برچسب‌زنی خودکار نقش‌های معنایی در جملات فارسی به‌کمک درخت‌های وابستگی

مرتضی رضائی شریف‌آبادی و پروانه خسروی‌زاده فروشانی
گروه زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، تهران، ایران

چکیده

تشخیص خودکار واژه‌های دارای نقش‌های معنایی در جملات و اختصاص صحیح نقش‌های معنایی (همچون کنش‌گر، کنش‌پذیر، منشأ، و ...) به آن‌ها توسط رایانه، می‌تواند موجب بهبود کیفیت در بسیاری از کاربردهای پردازش زبان طبیعی همچون استخراج اطلاعات، پرسش و پاسخ، خلاصه‌سازی، و ترجمهٔ ماشینی شود. در چنین پردازشی که برچسب‌زنی نقش معنایی و یا تجزیهٔ معنایی سطحی خوانده می‌شود، به طور معمول از تجزیهٔ نحوی جملات به‌منظور تعریف ویژگی های نحوی استفاده می شود و نوع بازنمایی نحوی مورد استفاده در دقت سامانهٔ برچسب‌زنی نقش معنایی مؤثر است. در این پژوهش به ارائهٔ برچسب‌زن نقش معنایی مبتنی بر تجزیهٔ نحوی کامل می‌پردازیم. بدین منظور از تجزیه‌گر نحوی وابستگی و روش‌های یادگیری ماشینی استفاده می‌شود. در برچسب‌زن ارائه‌شده سعی شده است که مشکلات برچسب‌زن‌های قبلی ارائه‌شده برای زبان فارس، که همگی مبتنی بر تجزیهٔ نحوی سطحی بوده‌اند، رفع شود و معماری سامانه به برچسب‌زن‌های به‌روز دنیا نزدیک باشد. نتایج پژوهش نشان‌دهندهٔ دقت مناسب سامانهٔ ارائه‌شده است.

واژگان کلیدی: برچسب‌زنی نقش معنایی، تجزیهٔ معنایی سطحی، دستور وابستگی، زبان فارسی، پردازش زبان طبیعی، زبان‌شناسی رایانه‌ای.

۱- مقدّمه

در زبان‌شناسی نوین به‌طورمعمول سطوح مختلفی برای دانش و تحلیل زبانی در نظر گرفته می‌شود و معنا یکی از این سطوح است که به درک مفهوم واژه‌ها و جملات مربوط می‌شود (سعید، ۲۰۰۳). معنا را می‌توان قلب تپندهٔ زبان دانست. به عبارت دیگر زنجیرهٔ آواها و یا حروف، بدون داشتن معنا به‌هیچ‌وجه کارکرد زبان را نخواهند داشت (ریمر، ۲۰۱۰). بنابراین به‌منظور فهم صحیح چگونگی درک و تولید زبان، مطالعهٔ دقیق سطح معنا، که از آن به معناشناسی[1] یاد می‌شود، اهمیت ویژه‌ای برخوردار دارد.

متخصصان پردازش زبان طبیعی[2] نیز در سال‌های اخیر توجه ویژه‌ای به کاربرد تحلیل معنایی (در کنار سطوح پایین‌تر تحلیل مانند آواشناسی[3] و واج‌شناسی[4]، صرف[5] و نحو[6] و سطوح بالاتر مانند کاربردشناسی[7] و گفتمان[8] (جورافسکی و دیگران، ۲۰۰۹) در سامانه‌های رایانه‌ای درک و تولید زبان داشته‌اند. وجود همایش‌های متعدد در حوزهٔ معناشناسی رایانه‌ای که توسط نهادهای معتبر بین‌المللی همچون انجمن زبان‌شناسی رایانه‌ای[9] برگزار می‌شوند، خود شاهدی بر این مدعاست.

تحلیل معنایی، جملات اطلاعاتی را در اختیار رایانه قرار می‌دهد که تمام آن‌ها به‌طور مستقیم از طریق تحلیل نحوی قابل دستیابی نیست. برای مثال دو جملهٔ زیر را در نظر بگیرید:

- علی پنجره را شکست.
- پنجره شکست.

تحلیل نحوی جملات بالا به ما می‌گوید که واژه «پنجره» در جملهٔ نخست مفعول و در جملهٔ دوم فاعل است؛

[1] semantics
[2] natural language processing
[3] phonetics
[4] phonology
[5] morphology
[6] syntax
[7] pragmatics
[8] discourse
[9] Association for Computational Linguistics

اما این اطلاعات نحوی به‌طور مستقیم نشان نمی دهد که رابطهٔ مفهومی موجود برای این واژه در هر دوی این جملات یکسان است. به‌منظور روشن‌کردن این موضوع می توان گفت که «پنجره» در هر دو جمله نقش معنایی[1] «کنش پذیر» (شرکت‌کننده‌ای که تحت تاثیر رویداد جمله تغییر حالت می‌دهد) را دارد (پالمر و دیگران، ۲۰۱۰).

منظور از نقش های معنایی، نقش هایی است که برای نشان‌دادن روابط معنایی یک محمول[2] (به طور معمول فعل) و موضوع های[3] آن (به‌طورمعمول متمم‌ها و افزوده‌های فعل) به کار می رود. پژوهشگران، با توجه به هدف مورد نظرشان، مجموعه نقش های مختلفی را به‌عنوان نقش های معنایی ارائه کرده و به کار گرفته اند. برخی برای هر فعل مشخص نقش های معنایی مخصوصی در نظر می گیرند (برای مثال نقش های «خورنده» و «خورده شده» برای متمم های فعل «خوردن») و برخی دیگر - در سر دیگر طیف - تنها دو نقش عمده که دارای بیشترین ویژگی های کنش گر یا کنش پذیر هستند، قائل می شوند. در میان این دو وضعیت، عدهٔ زیادی از پژوهش‌گران در حدود ده نقش معنایی برمی‌شمارند (گیلدِی و دیگران، ۲۰۰۲). گروهی از نقش های معنایی شناخته‌شده عبارتند از کنش‌گر[4] (آغازکنندهٔ عمل، دارای اراده)، کنش پذیر[5] (تحت تأثیر عمل، اغلب دچار تغییر وضعیت)، کنش‌بر[6] (در حال حرکت و یا دارای مکان مشخص)، تجربه‌گر[7] (تجربه‌کننده عمل در افعالی چون دید، شنید، درک کرد ...)، بهره‌ور[8] (شخص یا چیزی که عمل فعل به نفع او انجام می شود)، ابزار[9] (واسطه یا ابزار مورد استفاده برای انجام عمل)، مکان[10] (مکان شیء یا عمل)، منشأ[11] (نقطه آغاز)، و مقصد[12] (نقطهٔ پایان) (پالمر، ۲۰۱۰).

برچسب‌زنی نقش معنایی[13] را، که عبارت است از اختصاص خودکار نقش‌های معنایی به واژه‌ها و گروه‌ها نسبت به یک محمول مشخص در جمله، می‌توان همچون بسیاری از فعالیت های دیگر در پردازش زبان طبیعی با روش مبتنی بر قاعده انجام داد؛ اما با تولید داده های دارای اطلاعات معنایی و ارائهٔ برچسب زن های آماری نقش های معنایی، تحولی بزرگ در این حوزه رخ داد و روش‌های آماری جایگزین روش‌های مبتنی بر قاعده شدند.

در صورت داشتن منابع زبانی دارای اطلاعات معنایی، فرایند ایجاد یک برچسب زن نقش معنایی با استفاده از روش‌های آماری را می‌توان مشابه سایر فعالیت‌های طبقه‌بندی[14] بر اساس یادگیری ماشینی[15] تعریف کرد (پالمر و دیگران، ۲۰۱۰). به این صورت که وظیفهٔ برچسب‌زن این خواهد بود که از مجموعهٔ نقش‌های معنایی ازپیش تعریف‌شده، نقشی را برای گروه های نحوی مرتبط با محمول مد نظر انتخاب کند. برای این منظور برچسب زن با درنظرگرفتن ویژگی های استخراج‌شده برای هر گروه نحوی و برچسب نقش معنایی تعلق‌گرفته به آن در دادهٔ آموزش، توسط یکی از روش‌های مرسوم یادگیری ماشینی یک طبقه بند را آموزش خواهد داد که در آن هر برچسب نقش معنایی یک طبقه به حساب می آید. حال با دادن یک جملهٔ ورودی جدید، سامانه ویژگی های مربوط به گروه های نحوی آن جمله را بررسی کرده و گروه مورد نظر را در یکی از طبقات تعریف‌شده قرار خواهد داد و یا به عبارت دیگر یکی از برچسب‌های نقش معنایی را به آن نسبت خواهد داد.

