دانشکده مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

ایجاد یک برچسب‌زن معنایی

با استفاده از فریم‌نت برای متون فارسی

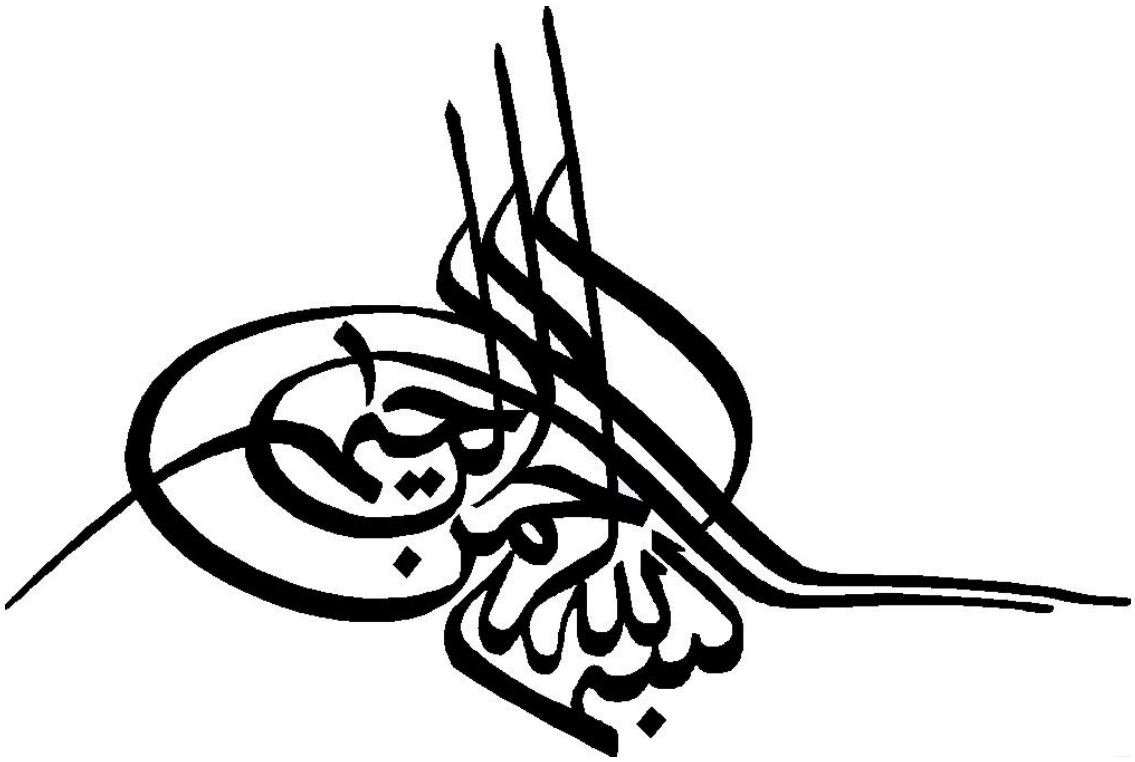
تهیه و تنظیم:

نجمه نوری

استاد راهنما:

دکتر محسن کاهانی

زمستان ۹۲


بسم الله الرحمن الرحيم

تقدیم

به گلهای زندگیم

پدر دلسوز و فداکارم که عشق ورزیدن را به من آموخت

مادر صبور و عطوفم که دعای خیرش همراهم بود

و همسر فداکار و مهربانم که همواره، صبورانه، دلسوزانه و عاشقانه یاریگرم بوده و هست.

و غنچه زندگی ام ، دختر کوچکم که با وجود کودکی اش به من انگیزه تلاش داد

سپاس

پروردگار مهربانم را سپاس‌گزارم به خاطر نعمت استادی گرانقدر و معلمی عالم ، جناب آقای دکتر محسن کاهانی، که در نگاهش جز شوق علم ، در کلامش جز دلسوزی و محبت و در کردارش جز انسانیت و شرافت ندیدم. پندهایش همواره زمزمه زبانم و رفتارش همواره الگوی زندگی علمی‌ام خواهد بود. از خداوند بزرگ تمنا دارم تا مجالی دهد ، گوشه‌ای از زحمات ایشان را جبران نمایم.

همچنین از جناب مهندس عسگریان به خاطر تمامی کمک ها و حمایت های بی دریغشان صمیمانه سپاسگزارم و توفیق روزافزون وی را از خداوند منان خواستارم.

در انتها نیز لازم می دانم از تمامی دانشجویان و اعضای آزمایشگاه فناوری وب دانشگاه فردوسی مشهد به خاطر کمک های و راهنمایی هایشان تشکر و قدردانی نمایم.

چکیده:

امروزه با رشد چشمگیر حجم مستندات منتشر شده در وب و نیاز اساسی به نگهداری، دسته‌بندی، بازیابی و پردازش ماشینی و سریع آنها، توجه به پردازش زبان طبیعی و بهره‌گیری از ابزارهایی نظیر خلاصه‌سازهای خودکار و مترجم‌های ماشینی، بیش از پیش خودنمایی می‌کند و فعالیت‌های زیادی برای طراحی چنین ابزارهایی در سرتاسر جهان انجام شده است. در زبان فارسی هم نظیر دیگر زبان‌ها تلاش‌هایی در این زمینه صورت گرفته است.

بدون تردید یکی از اهداف زبان‌شناسی رایانه‌ای که از ابتدای تکوین و گسترش این علم مد نظر بوده و پیوسته دنبال شده است، ارائه ابزارهایی برای تحلیل بر روی متون یا پیکره‌های زبانی می‌باشد. ورود رایانه به علم زبان‌شناسی، امکان وارد شدن به عرصه‌هایی را که تصور آنها نیز مشکل بود، فراهم کرده است.

مشکل اساسی در پردازش متون زبان بخصوص در مورد زبان شیرین فارسی، عدم وجود ابزارهای پایه‌ای استاندارد جهت انجام روال پیش‌پردازشی و تجزیه و تحلیل متون می‌باشد. در راستای انجام این پایان‌نامه با طراحی یک شبکه از قالب‌های جملات (FrameNet) ابزاری جهت برچسب‌زنی معنایی جملات زبان فارسی طراحی نمودیم که با دقت قابل قبولی به تطبیق جملات یک متن با فریم‌های موجود در پیکره تولید شده پرداخته و برچسب معنایی متناظر با آن جمله را به کلمات تشکیل دهنده‌ی جمله اختصاص می‌دهد. هدف این است که گستره وسیعی از ترکیبات معنایی و نحوی ممکن برای هر کلمه در هر کدام از معانی آن (ظرفیت‌ها) با استفاده از تفاسیر کامپیوتری از جملات نمونه و جدول‌بندی خودکار مستندسازی شود و نتایج تفاسیر نمایش داده شود.

برچسب‌زنی معنایی نقش کلمات یکی از مهم‌ترین ابزارهای پیش‌پردازشی برای اکثر کاربردهای پردازش متن مانند خلاصه‌سازی، ترجمه‌ی ماشینی، استخراج مفاهیم، سیستم‌های مکالمه‌ی معنایی، سیستم‌های پاسخگو، متن‌کاوی و ... می‌باشد. این ابزار به دنبال تعیین نقش معنایی هر عنصر جمله در ارتباط با فعل آن جمله می‌باشد. در حین انجام این عملیات نقش گرامری آن عنصر از قبیل فاعل بودن، مفعول بودن چه مستقیم و چه غیر مستقیم، انواع قیود و ... نیز مشخص می‌گردد.

در شبکه‌ی قالب‌های تولیدی برای افعال جملات زبان فارسی در مجموع ۴۵۹۳ ساختار برای افعال وجود دارد که در هر ساختار، چندین فریم وجود دارد که به کلمات آن، برچسب‌های معنایی تخصیص یافته‌اند.

کلمات کلیدی:

پردازش زبان طبیعی، زبان فارسی، شبکه قالب‌ها، برچسب‌زن معنایی، SRL، FrameNet.

فصل اول:

مقدمه

۱– مقدمه

۱–۱– انگیزه

امروزه با رشد چشمگیر حجم مستندات منتشر شده در وب و نیاز اساسی به نگهداری، دسته‌بندی، بازیابی و پردازش ماشینی و سریع آنها، توجه به پردازش زبان طبیعی و بهره‌گیری از ابزارهایی نظیر خلاصه‌سازهای خودکار و مترجم‌های ماشینی، بیش از پیش خودنمایی می‌کند. زبان‌شناسی رایانه‌ای در سال‌های اخیر به یکی از دغدغه‌های اساسی محققان و پژوهشگران حوزه کامپیوتر و زبان‌شناسی تبدیل شده است. استفاده از رایانه و ابزارهای هوشمند باعث شده است که بتوان بسیاری از کارهای مرتبط با پردازش متن را با سرعت و دقت قابل توجهی انجام داد.

مشکل اساسی در پردازش متون زبان طبیعی با بهره‌گیری از رایانه که زبانشناسی محاسباتی نامیده می‌شود، بخصوص در مورد زبان شیرین فارسی، عدم وجود ابزارهای پایه‌ای استاندارد جهت انجام روال پیش‌پردازشی و تجزیه و تحلیل متون می‌باشد.

یکی از ابزارهای مهم و اساسی در پردازش زبان طبیعی که کاربردهای بسیاری در الگوریتم‌های این حوزه دارد، ابزار برچسب‌زنی معنایی متون یا SRL می‌باشد که متاسفانه چنین ابزاری برای زبان فارسی موجود نمی‌باشد. به همین دلیل در راستای انجام این پایان‌نامه بر آن شدیم تا با طراحی شبکه‌ای از قالب‌های جملات زبان فارسی، ابزار برچسب‌زنی معنایی متون فارسی را طراحی و پیاده‌سازی نماییم.

برچسب‌زنی معنایی کلمات[۱] مشابه برچسب گذاری بخش‌های سخن بوده با این تفاوت که وارد عمق بیشتری از معنا می‌شود. نقش‌های معنایی رابطه اجزاء مختلف جمله را در ارتباط با فعل متناظرشان معرفی می‌نمایند. برچسب‌زنی معنایی وظیفه استخراج نقش‌های معنایی جملات نظیر فاعل، مفعول مستقیم، مفعول غیر مستقیم، فعل و ... را بر عهده دارد.

برای طراحی ابزار برچسب‌زنی معنایی کلمات ابتدا شبکه قالب[۲] جملات را برای زبان فارسی طراحی نموده و سپس از این شبکه قالب‌ها برای تولید ابزار SRL استفاده کردیم. فریم‌نت برای محققان پردازش زبان طبیعی، بیش از ۴۵۹۳ ساختار افعال جملات را که به صورت دستی حاشیه‌نویسی شده است یک مجموعه داده آموزشی منحصر به فرد برای برچسب‌گذاری نقش‌های معنایی فراهم کرده است، که در برنامه‌های کاربردی مانند استخراج اطلاعات، ترجمه ماشینی، تشخیص رخداد، تجزیه و تحلیل و غیره مورد استفاده واقع می‌شود.

---

[۱] Semantic Role Labeling

[۲] FrameNet

۲–۱– روش پیشنهادی

در این پایان‌نامه تلاش شده است تا در گام اول، یک شبکه از قالب‌های جملات (FrameNet) برای افعال زبان فارسی طراحی گردیده و در قالب یک پیکره، ذخیره‌سازی گردد. در راستای انجام این کار، از پیکره‌ی ظرفیت نحوی افعال زبان فارسی استفاده گردیده است. فرهنگ ظرفیت نحوی افعال فارسی مجموعه‌ای حاوی اطلاعات مربوط به ظرفیت نحوی بیش از ۴۵۰۰ فعل در زبان فارسی است. در این فرهنگ، متمم‌های اجباری و اختیاری انواع فعل‌های ساده، مرکب، پیشوندی و عبارات فعلی مشخص شده است. فراوانی فعل‌های مرکب در زبان فارسی، نیاز به فرهنگ ظرفیت فعل را در این زبان دوچندان می‌نماید. چرا که شناخت فعل‌های مرکب چه از لحاظ انسانی و چه از لحاظ پردازشی کاری دشوارتر از شناخت فعل‌های ساده است و به همین خاطر فراهم آوردن فهرستی از فعل‌های زبان (که شامل فعل‌های مرکب نیز می‌شود) به همراه ساخت‌های ظرفیتی افعال، کمکی شایان برای کارهای پردازشی است. از سوی دیگر، بر اساس نظریه وابستگی، ساخت بنیادین جمله را می‌توان از روی ساخت ظرفیتی فعل جمله به دست آورد و به همین دلیل بر اهمیت دانستن ساخت‌های ظرفیتی فعل در متن‌های زبانی افزوده می‌شود.

پس از طراحی و تکمیل شبکه‌ی قالب‌های جملات، ابزاری جهت برچسب‌زنی معنایی جملات زبان فارسی طراحی نمودیم که با دقت قابل قبولی به تطبیق جملات یک متن با فریم‌های موجود در پیکره تولید شده پرداخته و برچسب معنایی متناظر با آن جمله را به کلمات تشکیل دهنده‌ی جمله اختصاص می‌دهد.

در شبکه‌ی قالب‌های تولیدی برای افعال جملات زبان فارسی در مجموع ۴۵۹۳ ساختار برای افعال وجود دارد که در هر ساختار، چندین فریم وجود دارد که به کلمات آن، برچسب‌های معنایی تخصیص یافته‌اند.

در راستای طراحی ابزار برچسب‌زن معنایی زبان فارسی از ابزارهای پایه‌ای بسیاری استفاده گردید که می‌توان ابزار نرمال‌ساز یا یکسان‌ساز، ابزار تشخیص دهنده جملات، ابزار تشخیص دهنده لغات، ابزار ریشه‌یاب، ابزار برچسب‌زن اجزای واژگانی کلام، ابزار پاسر و غیره را نام برد. لازم به ذکر است اغلب این ابزارها توسط اعضای آزمایشگاه فناوری وب دانشگاه فردوسی مشهد تولید گردیده است.

- نرمال‌ساز (Normalizer): در ابتدا بایستی همه‌ی نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی گردند.
- جدا کننده جملات (Sentence Splitter): به کمک این پردازشگر می‌توان جملات متن را استخراج کرد.
- جداکننده کلمات (Tokenizer): به کمک این پردازشگر می‌توان کلمات متن را استخراج نمود.

- ریشه‌یاب (Stemmer): وظیفه ریشه‌یابی[۱] کلمات را برعهده دارد.
- برچسب‌زننده اجزای واژگانی کلام (POS): از این پردازشگر برای برچسب‌زنی اجزای واژگانی کلام استفاده می‌شود.

در روند هرگونه پردازش روی متن‌های زبان طبیعی انجام یک سری عملیات پیش‌پردازش، امری اجتناب‌ناپذیر است. دقت این پیش‌پردازش‌ها، تاثیر بسزایی در نتایج اعمال الگوریتم‌ها در فازهای بعدی دارد. هرچقدر که دقت پیش‌پردازش بیشتر باشد، الگوریتم‌ها به نتایج واقعی خود نزدیک‌تر خواهند شد.

در پردازش متون زبان فارسی با مشکلات بیشتری نسبت به سایر زبان‌ها مواجه خواهیم شد. عدم رعایت فاصله‌گذاری کلمات، فاصله و نیم‌فاصله، ضمایر متصل، ساختار گرامری خاص زبان فارسی و همچنین قرابت رسم‌الخط فارسی با زبان عربی از جمله مشکلات اساسی پردازش متون فارسی می‌باشد.

در روند طراحی ابزار برچسب‌زن معنایی، ابتدا جملات موجود در متن شناسایی گردیده و سپس برای هر کدام از جملات با شناسایی عبارت فعلی موجود در جمله، به شناسایی بن ماضی، بن مضارع، پیشوند، فعل‌یار و حرف اضافه فعلی می‌پردازیم. پس از شناسایی موارد ذکر شده برای هر فعل به شناسایی فریم منطبق با آن جمله پرداخته و سپس برچسب‌های معنایی موجود در فریم را به کلمات موجود در جمله تخصیص می‌دهیم.

۳-۱- ساختار پایان‌نامه

این نوشتار شامل چهار فصل دیگر به شرح زیر می‌باشد:

فصل دوم (مرور ادبیات): در این فصل در ابتدا مروری بر اطلاعات پایه و اولیه پردازش متن، فرهنگ ظرفیت نحوی افعال فارسی و همچنین چگونگی شناسایی افعال بخصوص افعال مرکب خواهیم داشت. سپس ساختار کلی فریم‌نت، کارهای انجام شده در راستای طراحی فریم‌نت در سایر زبان‌ها و همچنین فلسفه وجودی ابداع آن و در انتها ابزار برچسب‌زن معنایی متون تشریح می‌گردد.

فصل سوم (روش پیشنهادی): این فصل را می‌توان به سه بخش اصلی تقسیم نمود. در بخش اول، چگونگی طراحی و تولید فریم‌نت و در بخش دوم روش پیشنهادی برای برچسب‌زنی معنایی متون فارسی تشریح خواهد شد. در بخش سوم نیز روال پیاده‌سازی فریم‌نت و ابزار برچسب‌زنی معنایی یا SRL ذکر خواهد گردید.

فصل چهارم (نتیجه‌گیری): در این فصل به نتیجه‌گیری ایده‌ی پیشنهادی در این پایان‌نامه پرداخته شده و پیشنهادهایی برای کارهای آتی ارائه شده است.

۱- Steaming

فصل دوم:

مرور ادبیات

۲- مرور ادبیات

در این فصل به مرور مفاهیم مرتبط با پایان‌نامه، پیش‌نیازهای لازم برای درک مفاهیم مطرح شده و کارهای انجام شده در ارتباط با موضوع پایان‌نامه می‌پردازیم. ابتدا تعاریف پایه در پردازش زبان طبیعی ذکر گردیده است. در ادامه اشاره‌ای به فرهنگ ظرفیت نحوی افعال فارسی که در راستای تولید فریم‌نت زبان فارسی مورد استفاده قرار گرفت، می‌کنیم و سپس مروری بر مفاهیم مرتبط با شبکه قالب جملات و کارهای انجام شده در این حوزه برای سایر زبان‌ها خواهیم داشت.

۲-۱- تعاریف پایه زبان‌شناسی

پردازش متن از جمله مسائل اساسی در حوزه هوش مصنوعی و شناخت رایانشی است که در چند دهه اخیر، توجهات گسترده‌ای را در قالب‌های عدیده به خود معطوف کرده است. در پردازش متون زبان طبیعی با زبان نوشتاری سر و کار داریم. این مسأله باعث می‌شود گرچه به جهت از دست دادن اطلاعات گویشی مانند لحن گوینده، آهنگ صدا، تاکید و مکث، با مشکلات و ابهاماتی مواجه شویم، ولی در مقابل با شکل محدودتر و با قالب دستوری مشخص‌تری از زبان کار می‌کنیم. پردازش متون زبان فارسی در سطوح چهارگانه‌ی آوایی، ساخت‌واژی، نحو و معنایی و همچنین در حوزه‌های کاربردی متعددی امکان‌پذیر می‌باشد.

قبل از پرداختن به هر مطلبی در زمینه پردازش زبان طبیعی، برای آشنایی بهتر با مباحث مربوط به پردازش زبان طبیعی، بهتر است با مفاهیم پایه و تعاریف اولیه‌ی این حوزه که به منزله الفبای پردازش متن می‌باشند، آشنا شویم. اغلب اقدامات مربوط به این مفاهیم، در واقع نوعی پیش‌پردازش متن می‌باشد؛ بدین معنی که انجام این پردازش‌ها بر روی متن، در واقع آماده‌سازی متن به منظور اعمال فرآیندها و فعالیت‌های بعدی می‌باشد. در ادامه‌ی این بخش، تعاریف پایه و ابتدایی مورد نیاز، توضیح داده شده است. در کاربردهای مختلف پردازش زبان طبیعی عموما از این تعاریف پایه استفاده می‌شود.

۲-۱-۱- زبان فارسی

در تلاش برای ساخت یک سیستم پردازش و درک متون فارسی با مسائل و مشکلاتی مواجه می‌شویم که بعضی از آنها در بیشتر زبان‌ها بروز کرده و برخی خاص زبان فارسی می‌باشند.

همچنین برخی از این پیچیدگی‌ها به طبیعت زبان و نارسایی‌های دستورات زبان‌شناسی مربوط و برخی دیگر برخاسته از مشکلات ایجاد سیستم‌های هوش مصنوعی است [داد ۸۰]. در این بخش به برخی از این مسائل اشاره می‌شود.

زبان فارسی از نظر ساختاری دارای تفاوت‌های بسیاری با زبان انگلیسی است. برخی از تفاوت‌های مشهود بین

زبان فارسی و انگلیسی عبارتند از:

- تفاوت در ترتیب قرارگیری ارکان جمله. در اصطلاح، زبان‌هایی مثل انگلیسی را SVO[1] و زبان‌هایی مثل فارسی را SOV[2] می‌نامند که در واقع نشان‌دهنده‌ی ترتیب ارکان در جملات می‌باشد.
- زبان فارسی یک زبان اصطلاحاً بازتابی[3] نامیده می‌شود. یعنی کلمات براساس زمان و شخص موجود در جمله، می‌توانند حالت‌های مختلفی به خود بگیرند.
- در فارسی برخی ضمیرها وجود دارند که به اسم‌ها و افعال متصل می‌شوند (ضمیرهای متصل) که باعث بروز شکل‌های مختلف برای کلمات می‌شوند که این حالت هم در زبان انگلیسی وجود ندارد و تمامی‌ضمیرها منفصل می‌باشند.

با توجه به موارد ذکر شده و از آنجایی که زبان فارسی نوعی از زبان‌های غیرساختیافته است با مشکلات بسیار بیشتری نسبت به زبان انگلیسی مواجه خواهیم شد. متون غیرساختیافته، متونی هستند که پیش‌فرض خاصی در مورد قالب آنها نداریم و آنها را به صورت مجموعه‌ای مرتب از جملات و کلمات در نظر می‌گیریم.

