آستانه قوانین با امتیاز کمتر از این حد حذف می‌گردند. این قوانین به کلمات در مجموعه آزمون (کلمات با فراوانی یک در واژگان) اعمال می‌شود تا برچسب آن‌ها پیش‌بینی شود. در این مرحله اولیت اعمال قوانین بر اساس امتیاز می‌باشد. یعنی از بین قوانینی که قابلیت اعمال به یک کلمه را دارند قانون با امتیاز بالاتر اعمال می‌شود. پس از اعمال قوانین، صحت برچسب‌های پیش‌بینی شده تعیین می‌شود. این عملیات با حد آستانه‌های متفاوت تکرار می‌شود تا زیرمجموعه‌ای از قوانین که کارایی بهتری از خود نشان می‌دهد، انتخاب شود. نکته‌ای که باید توجه شود این است که با حذف برخی از قوانین ممکن است هیچ قانونی قابلیت اعمال به بعضی از کلمات را نداشته باشد. این موارد نباید در محاسبه صحت زیرمجموعه قوانین در نظر گرفته

شود و برای پیش‌بینی برچسب این کلمات در برچسب‌گذار می‌توان از روش‌های دیگر مانند روش شرح داده شده در ۶-۳-۱ استفاده کرد.

# نتیجه‌گیری

در این بخش به بررسی کلمات ناشناخته در زبان فارسی پرداختیم. مطالعه و بررسی رفتار توزیعی کلمات ناشناخته به تشخیص برچسب آن‌ها کمک می‌کند. دو روش برای به دست آوردن توزیع کلمات ناشناخته و هنچنین چند روش غلبه بر کلمات ناشناخته بیان شد. در نظر گرفتن محتمل‌ترین برچسب برای کلمات ناشناخته یک روش ناکارآمد برای کلمات ناشناخته است. روش دیگر که مورد توجه قرار گرفت استفاده از توزیع احتمالی کلمات ناشناخته است. توجه به پیشوندها و پسوندهای کلمات معمولا کارایی بهتری دارد که یک روش استخراج قوانین بر اساس پیشوند و پسوند کلمات تشریح شد. دقت و کارایی دو روش اخیر هنگام ارائه نتایج تجربی بررسی خواهد شد.

# هم‌نگاره‌ها در زبان فارسی

## مقدمه

هم‌نگاره‌ها کلمات با ساختار نوشتاری یکسان و تلفظ متفاوت در زبان‌های مختلف یکی از مهم‌ترین لایه‌های ابهام را در متن ایجاد می‌کنند. یکی از بارزترین کاربرد ابهام‌زدایی از هم‌نگاره‌ها، در سیستم‌های تبدیل متن به گفتار است. اگرچه تعداد کلمات هم‌نگاره نسبت به کل کلمات موجود و مورد استفاده در یک زبان کم است ولی تلفظ ناصحیح هم‌نگاره‌ها به جای یکدیگر موجب ابهام زیادی در درک متن می‌گردد. در برخی از زبان‌ها مانند زبان فارسی به دلیل ساختار خاص آن، تعداد هم‌نگاره‌ها نسبت به دیگر زبان‌ها زیاد است و مشکلی که ایجاد می‌کنند قابل اغماض نیست. تعداد زیادی از هم‌نگاره‌ها از ساختار زبان ناشی می‌شوند به عبارت دیگر هم‌نگاره‌ها در هر زبان به آن زبان خاص وابسته‌اند و هرچه ساختواژه اشتقاقی و تصریفی زبان پیچیده‌تر باشد می‌توان انتظار داشت هم‌نگاره‌هایی بیشتری وجود دارد و ابهام‌زدایی از آن‌ها نیز مشکل‌تر است. با این اوصاف می‌توان دریافت بازشناسی هم‌نگاره‌ها نیز تا حد زیادی وابسته به زبان خواهد بود. در زبان‌های دیگر روش‌های متن‌کاوی زیادی برای ابهام‌زدایی از هم‌نگاره‌ها تجربه شده است ولی تلاش در این زمینه برای زبان فارسی مغفول مانده است.

چون هم‌نگاره‌ها در هر زبان وابسته به همان زبان است، دسته‌بندی هم‌نگاره‌ها و توصیف انواع هم‌نگاره‌ها و دلایل ایجاد آن‌ها باید با توجه به هر زبان صورت گیرد. لذا ابتدا دلایل ایجاد هم‌نگاره‌ها را در زبان فارسی بررسی می‌کنیم و پس از آن به دسته بندی هم‌نگاره‌های فارسی می‌پردازیم. پس از بیان مشکلات ابهام‌زدایی از هم‌نگاره‌ها روشی برای ابهام‌زدایی از هم‌نگاره‌ها یعنی لیست‌های تصمیم‌گیری ( Yarowsky,1994) را مورد بررسی قرار می‌دهیم.

# علل هم‌نگارگی

به طور کلی، هم‌نگارگی در زبان فارسی ناشی از بازنمایی واجی و صرفی عناصر زبانی در خط فارسی است، به طوری که یک رابطهٔ چند به چند و در بعضی موارد یک رابطهٔ غیرنظام‌مند بین عناصر واجی و صرفی زبان فارسی و تظاهر نوشتاری آنها در خط فارسی وجود دارد. در مجموع، هم‌نگارگی در خط فارسی را می‌توان ناشی از عوامل زیر دانست (بی‌جن‌خان و مرادزاده، ۱۳۸۳):

۱- عدم بازنمایی واکه‌های کوتاه در خط فارسی: به عنوان مثال کلمه "مرد" دارای دو معنی و دو تلفظ متفاوت /mard/، /mord/ می‌باشد.

۲- عدم تناظر یک به یک میان واج‌ها و حروف فارسی: در زبان فارسی، بعضی از حروف نوشتاری تنها یک واج متناظر و برخی از واج‌ها تنها یک حرف نوشتاری متناظر دارند. مورد متاخر (تناظر برخی از واج‌ها با یک حرف نوشتاری) می‌تواند منجر به ایجاد هم‌نگاره شود. جدول۷-۱ چند مثال از این نمونه را نشان می‌دهد.

چند نمونه از عدم تناظر یک به یک میان واج‌ها و حروف فارسی

| مثال | واج | حرف |
|---|---|---|
| تو | /o/ | و |
| تو | /u/ | |
| جو | /av/ | |
| جو | /ow/ | |
| اشراف | /e/ | ا (الف) |
| اشراف ،ارگ | /a/ | |
| ارگ | /o/ | |
| علم | /a/ | ع |
| علم، عقاب | /e/ | |
| عقاب | /o/ | |

۳- یکسانی تظاهر واجی و نوشتاری تکواژها: در زبان فارسی ممکن است چندین تکواژ متفاوت دارای یک شکل نوشتاری یکسان باشند. در زیر دو نمونه از این یکسانی را نام می‌بریم:

- یکسانی تکواژ یاء نکره با یاء اسم‌ساز با شناسه (دوم شخص مفرد) و با یاء نسبت.

مثال:

یاء نکره: جوانی را دیدم. /ja'vaani/

یاء اسم‌ساز: جوانی نعمتی است. /javaa'ni/

یاء شناسه: تو هنوز جوانی. / ja'vaani/

یاء نسبت: مشکلات جوانی.... / javaa'ni /

- یکسانی تکواژهای ضمیر متصل سوم شخص مفرد با تکواژ اسم‌ساز.

مثال:

ضمیر متصل سوم شخص مفرد: به رویش خندیدم. /'ruyash/ /

تکواژ اسم‌ساز: رویش گل‌ها را ببین. /ru'yesh/ /

۴- رابطه بین وزن کلمات عربی و بعضی پسوندهای فارسی: به عنوان مثال کلمهٔ "منزلت" در باب مفعله (maf?alat) یک معنی دارد و با تعبیر "منزل تو" دارای پی حسب ملکی "ت" (at- ) معنی دیگری دارد.

## طبقه‌بندی هم‌نگاره‌ها

یک طبقه‌بندی که برای هم‌نگاره‌ها در (مرادزاده، ۱۳۸۳) ارائه شده است در اولین سطح هم‌نگاره‌ها را به دو دسته هم‌نگاره‌های تکیه‌ای و غیر تکیه‌ای تقسیم‌بندی می‌کند. هم‌نگاره‌های تکیه‌ای و غیر تکیه‌ای خود به دو دسته هم‌نگاره‌های بسیط و مرکب تقسیم‌بندی می‌شوند. هم‌نگاره‌های بسیط هم‌نگاره‌هایی هستند که ذاتا هم‌نگاره هستند و از ساختواژه زبان ناشی نمی‌شوند.

تعداد هم‌نگاره‌های تکیه‌ای بسیط بسیار کم می‌باشد. "ولی" و "گویا" دو هم‌نگاره تکیه‌ای بسیط هستند (بی‌جن‌خان و مرادزاده، ۱۳۸۳):

سرپرست : va´li ؛ اما : ´vali

واضح : gu´yaa ؛ مثل این که : ´guyaa

گروه دیگر هم‌نگاره‌های بسیط، هم‌نگاره‌های غیرتکیه‌ای بسیط می‌باشد. تعداد هم‌نگاره‌ها در این

گروه زیاد است. این گروه شامل هم‌نگاره‌های پرکاربرد است، بنابراین بیشترین ابهام را می‌توانند در متن ایجاد می‌کنند و ابهام‌زدایی از آن‌ها بسیار مهم است. نمونه‌ای از این هم‌نگاره‌ها در جدول ۷-۲ آمده است.