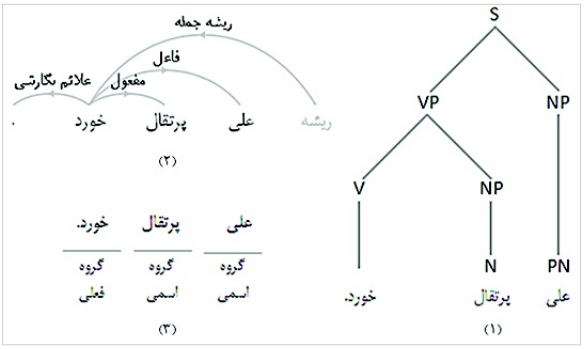
برخلاف اهمیت برچسب زنی خودکار نقش معنایی در پردازش زبان طبیعی و درک و تولید رایانه‌ای زبان، تعداد پژوهش های انجام گرفته با هدف برچسب‌زنی نقش های معنایی در زبان فارسی انگشت‌شمار است.

۲- اهمیت تجزیهٔ نحوی در برچسب‌زنی نقش معنایی

ساختار نحوی جملات، اطلاعات ارزشمندی را به‌منظور تشخیص صحیح نقش‌های معنایی در اختیار سامانه‌های برچسب‌زنی خودکار نقش‌های معنایی قرار می‌دهند. برای مثال اطلاع از انواع گروه‌های به کار رفته در جمله (گروه اسمی، فعلی، ...) و یا روابط نحوی حاکم در جمله (فاعل، مفعول، ...) بسیار مفید است و بنابراین اغلب این سامانه‌ها ابتدا از طریق یک تجزیه گر نحوی، ساختار نحوی جملات ورودی را مشخص می‌کند. با توجه به این موضوع، نوع بازنمایی نحوی مورد استفاده در برچسب‌زن نقش‌های معنایی، یکی از مسائل مهمی است که باید درخصوص آن تصمیم‌گیری شود.

[1] semantic role
[2] predicate
[3] arguments
[4] agent
[5] patient
[6] theme
[7] experiencer
[8] beneficiary
[9] instrument
[10] location
[11] source
[12] goal
[13] semantic role labeling
[14] classification
[15] machine learning

به‌طور کلی می‌توان تجزیهٔ نحوی مورد استفاده برای امر برچسب‌زنی نقش‌های معنایی را به دو دستهٔ عمده که عبارتند از تجزیهٔ نحوی کامل[1] و تجزیهٔ نحوی سطحی[2] تقسیم کرد. در تجزیه نحوی کامل، درخت‌های نحوی کامل مورد استفاده قرار می‌گیرند و ساختار و روابط تمام اجزای جمله مشخص است. این درخت‌ها یا مبتنی بر دستور ساخت‌سازه‌ای[3] و یا بر اساس دستور وابستگی[4] هستند. از سوی دیگر در تجزیهٔ نحوی سطحی، جملات تنها به گروه‌های نحوی (گروه‌های اسمی، فعلی،...) تقسیم می‌شوند. در این نوع تجزیه، نقش هر یک از گروه‌ها در جمله و همچنین روابط میان اجزای درونی گروه‌ها مشخص نمی‌شود. تفاوت میان انواع تجزیهٔ نحوی در شکل (۱) قابل ملاحظه است.



(شکل- ۱): انواع تجزیهٔ نحوی در برچسب‌زنی نقش معنایی (۱- تجزیهٔ نحوی کامل بر مبنای دستور ساخت‌سازه‌ای ۲- تجزیهٔ نحوی کامل بر مبنای دستور وابستگی ۳- تجزیهٔ نحوی سطحی)

پژوهش‌ها نشان داده است که سامانه‌های برچسب‌زنی نقش معنایی که از تجزیه‌گرهای نحوی کامل استفاده می‌کنند، نسبت به سامانه‌هایی که به تجزیهٔ نحوی سطحی بسنده می‌کنند، نتایج بهتری به همراه داشته‌اند (پانیاکونوک، ۲۰۰۵). با این حال با توجه به‌سرعت بیشتر تجزیه‌گرهای نحوی سطحی و کاربرد آن‌ها در مواقعی که عامل زمان اهمیت بالایی دارد، همچنین به‌دلیل عدم امکان استفاده از تجزیهٔ نحوی کامل در همهٔ زبان‌ها به‌علت نداشتن درخت‌بانک‌های مناسب، تلاش‌هایی به‌منظور بهبود دقت برچسب‌زنی نقش‌های معنایی با استفاده از این نوع از تجزیه‌گرها نیز انجام گرفته است.

در زبان انگلیسی اغلب از درخت‌های نحوی مبتنی بر دستور ساخت‌سازه‌ای به‌منظور برچسب‌زنی نقش معنایی استفاده می‌شود. مهم‌ترین دلیل این امر آن است که اصلی‌ترین منبع برای آموزش تجزیه‌گر خودکار نحوی در زبان انگلیسی، یعنی درخت‌بانک پن[5]، براساس این نوع دستور به وجود آمده است. در ضمن پیکرهٔ معنایی پراپ‌بنک[6] که یکی از منابع اصلی برای برچسب‌زنی نقش‌های معنایی در زبان انگلیسی است با افزودن برچسب‌های سطح معنا به درخت‌بانک پن تهیه شده است. با این حال، با توجه به ویژگی‌های دستور وابستگی، استفاده از تجزیهٔ نحوی کامل مبتنی بر این دستور به‌منظور برچسب‌زنی نقش معنایی اخیراً مورد توجه قرار گرفته است.

۳- دستور و تجزیهٔ وابستگی

نظریهٔ دستور وابستگی یکی از نظریه‌های ساخت‌گرا[7] و صورت‌گراست[8] که در آن به طور اساسی از طریق بررسی روابط وابستگی بین عناصر هسته و وابسته در زبان، به توصیف ساخت‌های نحوی در زبان‌های گوناگون پرداخته می‌شود. نخستین‌بار لوسین تنی‌یر[9] فرانسوی مبانی نظری این دیدگاه را در سال ۱۹۵۳ در کتاب کم‌حجمی با عنوان «گفتارهایی در نحو ساختاری» معرفی کرد. در دستور وابستگی، فرض بر این است که نخست این که هر جمله یک فعل مرکزی دارد و دوم این که بر اساس نوع و تعداد متمم‌های اجباری و اختیاری آن فعل مرکزی می‌توان ساخت بنیادین جمله‌هایی را که آن فعل در آن‌ها به کار رفته است، تعیین کرد. مهم‌ترین مبحث در دستور وابستگی عبارت است از مسئلهٔ ظرفیت نحوی که به بحث درباره وابسته‌های فعل و اسم و صفت اختصاص دارد (طبیب‌زاده، ۱۳۸۵).

تجزیهٔ وابستگی[10] رهیافتی برای تجزیهٔ نحوی زبان طبیعی به‌صورت خودکار است. عوامل مختلفی باعث شده است که در سال‌های اخیر این روش بیش از پیش مورد توجه قرار بگیرد. یک عامل این است که این نوع از دستور زبان (و تجزیهٔ مبتنی بر آن)، در مقایسه با دستور زبان مبتنی بر عبارات، سازگاری بیشتری با طبیعت زبان‌های بی‌ترتیب[11] دارد. اما دلیل مهم‌تر نتایج رضایت‌بخش حاصل از اعمال این روش در برخی از زبان‌ها با استفاده از روش‌های یادگیری خودکار بوده است (رسولی، ۱۳۸۹).

[1] full syntactic parse
[2] shallow syntactic parse
[3] Phrase Structure Grammar
[4] Dependency Grammar

[5] Penn Treebank
[6] PropBank
[7] structural
[8] formal
[9] Lucien Tesnière
[10] dependency parsing
[11] مقصود از زبان‌های بی‌ترتیب زبان‌هایی است که قابلیت جابه‌جایی اجزای جمله در آن‌ها وجود دارد. زبان فارسی از این دسته از زبان‌هاست.

پژوهش‌های اخیر در حوزهٔ برچسب‌زنی نقش‌های معنایی نشان داده است که استفاده از تجزیهٔ نحوی کامل مبتنی بر دستور وابستگی برای برچسب‌زنی معنا نتایج مطلوبی به همراه دارد. یک عامل اصلی این امر آن است که در دستور وابستگی، روابط نحوی میان اجزای جمله (مانند فاعل، مفعول و ...) به‌طور مشهود مشخص می‌شود و این روابط در برچسب‌زنی نقش‌های معنایی دارای اهمیت هستند. پژوهش‌ها همچنین نشان داده است که برچسب‌زن‌های نقش معنایی مبتنی بر دستور وابستگی، اتکای کمتری به ویژگی‌های واژگانی دارند و به همین دلیل با داشتن میزان کمتر دادهٔ آموزش نسبت به تجزیهٔ ساخت‌سازه‌ای، یادگیری بهتری دارد و با تغییر دامنهٔ موضوعی متن، افت دقت پایین‌تری دارد (یوهانسون، ۲۰۰۸a).