به طور کلی مشکلات اصلی در پردازش متون فارسی را می‌توان در چند دسته زیر، خلاصه نمود [داد ۸۰]:

- عدم وجود منابع زبانی مناسب و کافی برای زبان فارسی.
- مشکل تشخیص مرز کلمات (مسأله شیوه‌های نگارش متفاوت)
- مشکل تشخیص مرز گروه‌های اسمی (مسأله‌ی کسره‌ی اضافه نامرئی)
- از دست دادن اطلاعات گویشی
- مسأله‌ی ابهام
- افعال مرکب و اصطلاحات
- مسأله‌ی همنگاره‌ها و تحت آن مسأله‌ی حذف مصوت‌های کوتاه (اعراب) از نوشتار
- معناشناسی و مشکلات تحلیل معنایی.

۲-۱-۲- ایست واژه‌ها (Stop words)

ایست واژه‌ها لغاتی هستند که علی‌رغم تکرار فروان در متن، از لحاظ معنایی دارای اهمیت کمی هستند

---

[1] - Subject Verb Object
[2] - Subject Object Verb
[3] - reflective

مثل "اگر"، "و"، "ولی"،"که" و غیره. در نگاه اولیه کلمات ربط و تعریف، ایستواژه به نظر می‌آیند؛ در عین حال بسیاری از افعال، افعال کمکی، اسم‌ها، قیدها و صفات نیز ایستواژه شناخته شده‌اند. در اغلب کاربردهای متن، حذف این کلمات، نتایج پردازش را به شدت بهبود می‌دهد و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل این کلمات غالبا در فاز پیش‌پردازش، حذف می‌شوند. برای زبان فارسی چندین لیست از این کلمات منتشر شده است که بطور میانگین شامل ۵۰۰ کلمه می‌باشند.

۳-۱-۲- ریشه‌یابی

در این مرحله به منظور یکسان‌سازی اشکال مختلف یک کلمه، یکپارچه‌سازی و همچنین اعمال پردازش‌های بعدی بایستی کلمات، ریشه‌یابی شوند. ریشه‌یابی به فرآیند تبدیل کلمات به فرم ریشه‌ای و پایه‌ای آنها اشاره می‌نماید. بنابراین "*دانش‌آموز*" و "*دانشجو*"و "*دانشگاه*" به"*دان*" که ریشه‌ی اصلی است، کاهش می‌یابند. لازم به ذکر است که منظور از ریشه در این بخش، دقیقا ریشه‌ی کلمات که در زبان‌شناسی استفاده می‌شود، نیست. بلکه منظور از ریشه، یک نماینده برای کلماتی است که از لحاظ معنایی و نحوی در یک حوزه قرار می‌گیرند. این فرآیند در پردازش متن، اهمیت بسیاری دارد؛ چرا که باعث می‌شود ماشین با دو کلمه‌ی هم-خانواده اما ظاهرا متفاوت، مانند دو کلمه‌ای که از لحاظ ریشه‌ای هیچ ارتباطی با هم ندارند، برخورد ننماید. الگوریتم‌های مختلفی برای ریشه‌یابی لغات پیشنهاد شده است و مورد استفاده قرار می‌گیرد. الگوریتم پیشنهاد شده در [POR80] رایج‌ترین الگوریتم در زبان انگلیسی می‌باشد. نمونه‌های دیگری از الگوریتم‌های ریشه‌یابی، الگوریتم *کراوتز*[۱] در انگلیسی و الگوریتم کاظم تقوا در فارسی هستند [TAG05][KRO93][POR80]. اما از آنجا که خروجی ریشه‌یاب در فازهای بعد، مورد استفاده‌های گوناگون از جمله اندازه‌گیری شباهت معنایی بر مبنای شبکه واژگان قرار می‌گیرد، بایستی بررسی شود تا خروجی ریشه‌یاب، ورودی مناسبی برای آن فازها باشد. در ادامه روش‌های طراحی و پیاده‌سازی ریشه‌یاب تشریح می‌گردد.

۴-۱-۲- ریشه‌یابی مبتنی بر قاعده

در این روش، با توجه به ساختار و طول هر کلمه و همچنین قواعد و ویژگی‌های خاص زبان فارسی برای ساخت کلمات مشتق، مرکب، صفات، قیدها، افعال و ... ریشه‌یابی صورت می‌گیرد. وندهای مشخصی در زبان فارسی برای ساخت انواع اسم، صفت و قید بکار می‌روند. در طراحی ریشه‌یاب مبتنی بر قاعده، با توجه به طول کلمه قبل از ریشه‌یابی و طول کلمه بعد از ریشه‌یابی، از قواعد ساخت کلمات زبان فارسی برای تشخیص ریشه، استفاده شده است.

بسیاری از پیشوندها و پسوندها در روند ریشه‌یابی ، مورد استفاده و بررسی قرار گرفته است. در روش

۱- Krovetz

پیشنهادی در مورد میانوندها، عملیات خاصی صورت نمی‌گیرد. لازم به ذکر است اگر چه ریشه‌یاب مبتنی بر قاعده‌ی طراحی شده دارای کارکرد نسبتا خوبی است؛ ولی استفاده در این پایان‌نامه از روش ترکیبی که در ادامه ذکر می‌گردد، استفاده شده است. در بسیاری از الگوریتم‌های پردازش زبان طبیعی از آنجاییکه خروجی ریشه‌یاب در بسیاری از کاربردها در فازهای بعدی برای مقایسه یا اندازه‌گیری شباهت معنایی بر مبنای شبکه واژگان مورد استفاده قرار می‌گیرد، بایستی ریشه‌یابی طراحی و مورد استفاده قرار گیرد که خروجی آن ورودی مناسبی برای آن فازها باشد.

۵-۱-۲- ریشه‌یابی مبتنی بر قواعد و فرهنگ لغت

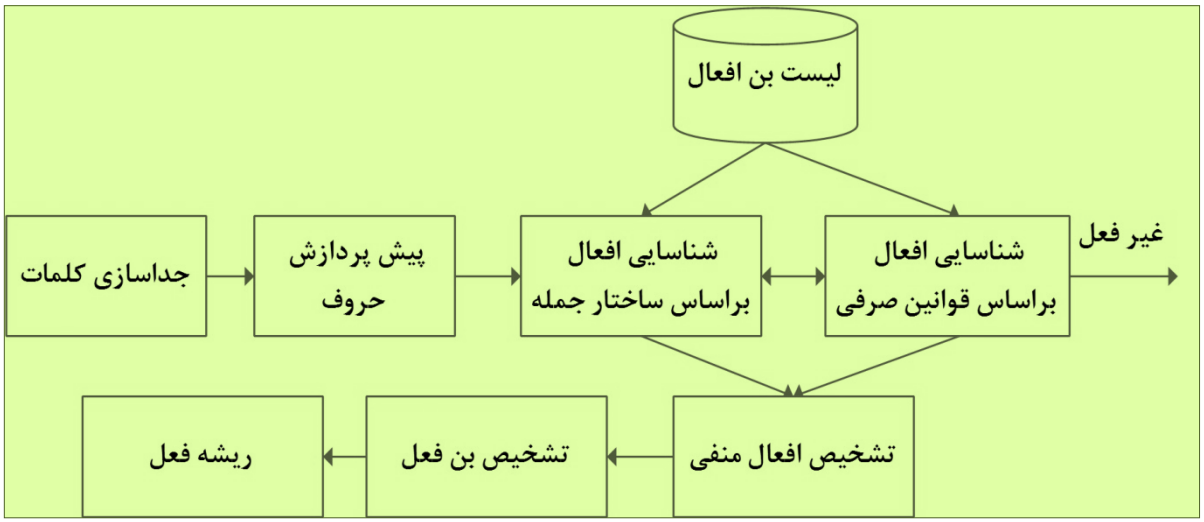
الگوریتم ریشه‌یابی مبتنی بر قواعد و فرهنگ لغت، در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد طراحی شده است. در الگوریتم طراحی شده برای ریشه‌یابی کلمات فارسی، از هر دو رویکرد استفاده از فرهنگ لغت و قوانین نحوی بهره گرفته شده است. برای اجرای مراحل الگوریتم از پنج فرهنگ لغت استفاده شده است.

۱. فرهنگ لغت برای نگه داشتن تمامی ریشه‌های مربوط به کلمات غیر فعلی (شامل اسامی، صفت‌ها، قیدها و ...).
۲. فرهنگ لغت برای نگهداری کلمات جمع مکسر و حالت‌های جمع بدون قاعده به همراه حالت مفرد این کلمات
۳. فرهنگ لغت شامل بن‌های گذشته و حال افعال زبان فارسی
۴. فرهنگ لغت شامل تمامی وند (پیشوند و پسوند)های زبان فارسی
۵. فرهنگ لغت برای کلمات استثناء

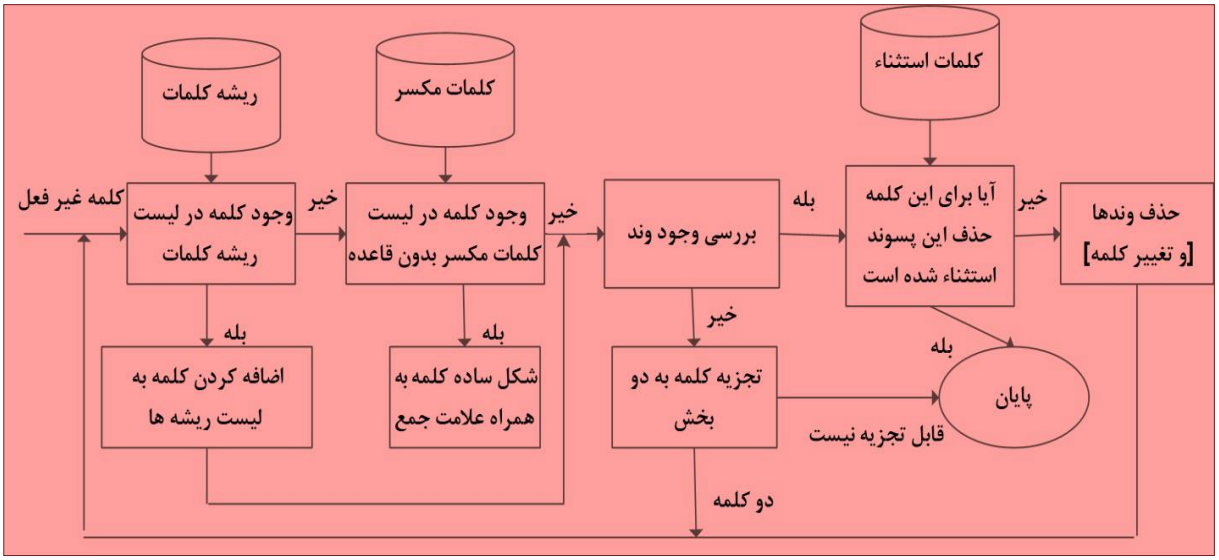
در الگوریتم ریشه‌یابی، ابتدا بایستی نوع عبارت (فعل یا غیرفعل)، تشخیص داده شود. برای انجام این عمل از دو روش استفاده شده است. در روش اول شناسایی صحیح انواع فعل پیشوندی، ساده و مرکب در زمان‌های مختلف در متن با بهره‌گیری از قواعد نحوی گرامر زبان فارسی، موقعیت لغات و تحلیل ریخت‌شناسی لغات (مطالعه ساختار لغات) صورت پذیرفته است. برای شناسایی افعال موجود در متن، احتمال ظهور فعل در موقعیت‌هایی خاص نظیر لغات قبل از جداکننده‌های جملات مثل علامت‌های "."، "؛"، "!"، "؟"، "?"، ":" و غیره، جداکننده‌های لغات مثل علامت‌های "،"، ","، "-" و غیره. همچنین یک سری حروف خاص آغازکننده جملات مثل "که"، "*اساسا*"، "*البته*"، "*تا*"، "*اما*"، "*اگر*"، "*ولی*"، "*زیرا*"، "سپس"، "همچنین"، "*و*"، "*یا*" و غیره؛ از طریق چک کردن تطبیق با ساختار یکی از زمان‌های افعال فارسی، مورد بررسی قرار می‌گیرد. احتمال فعل بودن لغات در موقعیت‌های ذکر شده با بررسی مطابقت با انواع ساختار افعال فارسی توسط عبارات منظم بررسی می‌گردد. در صورتی که لغات مورد ارزیابی با هر کدام از ساختارهای فعل توسط عبارات منظم، تطبیق پیدا کنند، گروه فعلی را تشکیل خواهند داد. سایر عباراتی که در این قالب نمی‌گنجند نیز به صورت غیرفعل

شناسایی می‌گردند.

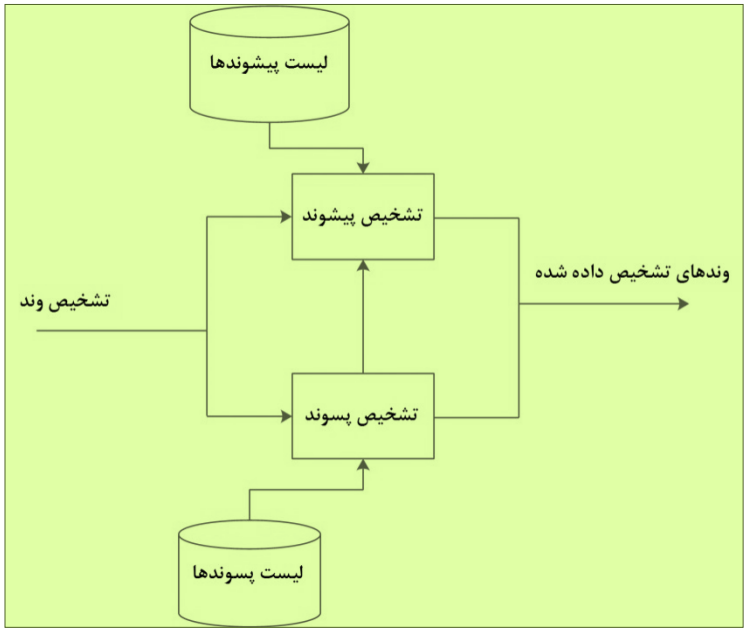
در روش دوم ابتدا از روی افعال استخراج شده از مجموعه دادگان، تمامی تصریف‌های مختلف افعال درون یک فرهنگ لغت ایجاد می‌شوند و سپس با جستجوی کلمه در این مجموعه، نوع عبارت تشخیص داده می‌شود. سپس در مرحله بعد برای هر کدام از عبارات فعل و غیرفعل به صورت جداگانه و براساس قوانین جداگانه، عمل می‌شود. در بیان روند ریشه‌یابی عبارات فعل و غیرغعل به چند شکل زیر (شکل ۳- ۲، شکل ۳- ۳ و شکل ۳- ۴) بسنده می‌شود.



شکل ۲-۱- مراحل ریشه‌یابی افعال



شکل ۲-۲- مراحل ریشه‌یابی کلمات غیرفعلی



شکل ۲-۳- مرحله تشخیص وندها

۲-۱-۶- Named Entity Recognition

ابزاری برای تشخیص اسامی و نوع آنها اعم از اسامی افراد، اماکن، مقادیر عددی و ... . برای تشخیص اینکه یک کلمه اسم است، راه‌های مختلفی وجود دارد که از جمله ی آنها مراجعه به لغت‌نامه، مراجعه به شبکه واژگان یا word-net، در نظر گرفتن ریشه‌ی کلمه، استفاده از قواعد نحوی ساخت‌واژه و ... می‌باشد. در این ابزار پس از تشخیص اسم‌ها با استفاده یک لغت‌نامه از اسامی افراد، مکان‌ها، مقادیر عددی و ... نوع اسم تشخیص داده می‌شود. به نظر می‌رسد که این لغت‌نامه در فارسی موجود نمی‌باشد.

از جمله نمونه‌های انگلیسی این ابزار می‌توان به Stanford NER و Illinois NER اشاره کرد.

۲-۱-۷- برچسب‌زنی بخش‌های سخن (POS)

در دستور زبان، بخش‌های سخن، طبقه‌بندی‌هایی زبانی از کلمات هستند که رفتار نحوی یک قسمت از جمله را بیان می‌دارند. به طور عموم، تمامی زبان‌ها دو بخش سخن فعل و اسم را دارند. بقیه بخش‌های سخن در زبان‌های مختلف، متفاوت می‌باشند. از جمله مهم‌ترین بخش‌های سخن در زبان فارسی اسم، ضمیر، صفت، قید و حرف اضافه را می‌توان نام برد.

از نمونه های فارسی آن می توان به ابزار آزمایشگاه آقای دکتر بیجن خان، و ابزار آزمایشگاه فناوری وب دانشگاه فردوسی مشهد اشاره کرد. از نمونه های انگلیسی آن می توان به Illinois Part Of Speech Tagger و Stanford POS Tagger اشاره کرد.

در زبان‌شناسی پیکره‌ای[۱] برچسب‌زن اجزای کلام[۲](POST یا POS tagging)، که همچنین برچسب‌زن دستوری[۳] یا ابهام‌زدایی لغت-دسته[۴] نامیده می‌شود، فرآیند نشانه‌گذاری لغت در یک متن است، که این نشانه، بیانگر وجه آن جزء از کلام می‌باشد. تشخیص این امر، مبتنی بر تعریف و نوع کاربرد در متن، انجام می‌شود. برای مثال رابطه‌ای که یک لغت با دیگر لغات در یک عبارت، جمله و یا پاراگراف دارد مشخص می-شود. شکل ساده شده‌ی این موضوع، همان مشخص کردن نوع لغت از لحاظ *اسم*، *فعل*، *صفت* و *قید* می‌باشد که در مدارس به آن پرداخته می‌شود. در شکل ۲-۱ نمونه‌ای فرضی از یک مجموعه برچسب (Tagset) برای زبان انگلیسی [CCG1] و همچنین در شکل ۲-۲ نمونه‌ای از یک مجموعه برچسب (Tagset) برای زبان فارسی معرفی شده است.

- # - Pound sign
- $ - Dollar sign
- '' - Close double quote
- `` - Open double quote
- ' - Close single quote
- ` - Open single quote
- , - Comma
- . - Final punctuation
- : - Colon, semi-colon
- -LRB- - Left bracket
- -RRB- - Right bracket
- CC - Coordinating conjunction
- CD - Cardinal number
- DT - Determiner
- EX - Existential there
- FW - Foreign word
- IN - Preposition
- JJ - Adjective
- JJR - Comparative adjective
- JJS - Superlative adjective
- LS - List Item Marker
- MD - Modal
- NN - Singular noun
- NNS - Plural noun
- NNP - Proper singular noun
- NNPS - Proper plural noun
- PDT - Predeterminer
- POS - Possesive ending
- PRP - Personal pronoun
- PP$ - Possesive pronoun
- RB - Adverb
- RBR - Comparative adverb
- RBS - Superlative Adverb
- RP - Particle
- SYM - Symbol
- TO - to
- UH - Interjection
- VB - Verb, base form
- VBD - Verb, past tense
- VBG - Verb, gerund/present participle
- VBN - Verb, past participle
- VBP - Verb, non 3rd ps. sing. present
- VBZ - Verb, 3rd ps. sing. present
- WDT - wh-determiner
- WP - wh-pronoun
- WP$ - Possesive wh-pronoun
- WRB - wh-adverb

شکل ۲-۴- نمونه‌ای فرضی از یک مجموعه برچسب برای زبان انگلیسی

۱- corpus linguistics

۲- part-of-speech tagging

۳- grammatical tagging

۴- word-category disambiguation

ADV	Adverb	قید
AJ	Adjective	صفت
CL	Classifier	شاخص
CONJ	Conjunction	حرف ربط
DET	Determiner	حرف تعریف
INT	Interjection	حرف صوت
N	Noun	اسم
NUM	Number	عدد
P	Preposition	حرف اضافه
POSTP	Postposition (را)	حرف اضافه پسین
PRO	Pronoun	ضمیر
PUNC	Punctuation	جدا کننده
RES	Residual:Arabic and Latin words, etc	متفرقه
V	Verb	فعل

شکل ۲-۵- نمونه‌ای فرضی از یک مجموعه برچسب برای زبان فارسی

برچسب‌زنی بخش‌های سخن[۱]عمل انتساب برچسب‌های نحوی به واژه‌ها و نشانه‌های تشکیل دهنده یک متن است به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جمله باشد. درصد بالایی از واژگان از نقطه نظر برچسب‌زنی نحوی دارای ابهام هستند زیرا کلمات در جایگاه‌های مختلف برچسب‌های نحوی متفاوتی دارند. بنابراین برچسب‌زنی نحوی، عمل ابهام زدایی از برچسب‌ها با توجه به زمینه (متن) مورد نظر است. برچسب‌زنی نحوی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می‌باشد.

مطالعات در حوزه زبانشناسی رایانه‌ای[۲] در جهتی است که فرآیند برچسب‌زنی را با بالاترین دقت و بطور ماشینی انجام دهد. بطور کلی الگوریتم‌های که تا کنون در این حوزه استفاده شده‌اند به دو گروه مبتنی بر قانون[۳] و آماری[۴] می‌باشند.

برچسب‌زنی اجزاء کلام، نسبتا کار مشکلی است، زیرا در خصوص بعضی از لغات، بیش از یک نقش برای یک لغت می‌توان نام برد که مشخص شدن آن تنها با در نظر گرفتن ویژگی‌های متنی که در آن واقع شده است، میسر است و این یک امر متداول در زبان‌های طبیعی است. در صد زیادی از اشکال لغت دارای ابهام می‌باشند. برای مثال حتی لغت "dogs" که معمولا تصور می‌شود که تنها یک اسم جمع است، می‌تواند بعنوان یک فعل نیز بکار رود.

Example: The sailor dogs the barmaid.

انجام برچسب‌زنی دستوری مشخص می‌کند که "dogs" یک فعل است و نه یک اسم جمع. انجام

۱- Part of Speech tagging

۲- computational linguistics

۳- rule-based

۴- stochastic

برچسب‌زنی POS شامل دو گام اصلی است: الف) پیدا کردن مجموعه نشانه‌های ممکن برای یک لغت با توجه به نقش آن در جمله. ب) انتخاب بهترین برچسب مبتنی بر محتوایی که در آن قرار گرفته است.

در زمینه برچسب‌زنی اجزای واژگانی کلام، روش‌های متعددی پیشنهاد شده است که می‌توان به مواردی همچون مدل مخفی مارکوف (HMM) (که روش‌های آماری هستند که ترتیب نشانه را با استفاده از احتمال Maximum Likelihood و ترتیب نشانه محاسبه می‌کنند) اشاره نمود.