چند هم‌نگاره‌ها غیرتکیه‌ای بسیط

| صورت نوشتاری | تلفظ ۱ | تلفظ ۲ |
|---|---|---|
| اشراف | ?esh´raaf | ?ash´raaf |
| اعراب | ?e´?raab | ?a? ´raab |
| اعمال | ?e? ´maal | ?a? ´maal |
| بعد | bo?d | ba?d |
| پر | por | par |

در کنار هم‌نگاره‌های بسیط هم‌نگاره‌های مرکب وجود دارند. هم‌نگاره‌های مرکب هم‌نگاره‌هایی هستند که ناشی از ساختواژه اشتقاقی و تصریفی زبان یا افزوده شدن واژه‌بست‌ها به کلمات می‌باشند. هم‌نگاره‌های تکیه‌ای مرکب و غیر تکیه‌ای مرکب به زیربخش‌های زیادی تقسیم می‌شوند که در این مجال کوتاه به بررسی آن‌ها نمی‌پردازیم.

علاوه بر این دو دسته هم‌نگاره‌های تکیه‌ای و غیر تکیه‌ای، دسته‌ای دیگر نیز در (بی‌جن‌خان و مرادزاده، ۱۳۸۳) تحت عنوان هم‌نگاره‌های ترکیبی آورده شده است که در مورد هم‌نگاره‌هایی که صورت نوشتاری آن‌ها برحسب روابط هم‌نگارگی متعدد دارای تلفظ و معنی متعددی است صادق می‌باشد. به‌عنوان مثال کلمهٔ " نبرد " چهار تلفظ و معنی متفاوت دارد:

(۱) nabard در نبرد پیروز شد
(۲) nabarad اگر بازی را نبرد . . .
(۳) naborad دستش را نبرد!
(۴) nabord چرا جایزه را نبرد؟

رابطهٔ هم‌نگارگی بین کلمهٔ (۱) و کلمات (۲)، (۳) و (۴) از نوع هم‌نگاره‌های غیرتکیه‌ای مرکب یک‌سویه و رابطهٔ هم‌نگارگی بین کلمات (۲) و (۳ ) و (۴) از نوع هم‌نگاره‌های بن یکسان منفی است (بی‌جن‌خان و مرادزاده، ۱۳۸۳).

# ابهام‌زدایی از هم‌نگاره‌ها

## ۷-۴-۱. بررسی مشکلات موجود

در بررسی انجام‌شده در بخش قبل مشخص شد که هم‌نگاره‌ها در زبان فارسی انواع گوناگونی دارند و به روش‌های متفاوتی ساخته می‌شوند. به عبارت دیگر علاوه بر وجود هم‌نگاره‌هایی که ذاتا هم‌نگاره هستند (مانند هم‌نگاره‌های غیرتکیه‌ای بسیط) زبان فارسی ظرفیت بالایی برای ساخت هم‌نگاره‌های جدید که از ساختواژه زبان ناشی می‌شود را نیز دارد که این خود مشکل دیگری را نیز به همراه می‌آورد و آن تشخیص این مطلب است که آیا یک کلمه خاص در جمله هم‌نگاره است یا خیر. برای روشن‌تر شدن بحث زبان انگلیسی را در نظر بگیرید. در این زبان اکثر هم‌نگاره‌ها به صورت بسیط یا اشتقاقی هستند و این موارد را نیز می‌توان از مدخل‌های واژگان استخراج کرد. هم‌نگاره‌های تصریفی بخش کوچکی از هم‌نگاره‌ها را در این زبان تشکیل می‌دهند و قوانین یافتن آن‌ها ساده می‌باشد. مواردی نیز در مورد اعداد و تاریخ‌های عددی وجود دارد که یافتن این موارد نیز مشکل نیست. بنابراین در زبانی مثل زبان انگلیسی یافتن هم‌نگاره‌ها مشکل نمی‌باشد و تمرکز بیشتر بر روی ابهام‌زدایی از آن‌هاست. ولی در زبان فارسی علاوه بر هم‌نگاره‌هایی که می‌توان از مداخل واژگان استخراج کرد، هم‌نگاره‌های بسیاری وجود دارد که از ساختواژه زبان ناشی می‌شوند و در واژگان حضور ندارند. پس تشخیص این‌که یک کلمه در یک جمله فارسی یک هم‌نگاره است یا خیر خود مشکلی است بزرگ. برای رفع این مشکل نیاز به یک تحلیل‌گر ساختواژی بسیار قوی است که بتواند با دقت بالا و به صورت خودکار کلمات یک جمله را تحلیل کند.

اگر خود را محدود به هم‌نگاره‌های بسیط کنیم باز هم مشکلاتی وجود دارد. اولین مسئله این است که ابهام‌زدایی از این هم‌نگاره‌ها نیاز به پیکره‌های بسیار بزرگی دارد که هم‌نگاره‌ها در آن برچسب‌گذاری تلفظی شده باشند یعنی تلفظ صحیح هر هم‌نگاره در زمینه آن مشخص شده باشد که دسترسی به پیکره‌هایی با این مشخصات نیز در زبان فارسی مشکل می‌باشد.

از جمله مسائل دیگری که می‌توان بیان کرد این است که معمولا فراوانی‌های تلفظات مختلف یک هم‌نگاره در پیکره تفاوت زیادی با هم دارند، به عنوان مثال برای هم‌نگاره "اسناد" فراوانی تلفظ

/as´naad?/ بسیار بیشتر از فراوانی تلفظ /es´naad?/ است. این مسئله برای روش‌های ابهام‌زدایی آماری مشکل ساز می‌باشد.

با وجود این مسائل می‌توان گفت که ابهام‌زدایی از هم‌نگاره‌ها یا حداقل بخش گسترده‌ای از هم‌نگاره‌ها در زبان فارسی نیاز به بررسی و مطالعه بسیار عمیق‌تر و فراهم آمدن پیش‌نیازهایی بسیاری است. در ادامه و در بخش بعد روشی برای ابهام‌زدایی از هم‌نگاره‌ها بررسی شده است که ما آن را برای ابهام‌زدایی از هم‌نگاره‌های غیر تکیه‌ای بسیط با فراوانی بالا به خدمت می‌گیریم.

## ۷-۴-۲. ابهام‌زدایی از هم‌نگاره‌های با فراوانی بالا

لیست‌های تصمیم‌گیری (Yarowsky, 1994) یک روش متن‌کاوی برای ابهام‌زدایی از هم‌نگاره‌هاست که حاصل ترکیب سه روش دیگر ارائه شده در این حوزه می‌باشد: رده‌بندی‌کننده Bayesian، برچسب‌گذار N-gram، و درخت‌های تصمیم‌گیری. این روش از زمینه محلی هم‌نگاره‌ها برای ابهام‌زدایی استفاده می‌کند. این روش متشکل از چند گام است که در زیر شرح داده می‌شود:

۱- جمع‌آوری و برچسب‌گذاری مجموعه آموزشی: برای هر هم‌نگاره، همه رخدادهای آن در یک پیکره جمع‌آوری می‌شود و سپس هر مورد با توجه به تلفظ صحیح آن در زمینه مورد نظر برچسب‌گذاری تلفظی می‌شود. به عنوان مثال در زیر هم‌نگاره "شرف" در چند جمله برچسب‌گذاری تلفظی شده است:

| مثال | برچسب تلفظی |
|---|---|
| مگر شرف و بزرگواری آدمی به کرامت نفس او نیست ؟ | sha´raf |
| او دیگر در شرف بازنشستگی بود . | sho´rof |
| و خلاصه این که ناحیه زرخیز خوزستان از بی فکری و ظلم حکمران در شرف انهدام است . | sho´rof |
| زمین ، می توانند پیوسته در رحمت و آسایش به سر برند ، مادامی که بشر دوست ، امین مال و آبرو و شرف و پیرو حق و عدالت باشند . | sha´raf |

۲- به دست آوردن توزیع‌های باهم‌آیی[1]: با تعریف چند قانون مانند "یک کلمه بعد"، "یک کلمه قبل"، "دو کلمه بعد"، "دو کلمه قبل" و ... رخداد دیگر کلمات با یک تلفظ خاص از یک هم‌نگاره به دست می‌آید. بر اساس این رخدادهای باهم‌آیی می‌توان به توزیع‌هایی دست یافت که برای ابهام‌زدایی از هم‌نگاره‌ها استفاده شود. هر کدام از این توزیع‌ها به منزله یک پیشامد احتمالی می‌باشد.

