دانشکده مهندسی کامپیوتر

ساز و کاری برای کشف تأثیر ویژگی‌های مختلف ساخت‌واژی و صرفی بر روی تجزیهٔ وابستگی زبان فارسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر - گرایش هوش مصنوعی و رباتیک

مجتبی خلاش

استاد راهنما:

دکتر بهروز مینایی بیدگلی

آبان ماه ۱۳۹۱

دانشکده مهندسی کامپیوتر

ساز و کاری برای کشف تأثیر ویژگی‌های مختلف ساخت‌واژی و صرفی بر روی تجزیهٔ وابستگی زبان فارسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیک

مجتبی خلاش

استاد راهنما:

دکتر بهروز مینایی بیدگلی

آبان ماه ۱۳۹۱


بسم الله الرحمن الرحيم

تأییدیه هیأت داوران جلسه دفاع از پایان‌نامه

نام دانشکده: مهندسی کامپیوتر

نام دانشجو: مجتبی خلاش

عنوان پایان‌نامه: ساز و کاری برای کشف تأثیر ویژگی‌های مختلف ساخت‌واژی و صرفی بر روی تجزیهٔ وابستگی زبان فارسی

تاریخ دفاع: آبان ماه ۱۳۹۱

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی و رباتیک

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضاء
۱	استاد راهنما	بهروز مینایی بیدگلی	استادیار	دانشگاه علم و صنعت ایران	
۲	استاد مدعو خارجی	هشام فیلی	استادیار	دانشگاه تهران	
۳	استاد مدعو داخلی	مرتضی آنالویی	استادیار	دانشگاه علم و صنعت ایران	

تأییدیه صحت و اصالت نتایج

باسمه تعالی

اینجانب مجتبی خلاش به شماره دانشجویی ۸۹۷۲۲۲۳۶ دانشجوی رشته کامپیوتر، گرایش هوش مصنوعی و رباتیک در مقطع تحصیلی کارشناسی ارشد تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض در خصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسئولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده اینجانب خواهد بود و دانشگاه هیچ‌گونه مسئولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: مجتبی خلاش

امضا و تاریخ:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

نام استاد یا اساتید راهنما:

تاریخ:

امضا:

با تشکر از:

آقای محمد صادق رسولی که کمک فراوانی در پیشبرد این پژوهش داشتند.
آقای دکتر حسن اصغریان و مهندس علی هادیان که نهایت همکاری را در اجرای آزمایش‌ها داشتند.
و تشکر ویژه از همه اعضای خانواده

چکیده

سامانه‌های مبتنی بر داده به راحتی می‌توانند به سایر زبان‌ها یا دامنه‌ها منتقل شود. به همین دلیل در تجزیهٔ وابستگی نیز اقبال به روش‌های «مبتنی بر داده» بیش از روش‌های «مبتنی بر دستور» بوده است. تنها پیش‌نیاز این روش‌ها وجود پیکرهٔ وابستگی شامل جملات و درخت وابستگی متناظر با آن است که تهیهٔ آن امری پرهزینه و زمان‌بر است. با این وجود تاکنون برای حدود ۳۰ زبان پیکرهٔ وابستگی تهیه شده که زبان فارسی نیز جزو این دسته از زبان‌هاست.

با وجود صحت بالای تجزیهٔ وابستگی در زبان انگلیسی، اعمال الگوریتم‌های موجود بر روی دسته‌ای از زبان‌ها اغلب منجر به افت صحت می‌شود که دلیل این امر را می‌توان در پررنگ‌تر بودن عامل بی‌ترتیبی و غنی بودن ساخت‌واژه‌ها در زبان مقصد نسبت به زبان انگلیسی دانست. این بدان معناست که سامانه‌های مبتنی بر داده نیازمند انتخاب خصوصیات و تنظیم دقیق پارامترها به منظور رسیدن به کارایی بهینه هستند. این امر کار پیچیده‌ای است که نیازمند دانش خاص از سامانه و خواص زبان مقصد است.

در این پایان‌نامه ابتدا مروری بر تلاش‌های انجام‌شده برای برخورد با این مسئله در سایر زبان‌ها خواهیم پرداخت و پس از ترسیم روند کاری، سعی شده تا این روند بر روی زبان فارسی انجام و عوامل تأثیرگذار بر کاهش صحت تجزیه شناسایی شوند. سپس مجموعه‌ای از ۱۰ خصوصیت ساخت‌واژی و مفهومی مورد بررسی قرار گرفته و ضمن بررسی تأثیر هر خصوصیت به صورت جداگانه، بهترین ترکیب این خصوصیات ارائه شده است.

واژه‌های کلیدی: تجزیهٔ وابستگی، خصوصیات ساخت‌واژی و صرفی، پیکرهٔ وابستگی.

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

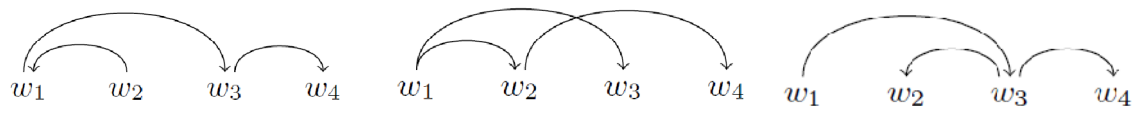
فصل ۱:

مقدمه

۱-۱- شرح مسئله

دستور وابستگی یکی از مکتب‌های دستورنویسی است که هدف آن توصیف ساخت‌های نحوی در زبان‌های گوناگون است. هر ساخت نحوی به صورت رابطهٔ وابستگی بین عناصر هسته و وابسته توصیف شده که در مجموع یک درخت وابستگی به دست می‌آید. فعل جمله در ریشهٔ درخت قرار دارد، وابسته‌ها و وابستهٔ وابسته‌های فعل سایر اجزای درخت را شکل می‌دهند. درخت‌های وابستگی به دو دستهٔ افکنشی[1] و غیرافکنشی تقسیم می‌شوند که در شکل (۱-۱) نشان داده شده است. تفاوت درخت‌های غیرافکنشی با درخت‌های افکنشی وجود یال‌هایی است که یکدیگر را قطع می‌کنند [۱].

تلاش‌هایی برای ارائهٔ ساختاری با قدرت بیان بیشتر از این دو ساختار صورت گرفته است که نمونه‌ای از این تلاش‌ها، مفهوم ساختار مسطح[2] است. در این ساختار، مشابه درخت افکنشی یال‌های متقاطع نباید وجود داشته باشد اما گره مصنوعی ریشه (که در مکان ω_0 اضافه می‌شود) در ساختار مسطح در نظر گرفته نمی‌شود [۲]. نکتهٔ اصلی در این مفهوم جدید قابلیت تعمیم آن به ساختار m-مسطح است که در آن ساختار وابستگی غیرافکنشی را می‌توان توسط حداقل m ساختار مسطح بازنمایی کرد. دو بازنمایی ۲-مسطح از ساختار غیرافکنشی (ب) شکل (۱-۱) در شکل (۱-۲) نشان داده شده است که ساختار مسطح اول توسط یال‌های توپر و ساختار مسطح دوم توسط یال‌های خط تیره مشخص شده است. در مرجع [۳] نشان داده شده است که در اکثر پیکره‌های وابستگی، ۹۹ درصد ساختارها ۲-مسطح و تقریباً کل ساختارها ۳-مسطح هستند. مسئلهٔ یافتن ساختار m-مسطح قابل کاهش به مسئلهٔ رنگ آمیزی گراف است: «گراف G از نوع m-مسطح است اگر هر کدام از یال‌های آن را بتوان به یکی از m رنگ منتسب کرد که در آن یال‌های با رنگ یکسان یکدیگر را قطع نکنند». این مسئله برای مقادیر m بزرگ‌تر از دو، غیرقطعی کامل[3] خواهد بود.



الف) ساختار وابستگی افکنشی ب) ساختار وابستگی غیرافکنشی ج) ساختار مسطح

شکل (۱-۱) انواع درخت‌های وابستگی [۲]

[1] Projective
[2] Planar
[3] NP-Complete

شکل (۱-۲) دو حالت بازنمایی ساختار ۲-مسطح برای ساختار غیرافکنشی (ب) در شکل (۱-۱)

تجزیهٔ وابستگی راهکاری برای تحلیل نحوی است که از دستور وابستگی الهام گرفته شده است. در حوزهٔ زبان‌شناسی رایانه‌ای روش‌های ارائه‌شده برای استنتاج دستور بر مبنای دستور وابستگی را به دو دستهٔ «مبتنی بر داده» و «مبتنی بر دستور» تقسیم می‌کنند. از این میان روش‌های مبتنی بر داده به دلیل ماهیت مستقل از زبان مورد اقبال بیشتری قرار گرفته‌اند. الگوریتم‌های این دسته از روش‌ها، برای اجرا نیازمند دادهٔ آموزشی نمادگذاری شده هستند و با کسب اطلاعات آماری به دست آمده از آن قادر به تجزیهٔ جملات خواهند بود. روش‌های مبتنی بر داده خود به دو دستهٔ «مبتنی بر گذار» و «مبتنی بر گراف» تقسیم می‌شوند:

- روش‌های مبتنی بر گذار: این راهکارها با تعریف یک سامانهٔ گذار یا ماشین حالت برای نگاشت جمله به درخت وابستگی شروع می‌شود. مسئلهٔ یادگیری معادل استنتاج الگو برای پیش‌بینی گذار بعدی بر اساس تاریخچهٔ گذار است و مسئلهٔ تجزیه معادل ساخت رشته‌ای از گذارهای بهینه برای جملهٔ ورودی توسط الگوی به دست آمده است. به روش‌های این دسته اصطلاحاً «تجزیهٔ وابستگی جابه‌جایی-کاهش»[1] می‌گویند.
- راهکارهای مبتنی بر گراف: در این راهکارها فضایی از گراف‌های وابستگی نامزد برای جمله تعریف می‌شود. مسئلهٔ یادگیری معادل ارائهٔ الگو برای انتساب امتیاز به گراف‌های وابستگی نامزد جمله است و مسئلهٔ تجزیه معادل یافت گراف وابستگی با بیش‌ترین امتیاز برای جملهٔ ورودی توسط الگو است. به روش‌های این دسته اصطلاحاً «تجزیهٔ درخت پوشای بیشینه»[2] گویند.

اکثر این روش‌ها با وجود مستقل از زبان بودن با فرض داده‌های زبان انگلیسی طراحی شدند و اعمال آن‌ها به سایر زبان‌ها اغلب منجر به کاهش صحت[3] می‌شود [۴]. در مورد این زبان‌ها تلاش شده اهمیت نسبی خصوصیات زبان‌شناختی مختلف برای تجزیهٔ وابستگی مبتنی بر داده بررسی شود.

[1] Shift-reduce dependency parse
[2] Maximum spanning tree parser
[3] Accuracy

۱-۲- انگیزه‌های پژوهش

پیکرهٔ وابستگی برای زبان فارسی به تازگی انتشار یافته [۵] و به زودی تحقیقات بر روی تجزیهٔ وابستگی فارسی آغاز خواهد شد. اولین گام در اجرای الگوریتم‌های تجزیهٔ وابستگی مبتنی بر داده در هر زبانی یافتن خصوصیات و تنظیم پارامترهاست. انگیزهٔ اصلی این پژوهش بررسی و ارائهٔ این خصوصیات برای زبان فارسی توسط دو تجزیه‌گر «MaltParser» و «MSTParser» که به ترتیب نمایندهٔ روش‌های مبتنی بر گذار و مبتنی بر گراف هستند.

۱-۳- ساختار پایان‌نامه

در فصل دوم ابتدا تعریفی از ساخت‌واژه و زبان‌هایی که از نظر ساخت‌واژی غنی هستند ارائه خواهد شد و پس از آن به طور خاص زبان فارسی را از منظر ساخت‌واژی و صرفی مورد بررسی قرار خواهیم داد. در پایان این فصل پیکرهٔ وابستگی زبان فارسی که مبنای تمام آزمایش‌های این پژوهش است، معرفی خواهد شد. برای بررسی کارهای پیشین در این حوزه، ابتدا فصل سوم را با توصیف روند طبقه‌بندی سطوح دشواری زبان‌ها و گرایش به بررسی زبان‌های از نظر ساخت‌واژی غنی آغاز کرده؛ سپس ضمن ترسیم چارچوب کلی برای بررسی این دسته از زبان‌ها، تلاش‌های انجام شده در سایر زبان‌ها مورد بررسی قرار گرفته است. چارچوب ارائه شده در این فصل مبنای بررسی‌های انجام شده بر روی زبان فارسی در این پژوهش است و بر همین اساس در فصل چهارم مراحل و ابزارهای لازم برای اجرای این چارچوب معرفی خواهد شد. در فصل پنجم به صورت مجزا نتایج آزمایش‌ها ارائه شده و به ارزیابی آن‌ها خواهیم پرداخت. نتایج ارائه‌شده در این فصل عوامل مؤثر در افت صحت تجزیه را آشکار می‌سازد. همچنین ضمن ارائهٔ مجموعه‌ای از خصوصیات، بهترین ترکیب آن‌ها برای ارائه به تجزیه‌گر معرفی خواهد شد. سرانجام در فصل پنجم به جمع‌بندی نتایج پژوهش پرداخته و پیشنهادهایی برای ادامهٔ کار تجزیهٔ وابستگی زبان فارسی ارائه خواهد شد.

فصل ۲:

تعاریف و مفاهیم مبنایی

۲-۱- مقدمه

در این فصل ابتدا تعریفی از ساخت‌واژه ارائه خواهد شد و سپس به معرفی زبان‌های از نظر ساخت‌واژی غنی[1] خواهیم پرداخت. به این دلیل که زبان فارسی جزو این دسته از زبان‌هاست، خصوصیات ساخت‌واژی و صرفی زبان فارسی بررسی می‌شود. برای بررسی خصوصیات ساخت‌واژی، مقوله‌های دستوری تعریف شده و مقادیر مجاز هر کدام از این مقوله‌ها در زبان فارسی معرفی خواهد شد. برای بررسی خصوصیات صرفی بی‌ترتیبی و ساخت‌های ظرفیتی مورد بررسی قرار می‌گیرند. در پایان این فصل پیکرهٔ وابستگی زبان فارسی معرفی شده و اطلاعات آماری و خصوصیات موجود در این پیکره توصیف خواهد شد.

۲-۲- تعریف ساخت‌واژه

ساخت‌واژه اشاره به روش‌های ساخت واژه‌ها از واحدهای با معنای واضح توسط تکواژ[2]ها دارد. تکواژ یا واژک کوچک‌ترین واحد صورت‌های زبانی است که دارای معنای دستوری یا واژگانی بوده و قابل تجزیه به واحدهای معنی‌دار دیگر نیست. تکواژها به دو دستهٔ کلی تقسیم می‌شوند [۶]:

۱) ریشه‌ها[3]: تکواژ مستقل دارای معنای اصلی

۲) وندها[4]: تکواژ وابسته برای افزودن معنای اضافه. وندها با افزوده شدن به آغاز، پایان یا میان واژه‌ها، واژه جدیدی پدید می‌آورند.

۲-۳- زبان‌های از نظر ساخت‌واژی غنی

عبارت زبان‌های از نظر ساخت‌واژی غنی اشاره به زبان‌هایی دارد که حاوی اطلاعات دستوری قابل توجهی هستند. به عبارت دیگر اطلاعات زیادی مرتبط با واحدهای نحوی دارند و روابط در سطح واژه مطرح می‌شود. اطلاعات روابط اجزای نحوی در قالب واژه‌ها نشان داده می‌شوند که این واژه‌ها می‌توانند آزادانه موقعیت

[1] MRLs: Morphologically Rich Languages
[2] Morpheme
[3] Stem
[4] Affix

خود را در جمله تغییر دهند. به این خصوصیت «بی‌ترتیبی[1]» گفته می‌شود. این خصوصیات در ساختار وابستگی منجر به تولید ساختارهای غیرافکنشی خواهد شد. شواهد فراوانی وجود دارد که کاربرد مدل‌های تجزیهٔ احتمالی به چنین زبان‌هایی مستعد کاهش کارایی است.

زبان انگلیسی، که در حوزهٔ پردازش زبان طبیعی بسیار مورد مطالعه قرار گرفته است، جزو این دسته از زبان‌ها به حساب نمی‌آید[2]. حتی با وجود بازتاب برخی خصوصیات نحوی در قالب واژه‌ها، اطلاعات ساخت‌واژی در این زبان اغلب نسبت به سایر عوامل نحوی مثل موقعیت واژه‌ها درجهٔ دوم اهمیت را دارد. زبان‌های هندی‌اروپایی مثل آلمانی با وجود ارتباط نزدیک با انگلیسی، دارای برخی خواص هستند که آن‌ها را جزو این دسته از زبان‌ها می‌کند. زبان‌های سامی مثل عربی و عبری با وجود غنی بودن حالت‌های ساخت‌واژی و انعطاف در ترتیب نحوی که از خود نشان می‌دهند، حد نهایی این دسته از زبان‌ها هستند [۴].

۲-۴- خصوصیات زبان فارسی

در این بخش به بررسی خصوصیات ساخت‌واژی و صرفی زبان فارسی خواهیم پرداخت:

◘ خصوصیات ساخت‌واژی زبان فارسی

در زبان فارسی بیش از ۱۲۰ تصریف مختلف فعل وجود دارد که اگر ضمایر پیوسته را نیز در نظر بگیریم این تعداد برای افعال گذرا به بیش از ۷۰۰ مورد تصریف نیز می‌رسد [۷]؛ این امر نشان‌دهندهٔ غنی بودن زبان فارسی از نظر ساخت‌واژی است. وجود ساخت‌واژهٔ غنی در یک زبان باعث تولید واژه‌های متمایز زیاد و نرخ بالای لغات خارج از واژگان می‌شود که این امر میزان اعتماد به پارامترهای لغوی را کاهش می‌دهد.

مقوله‌های دستوری[3]، ردهٔ تحلیلی درون دستور زبان است که اعضای آن دارای توزیع نحوی یکسان هستند و به عنوان واحد ساختاری در زبان تکرار شده و خواص مشترک مفهومی یا نحوی را به اشتراک می‌گذارند. در ادامه فهرستی از مقوله‌های دستوری که در اکثر زبان‌ها مشترک است، ارائه خواهد شد.

- جانداری[4]: یک مقولهٔ مفهومی یا دستوری است که دارا بودن درک یا زنده بودن را نشان می‌دهد. این

---

[1] Free word-order

[2] در موارد خاصی خصوصیات مطرح شده در زبان انگلیسی نیز رخ می‌دهند اما به علت چشمگیر نبودن وقوع آن‌ها، نمی‌توان انگلیسی را جزو زبان‌های از نظر ساخت‌واژی غنی دانست.

[3] Grammatical Category

[4] Animacy

مقوله دو مقدار جاندار و بی‌جان می‌تواند داشته باشد. در برخی زبان‌ها (مثل ترکی، اسپانیایی) تفاوتی در مقولهٔ جانداری وجود ندارد. در زبان انگلیسی جانداری تعریف شده که به عنوان مثال تفاوت بین he/she با it در جانداری است. در منابع مختلف از جمله [۸] سلسله مراتبی از جانداری ارائه شده که طیفی از جانداری تا بی‌جانی را پوشش می‌دهد.

اول شخص > دوم شخص > سوم شخص > اسامی خاص > جانداران غیر انسان > بی‌جان

- نمود[1]: این مقوله که برای فعل تعریف می‌شود، نشان‌دهندهٔ نوع احساس متکلم از ساختار زمانی است. در زبان فارسی چهار نمود (ساده، نقلی، استمراری و مستمر) وجود دارد. به عنوان مثال زمانی که عملی در یک دورهٔ زمانی به طور پیاپی انجام شده باشد «نمود استمراری» استفاده می‌شود که نشانهٔ آن در فارسی «می» و در انگلیسی «ing» است.
- حالت[2]: ابزاری است که نقش دستوری یک واژه در جمله را نشان می‌دهد. در فارسی از «حرف اضافه» برای نشان دادن حالت‌های گوناگون نحوی استفاده می‌شود اما به طور کلی حالت در زبان فارسی تعریف نشده است. در زبان‌های هندی‌اروپایی ۸ حالت وجود دارد:
  - حالت نهادی (فاعلی)[3]: این حالت نشان‌دهندهٔ این است که انجام‌دهندهٔ فعل کیست یا اینکه نهاد جمله کجاست (**ما** به بوستان رفتیم).
  - حالت مفعولی[4]: این حالت برای نشان دادن مفعول مستقیم یک فعل متعدی استفاده می‌شود. (شهریار **ما را** دید).
  - حالت کنش‌گری[5]: این حالت برای نشان دادن مفعول با واسطه است (شهریار **به ما** پول داد).
  - حالت از سویی[6]: این حالت برای نشان دادن حرکت از سوی چیزی یا دلیل چیزی را نشان می‌دهد (**از خانه** رفت).
  - حالت وابستگی[7]: این حالت نقش اضافه و شکل ملکی را نشان می‌دهد (**خانهٔ ما** بزرگ است).
  - حالت ندایی[8]: این حالت مخاطب جمله را نشان می‌دهد (**سعدیا** مرد نکونام نمیرد هرگز).

---

[1] Aspect
[2] Case
[3] Nominative Case
[4] Accusative Case
[5] Dative Case
[6] Ablative Case
[7] Genitive Case
[8] Vocative Case

- حالت مکانی[1]: این حالت مکان جمله را نشان می‌دهد (او **در کرج** زندگی می‌کند).
- حالت ابزاری[2]: این حالت ابزار انجام یک فعل را نشان می‌دهد (او زمین را **با جارو** رُفت).

- معرفگی[3]: خصوصیتی برای یک عبارت اسمی است که در بین هویت‌های خاص قابل تمایز یا در حوزهٔ داده‌شده قابل تشخیص باشد. در فارسی معمولاً حرف تعریف[4] وجود ندارد در حالی که اکثر اسامی انگلیسی با یک حرف تعریف (مثل the) ظاهر می‌شوند [۹].
- درجهٔ برتری[5]: این مقوله برای یک صفت یا قید است که مقدار نسبی را در جمله توصیف می‌کند. برای زبان فارسی سه درجه قابل تعریف است:
  - صفت مطلق[6]: صفتی که کیفیتی را توصیف کند (گل **سرخ**).
  - صفت تفضیلی[7]: کیفیت را با نوع دیگری مقایسه کند (لیگ **برتر**).
  - صفت عالی[8]: کیفیت را با تعداد زیاد یا همه مقایسه کرد (**بهترین** زمان).
- جنسیت[9]: نماد جنسیت در دستور وابستگی متفاوت از نماد اجتماعی و بیولوژیک آن است. با این وجود به طور نزدیکی با هم در تعامل هستند. برخی زبان‌ها (مثل مجاری، فنلاندی و ترکی) جنس ندارند. برخی زبان‌ها (مثل هندی، عربی و عبری) دارای دو جنس مذکر و مؤنث هستند. برخی زبان‌ها مثل (بلغاری و آلمانی) دارای جنس مذکر، مؤنث و خنثی هستند. در برخی دیگر از زبان‌ها (مثل چک و روسی) بیش از سه دسته جنس دارند. در فارسی بر خلاف انگلیسی تمایزی بین مذکر و مؤنث وجود ندارد [۹]، [۱۰]. به جز برای تعداد اندکی از کلمات عاریتی جاندار عربی که برای نشان دادن تأنیث«ة» می‌گیرند [۱۱].
- وجه[10]: این مقوله برای افعال تعریف می‌شود. در زبان فارسی سه وجه وجود دارد:
  - وجه اخباری[11]: وقوع کاری را به طور قطع و یقین خبر می‌دهد (رفتم، زدم، خواهم رفت).
  - وجه التزامی[1]: کار را از طریق شک و دودلی، آرزو و خواهش بیان می‌کند (می‌خواهم بروم، شاید

---

[1] Locative Case
[2] Instrumental Case
[3] Definiteness
[4] Definite Article
[5] Degree of comparison
[6] Positive
[7] Comparative degree
[8] Superlative degree
[9] Gender
[10] Mood
[11] Indicative mood

بروم، گمان کنم بروم).

- وجه امری[۲]: کار را به صورت حکم، خواهش و فرمان بیان می‌کند (برو، بروید، بگو، بگویید).

امر منفی را نهی گویند که جزو وجه امری به حساب می‌آید (مرو، نشنو).

در فعل امری به جهت تاکید یا استمرار «می» اضافه می‌شود.