راه‌حل دیگری که استفاده می‌شود، مبتنی بر قانون[۱]است که از قوانین و مخزن مخزن لغات[۲]برای مشخص کردن نشانه لغات استفاده می‌کند. این قوانین می‌توانند مبتنی بر دستور نگارش زبان نوشته شده باشند و یا با سامانه نشانه گذاری آموزش داده شده باشند و یا روش مبتنی بر تغییر شکل، هدایت شده با خطا باشند[BRI92]. دیگر مدل‌های یادگیری ماشین که برای نشانه گذاری بکار می‌روند شامل maximum entropy و دیگر مدل‌های log-linear، درخت‌های تصمیم[۳] یادگیری مبتنی بر حافظه[۴]و یادگیری مبتنی بر تغییر شکل[۵]می‌باشد.

برچسب‌زن بریل[۶][BRI92]، یکی از قدیمی‌ترین برچسب‌زن POS در زبان انگلیسی می‌باشد، که از الگوریتم-های مبتنی بر قانون استفاده می‌کند. تقریبا تمامی روش‌های ذکر شده در زبان فارسی نیز مورد استفاده قرار گرفته‌اند و نتایج به دست آمده دقتی بین ۹۵٪ تا حدود ۹۷٪ را نشان می‌دهند. از طرفی تقریبا در تمام موارد این دقت‌ها برای لغات یافت شده در پیکره بوده و دقت بدست آمده در شرایطی که در پیکره لغت یافت نشود حدود ۵۰ تا ۸۰٪ بر آورد شده است.

در ادامه به بررسی چند روش برچسب‌زنی POS مشهور به همراه میزان دقت هر کدام در زبان فارسی و انگلیسی پرداخته شده است.

۸-۱-۲- مروری بر کارهای انجام شده پیرامون برچسب‌زنی بخش‌های سخن

۱-۸-۱-۲- برچسب‌زن TnT (Trigrams'n'Tags)

برچسب‌زن TnT برنتس [BRA00] یک نشانه گذار POS آماری است، که قابلیت یادگیری زبان‌های متفاوت و تقریبا هر مجموعه نشانه‌ای را دارا است. این سامانه دارای روش‌های مختلفی برای برخورد با لغات

۱- rule-based

۲- lexicon

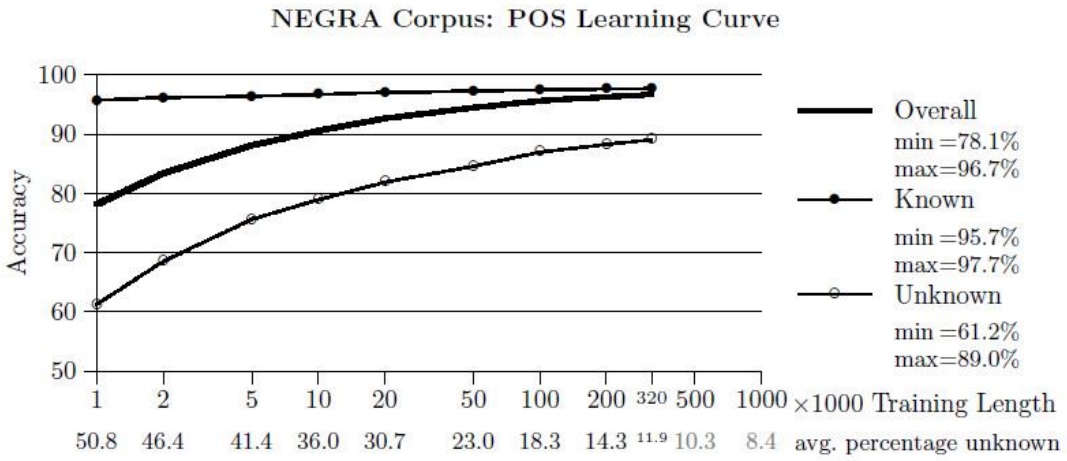
۳- decision trees

۴- Memory-based learning

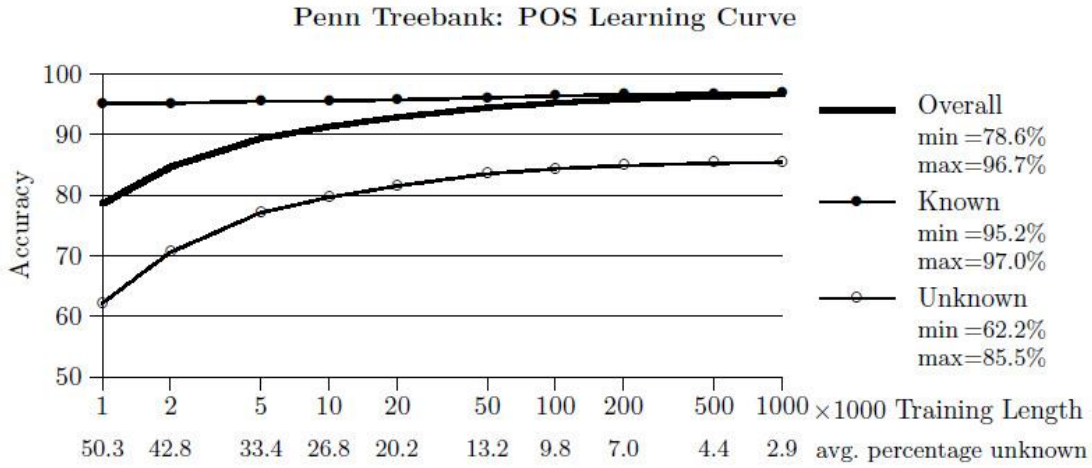
۵- transformation based learning

۶- E. Brill's tagger

ناشناس است. این برچسب‌زن، در واقع یک پیاده‌سازی از الگوریتم ویتربی برای مدل‌های ماکوف مرتبه دوم است. دقت این برچسب‌زن برای زبان‌های مختلف تقریبا بین ۹۶٪ تا ۹۷٪ است. لازم به ذکر است که دقت حاصل از این برچسب‌زن به شدت به پیکره مورد استفاده و همچنین متنی که برای ارزیابی به آن داده می‌شود بستگی دارد، بطوری که در ارزیابی انجام شده در [BRA00] بر روی پیکره NEGRA و Penn Treebank به ترتیب در شکل ۶- ۱ و شکل ۶- ۲نمایش داده شده است.



شکل ۲-۶- بررسی دقت **TnT** برروی پیکره **NEGRA**



شکل ۲-۷- بررسی دقت **TnT** برروی پیکره **Penn Treebank**

در ارزیابی صورت گرفته در [ORO07] نتیجه بررسی این برچسب‌زن بر روی زبان فارسی با استفاده از پیکره Bijankhan در جدول ۶- ۱ آمده است.

جدول ۲-۱- نتیجه ارزیابی **Tnt** برروی زبان فارسی

Run	Known words	Unknown words	Overall

1	96.94%	75.12%	96.52%
2	97.18%	80.09%	96.86%
3	96.96%	77.34%	96.57%
4	96.96%	77.69%	96.58%
5	97.03%	78.62%	96.67%
Avg.	97.01%	77.77%	96.64%

۲-۸-۱-۲- برچسب‌زن HunPoS

HunPoS [HAL07] یک پیاده‌سازی مجدد متن باز از TnT است که مبتنی بر مدل‌های مخفی مارکوف می‌باشد(HMM) که با مدل زبان trigram کار می‌کند و اجازه می‌دهد که کاربر با توجه به ویژگی‌های خاص زبان برچسب‌زنی را دقیق‌تر کند.

این برچسب‌زن مشابه TnT بوده و تنها در نحوه محاسبه احتمال emission/lexical متفاوت می‌باشد. در HunPoS برآورد احتمال emission/lexical مبتنی بر نشانه جاری و قبلی انجام می‌شود. تفاوت دیگری که آن را از TnT متمایز می‌کند متن باز بودن آن است و این در حالی است که TnT اینگونه نیست. این سامانه مشابه TnT دارای مکانیز تشخیص لغات ناشناس مبتنی بر پسوند است. همچنین HunPoS دارای مکانیزم تحلیل گر Morphological است [SER11].

نتایج ارزیابی این سامانه بر روی زبان انگلیسی [HAL07] با استفاده از داده‌های WSJ از Penn Treebank II در جدول زیرآمده است.

جدول ۲-۲- دقت برچسب‌زنی **WSJ** با احتمالات مرتبه اول و دوم **emission/lexicon**

	seen	unseen	Overall
TnT	96.77	85.91	96.46%
HunPoS 1	96.76	86.90%	96.49%
HunPoS 2	96.88	86.13%	**96.58%**

در مطالعه‌ای دیگر [SER11] که بررسی سامانه HunPoS برروی زبان فارسی است، نتایج ارزیابی به شرح جدول زیر گزارش شده است.

جدول ۲-۳- میزان دقت **HunPoS** بر روی زبان فارسی

Total Tokens	Tokens Correctly Tagged	Tokens Incorrectly Tagged	Accuracy
268008	259618	8390	96.90%

۳-۸-۱-۲- برچسب‌زنی مبتنی بر حافظه (Memory-Based POS Tagging)

برچسب‌زنی مبتنی بر حافظه POS [DAE96] با استفاده از ویژگی‌هایی نظیر نشانه‌های ممکن برای یک لغت و محتوای با عرض مشخص (نشانه‌های لغات قبلی که بدون ابهام هستند) کار می‌کند. این سامانه از روش یادگیری ماشین مبتنی بر حافظه استفاده می‌کند. یادگیری مبتنی بر حافظه با نام‌های دیگری همچون یادگیری Lazy، یادگیری Example-Based و یا یادگیری Case-Based شناخته می‌شود. یادگیری-های مبتنی بر حافظه معمولا درختی شبیه ساختار داده نمونه‌های یادگرفته شده ایجاد می‌کند و در حافظه نگه می‌دارد و وقتی یک نمونه اضافه می‌شود، با استفاده از معیارهای اندازی گیری شباهت، فاصله‌آن با نمونه-ها موجود بررسی شده و آن را طبقه‌بندی کرده و در محل مناسب خود در ساختمان داده تولید شده قرار می-دهند.[DAE96]. برای یادگیری مبتنی بر حافظه دو الگوریتم اصلی وجود دارد.

أ. Wieghted MBL: IB1-IG

ب. Optimized weighted MBL: IGTREE

روش IB1-IG، یک الگوریتم یادگیری مبتنی بر حافظه است که در طول یادگیری، پایگاه داده‌ای از نمونه‌ها تولید می‌کند. بعد از آنکه نمونه پایه تولید شد، طبقه‌بندی نمونه‌های جدید با انطباق آن‌ها با تمامی نمونه‌های موجود پایه و محاسبه فاصله نمونه جدید X با نمونه موجود در حافظه Y انجام می‌شود.

با توجه به اینکه جستجو برای یافتن نزدیک ترین همسایه‌ها در IB1-IG نیاز به صرف زمان دارد و این در حالی است که برچسب‌زن‌های POS بایستی سریع کار کنند از این رو IGTREE از جستجو در درخت-های تصمیم استفاده می‌کند. در IGTREE، حافظه نمونه دوباره طراحی شده با این هدف که شامل اطلاعات مشابه قبل باشد. میزان دقت گزارش شده [DAE96] این روش در زبان انگلیسی به همراه مقایسه الگوریتم‌های مختلف آن در جدول ۶- ۴ آمده است.

جدول ۲-۴- دقت برچسب‌زنی مبتنی بر حافظه در زبان انگلیسی

Algorithm	Accuracy
IB1	92.5%

IB1-IG	96.0%
IGTREE	96.4%

نتیجه ارزیابی این روش بر روی زبان فارسی [ORO07] به شرح جدول ۶- ۵ می‌باشد.

جدول ۲-۵- ارزیابی **IGTREE** برروی زبان فارسی

Run	Known words	Unknown words	Overall
1	96.43%	88.55%	96.27%
2	96.72%	91.80%	96.62%
3	96.98%	64.23%	96.30%
4	97.04%	66.18%	96.39%

۴-۸-۱-۲- برآورد حداکثر احتمال وقوع (Maximum Likelihood Estimation)

در این روش با توجه به پیکره برای هر لغت حداکثر احتمال وقوع برچسب آن محاسبه شده و به آن لغت مربوط می‌گردد. در ساده‌ترین مدل پیاده‌سازی، برچسبی که بیشترین تکرار را برای یک لغت داشته باشد، به عنوان نشانه منتخب بر گزیده می‌شود. در جدول ۶- ۶ میزان دقت حاصل از این روش بر روی زبان فارسی و استفاده از پیکره بی‌جن خان نمایش داده شده است که در خصوص لغاتی که در مخزن نبوده‌اند نشانه‌ای اختصاص نیافته است.

جدول ۲-۶- دقت **MLE** بر روی زبان فارسی

Run	Known words	Unknown words	Overall
1	96.50%	12%	94.55%
2	96.78%	16%	94.91%
3	96.53%	18%	94.53%
4	96.53%	9%	94.51%
5	96.64%	23%	94.68%
Avg.	96.60%	15%	94.63%

همانطور که درجدول ۶- ۶ مشخص است، میزان دقت در لغات ناشناس بسیار پایین است و این به دلیل عدم انتخاب نشانه برای آن لغت است. در ارزیابی دیگری برای لغات ناشناس از برچسب اسم مفرد (N_SING) استفاده شده است، به این دلیل که این نشانه بیشترین تعداد تکرار را در پیکره داشته است و در پیکره بی‌جن خان ۹۶۷۵۴۶ بار تکرار شده است. نتیجه این ارزیابی در ۶- ۷ ارائه شده است. همانطور که در جدول ۶- ۷

مشخص است این روش باعث بهبود چشمگیری در میزان دقت در لغات ناشناس می‌شود.

جدول ۲-۷- دقت **MLE** بر روی زبان فارسی با استفاده از **N_SING**

Run	Unknown words	Overall
1	52.60%	95.61%
2	56.63%	96.00%
3	51.49%	95.59%
4	55.48%	95.67%
5	54.34%	95.78%
Avg.	54.11%	95.73%

۹-۱-۲- تجزیه‌گر

به موازات پیشرفت و تحولات نظری در زبان‌شناسی جدید، روش‌های تحلیل متون و دستورات زبان بوسیله‌ی رایانه نیز تحول یافته است. منظور از گرامر هر زبان، در دست داشتن یک سری دستورات زبانی قابل فهم برای رایانه است که به کمک آنها بتوان اجزای نحوی یک جمله را به طور صحیح، تفکیک نمود. تجزیه و تحلیل جمله و شکستن آن به اجزای تشکیل دهنده مانند *گروه‌های اسمی*، *فعلی*، *قیدی* و غیره توسط ابزاری به نام تجزیه‌گر[۱] یا پارسر صورت می‌گیرد که نقش اساسی در طراحی و یا افزایش دقت سایر ابزارهای پردازش متن دارد.

تجزیه‌گرها با بهره‌گیری از دستورات گرامری زبان به تفکیک جملات متون به اجزای تشکیل‌دهنده‌ی آن، مشخص کردن نقش هر عبارت و لغت در متن و همچنین تشکیل درخت تجزیه برای جملات متن می‌پردازند.

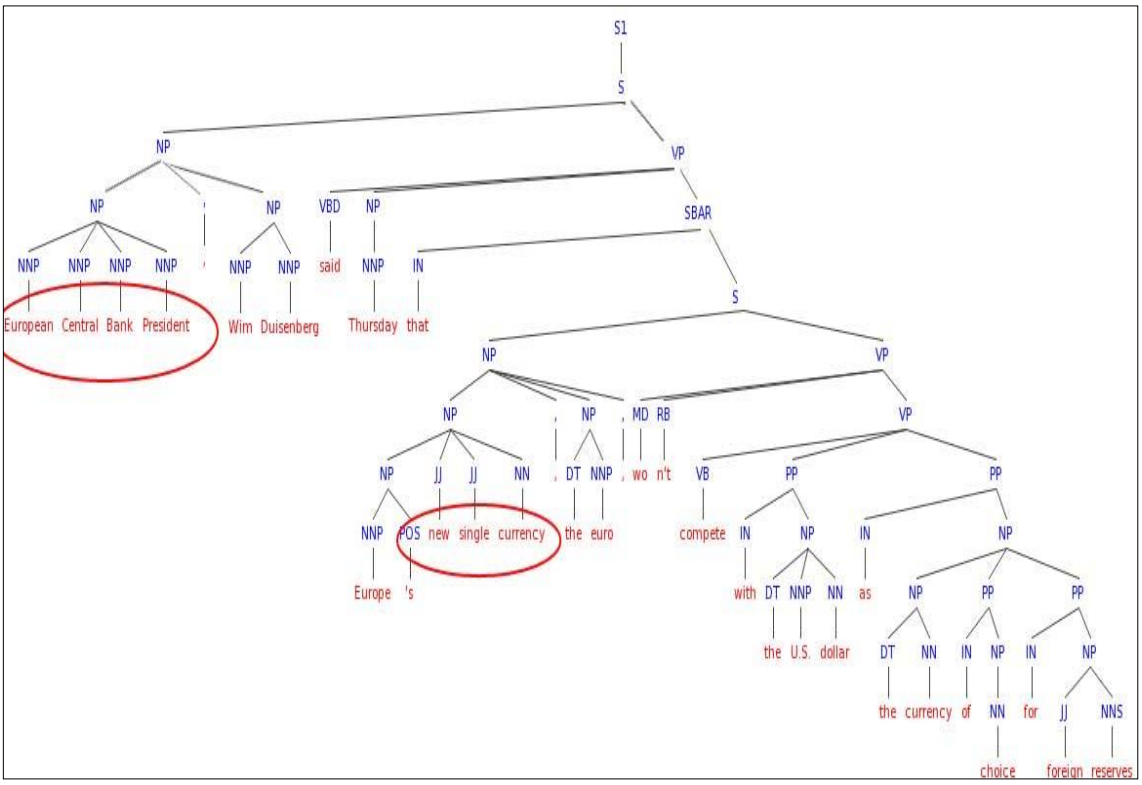
تجزیه‌گر نقش پایه‌ای و مهمی را در بهبود ابزارهای پردازش متن ایفا می‌کند. به عنوان مثال جهت تقویت الگوریتم‌های وابسته به برچسب‌زن معنایی لغات (SRL) علاوه بر نقش‌های کلمات، وابستگی‌های کلمات به لحاظ نقشی در جمله نیز باید مشخص گردد.

۱۰-۱-۲- درخت تجزیه

اجزای هر جمله را می‌توان در قالب گروه‌های اسمی، فعلی، حرف اضافه‌ای و ... تقسیم‌بندی نمود. گاه هر کدام از این گروه‌ها خود شامل زیرگروه دیگری می‌باشند. علاوه بر این، هر کدام نیز دارای روابطی می‌باشند، مثلا

۱- parser

یک گروه اسمی می‌تواند متعلق به یک گروه فعلی باشد. در نتیجه‌ی این تقسیم‌بندی‌های سلسله مراتبی، می‌توان یک ساختار درخت‌گونه از جمله داشت که درخت تجزیه نام دارد. درخت تجزیه، درختی است که ساختار نحوی یک جمله را بر اساس برخی روابط گرامری موجود در آن به شکلی ساده و قابل فهم برای کسانی که دانش عمیق زبان‌شناسی ندارند، نمایان می‌سازد [MAR08]. ابزارهای مختلفی برای تجزیه جمله توسعه یافته‌اند که خروجی اغلب آنها به صورت رشته‌ای شامل پرانتزهای تو در تو به همراه برچسب‌ها و کلمات می‌باشند. این مدل نمایش برای ورودی سیستم‌ها مناسب است، اما برای انسان خوانایی چندانی ندارد. در ابزار lfgParser[۱] شاهد نمایش گرافیکی و درخت‌گونه‌ی درخت تجزیه خواهیم بود. در شکل ۲- ۳ نمونه‌ای از یک درخت تجزیه، نشان داده شده است.



شکل ۲-۸- نمونه‌ای از یک درخت تجزیه جمله

۲-۲- فرهنگ ظرفیت نحوی افعال فارسی

فرهنگ ظرفیت نحوی افعال فارسی مجموعه‌ای است حاوی اطلاعات مربوط به ظرفیت نحوی بیش از ۴۵۰۰

[۱]- Http://lfg-demo.computing.edu.ie

فعل در زبان فارسی. در این فرهنگ، متمم‌های اجباری و اختیاری انواع فعل‌های ساده، مرکب، پیشوندی و عبارات فعلی مشخص شده است.

۱-۲-۲- معرفی فرهنگ ظرفیت نحوی افعال فارسی

فرهنگ‌های ظرفیت منابعی با ارزش برای پردازش زبان طبیعی تلقی می‌شوند. در این فرهنگ‌ها، اطلاعاتی در مورد متمم‌های اجباری و اختیاری واژه‌ها در سطح نحوی و یا معنایی برچسب می‌خورد.

فراوانی فعل‌های مرکب در زبان فارسی، نیاز به فرهنگ ظرفیت فعل را در این زبان دوچندان می‌نماید. چراکه شناخت فعل‌های مرکب چه از لحاظ انسانی و چه از لحاظ پردازشی کاری دشوارتر از شناخت فعل‌های ساده است و به همین خاطر فراهم آوردن فهرستی از فعل‌های زبان (که شامل فعل‌های مرکب نیز می‌شود) به همراه ساخت‌های ظرفیتی افعال، کمکی شایان برای کارهای پردازشی محسوب می‌شود. از سوی دیگر، بر اساس نظریه وابستگی، ساخت بنیادین جمله را می‌توان از روی ساخت ظرفیتی فعل جمله به دست آورد و به همین دلیل بر اهمیت دانستن ساخت‌های ظرفیتی فعل در متن‌های زبانی افزوده می‌شود.

مسائل مطرح‌شده گروه پژوهش زبان‌شناسی رایانه‌ای دکتر محمدصادق رسولی در دانشگاه علم و صنعت ایران را بر آن داشت که «فرهنگ ظرفیت نحوی افعال زبان فارسی» را که نخستین فرهنگ ظرفیتی زبان فارسی است، تولید و عرضه نمایند.