تولید داده‌های زبانی مبتنی بر دستور وابستگی برای زبان فارسی در سال‌های اخیر باعث شده است که امکان تجزیهٔ نحوی کامل جملات فارسی با استفاده از این دستور فراهم و در نتیجه استفاده از این نوع تجزیه، به‌منظور برچسب‌زنی نقش معنایی امکان‌پذیر شود.

از منابع مبتنی بر دستور وابستگی که برای زبان فارسی تهیه شده‌اند می‌توان به پیکرهٔ نحوی وابستگی زبان فارسی (رسولی و دیگران، ۲۰۱۳) که شامل حدود ۳۰ هزار جملهٔ برچسب‌خورده با اطلاعات نحوی و ساخت‌واژی است و فرهنگ ظرفیت نحوی افعال فارسی (رسولی و دیگران، ۲۰۱۱) که مجموعه‌ای است حاوی اطلاعات مربوط به ظرفیت نحوی[1] بیش از ۴۵۰۰ فعل در زبان فارسی اشاره کرد. این دو داده در پژوهش حاضر مورد استفاده قرار گرفته‌اند.

۴- مروری بر پژوهش‌های پیشین

نخستین اثر جدی در حوزه برچسب‌زنی خودکار نقش‌های معنایی، کار سال ۲۰۰۲ *گیلدیِ* و همکارانش است که از مجموعه فریم‌نت[2] که در آن زمان شامل حدود پنجاه هزار جمله بود که به‌صورت دستی برچسب معنایی خورده بودند، استفاده کرده است (گیلدِی و دیگران، ۲۰۰۲). از آنجا که جملات مجموعه فریم‌نت از پیکره ملی بریتانیا[3] که تنها دارای برچسب اجزای سخن[4] است انتخاب شده اند و بنابراین فاقد درخت‌های نحوی کامل هستند، در پژوهش بالا پژوهش‌گران به‌منظور استخراج ویژگی‌های نحوی، درخت‌های نحوی جملات را با استفاده از یک تجزیه‌گر که بر مبنای درخت بانک پن آموزش داده شده بود، استخراج می‌کنند.

از زمان تولید مجموعه پراپ‌بنک در سال ۲۰۰۴، بسیاری از سامانه‌های برچسب‌زنی نقش معنایی با بهره‌گیری از این مجموعه تولید شده‌اند. علت توجه پژوهش‌گران به مجموعه پراپ‌بنک برای برچسب‌زنی نقش معنایی این است که در پراپ‌بنک - که از ابتدا با هدف برچسب‌زنی نقش معنایی تهیه شده است - از جملات موجود در درخت بانک پن که دارای درخت‌های کامل نحوی هستند، استفاده شده و برچسب‌های معنایی به آن‌ها اضافه شده است. همین امر موجب شده است که در انتصاب نقش‌های معنایی به گروه‌های نحوی، توجه بیشتری به ساختار نحوی جملات شود و در نتیجه استفاده از ویژگی‌های نحوی به‌منظور آموزش برچسب‌زن معنایی نتایج بهتری به همراه داشته باشد. این در حالی است که در مجموعه فریم نت - که شامل مجموعه‌ای از قاب‌های معنایی است که افعال مختلف را با توجه به وابسته‌های معنایی‌شان در داخل این قاب‌ها طبقه‌بندی می‌کند و جملات را تنها به‌عنوان شاهد برای این قاب‌ها استفاده می‌کند – برچسب‌های معنایی بدون درنظرگرفتن ساخت‌های نحوی به اجزای جمله تعلق می‌گیرند و همین امر موجب می‌شود که انطباق میان برچسب‌های معنایی و ساخت‌های نحوی کاهش یابد و امکان استفاده از ویژگی‌های نحوی به‌منظور آموزش برچسب‌زن محدود شود. از میان برچسب‌زن‌های نقش معنایی تولیدشده بر اساس پراپ‌بنک می‌توان به حدود پنجاه سامانهٔ برچسب‌زنی اشاره کرد که در کنفرانس‌های معتبر یادگیری زبان طبیعی[5] در سال‌های ۲۰۰۴، ۲۰۰۵ و ۲۰۰۸ ارائه شدند.

از نخستین کارهای انجام‌شده در حوزه برچسب‌زنی نقش‌های معنایی با استفاده از دستور وابستگی، پژوهشی است که هاچی‌اوغلو در مقاله سال ۲۰۰۴ خود گزارش کرده است (هاچی‌اوغلو، ۲۰۰۴). پیکرهٔ معنایی مورد استفاده در این اثر پیکرهٔ پراپ‌بنک است و به‌منظور استفاده از بازنمایی نحوی وابستگی، درخت‌های نحوی ساخت‌سازه‌ای پراپ‌بنک به‌صورت خودکار به درخت‌های وابستگی تبدیل و برچسب‌های نحوی با درخت‌های جدید انطباق داده شده‌اند.

[1] syntactic valency
[2] FrameNet
[3] British National Corpus (BNC)
[4] part of speech (POS)
[5] The Conference on Natural Language Learning(CONLL)

علت استفاده از درخت‌های حاصل از تبدیل خودکار این بود که پیکرهٔ درختی استاندارد مبتنی بر دستور وابستگی در زبان انگلیسی وجود نداشت. در این پژوهش آموزش برچسب‌زن نقش‌های معنایی براساس ویژگی‌های قابل تعریف برای روابط وابستگی انجام می‌گیرد. این ویژگی‌ها عبارتند از نوع رابطهٔ وابستگی، نوع ارتباط تعیین‌شده توسط رابطهٔ وابستگی نسبت به فعل مورد نظر (رابطهٔ والد/فرزندی)، جایگاه گرهٔ والد در رابطهٔ نحوی مورد نظر نسبت به محمول (قبل یا بعد)، واژهٔ والد، واژهٔ فرزند، برچسب اجزای سخن واژه والد، برچسب اجزای سخن واژهٔ فرزند، زنجیرهٔ روابط از رابطهٔ مورد نظر به فعل. همچنین تعدادی ویژگی مختص به محمول‌ها تعریف شده است که شامل مواردی چون الگوی برچسب‌های اجزای سخن فرزندان محمول، الگوی روابط وابستگی فرزندان محمول و ... می‌شوند. با توجه به تنوع ویژگی‌های به کار رفته، پژوهش‌گران به‌منظور آموزش برچسب‌زن از روش یادگیری ماشین بردار پشتیبان[1] که می‌تواند به‌خوبی با تعداد زیادی از ویژگی‌ها کار کند، استفاده می‌کنند. آزمایش‌های انجام‌شده بر روی این سامانه نتایج امیدوارکننده‌ای به همراه داشت.

در کنفرانس یادگیری زبان طبیعی سال ۲۰۰۸ تعداد ۱۹ سامانه برچسب‌زنی نقش معنایی مبتنی بر دستور وابستگی ارائه شد. از این میان، سامانهٔ ارائه‌شده در پژوهش یوهانسون و همکارانش بهترین نتایج را به‌دست آورد (یوهانسون و دیگران، ۲۰۰۸b). در این پژوهش اشاره شده است که یکی از دلایل دقت پایین سامانه‌های پیشین برچسب‌زن نقش معنایی مبتنی بر دستور وابستگی در مقایسه با سامانه‌های مبتنی بر دستور ساخت‌سازه‌ای، استفاده از تجزیه‌گرهای وابستگی مبتنی بر قاعده بوده است. این تجزیه‌گرهای نحوی قابلیت رقابت با تجزیه‌گرهای پیشرفتهٔ آماری دستور ساخت‌سازه‌ای را نداشته و درنتیجه دقت برچسب‌زنی معنایی نیز کاهش می‌یافته است. از دلایل موفقیت سامانهٔ مورد بحث نسبت به سایر سامانه‌های ارائه‌شده در کنفرانس، دقت بالای تجزیه‌گر نحوی به‌کاررفته در آن است. در این پژوهش دو منبع پراپ‌بنک و نام‌بنک[2] به کار رفته‌اند و دو دسته ویژگی برای برچسب‌زنی نقش معنایی مورد استفاده قرار گرفته است. دستهٔ نخست شامل ویژگی‌های مورد نظر برای تشخیص محمول‌ها و ابهام‌زدایی از آن‌هاست که عبارتند از صورت و ریشهٔ محمول، صورت و برچسب اجزای سخن والد محمول، ارتباط محمول با والد خود و ... . دستهٔ دوم شامل ویژگی‌هایی می‌شود که برای تشخیص موضوع‌ها و طبقه‌بندی آن‌ها مد نظر است. ویژگی‌هایی چون ریشه و معنای محمول، جهت (معلوم یا مجهول‌بودن) محمول، جایگاه موضوع نسبت به محمول، برچسب اجزای سخن محمول و ... در این دسته قرار می‌گیرند. همچنین سامانه ارائه‌شده در این پژوهش شامل مراحل پس‌پردازش است که بعد از تشخیص محمول‌ها و موضوع‌ها مورد استفاده قرار می‌گیرد و براساس قواعدی که در این مراحل تعریف شده است نقش‌های اختصاص‌داده‌شده مورد ارزیابی مجدد قرار می‌گیرد، قواعدی چون «برچسب‌های موضوع‌های اصلی نباید بیش از یک بار ظاهر شده باشند». تمام مراحل بالا با بهره‌گیری از طبقه‌بندی‌کننده‌های مبتنی بر رگرسیون لجستیک[3] انجام می‌شود.