۳- محاسبه نرخ درست‌نمایی[2]: با استفاده از لگاریتم نرخ درست‌نمایی پیشامدها می‌توان به قوانینی برای ابهام‌زدایی از هم‌نگاره‌ها دست یافت. لگاریتم نرخ درست‌نمایی با فرمول زیر به دست می‌آید:

$$\left| Log\left( \frac{P(pronunciation_1 \mid collocation_i)}{P(pronunciation_2 \mid collocation_i)} \right) \right|$$

در این فرمول لگاریتم نرخ درست‌نمایی بر اساس نسبت دو تلفظ متفاوت یک هم‌نگاره با توجه به یک پیشامد باهم‌آیی خاص اندازه‌گیری می‌شود.

۴- مرتب‌سازی پیشامدها بر اساس نرخ درست‌نمایی و قرار دادن آن‌ها در لیست‌های تصمیم‌گیری: توزیع‌های باهم‌آیی که می‌توانند به عنوان تمیز دهنده یک تلفظ خاص به کار روند دارای لگاریتم بزرگ‌نمایی بالا هستند. با مرتب‌سازی نزولی این مقادیر قوانین با قدرت تمایزدهی بیشتر در ابتدای لیست قرار می‌گیرند. بنابراین برای هر هم‌نگاره می‌توان لیستی تحت عنوان لیست تصمیم‌گیری به دست آورد. برای مثال برای هم‌نگاره "شرف"، برخی از پیشامدهای با درست‌نمایی بالا و قانون استفاده شده در زیر لیست شده است:

| قانون | رخداد باهم‌آیی | مقدار درست‌نمایی | تلفظ |
|---|---|---|---|
| دو کلمه قبل | حیثیت و | ۶.۴ | shá raf |
| دو کلمه قبل | عزت و | ۶.۴ | shá raf |
| یک کلمه قبل و یک کلمه بعد | در ... تأسیس | ۶.۲ | sho´rof |
| دو کلمه بعد | وقوع است | ۵.۲ | sho´rof |

[1] Collocational distributions
[2] Likelihood ratio

۵- استفاده از لیست‌های تصمیم‌گیری: پس از ساخت لیست‌های تصمیم‌گیری، می‌توان از آن‌ها برای تعیین تلفظ کلمات مبهم در یک زمینه جدید استفاده کرد. با فرض حضور یک هم‌نگاره در یک زمینه جدید بالاترین پیشامد، در لیست تصمیم‌گیری که در این زمینه حضور داشته باشد ملاک تعیین تلفظ هم‌نگاره قرار می‌گیرد.

اگر هم‌نگاره‌ای در زمینه‌ای رخ دهد و پیشامدهای موجود در لیست تصمیم‌گیری در آن زمینه رخ ندهد می‌توان فراوانی را ملاک تصمیم‌گیری در نظر گرفت و تلفظ با فراوانی بیشتر را برای آن هم‌نگاره در آن زمینه به عنوان تلفظ صحیح لحاظ کرد.

کاربرد اولیه روش لیست‌های تصمیم‌گیری برای ابهام‌زدایی هم‌نگاره‌های غیر اشتقاقی و غیر تصریفی است، زیرا به صورت آماری عمل می‌کند و به مواردی همچون اشتقاق و تصریف که مختص به زبان می‌باشد وابسته نمی‌باشد. ما نیز از این روش برای ابهام‌زدایی از تعدادی از هم‌نگاره‌های غیر تکیه‌ای بسیط که فراوانی بالایی در متن دارند استفاده خواهیم کرد. دلیل انتخاب هم‌نگاره‌های با فراوانی بالا این است که بزرگی پیکره مورد استفاده برای این امر چندان قابل توجه نبوده است.

## نتیجه‌گیری

در این بخش مسئله هم‌نگاره‌ها در زبان فارسی مورد بررسی قرار گرفت. ابتدا به بررسی دلایل هم‌نگارگی پرداخته شد و سپس نحوه طبقه‌بندی هم‌نگاره‌ها بیان شد. مفهوم هم‌نگاره‌های تکیه‌ای و غیرتکیه‌ای، هم‌نگاره‌ها بسیط و مرکب تبیین شد. بعد از آن به بررسی مشکلات ابهام‌زدایی از هم‌نگاره‌ها در زبان فارسی پرداخته شد. یک روش ابهام‌زدایی از هم‌نگاره‌ها روش لیست‌های تصمیم‌گیری است که مراحل آن با بیان مثال شرح داده شد. نتایج اعمال این روش بر روی هم‌نگاره‌های با فراوانی بالا در بخش نتایج تجربی ارائه می‌گردد.

# نتایج تجربی

# مقدمه

در فصول پیشین مباحث مختلفی در راستای برچسب‌گذاری در زبان فارسی بیان شد. حال در این فصل به ارائه نتایج به‌دست آمده در برچسب‌گذاری و ابهام‌زدایی از هم‌نگاره‌های زبان فارسی پرداخته می‌شود. ابتدا روش ارزیابی در برچسب‌گذاری شرح داده می‌شود و سپس نتایج حاصل از اعمال دو برچسب‌گذاری ماکوفی bigram و trigram، و برچسب‌گذار مبتنی بر حافظه بر روی بخشی از پیکره ارائه خواهد شد. پس از آن نتایج حاصل از ابهام‌زدایی از هم‌نگاره‌های با فراوانی بالا ارائه می‌گردد. در هر مورد سعی می‌شود تحلیلی از نتایج به‌دست آمده بیان گردد.

# نتایج برچسب‌گذاری

## ۸-۲-۱. روش ارزیابی

برای ارزیابی سیستم برچسب‌گذاری مشابه دیگر کارهای ارائه شده در این حوزه عمل می‌کنیم (Thede and Harper, 1999) (Lee et al., 2000) (Brants, 1999). ارزیابی برچسب‌گذارها معمولا با توجه به دقت[1] آن‌ها صورت می‌گیرد. برای ارزیابی از تکنیک اعتبارسنجی متقابل ۵ قسمتی[2] (Tan et al., 2005) استفاده می‌کنیم تا ارزیابی هرچه دقیق‌تر انجام شود. در این روش ارزیابی، ابتدا داده

[1] Accuracy
[2] 5-fold cross validation

آموزشی به ۵ قسمت تقسیم می‌شود، سپس برچسب‌گذارهای شرح داده شده در فصول قبل، ۵ مرتبه اجرا می‌شوند به طوری که در هر مرتبه ۴ قسمت برای آموزش و ۱ قسمت برای آزمون مورد استفاده قرار می‌گیرد و در نهایت دقت برچسب‌گذاری با توجه به نتایج ۵ مرتبه اجرای متفاوت به دست می‌آید. همچنین دقت‌های جداگانه برای کلمات شناخته شده و ناشناخته و دقت کلی ارائه خواهد شد.

به دلیل محدودیت‌های محاسباتی (پیچیدگی زمانی و مکانی) نمی‌توان همه پیکره را برای ارزیابی مورد استفاده قرار داد. به همین علت ما تنها بخشی از پیکره را در نظر می‌گیریم و به عنوان ورودی به برچسب‌گذارها می‌دهیم. همان‌گونه که در فصل ۲ ذکر شد متون موجود در پیکره به دو صورت رسمی و غیر رسمی و محاوره‌ای می‌باشد. در این‌جا سعی می‌شود بخشی از پیکره که برای برچسب‌گذاری استفاده می‌شود، شامل متون رسمی باشد و متون غیر رسمی و محاوره‌ای را شامل نشود. علت این امر این است که متون غیر رسمی و محاوره‌ای نسبت به متون رسمی بخش کمی از پیکره را به خود اختصاص می‌دهند، به علاوه کلمات و جملات فارسی در محاوره دچار تغییر زیادی می‌شوند و این تغییرات معمولا از قاعده خاصی پیروی نمی‌کند و بررسی رفتار کلمات هنگام افزودن شدن تکواژهای دیگر به آن‌ها را بسیار مشکل می‌کند و این مسائل باعث مشکلاتی در سیستم‌های برچسب‌گذاری می‌شود، بنابراین برای برچسب‌گذاری متون غیر رسمی و محاوره‌ای نیاز به پیکره‌های مناسب و بررسی تغییرات کلمات و ساختار جملات محاوره‌ای می‌باشد.

بخشی از پیکره را که برای برچسب‌گذاری و ارائه نتایج انتخاب کرده‌ایم برای همه آزمایشات به طور یکسان به کار می‌بریم تا مقایسه نتایج به دست آمده بهتر انجام گیرد. به علاوه از دو نسخه پیکره (رک.، فصل ۲) فایل‌های یکسانی را انتخاب می‌کنیم تا امکان مقایسه نتایج برچسب‌گذاری بر روی یک مجموعه متن یکسان با دو مجموعه برچسب متفاوت فراهم آید. در بخش بعد هنگام ارائه نتایج اختلاف اندکی در تعداد کلمات فایل‌های انتخاب شده نسخه یک و دو مشاهده می‌شود که به دلیل دو است: یا اصلاحاتی در برخی از فایل‌ها انجام شده، به عنوان مثال برخی قسمت‌های تکراری حذف شده، یا در بیان اطلاعات افزوده شده به ابتدای فایل‌ها که معرف مشخصات منبع و موضوع متن می‌باشد تغییراتی صورت گرفته است.