علاوه بر موارد ذکرشده، یکی از اهداف اصلی در ایجاد فرهنگ ظرفیت افعال زبان فارسی، استفاده از این فرهنگ به عنوان راهنمایی برای زبان‌شناسان در برچسب‌زنی جملات پیکره وابستگی نحوی زبان فارسی بود. با ایجاد ابزار جستجو در فرهنگ ظرفیت، این هدف در طی برچسب‌زنی پیکره‌ی وابستگی محقق شد.

۲-۲-۲- ویژگی‌های فرهنگ ظرفیت نحوی افعال فارسی

این فرهنگ، دارای اطلاعات زیر می‌باشد:

- دارای اطلاعات مربوط به انواع متمم‌ها: فاعل (+ فاعل تهی)، مفعول، مفعول دوم، مفعول حرف اضافه‌ای، مسند، تمیز، بند متممی فعل، متمم قیدی، مفعول نشانه اضافه‌ای.

- دارای اطلاعات جزئی در خصوص متمم‌ها؛ برای مثال، متمم قیدی فعل مورد نظر از چه نوعی است (حالت، مکان، مقدار)؟ مفعول حرف اضافه‌ای مورد نظر با چه حرف اضافه‌ای می‌آید (از، به، در ...)؟

- عرضه شده در دو قالب (فایل متنی/فایل اکسل) با قالب‌بندی بن ماضی، بن مضارع، پیشوند، فعل‌یار، حرف اضافه فعلی، ساخت ظرفیتی.

۳-۲-۲- افعال مرکب

«فعل مرکب به فعلی اطلاق می‌شود که از دو واژهٔ مستقل ترکیب یافته است؛ واژهٔ اول اسم یا صفت یـا قیـد و واژهٔ دوم فعل است. مانند «اجرا کردن» و «حدس زدن». جزء اول را فعل‌یار و جزء دوم را همکرد می‌نـامیم. از مجموع فعل‌یار و همکرد معنی واحدی دریافته می شود. هرگاه دو کلمه از این انواع ، دو معنی را به ذهـن القـا کنند، یعنی هر یک از اجزا، معنی مستقل و اصلی خود را حفظ کرده باشد، اطلاق فعل مرکب به آن‌ها درسـت نیست. وجه مشخص همکردها این است که معنی اصلی خـود را از دسـت مـی‌دهنـد یـا معنـی‌شـان کمرنـگ می‌شود و عمدتاً همچون عنصری صرفی که به ترکیب هویت فعلی می‌بخشد به کار مـی‌رونـد. بـرای مثـال در فعل‌های «اتو کردن» و «رنگ کردن»، جزء فعلی از نظر معنی تهی است و نقش اصلی آن این است که به کـل عبارت هویت فعلی می‌دهد» (طباطبایی ۱۳۸۴).

باید به این نکته اشاره کرد که « رابطهٔ میان اجزاء فعل‌یار و همکرد از نوع رابطهٔ نحـوی نیسـت، و هـیچ‌یـک از آن‌ها وابسته یا متمم دیگری محسوب نمی شود» (طبیـب‌زاده ۱۳۸۵). بنـابراین عبـارت‌هـای «غـذا خـورد» و «فیلم دید» فعل مرکب محسوب نمی‌شوند؛ زیرا در این ترکیب‌ها، «غذا» و «فیلم» مفعول‌هایی بدون نقش‌نمـا برای فعل‌های متعدی «خورد» و «دید» هستند. برای تشخیص فعل مرکـب در جملـه مـی‌تـوان از معیارهـای صوری زیر بهره جست:

۴-۲-۲- معیارهای صوری برای تشخیص فعل مرکب

۱) به « فعل‌یار» نمی‌توان نقش‌نمای «را» افزود.انـوری و گیـوی (انـوری و گیـوی ۱۳۸۵)معتقدنـد کـه بـرای تشخیص فعل مرکب از غیر آن، می‌توان به عنصر غیرفعلی (فعل‌یار) نقش‌نمای «را» داد، اگر جملـه بـا گـرفتن «را» صورت و معنای درست و رسایی پیدا کرد، آن کلمه مفعول است و فعل جمله سـاده؛ امـا اگـر بـا افـزودن «را» صورت ومعنای جمله نارسا باشد، آن کلمه فعل‌یار است. به طور مثال در جمله‌های زیر با افزودن «را» بـه عنصر غیرفعلی در جملهٔ اول، جمله خوش‌ساخت باقی می‌ماند. بنابراین «کتاب» مفعـول اسـت و «دادم» فعـل آن محسوب می‌شود؛ اما در جملهٔ دوم، پس از اضافه کردن «را» به عنصر غیرفعلی، جمله بدسـاخت مـی‌شـود. در نتیجه «سلام» فعل‌یار و فعل جمله «سلام دادم» است.

الف) به ساسان کتاب دادم.

ب) به ساسان سلام دادم.

۲) به «فعل‌یار» نمی‌توان پسوند صفت برترِ «تر» اضافه کرد. اگر بتوان در زنجیرهٔ صفت + کردن، پسـوند «تـر» را به صفت اضافه کرد و عبارت به دست آمده خوش‌ساخت باشد، در این صورت با ساخت بنیادین || فـا، مـف، مس|| روبرو خواهیم بود؛ به طور مثال از آنجا که می‌توان در جملهٔ «مینا اتاق را مرتب کرد»، پسوند «تر» را به صفت «مرتب» اضافه کرد، «مرتب» نه یک فعل‌یار بلکه مسـندی بـرای فعـلِ سـببیِ «کـرد» محسـوب شـده،

ساختِ || فا، مف، مس || به آن اختصاص می‌یابد. ذکر این نکته ضروری است که از آنجایی که در زنجیره‌هایی چون «مرتب کردن، زیبا کردن، خوشحال کردن، کلفت کردن»، می‌توان «کردن» را به «بودن» تبدیل کرد و با وجودِ این تبدیل، زنجیرهٔ به دست آمده همچنان خوش‌ساخت باقی می‌ماند، کلمات «مرتب، زیبا، خوشحال، وکلفت» در عبارت‌های فوق، نه یک فعل‌یار، بلکه مسندی برای فعل سببیِ «کرد» محسوب می‌شوند:

او اتاق را مرتب کرد.

اتاق مرتب است.

او مریم را خوشحال کرد.

مریم خوشحال است.

در حالیکه در مواردی که «کردن» نه یک فعل سببی، بلکه یک همکرد است، نمی‌توان آن را به «بودن» تبدیل کرد.

او دختر را آرایش کرد.

*دختر آرایش است/ بود.

او مادرش را آگاه کرد که میهمان‌ها دیر می‌آیند.

مادرش آگاه است/ بود که میهمان‌ها دیر می‌آیند.

۳) قبل از فعل‌یار، صفت‌های اشاره به‌کار نمی‌رود؛ البته ممکن است در موارد بسیار محدودی ضمیر اشاره به‌کار رود که همان‌طور که اشاره شد این‌گونه موارد بسیار محدود است. به طور مثال جمله‌های «*این تهدید کرد» و «*این گول خورد» بدساخت هستند. البته در جمله‌ای نظیر «او این تهدید را کرد»، نقش‌نمای «را» نه نشانهٔ مفعول بلکه علامت معرفگی است و باعث نمی‌شود که «تهدید» را مفعولِ فعلِ «کرد» به حساب آوریم؛ به عبارت دیگر «این» و «را» برای معرفه کردنِ «تهدید» به کار می‌روند.

۴) بلافاصله قبل از فعل‌یار، «یکِ نکره»، صفت‌های پرسشی «چندمین» و «کدام»، صفت‌های شمارشی اصلی (مانند: یک، دو و سه)، صفت‌های شمارشی ترتیبی با پسوند ُمین (مانند: دومین و سومین) صفت‌های عالی (مانند: بهترین، زیباترین و باهوش‌ترین) و شاخص (مانند: خواهر، دایی، سرگرد و آقا) به کار نمی‌رود زیرا حضور اکثر این وابسته‌های پیشین با اضافه شدن «را» بلافاصله بعد از اسم همراه خواهد بود و همان‌طور که در بالا اشاره شد به فعل‌یار نمی‌توان نقش‌نمای «را»ی مفعولی افزود.

۵) «ظرفیت نحوی هر فعل مرکب نه براساس فعل‌یار و نه براساس همکرد به تنهایی، بلکه براساس مجموع فعل‌یار و همکرد شکل می‌گیرد. مثلاً فعل مرکب «صدا زدن» یک مفعول مستقیم می‌گیرد، و فعل مرکب «حرف زدن» یک مفعول حرف اضافه‌ای بایی:

الف) او مرا صدا زد.

ب) او با من حرف زد.

اگر این دو فعل مرکب واحدهای واژگانی یکدستی نبودند، یعنی اگر ظرفیت نحوی آنها براساس همکردشان تعیین می‌شد باید هر دو، متمم‌های واحدی می‌گرفتند، حال آنکه چنین نیست» (طبیب‌زاده ۱۳۸۵).

۶) قیود می‌توانند کل فعل مرکب را به صورت یک واحد همبسته توصیف کنند اما نمی‌توانند بین دو سازهٔ افعال مرکب گسستگی ایجاد کنند. این وضعیت در مورد کلیهٔ افعال مرکب صادق است. البته باید توجه داشت که برای انجام این آزمون نباید بعد از فعل‌یار کسرهٔ اضافه بیاید. (علامت * نشان‌دهندهٔ بدساختی جمله است.)

الف) علی مینا را بی‌موقع تهدید کرد.

ب) * علی مینا را تهدید بی‌موقع کرد.

۷) فرایند مصدرساز:

زبان فارسی دارای نوعی فرایند مصدرساز است که فعل زماندار جمله را به مصدر تبدیل می‌کند و آن را هستهٔ ساخت اضافه قرار می‌دهد.

مثال: علی رفت..... رفتن علی

در مورد افعال مرکب جزء غیرفعلی و فعل مشترکاً هستهٔ ساخت اضافه را به وجود می‌آورند.

مثال: علی مینا را دلخور کرد..... دلخور کردن مینا توسط علی

۸)قلب نحوی

در زبان فارسی فرایندی وجود دارد که در نتیجهٔ عملکرد آن، توالی سازه‌های تشکیل‌دهندهٔ جمله با حفظ نشانه‌های دستوری آن سازه‌ها دگرگون می‌شود. چنین دگرگونی به دلایل کاربردشناختی و کلامی در زبان محاوره (و البته با گستردگی و انعطاف بیش‌تری در زبان ادبی و شعر) صورت می‌گیرد. در مثال‌های زیر جمله‌های ۵ و ۶ غیردستوری‌اند؛ زیرا در جملهٔ ۵ فقط همکرد و در جملهٔ ۶ فقط فعل‌یار قلب شده است.

۱) علی مینا را تهدید کرد.

۲) مینا را علی تهدید کرد.

۳) تهدید کرد علی مینا را.

۴) تهدید کرد مینا را علی.

۵) * کرد علی مینا را تهدید.

۶) *علی مینا را کرد تهدید.

سه معیار آخر از دبیر مقدم (دبیرمقدم ۱۳۷۴) اتخاذ شده است.

۹) عدم حضور علائم نگارشی بین دو جزء فعل مرکب:

نمی‌توان بین دو جزء فعل مرکب از علائم نگارشی استفاده کرد.

*علی مریم را صدا، زد.

*علی مریم را صدا؟ زد

*علی مریم را صدا. زد

*علی مریم را صدا؛ زد

بایستی خاطر نشان کرد که از آنجا که ما به دنبال ملاک‌های صوری برای تعیین فعل مرکب هستیم، ملاک‌های معنایی را که برخی از پژوهشگران برای تشخیص فعل مرکب ذکر کرده‌اند، در نظر نمی‌گیریم. به طور مثال طباطبائی (طباطبایی ۱۳۸۴)در مقالهٔ خود، به عنوان نخستین ملاک برای تشخیص فعل مرکب به این نکته اشاره می‌کند که فعل مرکب یک واحد معنایی است و جزء فعلی آن از محتوای معنایی خود تهی شده است و بخش اعظم معنی را جزء غیرفعلی حمل می‌کند. البته طباطبائی به عنوان ملاک صوری برای تشخیص فعل مرکب به این مسئله اشاره می‌کند که به « فعل‌یار» نمی‌توان نقش‌نمای «را» افزود. همچنین دبیرمقدم (دبیرمقدم ۱۳۷۴)معتقد است که فعل مرکب فعلی است که ساختمان واژی آن بسیط نیست، بلکه از پیوند یک سازهٔ غیرفعلی همچون اسم (تهدید کردن)، صفت (دلخور بودن)، اسم مفعول (کشته شدن)، گروه حرف اضافه‌ای (به دنیا آمدن)، یا قید (فرا گرفتن) با یک سازهٔ فعلی تشکیل شده است. به عقیدهٔ ایشان چندین فرایند ساخت‌واژی با درجات زایایی متفاوت در تشکیل فعل مرکب زبان فارسی دخیل‌اند که ایشان آن‌ها را تحت دو عنوان کلی ترکیب و انضمام طبقه‌بندی می‌کند. «در فرایند انضمامِ مفعول صریح، مفعول صریح نشانه‌های دستوری وابسته به خود را که می‌تواند شامل حرف نشانهٔ «را»، حرف تعریف نکرهٔ - ی، نشانهٔ جمع، ضمیر ملکی متصل و ضمیر اشاره باشد، از دست می‌دهد و به فعل منضم می‌شود. طبق نظر ایشان ماحصل فرایند انضمام به لحاظ ساختی فعل مرکب لازم است و از نظر معنایی یک کل واحد معنایی است. مثلاً در جملهٔ «بچه‌ها غذا خوردند»، طبق نظر ایشان « غذا خوردند» یک فعل مرکب انضمامی است» (دبیرمقدم ۱۳۷۴). در حالیکه به نظر نگارندهٔ این سطور، این نوع فعل‌ها فعل مرکب محسوب نمی‌شوند؛ زیرا افعال گذرا (متعدی) مانند «خوردن» لزوماً باید مفعول بگیرند و این مفعول ممکن است بر حسبِ ضرورت‌های بافتی، نشان‌دار یا بی‌نشان باشد. همچنین دبیرمقدم (دبیرمقدم ۱۳۷۴) برای تایید این فرضیه که افعالی که از طریق ترکیب و انضمام ساخته می‌شوند به لحاظ ساخت‌واژی مرکب‌اند، از برخی استدلال‌های واجی، نحوی و معنایی استفاده می‌کند. با توجه به اهداف مورد نظر ما، تنها ذکر شواهد نحوی ایشان کافی به نظر می‌رسید که در بالا به آنها اشاره شد.

۵-۲-۲- انواع فعل مرکب

فعل مرکب: ترکیب اسم + همکرد (در صورت داشتن معیارهای فعل مرکب). مانندِ: حرف زد (اسم + همکرد).

فعل مرکب پیشوندی: فعل‌های پیشوندی با کلمه‌ای ترکیب می‌شوند و معنای واحدی را بیان می‌کنند. معنی مزبور نسبت به معنی لغوی کلمه‌های سازنده غالباً مجازی است. مانند: تن در داد (= تسلیم شد)، سر در آورد (= فهمید).

عبارت فعلی: ترکیب حرف اضافه + اسم (در صورت داشتن معیارهای فعل مرکب). مانند: به دنیا آمد، در میان نهاد.

۳-۲- شبکه‌ی قالب‌ها (FrameNet)

۱-۳-۲- معرفی

پروژه Berkeley FrameNet یک منبع الفبایی آنلاین برای انگلیسی بر اساس قاب معنایی و شواهد نوشتاری به وجود می‌آورد. هدف این است که گستره وسیعی از ترکیبات معنایی و نحوی ممکن برای هر کلمه در هر کدام از معانی آن (ظرفیت‌ها) با استفاده از تفاسیر کامپیوتری از جملات نمونه و جدول‌بندی خودکار مستندسازی شود و نتایج تفاسیر نمایش داده شود. محصول اصلی این کار یعنی پایگاه داده الفبایی FrameNet، در حال حاضر شامل بیش از ۱۰۰۰۰ واحد الفبایی است که بیش از ۶۰۰۰ واحد آن کاملا تفسیر شده‌اند و در بیش از ۱۳۵۰۰۰ جملات تفسیری به عنوان مثال آورده شده‌اند. در ابتدای Release ۱,۳، کیفیت داده‌های FrameNet با استفاده از یک سیستم مدیریت ثبات بررسی شده است. این پایگاه داده پس از سه انتشار، اکنون توسط صدها محقق، معلم و دانش‌آموز در سراسر جهان استفاده می‌شود. پروژه‌های تحقیقاتی به دنبال ایجاد لغتنامه‌های معنایی برای سایر زبانها و اختراع ابزارهایی برای برچسب‌گذاری خودکار متن با استفاده از اطلاعات معنایی هستند.

پروژه FrameNet ساخت یک پایگاه داده واژگان زبان انگلیسی است که قابل خواندن بوسیله انسان و ماشین می باشد که بر اساس حاشیه‌نویسی این که کلمات چگونه در متن واقعی استفاده می‌گردد، می‌باشد. از دیدگاه یک دانش آموز یک فرهنگ لغت بیش از ۱۰۰۰۰ کلمه ای، همراه با مثال‌های حاشیه‌نویسی شده که معنی و چگونگی استفاده آنها را نشان می‌دهد. برای محققان پردازش زبان طبیعی، بیش از ۱۷۰،۰۰۰ جمله که به صورت دستی حاشیه‌نویسی شده است یک مجموعه داده آموزشی منحصر به فرد برای برچسب‌گذاری نقش-های معنایی فراهم کرده است، که در برنامه‌های کاربردی مانند استخراج اطلاعات، ترجمه ماشینی، تشخیص رخداد، تجزیه و تحلیل، و غیره مورد استفاده واقع می‌شود. این پروژه به صورت عملی در موسسه بین المللی علوم کامپیوتر در دانشگاه برکلی از سال ۱۹۹۷ به اجرا در آمده است، در درجه اول توسط بنیاد ملی علوم

حمایت می شود، و داده‌ها به صورت رایگان برای دانلود در دسترس است، توسط محققان در سراسر جهان برای طیف گسترده‌ای از اهداف، دریافت و مورد استفاده می‌گردد.

FrameNet بر حسب یک نظریه به معنی معناشناسی فریم، ناشی از کار چارلز J. Fillmore و همکاران (Fillmore 1976، 1977، 1982، 1985، Fillmoreو بیکر ۲۰۰۱، ۲۰۱۰) است.

۲-۳-۲- ایده کلی

ایده اساسی ساده: توصیف یک نوع رویداد، ارتباط، یا نهاد و شرکت کنندگان در آن : که معانی بسیاری از بهترین کلمات می تواند بر اساس یک چارچوب معنایی قابل درک باشد .به عنوان مثال، مفهوم پخت و پز به طور معمول شامل یک فرد انجام دهنده پخت و پز (cook)، غذایی که پخته می شود (Food)، چیزی که برای نگهداری مواد غذایی در حالی که پخت و پزبه کار می رود (Container) و یک منبع حرارتی(Heating_instrument) می باشد. در پروژهFrameNet ، این مفهوم توسط یک فریم به نام Apply_heat نشان داده می شود و cook، Food، Heating_instrumentو Container عناصرفریم (FEsنامیده می شوند .واژه‌هایی که این فریم را فرا خوانی می کنند، مانند fry، bake، boil، و broil واحدهای لغوی (LUs) از قاب Apply_heat نامیده می شوند .دیگر فریم‌های پیچیده‌تر، مانند Revenge، که شامل Fes های بیشتر (Offender, Injury, Injured_Party, Avenger, and Punishment) و دیگران ساده‌تر هستند. کار FrameNet تعریف فریم‌ها و حاشیه‌نویسی جملات برای نشان دادن چگونگی FES مناسب در کلمات موجود در فریم می‌باشد، است. همانطور که در زیر نمونه-هایی از Apply_heat و Revenge دیده می‌شود :

- ... [Cook the boys] ... GRILL [Food their catches] [Heating_instrument on an open fire].
- [Avenger I] 'll GET EVEN [Offender with you] [Injury for this]!

واحد الفبایی فراخوانی کننده فریم (frame-evoking) می تواند هر یک از دسته بندی های اصلی لغوی مانند اسم ، فعل ، صفت و حتی قید یا حرف اضافه باشد . اما در ساده‌ترین مورد، کلمه فراخوانی کننده فریم (frame-evoking) یک "فعل" میباشد و Fes ها وابستگان نحوی آن فعل می باشند. در مثال بالا میتوان به این که boys فاعل فعل grill و their catches، مفعول مستقیم بوده و on an open fire یک عبارت حرف اضافه ای می باشد. اما LUs می تواند همچنین یک رخداد اسمی ازقبیل retaliation در فریم Revenge باشد.

- [Punishment This attack was conducted] [Support in] RETALIATION [Injury for the U.S. bombing raid on Tripoli...

یا صفت‌هایی مانند asleep در فریم Sleep :

- [Sleeper They] [Copula were] ASLEEP [Duration for hours]

ورودی لغوی برای هر LU از جمله حاشیه‌نویسی مشتق شده است، و راه‌هایی که در ساختارهای نحوی که توسط کلمه متوجه مشخص می‌شود .

بسیاری از اسم‌های رایج، مانند tree، hat و یا tower، معمولا به عنوان وابسته‌گان به FE های سر در نظر گرفته می‌شوند، نه اینکه به وضوح موجب فراخوانی فریم گردند، بنابراین تلاش کمتری به منظور حاشیه-نویسی این موارد اختصاص داده می‌شود، اطلاعات در مورد آنها در لغت‌نامه‌های دیگر دردسترس می‌باشد، مانند (Miller et al. 1990) WordNet . بهرحال شناسایی چنین اسم‌هایی نیز بخشی از ساختار فریم می-باشد، و در واقع در پایگاه داده‌ها FrameNet تعداد اسم‌ها کمی بیشتر از افعال می‌باشد.