- وجهیت[۳]: این مقوله برای تعیین فعل‌های وجهی است. فعل‌های وجهی یکی از ابزارهای بیان وجه هستند که در برخی از زبان‌ها از جمله زبان فارسی استفاده می‌شود. افعال «بایستن»، «شدن» و «توانستن» سه فعل وجهی در فارسی امروز هستند.
- شمار[٤]: در زبان فارسی دو شمار «مفرد» و «جمع» وجود دارد. این مقوله، مطابقهٔ[٥] شمار بین اسم، ضمیر، صفت و فعل را مشخص می‌کند. گرچه در فارسی مطابقهٔ شمار می‌تواند برای جمع‌های بی‌جان برقرار نباشد [۹].
- شخص[٦]: در فارسی سه شخص (اول شخص، دوم شخص، سوم شخص) وجود دارد.
- قطبیدگی[۷]: این خصوصیت برای افعال تعریف می‌شود و نشان می‌دهد که فعل در حالت منفی (مثل نرو یا مرو) یا در حالت مثبت قرار دارد.
- زمان[۸]: این مقوله برای افعال تعریف شده است. در فارسی سه زمان دستوری (گذشته، حال و آینده) وجود دارد.
- گذرایی[۹]: این خصوصیت برای افعال تعریف شده که وضعیت فعل را از نظر لازم و متعدی بودن نشان می‌دهد.
- جهت[۱۰]: این خصوصیت برای افعال تعریف شده که رابطهٔ بین عمل یا حالت فعل و شرکت کنندگان را نشان می‌دهد. در واقع وضعیت فعل از نظر معلوم و مجهول بودن را مشخص می‌کند (گربه موش را خورد – موش توسط گربه خورده شد).

---

[1] Subjunctive mood
[2] Imperative mood
[3] Modality
[4] Number
[5] Agreement
[6] Person
[7] Polarity
[8] Tense
[9] Transitivity
[10] Voice

◘ خصوصیات صرفی زبان فارسی

به صورت متعارف زبان فارسی دارای ترتیب واژهٔ «فاعل، مفعول، فعـل» یـا «SOV» اسـت، امـا اسـتثناهای فراوانی در ترتیب واژه‌ها وجود دارد که زبان فارسی را جزو زبان‌های بی‌ترتیب قرار داده است [۹]. به عنـوان مثال در یک جمله، قید می‌تواند در ابتدا، انتها یا وسط جمله ظاهر شود، بدون آنکه معنای جمله تغییر یابد.

۲-۴-۱- ساخت بنیادین

ساخت بنیادین هر جمله عبارتست از فعل اصلی یا مرکزی آن جمله به علاوهٔ متمم‌های اجباری و اختیـاری که این با جملهٔ ساده متفاوت است زیرا جملهٔ ساده یک فعل دارد اما برخی جمله‌های بنیادین بسته به نـوع فعلشان ممکن است بیش از یک فعل داشته باشند.

طبیب‌زاده [۱۲] با بررسی بیش از ۲۰۰ فعل فارسی، ۸ متمم نحوی یا متمم ظرفیتی[1] پیشنهاد داده است و ویژگی‌های کاربردی هر کدام از این متمم‌ها به همراه راه‌های تشخیص آن‌هـا از یکـدیگر را بررسـی کـرده است.

۱) فاعل[2] (فا - SBJ): فاعل در جملات زبان فارسی می‌تواند به صورت «آشکار»، «محذوف» و «صفر» بیاید. همچنین می‌تواند هر جای جمله ظاهر شده یا کاملاً حذف شود (**علی** به خانه آمد).

۲) مفعول[3] (مف - OBJ): مفعول در جملات به شکل کلمه یا گروهی مستقل ظاهر می‌شـود (مـن **کتـاب** خواندم، من **کتاب** را خواندم). طبق دستورهای سنتی افعالی که مفعول می‌پذیرنـد، «فعـل متعـدی» و بقیه را «فعل لازم» می‌نامند.

۳) مفعول حرف اضافه‌ای (مفح - VPP): به شکل «گروه حرف اضافه‌ای» در جمله ظاهر می‌شود که بسـته به فعـل مـی‌توانـد بـه صـورت «مـتمم اجبـاری» (او **بـه دختـرش** اجـازه داد کـه تحصـیل کنـد) یـا «متمم اختیاری» (او **با دست** غذا را خورد) ظاهر شود.

۴) مفعول نشانهٔ اضافه‌ای[4] (مفن - EZC): تنها متممی در زبان فارسی است کـه بـا افعـال بسـیط بـه کـار نمی‌رود، بلکه فقط با برخی از افعال مرکب می‌آید. توسط یک «واسطه» یـا «کسـرهٔ اضـافه» بـه مؤلفـهٔ

---

[1] valency slot
[2] Subject
[3] Object
[4] Ezafe Object

غیرفعلی در برخی افعال مرکب می‌پیوندد. به جای آن می‌توان از مفعول یا مفعول حرف اضافه‌ای استفاده کرد. این متمم نیز می‌تواند اجباری یا اختیاری باشد (او به دخترش اجازهٔ **تحصیل** داد) [۱۳].

۵) بند متممی[1] (بند - VCL): برخی افعال فقط با یک بند متممی می‌آیند که حذف از جمله، آن را بدساخت یا ناقص کرده و به جای آن از هیچ متمم دیگری نمی‌توان استفاده کرد. بند متممی اغلب به صورت متمم اجباری فعل در جمله ظاهر می‌شود (می‌دانم که می‌آید).

۶) مسند[2] (مس - MOS): اسم یا صفتی که با افعال ربطی به کار می‌رود (هوا **سرد** است).

۷) تمیز[3] (تم - TAM): صفت یا حالتی را به مفعول جمله نسبت می‌دهد. مفعول و تمیز به هم مرتبطند، به طوری که نام دیگر تمیز «متمم مفعول» است. در موارد معدودی تمیز به جای مفعول، با مفعول حرف اضافه‌ای به کار می‌رود (مثل، به قسمت‌های شمالی این منطقه آران اطلاق می‌شود). تمیز را می‌توان به صورت دو اسم یا دو صفت هم‌پایه استفاده کرد (مثل، آن‌ها را حسن و احمد پنداشتم)، اما هیچ‌گاه نمی‌توان یک اسم و صفت را به عنوان تمیز هم‌پایه کرد (مثل، آن‌ها را حسن و خوشحال پنداشتم).

۸) متمم قیدی[4] (مق - ADVC): به قید یا گروه قیدی گویند که در ساخت ظرفیتی فعل به عنوان یکی از متمم‌های آن ظاهر می‌شود. این افعال در فارسی چندان زیاد نیستند. از نظر نوع، متمم قیدی دارای انواع زیر است:

الف. متمم قید مکانی: به جای آن همواره می‌توان از ضمایری مثل «اینجا» و «آنجا» استفاده یا توسط کلمهٔ پرسشی «کجا» سؤالی کرد. (مثل، آریا **پیش** میناست). این نوع متمم اغلب اجباری است و حذف آن جمله را بدساخت یا غیردستوری می‌کند.

ب. متمم قید حالتی: به جای آن همواره می‌توان از ضمیرهایی مثل «این طور» و «این گونه» استفاده یا توسط کلمهٔ پرسشی «چگونه» سؤالی کرد (مثل، علی **عاقلانه** رفتار کرد). این نوع متمم نیز اغلب اجباری است.

ج. متمم قید مقداری: به جای آن همواره می‌توان از ضمیرهای «آنقدر» و «این اندازه» استفاده یا توسط کلمهٔ پرسشی «چقدر» سؤالی کرد (مثل، بچه **خیلی** رشد کرده است). این متمم از نوع اختیاری بوده و

---

[1] Complement Clause
[2] Mosnad
[3] Tamiz
[4] Adverbial Complement

می‌تواند از جمله حذف شود.

طبیب‌زاده برای استخراج این افعال از پیکره‌ای بالغ بر ۶۰۰۰ جمله استفاده کرده که از فرهنگ فارسی عامیانهٔ نجفی، فرهنگ سخن و کتاب دستور سال دوم و سوم آموزش متوسط عمومی استخراج شدند. با استفاده از این پیکره، ۲۳ ساخت یا جملهٔ بنیادین زبان فارسی استخراج شده است که فهرست کامل آن در جدول (۲-۱) آمده است. در این بین دو ساخت «‖فا، مف‖» و «‖فا، مس‖» پرکاربردترین ساخت‌های بنیادین در زبان فارسی نوشتاری و معیار امروز هستند.

◘ ساخت ظرفیتی

ویژگی‌های نحوی فعل را از حیث مجموعهٔ متمم‌هایی اجباری و اختیاری که در جملات گوناگون دریافت می‌کند با ساخت ظرفیتی نمایش می‌دهند. این در حالی است که ساخت بنیادین جمله، مهم‌ترین ویژگی‌های نحوی جمله را در اولین گام تحلیل نحوی بر اساس ساخت ظرفیتی فعل مرکزی جمله نمایش می‌دهد. در حقیقت سطح انتزاع «ساخت بنیادین جمله» کمتر از سطح انتزاع «ساخت ظرفیتی» است. به ازای هر «ساخت ظرفیتی» می‌تواند یک یا چند «ساخت بنیادین جمله» وجود داشته باشد. از یک سو تمام جملات زبان بر اساس «ساخت‌های بنیادین جمله» به وجود می‌آید و از سوی دیگر تمام جملات زبان قابل تقلیل به آن‌ها هستند.

◘ ظرفیت فعل

ظرفیت فعل تعداد کل مکمل‌هایی را نشان می‌دهد که یک فعل می‌تواند دریافت کند. این عدد نمادی انتزاعی بوده و متعلق به پیکرهٔ ذهنی سخنگوی محلی است. انواع ظرفیت فعل، توصیف کنندهٔ حالات ممکنی است که فعل می‌تواند بگیرد.

◘ فرهنگ ظرفیت فعل فارسی

این فرهنگ توسط گروه پژوهشی دادگان تحت حمایت دبیرخانهٔ شورای عالی اطلاع رسانی تهیه شده و به صورت رایگان قابل دستیابی است[1]؛ شامل ۴۲۸۲ بن فعل متمایز و ۵۴۲۹ جفت فعل-ظرفیت متمایز است. همچنین حاوی اطلاعات مکمل‌های اجباری و اختیاری افعال است. برای تولید این پیکره، یک تیم بیش از ۱۰ ماه افعال نامزد خام فارسی را جمع‌آوری کردند. هر فعل می‌تواند ۱ تا حداکثر ۵ ظرفیت مختلف

[1] http://dadegan.ir/download

داشته باشد که به طور متوسط هر فعل ۱.۲۶۸ ظرفیت متمایز می‌تواند دریافت کند [۵]. همچنین این افعال و ظرفیت متناظر با آن قابل جستجو به صورت آنلاین[1] می‌باشد.

جدول (۲-۱) ساخت‌های بنیادین زبان فارسی [۱۲]

ظرفیت	ساخت بنیادین	مثال
تگ ظرفیتی	‖فا‖	آمدن: مامان آمد – رفتن: آقا رفت
	‖Ø، بند‖	جاداشتن: جاداشت دیدن او برویم
دو ظرفیتی	‖فا، مف‖	خواندن: ما کتاب را خواندیم
	‖فا، مفح‖	تعریف کردن: مژده از آرایش ترانه تعریف کرد
	‖فا، مفن‖	صحبت کردن: امیر صحبت پول را کرد
	‖فا، بند‖	گفتن: او گفت برویم
	‖فا، مس‖	بودن: او زیباست
	‖فا، مق‖	بودن (به معنی وجود داشتن، قرار داشتن): آریا پیش مرجان است
	‖فا / Ø بند‖	ضروری بودن: ضروری است که او را ببینم (‖Ø بند‖) دیدن او ضروری است (‖فا‖)
سه ظرفیتی	‖فا، مف، مفح‖	پرسیدن: امیر راه را از مسعود پرسید
	‖فا، مف، مفن‖	تحویل دادن: مدرک را تحویل ما بدهید
	‖فا، مف، بند‖	مطلع کردن: او ما را مطلع کرد که جنگ شده است
	‖فا، مف، تم‖	نامیدن: همدان را هگمتانه نامیدند
	‖فا، مف، مق‖	گذاشتن: آریا را پیش مینا گذاشتم
	‖فا، مفح، مفح‖	رفتن: او از تهران به کرج رفت
	‖فا، مفح، مفن‖	اجازه دادن: او به ما اجازه رفتن داد
	‖فا، مفح، بند‖	اجازه گرفتن: آن‌ها از استاد اجازه گرفتند که بروند
	‖فا، مفح، مق‖	طول کشیدن: از اینجا تا تهران ۵ ساعت طول می‌کشد
	‖فا، مفح، تم‖	اطلاق کردن: آن‌ها به این کوچه‌ها خیابان اطلاق می‌کردند
	‖فا، مفن، بند‖	کمک کردن: او کمک کرد که برود
	‖فا / Ø بند، مفح‖	پیدا بودن: از چشم‌هایت پیداست که دروغ می‌گویی (‖Ø بند، مفح‖) اینکه دروغ می‌گویی از چشم‌هایت پیداست (‖فا، بند‖)
چهار ظرفیتی	‖فا، مف، مفح، مفح‖	فرستادن: او نامه را از ایران به آلمان فرستاد
	‖فا، مفح، مفح، مق‖	طول کشیدن: مسافرت از اینجا تا تهران پنج ساعت طول می‌کشد

[1] http://dadegan.sobhe.ir/

در این فرهنگ علاوه بر ۸ متمم نحوی تعریف شده توسط طبیب‌زاده، از مفعول دوم (مف۲ - OBJ2) نیز استفاده شده که بر این اساس ۲۵ ساخت بنیادین تعریف شده است. در جدول (۲-۲) جزئیات این ساخت‌های بنیادین نشان داده شده است.

جدول (۲-۲) ساخت‌های بنیادین استفاده شده در فرهنگ ظرفیت فارسی [۷]

‖فا‖	‖Ø، بند‖	‖فا، مف‖	‖فا، مفح‖	‖فا، مفن‖
‖فا، بند‖	‖فا، مس‖	‖فا، مق‖	‖فا، مف، مفح‖	‖فا، مف، مفن‖
‖فا، مف، بند‖	‖فا، مف، تم‖	‖فا، مف، مس‖	‖فا، مفح، مفح‖	‖فا، مفح، مق‖
‖فا، مفح، بند‖	‖فا، مفح، تم‖	‖فا، مفن، بند‖	‖فا، مق، مفن‖	‖فا / Ø بند، مفح‖
‖فا، مف، مفح، مفح‖	‖فا، مفح، مفح، مق‖	‖فا، مف، مفح، تم‖	‖فا، مف، مفح، مق‖	‖فا، مف، مفح، مف۲‖

۲-۵- پیکرهٔ وابستگی زبان فارسی

برای کلیهٔ آزمایش‌ها انجام شده در این پژوهش از نسخهٔ ۰.۱ پیکرهٔ وابستگی استفاده شده است[۱]. خصوصیات عمومی این پیکره در جدول (۲-۳) آمده است. توزیع طول جملات پیکره در شکل (۲-۱) نشان داده شده است که دارای جملات با طول ۱ تا ۱۹۳ واژه است.

جدول (۲-۳) خصوصیات عمومی پیکرهٔ وابستگی فارسی

خصوصیت	مقدار
تعداد جملات	۱۲۴۵۵
تعداد واژه‌ها	۱۸۹۵۷۲
متوسط طول جملات	۱۵.۲۲
درصد روابط وابستگی غیرافکنشی	٪۱.۷۷

یک پیکرهٔ وابستگی دیگر برای زبان فارسی به نام UPEDT[۲] با استفاده از نسخهٔ اصلاح شدهٔ پیکرهٔ

[۱] در حال حاضر نسخهٔ ۱.۰ قابل دریافت می‌باشد اما به این دلیل که نیمی از آزمایش‌ها قبل از عرضهٔ این نسخه انجام شده بود، کار را با نسخه ۰.۱ ادامه دادیم.

[2] Uppsala PErsian Dependency Treebank

بیجن‌خان در حال طراحی است [۱۴] که تنها نسخهٔ آزمایشی از آن شامل ۱۲۸۲ جمله (۲۶۰۶۵ واژه) ارائه شده است[1].

شکل (۲-۱) توزیع طول جملات پیکرهٔ وابستگی زبان فارسی

برای ارائهٔ پیکرهٔ وابستگی، قالب‌های مختلفی وجود دارد. در این میان قالب CoNLL[2] که در سال ۲۰۰۶ ارائه شده رایج‌ترین است. پیکرهٔ وابستگی فارسی نیز در همین قالب ارائه شده است. اطلاعات در آن به صورت ستونی بوده که معرفی ستون‌ها و کاربرد هر یک در جدول (۲-۴) آمده است.

جدول (۲-۴) قالب CoNLL برای نمایش پیکرهٔ وابستگی

ستون	مفهوم
ID	اندیس جملهٔ ورودی (شروع از ۱)
Word Form	واژهٔ خام ورودی
Lemma	ریشهٔ واژهٔ ورودی
CPOSTAG	برچسب سطح بالای اجزای سخن
POSTAG	برچسب سطح پایین اجزای سخن
FEATS	فهرستی از خصوصیات ساخت‌واژی و صرفی به فرمت key=value که با کارکتر | از هم جدا می‌شوند.
HEAD	اندیس واژهٔ پدر در درخت وابستگی
DEPREL	برچسب وابستگی به واژهٔ پدر
PHEAD	مانند HEAD با این تفاوت که تضمین می‌کند درخت حاصل افکنشی خواهد بود. معمولاً این ستون خالی (_) رها می‌شود.
PDEPREL	برچسب وابستگی برای PHEAD که معمولاً این ستون نیز خالی (_) رها می‌شود.

[1] http://stp.lingfil.uu.se/~mojgan/UPEDT.html
[2] http://ilk.uvt.nl/conll/#dataformat

در این فرمت ۶ ستون اول را اطلاعات ورودی و ۴ ستون آخر را اطلاعات خروجی تشکیل می‌دهد. در زمان آموزش از اطلاعات همه ستون‌ها استفاده شده و در زمان تجزیه اطلاعات خروجی در نظر گرفته نشده و توسط تجزیه‌گر پیش‌بینی و در این ستون‌ها نوشته می‌شود.

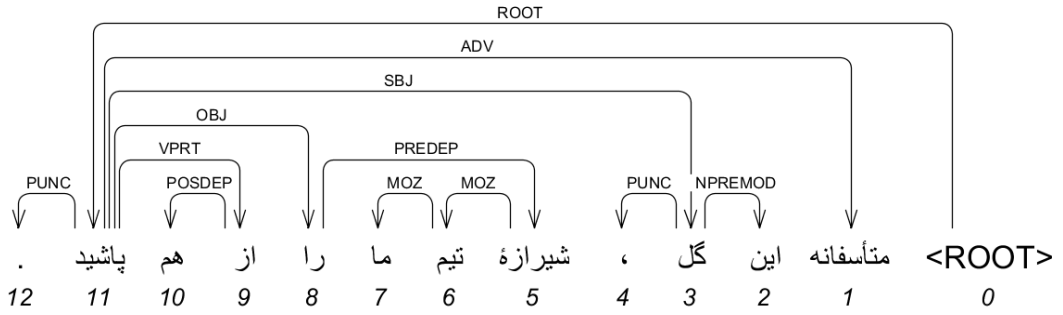
به عنوان مثال برای جملهٔ «متأسفانه این گل، شیرازهٔ تیم ما را از هم پاشید.» فرمت ورودی و درخت وابستگی به ترتیب در شکل (۲-۲) و شکل (۲-۳) نشان داده شده است.

1	متأسفانه	متأسفانه	ADV	SADV	separation=ISO	11	ADV	_	_
2	این	این	PREM	DEMAJ	separation=ISO	3	NPREMOD	_	_
3	گل	گل	N	IANM	separation=ISO|number=SING	11	SBJ	_	_
4	،	،	PUNC	PUNC	separation=ISO	3	PUNC	_	_
5	شیرازهٔ	شیرازه	N	IANM	separation=ISO|number=SING	8	PREDEP	_	_
6	تیم	تیم	N	IANM	separation=ISO|number=SING	5	MOZ	_	_
7	ما	ما	SPR	SPR	person=1|separation=ISO|number=PLUR	6	MOZ	_	_
8	را	را	POSTP	POSTP	separation=ISO	11	OBJ	_	_
9	از	از	PREP	PREP	separation=ISO	11	VPRT	_	_
10	هم	هم	UPR	UPR	person=3|separation=ISO|number=SING	9	POSDEP	_	_
11	پاشید	پاشید#پاش	V	ACT	person=3|separation=ISO|number=SING|tma=GS	0	ROOT	_	_
12	.	.	PUNC	PUNC	separation=ISO	11	PUNC	_	_

شکل (۲-۲) نمونه قالب CoNLL پیکرهٔ وابستگی

برای ارائهٔ اطلاعات ستون Lemma در پیکرهٔ زبان فارسی از قالب «(پیشوند#)(بن ماضی)#بن مضارع» استفاده شده است. در این قالب پیشوند و بن ماضی اختیاری هستند و به این ترتیب مقادیر این ستون به سه صورت زیر عرضه می‌شوند:

- برای فعل «هستند» ریشه «#است»
- برای فعل «شد» ریشه «کرد#کن»
- برای فعل «فرومی‌خورد» ریشه «فرو#خورد#خور»



شکل (۲-۳) درخت وابستگی متناظر با جملهٔ شکل (۲-۲)

در پیکرهٔ وابستگی فارسی برای ستون FEATS پنج نوع اطلاعات ساخت‌واژی در نظر گرفته شده است که جزئیات آن در جدول (۲-۵) آمده است. یکی از این خصوصیات «اتصال» است که برای تمامی لغات پیکره نمادگذاری شده، وضعیت استقلال یا وابستگی واژهٔ جاری به واژهٔ قبل و بعد از خود را نشان می‌دهد. خصوصیت «شناسهٔ جمله» برای تمام واژه‌های یک جمله مقدار یکسانی دارد ولی ارزشی از نظر زبان‌شناسی ندارد. دو خصوصیت شخص و شمار که در بخش مقوله‌های دستوری معرفی شدند نیز نمادگذاری شدند.

جدول (۲-۵) اطلاعات ساخت‌واژی موجود در ستون FEATS پیکرهٔ وابستگی فارسی

خصیصه	مقادیر مجاز	مفهوم	مثال	تعداد حضور	درصد حضور
Attachment	ISO	واژهٔ مستقل	رهایم نمی‌کند	۱۸۸۲۸۱	٪۹۹.۳۲
	PRV	وابسته به واژهٔ سمت راست	رهایم نمی‌کند	۶۳۶	٪۰.۳۴
	NXT	وابسته به واژهٔ سمت چپ	رهایم نمی‌کند	۶۴۷	٪۰.۳۴
senID	نشان دهندهٔ شمارهٔ جمله در پایگاه دادهٔ دادگان به منظور گزارش خطا				
Number	SING	مفرد	ترکمن	۸۱۲۳۲	٪۴۲.۸۵
	PLUR	جمع	ترکمن‌ها	۲۰۳۸۴	٪۱۰.۷۵
Person	۱	اول شخص	می‌کنم - می‌کنیم	۵۸۱۰	٪۳.۰۷
	۲	دوم شخص	می‌کنی - می‌کنید	۳۴۷۷	٪۱.۸۳
	۳	سوم شخص	می‌کند - می‌کنند	۲۰۲۶۳	٪۱۰.۶۹
Tma	HA	حال امری	بخور	۸۷۷	٪۰.۴۶
	AY	آیندهٔ اخباری	خواهم خورد	۱۰۰۸	٪۰.۵۳
	GNES	گذشتهٔ نقلی استمراری اخباری	می‌خورده‌ام	۴۶	٪۰.۰۲
	GBES	گذشتهٔ بعید استمراری اخباری	می‌خورده بودم	۰	٪۰.۰۰
	GES	گذشتهٔ استمراری اخباری	می‌خوردم	۱۸۴۹	٪۰.۹۷
	GN	گذشتهٔ نقلی اخباری	خورده‌ام	۲۲۴۵	٪۱.۱۸
	GB	گذشتهٔ بعید اخباری	خورده بودم	۱۱۵۵	٪۰.۶۱
	H	حال اخباری	می‌خورم	۷۴۶۸	٪۳.۹۴
	GS	گذشتهٔ ساده اخباری	خوردم	۵۲۱۹	٪۲.۷۵
	GBESE	گذشتهٔ بعید استمراری التزامی	می‌خورده بوده باشم	۱	٪۰.۰۱
	GESEL	گذشتهٔ استمراری التزامی	می‌خورده باشم	۱	٪۰.۰۱
	GBEL	گذشتهٔ بعید التزامی	خورده بوده باشم	۰	٪۰.۰۰
	HEL	حال التزامی	بخورم	۳۸۷۸	٪۲.۰۵
	GEL	گذشتهٔ التزامی	خورده باشم	۶۰۱	٪۰.۳۲

خصوصیت tma (زمان/وجه/نمود)[1] به عنوان ویژگی فعل، سامانه‌ای دستوری در زبان است که بیان زمان دستوری (محل وقوع رخداد فعل روی خط زمان)، وجه (میزان اجبار، ضرورت، توانایی و غیره) و نمود (تمام شده بودن یا جریان داشتن رویداد فعل) را بر عهده دارد. در زبان فارسی سه زمان دستوری (گذشته، حال و آینده)، سه وجه (اخباری، التزامی و امری) و چهار نمود (ساده، نقلی، استمراری و مستمر) وجود دارند. برای هر فعل، سه مقولهٔ دستوری زمان، وجه و نمود به نحوی به هم آمیخته شده‌اند که بهتر است به جای گزینش مجزای هر یک از آن‌ها به عنوان ارزش‌های جداگانه، ترکیب آن‌ها در هر صیغهٔ فعل به عنوان یک ارزش در نظر گرفته شود زیرا مشاهده شده است که در برخی از صیغه‌های افعال میان نمودها تداخل وجود دارد. به عنوان مثال صیغهٔ «گذشتهٔ نقلی استمراری اخباری» از لحاظ نمود هم نقلی و هم استمراری است.