همچنین در کنفرانس یادگیری زبان طبیعی سال ۲۰۰۹ تعداد ۳۷ سامانه به برچسب‌زنی نقش معنایی بر مبنای دستور وابستگی در هفت زبان (کاتالان، چینی، چکی، انگلیسی، آلمانی، ژاپنی و فرانسوی) پرداختند. سامانهٔ معرفی‌شده در مقالهٔ ژائو و همکارانش بهترین نتیجه را (به‌طور میانگین در تمام زبان‌ها) در این کنفرانس به‌دست آورد (دقت ۸۰/۴۷ درصد). در این سامانه پژوهش‌گران از روش یادگیری آنتروپی بیشینه[4] استفاده می‌کنند و تشخیص کلمات دارای نقش و اختصاص نقش معنایی مناسب به آن‌ها در یک مرحله انجام می‌پذیرد. از جمله ویژگی‌های به‌کاررفته به‌منظور آموزش برچسب‌زن ارائه‌شده در این پژوهش، به صورت، ریشه، برچسب اجزای سخن، رابطه وابستگی، والد، فرزندان و ... می‌توان اشاره کرد (ژائو و دیگران، ۲۰۰۸).

همانطور که پیش‌تر گفته شد، در زبان فارسی ، پژوهش‌های بسیار محدودی در زمینه برچسب‌زنی نقش‌های معنایی انجام گرفته است. همچنین اشاره شد که هیچ یک از این پژوهش‌ها از تجزیه‌گرهای نحوی کامل از جمله تجزیه‌گرهای مبتنی بر دستور وابستگی استفاده نمی‌کنند.

صدر موسوی و *شمس‌فرد* (۱۳۸۶) یکی از نخستین تلاش‌های انجام‌گرفته برای برچسب‌زنی نقش‌های معنایی در زبان فارسی را گزارش می‌کنند. در این پژوهش، هم در مرحله تجزیهٔ نحوی و هم در مرحلهٔ اختصاص نقش‌های معنایی، از روش‌های مبتنی بر قاعده استفاده می‌شود. تجزیهٔ

[1] support vector machine
[2] NomBank
[3] logistic regression
[4] maximum entropy

نحوی به‌کاررفته، تجزیهٔ نحوی سطحی است و برای تشخیص گروه‌های اسمی، قیدی و حرف اضافه‌ای حدود شصت قاعده به‌صورت دستی استخراج شده است. *کامل قالیباف* و همکاران (۱۳۸۸) برچسب‌زنی را برای تعیین نقش‌های معنایی تهیه کرده‌اند که در هر دو مرحلهٔ تحلیل نحوی و معنایی از روش یادگیری مبتنی بر حافظه[1] استفاده می‌کند. این روش براساس ذخیره‌سازی مجموعهٔ داده‌های آموزش در حافظه و محاسبهٔ میزان مشابهت دادهٔ جدید با داده‌های ذخیره‌شده انجام می‌شود. الگوریتم‌های مختلفی برای یادگیری مبتنی بر حافظه وجود دارد که درواقع تمامی این الگوریتم‌ها از الگوریتم K-نزدیک‌ترین همسایه[2] مشتق شده‌اند. جعفری‌نژاد و شمس‌فرد (۲۰۱۲) در پژوهش خود با استفاده از روش مبتنی بر قاعده، سامانه‌ای را برای تشخیص نقش‌های معنایی تعمیم‌یافته[3] تهیه کرده‌اند. این سامانه در مرحلهٔ تجزیهٔ نحوی از تجزیه‌گر نحوی سطحی استفاده کرده و گروه‌های اسمی و هستهٔ آن‌ها را با استفاده از تعدادی قاعده مشخص می‌کند. سعیدی و فیلی (۲۰۱۲) نیز در پژوهش خود از تجزیهٔ نحوی سطحی مبتنی بر قاعده استفاده نموده‌اند. روش مورد استفاده برای برچسب‌زنی نقش‌های معنایی در این پژوهش روش یادگیری مبتنی بر حافظه است. شمس‌فرد و جعفری‌نژاد (۱۳۹۱) در پژوهش خود سه روش را به‌منظور استخراج روابط معنایی ارائه می‌کنند. در روش نخست با تحلیل‌های صرفی[4] روابط بین یک فعل و وابسته‌های آن به دست می‌آیند. برای مثال مشاهده شده که استفاده از الگوهای ساخت صفت مفعولی و اسم مفعول (مصدر+ی، بن ماضی+ه، بن مضارع+اک) برای استخراج روابط معنایی موضوع و پذیرا از دقت نسبی مناسبی برخوردار است. در روش دوم با بهره‌گیری از هستان‌شناسی فارس‌نت[5] به‌عنوان ساختار سلسله‌مراتبی، محدودیت گزینشی هر یک از وابسته‌های معنایی یک مفهوم فعلی به‌صورت رابطهٔ معنایی میان مفهوم فعلی و مفاهیم اسمی بیان می‌شود. به‌عنوان مثال با مشاهدهٔ جملات «سیب را خورد»، «او گلابی می‌خورد»و «شیر خورده شد» و با دانستن این که در هر یک از این جملات به‌ترتیب سیب، گلابی و شیر تحت رابطهٔ معنایی موضوع برای معنای فعلی «خوردن، صرف‌کردن، میل‌کردن» قرار داشته‌اند، با بهره‌گیری از سلسله‌مراتب فارس‌نت و دانستن این که سیب، گلابی و شیر همگی زیرنوعی از (خوراک، خوراکی، خوردنی، ماده مغذی) هستند، این مفهوم (خوراک، خوراکی، خوردنی، ماده مغذی) را به‌عنوان وابستهٔ معنایی «موضوع» برای فعل خوردن به معنای «میل‌کردن، صرف‌کردن» تشخیص می‌دهد. روش سوم نیز که براساس تشخیص نقش‌های معنایی تعمیم‌یافته است تا حد زیادی مشابه همان روش معرفی‌شده در مقالهٔ ۲۰۱۲ جعفری‌نژاد و شمس‌فرد است با این تفاوت که در اینجا نقش تعمیم‌یافته شبه‌عامل[6] به یکی از نقش‌های موضوعی نیرو، ذی‌نفع، عامل، تجربه‌گر، و یا ابزار نگاشته می‌شود و نقش معنایی شبه‌پذیرا[7] نیز به یکی از نقش‌های موضوعی پذیرا و موضوع نگاشته می‌شود.

همان‌طور که ملاحظه می‌شود و در بخش‌های قبلی نیز اشاره شد، تمامی پژوهش‌های انجام‌شده در حوزهٔ برچسب‌زنی نقش‌های معنایی در زبان فارسی در مرحلهٔ تجزیهٔ نحوی از تجزیه‌گر نحوی سطحی بهره می‌گیرند. بسیاری از ویژگی‌هایی که با داشتن درخت‌های وابستگی و اطلاع از روابط حاکم میان اجزای جمله قابل تعریف است، با صرف دانستن مرز گروه‌های نحوی موجود در جمله قابل استفاده نیستند. به همین دلیل پژوهش‌گران برای استخراج ویژگی‌های مشابه به‌ناچار به تقدم و تأخر گروه‌های نحوی بسنده و یا به‌صورت دستی روابط میان برخی از اجزای جمله را مشخص کرده‌اند.