بعبارت دیگر، حاشیه‌نویسی FrameNet مجموعه ای از سه گانه هاست که نشان دهنده تحقق FE برای هر جمله می‌باشد، که هر کدام شامل یک نام عنصر فریم (به عنوان مثال، Food)، یک ساختار گرامری (مثلا Object) و یک نوع عبارت (مثلا عبارت اسم (NP)).ما می توانیم هرکدام از این سه نوع حاشیه نویسی بر روی هر FE به عنوان یک " لایه " درنظر بگیریم، اما ساختار گرامری و لایه نوع عبارت در سیستم گزارش مبتنی بر وب نشان داده نمی‌شود، برای جلوگیری از بهم ریختگی بصری .نسخه XML قابل دانلود از داده ها شامل این سه لایه (بیشتر از آنها اینجا مورد بحث نیست) بوده، همراه با فریم کامل و توصیفاتFE ، روابط فریم با فریم، و واژگان ورودی برای هر LU مشخص بوده است. بسیاری از حاشیه نویسی از احکام جداگانه مشخص تنها برای یک LU هستند، اما مجموعه‌ای از متون که در آن تمام فراخوانی کننده‌های فریم معین گردیده وجود دارد؛ فریم‌ها با هم تداخل دارند که موجب ارائه غنی از معنای کل متن می‌گردد. درتیم FrameNet بیش از ۱۰۰۰ فریم معنایی تعریف شده است و آنها را با یک سیستم به فرم‌های کلی‌تری مربوط می‌کند. که مربوط به فریم‌های کلی‌تری و مشخص‌تر برای ارائه استدلال در مورد وقایع و فعالیت‌های ملی می‌باشد.

از آنجا که فریم به صورت معنایی، اغلب در زبان ها مشابه بوده،و به عنوان مثال، فرم های مورد خرید و فروش شامل FESخریدار، فروشنده، محصولات و پول، صرف نظر از زبان که در آن بیان شده است .چندین پروژه در دست اقدام برای ساخت FrameNets به موازات پروژه انگلیسی FrameNet برای زبان در سراسر جهان، از جمله اسپانیایی، آلمانی، چینی، و ژاپنی، و فرم تجزیه و تحلیل معنایی و حاشیه‌نویسی بوده و در زمینه‌های تخصصی از اصطلاحات حقوقی، فوتبال و گردشگری می باشد.

۳-۳-۲- ساختار کلی FrameNet

یک واحد الفبایی ترکیبی از یک کلمه و معنای آن است. عموما هر معنایی از یک کلمه با معانی متعدد به یک فریم معنایی متفاوت تعلق دارد. این فریم یک ساختار ادراکی است که یک موقعیت، موضوع یا اتفاق خاص را

همراه با پایه‌ها و متعلقات آن توضیح می‌دهد. در ساده‌ترین حالت واحد الفبایی فراخوانی کننده فریم، یک فعل است و عناصر فریم ، وابسته‌های نحوی آن هستند.

ورودی الفبایی برای کلمه پیش بینی‌شده که از تفسیر به دست می‌آید، فریمی را که به یک معنی مشخص تاکید دارد و روشی که عناصر فریم را بر اساس ساختار کلمه معلوم می‌کند تشخیص می‌دهد.

اسم‌ها بیشتر از اینکه فریم خود را فرا بخوانند، وابسته هستند. هدف از تفسیر چنین کلماتی شناسایی خبرهایی هست که جمله را کنترل می‌کنند. بدین طریق روشهایی که این اسمها به عنوان عناصر فریم عمل می‌کنند نمایش داده می‌شوند.

در واقع تفاسیر FrameNet مجموعه‌ای از سه‌گانه‌هاست که ادراکی از عنصر فریم برای هر جمله تفسیر شده را به وجود می‌آورد. هر سه‌گانه شامل یک نام عنصر فریم، یک گرامر دستوری و یک نوع عبارت است. به هر کدام از این انواع تفسیر به عنوان یک لایه نگاه می‌شود، اما لایه گرامر دستوری و نوع عبارت در گزارش سیستم نشان داده نمی‌شود.

تفاسیر FrameNet از دو منبع به دست می‌آیند. برای دنبال کردن این هدف که ظرفیت‌های هر کلمه در هرکدام از معانی آن ضبط شود، معمولا بر روی یک واحد الفبایی تمرکز شده و جملات از متون مختلف که شامل آن واحد الفبایی هستند استخراج می‌شوند. سپس نمونه‌هایی از این جملات با توجه به واحد الفبایی تفسیر می‌شوند. در بخش دیگری از پروژه که قسمت کوچکی از تفاسیر را شامل می‌شودُ یک متن تفسیر می‌شود. تفسیر متن با تفسیر جمله از این جهت تفاوت دارد که جمله توسط نویسنده متن برای ما انتخاب می‌شود.

۴-۳-۲- مقایسه با WordNet و آنتولوژی‌ها

پایگاه داده FrameNet یک منبع الفبایی با ویژگی‌های خاص است که آن را از سایر لغت‌نامه‌ها و شناخته‌شده‌ترین منبع الفبایی برخط یعنی WordNet جدا می‌کند.

- مانند لغات لغتنامه، واحدهای الفبایی FrameNet با تعاریفی می‌آیند که یا از لغتنامه آکسفورد به دست آمده است و یا توسط یکی از اعضای FrameNet نوشته شده است.
- برخلاف لغت‌نامه‌های تجاری، مثال‌های تفسیری مختلفی از هر یک از معانی یک کلمه آورده شده است. علاوه بر این مجموعه مثال‌ها تمامی ظرفیت‌های یک واحد الفبایی را نشان می‌دهد.
- مثال‌ها شواهدی هستند که از شرکت‌های طبیعی گرفته شده اند و توسط زبانشناسان ساخته نشده‌اند.
- تحلیل الفبای انگلیسی فریم به فریم انجام می‌شود، برخلاف لغت‌نامه‌های مرسوم که کلمه به

کلمه پیشرفت می‌کنند. به همین دلیل است که لغتنامه‌نویسها میزان پیشرفت را بر اساس کلمات کامل شده اندازه می‌گیرند اما FrameNet بر اساس فریم های کامل شده اندازه می-گیرد. داشتن یک یا تعداد بیشتری واحد الفبایی برای یک کلمه در یک فریم کامل شده مانع یافتن سایر واحدهای الفبایی برای آن کلمه در فریم های آینده نمی‌شود.

- هر واحد الفبایی به یک فریم معنایی و در نتیجه کلماتی که آن فریم را فرا می‌خوانند متصل است. به این لحاظ FrameNet شبیه به لغتنامه است که کلمات هم‌معنی را گرد می‌آورد.
- WordNet و تمامی آنتولوژیها یک رابطه هرمی میان نقاطشان وجود دارد. به طور مشابه FrameNet شامل شبکه‌ای از روابط میان فریم هاست. انواع متعددی تعریف شده‌اند اما مهمترین آن‌ها عبارتند از:
  - وراثت: هر فریم فرزند یک زیرگونه از یک فریم پدر است و هر عنصر فریم در پدر به یک عنصر فریم در فرزند متصل است.
  - کاربرد: فریم فرزند، فریم پدر را به عنوان پیش زمینه پیش فرض می‌گیرد اما لزومی ندارد که تمامی عناصر فریم پدر به فریم فرزند متصل باشند.
  - زیرفریم: فریم فرزند اتفاقی است که زیرمجموعه یک اتفاق پیچیده که توسط پدر نمایش داده می‌شود است.
  - چشم‌انداز: فریم فرزند چشم‌انداز ویژه‌ای از فریم پدر را ارایه می‌کند.

روابط فریم به فریم در گزارش‌های فریم نمایش داده شده‌اند.

- از آنجا که بسیاری از اسم‌های مربوط به گونه‌های طبیعی تفسیر نشده‌اند، پایگاه داده FrameNet به عنوان یک آنتولوژی قابل استفاده نیست. در این زمینه FrameNet تسلیم WordNet است که بخشهای گسترده‌ای را پوشش می‌دهد.

منظور از کلمه چیست؟

وقتی گفته می‌شود کلمه «پختن» دارای معانی متعدد است به این مفهوم است که به سه فریم مختلف متصل است:

- اعمال گرما: مایکل سیب‌زمینی‌ها را ۴۵ دقیقه پخت.
- پختن به معنای ایجاد: مایکل برای تولد مادرش یک کیک پخت.
- جذب گرما: سیب‌زمینی‌ها باید بیش از ۴۵ دقیقه پخته شوند.

اینها سه واحد الفبایی مختلف با سه مفهوم مختلف ایجاد می‌کنند.

۵-۳-۲- توسعه فریم

هسته فرایند همیشه این بوده است که به دنبال گروهی از کلمات باشیم که به لحاظ معنایی با یکدیگر وجه مشترک دارند و آنها را به گروه‌هایی تقسیم کنیم. سپس گروه‌های کوچکتر تبدیل به گروه‌های بزرگتر می‌شوند تا فریم های قابل قبولی به وجود آورند تا آنجا که به آنها کلمات هدف، واحدهای الفبایی یا عناصر فراخواننده فریم گفته می‌شود. در گذشته قیود برای چنین گروه‌بندی‌هایی شهودی و غیررسمی بود اما اکنون این قیود روشن‌تر شده‌اند. کوچکترین گروه‌ها به صورت زیر تشکیل می‌شوند:

- تمامی واحدهای الفبایی در یک فریم باید به لحاظ انواع و تعداد عناصر فریم چه در محتواهای صریح و چه ناصریح یکسان باشند.
  - تعداد: اگر تعداد عناصر فریم موردنیاز از یک واحد الفبایی به یک واحد الفبایی دیگر یا از یک جمله به جمله دیگر متفاوت باشد، این بدان مفهوم است که فریم باید شکسته شود تا فریمی که این تفاوت را از میان ببرد به دست آید.
  - نوع: نوع معنایی پایه برای یک عنصر فریم باید برای کاربردهای گوناگون یکسان بماند. اگر چنین نباشد، باید عناصر فریمی مشخص فرض قرار داده شوند. جداسازی بر این اساس کمک می‌کند که عنصر فریمی صحیح روابط عنصر را در فریم قرار دهد. استفاده از عناصر فریم مشخص، یافتن اطلاعات را ساده و تفسیر را آسان‌تر می‌کند.
- در فریم هایی که به لحاظ ابعاد پیچیده هستند، واحدهای الفبایی باید مجموعه مشابهی از مراحل را در بر داشته باشند. اگر چنین نباشد، فریم ها شکسته می‌شوند.
- عناصر معنایی مشابه در تمامی واحدهای معنایی یک فریم نمایش داده می‌شوند. این بدین معناست که نقطه نظر یک شرکت‌کننده یکسان باید برای همه آن‌ها در نظر گرفته شود.
- روابط میانی بین عناصر فریم باید برای همه واحدهای الفبایی در یک فریم یکسان باشد.
- پیش‌فرض‌ها، انتظارات و همراهان اهداف در یک فریم مشترک هستند.
- معنی و مفهوم پایه‌ای اهداف در یک فریم باید مشابه باشد. به عنوان مثال آبی و شکسته با وجود آنکه به لحاظ توزیع یکسان هستند، در یک فریم قرار نمی‌گیرند زیرا به موقعیتهای متفاوتی اشاره دارند.
- پیش‌معانی که عناصر فراخواننده فریم به عناصر فریمی مختلف می‌دهند مشابه است.

علاوه بر فاکتورهایی که برای گروه‌بندی کلمات به فرم فریم ها وجود دارد، عوامل دیگری هستند که هیچ‌گاه

در نظر گرفته نمی‌شوند.

- در تمامی اوقات، گروه‌هایی که تفاوت‌های معنایی آن‌ها به دلیل ساختارهای عمومی زبان است گردآوری می‌شوند.
- کلمات متضاد نیز در یک گروه گردآوری می‌شوند. اما کلماتی که برعکس یکدیگرند مانند خریدن و فروختن در دو فریم مختلف قرار می‌گیرند زیرا شرکت‌کننده‌های مختلفی را نشان می‌دهند.
- تفاوتهایی که به محتوای سخن مرتبط است. این تفاوتها در قالب انواع معنایی نشان داده می‌شوند.

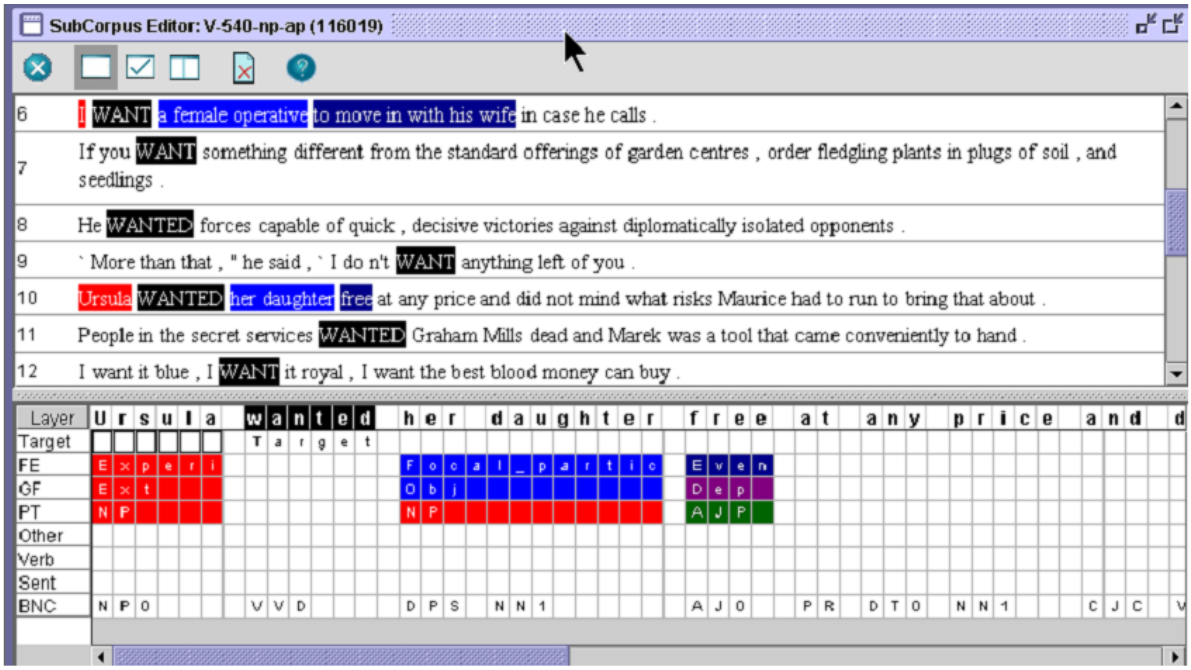
در نهایت توسعه فریم بر ترجمه و تفسیر کلمات و کلمات چندگانه تمرکز دارد. این بدان معناست که آیا می‌توان یک واحد الفبایی را با یک واحد دیگر جایگزین کرد و همچنان فریم یکسانی را فراخواند و با وابسته‌های نحوی واحد الفبایی جدید نقش‌های معنایی مشابهی را بیان کرد. توسعه فریم به‌طور مستقیم به قابلیت ترجمه و تفسیر که در کل گفتار ممکن است وجود داشته باشد اشاره نمی‌کند.

به عنوان مثال در بسیاری از گفتارها معنی یک عضو از تعداد زیادی از عناصر فراخوان فریم به دست می‌آید در حالی‌که سایر اعضا از تنها یک واحد الفبایی که در آن ساختارهای معنایی پیچیده باهم ترکیب شده‌اند به دست می‌آیند.

عموما FrameNet کلمات را نه به دلیل ترجمه و تفسیر گفتار بلکه تنها به دلیل ترجمه و تفسیر بین واحدهای الفبایی در یک گروه قرار می‌دهد.

۶-۳-۲- تفسیر FrameNet

به عنوان یک مساله تکنیکی، روشی که FrameNet برای تحلیل نمونه‌هایی از گزاره هدف دارد شامل اجرای لایه‌هایی موازی از تفسیر با استفاده از مجموعه‌های نام‌گذاری مناسب است که در شکل نشان داده شده است. لایه‌ها می‌توانند توسط مفسر انتخاب شوند. تعداد لایه‌ها و نوع اطلاعاتی که می‌توانند ذخیره کنند نامحدود است. در حال حاضر در FrameNet چهار لایه تفسیر اصلی وجود دارد که عبارتند از هدف، عنصر فریم، عملگر دستوری و نوع جمله‌ای. در اولین لایه گزاره هدف نشان داده می‌شود در حالی‌که در سه لایه دیگر نام-گذاری بر روی جزء اصلی که عناصر فریم برای هدف را بیان می‌کند، اجرا می‌شود.



لایه مهم بعدی شامل مجموعه‌ای از لایه‌هاست که دیگر لایه نامیده شده است. لایه‌ای که با توجه به قسمتی از گفتار هدف که در ان هستیم اسم، فعل، صفت، قید و یا حرف اضافه نامیده می‌شود. لایه مهم بعدی لایه جمله است. در دیگر لایه نام‌گذاری‌ها با توجه به ساختار محتوایی ویژه‌ای که هدف اتفاق می‌افتد نگهداری می‌شود. لایه جمله احتیاجی به نگهداری نام‌های تفسیری ندارد. هنگامی که لایه فراخوانده می‌شود، اطلاعات جمله به صورت یکجا می‌تواند بر روی فهرستی از چک باکس‌ها ذخیره شود.

آخرین لایه در بین لایه‌ها، لایه‌ای است که نام‌گذاری‌های مرتبط با بخشی از گفتار و به رسمیت شناختن موجودیت نام را نگه می‌دارد. این اطلاعات معمولا از بدنه و نرم‌افزار ثالث به دست می‌آید و معمولا توسط مفسران FrameNet تغییر داده نمی‌شود.

اکنون نوبت توضیح فرایند تفسیر FrameNet است. این کار می‌تواند با توجه به جمله‌ای که برای تفسیر انتخاب شده است به دو دسته تقسیم شود. در دسته تفسیر الفبایی تمرکز بر روی ضبط ظرفیت‌های هر کلمه در هرکدام از معانی آن است. برای این کار، جملاتی از متون مختلف که شامل یک واحد الفبایی مشخص باشند درآورده می‌شوند. سپس گلچینی از این جملات با توجه به واحد الفبایی مشخص شده درآورده می‌شوند.

در دسته دیگر که تفسیر متن کامل است، جملات برای FrameNet انتخاب می‌شوند. تفسیر متن کامل به دلیل تکنیک‌های لایه‌ای تفسیر ممکن شده است. فرهنگ‌نویسان FrameNet هر لغت در یک جمله را به عنوان یک هدف شناسایی کرده و سپس یک فریم متناسب با آن را برای تفسیر انتخاب می‌کنند. سپس لایه‌های جدید تفسیر گرفته می‌شوند و برچسب‌های عنصر فریم مناسب زده می‌شود و به دنبال آن اجزای اصلی مرتبط تفسیر می‌شوند.

دستورات زیر بدون توجه به نوع تفسیر برای تفسیر یک برداشت مشخص از یک کلمه هدف اجرا می‌شوند:

- تفاسیر FrameNet به سمت وابسته‌های کلمه هدف جهت دارد. اجزای اصلی که تنها با توجه به متن می‌توانند فهمیده شوند تفسیر نمی‌شوند.
- به جای تفسیر تنها کلمات سرگروه در وابسته‌های نحوی هدف، تمامی اجزای اصلی تفسیر می‌شوند.
- هر وابسته برای موجودیت عنصر فریم، نوع جمله‌ای و عملگر دستوری مرتبط با واحد الفبایی هدف تفسیر می‌شود.

تفاوت بین دو دسته تفسیر که در قبل معرفی شدند بدین قرار است:

- در دسته الفبایی تفسیر با توجه به تنها یک واحد الفبایی در هر جمله انجام می‌شود. در تفسیر متن کامل، تمامی واحدهای الفبایی که محتوایی دارند به عناون هدف در نظر گرفته می‌شوند و تمامی وابسته‌های آن‌ها تفسیر می‌شوند.
- در تفسیر متن کامل تمامی برداشتها از کلمه هدف چه به اطلاعات الفبایی در مورد هدف کمک کند و چه نکند باید تفسیر شوند.
- در تفسیر الفبایی، هدف آن است که مجموعه جملاتی که برای واحد الفبایی داده شده تفسیر می‌شوند نمایانگر کلیه امکانات ترکیبی برای آن واحد الفبایی باشند. در یک متن داده شده، غیرمعمول است که تمامی روندهایی که هدف ممکن است در آن اتفاق بیفتد بررسی شوند.
- تفاسیری که در پایگاه داده FrameNet وجود دارند هیچ اطلاعاتی در مورد فرکانس وقوع نمی‌دهد. تفسیر متن کامل به این معناست که تمامی برداشتها از یک هدف باید تفسیر شوند. در نتیجه حداقل در مورد یک متن مشخص، اطلاعات فرکانس برای یک واحد الفبایی می‌تواند استخراج شود.

تفسیر الفبایی را نیز می‌توان به دو قسمت با توجه نوع کلمه هدف تقسیم کرد:

- تفسیر با توجه به کنترل‌کننده نحوی فریم، که یک گزاره، تغییردهنده یا یک عبارت ارجاع دهنده است.
- تفسیر با توجه به یک فاصله‌پرکن که به این معناست که تفسیر با توجه به یک عبارت ارجاع‌دهنده که یک عنصر فریم از یک فریم که با استفاده از کنترل‌کننده تشخیص داده شده است، انجام می‌شود.

۴-۲- برچسب‌زنی نقش معنایی کلمات (SRL)

برچسب‌زنی معنایی کلمات[۱] مشابه برچسب‌گذاری اجزای واژگانی کلام بوده با این تفاوت که عمیق‌تر و پیچیده‌تر از آن می‌باشد. برچسب‌زنی معنایی، وظیفه‌ی استخراج نقش‌های معنایی جملات نظیر فاعل، مفعول مستقیم، مفعول غیرمستقیم، فعل و ... را بر عهده دارد. برچسب‌زنی معنایی کلمات هم عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی (NLP) از قبیل ترجمه ماشینی، خطایاب و شباهت معنایی می‌باشد.

از جمله نمونه های انگلیسی آن می توان به OpenNlP، Illinois SRL، Swirl و LTHSRL اشاره کرد. این ابزارها از الگوریتم پارسینگ charniak استفاده می کنند. درشکل ۲-۴ یک مجموعه برچسب بکار رفته توسط این ابزار معرفی شده است [CCG2].