دو ارزش حال مستمر (مثل «دارم می‌خورم») و گذشتهٔ مستمر (مثل «داشتم می‌خوردم») به عمد در میان سایر ارزش‌ها گنجانده نشده است. به منظور سهولت پردازش رایانه‌ای، رابطهٔ وابستگی میان دو کلمه سازندهٔ آن‌ها در نظر گرفته شده است.

۲-۶- نتیجه‌گیری

در این فصل ساخت‌واژه را روش‌های ساخت واژه‌ها از تکواژها تعریف کردیم. بر اساس این تعریف زبان‌های از نظر ساخت‌واژی غنی، زبان‌هایی هستند که میزان تصریف بالایی داشته و نرخ تولید واژگان جدید در آن‌ها بالاست. سپس زبان فارسی که جزئی از این زبان‌هاست را از منظر ویژگی‌های ساخت‌واژی و صرفی مورد مطالعه قرار دادیم. در پایان به معرفی فرهنگ ظرفیت فعل و پیکرهٔ وابستگی که برای پردازش رایانه‌ای زبان فارسی مفید هستند، پرداختیم. در فصل بعد به بررسی مشکلات اصلی زبان‌های از نظر ساخت‌واژی غنی در تجزیهٔ وابستگی خواهیم پرداخت.

[1] tense-mood-aspect

فصل ۳:

مروری بر کارهای مرتبط

۳-۱- مقدمه

در ابتدای این فصل روند کاری در تجزیهٔ وابستگی که منجر به تمرکز بر روی زبان‌های از نظر ساخت‌واژی غنی شده می‌پردازیم. ضمن بررسی علت‌های دشواری این دسته از زبان‌ها، مهم‌ترین چالش‌ها در تجزیهٔ این زبان‌ها معرفی خواهد شد. در پایان تلاش‌های انجام شده در زبان‌های مختلف برای حل هر یک از این چالش‌ها بیان خواهد شد.

۳-۲- مشکل کجاست؟

در دهمین کار مشترکی[1] همایش سالیانهٔ یادگیری رایانه‌ای زبان طبیعی[2] بر روی تجزیهٔ وابستگی که در سال ۲۰۰۶ [۱۵] انجام شد، تجزیه‌گرهای مختلف بر روی ۱۳ زبان مورد ارزیابی قرار گرفتند. نتایج بر روی زبان‌های مختلف متفاوت بود و در جمع‌بندی، این زبان‌ها به دو دستهٔ ساده و دشوار تقسیم شدند اما معیاری برای تعریف سادگی و دشواری زبان به طور دقیق مورد بررسی قرار نگرفت. این موضوع در کار مشترک سال ۲۰۰۷ [۱۶]، [۱۷] با افزودن بخش تطبیق تجزیه‌گر به سایر دامنه‌ها به صورت دقیق‌تر مورد بررسی قرار گرفت. در جمع‌بندی نتایج، زبان‌ها بر اساس خواص توپولوژیکی آن‌ها به سه دسته تقسیم شدند:

- زبان‌های با صحت بالا: این زبان‌ها از نظر ساخت‌واژی ضعیف بوده و میزان تصریف پایینی دارند. زبان‌های کاتالان، انگلیسی و چینی در این دسته قرار دارند.
- زبان‌های با صحت متوسط: در این زبان‌ها تنها، درجهٔ تصریف بالا دیده شده است. زبان‌های چک، مجاری و ترکی در این دسته قرار دارند.
- زبان‌های با صحت کم: زبان‌های عضو این دسته جزو سخت‌ترین زبان‌ها هستند. ویژگی مشترک این زبان‌ها، دارا بودن درجهٔ بالای تصریف (غنی بودن از نظر ساخت‌واژی) و بی‌ترتیبی نسبی به صورت همزمان است. زبان‌های عربی، باسک و یونانی در این دسته قرار دارند.

بر اساس این خصوصیات زبان چک جزو زبان‌های با صحت کم باید قرار بگیرد اما به دلیل دارا بودن مجموعهٔ آموزش بسیار بزرگ‌تر از زبان‌هایی مثل عربی، صحت بالاتری نسبت به این زبان‌ها کسب کرده است. اندازهٔ دادهٔ آموزش به تنهایی برای این امر کافی نیست. عامل دیگر درصد واژه‌های جدید در مجموعهٔ

[1] Shared Task
[2] Conference on Computational Natural Language Learning (CoNLL)

آزمون است. انتظار می‌رود درصد این واژه‌ها در زبان‌های با تصریف بالا زیاد و در زبان‌های با ساخت‌واژهٔ ضعیف کم باشد. در نتیجه توصیه شده برای زبان‌های با تصریف بالا و بی‌ترتیبی نسبی، دادهٔ آموزشی بیش‌تری استفاده شود. در پایان نیاز به بهبود روش‌های تجزیه برای زبان‌های از نظر ساخت‌واژی غنی تأکید شده است.

برای بررسی دقیق‌تر این دسته از زبان‌ها، «کارگاه بررسی تجزیهٔ آماری زبان‌های از نظر ساخت‌واژی غنی[1]» در سال‌های ۲۰۱۰ [۱۸] و ۲۰۱۱ [۱۹] برگزار شده است. عوامل زبان‌شناختی متناظر با زبان‌های از نظر ساخت‌واژی غنی (مثل فهرست بلند واژگان، درجهٔ بالای بی‌ترتیبی واژه‌ها و استفاده از اطلاعات ساخت‌واژی در نمایش روابط نحوی) دشوارتر از آن است که با مدل‌ها و تکنیک‌های توسعه داده شده با فرض داده‌های انگلیسی، تجزیه شوند. از دیدگاه تجزیهٔ نحوی، زبان‌های دنیا معمولاً بر اساس سطحی که قابل پیکره‌بندی است طبقه‌بندی می‌شوند. بر همین اساس طیفی از زبان‌ها تعریف می‌شود که در یک سمت آن زبان‌هایی مانند انگلیسی قرار دارند که به شدت قابل پیکره‌بندی هستند و در سمت دیگر زبان‌هایی مثل عربی قرار دارند که ساختارهای ثابت بسیار کمی در سطح جمله دارند [۲۰].

به منظور بررسی تأثیر خصوصیات ساخت‌واژی و صرفی برای تجزیهٔ وابستگی مبتنی بر داده، تلاش‌های زیادی در سایر زبان‌ها (شامل زبان‌های اسپانیایی، آلمانی، باسک، ترکی، چک، چینی، عبری، عربی، کره‌ای و هندی) صورت گرفته است که در ادامه به بررسی این روش‌ها می‌پردازیم. در مورد زبان فارسی به دلیل در دسترس نبودن پیکرهٔ وابستگی، تنها بررسی انجام شده [۲۱] به صورت بی‌ناظر از روی جملات خام و بدون کمک گرفتن از داده‌های نمادگذاری شده صورت گرفته است. همچنین در مرجع [۲۲] دستور پیوندی[2] که نمایش خاصی از دستور زایشی است در زبان فارسی مورد بررسی قرار گرفته است. این دستور بسیار شبیه به دستور وابستگی است با این تفاوت که دستور وابستگی شامل روابط هسته و وابسته به صورت جهت‌دار است، در حالی که دستور پیوندی جهت را در روابط بین واژه‌ها در نظر نمی‌گیرد.

بر اساس مرجع [۴] هنگام تلاش برای ترکیب اطلاعات ساخت‌واژی در مدل‌های تجزیه، با سه نوع چالش روبرو خواهیم بود: معماری و تنظیمات اولیه، بازنمایی و مدل‌کردن، تخمین و هموارسازی

[1] SPMRL: Statistical Parsing of Morphologically Rich Languages
[2] Link Grammar

۳-۲-۱- معماری و تنظیمات اولیه

هنگام تلاش برای تجزیه، با واژه‌های پیچیده‌ای مواجه خواهیم شد که شامل هر دو اطلاعات لغوی و کارکردی هستند. همچنین وندها می‌توانند توابع مستقل نحوی داشته باشند. در اکثر مدل‌های تجزیه قسمت‌بندی بخش‌های مستقل نحوی قبل از شروع مفروض است؛ این در حالی است که مسائل دنیای واقعی این فرض را ندارند. قسمت‌بندی ساخت‌واژی خود نیازمند ابهام‌زدایی است که امری ساده و بدیهی نخواهد بود. اطلاعات ساخت‌واژی ستون FEATS و همچنین برچسب‌های اجزای سخن به صورت خودکار باید تولید شوند و ممکن است دارای مقادیر نویزی و همراه با خطا باشند.

تلاش‌های انجام‌شده با هدف رسیدگی به این چالش شامل بررسی تأثیر استفاده از «اطلاعات استاندارد و بدون خطای اولیه[1]» در مقابل استفاده از «اطلاعات پیش‌بینی‌شده به صورت خودکار» قبل از شروع تجزیه است. افت صحت هنگام ارزیابی سامانه‌هایی که از داده‌های پیش‌بینی‌شده استفاده می‌کنند، تعجب‌آور نخواهد بود، اما نتیجهٔ جالب این است که در اکثر حالات استفاده از اطلاعات ساخت‌واژی نویزی بدتر از عدم استفاده از آن‌هاست.

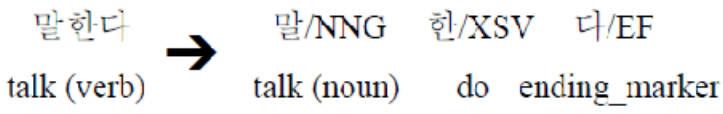
◘ یافتن مجموعه برچسب‌های بهینه اجزای سخن

در دسته‌ای از تلاش‌ها، سعی شده اثر مجموعه برچسب[2] بهینه اجزای سخن در تجزیهٔ وابستگی مورد بررسی قرار گیرد. این اثر تابعی از میزان اطلاعات رمزنگاری شده در آن است. اما نکتهٔ مهم این است که افزایش بی‌رویه در اندازهٔ مجموعه، می‌تواند منجر به تنکی داده و افزایش واژه‌های ناشناخته[3] شود. یکی از راهکارهای پیشنهادشده، غنی‌سازی مجموعه برچسب پایه به کمک اطلاعات ساخت‌واژی است.

این رویکرد بر روی زبان کره‌ای [۲۳] مورد استفاده قرار گرفته است. دلیل استفاده از این رویکرد ماهیت زبان کره‌ای بوده که در آن یک واژه می‌تواند بیش از یک برچسب دریافت کند. به عنوان مثال برای فعل صحبت کردن که در زبان انگلیسی برچسب فعل می‌گیرد، در زبان کره‌ای شامل سه تکواژ با برچسب‌های مختلف است که در شکل (۳-۱) نشان داده شده است. این امر در زبان‌های دیگر نیز اتفاق می‌افتد (به عنوان مثال در انگلیسی فعل buying متشکل از فعل buy و پسوند مستمرساز ing است)، اما این ساخت‌واژه‌ها معمولاً در تجزیه استفاده نمی‌شوند. واضح نیست که چه ترکیبی از این تکواژها بهترین بازنمایی واژه برای

[1] Gold Standard, Optimal, Humman Annotated
[2] Tagset
[3] Unknown word

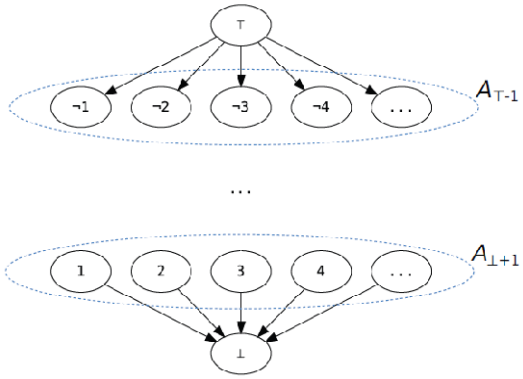
تجزیهٔ وابستگی است.



شکل (۳-۱) تحلیل ساختواژی فعل «صحبت کردن» در زبان کره‌ای [۲۳]

در مرجع [۲۴] بر روی زبان عربی بررسی شده که اطلاعات ساختواژی در قالب برچسب اجزای سخن چطور در صحت تجزیه اثر می‌گذارد. چندین روش خودکار مستقل از متن ارائه شده که به جستجوی ترکیب خصوصیات بهینه می‌پردازد. دو جهت برای این کار استفاده شده است که شمای کلی این دو جهت در شکل (۳-۲) نشان داده شده است:

- جهت اول: شروع از مجموعه برچسب کامل[1] و کاهش اندازه مجموعه با کنار گذاشتن خصوصیاتی که به صحت تجزیه آسیب می‌زنند.
- جهت دوم: شروع از مجموعه برچسب کمینه[2] و افزودن بهترین خصوصیت در هر مرحله به مجموعهٔ خصوصیات.



شکل (۳-۲) شمای کلی دو جهت یافت خصوصیات ساختواژی بهینه [۲۴]

◘ بررسی تأثیر برچسب‌های دستی و خودکار

در تعدادی از مقالات اثرات ناشی از جایگزینی اطلاعات ستون‌ها با نمادگذاری‌های تولید شده توسط ماشین به صورت خودکار بررسی شده است. در میان این اطلاعات، نقش برچسب اجزای سخن در تجزیهٔ وابستگی

[1] Top tagset
[2] Bottom tagset

غیرقابل چشم‌پوشی است که عمدتاً استفاده از برچسب‌هایی که به صورت خودکار تولید می‌شوند منجر به افتی حدوداً ۶ تا ۱۰ درصدی صحت تجزیه می‌شود. تلاش‌ها در این زمینه به دو رویکرد پیاپی و همزمان تقسیم می‌شود.

۱) **رویکرد پیاپی:** در این رویکرد که به صورت سنتی مورد استفاده قرار می‌گیرد، واژه‌ها ابتدا به تحلیل‌گر و ابهام‌زدای ساخت‌صرفی ارائه شده و با اجبار تنها یک تفسیر برای هر واژه به دست می‌آید؛ سپس این تفسیر به تجزیه‌گر ارائه می‌شود. این امر یک حد بالا روی افزایش نرخ خطا ناشی از تفاسیر اشتباه در ابهام‌زدایی نتایج تحلیل‌گر ساخت‌واژی می‌دهد.

در مرجع [۲۵] که بر روی زبان عبری انجام شده است، چند چالش این زبان مورد بررسی قرار گرفته است. یکی از این چالش‌ها انتخاب وندهای مناسب برای تجزیهٔ وابستگی است. بیان شده که جداسازی پیشوندها و پسوندها به جز در حوزه‌های خاصی، اغلب منجر به تولید ابهام بیش‌تر می‌شود. در حوزهٔ تجزیهٔ وابستگی این بدان معناست که روابط وابستگی نه تنها بین واژه‌هایی که با فاصله از هم جدا شدند رخ می‌دهد بلکه بین عنصرهای درون یک واژه نیز رخ می‌دهد. بنابراین اشتباه در قسمت‌بندی واژه‌ها در صحت تجزیه اثر می‌گذارد. آموزش بر روی قسمت‌بندی و برچسب اجزای سخن استاندارد انجام شده و در ادامه، آزمون بر روی دو حالت استفاده از قسمت‌بندی و برچسب اجزای سخن استاندارد و خودکار (به کمک ابهام‌زدای ساخت‌واژی قبل از تجزیه) مورد بررسی قرار گرفته است. نتایج بررسی‌های این مقاله نشان می‌دهد که در زبان عبری استفاده از خصوصیات ساخت‌واژی دستی بهبودی نسبت به عدم استفاده از آن نشان نمی‌دهد و استفاده از خصوصیات پیش‌بینی شده منجر به افت شدید صحت نسبت به حالت عدم استفاده از آن خواهد شد.

در مرجع [۲۶] که بر روی زبان باسک انجام شده، اثرات ناشی از جایگزین کردن خروجی تحلیل‌گر ساخت‌واژی[1] به جای داده‌های دستی نمادگذاری شده در پیکره مورد بررسی قرار گرفته است. (سه حالت آموزش و آزمون روی دادهٔ اصلی، آموزش روی دادهٔ اصلی و آزمون روی دادهٔ خودکار، آموزش و آزمون روی دادهٔ خودکار بررسی شده است). همچنین نشان داده شده که خطا در برچسب اجزای سخن، نسبت به خطا در برچسب ساخت‌واژی (با وجود برچسب اجزای سخن درست) صدمهٔ بیش‌تری به صحت خواهد زد. همچنین ابهام‌زدایی از خروجی تحلیل‌گر ساخت‌واژی به منظور به دست آوردن یک تفسیر برای هر واژه، مسئلهٔ مهمی است.

[1] Morphological analizer

در مراجع [۲۷]، [۲۸] اثر مجموعه برچسب‌های اجزای سخن با اندازه‌های مختلف از ۶ تا ۴۳۰ برچسب بر روی زبان عربی در دو حالت تولید دستی و خودکار بررسی شده است. نتایج این مقاله‌ها نشان می‌دهد، اگر برچسب‌ها به صورت دستی زده شده باشد، بزرگ‌تر شدن مجموعه برچسب‌ها منجر به بهبود صحت خواهد شد اما نتایج در مورد برچسب‌های خودکار کاملاً متفاوت خواهد بود. در این حالت مجموعه برچسب‌های کوچک‌تر کارایی بهتری خواهند داشت.

۲) **رویکرد همزمان:** تلاش‌های اخیر در این زمینه به ترکیب پردازش ساخت‌واژی و صرفی و همچنین برچسب‌زنی اجزای سخن با تجزیهٔ وابستگی معطوف شده است. در واقع مشکل اصلی الگوی پیاپی، انتشار خطا از مرحلهٔ پیش‌پردازش به مرحلهٔ تجزیه است.

در مرجع [۲۹] مدلی برای اجرای همزمان ابهام‌زدایی ساخت‌واژی و تجزیهٔ وابستگی ارائه شده است. در زبان‌هایی که از نظر ساخت‌واژی غنی هستند اغلب بین ساخت‌واژه‌ها و نحو تعامل قابل توجهی وجود دارد، به طوری که هیچ‌کدام را بدون دیگری نمی‌توان ابهام‌زدایی کرد. در واقع با مسئلهٔ مرغ و تخم مرغ روبرو هستیم که در آن بهبود برچسب‌زنی واژه‌های مبهم بدون در نظر گرفتن دانش نحوی منجر به افت صحت شده و از سوی دیگر تجزیه‌گرها، ابهام نحوی را نمی‌توانند بدون اطلاعات ساخت‌واژی دقیق به صورت قابل اعتماد رفع کند. کلید حل این مشکل انجام همزمان هر دو وظیفه است. برای حل این مشکل در این مقاله یک سامانهٔ تجزیهٔ مبتنی بر گراف طراحی شده که نتایج حاصل نشان می‌دهد، استفاده از مدل‌های توأم منجر به بهبود هر دو وظیفه خواهد شد.

بر اساس همین نتیجه در مراجع [۳۰]، [۳۱] سامانه‌های تجزیهٔ مبتنی بر گذار ارائه شدند که همزمان کار برچسب‌زنی اجزای سخن و تجزیهٔ وابستگی را انجام دهد. مدل ترکیب شده با دو چالش مواجه است:

۱) فضای جستجوی مدل ترکیب شده بزرگ‌تر هر کدام از وظایف به تنهایی است و رمزنگاری مؤثر اطلاعات دشوارتر خواهد بود. برای حل این مشکل در روش‌های مبتنی بر گراف راهبرد مؤثر برای هرس‌کردن[1] گراف و در روش‌های مبتنی بر گذار از الگوریتم برنامه‌سازی پویا برای تشخیص حالت‌های تکراری استفاده شده است.

۲) در سامانهٔ تجزیهٔ مبتنی بر گذار به دلیل الگوی چپ به راست گذر از واژه‌ها، مدل توأم نمی‌تواند از برچسب اجزای سخن واژه‌های بعدی برای انتخاب عمل جاری استفاده کند. برای حل این مشکل نیز

[1] Pruning

خصوصیات با تأخیر[1] پیشنهاد شده‌اند که کار تصمیم‌گیری را تا هنگام کسب اطلاعات لازم به تأخیر می‌اندازد.

در میان کارهای انجام شدهٔ رویکرد همزمان، تنها مرجع [۳۱] روی تجزیهٔ وابستگی برچسب‌دار اجرا شده و بقیهٔ روش‌ها برای تجزیهٔ بدون برچسب طراحی شدند. برای ارزیابی چنین سامانه‌ای یک معیار جدید به نام TLAS[2] تعریف شده که درصد واژه‌هایی که همزمان «واژهٔ پدر، برچسب وابستگی و اجزای سخن» را درست پیش‌بینی کرده باشد، نشان می‌دهد.

۳-۲-۲- بازنمایی و مدل‌کردن

ورودی سامانهٔ تجزیه باید بازتاب کننده اطلاعات ساخت‌واژی باشد. در این دسته از تلاش‌ها بررسی شده است که چه اطلاعات ساخت‌واژی باید در مدل تجزیه وارد شود. هدف از این آزمایش‌ها یافتن اطلاعات ساخت‌واژی بهینه برای ارائه به تجزیه‌گر است.

در مراجع [۲۷]، [۲۸] بعد از یافتن مجموعهٔ برچسب بهینه برای زبان عربی، اثر افزودن خصوصیات لغوی به برچسب‌های زده شده بررسی شده است. مجموعهٔ ۹ خصوصیت لغوی موسوم به MADA[3] (معرفگی، شخص، نمود، جهت، وجه، جنسیت، شمار، وضع[4]، حالت) به صورت مجزا و حریصانه به مجموعهٔ برچسب‌ها اضافه شده و بهترین ترکیب ارائه شده است.

در مجموع مفیدترین خصوصیات برای زبان مختلف به صورت زیر است:

۱) خصوصیت «حالت» در بسیاری از زبان‌ها (باسک، عبری، هندی و عربی) نشان داده شده که می‌تواند به بهبود تجزیه کمک کند.

۲) خصوصیت «معرفگی» و «وضع» نشان داده شده که اگر به صورت صریح در بازنمایی مدل شود، برای زبان عبری و عربی مفید خواهد بود.

۳) خصوصیت‌های «وضع»، «نمود» و «وجه» برای تجزیهٔ زبان هندی مفید خواهد بود.

۴) خصوصیت‌های «حالت» و «گونهٔ تبعی[5]» مفیدترین خصوصیات در تجزیهٔ مبتنی بر گذار زبان باسک

---

[1] Delayed feature
[2] Tagged Labled Attachment Score
[3] Morphological Analysis and Disambiguation for Arabic
[4] State
[5] Subordinate type

خواهد بود.

◘ افزودن خصوصیات مفهومی

در دسته‌ای از تلاش‌ها سعی شده تا خصوصیات مفهومی همراه با سایر خصوصیات به تجزیه‌گر ارائه شود. اثر این خصوصیات نیز به دو روش دستی و خودکار مورد بررسی قرار گرفته است.

۱) ارائهٔ دستی خصوصیات مفهومی

یکی از روابطی که تجزیه‌گرها مشکل زیادی در تشخیص آن دارند، مطابقه است. مطابقه به بازیابی ساختارهای داخلی عبارات اسمی کمک می‌کند. در این دسته از روش‌ها سعی شده که به صورت صریح مطابقه را در ستون FEATS اضافه کرده یا خصوصیاتی (مثل انسان/غیرانسان ، جاندار/غیرجاندار، زمان و مکان) که در کشف مطابقه مؤثر هستند را وارد کنند.

در مرجع [۳۲] نشان داده شده است که دو خصوصیت مفهومی جانداری و بی‌جانی، می‌تواند اشتباه بین فاعل و مفعول را کاهش دهد. بر همین اساس در مرجع [۳۳] شش برچسب مفهومی (انسان، غیرانسان، غیرجاندار، زمان، مکان و انتزاع) تعریف و به صورت دستی روی ۱۲۲۱ جملهٔ زبان هندی اعمال و به ستون FEATS اضافه شده است. نشان داده است مقدار کمی اطلاعات مفهومی می‌تواند صحت تجزیه را بهبود دهد. علاوه بر این یک خصوصیت مفهومی به نام GNP متشکل از سه خصوصیت «جنسیت، شمار و شخص» امتحان شده با این امید که نه تنها برای تمایز فاعل و مفعول بلکه برای تشخیص مطابقه مفید واقع شود؛ اما بر خلاف انتظار این خصوصیت نتوانست صحت را بهبود دهد. به منظور مشخص کردن اهمیت کشف مطابقه برای ابهام‌زدایی، به جای کشف مطابقه توسط تجزیه‌گر، سعی شده آن را به صورت صریح در اختیار تجزیه‌گر قرار دهند و همان‌طور که انتظار می‌رفت، این کار منجر به بهبود صحت تجزیه شد.