## ۸-۲-۲. نتایج برچسب‌گذاری مقولات اصلی

همان‌طور که ذکر شد دقت برچسب‌گذاری هر روش برای کلمات شناخته شده و ناشناخته جداگانه ارائه می‌شود. برای غلبه بر کلمات ناشناخته در دو برچسب‌گذار bigram و trigram روش شرح داده شده در بخش ۶-۳-۲ (توجه به وندها) مورد استفاده قرار می‌گیرد. در این روش باید سه پارامتر تعیین شود که مقدار بهینه را پس از آزمایشات گوناگون گزارش می‌دهیم: طول وندهای استخراجی قوانین غیر ساختواژی ۵ در نظر گرفته می‌شود، حد آستانه برای حذف قوانین ساختواژی با فراوانی کم ۴ و حد آستانه برای حذف قوانین غیر ساختواژی با فراوانی کم ۲۰ تعیین می‌گردد. پس از استخراج قوانین و امتیازدهی به قوانین در مرحله حذف قوانین غیر مفید، مقدار حد آستانه برای حذف قوانین با امتیاز کم به صورت تجربی ۵/. به دست می‌آید. مقادیر پارامترها در همه آزمایشات یکسان می‌باشد.

در صورت عدم امکان اعمال روش فوق (به دلیل عدم وجود قانون قابل اعمال) روش گفته شده در بخش ۶-۳-۱ (توزیع احتمالی کلمات ناشناخته) برای غلبه بر کلمات ناشناخته استفاده می‌شود. برای به دست آوردن توزیع احتمالی کلمات ناشناخته از روش متقابل قرار دادن مجموعه آزمون با مجموعه آموزش استفاده می‌کنیم. برای این‌که توزیع دقیق‌تری به‌دست آید از روش اعتبارسنجی متقابل ۵ قسمتی استفاده می‌کنیم.

جدول ۸-۱ توزیع احتمالی کلمات ناشناخته را در مقولات اصلی برای نسخه یک پیکره نشان می‌دهد.

توزیع احتمالی کلمات ناشناخته در مقولات اصلی (نسخه یک پیکره)

| **برچسب** | اسم | صفت | قید | فعل | عربی | بقیه برچسب‌ها |
|---|---|---|---|---|---|---|
| **احتمال** | ٪۶۱ | ٪۲۱ | ٪۱ | ٪۱۰ | ٪۴ | ٪۳ |

همان‌گونه که انتظار می‌رود کلمات ناشناخته بیشتر از مقولات باز مانند اسم، صفت و فعل هستند. به علاوه چون کلمات و عبارات عربی در متون فارسی معمولا زیاد استفاده می‌شود آن‌ها نیز درصدی از کلمات ناشناخته را به خود اختصاص می‌دهند.

علاوه بر برچسب‌های مارکوفی bigram و trigram که در فصل ۳ بیان شد، ما برچسب‌گذار مبتنی

بر حافظه را نیز در فصل ۴ شرح دادیم تا امکان مقایسه یک برچسب‌گذار دیگر با برچسب‌گذارهای مارکوفی bigram و trigram فراهم شود. در برچسب‌گذار مبتنی بر حافظه باید اطلاعاتی که پایگاه نمونه کلمات شناخته شده و پایگاه نمونه کلمات ناشناخته ذخیره می‌شود تعیین شود. برای برچسب‌گذاری مبتنی بر حافظه از ابزار MBT (Daelemans et al., 2003) استفاده می‌شود. برای تعیین این‌که چه اطلاعاتی مفیدتر است تا در پایگاه‌های نمونه ذخیره شود باید به صورت تجربی عمل کرد. پس از آزمایشات متعدد به نظر می‌رسد که قرار دادن این اطلاعات در پایگاه نمونه کلمات شناخته شده نتیجه بهتری ارائه می‌دهد: برچسب‌های دو کلمه 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 هنگام عمل آموزش به‌دست می‌آید). همچنین در پایگاه نمونه کلمات ناشناخته این اطلاعات ذخیره می‌شود: برچسب‌های دو کلمه قبل، برچسب مبهم کلمه بعد، حرف اول کلمه، سه حرف آخر کلمه. برای همه آزمایشات همین اطلاعات را در پایگاه‌های نمونه قرار خواهیم داد.

نتایج برچسب‌گذاری با برچسب‌گذارهای مارکوفی bigram و trigram و برچسب‌گذاری مبتنی بر حافظه (MBT) بر روی نسخه یک پیکره در جدول ۸-۲ قابل ملاحضه می‌باشد.

نتایج سه برچسب‌گذار bigram، trigram و MBT در برچسب‌گذاری مقولات اصلی (نسخه یک پیکره)

| | تعداد | Bigram | Trigram | MBT |
|---|---|---|---|---|
| کلمات شناخته‌شده | ۱۱۶۵۸۷۱ | ٪۹۶.۲ | ٪۹۶.۴ | ٪۹۴.۷ |
| کلمات ناشناخته | ۲۳۶۴۲ | ٪۷۳.۵ | ٪۷۳.۵ | ٪۴۴.۷ |
| کل کلمات | ۱۱۸۹۵۱۳ | ٪۹۵.۸ | ٪۹۵.۹ | ٪۹۳.۷ |

جدول ۸-۳ توزیع احتمالی کلمات ناشناخته را در مقولات اصلی برای نسخه دو پیکره نشان می‌دهد.

توزیع احتمالی کلمات ناشناخته در مقولات اصلی (نسخه دو پیکره)

| **برچسب** | اسم | صفت | قید | فعل | عدد | متفرقه | بقیه برچسب‌ها |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **احتمال** | ٪۵۶ | ٪۲۰ | ٪۱ | ٪۱۰ | ٪۲ | ٪۷ | ٪۴ |

چون مجموعه برچسب نسخه دو پیکره با مجموعه برچسب نسخه یک متفاوت می‌باشد جدول ۸-۳ با جدول ۸-۱ تفاوت دارد، یعنی توزیع کلمات ناشناخته در دو نسخه به دلیل تفاوت در مجموعه برچسب متفاوت می‌باشد. در نسخه دو، مقوله اصلی عدد بخشی از کلمات ناشناخته را به خود اختصاص می‌دهد. دلیل این‌که برچسب "متفرقه" درصد نسبتا بالایی از کلمات ناشناخته را به خود اختصاص داده است این است که برچسب "عربی" که در نسخه یک پیکره به عنوان یک مقوله اصلی بود، در نسخه دو پیکره به عنوان یک زیرمقوله برای مقوله اصلی "متفرقه" در نظر گرفته شده است.

جدول ۸-۴ نتایج برچسب‌گذاری با برچسب‌گذارهای bigram و trigram و برچسب‌گذار مبتنی بر حافظه (MBT) را در نسخه دو پیکره نشان می‌دهد.

نتایج سه برچسب‌گذار bigram، trigram و MBT در برچسب‌گذاری مقولات اصلی (نسخه دو پیکره)

| | 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 |

## ۸-۲-۳. نتایج برچسب‌گذاری با کمک تحلیل‌گر ساختواژی

برای افزایش کارایی روش‌های برچسب‌گذاری در برچسب‌گذاری پیکره متنی زبان فارسی روش شرح داده شده در بخش ۵-۵-۲ (استفاده از یک تحلیل‌گر ساختواژی با امکان تجزیه تصریفی برای برچسب‌گذاری) را به کار می‌بریم. در این‌جا گزارش نتایج استفاده از این روش را محدود به نسخه دو پیکره می‌کنیم. دلیل این کار این است که ما این روش را با توجه به مجموعه برچسب نسخه دو پیکره توسعه داده‌ایم. ولی انتظار می‌رود که اگر این روش بر روی نسخه یک پیکره نیز اعمال شود نتایج مشابهی به دست آید.

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. چند نمونه از برچسب‌های حذف شده را که قبلا نیز ذکر شد بیان می‌کنیم. در مقوله اسم برچسب‌های DAY (روز)، LOC (مکان)، DIR (جهت)، SES (فصل)، MON (ماه)، SURN (لقب) و TIME (زمان) حذف می‌شوتد. در مقوله قید نیز چون انواع مختلفی از قیدها وجود دارند حذفیاتی صورت می‌گیرد: TIME (زمان)، LOC (مکان)، EXM (مثال)، ORD (ترتیبی)، REPT (تکرار)، I (پرسشی)، NEGG (نفی).

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 عملا برچسب‌گذار را ناکارامد خواهد کرد.

در مرحله دو در بخش ۵-۵-۲ همان‌طور که گفته جدولی از تکواژهای تصریفی و واژه‌بست‌ها استخراج می‌شود که از توضیح مجدد آن خودداری می‌کنیم.

در مرحله سه هر کلمه به بن‌واژه و تکواژهای تصریفی و واژه‌بست‌های تشکیل‌دهنده 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. پس از ارائه نتایج برچسب‌گذاری به تحلیل نتایج می‌پردازیم.

ابتدا باید به مبحث کلمات ناشناخته بپردازیم. برای غلبه بر کلمات ناشناخته مشابه بخش قبل عمل می‌شود. جدول ۸-۵ توزیع احتمالی کلمات ناشناخته را برای نسخه دو پیکره پس از تجزیه تصریفی کلمات نشان می‌دهد. همان‌طور که مشاهده می‌شود بیشتر کلمات ناشناخته اسم عام، اسم خاص و یا صفت ساده می‌باشند.