ARGUMENTS	
A0	subject
A1	object
A2	indirect object
ADJUNCTS	
AM-ADV	adverbial modification
AM-DIR	direction
AM-DIS	discourse marker
AM-EXT	extent
AM-LOC	location
AM-MNR	manner
AM-MOD	general modification
AM-NEG	negation
AM-PRD	secondary predicate
AM-PRP	purpose
AM-REC	recipricol
AM-TMP	temporal
OTHER LABELS	
C-*arg*	continuity of an argument/adjunct of type *arg*
R-*arg*	reference to an actual argument/adjunct of type *arg*

شکل ۲-۹- نمونه‌ای از یک مجموعه برچسب نقش معنایی کلمات [CCG2]

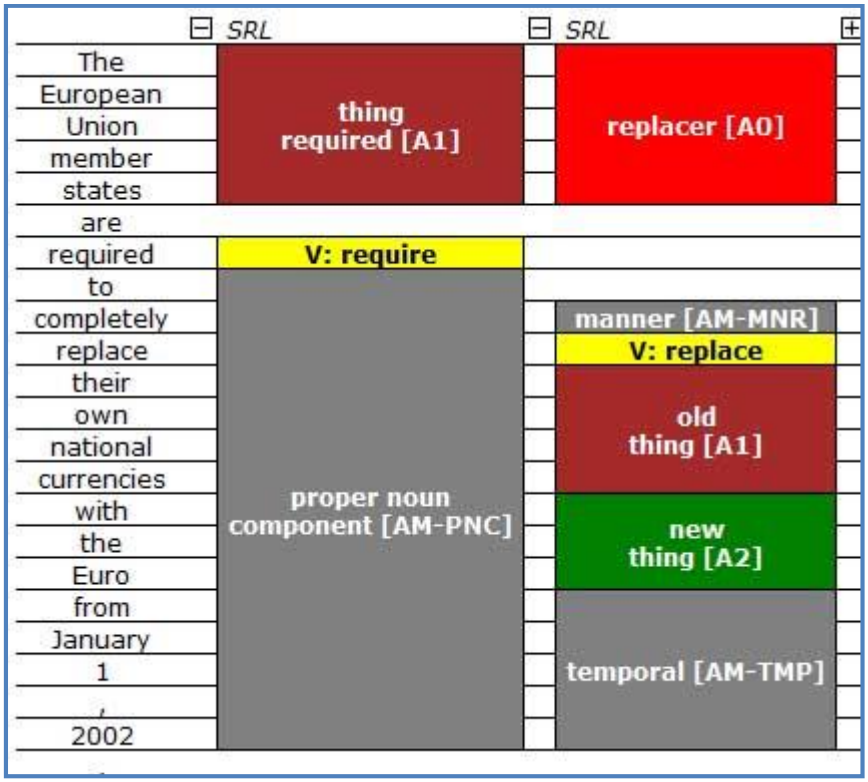
در واقع نقش معنایی نقشی است که یک جز در ارتباط با فعل جمله ایفا می‌کند. یکی از SRLهای معروف، ابزار برچسب زنی معنایی دانشگاه Illinois At Urbana Champaign[۲] که از قوی‌ترین ابزارهای برچسب زنی معنایی است، می‌باشد. با استفاده از این ابزار، نقش‌های معنایی جمله نظیر فاعل، مفعول مستقیم و غیر

۱- Semantic Role Labeling

۲- http://cogcomp.cs.illinois.edu/

مستقیم و ... در ارتباط با فعل جمله مشخص می‌شود. بعبارت دیگر هر فعل در جمله با نقش‌های تکمیل کننده‌اش برچسب زده می‌شود که در اینجا سطح نامیده شده است. بنابراین به تعداد فعل‌های جمله، سطوح مختلفی وجود دارد. در شکل ۳-۶ نمونه ای از خروجی این ابزار برای جمله ورودی زیر نشان داده شده است.

Example 1:The European Union member states are required to completely replace their own national currencies with the Euro from January 1, 2002.



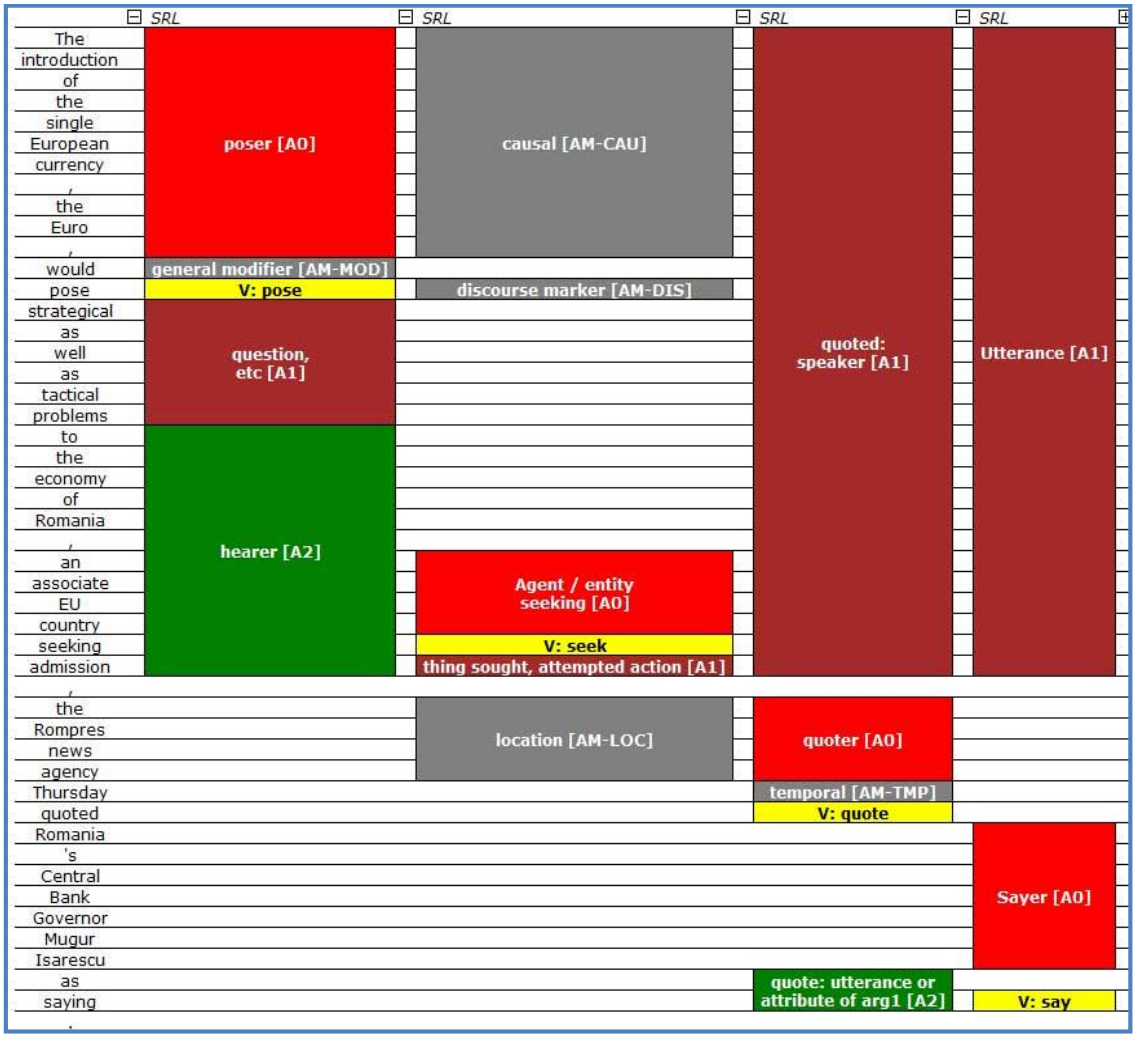
شکل ۲-۱۰- نمونه از خروجی ابزار برچسب زنی دانشگاه ایلینویز

در جمله داده شده دو فعل "*replace*" و "*require*" وجود دارد و بهمین دلیل متناسب با هر فعل یک سطح تعریف شده است. در سطح اول که متناسب با فعل "*require*" می باشد سه نقش معنایی تشخیص داده شده است. "*The European union member states*" به عنوان مفعول مستقیم، "*required*" به عنوان فعل و "*to completely replace their own national currencies with the euro from January 1, 2002*" به عنوان نقش قیدی. در سطح دوم هم متناسب با فعل "*replace*" پنج نقش معنایی تشخیص داده شده است.

لازم به ذکر است که در مواردی که عموما طول جملات طولانی‌تر می‌شود، تعداد این سطوح هم افزایش می-یابد. به عنوان مثال به جمله طولانی زیر را درنظر بگیرید:

Example2: The introduction of the single European currency, the Euro, would pose strategical as well as tactical problems to the economy of Romania, an associate EU country seeking admission, the Rompres news agency Thursday quoted Romania's Central Bank Governor Mugur Isarescu as saying.

خروجی برچسب زده آن در شکل ۳-۷ آمده است که شامل چهار سطح متناسب با چهار فعل آن می‌باشد.



شکل ۲-۱۱- نمونه خروجی ابزار SRL برای جملات طولانی

مهم‌ترین چالشی که در اندازه‌گیری شباهت کلمات در نقش‌های یکسان پیش می‌آید وجود سطوح برچسب زده متفاوت از هم می‌باشد که هرکدام شامل چندین نقش مجزا از یکدیگر هستند. به عنوان نمونه در جمله مثال ۲ چهار سطح وجود دارد که هر کدام برچسب "A0" یا همان فاعل را دارند. چالشی که در این جا مطرح می‌شود این است که کدام یک از این برچسب‌ها را برای مقایسه در نقش فاعلی باید انتخاب نمود. در

کاربردهایی که از برچسب‌زنی معنایی وجود دارد عموما این مشکل پیش می‌آید و معمولا سطحی که بیشترین کلمات را پوشش می‌دهد به عنوان سطح اصلی در نظر گرفته می‌شود[SHE07][AKS09].

۵-۲- خلاصه

در این فصل در ابتدا اطلاعات پایه و اولیه مورد نیاز شرح داده شد. پیش‌نیازهای لازم برای درک مفاهیم مطرح شده و کارهای انجام شده در ارتباط با موضوع پایان‌نامه ذکر گردید. در ادامه فرهنگ ظرفیت نحوی افعال فارسی که کار گروه پژوهش زبان‌شناسی رایانه‌ای دکتر محمدصادق رسولی می‌باشد و در راستای تولید فریم‌نت زبان فارسی مورد استفاده قرار گرفت، اشاره شد. سپس مفاهیم مرتبط با شبکه قالب جملات و کارهای انجام شده در این حوزه و کلیه مفاهیم و موضوعات این حوزه تشریح گردید. در انتها برچسب‌زن نقش معنایی در متون مختصرا مورد اشاره قرار گرفت.

فصل سوم :

روش پیشنهادی

۳- روش پیشنهادی

دو روش کلی برای برچسب‌زنی نقش‌های معنایی متون وجود دارد که شامل الگوریتم‌های با ناظر[۱] و بدون ناظر[۲] می‌باشد. از روش‌های با ناظر می‌توان به [GIL02] اشاره کرد. این روش‌ها از پیکره‌های برچسب‌زده شده‌ی انسانی به عنوان مجموعه‌ی یادگیری استفاده می‌کنند. از جمله‌ی این پیکره‌ها FrameNet و PropBank می‌باشند. از روش‌های دسته دوم نیز می‌توان به [SWI04] اشاره کرد.

در راستای انجام این پایان‌نامه، روش اول یعنی الگوریتم‌های با ناظر مدنظر قرار گرفته و به همین دلیل ابتدا بایستی شبکه‌ی قالب‌های جملات برای افعال زبان فارسی یا همان فریم‌نت تولید گردیده و سپس ابزار برچسب‌زنی نقش‌های معنایی متون برای زبان فارسی طراحی و پیاده‌سازی گردد.

پس از طراحی و تکمیل شبکه‌ی قالب‌های جملات، ابزار SRL زبان فارسی به تطبیق جملات یک متن با فریم‌های موجود در پیکره تولید شده پرداخته و برچسب معنایی متناظر با آن جمله را به کلمات تشکیل دهنده‌ی جمله با دقت قابل قبولی اختصاص می‌دهد. البته در نسخه فعلی ابزار پیاده‌سازی شده، گروه‌های تشکیل دهنده‌ی جملات، تک کلمه‌ای در نظر گرفته شده‌اند؛ چرا که برای شناسایی گروه‌های تشکیل‌دهنده‌ی جملات متون فارسی نیاز به ابزار پارسر زبان فارسی با دقت بسیار بالا می‌باشد.

در ادامه گام‌های سپری شده برای تولید فریم‌نت و ابزار SRL زبان فارسی به تفکیک تشریح می‌گردد.

۳-۱- تولید فریم‌نت

همانطور که به طور مفصل ذکر گردید فریم‌نت شبکه‌ای از قالب‌های اجزای جملات می‌باشد که بر اساس ظرفیت افعال تولید می‌گردد. در راستای تولید فریم‌نت برای زبان فارسی، از پیکره‌ی ظرفیت نحوی افعال زبان فارسی استفاده شده است. فرهنگ ظرفیت نحوی افعال فارسی مجموعه‌ای حاوی اطلاعات مربوط به ظرفیت نحوی بیش از ۴۵۰۰ فعل در زبان فارسی است که ظرفیت افعال را با قالب‌بندی بن ماضی، بن مضارع، پیشوند، فعل‌یار و حرف اضافه فعلی عرضه نموده است. حال بایستی روالی را برای استخراج قالب‌های جملات زبان فارسی در قالب عبارات منظم طراحی کرده و از این روال برای تولید فریم‌نت زبان فارسی استفاده نماییم.

۳-۱-۱- قالب پیکره‌ی ظرفیت نحوی افعال زبان فارسی

جهت استخراج فریم‌ها از پیکره‌ی ظرفیت نحوی افعال زبان فارسی لازم است که به طور کامل قالب بیان ظرفیت افعال درک گردیده و بر این اساس به تولید فریم‌های مورد نظر با توجه به ساختار افعال بپردازیم. در

۱- Supervised

۲- Unsupervised

پیکره‌ی ظرفیت نحوی افعال زبان فارسی، افعال با قالب زیر معرفی می‌گردند:

بن-ماضی بن-مضارع پیشوند فعلیار حرف-اضافهٔ-فعلی ساخت-ظرفیتی

در واقع قالب فوق بمنزله‌ی کلید برای شناسایی افعال مورد استفاده قرار می‌گیرد و در ادامه، ظرفیت فعل برای تولید جملات بدون ذکر فعل در یک قالب تعریف شده توسط تیم تولید کننده‌ی پیکره‌ی ظرفیت نحوی افعال زبان فارسی ذکر گردیده است.

در این پیکره عبارت و نمادهای متعددی بکار رفته است که برخی از آنها را ادامه آورده‌ایم.

<>: نمایندهٔ یک ساخت ظرفیتی

(): اختیاری

[]: جزئیات متمم ظرفیتی

|: یا

،: و

-: بدون (برای جزئیات متمم)

+: با (برای جزئیات متمم)

+/-: با یا بدون (برای جزئیات متمم)

التزامی: وجه التزامی فعل بند متممی(جزئیات فعل بند متممی)

مطابقت: مطابقت شخص و شمار فاعل بند اصلی با فعل بند متممی (جزئیات فعل بند متممی)

را: نشانگر "را" برای مفعول صریح

حالت: جزئیات متمم قیدی

مکان: جزئیات متمم قیدی

زمان: جزئیات متمم قیدی

مقدار: جزئیات متمم قیدی

۲-۱-۳- چگونگی تولید عبارات منظم در فریم‌ها

در پیکره‌ی ظرفیت نحوی افعال زبان فارسی، پس از بیان هر فعل و ویژگی‌های آن نظیر بن ماضی، بن مضارع،

پیشوند، فعل‌یار و حرف اضافه فعلی، ظرفیت فعل برای تشکیل جملات بیان گردیده است که برای تولید فریم‌نت بایستی ظرفیت بیان شده به درستی درک گردیده و در قالب یک عبارت منظم برای تولید فریم‌نت و بکارگیری در سایر ابزارهای پردازش متن از جمله ابزار برچسب‌زنی نقش‌های معنایی متون فارسی بیان گردد.

برای تولید عبارت منظم با توجه به ظرفیت هر فعل لازم به بیان معنی برخی از عبارات بکار رفته در پیکره‌ی ظرفیت نحوی افعال زبان فارسی و فریم‌نت تولید شده می‌باشد که در ادامه به ذکر این موارد می‌پردازیم.

فا: فاعل

مف: مفعول

مفد: مفعول دوم

مفح: مفعول حرف اضافه‌ای

Ø: فاعل تهی

مس: مسند

تم: تمییز

بند: بند متممی فعل

مق: متمم قیدی

مفن: مفعول نشانه اضافه‌ای

از آنجاییکه افعال در پیکره‌ی ظرفیت نحوی افعال زبان فارسی هر کدام در یک خط بیان گردیده‌اند، الگوریتمی را طراحی نموده‌ایم که متن این پیکره را خط به خط خوانده و مورد تجزیه و تحلیل قرار می‌دهد. همانطور که قبلا ذکر گردید در هر خط از این پیکره پس از بیان هر فعل و ویژگی‌های آن نظیر بن ماضی، بن مضارع، پیشوند، فعل‌یار و حرف اضافه فعلی، ظرفیت فعل برای تشکیل جملات بیان گردیده است که بایستی به یک عبارت منظم قابل استفاده در طراحی ابزار SRL تبدیل گردد.

در ابتدا دو نماد "<" و ">" از دو طرف عبارت حذف می‌شود. سپس به جای "،" کاراکتر فاصله در عبارت جایگزین می‌گردد. حروف اضافه تکی یا چندتایی در قالب یک پرانتز به قبل از مفعول حرف اضافه‌ای منتقل می‌شوند و با توجه به اجباری یا اختیاری بودن گروه قیدی یا حروف اضافه‌ای، حروف اضافه و مفعول حرف اضافه‌ای در نماد "(...)" یا "*(...)" قرار می‌گیرند. عبارت بیان شده در بین نماد "[]" در پیکره نیز بیانگر یک سری ویژگی‌ها و قیدهای خاص در مورد ساختار جمله هستند که نیازی به حضور در عبارت منظم ندارند و به همین دلیل از ساختار ظرفیت فعل حذف می‌گردند. عبارتی نظیر [را+/-] هم بیانگر وجود اختیاری یا اجباری نشانگر مفعولی "را" و یا حتی عدم حضور آن در جملات می‌باشد که در قالب عبارات منظم به صورت

"(را)*" یا "(را)" و یا "[را^]" بیان می‌گردد.

در پایان پس از تکمیل عبارت منظم برای برچسب‌های نقش‌های معنایی برای جملات، به ازای هر کدام از نقش‌های موجود، لغت "کلمه" در عبارت منظم قرار گرفته و جهت تطبیق با جملات مورد بررسی در قالب یک عبارت منظم دیگر ذخیره می‌گردد.

همچنین عبارت فعلی در انتهای هر دو بخش فریم به صورت زیر اضافه می‌گردد:

حرف-اضافهٔ-فعلی + Space + فعلیار + Space + پیشوند + (بن-ماضی|بن-مضارع)

تبدیل ظرفیت افعال به عبارت منظم در قالب چند مثال بیان می‌گردد.

افتاد افت - - -

<فا،(مفح) [از]،(مفح) [بر|به|بین|داخل|در|درون|روی|میان]>

عبارت منظم برای فعل فوق به صورت زیر درخواهد آمد:

کلمه ((از) کلمه)*((بر|به|بین|داخل|در|درون|روی|میان) کلمه)*((افتاد|افت))

فا ((از) مفح)*((بر|به|بین|داخل|در|درون|روی|میان) مفح)*((افتاد|افت))

مثال دوم:

افتاد افت - جا - <Ø،(مفح) [برای]،بند[التزامی+/-مطابقت+/-]>

عبارت منظم تولید شده:

((برای) کلمه)*کلمه (جا (افتاد|افت))

((برای) مفح)*بند (جا (افتاد|افت))

مثال سوم:

کرد کن پیدا تسکین - <فا،(مفح) [از]>

عبارت منظم تولید شده:

کلمه ((از) کلمه)*(تسکین پیدا(کرد|کن))

فا ((از) مفح)*(تسکین پیدا(کرد|کن))

مثال چهارم:

گزید گزین بر - - <فا،مف[را+/-]،(مفح)[از|از بین|از میان]>

عبارت منظم تولید شده:

کلمه کلمه (را)*((از|از کلمه)*بین|از میان) (بر(گزید|گزین))
فا مف (را)*((از|از مفح)*بین|از میان) (بر(گزید|گزین))

۳-۱-۳- نحوه ذخیره‌سازی فریم‌ها

پس از تشکیل عبارت منظم برای هر کدام از افعال موجود در پیکره‌ی فرهنگ ظرفیت نحوی افعال فارسی بایستی آنها را با ساختاری مناسب در پیکره‌ای دیگر که در واقع فریم‌نت زبان فارسی می‌باشد، ذخیره نماییم. برای ذخیره‌سازی فریم‌های تولیدی، همان ساختار ویژگی‌های استخراجی فعل را به صورت کلید هر ردیف فریم در نظر می‌گیریم. در واقع کلید هر ردیف فریم به صورت زیر درنظر گرفته شده و در مقابل آن فریم‌های موجود برای این فعل درج خواهند گردید.

بن-ماضی#بن-مضارع#پیشوند#فعلیار#حرف-اضافهٔ-فعلی

پس از بیان کلید هر فریم، فریم منطبق با آن فعل به صورت زیر در ادامه آن کلید درج می‌گردد.

[عبارت منظم برای تطبیق جملات ورودی@عبارت منظم دارای برچسب نقش‌های معنایی]

با توجه به ظرفیت هر فعل، آن فعل ممکن است دارای یک یا چندین قالب فریم با ساختارهای متفاوت باشد که بایستی به نحوی مناسب قالب‌بندی و ذخیره‌سازی گردد که در الگوریتم تشخیص برچسب معنایی متون براحتی قابل استفاده باشد.