۲) ارائهٔ خودکار خصوصیات مفهومی

- **خصوصیت مطابقه:** برای آموزش مطابقه به تجزیه‌گر می‌توان خصوصیات ساخت‌واژی مرتبط با مطابقه (جنسیت، شمار، شخص) را به صورت همزمان که معروف به خصوصیت Ø[1] است به تجزیه‌گر ارائه کرد. این روش در زبان عربی مفید بوده اما برای عبری و هندی چندان تأثیرگذار نبوده است.

روش دیگر، کشف خودکار خصوصیت مطابقه از اطلاعات ساخت‌واژی موجود توسط راهکار مستقل از

[1] Ø-feature

زبان است. در این روش، فهرست خصوصیات ساخت‌واژی برای واژه‌های هسته و وابسته مقایسه شده و در صورت داشتن مقادیر یکسان، اطلاعاتی مبتنی بر داشتن مطابقه بین آن‌ها به ستون FEATS اضافه می‌شود [۳۴]، [۳۵].

در مرجع [۳۶] برای زبان عبری نشان داده شده است که اگر الگوی مطابقه مستقیماً در یال‌های وابستگی بازنمایی شود منجر به بهبود تجزیه خواهد شد. ایدهٔ این روش افزودن دو خصوصیت دودویی هنگام ایجاد اتصال ساختارها برای نشان دادن وجود یا عدم خصوصیات «جنسیت» و «شمار» است.

- **خصوصیت جانداری:** در مرجع [۳۷] نشان داده که می‌توان به صورت خودکار خصوصیت جانداری را از داده‌ها استخراج کرد. به عنوان دادهٔ آموزشی ۴۰ اسم (۲۰ اسم جاندار و ۲۰ اسم بی‌جان) با تکرار بیش از ۱۰۰۰، به صورت دستی انتخاب شدند. این اسامی توسط مجموعه‌ای از خصوصیات بازتاب‌کنندهٔ توزیع ساخت‌صرفی آن‌ها بازنمایی و سپس توسط دو الگوریتم یادگیری «درخت تصمیم» و «نزدیک‌ترین همسایه» کار یادگیری انجام شده است.
- **شبکهٔ واژگانی:** شبکهٔ واژگانی[1] یا وردنت پایگاه دادهٔ لغوی است که در آن اسم‌ها، فعل‌ها، صفات و قیود به مجموعه‌ای از مترادف‌های ادراکی[2] گروه‌بندی می‌شوند و هر کدام بیان‌گر مفهوم مجزایی هستند. وردنت نام عمومی است که بر شبکه‌های واژگانی مختلفی در بسیاری زبان‌های جهان اطلاق می‌شود. این شبکه‌ها عموماً در نقش هستان‌شناسی[3] و یا واژگان معنایی محاسباتی در خدمت سامانه‌های هوشمند مبتنی بر دانش و معناگرا قرار دارند.

در شبکهٔ واژگانی زبان انگلیسی که «وردنت» نام دارد، واژه‌ها به مجموعه‌ای از مترادف‌های ادراکی (SS) قسمت‌بندی شدند که هر کدام متعلق به یک «فایل مفهومی»[4] (SF) یکتاست. در مجموع ۴۵ عدد فایل مفهومی (یکی برای قید، سه فایل برای صفت، ۱۵ فایل برای افعال و ۲۶ فایل برای اسامی) بر اساس رده‌های مفهومی و نحوی وجود دارد. فهرست کامل آن‌ها را در فایلی به نام «lexnames» می‌توان یافت که گوشه‌ای از آن در جدول (۳-۱) نشان داده شده است.

[1] WordNet
[2] Synset
[3] Ontology
[4] Semantic file

جدول (۳-۱) فهرست فایل‌های مفهومی موجود در وردنت

طبقه نحوی	فایل مفهومی	توضیح
قید	adv.all	تمام قیدها (Tomorrow)
صفت	adj.all	تمام صفت‌ها (Angry - Beautiful)
	adj.pert	صفت‌های رابطه‌ای
	adj.ppl	صفت‌های مشترک
اسم	noun.act	اسامی نشان‌دهندهٔ عامل یا عمل (Kill)
	noun.animal	اسامی نشان‌دهندهٔ حیوانات (Dog - Elephant)
	...	
	noun.substance	اسامی نشان‌دهندهٔ مواد (Wood – Oil)
	noun.time	اسامی نشان‌دهندهٔ زمان و روابط زمانی (Tomorrow)
فعل	verb.body	افعال آرایش کردن، پوشاندن، مراقبت از بدن (Dress)
	verb.change	افعال اندازه، تغییر دما، تشدید کردن و غیره (Warm)
	...	
	verb.stative	افعال بودن، داشتن (Be – Have)
	verb.weather	افعال باران و برف باریدن، گداختن و رعد و برق زدن (Snow)

در مقالات [۳۸]، [۳۹] از این خصوصیت برای بهبود تجزیهٔ وابستگی استفاده شده است. هر دو حالت شناسهٔ مترادف ادراکی (بازنمایی ریز[1] مفهوم) و فایل مفهومی (بازنمایی درشت[2] مفهوم) امتحان شده است. این دو بازنمایی، حالت‌های حدی از بازنمایی مفهومی هستند. به عنوان یک بازنمایی ترکیبی، اثر ترکیب واژه‌ها با فایل مفهومی متناظرشان (knife+ARTIFACT) نیز امتحان شده است.

از آنجایی که برای هر واژه ممکن است بیش از یک مترادف ادراکی وجود داشته باشد، ابهام در تشخیص آن وجود دارد. به منظور ابهام‌زدایی در مقاله‌های مختلف سه روش به کار برده شده است:

- استفاده از مترادفی که به صورت دستی برچسب خورده است. این خصوصیت حد بالای صحت در فرایند ابهام‌زدایی را داراست.
- استفاده از اولین مترادف: در وردنت معمولاً اولین مترادف ادراکی پرکاربردترین واژه در میان واژه‌های یافت شده آن زبان است.

[1] Fine-grained
[2] Coarse-grained

- استفاده از الگوریتم‌های «ابهام‌زدایی معنایی واژگان[1]»

در مرجع [۳۸] ضمن بررسی این سه روش ابهام‌زدایی، نشان داده شده که روش دوم در عین کم هزینه بودن، کاهش خطای مناسبی نیز دارد.

◘ شکستن جمله به واحدهای زبان‌شناسی مناسب

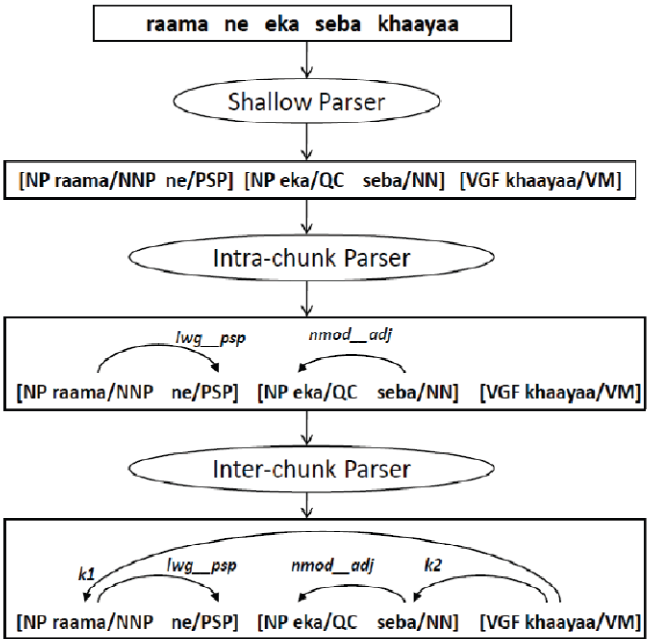
در دسته‌ای از مقالات تلاش شده به جای استفاده از واژه به عنوان واحد پردازشی، تأثیر واحدهای دیگری مورد بررسی قرار گیرد. یکی از انگیزه‌های رفتن به سمت واحدهای زبان‌شناسی جدید، کاهش فضای جستجوی تجزیه‌گرها برای یافتن وابسته‌ها ذکر شده است. در این دسته از راهکارها ساختارهای پیچیده با شکستن فرایند تجزیه به چندین گام مدیریت می‌شود.

- یکی از این انتخاب‌ها، استفاده از قطعه[2] است. در قطعه‌بندی یا تجزیهٔ کم‌عمق[3] جمله به رشته‌ای از واحدهای نحوی قسمت‌بندی می‌شود. این اطلاعات برای کاوش متن و تشخیص موجودیت‌های نامدار[4] مفید است. در مرجع [۴۰] از این اطلاعات برای تجزیهٔ وابستگی زبان انگلیسی استفاده شده است. به این دلیل که قطعه‌بندی می‌تواند صحیح و مؤثر انجام شود، می‌تواند به عنوان یک گام پیش‌پردازش انجام شود. بر همین اساس چهار خصوصیت اضافه شده است:
  - IOB: شامل سه برچسب شروع یک قطعه (B)، درون قطعه (I) و خارج از قطعه (O)
  - EOC: فاصله تا پایان قطعه
  - TYPE: نوع برچسب قطعه
  - NUMB: تعداد قطعه‌ها در یک جمله

در مرجع [۴۱] تلاش شده از اطلاعات قطعه‌ها برای بهبود تجزیهٔ زبان هندی استفاده شود. برای این منظور یک رویه تجزیهٔ دو مرحله‌ای طراحی شده است که در آن ابتدا روابط وابستگی واژه‌های دورن قطعه و ریشهٔ هر قطعه به دست می‌آید. حال تجزیهٔ کامل جمله توسط واژه‌های خارج از قطعات و ریشهٔ هر قطعه انجام می‌شود. شمای کلی این تجزیهٔ دو مرحله‌ای در شکل (۳-۳) نشان داده شده است.

---

[1] WSD: Word Sense Disambiguation
[2] Chunk
[3] Shallow parsing
[4] NER: Named Entity Recogmition



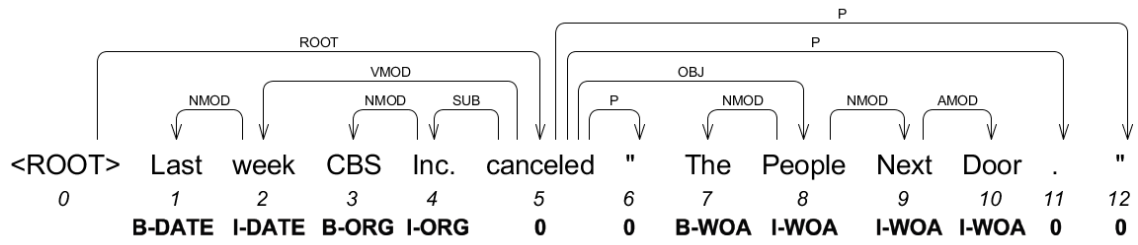
شکل (۳-۳) شمای کلی تجزیهٔ دو مرحله‌ای به کمک اطلاعات قطعات [۴۱]

- یکی دیگر از این واحدها بندها[1] هستند که در مرجع [۴۲] برای بهبود تجزیهٔ وابستگی زبان هندی مورد استفاده قرار گرفته است. بند، گروهی از واژه‌هاست که یک فاعل و یک گزاره دارد. فرایند شناسایی مرز بند را می‌توان یک گام تجزیهٔ جزئی[2] بعد از قطعه‌بندی دانست که سعی دارد جمله را به واحدهای معنادار تقسیم کند.
- از برچسب موجودیت‌های نامدار نیز می‌توان به عنوان واحد پردازشی استفاده کرد. در مرجع [۴۳] از اطلاعات حاصل از برچسب‌گذاری موجودیت‌های نامدار برای بهبود تجزیهٔ وابستگی استفاده شده است. همچنین در پاسخ به این سوال که چرا می‌توان از نمادگذاری‌های مفهومی (مثل برچسب گذاری موجودیت‌های نامدار، برچسب گذاری نقش‌های مفهومی و استخراج روابط) برای بهبود تجزیه استفاده کرد، ذکر شده است که دلیل امکان‌پذیر بودن این امر کارایی بالای آن‌ها بدون نیاز به تجزیهٔ جمله است. شکل (۳-۴) نمونه‌ای از درخت وابستگی با برچسب موجودیت‌های نامدار است. برای این منظور سه خصوصیات جدید بر روی موجودیت‌های نامدار نمادگذاری شده تعریف کرده است:
  - EOS: فاصله تا پایان بخش را نشان می‌دهد. برای واژهٔ Last و canceled به ترتیب ۱ و ۰ خواهد

[1] Clause
[2] Partial parsing

بود.

- BIO: اولین کارکتر برچسب هویت اسمی که نشان می‌دهد واژهٔ ابتدا (B)، داخل (I) یا خارج (O) بخش است. برای واژهٔ CBS و canceled به ترتیب B و O خواهد بود.
- TAG: برچسب کامل هویت اسمی که برای واژهٔ CBS و week به ترتیب B-ORG و I-DATE خواهد بود.



شکل (۳-۴) درخت وابستگی همراه با برچسب موجودیت‌های نامدار [۴۳]

۳-۲-۳- تخمین و هموارسازی

در مقایسه با انگلیسی، زبان‌های از نظر ساخت‌واژی غنی تمایل به تولید لغات بیشتر دارند که این امر منجر به نرخ بالای لغات خارج از واژگان[1] و تشدید مسئلهٔ تنکی داده‌های لغوی[2] می‌شود.

یک راه برخورد با این مسئله، خوشه‌بندی واژه[3] است. خوشه‌بندی واژه روشی برای تولید خوشه‌هایی از واژه‌های مشابه است. «برون[4]» در مرجع [۴۴] الگوریتمی برای خوشه‌بندی واژه ارائه کرد که به نام وی شناخته می‌شود. کلاس‌هایی که نشان‌دهنده گروه‌بندی مبتنی بر نحو یا مفهوم هستند، بسته به ماهیت آماری می‌تواند از متن خام استخراج شود. این روش مبتنی بر مدل زبانی n-گرام بوده و یک الگوریتم «متراکم شوندهٔ پایین به بالا»[5] با پیچیدگی $O(n^3)$ است. خوشه‌های تولیدشده سخت[6] است، به این معنی که هر واژه تنها متعلق به یک خوشه خواهد بود. این خوشه‌ها توسط خوشه‌بندی بی‌ناظر به دست می‌آید که امکان استفاده در پیکرهٔ خام تا میلیون‌ها واژه را دارد.

[1] OOV: out-of-vocabulary
[2] Lexical data sparseness
[3] Word Clustering
[4] Brown
[5] Bottom-up agglomerative
[6] Hard clustering

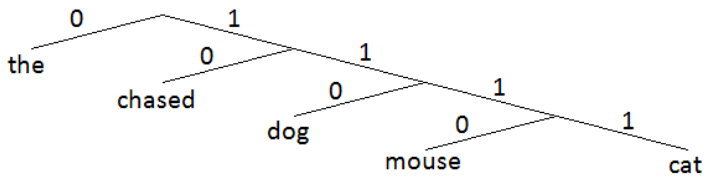
از خوشه‌بندی واژه در مراجع [۴۵]، [۴۶] برای «تشخیص موجودیت‌های نامدار» و در [۴۷]، [۴۸] بـرای تجزیهٔ وابستگی استفاده شده است. خوشه‌بندی واژه با فراهم آوردن بازنمایی با ابعاد کمتر بـه حـل مشـکل تنکی داده کمک می‌کند. در مرجع [۴۹] با بررسی بهبودهای به دست آمده از اعمال خوشه‌بندی واژه نسبت به بسامد واژه، نتیجه گرفته که جایگذاری ستون Word Form با خوشه‌ها، کارایی اتصال را بـرای واژه‌هـایی که در اصل ناشناخته یا با بسامد کم هستند بهبود می‌دهد.

در مرجع [۴۵] پیاده‌سازی «الگوریتم خوشه‌بندی برون» ارائه شده است[1]. ورودی این ابزار متن خـام بـه صورت زیر است.

The cat chased the mouse
The dog chased the cat
The mouse chased the dog

پس از اجرای الگوریتم، فایلی با سه ستون به عنوان خروجی تولید می‌شود که واژه‌های یکتا، رشته بیـت نسبت داده شده و دفعات تکرار آن را نشان می‌دهد. از روی رشته بیت‌های به دست آمده مـی‌تـوان درخـت دودویی متناظر را ساخت که در شکل (۳-۵) نشان داده شده است.

0	the	6
10	chased	3
110	dog	2
1110	mouse	2
1111	cat	2



شکل (۳-۵) خروجی الگوریتم خوشه‌بندی برون و درخت دودویی متناظر با رشته بیت

۳-۲-۴- بررسی تأثیر الگوی نمادگذاری پیکرهٔ وابستگی

یکی از تفسیرهای ارائه شده برای توضیح فاصلهٔ کارایی زبان انگلیسی و زبان‌های از نظـر سـاخت‌واژی غنـی، اختلاف الگوی نمادگذاری پیکره‌های وابستگی بوده است. در مرجع [۴] بیان شده است که هر چند نمی‌توان از تأثیر این عامل چشم‌پوشی کرد اما ماهیت غنای ساخت‌واژی، این دسته از زبان‌ها را مستعد کاهش کارایی کرده است. بنابراین انتظار می‌رود ساختار نمادگذاری زبان‌هایی که خـواص متفـاوت از انگلیسـی دارنـد نیـز متفاوت باشد.

[1] http://cs.stanford.edu/~pliang/software/

در مرجع [۵۰] مطرح شده است که اختلاف صحت تجزیه روی زبان‌های مختلف نمی‌تواند همواره ناشی از اختلاف زبان‌ها باشد. بلکه می‌تواند ناشی از اختلاف اندازه و الگوی نمادگذاری پیکره‌ها در زبان‌های مختلف باشد. اثر اندازهٔ داده را می‌توان با منحنی یادگیری آزمایش‌ها تخمین زد اما نرمال‌کردن الگوی نمادگذاری کار دشواری است. در این مقاله سعی شده است که نمادگذاری موجود در پیکره‌های وابستگی ۲۹ زبان مختلف (از جمله فارسی) به یک الگوی واحد تبدیل شود. برای این منظور فهرستی طولانی از اختلاف‌های نمادگذاری جمع‌آوری و برای هر کدام یک الگوی مشترک ارائه شده است. علاوه بر این مجموعهٔ برچسب‌های ساخت‌واژی و اجزای سخن و همچنین برچسب روابط وابستگی باید یکسان شوند. برای این منظور ابزار HamleDT[1] به زبان پرل تهیه شده است[2].

برای این منظور رویهٔ هماهنگ‌سازی[3] تعریف شده است که شامل شناسایی تمام ساختارهای نحوی است که حداقل در یک پیکره وجود داشته و نمادگذاری آن به صورت باقاعده از سایر پیکره‌ها متمایز باشد. نمادگذاری پیش‌فرض نشأت‌گرفته از الگوی نمادگذاری در پیکرهٔ وابستگی زبان چک است. در تبدیلات ساختاری و برچسب‌گذاری مجدد وابستگی انجام‌شده طی رویه هماهنگ‌سازی، تلاش شده که تبدیلات ساختاری تا حد امکان قابل برگشت باشند؛ اما در مورد برچسب‌گذاری مجدد این امر رعایت نشده است. مجموعه برچسب‌های وابستگی بین پیکره‌ها بسیار متفاوت بود و یکپارچه کردن آن جز با برچسب‌گذاری مجدد امکان پذیر نبود. تنها جنبه‌ای که قصد تغییر آن در رویه وجود نداشته واحدهای زبانی بوده که هدف آن داشتن تعداد گره‌های یکسان در درختان یکسان‌سازی شده و نمادگذاری اصلی است. اما برخی پیکره‌ها از جمله فارسی، عبارت‌هایی شامل چندین واژه را در یک گره قرار دادند. در برخی دیگر از پیکره‌ها یک گره می‌تواند حاوی یک قطعه به جای واژه باشد (مثلاً فعل با متمم‌ها یا اسم با پیش‌وندها و پسوندهایش).

به منظور یکسان کردن مجموعه برچسب‌های اجزای سخن در مرجع [۵۱] مجموعه‌ای شامل ۱۲ طبقه عمومی برچسب اجزای سخن همراه با نگاشت مجموعه برچسب پیکره‌های مختلف به این مجموعه ارائه شده است[4]. اعتقاد نویسنده بر این است که این ۱۲ طبقه پرتکرارترین برچسب‌ها در اکثر زبان‌هاست. از این مجموعه برچسب برای نرمال‌کردن برچسب اجزای سخن پیکره‌های وابستگی مختلف استفاده شده است [۳۵].

[1] HArmonized Multi-LanguagE Dependency Treebank
[2] http://ufal.mff.cuni.cz/hamledt
[3] Harmonization
[4] http://code.google.com/p/universal-pos-tags

۳-۳- نتیجه‌گیری

در این فصل چالش‌ها و دلایل اصلی افت صحت در زبان‌های از نظر ساخت‌واژی غنی مورد بررسی قرار گرفتند و برای حل هر یک راهکارهای ارائه شده در زبان‌های مختلف معرفی شدند. در فصل بعد سعی شده تا مجموعه‌ای از آزمایش‌ها طراحی شود تا چالش‌های مطرح شده در این فصل بر روی زبان فارسی مورد بررسی قرار گیرد.

فصل ۴:

بررسی تجزیهٔ وابستگی زبان فارسی

۴-۱- مقدمه

در این فصل به شرح آزمایش‌های انجام شده طی این پژوهش می‌پردازیم. به دلیل نبود پیشینهٔ کاری در حوزهٔ تجزیهٔ وابستگی باناظر زبان فارسی انتظار می‌رود ابتدا روش پایه‌ای[1] برای ادامهٔ آزمایش‌ها تعریف شود. در ادامه، تنظیمات و اطلاعات استفاده شده برای بررسی هر کدام از چالش‌های مطرح شده طی فصل قبل در زبان فارسی مطرح شده است. به این دلیل که دادهٔ آموزشی استفاده شده نسبتاً کوچک است در تمام آزمایش‌ها از اعتبارسنجی متقابل ۵ باره[2] استفاده شده است.

۴-۲- انتخاب الگوریتم تجزیه

قبل از بررسی روند اجرای آزمایش‌ها، الگوریتم‌های تجزیه و دلایل انتخاب آن‌ها را شرح می‌دهیم. راهکارهای تجزیهٔ وابستگی مبتنی بر داده به دو دستهٔ روش‌های مبتنی بر گذار و مبتنی بر گراف تقسیم می‌شوند. در آزمایش‌های انجام شده، یک الگوریتم به عنوان نمایندهٔ این دو دسته انتخاب شدند. در ادامه فهرستی از پیاده‌سازی‌های موجود در هر دسته ارائه می‌شود:

◘ پیاده‌سازی‌های موجود برای تجزیهٔ وابستگی مبتنی بر گذار

۱) MaltParser[3]: این ابزار شامل ۹ الگوریتم تجزیه و دو الگوریتم یادگیری است که به زبان جاوا پیاده‌سازی شده است. اکثر این الگوریتم‌ها دارای زمان خطی و در بدترین حالت درجهٔ دو هستند و قادر است در هر ثانیه، ۳ جمله را تجزیه کند. این تجزیه‌گر در کار مشترکی سال ۲۰۰۶ با اختلاف اندکی رتبهٔ دوم را کسب کرده است [۵۲]. برخی از الگوریتم‌های این ابزار تنها می‌توانند وابستگی‌های افکنشی را مدیریت کنند که برای حل این مشکل یک رویه پیش و پس‌پردازش قبل از یادگیری و بعد از تجزیه فراهم شده که نتیجهٔ کار را تجزیهٔ شبه‌افکنشی[4] نامند [۵۳].

---

[1] Baseline
[2] 5-fold cross-validation
[3] http://w3.msi.vxu.se/~nivre/research/MaltParser.html
[4] Pseudo-projective

۲) DeSR[1]: پیاده‌سازی به زبان C++ از الگوریتم تجزیهٔ قطعی پایین به بالا که امکان استفاده از الگوریتم‌های یادگیری مختلف را فراهم کرده است. همچنین قوانین خاصی ارائه کرده است که می‌تواند در یک گام و بدون هیچ پیش و پس‌پردازشی وابستگی‌های غیرافکنشی را مدیریت کند. این تجزیه‌گر در کار مشترکی سال ۲۰۰۶ شرکت داشته که تنها به دلیل استفاده از مقدار پیش‌فرض اشتباه برای برچسب‌گذاری یال وابستگی تمام واژه‌های متصل به ریشه، در انتهای فهرست شرکت کنندگان قرار گرفت و در کار مشترک سال ۲۰۰۷ جایگاه هفتم را کسب کرده است [۵۴].

۳) ClearParser[2]: این ابزار پیاده‌سازی از یکی از الگوریتم‌های موجود در MaltParser و به زبان جاوا است که دو راهکار برای بهبود تجزیهٔ وابستگی غیرافکنشی (یکی برای بهبود سرعت تجزیه و دیگری برای بهبود صحت تجزیه) به آن اضافه شده است. زمان تجزیهٔ این روش خطی است که می‌تواند یک جملهٔ انگلیسی با طول متوسط را در ۲.۲۹ میلی ثانیه تجزیه کند [۵۵].