توزیع احتمالی کلمات ناشناخته در مقولات اصلی پس از تجزیه تصریفی در نسخه دو پیکره

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

نتایج برچسب‌گذاری سه برچسب‌گذار bigarm، trigram و مبتنی بر حافظه (MBT) بر روی این

کلمات تجزیه شده در جدول ۸-۶ قابل مشاهده می‌باشد. (دلیل افزایش تعداد کلمات تجزیه آن‌هاست وگرنه داده استفاده شده همان داده استفاده شده در جدول ۸-۴ می‌باشد.)

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 |

برچسب‌گذاری کلمات ناشناخته در زبان فارسی به دلیل ساختار آن بسیار مشکل می‌باشد. به عنوان 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 در بعد اشتقاق بتواند نتایج را بهبود بخشد.

نتایج برچسب‌گذاری کلمات ناشناخته در برچسب‌گذار bigram و tirgram پس از تجزیه 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 جز موارد خطا می‌باشند. علاوه بر این مشکل در این‌جا، مقوله اسم در دو زیرمقوله جداگانه اسم عام و اسم خاص در نظر گرفته شده‌اند. یعنی کلماتی که در بخش قبل به طور صحیح در مقوله 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 کلمه خطا لحاظ گردد. یعنی عدم وجود نشانه‌های خاص برای تمایز اسامی عام از اسامی خاص در زبان فارسی باعث می‌شود که در تشخیص این دو مقوله خطای زیادی رخ دهد. برای مقایسه یک روش ساده را در زبان انگلیسی بیان می‌کنیم: به کلمات ناشناخته با حرف اول بزرگ برچسب اسم خاص و در غیر این صورت برچسب اسم عام منتسب شود. همین روش ساده باعث می‌شود که دقت بالایی حاصل شود. اگر پسوندها و پیشوندهای مشخصی که برای ساخت کلمات متفاوت مانند صفات و قیدها نیز به کار می‌روند لحاظ شود دقت

بسیار خوبی برای کلمات ناشناخته به‌دست می‌آید. در زبان فارسی عدم وجود چنین موارد مشخصی مسئله غلبه بر کلمات ناشناخته را بسیار مشکل می‌کند.

نتایج برچسب‌گذاری مبتنی بر حافظه حتی با در نظر گرفتن یک حرف اول کلمه، سه حرف آخر کلمه، برچسب‌های ابهام‌زدایی شده از دو کلمه قبل و برچسب مبهم کلمه بعد بسیار پایین است و حتی نتیجه عکس به دست آمده، زیرا اگر تمام کلمات ناشناخته اسم در نظر گرفته می‌شد دقت بالاتری به دست می‌آمد چون با توجه به جداول ۸-۱ و ۸-۳ 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 کاملا ناکارامد می‌باشد. ولی در جدول ۸-۶ مشاهده می‌کنیم که این دقت بسیار بهبود یافته و حتی از برچسب‌گذاری‌های مارکوفی نیز نتیجه بهتری به‌دست آمده است. این مسئله دلیلی دارد که پس از تحلیل نتایج کلی بیان خواهد شد.

مشکل دیگر، کلمات در عبارات عربی است که هم تاثیر بر دقت برچسب‌گذاری کلمات ناشناخته می‌گذارد و هم بر دقت برچسب‌گذاری کلمات شناخته شده. برای روشن‌تر شدن مطلب جمله زیرا را که از بخشی از پیکره استخراج شده در نظر بگیرد:

و CON پیام N,COM,SING,EZ الهی AJ,SIM فرمود V,PA,SIM,POS,3 : DELM 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

همان‌طور که قبلا نیز ذکر شد در پیکره متنی زبان فارسی به کلمات در عبارات و جملات عربی برچسب AR منتسب می‌شود. مثلا در جمله فوق همه کلمات در عبارت عربی برچسب RES,FW,AR دارند. در این حالات هیچ‌گاه برچسب‌گذار آماری نمی‌تواند ارتباط بین برچسب‌های جمله را یاد بگیرد و حتی این دنباله از برچسب‌های یکسان بر عملکرد برچسب‌گذار تاثیر منفی نیز دارد. کلمات الان، قد، قبل و من در عبارت عربی فوق چون شکل یکسانی با کلمات فارسی دارند و معمولا در فرآیند مجموعه آموزش حضور دارند در هنگام آموزش برچسب‌گذار به عنوان کلمات شناخته‌شده لحاظ می‌شود. بنابراین می‌توان انتظار داشت که هنگام برچسب‌گذاری به این کلمات برچسی غیر از AR منتسب شود و خطا صورت گیرد. ملاحظه کردیم که بخشی قابل توجهی از کلمات ناشناخته برچسب AR داشتند و این کلمات در اکثر قریب به اتفاق موارد اشتباه برچسب‌گذاری می‌شوند. مثلا اگر کلمات عصیت و المفسدین در فرآیند آموزش دیده نشده باشند

برچسب‌گذار سعی می‌کند با توجه به زمینه یا حروف ابتدا و انتهای آن‌ها برچسب صحیح را تشخیص دهد. ولی هنگامی که کلمات در زمینه همگی برچسب یکسان دارند نه تنها توجه به این زمینه مفید نخواهد بود بلکه گمراه کننده نیز می‌باشد و حروف ابتدا و انتهای این کلمات نیز نمی‌تواند کمک چندانی بکند زیرا به دلیل تفاوت کلمات عربی و ساختار آن‌ها با کلمات فارسی اطلاعات به دست آمده اثر منفی در غلبه بر کلمات ناشناخته فارسی می‌گذارد. بنابراین از بحث حاضر این نکته استنباط می‌شود که در برچسب‌گذاری در زبان فارسی باید سیستمی برای تشخیص عبارات عربی تعبیه شود. این سیستم با توجه به طرح کلی ارائه شده برای برچسب‌گذاری در فصل یک، باید در ادغام با بخش تشخیص کران کلمات و جملات عمل کند.

حال به نتایج کلی می‌پردازیم. نتایج کلی برچسب‌گذارهای مارکوفی در برچسب‌گذاری مقولات اصلی بر روی نسخه یک و نسخه دو پیکره تقریبا یکسان می‌باشد و در هر دو مورد دقت این برچسب‌گذارها از برچسب‌گذار مبتنی بر حافظه اندکی بیشتر است. ولی پس از تجزیه 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 برچسب کلمات بیشتر به برچسب کلمات قبلی است تا برچسب کلمات بعدی، که در برچسب‌گذارهای مبتنی بر مدل مارکوف این وابستگی در آرایه احتمال انتقالات خلاصه می‌گردد. به همین علت و با در نظر گرفتن نحوه عملکرد الگوریتم Viterbi برچسب‌گذارهای مارکوفی در مقولات اصلی نتیجه خوبی ارائه می‌دهند و دقت آن‌ها از برچسب‌گذار مبتنی بر حافظه بهتر است هرچند که این برچسب‌گذار به برچسب دو کلمه بعد برای کلمات شناخته شده و برچسب یک کلمه 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 می‌شوند و تعداد برچسب‌ها زیاد می‌شود وابستگی برچسب کلمات به برچسب کلمات بعدی افزایش می‌یابد. در ضمن با تجزیه کلمات اطلاعات زمینه افزایش می‌یابد. حال برچسب‌گذار مبتنی بر حافظه که به دلیل استفاده از روش نزدیکترین k همسایه نسبت به زمینه بسیار حساس می‌باشد و در مورد کلمات شناخته شده علاوه بر برچسب‌های دو کلمه قبلی و خود کلمه، به برچسب دو کلمه بعدی توجه

می‌کند؛ و در مورد کلمات ناشناخته علاوه بر برچسب‌های دو کلمه قبلی و یک حرف از اول و سه حرف از آخر کلمه، برچسب کلمه بعدی را 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 ۸-۶ بسیار بهتر از دو برچسب‌گذار مارکوفی بوده است.

# نتایج ابهام‌زدایی از هم‌نگاره‌ها

برای ابهام‌زدایی از هم‌نگاره‌ها روش لیست‌های تصمیم‌گیری را در فصل ۷ شرح دادیم. در این بخش از این روش برای ابهام‌زدایی از تعدادی هم‌نگاره با فراوانی بالا استفاده می‌کنیم و نتایج آن را می‌بینیم.

## ۸-۳-۱. جمع‌آوری داده آموزشی

لیست‌های تصمیم‌گیری برای یادگیری نیاز به داده آموزشی دارند و هرچه این داده آموزشی بیشتر باشد کارایی آن بهتر خواهد بود. برای جمع آوری داده آموزشی از همان بخش برچسب‌خورده پیکره متنی زبان فارسی استفاده می‌شود. برای این کار هم‌نگاره‌های مورد نظر در پیکره جستجو و هر هم‌نگاره‌ها به همراه ۲۰ کلمه قبل و ۲۰ کلمه بعد از آن استخراج می‌شود. چون الگوریتم، یادگیری را بر اساس زمینه انجام می‌دهد، این طول زیاد پنجره (۲۰ کلمه قبل و ۲۰ کلمه بعد از هم‌نگاره) امکان آزمایش الگوریتم بر روی زمینه‌های با طول متفاوت را فراهم می‌کند یعنی به عبارت دیگر با توجه به مطالب ارائه شده در فصل ۷ می‌توان قوانین مختلفی را آزمایش و مورد ارزیابی قرار داد.