کلید فریم [فریم]

کلید فریم [فریم ۱] [فریم ۲] [فریم ۳]

برای تخصیص فریم‌های متفاوت موجود برای یک ساختار یکسان از فعل در ابتدا از یک ساختار درهم (Hash) برای ذخیره‌سازی فریم‌ها استفاده می‌گردد که کلید آن کلید فریم و عناصر آن لیستی شامل فریم‌های منطبق با ظرفیت فعل می‌باشند. در انتها با خواندن تمامی عناصر موجود در ساختار درهم یا HashTable کلیه

کلیدها و فریم‌های آنها استخراج و در قالب یک فایل متنی ذخیره می‌گردد.

۴-۱-۳- بسط فریم‌ها و فریم‌های با کلید یکسان

همانطور که قبلا ذکرگردید در پیکره فرهنگ ظرفیت نحوی افعال فارسی پس از بیان هر فعل و ویژگی‌های آن نظیر بن ماضی، بن مضارع، پیشوند، فعل‌یار و حرف اضافه فعلی، ظرفیت فعل برای تشکیل جملات بیان گردیده است که بایستی به عبارت منظم قابل استفاده در طراحی ابزار SRL تبدیل گردد. یک عبارت منظم برای تطبیق جملات ورودی با ساختار فریم و یک عبارت منظم حاوی برچسب‌های معنایی برای تخصیص برچسب نقش‌های معنایی به کلمات موجود در جمله‌ی وردی تطبیق یافته با فریم مورد نیاز می‌باشد. گفته شد که با توجه به ظرفیت هر فعل، آن فعل ممکن است دارای یک یا چندین قالب فریم با ساختارهای متفاوت باشد.

فرض اولیه در طراحی فریم‌نت و برچسب‌زنی معنایی متون بر این بود که اجزای فعل به درستی تشخیص داده شوند و براحتی فعل موجود در جمله با کلید یکی از ساختارهای موجود در فریم‌نت منطبق گردد. این درحالیست که در بسیاری از اوقات با توجه به وجود ضمایر متصل و منفصل و یا سایر لغات بین حروف اضافه فعلی و همچنین فعل‌یار با هسته‌ی اصلی فعل، اجزای فعل یعنی حروف اضافه فعلی و همچنین فعل‌یار گاها به درستی تشخیص داده نمی‌شود و در اینصورت ساختار جمله با هیچ کدام از فریم‌ها منطبق نگردیده و جمله برچسب معنایی نمی‌خورد.

برای رفع این مشکل، هر کدام از کلیدهای فریم‌ها یک بار در شکل خود و یک بار دیگر هم با ساختاری که دارای فعل‌یار و حرف اضافه‌ی فعلی نیستند، ذخیره می‌گردد تا در مواردیکه در روال پردازش جمله ورودی پس از شناسایی عبارت فعلی، اجزای فعل یعنی حروف اضافه فعلی و فعل‌یار به درستی تشخیص داده نشدند باز هم با کلیدی در فریم‌نت منطبق گردیده و در بین فریم‌های موجود در آن فعل با یک فریم تطابق یافته و برچسب‌های متناظر با آن جمله به لغات تشکیل دهنده‌ی آن جمله، تخصیص یابد.

یعنی در واقع به ازای هر فعل یک بار آن فعل به صورت زیر بدون حروف اضافه فعلی و فعل‌یار هم در فریم‌نت ذخیره می‌گردد.

بن-ماضی#بن-مضارع#پیشوند#-#-

لازم به ذکر است در مواردیکه فعل دارای پیشوند، حرف اضافه فعلی یا فعل‌یار نباشد، نماد '-' به جای آن قرار می‌گیرد.

بدین ترتیب فریم‌های موجود برای هر فعل با اجزای آن بسط پیدا کرده و در دو قالب مجزا در فریم‌نت ذخیره می‌گردد. در صورتیکه در روال برچسب‌زنی معنایی متون، اجزای فعل به درستی تشخیص داده شود، مستقیما

به فریم منطبق با آن رفته و برچسب‌های معنایی متناظر با آن جمله را بـه کلمـات آن جملـه تخصـیص مـی-دهیم؛ ولی درصورتیکه اجزای فعل به درستی تشخیص داده نشود، بایستی در لیستی از فریم‌هـای منطبـق بـا بن فعل جستجو کرده تا به فریم منطبق با آن رسیده و برچسب‌های معنایی متناظر با آن جمله را بـه کلمـات آن جمله تخصیص می‌دهیم که روالی طولانی‌تر می‌باشد.

مثال مرتبط با این موضوع را می‌توانید در سه فریم زیر و بسط آنها بدون اجزای فعل مشاهده نمایید:

کرد#کن#-#مشاهده#- [کلمه کلمه (را)*(مشاهده (کرد|کن))@فا مف (را)*(مشاهده (کرد|کن))]

کرد#کن#-#مطالعه#- [کلمه (کلمه) (را)*(مطالعه (کرد|کن))@فا (مف) (را)*(مطالعه (کرد|کن))]

کرد#کن#-#رشد#- [کلمه ((در) کلمه)*(رشد (کرد|کن))@فا ((در) مفح)*(رشد (کرد|کن))]

کرد#کن#-#-#- [کلمه کلمه (را)*(مشاهده (کرد|کن))@فا مف (را)*(مشاهده (کرد|کن))]

[کلمه (کلمه) (را)*(مطالعه (کرد|کن))@فا (مف) (را)*(مطالعه (کرد|کن))]

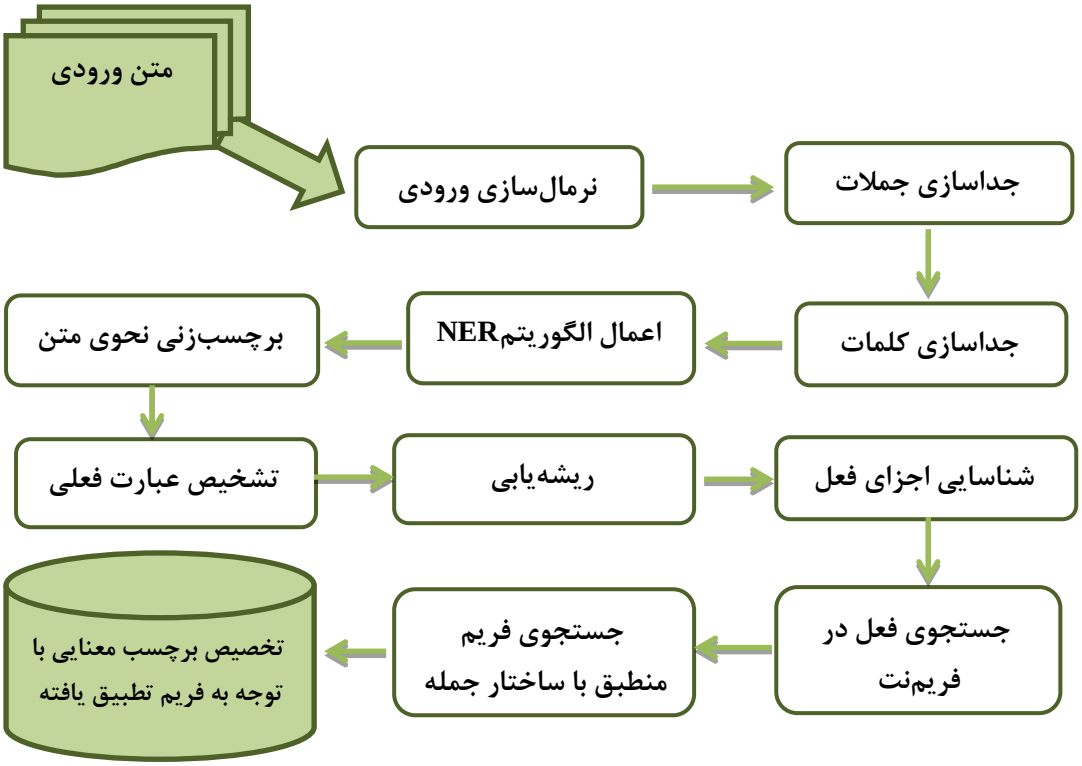
[کلمه ((در) کلمه)*(رشد (کرد|کن))@فا ((در) مفح)*(رشد (کرد|کن))]

۲-۳- برچسب‌زنی نقش‌های معنایی متون

در روند هرگونه پردازش روی متن‌های زبان طبیعی انجام یک سری عملیات پیش‌پردازش، امری اجتناب‌ناپذیر است. بطوریکه دقت این پیش‌پردازش‌ها تاثیر بسـزایی بـر نتـایج اعمـال الگـوریتم‌هـا در فازهـای بعـدی دارد. هرچقدر که دقت این پیش‌پردازش‌ها بیشتر باشد، الگوریتم‌ها به نتایج واقعی خود نزدیک‌تر خواهند شد.

۱-۲-۳- شمای کلی روش پیشنهادی

در این بخش شمای کلی روش پیشنهادی برای پردازش متون و اعمال الگوریتم‌های مورد نظر برای برچسـب-زنی نقش‌های معنایی متون فارسی آورده شده است. مراحل کلی روش پیشنهادی در شـکل ۳- ۱ نشـان داده شده است.



شکل ۳-۱- روش پیشنهادی برای پیش‌پردازش و اعمال الگوریتم برچسب‌زنی نقش‌های معنایی متون

۲-۲-۳- پیش‌پردازش‌ها

در ادامه ابتدا روند استفاده از ابزارهای مورد نیاز برای پیش‌پردازش متون فارسی ذکر گردیده و سپس روش پیشنهادی برای برچسب‌زنی نقش‌های معنایی متون تشریح خواهد شد. در روند انجام پایان‌نامه، ابزارهای زیر برای پردازش متون فارسی مورد استفاده قرار گرفت که به ترتیب بر روی هر کدام از متن‌های مورد پردازش، اعمال می‌گردند. لازم به ذکراست که این ابزارها در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد طراحی، تولید و ارائه گردیده است.

- استانداردکننده یا نرمال‌ساز: در ابتدا بایستی همه‌ی نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن یکسان‌سازی گردند.
- جداکننده جملات: با کمک این پردازشگر می‌توان جملات را از متن استخراج کرد.
- جداکننده کلمات: با کمک این پردازشگر می‌توان کلمات را استخراج نمود.

- ریشه‌یاب: وظیفه ریشه‌یابی کلمات را برعهده دارد.
- تشخیص دهنده‌ی عبارت فعلی: عبارت فعلی را در یک جمله تشخیص می‌دهد که این عملیات با بهره‌گیری از پیکره دادگان صورت می‌گیرد.
- برچسب‌زننده اجزای واژگانی کلام: از این پردازشگر برای برچسب‌زنی اجزای واژگانی کلام استفاده می‌شود.
- تجزیه‌گر: از این پردازشگر برای تجزیه نحوی جملات استفاده می‌شود. در حال حاضر و در نسخه‌ی فعلی ابزار تولیدی در راستای انجام این پایان‌نامه به دلیل دقت پایین این ابزار، از استفاده از آن در طراحی برچسب‌زنی نقش‌های معنایی متون صرف‌نظر گردید. در صورت طراحی ابزاری با دقت بالا برای تشخیص گروه‌های تشکیل دهنده جملات می‌توان به جای عبارت "کلمه" در عبارات منظم، عبارت "گروه" را قرار داد که منطبق با گروه‌های اسمی، فعلی، قیدی، صفتی و حرف اضافه‌ای می‌گردد.

۱-۲-۲-۳- استاندارد کننده (Normalizer)

پردازش زبان فارسی از جهاتی با پردازش زبان انگلیسی تفاوت دارد. در زبان انگلیسی تمامی حروف و تمامی کلمات جدا از هم و با قانونی مشخص نوشته می‌شوند و این در حالی است که در زبان فارسی بعضی از حروف به هم چسبیده‌اند، بعضی از حروف جدا از هم نوشته می‌شوند، بعضی از کلمات یکپارچه‌اند، بعضی از کلمات با فاصله یا نیم‌فاصله به دو یا چند بخش تقسیم می‌شوند. علاوه براین بعضی از حروف مانند "ی" در بعضی از نوشته‌ها با نسخه عربی مانند "ي عربی" نوشته می‌شوند که ممکن مشکلاتی را در الگوریتم‌های پردازش متون فارسی بوجود می‌آورد.

در ادامه ابتدا مشکلات و چالش‌های پیش رو در پردازش متون فارسی ذکرگردیده و سپس روش بکار رفته برای اصلاح و یکسان‌سازی متن آمده است.

۲-۲-۲-۳- چالش‌های ذاتی زبان فارسی

زبان شیرین فارسی دارای چالش‌های ذاتی‌ایی می‌باشد که در ذیل به تعدادی از آنها اشاره شده است:

- وجود کلماتی که معنای آنها در درون جملات مشخص می‌گردد. مثلا کلمه "*شیر*" می‌تواند به معنای "*شیر خوردنی*"، "*شیر آب*" و "*شیر جنگل*" بکار رود.
- عدم وجود علامت‌های آوایی می‌تواند باعث بروز ابهاماتی گردد. مثلا کلمه "کرم" می‌تواند به صورت‌های "*کرم kerem*"، "*کرم karam*"، "*کرم korom*" و "*کرم kerm*" بکار رود.
- عدم رعایت فاصله و نیم‌فاصله حتی در رسم‌الخط رسمی فارسی.

- در زبان فارسی مانند زبان انگلیسی حروف کوچک و بزرگ برای تشخیص اسامی خاص وجود ندارد.
- لغات مرکب چند قسمتی که تشخیص کلمات را مشکل می‌کنند، مانند: "*توپ‌جمع‌کن*" یا "*ابدارچی*".
- وجود ابهام در معنی جملات بدلیل عدم بکارگیری صحیح علائم سجاوندی و نحوه‌ی فاصله-گذاری. مثلا "*مردخوب می‌دود*" و "*مرد خوب می‌دود*".
- بی‌قاعده بودن قواعد تولید بن مضارع تعدادی زیادی از افعال.
- و سایر موارد مشابه.

۳-۲-۲-۳- چالش‌های نگارشی زبان فارسی

به دلیل ویژگی‌های خاص زبان فارسی، چالش‌هایی در نگارش کلمات فارسی وجود دارد که در ذیل بـه برخـی از آنها اشاره شده است:

- وجود کاراکترهای مختلف که دارای تلفظ یکسان در زبان فارسی هستند باعث می‌شود تا غلط‌های املایی زیادی رخ دهد. مثلا کلمه "صابون" ممکن است به اشتباه به صورت "*ثابون*" یا "*سابون*" نوشته شود. یا کلمه‌ای مثل "*تهران*" ممکن است به صورت "*طهران*" نوشته شود که البته در قدیم بیشتر به این صورت بکار می‌رفت. همچنین وجود حرف‌هایی که ممکن است تلفظ نشوند، باعث پیچیدگی نگارش زبان فارسی می‌گردد مثل کلمه "خواهر".
- استفاده از کاراکتر "ا" به جای "آ" و بالعکس در بعضی از کلمات مانند "*ابادان*".
- عدم رعایت درست فاصله و نیم‌فاصله در نگارش کلمات.
- استفاده از هر دو شکل "*می چسبان*" و "*می غیر چسبان*" در کلمات مختلف.، مانند "*می‌روم*" یا "*میروم*".
- استفاده از کلمات زبان‌های دیگر به صورت نگارش فارسی و نگارش انگلیسی، مثلا کلمه "*انتولوژی*" و "*Antology*".
- نگارش مختلف کلمات مختوم به ه که با ی به کلمه بعدی متصل می‌شوند، نظیر "*چاله فضایی*" و "*چاله‌ی فضایی*"
- استفاده از چندین فاصله بین کلمات.
- کاربردهای مختلف از "*ها*" به صورت چسبان و غیرچسبان نظیر "*کتاب هایش*" و "*کتابهایش*".

- نسبت به زبان انگلیسی و یا عربی، قوانین بسیار کمتری دارد و در حقیقت استثنائات زیادی در زبان فارسی در قسمت‌های مختلف وجود دارد.

۴-۲-۲-۳- روش بکار رفته برای یکسان‌سازی متن و اصلاحات اولیه

در اولین گام باید متون برای استفاده در گام‌های بعدی به شکلی استاندارد درآیند. به همین دلیل سعی شده است تفاوت‌های ساده‌ی ظاهری برطرف گردد. برای رسیدن به این هدف، قبل از اعمال الگوریتم‌های مورد نظر بر روی متون، پیش‌پردازش‌هایی روی آنها آنجام می‌شود. طبیعتا هر چه این پیش‌پردازش‌ها قوی‌تر باشد، نتایج حاصل ازمقایسه متون قابل اطمینان‌تر خواهد بود. لازم به ذکر است که از آن جایی که زبان فارسی جزو زبان‌های غیر ساختیافته است با مشکلات بسیار بیشتری نسبت به سایر زبان‌ها مواجه خواهیم شد. متون غیرساخت‌یافته، متونی هستند که پیش فرض خاصی در مورد قالب آنها نداریم و آنها را به صورت مجموعه‌ای مرتب از جملات در نظر می‌گیریم.

در ابتدا بایستی همه‌ی نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی گردند. در پردازش رسم‌الخط زبان فارسی، با توجه به قرابتی که با رسم‌الخط عربی دارد، همواره در تعدادی از حرف‌ها مشکل وجود دارد که از جمله آن‌ها می‌توان به حروف "ک"، "ی"، همزه و ... اشاره نمود. در اولین گام باید مشکلات مربوط به این حروف را برطرف ساخت. علاوه بر این، اصلاح و یکسان‌سازی نویسه‌ی نیم‌فاصله و فاصله در کاربردهای مختلف آن و همچنین حذف نویسه‌ی «ـ» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرد و مواردی مشابه برای یکسان‌سازی متون، از اقدامات لازم قبل از شروع فازهای بعدی می‌باشد. در این فاز مطابق با یک سری قاعده‌ی دقیق و مشخص، فاصله‌ها و نیم‌فاصله‌های موجود در متن برای علاماتی نظیر "ها" و "ی" غیرچسبان در انتهای لغات و همچنین پیشوندها و پسوندهای فعل‌ساز نظیر "می"، "ام"، "ایم"، "اید" و موارد مشابه جهت استفاده در فازهای بعدی، اصلاح می‌گردند. در ضمیمه (الف) به چند نمونه از این اصلاحات، اشاره شده است.

۵-۲-۲-۳- جداکننده جملات (Sentence Splitter)

پس از پایان مرحله‌ی نرمال سازی متون، ابزار تشخیص‌دهنده‌ی جملات با استفاده از علامت‌های ".“، ”؛“، ”!“، ”؟“، ”?“، ”:“ و بکارگیری برخی دستورات گرامری زبان فارسی و در نظرگرفتن برخی لغات آغاز کننده‌ی جملات، مرز جمله‌ها را برای استفاده در گام‌های بعدی تعیین می‌نماید.

باید توجه داشت که جستجو برای یافتن علائم جداکننده‌ی جملات به تنهایی کافی نیست و در بسیاری موارد مشکلاتی را به همراه دارد. به عنوان مثال، در زبان فارسی بعضی از واژه‌های اختصاری به صورت چند حرف که با نقطه از هم جدا شده‌اند، ظاهر می‌شوند (مانند "ه.ق" که مخفف "هجری قمری" است). در صورتی که

تعداد این حالات مشکل‌ساز در متن مورد نظر زیاد باشد، باید از روش‌های پیچیده‌تری نظیر بکارگیری برخی دستورات گرامری زبان فارسی و در نظرگرفتن برخی لغات آغاز کننده‌ی جملات استفاده نمود که قادر به شناسایی چنین مواردی باشد.

این ابزار بایستی توانایی تشخیص صحیح و دقیق جملات را در متن ورودی داشته باشد. برای ایجاد این ابزار باید ابتدا تمامی کاراکترها، نمادها و احیاناً قواعد دستوری که باعث شکسته شدن جملات می‌شوند، شناسایی گردند. با توجه به پایه بودن جمله در بسیاری از پردازش‌های زبانی، خروجی دقیق این ابزار از درجه‌ی اهمیت بالایی برخوردار است.

۶-۲-۲-۳- جداکننده کلمات (Tokenizer)

تشخیص‌دهنده‌ی لغات نیز با استفاده از علامت‌های فضای خالی، "،"، ","، "-" و در نظر گرفتن اصلاحات اعمال شده در مورد پیشوندها و پسوندها در فاز قبلی، واژه‌ها را شناسایی می‌نماید.

لازمه‌ی ایجاد این ابزار، جمع‌آوری واحدهایی است که در زبان به عنوان واحدهای مستقل معنایی شناخته می‌شوند. سپس بر اساس انتخاب هر کدام از این واحدها، متن بر اساس آن، شکسته خواهد شد.

۷-۲-۲-۳- ریشه‌یاب (Stemmer)

ریشه‌یابی عبارتست از بدست آوردن ریشه کلمات با حذف پسوندها و پیشوندها بطوری که کلمات با ریشه یکسان دارای شکل یکسان گردند. در زبان‌هایی مثل زبان انگلیسی به دلیل ماهیت زبان و شباهت ریشه با کلمه اصلی، انجام عمل ریشه‌یابی کار نسبتاً آسانی است. اما در برخی زبان‌های دیگر مانند زبان فارسی به دلیل ماهیت تصریفی[۱] زبان، عملیات ریشه‌یابی، فرآیند پیچیده‌ای دارد.[۲] [TAS02][FAR06] [TAG05]

در راستای انجام پایان‌نامه از آخرین نسخه‌ی ریشه‌یاب تولید شده در آزمایشگاه فناوری وب دانشگاه فردوسی که یک ریشه‌یاب ترکیبی مبتنی بر قواعد و فرهنگ لغت می‌باشد، استفاده گردید.

۸-۲-۲-۳- برچسب‌زننده اجزای واژگانی کلام (POS)

تشخیص هر کدام از لغات بکار رفته در متن به عنوان *فعل*، *اسم*، *قید*، *صفت* و غیره بر عهده برچسب‌زن نحوی می‌باشد. در واقع برچسب‌زن نحوی علاوه بر تعیین مقوله‌های فوق، می‌تواند *اسم مکان*، *اسم زمان*، *اسامی*

---

۱ inflective

۲ - به زبان‌هایی مانند زبان فارسی که شخص و زمان بر روی کلمات تاثیر می‌گذارند، زبان‌های تصریفی گفته می‌شود.