۴) ZPar[3]: هدف از طراحی این تجزیه‌گر بازنمایی غنی خصوصیات ساخت‌واژی با وارد کردن خصوصیات غیرمحلی در تجزیهٔ مبتنی بر گذار بوده است (پیاده‌سازی به زبان C++). این مجموعه شامل خصوصیات تعریف شده از یک واژه، جفت واژه و سه واژه به همراه خصوصیاتی برای بازنمایی فاصله، ظرفیت (تعداد فرزندانی که هر واژه در سمت راست یا چپ خود می‌تواند دریافت کند) است [۵۶].

◘ پیاده‌سازی‌های موجود برای تجزیهٔ وابستگی مبتنی بر گراف

۱) MSTParser: این ابزار شامل دو الگوریتم تجزیه برای درخت‌های افکنشی و غیرافکنشی برای مرتبهٔ اول و دوم است که به زبان جاوا پیاده‌سازی شده است. این الگوریتم در مرتبهٔ اول و دوم به ترتیب دارای پیچیدگی زمانی $O(n^2)$ و $O(n^3)$ است. الگوریتم این ابزار در اصل بدون برچسب طراحی شده و برای برچسب زدن یک رویه دو مرحله‌ای پیش‌بینی شده که ابتدا درخت وابستگی بدون برچسب تولید کرده و سپس برچسب وابستگی را پیش‌بینی می‌کند. این تجزیه‌گر در کار مشترکی سال ۲۰۰۶ رتبهٔ نخست را کسب کرده است [۵۷].

۲) Max-MSTParser[4]: پیاده‌سازی نسخه‌ای از MSTParser به زبان C++ که علاوه بر مرتبهٔ اول و دوم، مرتبهٔ سوم و چهارم تنها در نسخهٔ افکنشی نیز به آن اضافه شده است (در مرجع [۵۸] نشان داده شده

[1] http://sites.google.com/site/desrparser
[2] http://code.google.com/p/clearparser
[3] http://sourceforge.net/projects/zpar
[4] http://max-mstparser.sourceforge.net

که مرتبهٔ دوم غیرافکنشی مسئلهٔ غیر قطعی کامل است و تنها در حالت تقریبی قابل حل است).

۳) Mate-Tools[1]: پیاده‌سازی به زبان جاوا با هدف بهبود سرعت تجزیه نسبت به MSTParser است. امکان موازی‌سازی در استخراج خصوصیات فراهم آورده شده که می‌تواند بر روی چهار هسته تا ۳.۴ برابر سریع‌تر اجرا شود، به طوری که نشان داده شده می‌تواند یک جمله با طول متوسط در زبان انگلیسی را در ۲۷ میلی ثانیه تجزیه کند. همچنین الگوریتم امکان تجزیهٔ وابستگی برچسب‌دار در یک مرحله را داراست [۵۹].

از میان این الگوریتم‌ها، دو ابزار MaltParser و MSTParser چندین مزیت را نسبت به سایر انتخاب‌ها دارند که برای انجام آزمایش‌ها این پژوهش انتخاب شدند. هر دو ابزار دارای الگوریتم‌ها با تنوع مختلف هستند در حالی که سایر ابزارها تنها دارای یک الگوریتم تجزیه بوده که در اکثر موارد توسعه‌ای بر الگوریتم‌های موجود در این دو ابزار هستند. از سوی دیگر به دلیل کسب رتبه‌های نخست در کار مشترک سال ۲۰۰۶ توجه زیادی به خود جلب کردند و اکثر تحقیقات بر روی زبان‌های از نظر ساخت‌واژی غنی به منظور امکان مقایسه با کارهای گذشته، بر روی این دو تجزیه‌گر صورت گرفته است.

علاوه بر این دو دسته الگوریتم، روش دیگری به نام «وابستگی‌های نوعدار»[2] وجود دارد که طی دو مرحله درخت وابستگی را تولید می‌کند. ابتدا جمله را با استفاده از یک تجزیه‌گر مبتنی بر دستور مستقل از متن (مثل تجزیه‌گر استنفورد) به درخت زایشی تبدیل و سپس با قواعد یافت واژهٔ سر[3] به درخت وابستگی تبدیل می‌کند. در مرحلهٔ بعد برچسب این روابط وابستگی را پیش‌بینی می‌کند [۶۰]. این روش در مواقعی که پیکرهٔ وابستگی در دسترس نباشد مطلوب خواهد بود، همچنین از نظر پردازشی (پیچیدگی زمانی حداقل $O(n^5)$) در مقایسه با دو دسته الگوریتم ذکر شده، استفاده از این روش‌ها توصیه نمی‌شود.

به دلیل وجود الگوریتم‌های متنوع در این دو ابزار، آزمایش‌ها ارائه شده فصل بعد را با انتخاب الگوریتم مناسب برای تجزیهٔ زبان فارسی آغاز خواهیم کرد.

ابزار MaltParser شامل الگوریتم‌های مختلفی است که هر کدام پارامترهای خود را دارند. علاوه بر این اطلاعاتی که این ابزار از ستون‌های پیکره استفاده می‌کند، نقش مهمی در کارایی تجزیه‌گر خواهد داشت. نیور[4] [۶۱] اطلاعاتی که MaltParser در مدل خصوصیات خود می‌تواند از آن‌ها استفاده کند را به دو دسته تقسیم کرده است:

---

[1] http://code.google.com/p/mate-tools
[2] Typed dependency
[3] Head Percolation
[4] Nivre

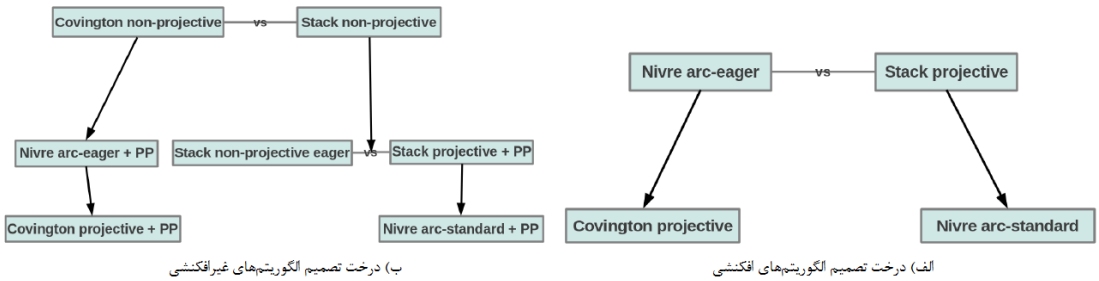
۱) خصوصیات ایستا: این خصوصیات همواره برای ورودی یکسان، خروجی یکسانی تولید می‌کنند. خصوصیاتی مثل Word Form، Lemma، POS، CPOS و FEATS از این دسته هستند.
۲) خصوصیات پویا: خروجی این نوع خصوصیات در طول پردازش می‌تواند تغییر کند. خصوصیاتی مثل HEAD و Deprel از این دسته هستند که در ابتدای تجزیه مقداری ندارند ولی به محض انجام تجزیه دارای مقدار خواهند شد و تا پایان تجزیه می‌توان از مقادیر آن‌ها استفاده کرد. به این دسته از خصوصیات که از روی درخت در حال ساخت وابستگی استخراج می‌شوند، اطلاعات ساخت‌صرفی گویند. در مقالات مختلفی سعی شده اطلاعاتی مثل فاصله تا هسته و ظرفیت واژه را به مجموعهٔ این خصوصیات اضافه کنند.

برای بهینه‌سازی و انتخاب پارامترهای مناسب MaltParser در مراجع [۶۲]، [۶۳] ابزاری به نام MaltOptimizer ارائه شده است[1] که ابتدا تحلیلی بر روی مجموعهٔ آموزش انجام می‌دهد تا نقطهٔ شروع مناسب برای بهینه‌سازی پیدا کند. پس از آن کاربر را از طریق بهینه‌سازی «الگوریتم تجزیه»، «مدل خصوصیات» و «الگوریتم یادگیری» هدایت می‌کند. مراحل کار این ابزار شامل سه فاز زیر است:

- ارزیابی و تحلیل داده: در این فاز اطلاعات آماری از دادهٔ آموزش جمع‌آوری می‌شود که برای تصمیم‌گیری در فازهای بعدی مورد استفاده قرار می‌گیرد.
- انتخاب الگوریتم تجزیه: بر اساس میزان یال‌های غیرافکنشی محاسبه شده در فاز اول، الگوریتم تجزیهٔ مناسب انتخاب و پس از آن پارامترهای الگوریتم انتخاب شده را تنظیم می‌کند.
  - اگر وابستگی‌های غیرافکنشی در دادهٔ آموزشی موجود نباشد، تنها الگوریتم‌های افکنشی که در درخت تصمیم (الف) شکل (۴-۱) نشان داده شده، مورد بررسی قرار می‌گیرند.
  - اگر دادهٔ آموزشی شامل مقداری وابستگی غیرافکنشی باشد، سامانه تنها الگوریتم‌های افکنشی و نسخهٔ شبه‌افکنشی الگوریتم‌های افکنشی درخت تصمیم (ب) شکل (۴-۱) را بررسی می‌کند.
  - اگر تعداد درخت‌ها با یال‌های غیرافکنشی کم باشد اما صفر نباشد، سامانه هر دو درخت تصمیم (الف) و (ب) را بررسی و بهترین الگوریتم را انتخاب می‌کند.
- انتخاب خصوصیات و بهینه‌سازی پارامترهای الگوریتم یادگیری: در این فاز تلاش می‌کند، برای الگوریتمی که در فاز ۲ انتخاب و پارامترهای آن تنظیم شده است، مدل خصوصیات مناسب را پیدا کند. ابتدا از مدل پیش‌فرض شروع کرده و گزینش رو به عقب[2] انجام می‌شود تا تضمین کند تمام خصوصیات

[1] http://nil.fdi.ucm.es/maltoptimizer/install.html
[2] Backward selection

مدل پیش‌فرض را که برای الگوریتم مورد نظر بیشترین سهم دارند، حفظ کند. پس از آن گزینش رو به جلو[1] انجام می‌دهد تا خصوصیات بالقوه مفید و ترکیبات آن‌ها را بیابد. لازم به ذکر است که از حد آستانهٔ ۰.۰۵٪ برای تعیین بهبود صحت تجزیه استفاده شده است.



شکل (۴-۱) درخت‌های تصمیم استفاده شده در MaltOptimizer [۶۳]

در کنار دو دسته روش‌های مبتنی بر گذار و مبتنی بر گراف، راهکارهای ترکیبی نیز وجود دارند که کار ترکیب چندین تجزیه‌گر پایه‌ای را انجام می‌دهد. در ادامه فهرستی از پیاده‌سازی‌های موجود در این دسته ارائه می‌شود:

۱) Ensemble[2]: پیاده‌سازی سه روش ترکیب تجزیه‌گرهای پایه‌ای در زمان تجزیه است که به زبان جاوا نوشته شده است [۶۴]. سه روش ترکیب موجود در این ابزار عبارتند از:

- رأی اکثریت: تمام یال‌ها و برچسب‌های وابستگی پیش‌بینی شده را به عنوان رأی در نظر گرفته و یال‌ها و برچسب‌هایی که بیشترین رأی را دریافت کرده باشند، برای ساخت گراف وابستگی انتخاب می‌کند. نتیجهٔ خروجی، گراف وابستگی خواهد بود به این دلیل که ممکن است همبند نبوده و درختی تولید نشود. این روش با وجود صحت بالا، میزان درخت‌های خوش‌ساخت کم‌تری دارد [۶۵].
- الگوریتم تجزیهٔ مجدد آیزنر[3]: الگوریتم آیزنر در MSTParser برای تجزیهٔ ساختار افکنشی استفاده شده است. در این ابزار نیز از همین الگوریتم برای تجزیهٔ مجدد گراف حاصل از ترکیب پیش‌بینی تجزیه‌گرهای پایه‌ای استفاده شده است. پیچیدگی زمانی این روش $O(n^3)$ است

[1] Forward selection
[2] http://www.surdeanu.name/mihai/ensemble
[3] Eisner

[۶۶].

- الگوریتم تجزیهٔ مجدد اتاردی[1]: این الگوریتم طی یک رویه بالا به پایین و به صورت حریصانه کار ترکیب تجزیه‌گرهای پایه‌ای را با زمان خطی انجام می‌دهد. مراحل کار این الگوریتم در شکل (۴-۲) نشان داده شده است [۶۷].

۲) MaltBlender[2]: پیاده‌سازی یک روش ترکیب تجزیه‌گرهای پایه‌ای در زمان تجزیه است که به زبان جاوا نوشته شده است. برای تجزیهٔ مجدد از الگوریتم چو-لیو-ادموندز[3] استفاده شده است. این الگوریتم نیز درMSTParser برای تجزیهٔ ساختارهای غیرافکنشی استفاده شده است که دارای پیچیدگی زمانی $O(n^3)$ است [۶۸].

۳) MSTParserStacked[4]: پیاده‌سازی ترکیب تجزیه‌گرهای پایه‌ای در زمان آموزش است که به زبان جاوا نوشته شده است. این ابزار توسعه‌ای بر MSTParser است که تجزیهٔ پیش‌بینی شده توسط تجزیه‌گر سطح اول را به دادهٔ آموزشی و آزمون اضافه کرده و تجزیه‌گر سطح دوم با استفاده از داده‌های اصلی و پیش‌بینی شده، کار تجزیه را انجام خواهد داد [۶۹].

در آزمایش‌های انجام شده علاوه بر دو تجزیه‌گر مبتنی بر گذار و گراف معرفی شده، الگوریتم تجزیهٔ مجدد اتاردی نیز استفاده شده تا ترکیب نتایج این دو تجزیه‌گر نیز مورد بررسی قرار گیرد.

۴-۳- شرح آزمایش‌ها

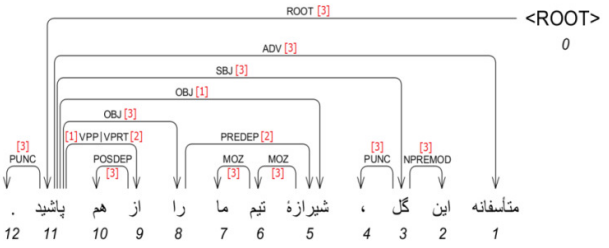
در فصل قبل چالش‌های موجود هنگام تجزیهٔ زبان‌های از نظر ساخت‌واژی غنی شرح داده شد. در این بخش، آزمایش‌های طراحی شده برای بررسی هر کدام از این چالش‌ها در زبان فارسی شرح داده می‌شود.

۴-۳-۱- معماری و تنظیمات اولیه

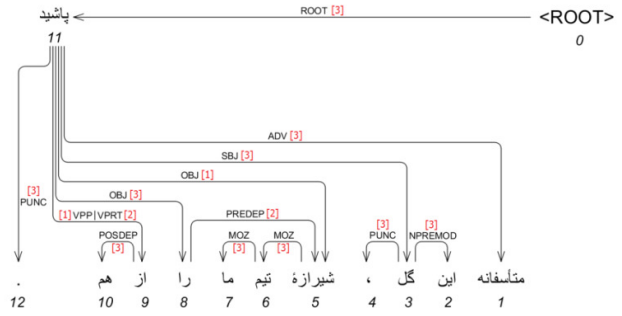
هدف از آزمایش‌های انجام شده در این بخش بررسی تأثیر دو مجموعه برچسب اجزای سخن در حالت دستی

[1] Attardi
[2] http://w3.msi.vxu.se/users/jni/blend
[3] Chu-Liu-Edmonds
[4] http://www.ark.cs.cmu.edu/MSTParserStacked

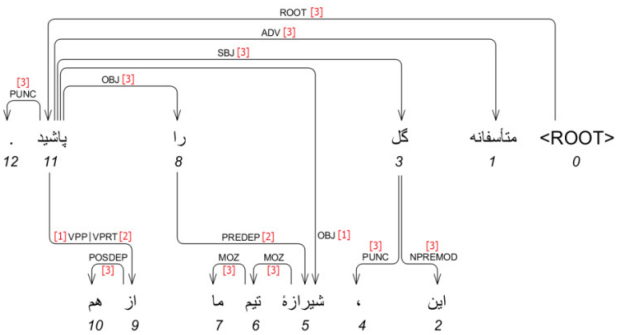
و خودکار است. در این آزمایش‌ها تنها از خصوصیت استفاده شده، برچسب اجزای سخن بوده و ستون FEATS در این آزمایش‌ها پاک شده‌اند.



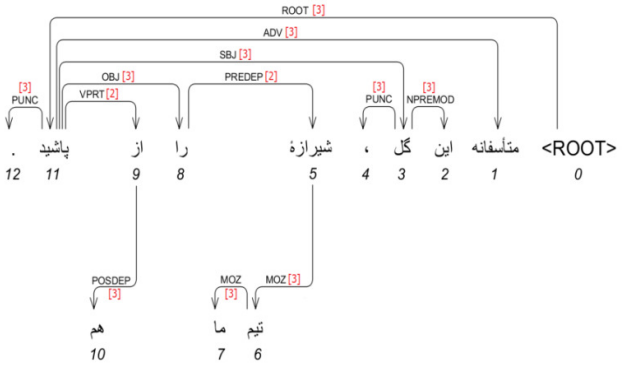
(۱) انتخاب برچسب وابستگی ROOT



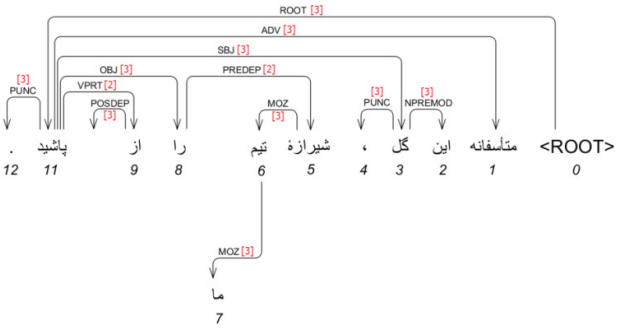
(۲) انتخاب تمام یال‌های وابستگی «پاشید» با بیشترین تعداد رأی



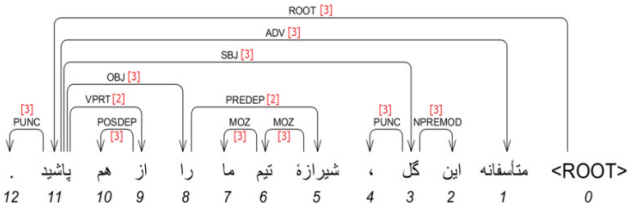
(۳) حذف رابطهٔ OBJ با یک رأی پس از انتخاب رابطهٔ PREDEP



(۴) انتخاب دو برچسب MOZ و POSDEP



(۵) پایان کار الگوریتم با انتخاب آخرین برچسب وابستگی



(۶) درخت وابستگی نهایی

شکل (۴-۲) مراحل کار الگوریتم تجزیهٔ مجدد اتاردی

دو نوع برچسب اجـزای سـخن در سـتون‌هـای CPOS و POS پیکـره وجـود دارد کـه هـر دو الگـوریتم MaltParser و MSTParser تنها از اطلاعات ستون POS استفاده می‌کنند. فهرست کامل این دو برچسب به همراه دفعات تکرار آن‌ها در جدول (۴-۱) آمده است. برخی از این مقادیر در پیکره وجود ندارند که در نتیجه برای ستون‌های CPOS و POS به ترتیب ۱۷ و ۳۰ مقدار مختلف وجود خواهد داشت.

برای تولید خودکار برچسب‌ها از ابزار MXPOST[1] استفاده شده است که پیاده‌سازی برچسب‌زن اجـزای سخن با مدل بیشینهٔ آنتروپی مرجع [۷۰] است. در این آزمایش‌ها دادهٔ آموزشی حاوی برچسب‌های دسـتی هستند و MXPOST بر روی این داده‌ها آموزش داده شـده و برچسـب اجـزای سـخن را بـرای دادهٔ آزمـون پیش‌بینی خواهد کرد.

جدول (۴-۱) برچسب‌های اجزای سخن ستون‌های POS و CPOS پیکرهٔ وابستگی فارسی

برچسب CPOS	دفعات تکرار	برچسب POS	دفعات تکرار
صفت (ADJ)	۱۳۴۴۸	صفت مطلق (AJP)	۱۲۵۸۷
		صفت تفضیلی (AJCM)	۴۲۴
		صفت عالی (AJSUP)	۲۰۰
نقش‌نمای ندا (ADR)	۸۶	نقش‌نمای ندا پیشین (PRADR)	۶۸
		نقش‌نمای ندا پسین (POSADR)	۱۸
قید (ADV)	۴۷۰۸	قید مختص (SADV)	۳۳۸۱
نقش‌نمای همپایگی (CONJ)	۸۶۱۶	نقش‌نمای همپایگی (CONJ)	۸۶۱۴
شاخص (IDEN)	۲۹۲	شاخص (IDEN)	۲۹۲
اسم (N)	۷۱۲۴۷	جاندار (ANM)	۱۰۵۷۰
		بی‌جان (IANM)	۶۱۲۰۸
جزء دستوری (PART)	۱۹۰	جزء دستوری (PART)	۱۹۰
صفت شمارشی پسین (POSNUM)	۱۸۹	صفت شمارشی پسین (POSNUM)	۱۸۹
حرف اضافهٔ پسین (POSTP)	۵۸۴۲	حرف اضافهٔ پسین (POSTP)	۵۷۹۶
ضمیر (PR)	۷۷۱۲	ضمیر شخصی جدا (SEPER)	۵۱۴۹
		شخصی پیوسته (JOPER)	۱۴۴
		اشاره (DEMON)	۸۲۰
		پرسشی (INTG)	۶۸
		بازتابی مشترک (CREFX)	۱۲۱۹

[1] http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html

		ضمیر بازتابی غیرمشترک (UCREFX)	۰
		ضمیر متقابل (RECPR)	۰
پیش توصیف‌گر (PREM)	۴۴۰۰	صفت تعجبی (EXAJ)	۲۲
		صفت پرسشی (QUAJ)	۲۲۷
		صفت اشاره (DEMAJ)	۲۸۸۳
		صفت مبهم (AMBAJ)	۱۲۶۸
صفت شمارشی پیشین (PRENUM)	۲۴۳۷	صفت شمارشی پیشین (PRENUM)	۲۴۱۳
حرف اضافهٔ پیشین (PREP)	۲۱۵۹۱	حرف اضافهٔ پیشین (PREP)	۲۱۵۸۶
شبه‌جمله (PSUS)	۲۱۷	شبه‌جمله (PSUS)	۲۱۷
علامت نگارشی (PUNC)	۱۹۰۱۹	علامت نگارشی (PUNC)	۱۸۸۴۸
فعل (V)	۲۴۳۴۸	معلوم (ACT)	۲۱۶۶۷
		مجهول (PASS)	۲۰۵۶
		وجهی (MODL)	۶۲۷
نقش‌نمای وابستگی (SUBR)	۵۲۲۸	نقش‌نمای وابستگی (SUBR)	۵۱۸۴

۴-۳-۲- بازنمایی و مدل‌کردن

در فصل دوم مقوله‌های دستوری را معرفی کردیم. در این میان مقوله‌های دستوری زیر در پیکـرهٔ وابسـتگی فارسی وجود دارد:

- خصوصیات شخص، شمار، زمان/وجه/نمود در ستون FEATS وجود دارد که دو خصوصیت زمان و وجـه را نیز می‌توان با استفاده از نگاشت جدول (۴-۲) به مجموعهٔ خصوصیات اضافه کرد.

جدول (۴-۲) نگاشت تولید خصوصیات زمان و وجه از روی خصوصیت زمان/وجه/نمود

خصوصیت جدید	مقادیر	مقادیر tma مبدأ	مفهوم	تعداد حضور	درصد حضور
Tense	PA	GEL – GESEL – GBEL – GS – GES – GN – GNES – GB – GBES – GBESE	زمان گذشته	۱۱۱۱۷	٪۵.۸۶
	PR	HEL – HA – H	زمان حال	۱۲۲۲۳	٪۶.۴۵
	FU	AY	زمان آینده	۱۰۰۸	٪۰.۵۳
Mood	Indicative	GB – GNES – GN – GES – GS – H – AY – GBES	وجه اخباری	۱۸۹۹۰	٪۱۰.۰۲
	Subjunctive	GESEL – GEL – HEL – GBESE – GBEL	وجه التزامی	۴۴۸۱	٪۲.۳۶
	Imperative	HA	وجه امری	۸۷۷	٪۰.۴۶

- خصوصیت جانداری (ANM، IANM) برای واژه‌هایی با برچسب اسم در ستون POS موجود است.
- خصوصیت برتری برای واژه‌های با برچسب صفت در ستون POS وجود دارد که سه مقدار صفت مطلق (AJP)، صفت تفضیل (AJCM) و صفت عالی (AJSUP) می‌تواند داشته باشد.
- خصوصیت جهت و وجهیت برای افعال در ستون POS وجود دارد که سه مقدار معلوم (ACT)، مجهول (PASS) و وجهی (MODL) می‌تواند داشته باشد.