بعد از استخراج هم‌نگاره‌ها از پیکره، با توجه به زمینه‌ای که هم‌نگاره در آن رخ داده است برچسب تلفظی به کلمات هم‌نگاره نسیت داده می‌شود. پس از انتساب برچسب تلفظی به هم‌نگاره‌ها مجموعه داده‌ای به دست می‌آید که می‌توان از آن برای ارزیابی الگوریتم استفاده کرد.

## ۸-۳-۲. نتایج ابهام‌زدایی از هم‌نگاره‌های با فراوانی بالا

مشکل اساسی در الگوریتم‌های یادگیری آماری نیاز به داده آموزشی بسیار بالاست تا ارتباطات بین نمونه‌ها به خوبی استخراج شود. لیست‌های تصمیم‌گیری نیز که بر اساس توزیع رخداد کلمات عمل می‌کنند از این مشکل رنج می‌برند. از طرف دیگر نیز باید توجه داشت که کلمات هم‌نگاره بخش نسبتا کوچکی از کلمات زبان را تشکیل می‌دهند و احتمال رخداد بسیاری از هم‌نگاره‌ها در جملات کم می‌باشد، لذا نیاز به پیکره‌های بسیار بزرگ می‌باشد تا تعداد رخداد هم‌نگاره‌ها در حد مطلوبی باشد و بتوان از روش‌های آماری برای رفع ابهام از آن‌ها استفاده کرد.

در زبان فارسی پیکره‌های بسیار بزرگ وجود ندارد و همان‌گونه که ذکر شد پیکره مورد استفاده در این تحقیق نسبتا کوچک بوده است، در صورتی که در کارهای مشابه در زبان‌های دیگر پیکره مورد استفاده در حدود ۴۰۰ میلیون کلمه داشته است (Yarowsky, 1994). چون بسیاری از هم‌نگاره‌های فارسی کاربرد کمتری نسبت دیگر کلمات دارند، فراوانی بسیار کمی در پیکره داشته‌اند و بنابراین برای نمایش کارایی این الگوریتم در زبان فارسی این روش را بر روی هم‌نگاره‌های پرکاربردتر و با فراوانی بالا اعمال کنیم.

برای ارزیابی الگوریتم لیست‌های تصمیم‌گیری برای ابهام‌زدایی از هم‌نگاره‌های فارسی از روش اعتبارسنجی متقابل ۱۰ قسمتی[3] استفاده می‌کنیم تا از تمام داده آموزشی بهره برده باشیم و ارزیابی هرچه دقیق‌تر انجام شود. در این روش ارزیابی، ابتدا داده آموزشی به ۱۰ قسمت تقسیم می‌شود، سپس الگوریتم لیست‌های تصمیم‌گیری شرح داده شده در بالا ۱۰ مرتبه اجرا می‌شود به طوری که در هر مرتبه ۹ قسمت برای یادگیری الگوریتم و یک قسمت برای آزمون مورد استفاده قرار می‌گیرد و در نهایت دقت الگوریتم بر روی داده آموزشی با توجه به نتایج ۱۰ مرتبه اجرای متفاوت به دست می‌آید. ما روش لیست‌های تصمیم‌گیری را بر روی چند هم‌نگاره غیر تکیه‌ای با فراوانی بالا در زبان فارسی اعمال می‌کنیم. برای این کار مجموعه قوانین زیادی آزمایش می‌شود تا مفیدترین قوانین پیدا شود. سرانجام مجموعه قوانینی که بهترین نتیجه را ارائه کردند به این صورت است: "یک کلمه قبل"، "یک

[3] 10-fold cross validation

کلمه بعد"، "دو کلمه قبل"، "دو کلمه بعد"، "یک کلمه قبل و یک کلمه بعد"، "پنجره‌ای به طول ۵± کلمه" (یعنی رخداد یک کلمه در فاصله حداکثر ۵ کلمه قبل یا بعد از هم‌نگاره). جدول۶ نتایج حاصل از این روش را بر روی چند هم‌نگاره غیر تکیه‌ای با فراوانی بالا نشان می‌دهد.

نتایج روش لیست‌های تصمیم‌گیری برای چند هم‌نگاره با فراوانی بالا

| هم‌نگاره | تلفظ ۱ | تلفظ ۲ | تعداد صحیح | تعداد غلط | دقت |
|---|---|---|---|---|---|
| شرف | /sha´raf/ | /sho´rof/ | ۱۳۴ | ۳ | ٪۹۷.۸ |
| اعمال | /?a?`mal/ | /?e?`mal/ | ۱۵۸۸ | ۱۸۴ | ٪۸۹.۶ |
| گرم | /garm/ | /ge´ram/ | ۱۰۴۷ | ۴۸ | ٪۹۵.۶ |
| حسن | /hosn/ | /ha´san/ | ۱۳۲۰ | ۷۶ | ٪۹۵.۵ |
| دور | /dowr/ | /dur/ | ۲۸۵۹ | ۱۶۴ | ٪۹۴.۵ |
| اشراف | /?ash´raaf/ | /?esh´raf/ | ۹۲ | ۱۵ | ٪۸۳.۶ |
| تن | /ton/ | /tan/ | ۴۰۱۴ | ۳۴۰ | ٪۹۲.۱ |
| فرق | /fargh/ | /fe´ragh/ | ۳۶۶ | ۱۳ | ٪۹۶.۵ |
| عالم | /?aa´lam/ | /?aa`lem/ | ۱۳۸۵ | ۱۳۰ | ٪۹۱.۴ |
| فوت | /fowt/ | /fut/ | ۲۹۹ | ۲۳ | ٪۹۲.۸ |
| ده | /dah/ | /deh/ | ۱۴۸۸ | ۸۴ | ٪۹۴.۶ |
| سرور | /sar´var/ | /so´rur/ | ۱۰۴ | ۱۱ | ٪۹۰.۴ |

ستون اول از سمت راست هم‌نگاره‌ها را نشان می‌دهند. این هم‌نگاره‌ها با توجه به تقسیم‌بندی ارائه شده در فصل ۷ در دسته هم‌نگاره‌های غیر تکیه‌ای بسیط قرار می‌گیرند. ستون دوم تعداد موارد صحیح تشخیص داده شده توسط الگوریتم و ستون سوم موارد خطا را نشان می‌دهد. در ستون آخر دقت روش یعنی نسبت تعداد موارد صحیح به تعداد کل موارد ارائه شده است. دقت بالای روش برای موارد ذکر شده نشان از کارایی لیست‌های تصمیم‌گیری برای ابهام‌زدایی از هم‌نگاره‌های فارسی می‌باشد.

نکته قابل ذکر دیگر تعداد موارد رخداد این هم‌نگاره‌هاست. چون روش استفاده شده در ارزیابی روش اعتبارسنجی متقابل ۱۰ قسمتی است، بنابراین مجموع ستون دوم و ستون سوم در جدول ۶ به

عبارت دیگر مجموع تعداد موارد صحیح و غلط، برابر تعداد رخداد هر هم‌نگاره در کل پیکره مورد استفاده است. با محاسبه این مجموع و مقایسه آن با اندازه پیکره (یعنی ۷.۵ میلیون کلمه) می‌توان دریافت که حتی هم‌نگاره‌های با فراوانی بالا بخش کوچکی از پیکره را به خود اختصاص داده‌اند. بنابراین با به کار بردن پیکره‌های بزرگتر می‌توان امید آن داشت که کارایی این روش به مراتب بهبود یابد.

# نتیجه‌گیری

حاصل تمام مطالب ارائه شده در فصول قبل در قالب ارائه نتایج در این فصل عرضه گردید. پس از لحاظ کردن یک روش ارزیابی مشخص به بیان نتایج حاصل از اعمال سه برچسب گذار bigram، trigram و مبتنی بر حافظه پرداخته شد. نتایج این سه برچسب‌گذار در برچسب‌گذاری مقولات اصلی بر روی بخش یکسانی از دو نسخه پیکره متنی زبان فارسی ارائه و نتایج آن‌ها با هم مقایسه گردید. روش تجزیه تصریفی کلمات با کمک یک تحلیل‌گر ساختواژی برای پوشش دادن تعداد زیادی از برچسب‌های کلمات در پیکره بر روی همان بخش از پیکره در نسخه دو آزمایش شد. نتایج حاصل نشان از کارایی بسیار بالایی این روش برای برچسب‌گذاری در زبان فارسی می‌باشد. برای ابهام‌زدایی از هم‌نگاره‌های فارسی که کاری بسیار دشوار می‌باشد و نیازمند پیکره‌های بسیار بزرگ است خود را به ابهام‌زدایی از هم‌نگاره‌های با فراوانی بالا محدود کردیم. نتایج حاصل کارایی روش استفاده شده را در این حوزه نشان می‌دهد.