خاص، *ضمایر*، *نوع صفات* و موارد مشابه را تعیین نماید.
در راستای انجام پایان‌نامه از آخرین نسخه‌ی ابزار برچسب‌زنی نحوی متون فارسی تولید شده در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد استفاده گردید.

۳-۲-۳- تشریح روش پیشنهادی

ابزاری که در راستای انجام این پایان‌نامه طراحی و پیاده‌سازی گردیده است، اولین ابزاری است که در زمینه‌ی برچسب‌زنی نقش‌های معنایی متون فارسی ارائه شده است که در ادامه جزئیات پیاده‌سازی آن، تشریح خواهد گردید.

همانطور که گفته شد در روش بکارگرفته شده، پس از پایان مرحله پیش‌پردازش و یکسان‌سازی متون، ابزارهای *تشخیص دهنده‌ی جملات*، *تشخیص دهنده‌ی کلمات* ، *ریشه‌یاب* و همچنین *NER* و *POS* بایستی بر روی متون اعمال شوند. تشخیص دهنده‌ی جملات با استفاده از علامت‌های "."، "؛"، "!"، "؟"، "?"، ":" و بکارگیری برخی قواعد گرامری زبان فارسی، مرز جمله‌ها را برای استفاده در گام‌های بعدی تعیین می‌نماید. تشخیص دهنده‌ی کلمات با استفاده از علامت‌هایی نظیر فضای خالی، "،"، "," و "-" واژه‌ها را شناسایی می‌نماید. NER اسامی خاص نظیر اسامی افراد، اماکن، مقادیر عددی و ... را با بهره‌گیری از یک لغت‌نامه تشخیص می‌دهد. ابزار برچسب‌زنی نحوی متون نیز که در خود ابزار تشخیص عبارت فعلی را نیز جای داده است، عبارت فعلی جمله را به طور کامل شناسایی کرده و همچنین مهم‌ترین بخش‌های سخن نظیر اسم، ضمیر، صفت، قید و حرف اضافه را تشخیص می‌دهد. ریشه‌یاب نیز ریشه‌ی واژه‌های شناسایی شده از مرحله‌ی قبل را بدست می‌آورد.

پس از آن از روی عبارت فعلی تشخیص داده شده توسط ابزار برچسب‌زنی نحوی متون، بن ماضی فعل و همچنین سایر اجزای فعل نظیر پیشوند فعل، فعل‌یار و حرف اضافه‌ی فعلی در صورت وجود تشخیص داده می‌شوند. با مراجعه به فرهنگ لغتی که حاوی بن ماضی و مضارع افعال است و از پیکره‌ی ظرفیت نحوی افعال زبان فارسی تولید کرده‌ایم، بن مضارع فعل را نیز یافته و عبارتی مطابق با ساختار کلید فریم‌ها یعنی به صورت زیر تشکیل می‌دهیم. در صورتیکه هر کدام از موارد پیشوند فعل، فعل‌یار و حرف اضافه‌ی فعلی در عبارت فعلی وجود نداشته و یا تشخیص داده نشود، عبارت '-' به جای آن قرار می‌گیرد.

بن-ماضی#بن-مضارع#پیشوند#فعلیار#حرف-اضافهٔ-فعلی

در ادامه کلمات تشکیل دهنده‌ی جمله را مورد بررسی قرار داده و به جز حروف اضافه و فعل‌یارها، سایر کلمات بکار رفته در جمله را با لغت "کلمه" جایگزین می‌کنیم تا با عبارت منظم موجود در فریم‌ها تطبیق پیدا کند. همچنین در ساختار عبارت فعلی به جای فعل در زمان‌های مختلف بکار رفته در جمله، فقط بن

ماضی فعل قرار می‌گیرد.

پس از ساخت کلید اجزای فعل و عبارت منظم بیانگر ساختار جمله، از روی کلید- عبارت فعلی تولیدی، به جستجوی ردیف-فریم کلید در فریم‌نت که در قالب یک ساختار درهم قابل دستیابی است، می‌پردازیم. جدول درهمی که کلید آن، اجزای فعل به صورت فوق و مقدار آن، لیستی از فریم‌های تولیدی با توجه به ظرفیت فعل می‌باشد.

پس از یافتن ردیف- فریم فعل بکار رفته در جمله‌ی ورودی در صورت وجود یک فریم در لیست فریم‌های آن فعل، آن فریم با جمله ورودی تطبیق یافته و برچسب‌های معنایی موجود در آن فریم به کلمات جمله، تخصیص می‌یابد و درصورتیکه چندین فریم در لیست فریم‌های آن فعل، موجود باشد؛ در بین عبارت منظم تولیدی جهت تطبیق جملات ورودی (یعنی بخش اول فریم) جستجو نموده و هر عبارت منظمی که با عبارت منظم تولید شده از جمله ورودی تطبیق پیدا کند، برچسب‌های معنایی موجود در آن فریم به کلمات جمله، تخصیص می‌یابد. بدین صورت که به ازای هر کدام از لغات "کلمه" در عبارت منظم بخش اول فریم، برچسب معنایی متناظر با آن از بخش دوم فریم تخصیص می‌یابد.

۳-۳- روال پیاده‌سازی فریم‌نت و SRL زبان فارسی

برای پیاده‌سازی فریم‌نت و ابزار برچسب‌زنی نقش‌های معنایی جملات متون فارسی از زبان Visual C# استفاده گردید و پیکره‌ها و فایل‌های مورد نیاز به صورت فایل متنی با پسوند txt. ذخیره گردیده و در برنامه مورد استفاده قرار می‌گیرند.

در روال اجرای برنامه ابتدا روال استخراج فریم‌ها و تولید فریم‌نت و همچنین تشکیل دو فرهنگ لغت یکی حاوی حروف اضافه و دیگری حاوی بن ماضی و بن مضارع افعال از پیکره ظرفیت نحوی افعال زبان فارسی اجرا می‌گردد و نتایج در قالب چند فایل متنی ذخیره می‌گردد.

در ابزار طراحی شده گزینه‌های زیر قرار داده شده است.

- **ورودی از فایل متنی:** می‌توان متن ورودی را از یک فایل متنی وارد نمود.
- **NER:** برای شناسایی اسامی خاص و همچنین برخی عبارت خاص نظیر اعداد، زمان‌ها و واحدها و ... مورد استفاده قرار می‌گیرد.
- **POSTagger:** برای برچسب‌زنی نحوی متون قابل استفاده است.
- **Lematizer:** کلمات موجود در متن ورودی را ریشه‌یابی می‌کند.
- **SRL:** با بهره‌گیری از فریم‌نت و روش پیشنهادی ذکرشده به شناسایی برچسب‌های معنایی در متن ورودی می‌پردازد.
- **بستن پنجره:** برای خروج از محیط نرم‌افزار بکار می‌رود.

خروجی نرم افزار بدین صورت نمایش داده می‌شود که به ازای هر کلمه، ریشه، برچسب NER، برچسب POS و برچسب SRL به صورت زیر پس از کلمه ذکر می گردند.

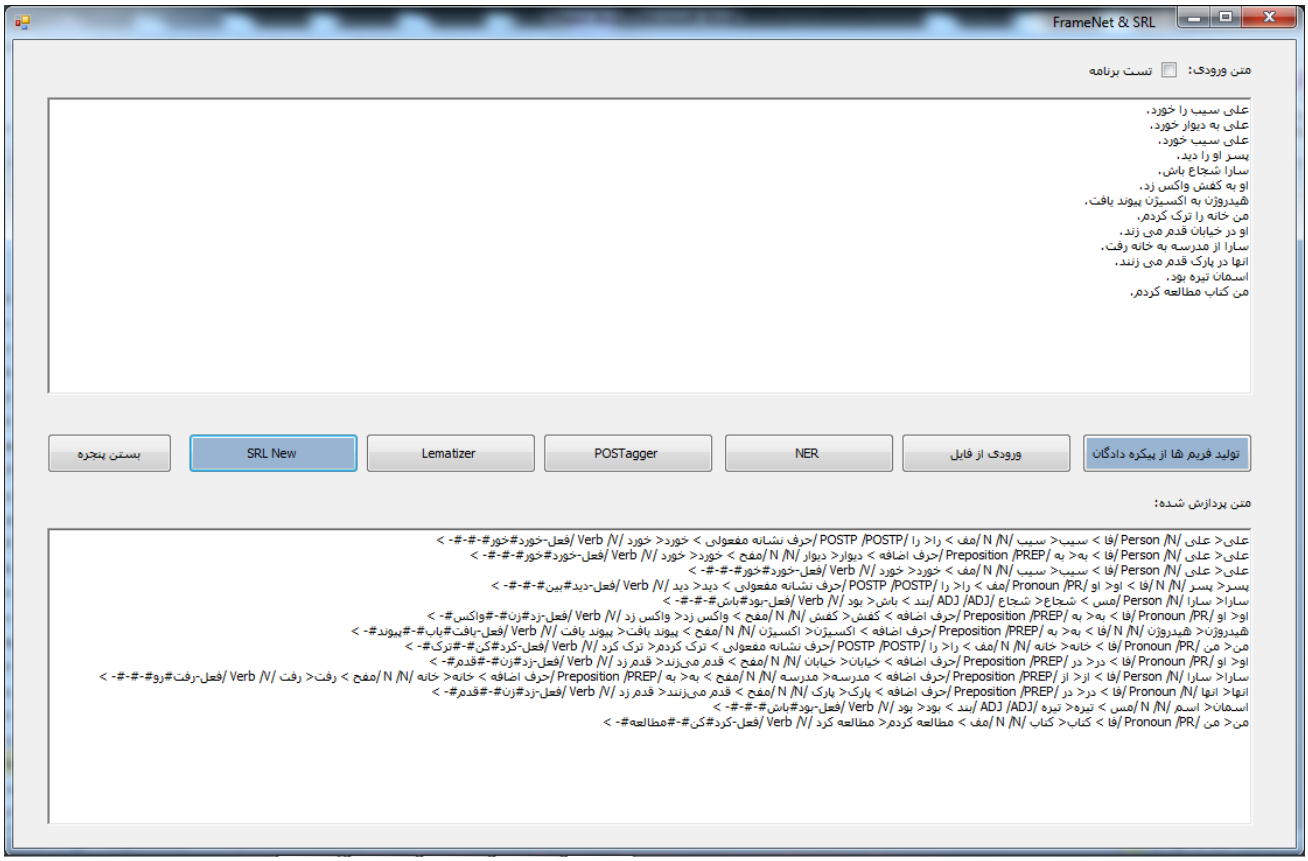
< ریشه / برچسبNER/ برچسب POS / برچسب SRL >

در مورد عبارت فعلی نیز برچسب فعل به صورت زیر خواهد بود:

< ریشه/برچسبNER/برچسب POS/برچسبSRL-بن ماضی#بن مضارع#پیشوند#فعلیار#حرف اضافهٔ فعلی >

۴-۳- شمای گرافیکی ابزار طراحی شده

شمای گرافیکی ابزار طراحی شده جهت تولید فریم‌نت و برچسب‌زنی معنایی متون فارسی در زیر آورده شده است:



۵-۳- چند نمونه جمله‌ی ورودی و خروجی ابزار

چند نمونه جمله‌ی ورودی داده شده به ابزار و همچنین خروجی تولید شده توسط ابـزار در ادامـه آورده شـده است:

جمله ورودی ۱:

علی سیب خورد.

خروجی ۱:

علی< علی /N/ Person /فا > سیب< سیب /N/ N /مف > خورد< خورد /V/ Verb /فعل-خورد#خور#-#-#- >

جمله ورودی ۲:

علی سیب را خورد.

خروجی ۲:

علی< علی /N/ Person /فا > سیب< سیب /N/ N /مف > را< را /POSTP/ POSTP /حرف نشانه مفعولی > خورد< خورد /V/ Verb /فعل-خورد#خور#-#-#- >

جمله ورودی ۳:

علی به دیوار خورد.

خروجی ۳:

علی< علی /N/ Person /فا > به< به /PREP/ Preposition /حرف اضافه > دیوار< دیوار /N/ N /مفح > خورد< خورد /V/ Verb /فعل-خورد#خور#-#-#- >

جمله ورودی ۴:

علی زمین خورد.

خروجی ۴:

علی< علی /N/ Person /فا > زمین خورد< زمین خورد /V/ Verb /فعل-خورد#خور#-#زمین#- >

جمله ورودی ۵:

سارا شجاع باش.

خروجی ۵:

سارا< سارا /N/ Person /مس > شجاع< شجاع /ADJ/ ADJ /بند > باش< بود /V/ Verb /فعل-بود#باش#-#-#- >

جمله ورودی ۶:

او به کفش واکس زد.

خروجی ۶:

او< او /PR/ Pronoun /فا > به< به /PREP/ Preposition /حرف اضافه > کفش< کفش /N/ N /مفح > واکس زد< واکس زد /V/ Verb /فعل-زد#زن#-#واکس#- >

جمله ورودی ۷:

هیدروژن به اکسیژن پیوند یافت.

خروجی ۷:

هیدروژن< هیدروژن /N/ N /فا > به< به /PREP/ Preposition /حرف اضافه > اکسیژن< اکسیژن /N/ N /مفح > پیوند یافت< پیوند یافت /V/ Verb /فعل-یافت#یاب#-#پیوند#- >

جمله ورودی ۸:

من خانه را ترک کردم.

خروجی ۸:

من< من /Pronoun /PR/فا > خانه< خانه /N /N/مف > را< را /POSTP /POSTP/حرف نشانه مفعولی > ترک کردم< ترک کرد /Verb /V/فعل-کرد#کن#-#ترک#- >

جمله ورودی ۹:

او در خیابان قدم می زند.

خروجی ۹:

او< او /Pronoun /PR/فا > در< در /Preposition /PREP/حرف اضافه > خیابان< خیابان /N /N/مفح > قدم می‌زند< قدم زد /Verb /V/فعل-زد#زن#-#قدم#- >

جمله ورودی ۱۰:

سارا از مدرسه به خانه رفت.

خروجی ۱۰:

سارا< سارا /Person /N/فا > از< از /Preposition /PREP/حرف اضافه > مدرسه< مدرسه /N /N/مفح > به< به /Preposition /PREP/حرف اضافه > خانه< خانه /N /N/مفح > رفت< رفت /Verb /V/فعل-رفت#رو#-#-#- >

جمله ورودی ۱۱:

انها در پارک قدم می زنند.

خروجی ۱۱:

انها< انها /Pronoun /N/فا > در< در /Preposition /PREP/حرف اضافه > پارک< پارک /N /N/مفح > قدم می‌زنند< قدم زد /Verb /V/فعل-زد#زن#-#قدم#- >

جمله ورودی ۱۲:

آسمان تیره بود.

خروجی ۱۲:

آسمان< آسمان /N/ Person /مس > تیره< تیره /ADJ/ ADJ /بند > بود< بود /V/ Verb /فعل-بود#باش#-#-#- >

جمله ورودی ۱۳:

من کتاب مطالعه کردم.

خروجی ۱۳:

من< من /PR/ Pronoun /فا > کتاب< کتاب /N/ N /مف > مطالعه کردم< مطالعه کرد /V/ Verb /فعل-کرد#کن#-#مطالعه#- >

۶-۳- خلاصه

در این فصل، روش پیشنهادی برای تولید فریم‌نت و ابزار برچسب‌زنی معنایی متون فارسی توضیح داده شد. در بخش اول این فصل، روال تولید فریم‌نت از روی پیکره ظرفیت نحوی افعال زبان فارسی تشریح شد. در بخش بعد، چگونگی بهره‌گیری از ابزارهای مورد نیاز برای برچسب‌زنی معنایی متون نظیر استانداردکننده‌ی متون، تشخیص‌دهنده جملات، تشخیص‌دهنده لغات، NER، ابزار برچسب‌زن بخش‌های کلام زبان فارسی و ریشه‌یاب بیان گردیده و در نهایت چگونگی طراحی و پیاده‌سازی ابزار برچسب‌زنی معنایی با استفاده از فریم-نت زبان فارسی تشریح گردید.

در بخش سوم روال پیاده‌سازی ابزار توضیح داده شده و در نهایت شمای گرافیکی ابزار و همچنین چند نمونه جمله‌ی داده شده به ابزار و نتایج خروجی ابزار، آورده شده است.

فصل چهارم:

نتیجه گیری و کارهای

آتی

۴- نتیجه‌گیری و کارهای آتی

۴-۱- نتیجه‌گیری

همانطور که بیان شد یکی از مشکلات اساسی در پردازش متون زبان فارسی، عدم وجود ابزارهای پایه‌ای استاندارد جهت انجام روال پیش‌پردازشی و تجزیه و تحلیل متون می‌باشد. برچسب‌زنی معنایی متون فارسی یکی از حوزه‌هایی است که تاکنون در زبان فارسی بسیار غریب و مهجور مانده و به دلیل مشکلات و دشواری-هایی که در طراحی این ابزار وجود دارد، کسی سراغ این حوزه نرفته و یا تلاش‌های موجود در این حوزه ناکام مانده است.

در راستای انجام این پایان‌نامه با بهره‌گیری از پیکره‌ی ظرفیت نحوی افعال زبان فارسی، نخستین فریم‌نت (FrameNet) زبان فارسی تولید گردید. همچنین با استفاده از ابزارهای تشخیص فعل در جملات، برچسب-زن نحوی جملات و همچنین ابزار ریشه‌یاب آزمایشگاه فناوری وب دانشگاه فردوسی مشهد، ابزاری جهت برچسب‌زنی معنایی جملات زبان فارسی طراحی نمودیم که با دقت قابل قبولی به تطبیق جملات یک متن با فریم‌های موجود در پیکره تولید شده پرداخته و برچسب معنایی متناظر با آن جمله را به کلمات تشکیل دهنده‌ی جمله اختصاص می‌دهد.

تخصیص نقش‌های معنایی کلمات در جمله در لایه‌ی اول شامل فاعل، مفعول، مسند، قید، بند و ... و همچنین در لایه‌ی دوم با توجه به نوع فعل مثلا در مورد فعل "قتل" شامل قاتل، مقتول، آلت قتل، دلیل قتل و ...، یکی از مهم‌ترین ابزارهای پیش‌پردازشی برای اکثر کاربردهای پردازش متن مانند خلاصه‌سازی، ترجمه‌ی ماشینی، استخراج مفاهیم، سیستم‌های مکالمه‌ی معنایی، سیستم‌های پاسخگو، متن‌کاوی و ... می‌باشد.

SRL با تعیین نقش معنایی هر عنصر جمله در ارتباط با فعل آن جمله، نقش گرامری آن عنصر از قبیل فاعل بودن، مفعول بودن چه مستقیم و چه غیر مستقیم، انواع قیود و ... را تشخیص می‌دهد.

در شبکه‌ی قالب‌های تولیدی برای افعال جملات زبان فارسی در مجموع ۴۵۹۳ ساختار برای افعال وجـود دارد که در هر ساختار، چندین فریم وجود دارد که به کلمات آن، برچسب‌های معنایی تخصیص یافته‌اند.

۴-۲- کارهای آتی

همانطور که ذکر گردید برچسب‌زنی معنایی کلمات مشابه برچسب‌گذاری بخش‌های سخن بوده با این تفاوت که وارد عمق بیشتری از معنا می‌شود. نقش‌های معنایی رابطه اجزاء مختلف جمله را در ارتباط با فعل متناظرشان معرفی می‌نمایند. در حال حاضر ابزار برچسب‌زنی معنایی پیاده‌سازی شده در راستای انجام این پایان‌نامه، به استخراج نقش‌های معنایی جملات نظیر فاعل، مفعول مستقیم، مفعول غیر مستقیم، قید، مسند، بند، فعل و ... می‌پردازد.

با توجه به اینکه ابزار طراحی شده جهت برچسب‌زنی معنایی کلمات در این پایان‌نامه به کمک فریم‌نت به تخصیص برچسب معنایی به کلمات می‌پردازد، می‌توان بخش فریم‌های فریم‌نت را توسعه داده و علاوه بر نقش‌های فاعل، مفعول مستقیم، مفعول غیر مستقیم، قید، مسند، بند و ... با توجه به نوع فعل به بیان نقش-های مرتبط با فعل پرداخته و فریم‌ها را در سه بخش عبارت منظم جهت تطبیق با جمله ورودی، برچسب‌های معنایی عمومی و برچسب‌های اختصاصی مرتبط با هر فعل تولید نمود. مثلا در مورد فعل "قتل" می‌توان برچسب‌های قاتل، مقتول، آلت قتل، دلیل قتل و ... را به بخش‌های مختلف جمله تخصیص داد.

لازمه‌ی این امر، آنست که فریم‌نت توسط یک فرد خبره در رشته ادبیات و زبانشناسی مورد بررسی قرار گرفته و با توجه به نوع فعل، برچسب‌های اختصاصی هر فعل به نقش‌های معنایی جمله‌ی متناظر با آن فعل تخصیص یابد.

از جمله کارهای دیگری که می‌تواند باعث دقت ابزار برچسب‌زنی معنایی متون گردد و بسیار ضروری و حیاتی نیز می‌باشد طراحی و پیاده‌سازی یک پارسر با دقت بسیار بالا برای تشخیص گروه‌های تشکیل دهنده جملات متن می‌باشد. در نسخه فعلی ابزار پیاده‌سازی شده، گروه‌های تشکیل دهنده‌ی جملات، تک کلمه‌ای در نظر گرفته شده‌اند؛ چرا که برای شناسایی گروه‌های تشکیل‌دهنده‌ی جملات متون فارسی نیاز به ابزار پارسر زبان فارسی با دقت بسیار بالا می‌باشد.

در صورت طراحی ابزاری با دقت بالا برای تشخیص گروه‌های تشکیل دهنده جملات می‌توان به جای عبارت "کلمه" در عبارات منظم، عبارت "گروه" را قرار داد که منطبق با گروه‌های اسمی، فعلی، قیدی، صفتی و حرف اضافه‌ای می‌گردد و برچسب‌زنی معنایی جملات متن به صورت کامل و برای تمامی جملات صورت خواهد گرفت.

منابع

Abstract*:*

Keywords:

Department of Computer Engineering

Ferdowsi University of Mashhad

M.Sc. Thesis

Implementation of Semantic Role Labeling Tool With using a Persian FrameNet

By:

Najmeh Noori

Supervisor:

Dr. Mohsen Kahani

January 2014