۵- روش پیشنهادی

۵-۱- تهیهٔ داده

نخستین گام برای تولید برچسب‌زن نقش معنایی آماری، تهیهٔ داده‌ای است که بتوان برچسب‌زن را بر مبنای آن آموزش داد. دادهٔ خام مورد استفاده برای این منظور در پژوهش حاضر جملاتی از پیکرهٔ نحوی وابستگی زبان فارسی هستند که توضیحات کامل مربوط به شیوهٔ استخراج آن‌ها در پایان‌نامه نگارندهٔ نخست این مقاله درج شده است (رضائی، ۱۳۹۳). دادهٔ استخراج شده شامل هزار جمله از پیکرهٔ وابستگی است که حاوی پنجاه عدد از پربسامدترین افعال این پیکره هستند. لازم به ذکر است که جملات انتخاب‌شده شامل جملات ساده و مرکب است. اطلاعات مربوط به دادهٔ خام مورد استفاده در جدول زیر قابل ملاحظه است:

[1] memory-based learning
[2] k-nearest neighbors
[3] generalized semantic roles
[4] morphologic
[5] FarsNet
[6] proto-agent
[7] proto-patient

تعداد کل جملات	۱۰۰۰
میانگین طول جملات	۱۱/۶۴۵ واژه
تعداد کل افعال	۲۰۰۶

(جدول- ۱): آمار داده خام

پس از تهیهٔ دادهٔ خام، باید برچسب‌های معنایی به دادهٔ تهیه‌شده، اضافه می‌شدند. با توجه به زمان‌بربودن فرایند برچسب‌زنی دستی پیکره، ابتدا هجده قاعدهٔ ساده به‌منظور برچسب‌زنی اولیهٔ پیکره تهیه شد. قواعدی چون قاعدهٔ زیر:

واژه ای که با رابطه وابستگی «فاعل» به فعل معلوم متصل است برچسب «کنش‌گر» می‌گیرد.

همان‌طور که گفته شد قواعد ارائه‌شده به‌منظور برچسب‌زنی اوّلیه و تسریع فرایند برچسب‌زنی تهیه شده‌اند و تمامی کاربردهای نقش‌های معنایی مورد نظر را پوشش نمی‌دهند. بنابراین به‌منظور تعیین دقیق برچسب‌های معنایی پیکره، تک‌تک جملات توسط پژوهش‌گر بررسی و برچسب‌های آن‌ها در صورت نیاز اصلاح شدند.

نقش‌های معنایی درنظرگرفته‌شده و تعریف مورد نظر برای آن‌ها در جدول ۲ قابل ملاحظه است[1].

۵-۲- تجزیهٔ نحوی

همان‌طور که در فصل‌های پیشین توضیح داده شد، پیش از برچسب‌زنی نقش‌های معنایی، لازم است به‌منظور استخراج ویژگی‌ها، تجزیهٔ نحوی انجام پذیرد. به‌منظور انجام تجزیهٔ وابستگی بر روی جملات از ابزار «هضم»[2] کمک گرفتیم. این ابزار متن‌باز که برای پردازش زبان فارسی با زبان برنامه‌نویسی پایتون[3] تهیه شده است، امکاناتی از قبیل مرتب کردن متن، تقطیع جمله‌ها و واژه‌ها، ریشه‌یابی واژه‌ها، تعیین برچسب اجزای سخن و تجزیهٔ نحوی جملات را فراهم می‌آورد. این ابزار در بخش تجزیهٔ نحوی از الگوریتم مالت‌پارسر[4] که از طریق کتابخانهٔ ان.ال.تی.کی.[5] قابل دسترسی است استفاده می‌کند. ارزیابی تجزیه‌گر نحوی ارائه‌شده در این ابزار بر روی داده‌های آموزش و آزمایش پیکرهٔ وابستگی نشان‌دهنده دقت ۸۳ درصدی آن است.

نقش	تعریف
کنش‌گر	انجام‌دهندهٔ کار، سبب انجام کار، دارندهٔ چیزی
کنش‌پذیر	تحت تأثیر عمل (همراه با تغییر ماهیت یا وضعیت)
کنش‌بَر	تحت تأثیر عمل (بدون تغییر ماهیت یا وضعیت)، دارای وضعیت مشخص (فاعل فعل‌های اسنادی)، در حال حرکت و یا دارای مکان مشخص، دارایی
تجربه‌گر	تجربه‌کنندهٔ عمل (در فعل‌های ادراکی و احساسی)
بهره‌ور	عمل فعل به نفع (یا ضرر) او انجام می‌شود
منشأ	نقطهٔ آغاز، مبدأ، منشأ
مقصد	نقطهٔ پایان، مقصد
نسبت	ویژگی، مشخصه، نسبت
مکان	مکان فیزیکی و یا انتزاعی
روش	روش، ابزار، چگونگی انجام کار
قید	قید، شامل توضیح اضافه در خصوص فعل (شامل قیدهایی که برچسب های دیگر چون مکان، زمان، مقدار و ... به آن‌ها اختصاص داده نشده است.)
وجه‌نما	بیان‌کننده وجهیّت فعل (باید، شاید، حتماً ...)
شرط	شرط (اگر، در صورتی که، ...)
زمان	زمان (فردا، پارسال، ...)
مقدار	مقدار (زیاد، به اندازه کافی، ...)
تکرار	تکرار (دوباره، هرگز، ...)
هدف	هدف، منظور، مقصود از انجام کار
علت	علت (زیرا، چرا، ...)

(جدول- ۲): نقش‌ها معنایی مورد نظر

۵-۳- آموزش برچسب‌زن

ویژگی‌های مورد استفاده در برچسب‌زن نقش معنایی تهیه‌شده، بیش‌تر با بررسی پژوهش‌های پیشین و به‌خصوص ویژگی‌های معرفی‌شده در پژوهش‌های *گیلدیِ* (۲۰۰۲) و *یوهانسون* (۲۰۰۸) به دست آمده‌اند. روش عملکرد سامانه به این صورت است که ابتدا محمول فعلی جمله را پیدا می‌کند؛ سپس برای تک‌تک واژه‌های جمله یک بردار ویژگی تعریف می‌کند که شامل ویژگی‌هایی است که در ادامه به آن‌ها خواهیم پرداخت. این بردار ویژگی در کنار برچسب معنایی که از قبل به‌صورت دستی برای واژهٔ مورد بررسی نسبت به

[1] نگارندهٔ نخست مقاله در حین انجام این پژوهش مشغول به فعالیت در پروژهٔ «پیکرهٔ معنایی زبان فارسی» در مرکز تحقیقات کامپیوتری علوم اسلامی (نور) بوده است. انتخاب نقش‌ها و معادل برای آن‌ها متأثر از مجموعه نقش‌های معنایی به کار رفته در پیکرهٔ نامبرده است.

[2] در این پژوهش از نسخه ۰.۱ این ابزار متن‌باز که از آدرس اینترنتی www.sobhe.ir/hazm قابل دریافت است استفاده شده است.

[3] Python

[4] MaltParser

[5] NLTK (Natural Language Toolkit)

محمول مورد نظر تعیین شده است، قرار گرفته و ذخیره می‌شود. ویژگی‌های مورد بررسی عبارتند از:

ریشهٔ فعل: ریشهٔ فعل مورد نظر در انتخاب برچسب‌های معنایی تأثیر دارد. برای مثال فاعل فعل از مصدر «کتک‌خوردن» در حالت معلوم کنش‌پذیر است و نه کنش‌گر. بنابر این ریشه این فعل (کتک خورد#خور) می‌تواند در تشخیص صحیح نقش‌های معنایی مرتبط با آن کمک کند.

جهت فعل: معلوم یا مجهول‌بودن فعل اهمیت زیادی در تشخیص نقش‌های معنایی – به‌خصوص کنش‌گر و کنش‌پذیر – دارد. نقش معنایی مفعول فعل معلوم، به‌طور معمول با نقش معنایی فاعل همان فعل در حالت مجهول یکسان است. برای مثال در هر دو جملهٔ «علی نامه را نوشت.» و «نامه نوشته شد.» واژهٔ «نامه» دارای نقش معنایی «کنش‌پذیر» است.

برچسب اجزای سخن واژه: احتمال آنکه واژه‌ای با برچسب اجزای سخن «حرف اضافه» منشأ یا مقصد باشد، بسیار بالاتر از آن است که کنش‌گر باشد؛ حال آنکه درخصوص واژه‌ای با برچسب اجزای سخن «اسم» عکس این موضوع صادق است. درصورت حجیم‌بودن داده می‌توان خود واژه‌ها را نیز به عنوان ویژگی در نظر گرفت.

مسیر درخت نحوی: این ویژگی مسیر موجود میان فعل مورد بررسی و واژهٔ مورد نظر را نشان می‌دهد. در مواردی که واژهٔ مدّ نظر فرزند مستقیم فعل باشد، مقدار این ویژگی همان رابطهٔ نحوی واژه با فعل است. اهمیت این ویژگی در این است که برای مثال واژه‌ای که با رابطهٔ نحوی «فاعل» به فعل معلوم متصل است، دارای نقش معنایی «کنش‌گر» است.