علاوه بر این خصوصیات، سه خصوصیت مفهومی دیگر نیز به این مجموعه اضافه شده است.

۱) خوشه‌بندی معنایی فعل: با استفاده از خوشه‌بندی مفهومی افعال به صورت خودکار، از یک سو افزونگی معنایی بین افعال کاهش یافته و از سوی دیگر مشکل تنک بودن داده‌ها را تا حدی حل خواهد شد. چرا که به این ترتیب می‌توان از اطلاعات ردهٔ معنایی فعل برای پیش‌بینی برخی ویژگی‌های افعال که شواهد و داده کافی در مورد آن‌ها در دست نیست، بهره برد. خوشه‌بندی معنای افعال در زبان آلمانی [۷۱] انجام شده که در آن ۴۳ خوشه با متوسط ۳.۹ فعل در هر کلاس، تعریف شده است. برای زبان فارسی [۷۲] ۱۰۸۲ فعل پیکرهٔ بیجن‌خان با فراوانی بالاتر از ۵۰ به ۴۳ خوشه تقسیم شدند که با استفاده از این اطلاعات می‌توان یک خصوصیت به تمام افعال پیکره اضافه کرد.

۲) خصوصیات معنایی با استفاده از شبکهٔ واژگانی: برای تولید شبکهٔ واژگانی در زبان فارسی تلاش‌هایی صورت گرفته است [۷۳–۷۷] که در این پژوهش از ابزار فارس‌نت[1] استفاده شده است. فارس‌نت شامل ۱۰۰۰۰ مجموعهٔ هم‌معنا و ۱۸۰۰۰ واژهٔ فارسی است. کلمات تحت پوشش فارس‌نت دارای ۳ نوع مقولهٔ نحوی واژگانی (اسم، فعل و صفت) هستند که از بین پرخدادترین کلمات زبان فارسی انتخاب شده‌اند. دو خصوصیت مفهومی (شناسهٔ اولین مترادف ادراکی و فایل مفهومی متناظر با آن در وردنت) را می‌توان با استفاده از فارس‌نت به دست آورد.

۳) خوشه‌بندی واژه‌ها: با استفاده از الگوریتم خوشه‌بندی واژهٔ برون، کلیه لغات موجود در پیکره در دو حالت استفاده از واژه و ریشه برای طول بیت‌های مختلف مورد بررسی قرار گرفتند.

نتایج حاصل از اعمال این خصوصیات جدید در جدول (۴-۳) آمده است. افزودن هر کدام از این خصوصیت‌ها منجر به بهبود صحت شده است. در مورد خوشه‌بندی واژه، استفاده از ریشه با طول ۵ بیت بهترین عملکرد را داشته که از این خصوصیت در آزمایش‌ها فصل بعد استفاده شده است.

[1] http://nlp2.sbu.ac.ir/farsnet/

جدول (۴-۳) اثر خصوصیات مفهومی جدید بر صحت تجزیه

روش		MaltParser	MSTParser	Ensemble
حالت پایه‌ای		۸۴.۷۶	۸۴.۵۹	۸۴.۹۸
شناسهٔ خوشهٔ معنایی فعل		۸۴.۸۳	۸۴.۵۷	۸۵.۰۶
شناسهٔ اولین مترادف ادراکی		۸۴.۸۲	۸۴.۵۳	۸۵.۰۴
فایل مفهومی وردنت		۸۴.۸۳	۸۴.۵۳	۸۵.۰۳
استفاده از واژه	۳ بیت	۸۴.۷۸	۸۴.۴۰	۸۴.۹۴
	۵ بیت	۸۴.۸۱	۸۴.۴۷	۸۴.۹۶
	۷ بیت	۸۴.۸۲	۸۴.۵۳	۸۵.۰۲
	۹ بیت	۸۴.۸۲	۸۴.۵۵	۸۵.۰۰
استفاده از ریشه	۳ بیت	۸۴.۸۰	۸۴.۴۲	۸۴.۹۴
	۵ بیت	**۸۴.۸۵**	۸۴.۵۳	۸۴.۹۹
	۷ بیت	۸۴.۸۰	۸۴.۵۴	۸۵.۰۰
	۹ بیت	۸۴.۸۲	**۸۴.۶۳**	۸۵.۰۸

۴-۳-۳- تخمین و هموارسازی

به این دلیل که زبان فارسی از نظر ساخت‌واژی غنی است، درصد لغات خارج از واژگان موجود در دادهٔ آزمون نیز بالا خواهد بود. بنابراین برای استفاده بهتر از اطلاعات لغوی موجود در ستون Word Form و کاهش تنکی داده‌های موجود در آن باید تدبیری اندیشیده شود. در آزمایش‌های این بخش، تأثیر راهکارهای زیر بر صحت تجزیه‌گرها مورد بررسی قرار گرفته است.

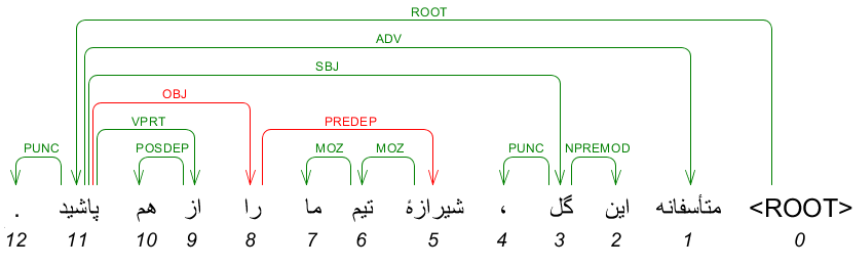
- کپی کردن اطلاعات ستون Lemma: یکی از روش‌ها برای حل این مشکل استفادهٔ مجدد از ریشهٔ کلمات در ستون Word Form است.
- نرمال کردن اعداد: در این روش که به صورت پیش‌فرض توسط MSTParser انجام می‌شود، کلیه اعداد، با برچسب <num> جایگزین خواهند شد. به دلیل وجود اعداد با نوشتار فارسی و انگلیسی در پیکره، این آزمایش در هر دو حالت مورد بررسی قرار گرفته است.
- بلوکه‌بندی اعداد: در مرجع [۲۸] راهکار قسمت‌بندی اعداد به بلوک‌ها استفاده شده است. به عنوان مثال اعدادی که به صفر، ۱ و ۲ ختم می‌شوند هر کدام را می‌توان در بلوک‌های مجزایی قرار داد. هر دو حالت بلوکه‌بندی اعداد فارسی و انگلیسی نیز مورد بررسی قرار گرفته است.

- استفاده از اطلاعات مفهومی: در مراجع [۳۸]، [۳۹] از فایل‌های مفهومی موجود در وردنت برای حل این مشکل استفاده شده است.

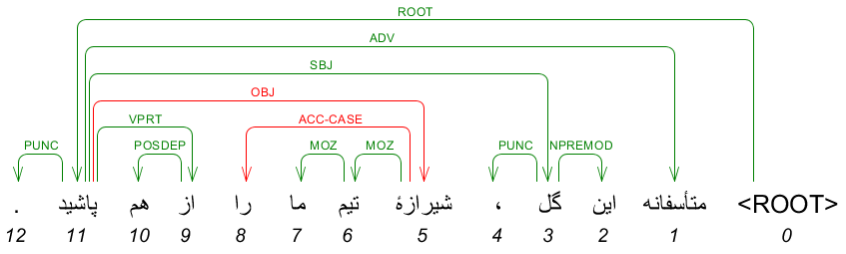
۴-۳-۴- بررسی تأثیر الگوی نمادگذاری پیکرهٔ وابستگی

◘ تغییر نمادگذاری «را»

جملاتی مثل جملهٔ قسمت (الف) شکل (۴-۳) که در آن‌ها مفعول با نشانه «را» در جمله حاضر می‌شود، حرف «را» به عنوان مفعول و واژهٔ «شیرازه»، وابستهٔ پیشین «را» نمادگذاری شده است. تغییر این الگو در قسمت (ب) شکل (۴-۳) نشان داده شده که در آن حرف «شیرازه» مفعول و حرف «را» نشانه مفعولی و وابسته به مفعول نمادگذاری شده است. دو دلیل برای استفاده از این نمادگذاری جدید قابل ذکر است. از یک سو نمادگذاری جدید از نظر زبان‌شناسی صحیح‌تر است و از سوی دیگر می‌تواند منجر به کاهش روابط غیرافکنشی شود زیرا تمام واژه‌های بین فاعل و مفعول زیر مجموعهٔ رابطهٔ آن‌ها خواهند بود.



الف) نمادگذاری پیش‌فرض



ب) نمادگذاری جدید

شکل (۴-۳) نمونهٔ تغییر نمادگذاری «را»

◘ یکسان کردن برچسب وابستگی به ریشه

همان‌طور که در جدول (۴-۴) نشان داده شده است، یال‌های وابستگی به ریشهٔ درخت در نسخهٔ ۰.۱ پیکرهٔ وابستگی، ۱۲ برچسب مختلف دریافت کرده‌اند[۱]. یکی دیگر از تغییرات نمادگذاری بررسی شده، یکسان‌سازی کلیهٔ این برچسب‌ها به ROOT است.

جدول (۴-۴) دوازده برچسب وابستگی به ریشهٔ درخت

نوع برچسب	دفعات وقوع	درصد استفاده
ROOT	۱۲۴۲۸	۹۹.۷۸
VCL	۶	۰.۰۴۸
VCONJ	۴	۰.۰۳۲
NVE	۲	۰.۰۱۶
SBJ	۲	۰.۰۱۶
PREDEP	۳	۰.۰۲۴
NPOSTMOD	۳	۰.۰۲۴
MOZ	۲	۰.۰۱۶
PRD	۲	۰.۰۱۶
PUNC	۱	۰.۰۰۸
POSDEP	۱	۰.۰۰۸
NCONJ	۱	۰.۰۰۸

◘ تغییر نمادگذاری افعال مرکب

فعل مرکب، فعلی است که دارای ساختار ساخت‌واژی ساده نیست بلکه از دو مؤلفهٔ مستقل تشکیل شده است:

- مؤلفهٔ اول (مؤلفهٔ غیرفعلی) را فعل‌یار می‌نامند که می‌تواند اسم، صفت یا قید باشد.
- مؤلفهٔ دوم (مؤلفهٔ فعلی) را همکرد می‌نامند که یک فعل ساده است.

رسولی [۷۸] توزیع فاصلهٔ بین فعل‌یار و همکرد «کردن» در پیکرهٔ بیجن‌خان را مورد بررسی قرار داده و نشان داده که در ۹۱٪ موارد فاصلهٔ بین آن‌ها یک است [۷].

به صورت پیش‌فرض مؤلفه‌های افعال مرکب جداگانه نمادگذاری شدند، اما با توجه به درصد بالای افعال مرکب که فاصله فعل‌یار و همکرد آن‌ها یک است، می‌توان هر دو مؤلفهٔ آن‌ها را با هم به عنوان یک واژه نمادگذاری کرد. برای این منظور از فرهنگ ظرفیت فارسی استفاده شده است. یکی از اطلاعاتی که از این

[۱] این مسئله در نسخهٔ ۱.۰ رفع شده و تمام یال‌های وابستگی به ریشه، برچسب ROOT دریافت کرده‌اند.

پیکره می‌توان استخراج کرد، فهرست فعل‌یارها و حروف اضافهٔ فعلی است که هر فعل می‌تواند دریافت کند. نمونه‌ای از این اطلاعات در جدول (۴-۵) آمده است که بر اساس این اطلاعات برای فعل «افکندن» دو فعل مرکب «پنجه افکندن» و «پی افکندن» می‌توان داشت. در نمادگذاری پیش‌فرض حرف‌اضافهٔ فعلی با برچسب VPRT و فعل‌یار با برچسب NVE به فعل متصل شده‌اند که در نمادگذاری جدید اگر تمام بخش‌های فعل مرکب پشت فعل قرار گرفته باشند به عنوان یک واژه، نمادگذاری می‌شوند. مزیتی که برای این نمادگذاری جدید قابل تصور است، کاهش حداکثر ۲ وابستگی است و به این دلیل که تجزیه‌گرها جملات کوتاه‌تر را با صحت بیش‌تری تجزیه می‌کنند، می‌توان انتظار داشت این نوع نمادگذاری منجر به بهبود صحت شود.

جدول (۴-۵) بخشی از پیکرهٔ ظرفیت فارسی

بن ماضی	بن مضارع	پیشوند	فعل‌یار	حرف‌اضافهٔ فعلی
افکند	افکن	-	-	-
افکند	افکن	-	پنجه	-
افکند	افکن	-	پی	-
پاشید	پاش	-	-	-
پاشید	پاش	-	هم	از

۴-۴- نتیجه‌گیری

در این فصل مفروضات و اطلاعات مورد نیاز برای اجرای آزمایش‌های این پژوهش شرح داده شد و نتایج اجرای آن‌ها در فصل بعد ارائه خواهد شد. ابتدا الگوریتم مناسب برای انجام آزمایش‌ها انتخاب و سپس پارامترهای آن تنظیم خواهد شد. سپس اثر استفاده از دو مجموعه برچسب اجزای سخن در حالت دستی و خودکار بررسی می‌شود. در ادامه مجموعه‌ای از ۱۰ خصوصیت ساخت‌واژی و مفهومی (شخص، شمار، زمان، وجه، زمان/وجه/نمود، اتصال، خوشه‌بندی فعل، خوشه‌بندی واژه، شناسهٔ اولین مترادف ادراکی و فایل مفهومی) تعریف شدند که در آزمایش‌های انجام‌شده تأثیر هر کدام از این خصوصیات به تنهایی و همچنین بهترین ترکیب آن‌ها معرفی خواهد شد. پس از آن چند عامل برای کاهش تنکی داده‌های لغوی پیشنهاد شده است. در پایان نیز پیشنهادهایی برای تغییر نمادگذاری پیکره ارائه شده است.

فصل ۵:

ارائهٔ نتایج و ارزیابی

۵-۱- مقدمه

این فصل را با معرفی معیارهای مختلف ارزیابی در تجزیهٔ وابستگی و معیار استفاده شده در تمام آزمایش‌های این فصل آغاز کرده و سپس به شرح آزمایش‌های انجام شده و ارزیابی نتایج می‌پردازیم. ترتیب آزمایش‌ها و نتایج ارائه شده بر اساس روندی است که در فصل قبل شرح داده شده است.

۵-۲- معیار ارزیابی

برای ارزیابی کارایی یک سامانهٔ تجزیهٔ وابستگی، معیارهای مختلفی وجود دارد که مهم‌ترین آن‌ها عبارتند از:

- امتیاز اتصال بدون برچسب[1] (UAS): درصد یال‌هایی که واژهٔ هسته در آن‌ها به درستی پیش‌بینی شده را نشان می‌دهد.

$$\mathrm{UAS} = \frac{\#\mathrm{ArcsWithCorrectHead}}{\#\mathrm{TotalArcs}} \qquad (۱-۵)$$

- امتیاز اتصال با برچسب[2] (LAS): درصد یال‌هایی که به صورت همزمان واژهٔ هسته و برچسب وابستگی در آن‌ها به درستی پیش‌بینی شده را نشان می‌دهد.

$$\mathrm{LAS} = \frac{\#\mathrm{ArcsWithCorrectHeadAndDeprel}}{\#\mathrm{TotalArcs}} \qquad (۲-۵)$$

- تطبیق کامل[3] (CM): درصد درختانی که به طور کامل درست پیش‌بینی شدند را نشان می‌دهد. این معیار معمولاً مقدار کمتری نسبت به دو معیار قبلی خواهد داشت.

$$\mathrm{CM} = \frac{\#\mathrm{CorrectParsedSentences}}{\#\mathrm{TotalSentences}} \qquad (۳-۵)$$

در تمامی آزمایش‌های انجام شده در این فصل از معیار LAS استفاده شده است.

[1] Unlabeled Attachment Score
[2] Labeled Attachment Score
[3] Complete Match

۵-۳- انتخاب الگوریتم و تنظیم پارامترها

۱) تجزیه‌گر مبتنی بر گذار MaltParser: این تجزیه‌گر شامل ۹ الگوریتم مختلف است که یکی از این الگوریتم‌ها در دو حالت افکنشی و غیرافکنشی قابل اجراست. همچنین به منظور شبه‌افکنشی سازی سه روش پیشنهاد شده که فضای انتخاب‌ها را بسیار بزرگ خواهد ساخت. در جدول (۵-۱) نوع، سال ارائه، مدت زمان یادگیری همراه با صحت پایه‌ای و سه روش شبه‌افکنشی‌سازی ارائه شده است. لازم به ذکر است که اجرای تمامی این الگوریتم‌ها با تنظیمات پیش‌فرض مربوط به خود آن‌ها صورت گرفته است.

جدول (۵-۱) نتایج الگوریتم‌های مختلف موجود در ابزار MaltParser بر روی زبان فارسی

نوع	الگوریتم	سال	صحت پایه‌ای	صحت شبه‌افکنشی‌سازی			زمان یادگیری
				Head	Path	Head+Path	
افکنشی	Arc Eager [79]	۲۰۰۳	۸۱.۶۸	۸۲.۷۸	۸۲.۷۴	۸۲.۶۹	۰۰:۰۰:۵۶
	Arc Standard [80]	۲۰۰۴	۷۹.۸۲	۸۱.۷۴	۸۱.۷۴	۸۱.۷۵	۰۰:۰۱:۲۹
	Covington [81], [82]	۲۰۰۱	۸۱.۶۱	۸۲.۸۴	۸۲.۵۸	۸۲.۶۲	۰۰:۰۱:۰۲
	Stack [83]	۲۰۰۹	۷۹.۹۴	۸۱.۸۴	۸۱.۸۷	۸۱.۸۶	۰۰:۰۰:۵۳
غیرفکنشی	Covington [81], [82]	۲۰۰۱	۸۳.۱۶	۸۲.۹۰	۸۲.۷۷	–	۰۰:۰۵:۱۸
	Stack Eager [83]	۲۰۰۹	۷۷.۲۱	۸۲.۰۶	۸۱.۹۵	۸۱.۹۷	۰۰:۰۱:۱۷
	Stack Lazy [84]	۲۰۰۹	۸۱.۳۱	۸۱.۸۴	۸۱.۸۶	۸۱.۸۶	۰۰:۰۱:۳۳
مسطح	Planar [3]	۲۰۱۰	۸۰.۵۰	۸۱.۱۶	۸۱.۰۴	۸۱.۰۹	۰۰:۰۱:۱۸
	2-planar [3]	۲۰۱۰	۸۰.۶۸	۸۱.۲۸	۸۱.۱۴	۸۱.۲۵	۰۰:۰۱:۲۶

نتایج ارائه شده در این جدول نشان می‌دهد الگوریتم کاوینگتون [۸۱] در حالت غیرافکنشی بهترین کارایی را ارائه کرده است. به همین دلیل در سایر آزمایش‌های انجام شده از این الگوریتم استفاده شده است. روش‌های شبه‌افکنشی می‌توانند الگوریتم‌های افکنشی و مسطح را بیش از یک درصد بهبود بخشند. این نتیجه در مورد الگوریتم‌های غیرافکنشی کمی متفاوت است. الگوریتم Stack Eager را بیش از چهار درصد بهبود داده، الگوریتم Stack Lazy را کم‌تر از یک درصد بهبود داده و در مورد الگوریتم کاوینگتون منجر به افت صحت شده است. از میان روش‌های شبه‌افکنشی‌سازی، روش Head در اکثر موارد نتیجهٔ بهتری نسبت

به دو روش دیگر ارائه می‌دهد.

در ادامه به منظور اجرای بهینه‌سازی، کل دادهٔ آموزشی را در اختیار MaltOptimizer قرار دادیم. نتایج هر یک از فازهای بهینه‌سازی در جدول (۵-۳) آمده است.

- در طی فاز اول ضمن محاسبه برخی اطلاعات آماری، به دلیل کوچک بودن دادهٔ آموزش، آن را برای اجرای اعتبارسنجی متقابل ۵ باره تقسیم می‌کند.
- در فاز دوم به دلیل وجود ساختارهای غیرافکنشی، درخت تصمیم (ب) شکل (۴-۱) مورد بررسی قرار گرفته و در پایان بهترین الگوریتم «کاوینگتون غیرافکنشی» انتخاب شده است. با وجود بررسی و انتخاب بهترین الگوریتم توسط MaltOptimizer، به این دلیل در بخش قبلی الگوریتم‌های مختلف را جداگانه مورد بررسی قرار دادیم که این ابزار تمام الگوریتم‌های موجود در MaltParser را پوشش نداده است.
- در فاز سوم با الگوی خصوصیات پایه‌ای جدول (۵-۲) شروع و سعی شده خصوصیاتی که اثر منفی بر صحت تجزیه دارند از مجموعه حذف شوند که هیچ‌کدام حذف نشدند. سپس سعی شده خصوصیات مفید دیگر به این مجموعه اضافه شود که نتیجهٔ آن افزودن الگوی خصوصیات توسعه‌یافته است.

جدول (۵-۲) الگوی خصوصیات پایه‌ای و توسعه‌یافته برای MaltParser

	پایه‌ای	**توسعه‌یافته**
از یک واژه	$P_{L0}, P_{L1}, P_{R0}, P_{R1}, P_{R2}, P_{R3}, P_{LC0}, P_{RC0}$ $D_{L0}, l(D_{L0}), l(D_{R0}), r(D_{L0}), D_{R0}$ $FM_{L0}, FM_{R0}, FM_{R1}, h(FM_{L0})$	$LM_{L0}, LM_{L1}, LM_{R0}, LM_{R1}$ $s(F_{R0}, \|)$
از دو واژه	$P_{L0} + P_{R0}, P_{L0} + D_{L0}$ $P_{R0} + l(D_{R0}), P_{R0} + D_{R0}$	$P_{R0} + FM_{R0}$
از سه واژه	$P_{L1} + P_{L0} + P_{R0}$ $P_{L0} + P_{R0} + P_{R1}$ $P_{R0} + P_{R1} + P_{R2}$ $P_{R1} + P_{R2} + P_{R3}$ $D_{L0} + l(D_{L0}) + r(D_{L0})$	$P_{R0} + P_{L0} + FM_{R0}$

ساختمان داده‌ها:

Li: عنصر i از بالای پشتهٔ Left — Ri: عنصر i از بالای پشتهٔ Right

LCi: عنصر i از بالای پشتهٔ LeftContext — Ri: عنصر i از بالای پشتهٔ RightContext

توابع:

$l(.)$: وابستهٔ سمت چپ — $r(.)$: وابستهٔ سمت راست

$h(.)$: واژهٔ هسته — $s(., C)$: تقسیم رشته بر اساس کارکتر C

ستون‌ها:

F: FEATS — LM: LEMMA — FM: FORM — D: DEPREL — P: POSTAG

جدول (۵-۳) نتایج سه فاز بهینه‌سازی MaltOptimizer بر روی پیکرهٔ وابستگی زبان فارسی

فاز ۱	فاز ۲	فاز ۳	آزمون
۸۲.۲۵	۸۳.۶۴	۸۵.۰۶	۸۴.۷۶

۲) تجزیه‌گر مبتنی بر گراف MSTParser: این ابزار از دو الگوریتم آیزنر و چو-لیو-ادموندز به ترتیب برای تجزیهٔ روابط افکنشی و غیرافکنشی استفاده می‌کند. همچنین الگوریتم آیزنر در حالت مرتبهٔ دوم توسعه یافته، اما به دلیل غیر قطعی کامل بودن حل مسئله یافت درخت پوشای کمینه غیرافکنشی در حالت مرتبهٔ دوم، تنها تقریبی از الگوریتم چو-لیو-ادموندز برای مرتبهٔ دوم پیاده‌سازی شده است. در جدول (۵-۴) نتایج صحت و زمان یادگیری الگوریتم‌های مختلف موجود در این ابزار نشان داده شده است. با توجه به نتایج ارائه شده، الگوریتم غیرافکنشی در حالت مرتبهٔ دوم بهترین عملکرد را داشته که در سایر آزمایش‌های انجام شدهٔ این فصل از این الگوریتم استفاده شده است.

جدول (۵-۴) نتایج الگوریتم‌های مختلف موجود در ابزار MSTParser بر روی زبان فارسی

مرتبه	نوع	صحت	زمان یادگیری
اول	افکنشی	۸۲.۹۴	۸:۲۲:۳۳
	غیرافکنشی	۸۳.۵۹	۸:۳۷:۳۱
دوم	افکنشی	۸۳.۷۴	۲۱:۳۰:۲۵
	غیرافکنشی	۸۴.۵۹	۲۳:۰۱:۰۴

۵-۳-۲- معماری و تنظیمات اولیه

در جدول (۵-۵) تأثیر دو مجموعه برچسب اجزای سخن در حالت دستی و خودکار نشان داده شده است. صحت پیش‌بینی برچسب CPOS و POS به ترتیب ۹۲.۹۰ و ۸۸.۸۴ درصد است. بیشترین علت افت در پیش‌بینی POS دو مقدار ANM و IANM است که دلیل آن ماهیت مفهومی این برچسب‌هاست.