# نتیجه‌گیری و کارهای آتی

برچسب‌گذاری خودکار اجزای واژگانی کلام از مباحث بنیادی در حوزه پردازش زبان طبیعی است که در کاربردهای مختلف مورد استفاده قرار می‌گیرد. ما نیز در این پایان‌نامه به برچسب‌گذاری خودکار اجزا واژگانی کلام در پیکره متنی زبان فارسی پرداختیم. در این راستا فعالیت‌های زیادی انجام شد که در نهایت پس از آزمایشات فراوان در قالب یک طرح در فصل اول ارائه گردید. این طرح با توجه به تجربیات به‌دست آمده و تحلیل نتایج حاصل از آن آزمایشات، ارائه شد و به نظر می‌رسد که برای نیل به یک سیستم برچسب‌گذاری کامل نیاز است که بخش‌های مختلف آن در قالب چندین پروژه انجام پذیرد. ما در بخش‌های مختلفی از این فعالیت‌ها وارد شده و با استفاده از روش‌ها و مدل‌های گوناگون سعی کردیم که مشکلات هر بخش را تشخیص داده و در مواردی به حل آن‌ها بپردازیم.

بعد از ارائه این طرح، چون تکیه ما بر برچسب‌گذاری در پیکره متنی زبان فارسی بوده است در فصل ۲ مشخصاتی از این پیکره را بیان کردیم. بعد از آن در فصل ۳ برچسب‌گذارهای مارکوفی که از گسترده‌ترین برچسب‌گذارهای استفاده شده در زبان‌های مختلف هستند را بررسی کردیم. روش‌های مبتنی بر حافظه نیز برای کاربردهای یادگیری ماشین به طور گسترده استفاده می‌شوند. ما نیز برچسب‌گذاری مبتنی بر حافظه را به خاطر دارا بودن مزایایی مانند سرعت، حساسیت بالا نسبت به زمینه (به دلیل استفاده از روش نزدکترین k همسایه) و همچنین برای مقایسه نتایج آن با برچسب‌گذارهای مارکوفی bigram و trigram، در فصل ۴ شرح دادیم.

تجربه نشان داده است که مهمتر از روش‌های یادگیری ماشین مورد استفاده برای کاربردهای مختلف، روشی که برای پیش‌پردازش داده انجام می‌گیرد بر نتایج تاثیر می‌گذارد. ما در فصل ۵ مباحثی مخصوص زبان فارسی، ساختواژه کلمات فارسی و ویژگی‌های ساختواژی مقولات مهم کلمات در زبان فارسی را بیان کردیم و سپس یک روش پیش‌پردازشی بر مبنای تجزیه تصریفی کلمات ارائه دادیم. نتیجه این روش پیش‌پردازشی آن بود که تعداد بسیاری از برچسب‌های متمایز کلمات در پیکره متنی پوشش داده شد در حالی که نه تنها دقت برچسب‌گذاری از دقت برچسب‌گذاری در مقولات اصلی کاهش نیافت بلکه دقت برچسب‌گذاری افزایش نیز یافت.

در فصل ۶ به روش‌های غلبه بر کلمات ناشناخته پرداختیم و روشی را که در برچسب‌گذاری

کلمات ناشناخته در دو برچسب‌گذار ماکوفی استفاده کردیم شرح دادیم. غلبه بر کلمات ناشناخته در برچسب‌گذاری زبان فارسی بسیار مشکل می‌باشد و دلایل این مسئله در فصل ۸ به طور مبسوط بحث کردید که از آن جمله می‌توان به مواردی همچون کلمات در عبارات عربی، عدم وجود نشانه‌های بارز در کلمات فارسی که نشان‌گر برچسب کلمات باشد (مثلا در اسامی خاص) اشاره کرد.

در فصل ۷ به مبحث هم‌نگاره‌ها پرداختیم و مطالبی راجع به علل هم‌نگارگی در زبان فارسی، انواع هم‌نگاره‌ها و یک روش ابهام‌زدایی از هم‌نگاره‌ها بیان شد. در زبان فارسی کلمات بسیاری وجود دارند که ذاتا هم‌نگاره‌اند یا در اثر عمل اشتقاق یا تصریف تشکیل هم‌نگاره می‌دهند. ابهام‌زدایی از هم‌نگاره‌ها خصوصا در زبان فارسی به دلیل عدم وجود پیکره‌های بسیار بزرگ مشکل می‌باشد. در ابهام‌زدایی از هم‌نگاره‌ها مسائلی وجود دارد که عملا روش‌های آماری را به چالش می‌کشد. از آن جلمه می‌توان به فراوانی پایین بسیاری از هم‌نگاره‌ها، عدم فراوانی یکسان تلفظات مختلف یک هم‌نگاره، تشخیص هم‌نگارگی یک کلمه (این مورد در فارسی یک مشکل اساسی می‌باشد زیرا به دلیل وجود ساختواژه اشتقاقی و تصریفی پیچیده نمی‌توان لیست مشخصی از آن‌ها تهیه کرد)، نیازمندی به پیکره‌های بسیار بزرگ و هزینه‌بر بودن تهیه مجموعه‌های آموزشی اشاره کرد. در فصل ۷ به دلیل کوچکی پیکره در دسترس (بخش برچسب‌خورده پیکره زبان فارسی که شامل ۷.۵ میلیون کلمه می‌باشد) و هزینه‌بر بودن تهیه مجموعه آموزشی، خود را محدود به ابهام‌زدایی از هم‌نگاره‌های با فراوانی بالا کردیم.

در نهایت نتایج روش‌های بیان شده و پیشنهاد شده در برچسب‌گذاری کلمات پیکره فارسی و ابهام‌زدایی از هم‌نگاره‌های با فراوانی بالا در فصل ۸ ارائه شد و نتایج به‌دست آمده تحلیل گردید. نتایج به‌دست آمده و تحلیل‌های ارائه شده در موارد مختلفی مانند کلمات ناشناخته، استفاده از تحیل‌گر ساختواژی در برچسب‌گذاری، استفاده از برچسب‌گذارهای استفاده شده و ابهام‌زدایی انجام شده از هم‌نگاره‌ها، می‌تواند روشن‌گر ادامه راه برای غلبه بر موانع موجود در فرآیند نیل به یک برچسب‌گذار کامل در زبان فارسی می‌باشد. برای نیل به این هدف تلاش‌های بسیاری باید انجام گیرد که مواردی چند پیشنهاد می‌گردد:

- تشخیص کلمات، عبارات و جملات غیر رسمی و عامیانه و بررسی رفتار آن‌ها
- تشخیص عبارات و جملات عربی در بین متون فارسی
- تشخیص کران جملات و کلمات در متون فارسی
- طراحی یک تحلیل‌گر ساختواژی خودکار در حوزه اشتقاق، تصریف و ترکیب، با کارایی

بالا

- تشخیص هم‌نگاره بودن یک کلمه مخصوصا در هم‌نگاره‌های مرکب و ارائه روش‌هـایی ابهام‌زدایی از هم‌نگاره‌های فارسی
- بررسی دقیق کلمات ناشناخته در زبان فارسی و ارائه روش‌هایی برای غلبه بر آن‌ها
- تشخیص محل کسره اضافه در عبارات فارسی با دقت بالا

# مراجع و مآخذ

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.

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

مرادزاده، شهروز. (۱۳۸۳). *طبقه‌بندی هم‌نویسه‌های خط فارسی*. دبیرخانه شورای‌عالی اطلاع‌رسانی (کارگروه خط و زبان فارسی).

مشکوةالدینی، مهدی. (۱۳۸۴). *دستور زبان فارسی: واژگان و پیوندهای ساختی*. سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت).

Aha, D. W., Kibler, D. and Albert, M. (1991). *Instance-based learning algorithm. Machine Learning*. 7,pp.37-66

Assi, S. M. (1997). *Farsi Linguistic Database (FLDB)*. International Journal of Lexicography. Vol10,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).

Atkins, S., Clear, J. H and Ostler N. (1992). *Corpus Design Criteria*. Literary & Linguistic Computing,Vol7,No.1,pp1-16

Brants, T. (1999). *Tagging and Parsing with Cascaded Markov Models, Automation of Corpus Annotation*. Doctoral thesis in Universitat des Saarlandes.

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

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

Brill, E. (1994b). *Some advances in transformation-based part of speech tagging*. Proceeding of the Twlfth National Conference on Artificial Intelligence (AAAAI-94).

Brill, E. (1995). *Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging*. Computational Linguistic, 21(4):543-565

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

Cloern, J. (1999) . *Tagsets*. In Halteren, H. Syntactic Wordclass Tagging. London: Kluwer Academic Publishers,pp47- 54.

Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. (1992) . *A Practical Part-of-Speech Tagger*.Proc .3rd ANLP , Trento,Italy , pp133- 140

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,pp14- 27

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

Dematas, E. and Kokkinakis, G. (1995) . *Automatic stochastic tagging of natural language texts*.Computational Linguistics , 20(2 ):137- 164

Gale, W., Church, K. and Yarowsky, D. (1992) . *A Method for Disambiguating Word Senses in a Large Corpus*.Computer and the Humanities,26:415 -439

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 ,40:237 -264.