حضور فاعل: در مواقعی که دو جمله، همپایه شده باشند و فاعل در جمله دوم قید نشده باشد، ممکن است فاعل جملهٔ نخست برای افعال هر دو جمله، برچسب معنایی بخورد. برای تعریف این ویژگی، درصورتی‌که هر دو جمله دارای فاعل باشند، مقدار صفر و درصورتی‌که تنها یک جمله دارای فاعل باشد، مقدار یک را نسبت می‌دهیم.

ساخت ظرفیتی فعل: ساخت ظرفیتی[1] فعل در انتخاب نقش های معنایی مؤثر است. برای مثال فعل «شکستن» دو ساخت ظرفیتی دارد که عبارتند از ساخت گذرا «فاعل، مفعول» (علی شیشه را شکست) و ساخت ناگذر «فاعل» (شیشه شکست). اطلاع از این که فعل «شکستن» در جمله مد نظر دارای کدام یک از این ساخت‌های ظرفیتی است، سامانه برچسب‌زنی نقش معنایی را در تشخیص نقش مناسب برای فاعل نحوی جمله (شیشه) کمک می‌کند. در برچسب‌زن ارائه‌شده این ویژگی با بررسی خودکار ساخت ظرفیتی فعل مورد نظر (تعداد وابسته‌ها نوع وابسته‌ها و ..) و اختصاص مقدار ۱ یا ۲ (ساخت ظرفیتی نخست یا دوم) استخراج و به بردار ویژگی واژه‌ها افزوده می‌شود.

نوع موجودیت اسمی: این ویژگی نیز در تشخیص صحیح نقش‌های معنایی می‌تواند کمک کند. برای مثال اگر اسم پس از حرف اضافه «به» از نوع «مکان» باشد، برچسب معنایی صحیح برای آن برچسب «مقصد» است. برای بررسی این ویژگی از اطلاعات موجود در واژگان زایای زبان فارسی کمک گرفتیم (اسلامی و دیگران، ۱۳۸۳) و اسامی خاص مکان و اسامی خاص اشخاص را مشخص کردیم.

صورت واژهٔ حرف اضافه: از آنجا که انواع حرف اضافه پیشین (از، به، در، تا، روی، ...) مجموعه‌ای محدود را تشکیل می‌دهند، می‌توان از صورت واژهٔ آن‌ها به‌عنوان ویژگی استفاده کرد. بنابراین درصورتی‌که برچسب اجزای سخن واژه مورد بررسی «حرف اضافهٔ پیشین» باشد، صورت آن را به‌عنوان مقدار این ویژگی در نظر می‌گیریم و در غیر این صورت مقدار آن تهی خواهد بود.

الگوریتم‌های انتخاب‌شده به‌منظور آموزش برچسب‌زن الگوریتم‌های بیز ساده[2] و انتروپی بیشینه هستند. همان‌طورکه در پیشینهٔ پژوهش توضیح داده شد، الگوریتم انتروپی بیشینه از پرکاربردترین الگوریتم‌ها در برچسب‌زنی نقش معنایی است و الگوریتم بیز ساده را نیز با توجه به ویژگی هایی چون سرعت بالای یادگیری و عملکرد مناسب هنگام داشتن دادهٔ کم مورد استفاده قرار داده‌ایم. دو الگوریتم بالا، که در این پژوهش از کتابخانهٔ ان.ال.تی.کی. به‌منظور پیاده‌سازی آن‌ها استفاده شده است، در ادامه تشریح می‌شوند.

الگوریتم بیز ساده، یکی از رایج‌ترین الگوریتم‌هایی است که در انواع طبقه‌بندی مورد استفاده قرار می‌گیرد. این الگوریتم مبتنی بر قانون بیز است و فرضیاتی نیز به‌منظور ساده‌سازی آن لحاظ می‌شود. در ادامه به معرفی قانون بیز و فرضیات ساده‌سازی آن می‌پردازیم.

[1] valency structure

[2] naive bayes

براساس قانون بیز، اگر واژه w و برچسب L را داشته باشیم، احتمال وقوع برچسب L به شرط دیدن واژهٔ w را می‌توان به شکل زیر محاسبه نمود:

$$P(L|w) = \frac{P(L) \times P(w|L)}{P(w)} \quad (۱)$$

از احتمال بالا به‌منظور انتخاب بهترین برچسب برای هر واژه استفاده می‌شود. بدین معنا که بهترین برچسب (L_{MAP}) برای واژهٔ مورد نظر عبارت است از برچسبی که احتمال $P(L|w)$ را بیشینه کند یا به عبارت دیگر برچسبی که احتمال وقوع آن به شرط دیدن واژهٔ w بیشینه است.

$$L_{MAP} = \underset{L \in \mathbf{L}}{\operatorname{argmax}}\ P(L|w) \quad (۲)$$

بر اساس قانون بیز، رابطهٔ بالا را می‌توان به‌صورت زیر بازنویسی کرد:

$$L_{MAP} = \underset{L \in \mathbf{L}}{\operatorname{argmax}}\ \frac{P(L) \times P(w|L)}{P(w)} \quad (۳)$$

هنگام بررسی احتمالات مربوط به برچسب‌های مختلف نسبت به واژه‌ای مشخص(w)، مخرج کسر در رابطهٔ بالا در تمام موارد عددی ثابت و مشخص خواهد بود و تأثیری در نتیجه مقایسه احتمالات نخواهد داشت؛ بنابراین می‌توان آن را حذف کرد.

$$L_{MAP} = \underset{L \in \mathbf{L}}{\operatorname{argmax}}\ P(L) \times P(w|L) \quad (۴)$$

محاسبهٔ $P(L)$ درواقع به معنای محاسبهٔ بسامد نسبی وقوع برچسب L یا نسبت تعداد برچسب مورد نظر در دادهٔ آموزش به تعداد کل برچسب‌ها است:

$$P(L) = \frac{N(L)}{N} \quad (۵)$$

به‌منظور محاسبهٔ $P(w|L)$ نیز w در عمل به‌صورت مجموعه‌ای از ویژگی‌ها در نظر گرفته می‌شود:

$$P(f_1 L_1, f_2 L_2, \ldots, f_n L_n \mid L) \quad (۶)$$

در اینجا دو فرض به‌منظور ساده‌سازی الگوریتم در نظر گرفته می‌شود. نخست اینکه ترتیب ویژگی‌ها اهمیت ندارد[1] و بعد اینکه ویژگی‌ها هیچ‌گونه وابستگی به یکدیگر ندارند. بدیهی است که هر دوی این فرض‌ها با واقعیت موجود فاصله دارند (درواقع هم ترتیب و هم وابستگی‌ها اهمیت دارد)؛ اما استفاده از آن‌ها به‌عنوان فرض‌های ساده‌سازی، محاسبات را ساده و دست‌یابی به نتایج مد نظر را امکان‌پذیر می‌سازد. بر این اساس می‌توان رابطهٔ بالا را به‌صورت حاصل‌ضرب احتمالات وقوع هر ویژگی به شرط دیدن برچسب مورد نظر بازنویسی کرد:

$$L_{MAP} = \underset{L \in \mathbf{L}}{\operatorname{argmax}}\ P(L) \times P(f_1|L) \times P(f_2|L) \times \ldots \times P(f_n|L)$$

$$L_{MAP} = \underset{L \in \mathbf{L}}{\operatorname{argmax}}\ P(L) \times \prod_{f \in \mathbf{F}} P(f|L) \quad (۷)$$

به‌منظور محاسبهٔ هر یک از احتمالات $P(f|L)$ نسبت تعداد وقوع ویژگی f در واژه‌های دارای برچسب L را به تعداد تمام واژه‌های دارای برچسب L به دست می‌آوریم:

$$P(f|L) = \frac{N(F=f, L=L)}{N(L=L)} \quad (۸)$$

از سوی دیگر در الگوریتم انتروپی بیشینه، وزن مشخصی به هر یک از ویژگی‌های مربوط به داده ورودی اختصاص داده و با داشتن ویژگی‌ها و وزن مربوط به آن‌ها محتمل‌ترین برچسب برای دادهٔ ورودی انتخاب می‌شود. براساس این الگوریتم احتمال برچسب L برای دادهٔ ورودی x برابر است با:

$$P(L|x) = \frac{1}{Z} \exp\left(\sum_i w_i f_i\right) \quad (۹)$$

در این رابطه Z عدد ثابتی است که استفاده می‌شود تا جمع احتمالات برابر یک باشد و exp هم نشانه تابع نمایی است ($e \simeq ۲.۷۱$):

$$\exp(x) = e^{x} \quad (۱۰)$$

۶- آزمایش‌ها و نتایج

۶-۱- روش ارزیابی

جهت انجام ارزیابی‌ها، لازم بود که دادهٔ برچسب‌خورده به دو مجموعهٔ آموزش و آزمایش تقسیم شود. برای این منظور از جملات مربوط به هر فعل، هجده جمله برای مجموعهٔ آموزش و دو جمله (غیر تکراری و با طول متوسط) برای مجموعهٔ آزمایش انتخاب شد. بدین ترتیب نهصد جمله به آموزش و یکصد جمله به آزمایش اختصاص داده شد.