نتایج به دست آمده در حالت دستی نشان می‌دهد که کاهش تعداد برچسب‌ها منجر به کاهش صحت در MaltParser شده اما صحت MSTParser را نه تنها کاهش نداده بلکه مقدار ناچیزی بهبود نیز داده است.

در حالت خودکار افت ۱۰ درصدی صحت در هر دو تجزیه‌گر مشاهده شده که در این بین استفاده از CPOS به دلیل صحت بالاتر برچسب‌گذاری عملکرد بهتری نسبت به استفاده از POS داشته است. میزان

افت در MSTParser اندکی کمتر از MaltParser بوده است. این بدان معناست که MSTParser اندکی نسبت به رقیب مبتنی گذار خود، در برابر اطلاعات نویزی مقاوم‌تر است. نتایج حاصل از Ensemble نشان می‌دهد که هنگام استفاده از CPOS در هر دو حالت دستی و خودکار صحت را بهبود می‌بخشد. هنگام استفاده از POS تنها در حالت دستی صحت را بهبود می‌دهد اما در حالت خودکار صحت را نسبت به MaltParser بهبود و نسبت به MSTParser اندکی کاهش می‌دهد.

جدول (۵-۵) نتایج دو مجموعه برچسب اجزای سخن در حالت‌های دستی و خودکار

روش		MaltParser	MSTParser	Ensemble
کپی ستون POS در ستون CPOS (۳۰ برچسب)	برچسب دستی	۸۴.۷۳	۸۴.۵۶	۸۴.۸۴
	برچسب خودکار	۷۴.۰۶	۷۴.۷۹	۷۴.۷۸
کپی ستون CPOS در ستون POS (۱۷ برچسب)	برچسب دستی	۸۴.۴۸	۸۴.۵۷	۸۴.۷۰
	برچسب خودکار	۷۴.۳۰	۷۴.۴۶	۷۴.۵۲

۵-۳-۳- بازنمایی و مدل‌کردن

در جدول (۵-۶) میزان تأثیر هر کدام از ۱۰ خصوصیت معرفی شده در فصل قبل بررسی شده است. نتایج نشان می‌دهد که برترین خصوصیت «شمار» است و نکتهٔ قابل توجه حضور دو خصوصیت مفهومی در رتبه‌های بعدی است. همچنین خصوصیت «زمان/وجه/نمود» برتری نسبی در مقایسه با وجه و زمان دارد.

جدول (۵-۶) تأثیر هر یک از ۱۰ خصوصیت معرفی شده بر صحت تجزیه

روش	LAS
شمار (N)	۸۴.۸۰
شناسهٔ خوشه معنایی فعل (VC)	۸۴.۷۷
شناسهٔ اولین مترادف ادراکی (SID)	۸۴.۷۵
شخص (P)	۸۴.۷۵
زمان/وجه/نمود (TMA)	۸۴.۷۵
وجه (M)	۸۴.۷۵
خوشه‌بندی واژه (WC)	۸۴.۷۵
زمان (T)	۸۴.۷۴
فایل مفهومی (SF)	۸۴.۷۴
اتصال (AT)	۸۴.۷۲

در ادامه دو رویهٔ گزینش رو به جلو و رو به عقب به منظور یافت برترین ترکیب این ۱۰ خصوصیت انجام شده که نتایج آن به ترتیب در جدول (۵-۷) و جدول (۵-۸) آمده است.

جدول (۵-۷) گزینش رو به جلو ۱۰ خصوصیت ساخت‌واژی و مفهومی

دو خصوصیت (N)		سه خصوصیت (N + SF)		چهار خصوصیت (N + SF + VC)		پنج خصوصیت (N + SF + VC + SID)	
SF	**۸۴.۸۱**	**VC**	**۸۴.۸۵**	**SID**	**۸۴.۸۷**	M	۸۴.۸۶
WC	۸۴.۸۱	SID	۸۴.۸۴	AT	۸۴.۸۶	TMA	۸۴.۸۶
SID	۸۴.۷۹	WC	۸۴.۸۳	P	۸۴.۸۶	T	۸۴.۸۶
TMA	۸۴.۷۹	M	۸۴.۸۲	M	۸۴.۸۵	P	۸۴.۸۴
M	۸۴.۷۸	P	۸۴.۸۲	T	۸۴.۸۵	AT	۸۴.۸۳
VC	۸۴.۷۷	AT	۸۴.۸۱	TMA	۸۴.۸۴	WC	۸۴.۸۳
P	۸۴.۷۷	TMA	۸۴.۸۱	WC	۸۴.۸۳		
T	۸۴.۷۷	T	۸۴.۸۱				
AT	۸۴.۷۷						

نتایج حاصل از گزینش رو به جلو نشان می‌دهد که چهار خصوصیت «شمار، فایل مفهومی، خوشه‌بندی فعل و شناسهٔ مترادف ادراکی» بهترین ترکیب این ۱۰ خصوصیت است و افزودن سایر خصوصیات منجر به کاهش صحت خواهد شد. نکتهٔ قابل تأمل، حضور سه خصوصیت مفهومی در بین چهار خصوصیت برتر است. نکتهٔ دیگر عدم راه‌یابی خوشه‌بندی واژه به مجموعهٔ این خصوصیات است. برای تولید این خوشه‌بندی از واژه‌های خود پیکره استفاده شده که به دلیل کوچک بودن، خوشه‌های به دست آمده کیفیت لازم را ندارند. با توجه به بی‌ناظر بودن و عدم نیاز به دادهٔ نمادگذاری شده برای خوشه‌بندی واژه، می‌توان آن را بر روی داده‌های بزرگ‌تر اجرا کرده تا نتایج بهتری به دست آید.

صحت به دست آمده از اجرای گزینش رو به عقب اندکی بهتر از گزینش رو به جلوست. در این حالت تنها کنار گذاشتن دو خصوصیت «وجه و شناسهٔ مترادف ادراکی» باعث بهبود کارایی شده است. نکته قابل توجه حضور شناسهٔ مترادف ادراکی در چهار خصوصیت برتر گزینش رو به جلو و دو خصوصیت نامناسب در گزینش رو به عقب است. می‌توان این طور تعبیر کرد که در نبود اطلاعات ساخت‌واژی دستی این خصوصیت اثر مثبت و در حضور آن‌ها اثر منفی بر صحت تجزیه گذاشته است.

جدول (۵-۸) گزینش رو به عقب ۱۰ خصوصیت ساخت‌واژی و مفهومی

استفاده از کل خصوصیات: ۸۴.۸۹					
حذف یک خصوصیات		حذف دو خصوصیات (M)		حذف سه خصوصیات (M + SID)	
M	**۸۴.۹۱**	**SID**	**۸۴.۹۲**	T	۸۴.۹۱
SID	۸۴.۹۱	T	۸۴.۹۰	WC	۸۴.۸۹
T	۸۴.۹۰	VC	۸۴.۸۹	VC	۸۴.۸۹
WC	۸۴.۹۰	AT	۸۴.۸۹	AT	۸۴.۸۹
SF	۸۴.۹۰	WC	۸۴.۸۸	TMA	۸۴.۸۸
TMA	۸۴.۸۹	TMA	۸۴.۸۸	SF	۸۴.۸۸
AT	۸۴.۸۸	SF	۸۴.۸۸	P	۸۴.۸۷
VC	۸۴.۸۷	P	۸۴.۸۷	N	۸۴.۸۳
P	۸۴.۸۵	N	۸۴.۸۴		
N	۸۴.۸۴				

۵-۳-۴- تخمین و هموارسازی

در جدول (۵-۹) نتایج اجرای سه روش معرفی شده در فصل قبل برای کاهش تنکی داده‌های لغوی ارائه شده است. نتایج به دست آمده برای دو تجزیه‌گر متفاوت بود. بهترین روش برای MaltParser و MSTParser به ترتیب بلوکه‌بندی اعداد و کپی اطلاعات ستون Lemma بوده است. نرمال یا بلوکه‌بندی اعداد فارسی در کنار اعداد انگلیسی نه تنها کمکی به بهبود نکرد، بلکه منجر به کاهش صحت نیز شده است. فایل مفهومی وردنت نیز در هر دو تجزیه‌گر صحت تجزیه را کاهش داد که می‌تواند ناشی از نویزی بودن استفاده اولین مترادف ادراکی باشد.

جدول (۵-۹) تأثیر روش‌های مختلف برای کاهش تنکی داده‌های لغوی

روش	MaltParser	MSTParser	Ensemble
کپی کردن اطلاعات ستون Lemma	۸۴.۵۹	**۸۴.۷۵**	۸۴.۸۷
نرمال کردن اعداد انگلیسی	۸۴.۷۹	۸۴.۵۹	۸۵.۰۰
نرمال کردن اعداد فارسی و انگلیسی	۸۴.۷۹	۸۴.۵۰	۸۴.۸۲
بلوکه‌بندی اعداد انگلیسی	**۸۴.۸۰**	۸۴.۵۵	**۸۵.۰۰**
بلوکه‌بندی اعداد فارسی و انگلیسی	۸۴.۸۰	۸۴.۵۲	۸۴.۹۸
استفاده از فایل مفهومی وردنت	۸۴.۵۸	۸۴.۳۱	۸۴.۸۸

در تمام موارد Ensemble صحت را نسبت به هر دو تجزیه‌گر بهبود داده است اما بیشترین میـزان بهبـود هنگام استفاده از نرمال کردن و بلوکه‌بندی به دست آمده است.

۵-۳-۵- تأثیر الگوی نمادگذاری پیکرهٔ وابستگی

در جدول (۵-۱۰) نتایج حاصل از تغییر الگوهای نمادگذاری مطرح شده در فصل قبـل نشـان داده شـده است.

- تغییر نمادگذاری «را»: این تغییر نمادگذاری در هر دو تجزیه‌گر منجر به افت صحت شده است که میزان آن در MSTParser بیش‌تر بوده است. دلایل زیر را می‌توان برای توضیح این افت صحت ذکر کرد:
  - نمادگذاری جدید منجر به تولید روابط وابستگی طولانی‌تر شده که صـحت کمتـری نسـبت بـه وابستگی‌های کوتاه‌تر دارد.
  - یکی از اصلی‌ترین دلایل افت صحت در هر دو تجزیه‌گر عدم توانایی آن‌ها در شناسـایی مفعـول (در MaltParser از ۸۶٪ بـه ۷۶٪ و در MSTParser از ۸۲٪ بـه ۶۸٪ افـت صـحت) اسـت. در نمادگذاری قبلی واژهٔ «را» مفعول بوده و به راحتی قابل شناسایی بود.
- یکسان کردن وابستگی ریشه: این تغییر نمادگذاری اندکی صحت MaltParser را بهبود بخشید اما منجر به افت صحت MSTParser شد.
- تغییر نمادگذاری افعال مرکب: این روش نمادگذاری نیز منجر به افت هر دو تجزیه‌گر شد. یکی از دلایلی که برای این افت صحت می‌توان بیان کرد، افزایش میزان تنکی داده‌های لغوی است. دفعات تکرار افعال مرکب تولید شده بسیار کمتر از اجزایشان است. این امر زمانی باعث بروز مشکل خواهـد شـد کـه فعـل مرکب تولید شده تنها در دادهٔ آزمون ظاهر شده باشد، در حالی کـه اجـزای آن بارهـا در دادهٔ آموزشـی وجود داشتند.

جدول (۵-۱۰) تأثیر تغییر چند الگوی نمادگذاری بر صحت تجزیهٔ وابستگی

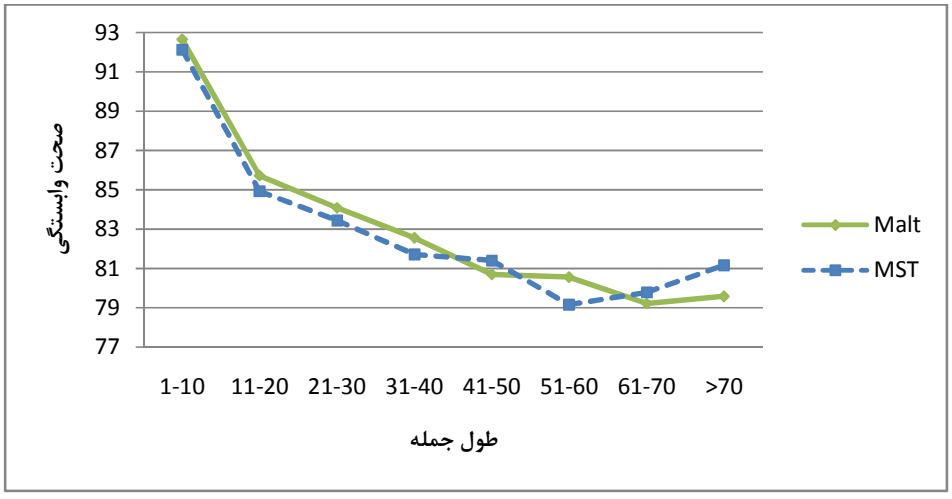
روش	MaltParser	MSTParser	Ensemble
تغییر نمادگذاری «را»	۸۴.۲۹	۸۳.۵۳	۸۴.۳۶
یکسان کردن وابستگی ریشه	**۸۴.۷۸**	۸۴.۵۲	۸۴.۹۷
تغییر نمادگذاری افعال مرکب	۸۳.۸۹	۸۳.۴۶	۸۴.۲۴

۵-۴- تحلیل خطا

به منظور تحلیل خطا، پیکره را به ۸۰ درصد دادهٔ آموزشی و ۲۰ درصد دادهٔ آزمون تقسیم کرده و دو تجزیه‌گر را توسط بهترین تنظیمات کسب شده در آزمایش‌های بخش‌های قبل اجرا کردیم. نتیجهٔ به دست آمده برای MSTParser و MaltParser به ترتیب ۸۴.۸۰ و ۸۵.۳۸ درصد و ترکیب آن‌ها ۸۵.۴۰ درصد بوده است. نتایج به دست آمده را در دو بخش «عوامل مرتبط با طول» و «عوامل زبان‌شناسی» مورد بررسی قرار دادیم.

۵-۴-۱- عوامل مرتبط با طول

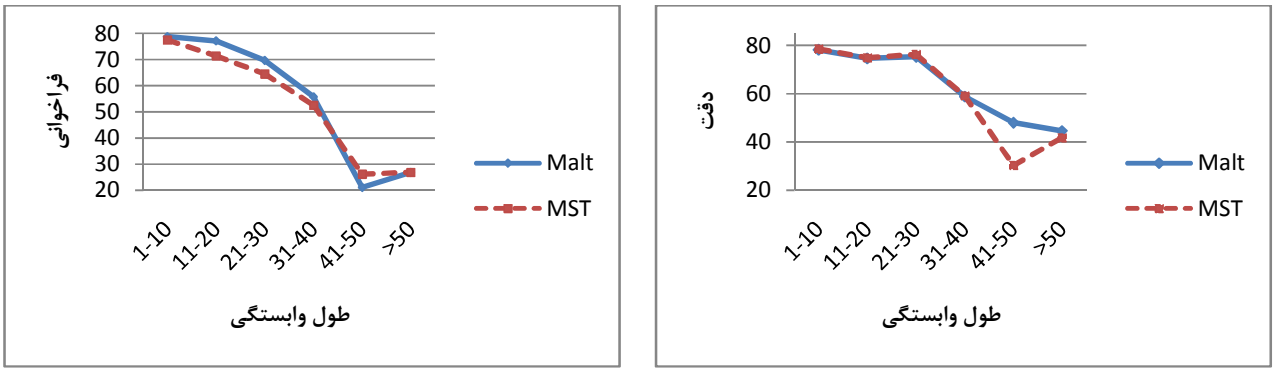
به طور کلی سامانه‌های تجزیه تمایل دارند در جملات طولانی‌تر صحت کمتری داشته باشند که دلیل اصلی آن افزایش حضور ساخت‌های نحوی پیچیده در جملات است. تأثیر این عامل بر صحت وابستگی هر کدام از دو تجزیه‌گر در شکل (۵-۱) نشان داده شده است. در جملات با طول کوتاه‌تر MaltParser اندکی بهتر از MSTParser عمل می‌کند اما با افزایش طول جمله به دلیل ماهیت حریصانهٔ MaltParser خطا منتشر شده و صحت کاهش می‌یابد. در جملات با طول بیش‌تر MSTParser بهتر از MaltParser عمل کرده، به طوری که برای جملات با طول بیش‌تر از ۷۰ این اختلاف به حدود ۱.۵ درصد رسیده است.



شکل (۵-۱) تأثیر طول جمله در صحت وابستگی

با استفاده از نتایج شکل (۵-۱) می‌توان انتظار داشت که استفاده از حد آستانهٔ مناسب در بازهٔ ۶۰ تا ۷۰ برای ترکیب نتایج می‌تواند منجر به بهبود صحت شود. پس از بررسی‌های انجام شده طول ۶۲ به عنوان بهترین حد آستانه انتخاب شد. به این ترتیب برای جملات با طول ۶۲ و کمتر از آن خروجی MaltParser و برای جملات با طول بیشتر از ۶۲ خروجی MSTParser استفاده شد. این ترکیب صحت ۸۵.۴۲ درصدی کسب کرد که اندکی بهتر از صحت ترکیب الگوریتم تجزیهٔ مجدد اتاردی است.

عامل دیگر مؤثر بر صحت، طول وابستگی است. انتظار می‌رود روابط با طول کوتاه‌تر صحت بهتری نسبت به روابط طولانی‌تر داشته باشند. در شکل (۵-۲) تأثیر این عامل بر صحت تجزیهٔ وابستگی نشان داده شده است. درصد یال‌های وابستگی درست با طول l در درخت پیش‌بینی شده را دقت[1] و همین درصد در درختی که به صورت دستی نمادگذاری شده را فراخوانی[2] نامند.



شکل (۵-۲) تأثیر طول وابستگی بر دقت و فراخوانی

۵-۴-۲- عوامل زبان‌شناختی

یکی از عوامل مهم در تحلیل سامانه‌های تجزیهٔ وابستگی، ارتباط آن‌ها با مقوله‌های زبان‌شناختی مثل برچسب اجزای سخن و برچسب وابستگی است. در جدول (۵-۱۱) صحت تجزیه برای برچسب‌های وابستگی درشت نشان داده شده است. دو تجزیه‌گر در برچسب‌های مختلف صحت‌های متفاوتی داشتند.

- MSTParser عملکرد بهتر برای فعل، صفت‌های شمارشی و شاخص
- MaltParser عملکرد بهتر برای اسم، ضمیر، قید و صفت

[1] Precision
[2] Recall

جدول (۵-۱۱) صحت تجزیه در برچسب‌های اجزای سخن درشت

برچسب وابستگی درشت	MaltParser	MSTParser	برچسب وابستگی درشت	MaltParser	MSTParser
صفت شمارشی پسین (POSNUM)	٪۹۱.۷	٪۱۰۰	صفت (ADJ)	٪۸۷.۶	٪۸۶.۴
حرف اضافه پسین (POSTP)	٪۹۵.۱	٪۹۵.۸	ضمیر (PR)	٪۹۲.۵	٪۹۱.۹
پیش توصیف‌گر (PREM)	٪۹۴.۹	٪۹۴.۸	نقش‌نمای همپایگی (CONJ)	٪۸۴.۷	٪۸۲.۷
صفت شمارشی پیشین (PRENUM)	٪۹۴.۳	٪۹۴.۵	قید (ADV)	٪۷۴.۳	٪۷۳.۱
علامت نگارشی (PUNC)	٪۹۲.۹	٪۹۲.۷	نقش‌نمای وابستگی (SUBR)	٪۷۲.۰	٪۷۲.۱
جزء دستوری (PART)	٪۸۸.۹	٪۸۳.۳	شبه‌جمله (PSUS)	٪۶۲.۸	٪۶۲.۸
شاخص (IDEN)	٪۸۸.۹	٪۹۰.۷	حرف اضافه پیشین (PREP)	٪۶۴.۴	٪۶۲.۷
فعل (V)	٪۸۹.۹	٪۹۰.۶	نقش‌نمای ندا (ADR)	٪۷۹.۲	٪۵۸.۳
اسم (N)	٪۸۷.۲	٪۸۶.۶			

در جدول (۵-۱۲) صحت تجزیه برای برخی برچسب‌های وابستگی (اکثراً وابسته‌های فعل) نشان داده شده است. در مجموع MaltParser در تمام این برچسب‌ها نسبت به MSTParser عملکرد بهتری داشته است که دلیلی بر اختلاف عمومی صحت دو تجزیه‌گر است.

جدول (۵-۱۲) صحت تجزیه در برچسب‌های وابستگی

برچسب وابستگی	(Recall – Precision) MaltParser	(Recall – Precision) MSTParser
ریشه جمله (ROOT)	٪۸۹.۴ – ٪۹۵.۹	٪۹۶.۰ – ٪۹۵.۹
فاعل (SBJ)	٪۸۱.۷ – ٪۸۳.۰	٪۸۰.۲ – ٪۸۱.۶
مفعول (OBJ)	٪۸۸.۲ – ٪۸۵.۸	٪۸۸.۳ – ٪۸۳.۵
مسند (MOS)	٪۷۳.۵ – ٪۷۵.۰	٪۷۷.۱ – ٪۷۰.۱
مفعول حرف اضافه‌ای (VPP)	٪۶۹.۳ – ٪۶۲.۸	٪۵۹.۹ – ٪۶۲.۱
تمیز (TAM)	٪۵۵.۸ – ٪۴۰.۰	٪۶۷.۹ – ٪۳۱.۷

۵-۵- نتیجه‌گیری

در این فصل نتایج حاصل از اجرای آزمایش‌های مختلف بر روی تجزیهٔ وابستگی زبان فارسی ارائه شد. ابتدا به تعیین بهترین الگوریتم تجزیه در هر کدام از دو ابزار MaltParser و MSTParser پرداخته شد. پس از آن نشان داده شد که هنگام استفاده از برچسب اجزای سخن، برچسب ریز برای داده‌های دستی و برچسب

درشت برای داده‌های خودکار مناسب است. در گام بعد ۱۰ خصوصیت معرفی شده در فصل قبل بـه صـورت منفرد مورد بررسی قرار گرفت که در نتیجه خصوصیت «شمار» به عنوان با ارزش‌تـرین خصوصـیت معرفـی شد. همچنین به کمک گزینش رو به جلو و عقب بهترین ترکیب این خصوصیات به دسـت آمـد. بـه منظـور کاهش مشکل ناشی از تنکی داده‌های لغوی راهکار بلوکه‌بندی اعداد و استفاده از ریشه کلمه به ترتیب بـرای MaltParser و MSTParser پیشنهاد شدند. در میان روش‌های تغییر نمادگذاری پیکره نیز تنها یکسان‌سازی نمادگذاری وابستگی به ریشه مؤثر واقع شد.

فصل ۶:

جمع‌بندی و کارهای آینده

۶-۱- جمع‌بندی

در آزمایش‌های انجام شده در فصل قبل نشان دادیم که مهم‌ترین چالش در استفادهٔ عملی از تجزیهٔ وابستگی زبان فارسی پیش‌بینی برچسب اجزای سخن است که می‌تواند تا ۱۰ درصد صحت را کاهش دهد. دو راهکار برای حل این مشکل وجود دارد. استفاده از دادهٔ آموزشی بزرگ‌تر به منظور افزایش صحت برچسب‌ها در رویکرد متوالی و یا رفتن به سمت استفاده از رویکرد همزمان تجزیه و برچسب‌زنی که نشان داده شده می‌تواند صحت هر دو وظیفه را بهبود بخشد. همچنین ۱۰ خصوصیت ساخت‌واژی و مفهومی را مورد بررسی قرار دادیم که در مجموع خصوصیت «شمار» به عنوان بهترین خصوصیت معرفی شد. این امر ضرورت تضمین صحت پیش‌بینی شمار در کاربردهای واقعی را پر رنگ‌تر می‌کند. مزیت اصلی خصوصیات مفهومی پیشنهاد شده عدم نیاز به نمادگذاری دستی آن‌هاست. از میان این خصوصیات خوشه‌بندی فعل در هر دو رویکرد گزینش رو به جلو و رو به عقب مؤثر واقع شد. صحت دو خصوصیت مفهومی «شناسهٔ مترادف ادراکی» و «فایل مفهومی» را نیز می‌توان توسط الگوریتم‌های ابهام‌زدایی معنایی واژگان بهبود داد. برای بهبود صحت خوشه‌بندی واژه نیز می‌توان از داده‌های بزرگ‌تر استفاده کرد که به دلیل بی‌ناظر بودن نیازی به نمادگذاری ندارند.