Jurafsky, D. and James, M,. (2000) . *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Hayslett, H. (1981) . *Frequency Analysis of English Usage: Lexicon and Grammar*. London: Heinemann.

Hearst, M. (1991) . *Noun Homograph Disambiguation Using Local Context in Large Text Corpora*. In the Proceeding of the 7th Annual Conference of the University of Waterloo Center for the New OED and Text Research, Oxford.

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 ):400- 401

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

Lee, S., Tsuji, J. and Rim, H. (2000) .*Part -Of- Speech Tagging Based on Hidden Markov Model Assuming Joint Independence*. In Proceedings of the 38th Annual Meeting of the ACL.

Leech, G. and Wilson, A. (1999) . *Standards for Tagsets*. In Halteren, H. Syntactic Wordclass Tagging. London: Kluwer Academic Publishers, pp. 55 - 80

Manning, C. and Schutze, H. (1999) . *Foundations of Statistical Natural Language Processing*. MIT Press.

Markov, A. (1913) . *An example of statistical investigation in the text of 'Eugene Onyegin' illustrating coupling of tests in chains*. In Proceedings of the Academy of Sciences,St .Petersburg , volume 7 of VI , pp. 153 - 162

Megerdoomian, K. (2000) . *Unification-Based Persian Morphology*. In Proceedings of CICLing 2000 Alexander Gelbukh (ed.). Centro de Investigacion en Computacion-IPN, Mexico.

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

Merialdo, B. (1990) . *Tagging Text with a Probabilistic Model*. In Proceedings of the IBM Natural Language ITL, Paris,France , pages 161 - 172

Mikheev, A. (1997) . *Automatic Rule Induction for Unknown-Word Guessing*. Computational Linguistics 26(3 ): 405 - 423

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

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

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.

Sejnowski, T. and Rosenberg, C. (1987) . *Parallel networks that learn to pronounce English text*Complex Systems 1, pp. 145 - 168

Sproat, R. (1992) .*Morphology and Computation* . Cambridge, MA: MIT Press.

Sproat, R., Hischberg, J. and Yarowsky, D. (1992) . *A Corpus-Based Synthesizer*. In Proceeding International Conference on Spoken Language Processing, Banff.

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. 175 - 182.

Weischedel, R., Meteer, M. and Schwartz, R. (1993). *Coping with Ambiguity and Unknown Words through Probabilistic Models*. Computational Linguistics, vol. 19, No.2, pp. 359 - 382.

Wilks, Y. and Stevenson, M. (1997). *The Grammar of Sense: Using part-of-speech tags as a first step in semantic disambiguation*. Natural Language Engineering 1 (1). Cambridge University Press.

Yarowsky, D. (1994). *Homograph Disambiguation in Speech Synthesis*. In Proceedings, 2nd ESCA/IEEE Workshop on Speech Synthesis.

# واژه‌نامه فارسی-انگلیسی

| | |
|---|---|
| اعتبارسنجی متقابل k قسمتی | k-fold cross validation |
| آموزش | Training |
| اشتقاق | Derivation |
| اطمینان | Confidence |
| افق محدود | Limited Horizon |
| ایستانی | Stationary |
| باناظر | Supervised |
| برچسب خاص | Unique tag |
| برچسب سلسله‌مراتبی | Hierarchical tag |
| برچسب متفرقه | Residual /Miscellaneous tag |
| برچسب نحوی-ساختواژی | Morphosyntactic tag |
| برچسب‌گذاری کاربردشناختی | Pragmatic tagging |
| برچسب‌گذار | Tagger |
| برچسب‌گذاری اجزای واژگانی کلام | Part-Of-Speech tagging |
| برچسب‌گذاری مبتنی بر تبدیل | Transformation-based tagger |
| بن‌واژه | Lemma |
| بهره اطلاعاتی | Information Gain |
| پراکندگی داده | Data sparsity |
| پیکره برچسب‌خورده | Tagged corpus |
| تحلیل‌گر ساختواژی | Morphological analyzer |
| ترکیب | Compounding |
| تصریف | Inflection |
| تکواژ آزاد | Free morpheme |
| تکواژ اشتقاقی | Derivational Morpheme |
| تکواژ تصریفی | Inflectional Morpheme |
| تکواژ مقید | Bound morpheme |
| توزیع باهم‌آیی | Collocational distribution |
| جنس | Genre |
| حد آستانه | Threshold |
| حد پایین اطمینان | Lower Confidence Limit |
| دسته | Bag |
| دقت | Accuracy |
| رده‌بند | Classifier |
| رمزگشایی | Decoding |
| روش افزایشی | Additive method |
| روش پنجره | Windowing approach |
| روش هموارسازی | Smoothing method |
| زمینه | Context |
| زنجیر مارکوف | Markov chain |

| | |
|---|---|
| ساختواژه | Morphology |
| ساختواژه اشتقاقی | Derivational Morphology |
| ساختواژه تصریفی | Inflectional Morphology |
| سیستم‌های ماکزیمم انتروپی | Maximum entropy systems |
| سیستم‌های مبتنی بر حافظه | Memory-based systems |
| شخص | Person |
| شمار | Number |
| فرآیند ماکوف | Markov process |
| کلمات ناشناخته | Unknown words |
| مبتنی بر قانون | Rule based |
| مجموعه آزمون | Test set |
| مجموعه آموزش | Training set |
| مدل مارکوف | Markov model |
| مدل مخفی مارکوف | Hidden Markov Model |
| مستقل از زمان بودن | Time invariant |
| معیار شباهت | Similarity measure |
| مقولات باز | Open category/wordclass |
| مقوله بسته | Closed category/wordclass |
| نرخ درست‌نمایی | Likelihood ratio |
| نزدیکترین k همسایه | k-nearest neighbors (knn) |
| نوع | Type |
| واحد | Token |
| واژگان | Lexicon |
| واژه‌بست | Clitic |
| هم‌نگاره | Homograph |
| یادگیری قیاسی | Inductive learning |

# واژه‌نامه انگلیسی- فارسی

| | |
|---|---|
| Accuracy | دقت |
| Additive method | روش افزایشی |
| Bag | دسته |
| Bound morpheme | تکواژ مقید |
| Classifier | رده‌بند |
| Clitic | واژه‌بست |
| Closed category/wordclass | مقوله بسته |
| Collocational distribution | توزیع باهم‌آیی |
| Compounding | ترکیب |
| Confidence | اطمینان |
| Context | زمینه |
| Data sparsity | پراکندگی داده |
| Decoding | رمزگشایی |
| Derivation | اشتقاق |
| Derivational Morpheme | تکواژ اشتقاقی |
| Derivational Morphology | ساختواژه اشتقاقی |
| Free morpheme | تکواژ آزاد |
| Genre | جنس |
| Hidden Markov Model | مدل مخفی مارکوف |
| Hierarchical tag | برچسب سلسله‌مراتبی |
| Homograph | هم‌نگاره |
| Inductive learning | یادگیری قیاسی |
| Inflection | تصریف |
| Inflectional Morpheme | تکواژ تصریفی |
| Inflectional Morphology | ساختواژه تصریفی |
| Information Gain | بهره اطلاعاتی |
| k-fold cross validation | اعتبارسنجی متقابل k قسمتی |
| k-nearest neighbors (knn) | نزدیکترین k همسایه |
| Lemma | بن‌واژه |
| Lexicon | واژگان |
| Likelihood ratio | نرخ درست‌نمایی |
| Limited Horizon | افق محدود |
| Lower Confidence Limit | حد پایین اطمینان |
| Markov chain | زنجیر مارکوف |
| Markov model | مدل مارکوف |
| Markov process | فرآیند ماکوف |
| Maximum entropy systems | سیستم‌های ماکزیمم انتروپی |

| | |
|---|---|
| Memory-based systems | سیستم‌های مبتنی بر حافظه |
| Morphological analyzer | تحلیل‌گر ساختواژی |
| Morphology | ساختواژه |
| Morphosyntactic tag | برچسب نحوی-ساختواژی |
| Number | شمار |
| Open category/wordclass | مقولات باز |
| Part-Of-Speech tagging | برچسب‌گذاری اجزای واژگانی کلام |
| Person | شخص |
| Pragmatic tagging | برچسب‌دهی کاربردشناختی |
| Residual /Miscellaneous tag | برچسب متفرقه |
| Rule based | مبتنی بر قانون |
| Similarity measure | معیار شباهت |
| Smoothing method | روش هموارسازی |
| Stationary | ایستانی |
| Supervised | با ناظر |
| Tagged corpus | پیکره برچسب‌خورده |
| Tagger | برچسب‌گذار |
| Test set | مجموعه آزمون |
| Threshold | حد آستانه |
| Time invariant | مستقل از زمان بودن |
| Token | واحد |
| Training | آموزش |
| Training set | مجموعه آموزش |
| Transformation-based tagger | برچسب‌گذاری مبتنی بر تبدیل |
| Type | نوع |
| Unique tag | برچسب خاص |
| Unknown words | کلمات ناشناخته |
| Windowing approach | روش پنجره |