در این پژوهش از سه معیار ارزیابی متداول در پردازش زبان طبیعی که عبارتند از دقت[2]، فراخوانی[3] و معیار F1 که میانگین توافقی[4] دو معیار دیگر است استفاده می‌شود. در اکثریت قریب به اتفاق پژوهش‌های مرتبط با برچسب‌زنی نقش‌های معنایی این سه معیار به‌منظور ارزیابی برچسب‌زن به کار گرفته می‌شوند. روش محاسبهٔ این معیارها به‌صورت زیر است:

دقت:

$$P = \frac{\text{تعداد برچسب‌های صحیح اختصاص داده شده توسط برچسب‌زن}}{\text{تعداد کل برچسب‌های اختصاص داده شده توسط برچسب‌زن}}$$

فراخوانی:

[1] Bag of words representation

[2] Precision
[3] Recall
[4] Harmonic mean

$$R=\frac{\text{تعداد برچسب‌های صحیح اختصاص داده شده توسط برچسب‌زن}}{\text{تعداد کل برچسب‌های موجود در داده استاندارد}}$$

F1:

$$F1=\frac{2\times P\times R}{P+R}$$

به‌منظور ارزیابی دقیق سامانهٔ برچسب‌زنی نقش معنایی و بررسی تأثیر بخش‌های مختلف سامانه بر عملکرد کلی آن، سه سطح از ارزیابی را می‌توان تعریف کرد. نخست، ارزیابی تجزیهٔ نحوی که به‌طورمعمول در تجزیهٔ وابستگی با معیار امتیاز وابستگی با برچسب[1] انجام می‌پذیرد. این معیار عبارت است از درصد واژه‌هایی که تجزیه‌گر نحوی، والد و رابطهٔ وابستگی صحیح را برای آن واژه‌ها تشخیص داده است. علاوه‌بر این می‌توان امتیاز وابستگی بدون برچسب[2] را نیز مورد بررسی قرار داد. این معیار درصد واژه‌هایی را نشان می‌دهد که تجزیه‌گر نحوی، والد آن واژه‌ها را صحیح تشخیص داده است (بدون در نظر گرفتن رابطهٔ وابستگی). ارزیابی دیگر ارزیابی برچسب‌زنی نقش معنایی است که با در نظر گرفتن درخت‌های نحوی موجود در دادهٔ استاندارد، معیارهای دقت، فراخوان و F1 مطابق آنچه در بخش ۴-۲-۳ توضیح داده شد محاسبه می‌شود. درنهایت عملکرد کلی سامانهٔ برچسب‌زن نقش معنایی با ارزیابی برچسب‌زنی نقش معنایی با داشتن درخت‌های نحوی حاصل از تجزیهٔ خودکار محاسبه می‌شود. بررسی نوع آخر درواقع میزان دقت سامانه را در مواجهه با جملات جدید نشان می‌دهد.

۶-۲- نتایج ارزیابی

همان‌طورکه در جدول زیر پیداست، روش یادگیری انتروپی بیشینه (ME) نتایج بهتری را هنگام استفاده از درخت‌های نحوی استاندارد به همراه دارد. البته لازم به ذکر است که این روش نسبت به روش بیز ساده (NB) به زمان بیشتری برای آموزش نیاز دارد.

روش یادگیری ماشینی	دقت	فراخوان	F1
ME	۰/۸۴	۰/۸۱	۰/۸۲
NB	۰/۷۶	۰/۸۰	۰/۷۸

(جدول- ۳): ارزیابی برچسب‌زنی نقش معنایی با درخت‌های نحوی استاندارد

در جدول زیر ملاحظه می‌شود که درصد دقت برچسب‌زنی نحوی بر اساس معیار LAS (با در نظر گرفتن برچسب وابستگی) ۹۱/۳ درصد و بر اساس معیار UAS (بدون در نظر گرفتن برچسب وابستگی) ۹۳/۱ درصد است. علاوه‌بر این دو مورد، دقت‌های مربوط به تشخیص ریشهٔ واژه، برچسب اجزای سخن ریز، و برچسب اجزای سخن درشت در جدول زیر ارائه شده است.

	دقت
LAS	۹۱/۳
UAS	۹۳/۱
ریشه	۷۷/۰
POS درشت	۸۳/۹
POS ریز	۲۲/۹

(جدول- ۴): دقت تجزیهٔ نحوی خودکار

همان‌طورکه در جدول (۵) پیداست، دقت مربوط به عملکرد کلی سامانه در بهترین حالت نزدیک به ۷۰ درصد است:

روش یادگیری ماشینی	دقت	فراخوان	F1
ME	۰/۶۹	۰/۷۰	۰/۶۹
NB	۰/۶۳	۰/۶۸	۰/۶۶

(جدول- ۵): ارزیابی برچسب‌زنی نقش معنایی با تجزیه نحوی خودکار

برای بررسی تأثیر حجم داده آموزش بر عملکرد کلی سامانه‌ای، آزمایش‌هایی با در نظر گرفتن حجم‌های مختلف برای داده آموزش انجام گرفته که نتایج آن در جدول زیر قابل رؤیت است:

حجم داده آموزش	F1
۳۰۰ جمله	۰/۶۲
۶۰۰ جمله	۰/۶۷
۹۰۰ جمله	۰/۶۹

(جدول- ۶): ارزیابی تاثیر حجم داده آموزش بر عملکرد برچسب‌زن

همان‌طورکه در جدول (۶) ملاحظه می‌شود حجم داده آموزش و معیار F1 سامانهٔ برچسب‌زنی نقش معنایی رابطهٔ مستقیمی با یکدیگر دارند. این به آن معناست که با داشتن حجم بیشتری از داده آموزش برچسب‌زن بهتری تولید خواهد شد.

به‌دلیل عدم استفاده از داده‌های مشترک، به کارگیری مجموعه برچسب‌های متفاوت و ارزیابی با روش‌های مختلف، امکان مقایسه مستقیم برچسب زن نقش معنایی تولیدشده در این پژوهش و برچسب‌زن‌های فارسی تولیدشده در پژوهش‌های پیشین وجود ندارد.

یکی از مهم‌ترین برجستگی‌های برچسب‌زن نقش معنایی ارائه‌شده در این پژوهش در مقایسه با برچسب‌زن‌های پیشین زبان فارسی، امکان برچسب‌زنی نقش

[1] Labeled Attachement Score (LAS)
[2] Unlabeled Attachement Score (UAS)

معنایی نسبت به تمام افعال جمله است. این بدان معناست که به ازای هر یک از افعال موجود در جمله، نقش معنایی احتمالی هر واژه از جمله نسبت به همان فعل بررسی می‌شود. یکی از نتایج این امر آن است که ممکن است، واژه‌ای نسبت به یکی از افعال جمله دارای نقش معنایی خاصی و نسبت به فعل دیگر جمله نقش معنایی دیگری داشته باشد. برای مثال در جمله «او ترسید و فرار کرد.» که جمله‌ای مرکب است، «او» نسبت به فعل «ترسید» تجربه‌گر و نسبت به فعل «فرار کرد» کنش‌گر است. به‌طوراساسی به نظر می‌رسد که در پژوهش‌های پیشین تنها جملات دارای یک فعل مورد بررسی قرار گرفته‌اند، هر چند این موضوع تصریح نشده است.

جا دارد که سامانه‌های مختلف برچسب‌زنی ارائه‌شده برای زبان فارسی با استفاده از مجموعه دادهٔ مشترک، مجموعهٔ برچسب‌های مشترک و روش‌های ارزیابی مشترک مورد مقایسه قرار بگیرند تا نقاط ضعف و قوت هر یک مشخص شود. این همان کاری است که در کنفرانس‌هایی چون کنفرانس یادگیری زبان طبیعی انجام می‌گیرد.

۷- نتیجه‌گیری

در این پژوهش سامانه‌ای برای برچسب‌زنی خودکار نقش‌های معنایی در زبان فارسی ارائه شد. این سامانه، طبق اطلاع نگارندگان، نخستین برچسب‌زن نقش معنایی در زبان فارسی است که از نحو مبتنی بر دستور وابستگی در مرحلهٔ تجزیهٔ نحوی استفاده می‌کند. پیش از این تمام برچسب‌زن‌های نقش معنایی فارسی از تجزیهٔ نحوی سطحی استفاده می‌کرده‌اند. به کارگیری این دستور امکان به کارگیری ویژگی‌هایی را فراهم کرده است که تا پیش از این در دسترس نبوده‌اند. از سوی دیگر در برچسب‌زن ارائه‌شده محدودیتی از نظر طول جملات و تعداد افعال و یا وجود بندها اعمال نشده است و این برچسب‌زن امکان کار با انواع مختلف جملات با ساخت‌های متنوع را دارد.