۶-۲- کارهای آینده

در مورد MaltParser سه عامل پارمترهای الگوریتم یادگیری و تجزیه، تهیه الگوی خصوصیات و ارائه بازنمایی اطلاعات به صورت مناسب در فایل ورودی بر صحت تجزیه‌گر تأثیر می‌گذارد. اما در مورد MSTParser تنها چند پارامتر قابل تنظیم وجود دارد و آزمایش‌های انجام شده نشان داد تغییر بازنمایی اطلاعات تأثیر چندانی بر صحت ندارد. به منظور افزایش صحت تجزیه در این ابزار تنها راه تغییر الگوی خصوصیات با تغییر کد آن است و راهی برای تغییر آن بدون تغییر کد منبع وجود ندارد. در مجموع هیچ‌کدام از این دو تجزیه‌گر به طور مناسب از اطلاعات ساخت‌واژی، ساخت‌صرفی و مفهومی که در اختیار آن‌ها قرار داده شده استفاده مناسب نکردند و نیازمند اصلاح الگوی استفاده از خصوصیات با تغییر کد منبع آن‌هاست. همچنین می‌توان روش‌های مختلف هموارسازی را در مدل یادگیری آن‌ها اعمال کرد. در مرجع [۸۵] هموارسازی تجزیهٔ وابستگی بی‌ناظر زبان انگلیسی مورد بررسی قرار گرفته است.

به منظور بهبود صحت برچسب‌های اجزای سخن پیش‌بینی شده، می‌توان از پیکرهٔ بیجن‌خان استفاده کرد که شامل ۲.۶ میلیون واژهٔ برچسب‌گذاری شده است. اگر بخواهیم برچسب‌های دادهٔ آموزش پیکرهٔ وابستگی را حفظ کرده و برچسب‌های دادهٔ آزمون را از روی پیکرهٔ بیجن‌خان پیش‌بینی کنیم نیازمند تعریف نگاشتی از مجموعه برچسب بیجن‌خان به وابستگی هستیم. حتی می‌توان هر دو مجموعهٔ آموزش و آزمون را با برچسب‌های بیجن‌خان پیش‌بینی کرد که در این صورت نیازی به تولید نگاشت نخواهد بود.

در هنگام بررسی عامل طول جمله بر صحت تجزیه، راهکار سادهٔ استفاده از حد آستانه طول جمله برای ترکیب نتایج دو تجزیه‌گر پایه‌ای پیشنهاد و نشان داده شد که اندکی بهتر از روش تجزیهٔ مجدد اتاردی عمل می‌کند. این آزمایش نشان می‌دهد می‌توان به صورت هدفمند از توانایی‌های تجزیه‌گرهای پایه‌ای استفاده کرد. نمونه‌ای از این تلاش در زبان اسپانیایی تحت عنوان تجزیه‌گر n-نسخه‌ای انجام شده است [۸۶] که در آن سعی شده n تجزیه‌گر که هر کدام در پیش‌بینی ساختار وابستگی خاصی مهارت دارد طراحی و آن‌ها را طوری ترکیب کند تا نظر هر تجزیه‌گر در حوزهٔ تخصصی خود بر دیگران ارجحیت داشته باشد. همچنین می‌توان از تجزیه‌گرهای پایه‌ای اطلاعات ساخت‌صرفی بیشتری استخراج کرد تا با استفاده از این خصوصیات صحت تجزیه‌گر ترکیبی را بهبود بخشید.

اخیراً تلاش‌هایی برای برچسب‌گذاری موجودیت‌های نامدار پیکرهٔ بیجن‌خان انجام شده است که می‌تواند برای بهبود تجزیهٔ وابستگی مورد استفاده قرار گیرد. در مراجع [۴۱]، [۴۳] نشان داده شده با استفاده از اطلاعات برچسب و مرز یک موجودیت یا تولید یک تجزیه‌گر دو مرحله‌ای، می‌توان صحت تجزیهٔ وابستگی را بهبود داد. همچنین با وجود این که قطعه‌یاب برای زبان فارسی نداریم، می‌توان هر کدام از تجزیه‌گرهای MaltParser یا MSTParser را ابتدا برای اجرای تجزیهٔ کم‌عمق استفاده نمود و از نتایج به دست آمده برای تجزیهٔ کامل جملات استفاده نمود.

مراجع

[1] S. Kübler, R. McDonald, and J. Nivre, *Dependency parsing*, vol. 1, no. 1. A Publication in the Morgan & Claypool Publishers series, pp. 1–127, 2009.

[2] A. Wróblewska and M. Woliński, "Preliminary experiments in polish dependency parsing", in *Proceedings of the 2011 international conference on Security and Intelligent Information Systems (SIIS 2011)*, pp. 279–292, 2011.

[3] C. Gómez-Rodríguez and J. Nivre, "A transition-based parser for 2-planar dependency structures", in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 1492–1501, 2010.

[4] R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kübler, M. Candito, J. Foster, Y. Versley, I. Rehbein, and L. Tounsi, "Statistical parsing of morphologically rich languages (SPMRL): what, how and whither", in *Proceedings of NAACL HLT 2010 First workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, pp. 1–12, 2010.

[5] M. S. Rasooli, A. Moloodi, M. Kouhestani, and B. Minaei-bidgoli, "A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank", *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 227–231, 2011.

[6] D. Jurafsky and J. H. Martin, *Speech & Language Processing*. Pearson Education India, 2000.

[۷] م. ص. رسولی, "استنتاج بی‌ناظر ظرفیت فعل در زبان فارسی بر مبنای دستور وابستگی"، پایان‌نامهٔ کارشناسی ارشد، دانشگاه علم و صنعت ایران, تهران, ۱۳۹۰.

[8] W. Croft, *Typology and universals*. Cambridge University Press, 2002.

[9] M. Shamsfard, "Challenges and Open Problems in Persian Text processing", in *LTC 2011*, pp. 65–69, 2011.

[10] M. Seraji, B. Megyesi, and J. Nivre, "A Basic Language Resource Kit for Persian", in *Proceedings of Language Resources and Evaluation (LREC 2012)*, 2012.

[11] B. Sagot and G. Walther, "A morphological lexicon for the Persian language", in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[۱۲] ا. طبیب زاده، *ظرفیت فعل و ساخت‌های بنیادین جمله در فارسی امروز*، نشر مرکز، ۱۳۸۵.

[13] ا. طبیب زاده، "مفعول نشانهٔ اضافه‌ای در زبان فارسی"، *نامهٔ فرهنگستان*، شمارهٔ ۲۱، صفحات ۱۰۳-۱۱۷، ۱۳۸۲.

[14] M. Seraji, B. Megyesi, and J. Nivre, "Bootstrapping a Persian Dependency Treebank", *Linguistic Issues in Language Technology*, vol. 7, no. 18, pp. 1–10, 2012.

[15] S. Buchholz and E. Marsi, "CoNLL-X shared task on multilingual dependency parsing", in *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, pp. 149–164, 2006.

[16] J. Nilsson, S. Riedel, and D. Yuret, "The CoNLL 2007 shared task on dependency parsing", in *Proceedings of EMNLP-CoNLL 2007*, pp. 915–932, 2007.

[17] J. Nivre, "Data-driven dependency parsing across languages and domains: perspectives from the CoNLL 2007 shared task", in *Proceedings of the Tenth International Conference on Parsing Technologies*, pp. 168–170, 2007.

[18] S. Kübler, D. Seddah, and R. Tsarfaty, "The First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)", 119 pages, 2010.

[19] D. Seddah, R. Tsarfaty, and J. Foster, "The Second Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2011)", 79 pages, 2011.

[20] R. Farkas, V. Vincze, and H. Schmid, "Dependency Parsing of Hungarian: Baseline Results and Challenges", in *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pp. 55–65, 2012.

[21] M. S. Rasooli and H. Faili, "Fast unsupervised dependency parsing with arc-standard transitions", in *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP (EACL 2012)*, pp. 1–9, 2012.

[22] J. Dehdari and D. Lonsdale, "A link grammar parser for Persian", *Aspects of Iranian Linguistics*, vol. 1, 2008.

[23] J. D. Choi and M. Palmer, "Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing", in *Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2011)*, pp. 1–11, 2011.

[24] J. Dehdari, L. Tounsi, and J. van Genabith, "Morphological Features for Parsing Morphologically-rich Languages: A Case of Arabic", in *Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2011)*, pp. 12–21, 2011.

[25] Y. Goldberg and M. Elhadad, "Hebrew dependency parsing: Initial results", in *Proceedings of the 11th International Conference on Parsing Technologies (IWPT '09)*, pp. 129–133, 2009.

[26] K. Bengoetxea, K. Gojenola, and A. Casillas, "Testing the Effect of Morphological Disambiguation in Dependency Parsing of Basque", in *Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2011)*, pp. 28–33, 2011.

[27] Y. Marton, N. Habash, and O. Rambow, "Improving Arabic dependency parsing with lexical and inflectional morphological features", in *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL)*, pp. 13–21, 2010.

[28] Y. Marton, N. Habash, and O. Rambow, "Improving Arabic dependency parsing with form-based and functional morphological features", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pp. 1586–1596, 2011.

[29] J. Lee, J. Naradowsky, and D. A. Smith, "A discriminative model for joint morphological disambiguation and dependency parsing", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, vol. 1, pp. 885–894, 2011.

[30] J. Hatori, T. Matsuzaki, Y. Miyao, and J. Tsujii, "Incremental joint pos tagging and dependency parsing in chinese", in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2011)*, pp. 1216–1224, 2011.

[31] B. Bohnet and J. Nivre, "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing", in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pp. 1455–1465, 2012.

[32] A. Bharati, S. Husain, B. R. Ambati, S. Jain, D. M. Sharma, and R. Sangal, "Two semantic features make all the difference in parsing accuracy", in *Proceedings of the 6th International Conference on Natural Language Processing (ICON-08)*, vol. 8, 2008.

[33] B. R. Ambati, P. Gade, S. Husain, and G. S. K. Chaitanya, "Effect of Minimal Semantics on Dependency Parsing", in *Proceedings of the Student Research Workshop*, pp. 1–5, 2009.

[34] M. Hohensee and E. M. Bender, "Getting More from Morphology in Multilingual Dependency Parsing", in *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 315–326, 2012.

[35] M. Hohensee, "It's Only Morpho-Logical: Modeling Agreement in Cross-Linguistic Dependency Parsing", University of Washington, 2012.

[36] Y. Goldberg and M. Elhadad, "Easy first dependency parsing of modern Hebrew", in *SPMRL-2010 – a NAACL/HLT workshop on Statistical Parsing of Morphologically Rich Languages*, pp. 103–107, 2010.

[37] L. Øvrelid, *Argument Differentiation. Soft constraints and data-driven models*. University of Gothenburg, 2008.

[38] E. Agirre, T. Baldwin, and D. Martinez, "Improving parsing and PP attachment performance with sense information", *Proceedings of ACL-08: HLT*, pp. 317–325, 2008.

[39] E. Agirre, K. Bengoetxea, K. Gojenola, and J. Nivre, "Improving dependency parsing with semantic classes", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11): shortpapers*, vol. 2, pp. 699–703, 2011.

[40] G. Attardi and F. Dell'Orletta, "Chunking and Dependency Parsing", in *Proceedings of LREC 2008 Workshop on Partial Parsing*, 2008.

[41] B. R. Ambati, S. Husain, S. Jain, D. M. Sharma, and R. Sangal, "Two methods to incorporate local morphosyntactic features in Hindi dependency parsing", in *Proceedings of NAACL HLT 2010 First workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, pp. 22–30, 2010.

[42] P. Gadde, K. Jindal, S. Husain, D. M. Sharma, and R. Sangal, "Improving data driven dependency parsing using clausal information", in *Proceedings Human Language Technologies: The 2010 Annual Conference of the North American Chapter of rhe ACL (HLT 2010)*, pp. 657–660, 2010.

[43] M. Ciaramita and G. Attardi, "Dependency parsing with second-order feature maps and annotated semantic information", in *Proceedings of the 10th Conference on Parsing Technologies*, pp. 133–143, 2011.

[44] P. F. Brown, P. V Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language", *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[45] P. Liang, "Semi-supervised learning for natural language", Massachusetts Institute of Technology, 2005.

[46] S. Miller, J. Guinness, and A. Zamanian, "Name tagging with word clusters and discriminative training", in *In Proceedings of 2004 Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004)*, vol. 4, pp. 337–342, 2004.

[47] T. Koo, X. Carreras, and M. Collins, "Simple semi-supervised dependency parsing", in *Proceedings ACL/HLT*, vol. 8, pp. 595–603, 2008.

[48] K. Sagae and A. S. Gordon, "Clustering words by syntactic similarity improves dependency parsing of predicate-argument structures", in *Proceedings of the 11th International Conference on Parsing Technologies (IWPT '09)*, pp. 192–201, 2009.

[49] M. Candito and D. Seddah, "Parsing word clusters", in *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL '10)*, pp. 76–84, 2010.

[50] D. Zeman, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič, "HamleDT: To Parse or Not to Parse?", in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1–7, 2012.

[51] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset", in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2011.

[52] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi, "MaltParser: A language-independent system for data-driven dependency parsing", *Natural Language Engineering*, vol. 13, no. 02, pp. 95–135, 2007.

[53] J. Nivre and J. Nilsson, "Pseudo-projective dependency parsing", in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pp. 99–106, 2005.

[54] G. Attardi, F. Dell'Orletta, M. Simi, A. Chanev, and M. Ciaramita, "Multilingual dependency parsing and domain adaptation using DeSR", in *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pp. 1112–1118, 2007.

[55] J. D. Choi and M. Palmer, "Getting the most out of transition-based dependency parsing", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11): shortpapers*, vol. 2, pp. 687–692, 2011.

[56] Y. Zhang and J. Nivre, "Transition-based dependency parsing with rich non-local features", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11): shortpapers*, pp. 188–193, 2011.

[57] R. McDonald, K. Lerman, and F. Pereira, "Multilingual dependency analysis with a two-stage discriminative parser", in *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pp. 216–220, 2006.

[58] R. McDonald and F. Pereira, "Online learning of approximate dependency parsing algorithms", in *Proceedings of EACL*, vol. 6, pp. 81–88, 2006.

[59] B. Bohnet, "Very high accuracy and fast dependency parsing is not a contradiction", in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pp. 89–97, 2010.

[60] M. C. De Marneffe, B. MacCartney, and C. D. Manning, "Generating typed dependency parses from phrase structure parses", in *Proceedings of LREC*, vol. 6, pp. 449–454, 2006.

[61] J. Nivre, *Inductive dependency parsing*. Springer Verlag, 2006.

[62] M. Ballesteros and J. Nivre, "MaltOptimizer: An Optimization Tool for MaltParser", in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 58–62, 2012.

[63] M. Ballesteros and J. Nivre, "MaltOptimizer: A System for MaltParser Optimization", in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 23–27, 2012.

[64] M. Surdeanu and C. D. Manning, "Ensemble models for dependency parsing: cheap and good?", in *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference (NAACL-2010)*, pp. 649–652, 2010.

[65] D. Zeman and Z. Žabokrtský, “Improving parsing accuracy by combining diverse dependency parsers”, in *Proceedings of the 9th International Workshop on Parsing Technologies*, pp. 171–178, 2005.

[66] J. M. Eisner, “Three new probabilistic models for dependency parsing: An exploration”, in *the 16th International Conference on Computational Linguistics*, pp. 340–345, 1996.

[67] G. Attardi and F. Dell’Orletta, “Reverse revision and linear tree combination for dependency parsing”, in *Proceedings of NAACL HLT 2009: Short Papers*, pp. 261–264, 2009.

[68] J. Hall, J. Nilsson, and J. Nivre, “Single malt or blended? A study in multilingual parser optimization”, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 933–939, 2007.

[69] A. F. T. Martins, D. Das, N. A. Smith, and E. P. Xing, “Stacking dependency parsers”, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pp. 157–166, 2008.

[70] A. Ratnaparkhi, “A maximum entropy model for part-of-speech tagging”, in *Proceedings of the Empirical Methods in Natural Language Processing Conference*, vol. 1, pp. 133–142, 1996.

[71] S. S. I. Walde, “Experiments on the automatic induction of german semantic verb classes”, *Computational Linguistics*, vol. 32, no. 2, pp. 159–194, 2006.

[٧٢] م. امینیان، ”خوشه‌بندی معنایی افعال زبان فارسی“، پایان‌نامهٔ کارشناسی ارشد، دانشگاه صنعتی شریف، تهران، ١٣٩١.

[73] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi, “Semi Automatic Development Of FarsNet: The Persian Wordnet”, in *Proceedings of 5th Global WordNet Conference (GWA2010)*, vol. 358, 2010.

[74] A. Famian and D. Aghajaney, “Towards Building a WordNet for Persian Adjectives”, *International Journal of Lexicography*, vol. 8, no. 4, pp. 281–303, 2000.

[75] C. I. Davis and D. Moldovan, “Feasibility of Automatically Bootstrapping a Persian WordNet”, vol. 7, no. 6, p. 6, 2010.

[76] F. Keyvan, H. Borjian, M. Kasheff, and C. Fellbaum, “Developing persianet: The persian wordnet”, pp. 315–318, 2006.

[77] M. Montazery and H. Faili, “Automatic Persian WordNet Construction”, in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pp. 846–850, 2010.

[78] M. S. Rasooli, H. Faili, and B. Minaei-bidgoli, "Unsupervised Identification of Persian Compound Verbs", in *Proceedings of the 10th Mexican international conference on Advances in Artificial Intelligence (MICAI 2011)*, vol. Part I, pp. 394–406, 2011.

[79] J. Nivre, "An efficient algorithm for projective dependency parsing", in *Proceedings of IWPT*, 2003.

[80] J. Nivre, "Incrementality in deterministic dependency parsing", in *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together (ACL)*, pp. 50–57, 2004.

[81] M. A. Covington, "A fundamental algorithm for dependency parsing", in *Proceedings of the 39th Annual ACM Southeast Conference*, pp. 95–102, 2001.

[82] J. Nivre, "Constraints on non-projective dependency parsing", in *Proceedings EACL*, pp. 73–80, 2006.

[83] J. Nivre, "Non-projective dependency parsing in expected linear time", in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (IJCNLP) of the AFNLP*, vol. 1, pp. 351–359, 2009.

[84] J. Nivre, M. Kuhlmann, and J. Hall, "An improved oracle for dependency parsing with online reordering", in *Proceedings of the 11th International Conference on Parsing Technologies (IWPT '09)*, pp. 73–76, 2009.

[85] W. P. Headden III, M. Johnson, and D. McClosky, "Improving unsupervised dependency parsing with richer contexts and smoothing", in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*, pp. 101–109, 2009.

[86] M. Ballesteros, J. Herrera, V. Francisco, and G. Pablo, "Giving Shape to an N-Version Dependency Parser", in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010)*, pp. 336–341, 2010.

واژه نامه

بخش الف: واژه نامه فارسی به انگلیسی

ابهام‌زدایی معنایی واژگان ..... Word Sense Disambiguation
امتیاز اتصال با برچسب ..... Labeled Attachment Score
امتیاز اتصال بدون برچسب ..... Unlabeled Attachment Score
بند ..... Clause
بند متممی ..... Complement Clause
بی‌ترتیبی ..... Free word-order
تحلیل گر ساخت‌واژی ..... Morphological Analizer
تشخیص موجودیت‌های نامدار ..... Named Entity Recognition
تطبیق کامل ..... Complete Match
تکواژ ..... Morpheme
تمیز ..... Tamiz
تنکی داده لغوی ..... Lexical Data Sparseness
جانداری ..... Animacy
جنسیت ..... Gender
جهت ..... Voice
حالت ..... Case
حالت ابزاری ..... Instrumental Case
حالت از سویی ..... Ablative Case
حالت کنش‌گری ..... Dative Case
حالت مفعولی ..... Accusative Case
حالت مکانی ..... Locative Case
حالت ندایی ..... Vocative Case
حالت نهادی ..... Nominative Case
حالت وابستگی ..... Genitive Case
خصوصیت با تأخیر ..... Delayed Feature
درجه برتری ..... Degree of comparison
درخت افکنشی ..... Projective Tree
درخت غیرافکنشی ..... Non-Projective Tree
درخت مسطح ..... Planar Tree
دستور پیوندی ..... Link Grammar
دقت ..... Precision

ریشه ..... Stem

زبان‌های از نظر ساخت‌واژی غنی ..... Morphologically Rich Languages

زمان ..... Tense

شبکه واژگانی ..... WordNet

شخص ..... Person

شمار ..... Number

صحت ..... Accuracy

صفت تفضیلی ..... Comparative Degree

صفت عالی ..... Superlative Degree

صفت مطلق ..... Positive Adjective

قطعه ..... Chunk

فاعل ..... Subject

فایل مفهومی ..... Semantic File

فراخوانی ..... Recall

قطبیدگی ..... Polarity

گذرایی ..... Transitivity

متمم ظرفیتی ..... Valency Slot

متمم قیدی ..... Adverbial Complement

مسند ..... Mosnad

مطابقه ..... Agreement

معرفگی ..... Definiteness

مفعول ..... Object

مفعول نشانه اضافه‌ای ..... Ezafe Object

مقوله دستوری ..... Grammatical Category

نمود ..... Aspect

واژه ناشناخته ..... Unknown Word

وجه ..... Mood

وجه اخباری ..... Indicative Mood

وجه التزامی ..... Subjunctive Mood

وجه امری ..... Imperative Mood

وجهیت ..... Modality

وند ..... Affix

هستان‌شناسی ..... Ontology

بخش ب: واژه نامه انگلیسی به فارسی

Ablative Case ..... حالت از سویی
Accuracy ..... صحت
Accusative Case ..... حالت مفعولی
Adverbial Complement ..... متمم قیدی
Agreement ..... مطابقه
Affix ..... وند
Animacy ..... جانداری
Aspect ..... نمود
Case ..... حالت
Chunk ..... قطعه
Clause ..... بند
Comparative Degree ..... صفت تفضیلی
Complement Clause ..... بند متممی
Complete Match ..... تطبیق کامل
Dative Case ..... حالت کنش‌گری
Definiteness ..... معرفگی
Degree of comparison ..... درجه برتری
Delayed Feature ..... خصوصیت با تأخیر
Ezafe Object ..... مفعول نشانه اضافه‌ای
Free word-order ..... بی‌ترتیبی
Gender ..... جنسیت
Genitive Case ..... حالت وابستگی
Grammatical Category ..... مقوله دستوری
Imperative Mood ..... وجه امری
Indicative Mood ..... وجه اخباری
Instrumental Case ..... حالت ابزاری
Labeled Attachment Score ..... امتیاز اتصال با برچسب
Lexical Data Sparseness ..... تنکی داده لغوی
Link Grammar ..... دستور پیوندی
Locative Case ..... حالت مکانی
Modality ..... وجهیت

وجه ..... Mood

تکواژ ..... Morpheme

تحلیل‌گر ساخت‌واژی ..... Morphological Analizer

زبان‌های از نظر ساخت‌واژی غنی ..... Morphologically Rich Languages

مسند ..... Mosnad

تشخیص موجودیت‌های نامدار ..... Named Entity Recognition

حالت نهادی ..... Nominative Case

درخت غیرافکنشی ..... Non-Projective Tree

شمار ..... Number

مفعول ..... Object

هستان‌شناسی ..... Ontology

شخص ..... Person

درخت مسطح ..... Planar Tree

قطبیدگی ..... Polarity

صفت مطلق ..... Positive Adjective

دقت ..... Precision

درخت افکنشی ..... Projective Tree

فراخوانی ..... Recall

فایل مفهومی ..... Semantic File

ریشه ..... Stem

فاعل ..... Subject

وجه التزامی ..... Subjunctive Mood

صفت عالی ..... Superlative Degree

تمیز ..... Tamiz

زمان ..... Tense

گذرایی ..... Transitivity

واژه ناشناخته ..... Unknown Word

امتیاز اتصال بدون برچسب ..... Unlabeled Attachment Score

متمم ظرفیتی ..... Valency Slot

حالت ندایی ..... Vocative Case

جهت ..... Voice

ابهام‌زدایی معنایی واژگان ..... Word Sense Disambiguation

شبکه واژگانی ..... WordNet

Abstract

Data-driven systems can be adapted to different languages and domains easily. Therefore, there are more trends to apply data-driven approaches compared to grammar-based approaches in dependency parsing task. Existence of appreciate corpus which contains sentences and theirs associated dependency trees is the only pre-requirement to use this approaches. Making of such corpus is costly and time consuming. Dependency parsing corpus is existed for about 30 languages currently and Persian language is one of them.

Despite obtaining high accurate results for dependency parsing task in English language, for many of other languages with high free-word order and rich morphology, most applying algorithms lead to drop in accuracy compared to English language. This means that data-driven systems require careful selection of features and tuning of parameters to reach optimal performance. This task is not straightforward enough and needs specific knowledge of system and characteristics of target language.

In this thesis, firstly we have reviewed efforts which are done to resolve this problem in other languages; then, we have tried to apply the algorithms in Persian language to detect effective factors for decreasing parsing accuracy. At the end, we have evaluated a set of 10 morphological and semantic features and we have shown influence of each feature individually to obtain the best combination of them.

Keywords: Dependency Parsing, Morphological and Morphosyntactic Features, Dependency Treebank

Iran University of Science and Technology
School of Computer Engineering

A mechanism for exploring of the effect of different morphologic and morphosyntactic features on Persian dependency parsing

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Engineering - Artificial Intelligence and Robotic

By:
Mojtaba Khallash

Supervisor:
Dr. Behrouz Minaei-Bidgoli

November 2012