مهم‌ترین موانع پیش رو در تهیهٔ برچسب‌زن‌های نقش معنایی با کاربرد عمومی عبارتند از عدم وجود پیکره استاندارد دارای نقش‌های معنایی و عدم وجود رده‌بندی معنایی افعال زبان فارسی. خوشبختانه یکی از مراکز پژوهشی[1] اقدام به افزودن برچسب های سطح معنا به پیکرهٔ نحوی وابستگی زبان فارسی کرده است. این اتفاق می‌تواند نقطهٔ عطفی در برچسب‌زنی خودکار نقش‌های معنایی در زبان فارسی و به‌طورکلی پردازش زبان فارسی باشد.

با در اختیار داشتن داده های معنایی استاندارد برای زبان فارسی، می‌توان سامانهٔ فعلی را از جنبه‌های مختلف توسعه و دامنهٔ کاربرد آن را گسترش داد. برای مثال می توان افعال با تعداد بالای ساخت ظرفیتی را پوشش داد. همچنین با افزایش حجم داده، می توان تعداد ویژگی های بیشتری را به‌منظور آموزش برچسب‌زن مورد استفاده قرار داد.

از جمله کارهایی که می توان در آینده انجام داد، افزودن مرحلهٔ رتبه‌بندی برچسب‌ها پس از مشخص‌شدن تمام برچسب‌های نقش معنا در یک جمله است. پژوهش‌ها نشان داده است که اضافه کردن این مرحله که در آن با توجه به احتمال وقوع برچسب‌ها در کنار هم برچسب‌های معنایی رتبه‌بندی می‌شوند، موجب بهبود کارایی سامانه می‌شود. همچنین امروزه در بسیاری از سامانه‌ها کار برچسب‌زنی نقش معنایی به دو مرحلهٔ تشخیص گروه‌های دارای نقش معنایی و اختصاص نقش های معنایی تقسیم می‌شود. می‌توان با اعمال این تقسیم‌بندی تأثیر آن را در دقت برچسب‌زنی بررسی نمود.

برچسب‌زنی نقش‌های معنایی برای موضوع‌های محمول های غیر فعلی همچون اسم و صفت از دیگر مواردی است که می‌توان در ادامهٔ این پژوهش انجام داد. توضیح آنکه در سامانه فعلی تنها برچسب معنایی موضوع های مرتبط با محمول های فعلی مد نظر هستند حال آن‌که در عبارتی مانند «برداشت پول از حساب» واژه های «پول» و «حساب» را می توان به‌ترتیب کنش‌بر و منشأ برای اسم «برداشت» در نظر گرفت.

همچنین جا دارد تأثیر به‌کارگیری اطلاعات حاصل از برچسب‌زنی نقش‌های معنایی در کاربردهای مختلف پردازش زبان طبیعی مورد بررسی قرار گیرد.

۸- مراجع

اسلامی، م. شریفی آتشگاه، م. علیزاده لمجیری، ص. و زندی، ط. واژگان زایای زبان فارسی. مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه، ۱۳۸۳.

رسولی، م. تجزیه نحوی با استفاده از دستور وابستگی. گزارش پژوهشی. تهران، مرکز تحقیقات کامپیوتری علوم اسلامی. ۱۳۸۹.

رضائی، م. برچسب‌زنی نقش‌های معنایی با استفاده از درخت‌های وابستگی جملات فارسی. پایان‌نامه کارشناسی ارشد، تهران، دانشگاه صنعتی شریف. اردیبهشت ۱۳۹۳.

[1] مرکز تحقیقات کامپیوتری علوم اسلامی (نور)

شمس‌فرد، م. جعفری‌نژاد، ف. استخراج روابط معنایی میان فعل و وابسته‌های آن از متون زبان فارسی. فصلنامه پازند، ۱۳۹۱، شماره ۸ (۳۰).

صدر موسوی، م. شمس‌فرد، م. برچسب‌زنی نقش‌های معنایی با استفاده از تجزیه سطحی جملات فارسی. سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، ۱۳۸۶، جزیره کیش.

طبیب زاده، م. ظرفیت فعل و ساخت های بنیادین جمله در فارسی امروز، پژوهشی بر اساس نظریه دستور وابستگی. تهران، نشر مرکز. ۱۳۸۵.

کامل قالیباف، آ. راحتی قوچانی، س. استاجی، ا. برچسب زنی نقش معنایی جملات فارسی با رویکرد یادگیری مبتنی بر حافظه. دوفصل‌نامه پردازش هوشمند علائم و داده‌ها، ۱۳۸۸، شماره ۱ (۱۱)، ۱۳-۲۲.

Gildea, D., & Jurafsky, D. Automatic Labeling of Semantic Roles. Computational linguistics, 2002, 28(3), 245-288.

Hacioglu, K. Semantic Role Labeling Using Dependency Trees. In Proceedings of the 20th International Conference on Computational Linguistics, 2004, No.1273.

Jafarinejad, F., & Shamsfard, M. Extracting Generalized Semantic Roles from Corpus. International Journal of Computer Science Issues (IJCSI), 2002, 9(2), 200-206.

Johansson, R., & Nugues, P. The Effect of Syntactic Representation on Semantic Role Labeling. In Proceedings of the 22nd International Conference on Computational Linguistics, 2008a, Vol. 1, 393-400.

Johansson, R., & Nugues, P. Dependency-Based Syntactic-Semantic Analysis with PropBank and NomBank. In Proceedings of the Twelfth Conference on Computational Natural Language Learning, 2008b, 183-187.

Jurafsky, D., & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.). Upper Saddle River, NJ: Prentice Hall, 2009.

Palmer, M., Gildea, D., & Xue, N. Semantic role labeling. Synthesis Lectures on Human Language Technologies, 2010, 3(1), 1-103.

Punyakanok, V., Roth, D., & Yih, W. T. The Necessity of Syntactic Parsing for Semantic Role Labeling. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2005, Vol. 5, 1117-1123.

Rasooli, M. S., Kouhestani, M., & Moloodi, A. Development of a Persian Syntactic Dependency Treebank. In Proceedings of NAACL-HLT, 2013, 306-314.

Rasooli, M. S., Moloodi, A., Kouhestani, M., & Minaei-Bidgoli, B. A Syntactic Valency Lexicon for Persian Verbs: The First Steps Towards Persian Dependency Treebank. In the 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics, 2011, 227-231.

Riemer, Nick. Introducing Semantics. Cambridge University Press, 2010.

Saeed, John I. Semantics (2nd ed.). Blackwell Publishing, 2003.

Saeedi, P., & Faili, H. Feature Engineering Using Shallow Parsing in Argument Classification of Persian Verbs. In Proceedings of the 16th Artificial Intelligence and Signal Processing (AISP), 2012, 333-338. Shiraz, Iran.

Zhao, H., Chen, W., Kit, C., & Zhou, G. (2009). Multilingual dependency learning: a huge feature engineering method to semantic dependency parsing. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task (pp. 55-60). Association for Computational Linguistics.

مرتضی رضائی شریف‌آبادی

دانش‌آموختهٔ مقطع کارشناسی ارشد در رشتهٔ زبان‌شناسی رایانشی از دانشگاه صنعتی شریف است. وی همچنین دورهٔ کارشناسی خود را در دانشگاه علامه طباطبایی (ره) گذرانده است. ایشان از سال ۱۳۹۰ تا کنون با مرکز تحقیقات کامپیوتری علوم اسلامی (نور) در پروژه‌هایی چون دادگان نحوی زبان فارسی بر اساس دستور وابستگی، دادگان معنایی زبان فارسی، مرجع دادگان زبان فارسی و ... همکاری داشته است.

نشانی رایانامهٔ ایشان عبارت است از:

mrezaeis@gmail.com

پروانه خسروی‌زاده فروشانی

عضو هیأت علمی دانشگاه صنعتی شریف است. ایشان دکترای خود را در رشتهٔ زبان‌شناسی همگانی از دانشگاه تهران دریافت کرده است. وی همچنین عضو هیئت مؤسس انجمن زبان‌شناسی ایران و سه دوره عضو هیئت مدیرهٔ این انجمن بوده است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از معنی‌شناسی، کاربردشناسی و ترجمهٔ ماشینی.

نشانی رایانامه ایشان عبارت است از:

khosravizadeh@sharif.edu