علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۱۱، شماره ۳ (الف)، ۱۳۹۲
صفحات ۲۲-۳۰
مقاله عادی

شناسایی افعال در جملات زبان فارسی

آرش چاقری مهرنوش شمس‌فرد

دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران

چکیده

در این مقاله دو روش، یکی مبتنی بر رویکرد بدون نظارت و دیگری مبتنی بر رویکرد بانظارت برای شناسایی افعال ساده و مرکب در متون فارسی ارائه می‌گردد. در هر دو روش ابتدا افعال ساده و افعال سبک جمله با استفاده از برچسب نحوی کلمات و ریشه‌یابی مشخص می‌شوند. سپس برای پیدا کردن فعل جمله که می‌تواند ساده یا مرکب باشد، به دنبال کلمات کاندید می‌گردیم. کلمات کاندید کلماتی هستند که می‌توانند به عنوان جزء غیرفعلی فعل اصلی درنظر گرفته شوند. برای انتخاب کلمات کاندید در روش بدون نظارت ابتدا براساس متون پیکره بیجن‌خان ویژگی‌ها و قواعدی استخراج می‌شوند که با استفاده از آنها یک معادله‌ی امتیازدهی تعریف می‌شود. در روش با نظارت با استفاده از بخشی از پیکره‌ی دارای برچسب فعل مرکب، ویژگی‌های خاص چنین پیکره‌ای مقداردهی می‌شود. در ادامه برای هر دو روش، امتیاز هر یک از کلمات کاندید براساس این معادله به‌دست می‌آید و در نهایت اگر بیش‌ترین امتیاز به‌دست آمده بیشتر از حد آستانه از پیش تعریف شده باشد، پرامتیازترین کلمه‌ی کاندید، به عنوان جزء غیرفعلی منظور می‌شود و در غیر این صورت فعل اصلی به عنوان فعل ساده در نظر گرفته می‌شود. از محاسن روش پیشنهادی می‌توان به دقت بالای آن در مقایسه با سایر کارهای مشابه و تشخیص افعال کمکی وابسته به افعال اصلی اشاره کرد.

کلمات کلیدی: تشخیص فعل، فعل مرکب، رویکرد بدون نظارت، زبان فارسی، عبارت چند کلمه‌ای، جزء فعلی و غیرفعلی، پیکره‌ی دارای برچسب فعل مرکب، معادله‌ی امتیازدهی، ویژگی‌ها، برچسب نحوی.

۱- مقدمه

در بیشتر زبان‌ها از جمله زبان فارسی، افعال را می‌توان به دو گروه ساده و مرکب دسته‌بندی کرد. در زبان فارسی افعال ساده هم می‌توانند یک‌جزئی باشند و هم می‌توانند چندجزئی باشند. فعل "خرید" یک نمونه از افعال ساده یک‌جزئی و "خریده بود" یک نمونه از افعال ساده چندجزئی است. در زبان فارسی تعداد رخداد افعال ساده چندجزئی به مراتب بیشتر از افعال ساده یک‌جزئی است. در این مقاله به افعال ساده صرف‌نظر از یک‌جزئی یا چندجزئی بودن، افعال تک واحدی اطلاق می‌گردد زیرا با اینکه از چند جزء تشکیل شده است ولی دارای یک معنی واحد است مانند "گفته شد". همچنین فعل "بزند" در جمله‌ی "شاید روزی او را به خاطر این تهمت بزند" فعل ساده است.

افعال مرکب در زبان فارسی از یک جزء غیرفعلی و جزء فعلی تشکیل می‌شود [۱]. جزء غیرفعلی می‌تواند اسم (مانند "کمک کردن")، صفت (مانند "باز کردن")، عبارتی که با حرف اضافه شروع شود (مانند "درجا زدن") و حرف اضافه یا قید (مانند "بار آوردن") باشد [۱]. به جزء فعلی در افعال مرکب فعل سبک نیز گفته می‌شود. افعال مرکب بر دو نوع می‌باشند؛ پیوسته و گسسته. فعل مرکب پیوسته فعلی است که مابین دو جزء آن کلمه یا کلماتی قرار نگرفته باشد مانند فعل "تلاش کرد" در جمله "او تلاش کرد نامه را بدست آورد." و یا فعل مرکب "پر می‌زد" در جمله‌ی "دلش برای ساندویچ پر می زد". فعل مرکب گسسته فعلی است که مابین دو جزء آن کلمه یا کلماتی قرار گرفته باشد مانند فعل "تلاش کرد" در جمله "تلاش زیادی کرد تا نامه را بدست آورد" و یا فعل مرکب "تهمت زدن" در جمله‌ی "برای اینکه تهمت به این بزرگی بزنید لازم است دلیل کافی داشته باشید".

افعال مرکب که در برخی زبان‌های دیگر به آن گزاره‌های مرکب نیز گفته می‌شود، به طور فراوان در زبان‌های آسیایی از جمله زبان فارسی، هندی و زبان‌های دیگر از خانواده‌ی زبان‌های هندی آریایی مورد استفاده قرار می‌گیرند [۲]. تاکنون

مطالعات بسیاری در زمینه افعال مرکب برای استخراج مدلی از ساختار افعال مرکب و قواعد معنایی مرتبط با آن انجام شده است [۳]، [۴]. افعال مرکب قدرت زبان را افزایش می‌دهند. تشخیص و ترجمه‌ی افعال مرکب برای کارهای مختلفی که در پردازش زبان طبیعی صورت می‌گیرد، از اهمیت خاصی برخوردار است. از جمله کاربردهای تشخیص افعال در حوزه پردازش زبان طبیعی می‌توان به ترجمه ماشینی، بازیابی اطلاعات، درک متن و خلاصه سازی اشاره کرد. همچنین تهیه‌ی پایگاه داده‌ای از همه‌ی افعال موجود در یک زبان، برای لغت‌نویسان، طراحان وردنت و دیگر طراحانی که در حوزه پردازش زبان طبیعی فعالیت می‌کنند، یک منبع داده‌ی با ارزش است [۲].

برای تحقق موارد مذکور نیاز به راهکاری وجود دارد که بتواند افعال مخصوصاً افعال مرکب را در جملات به درستی تشخیص دهد.

روش‌هایی که برای تشخیص افعال در زبان‌های طبیعی ارائه می‌شوند به دو دسته‌ی روش‌های با نظارت و روش‌های بدون نظارت تقسیم بندی می‌شوند. در روش‌های با نظارت از دو پیکره‌ی آموزش و آزمایش استفاده می‌شود. پیکره‌ی آموزش دارای برچسب فعل مرکب است. الگوریتم با پردازش این پیکره دانشی درباره‌ی افعال موجود در پیکره به‌دست می‌آورد و در مرحله‌ی ارزیابی با استفاده از این دانش، افعال موجود در پیکره‌ی آزمایش را شناسایی می‌کند. هرچند در مرحله‌ی آزمون نیز می‌توان دانشی از همان پیکره‌ی مورد استفاده یعنی پیکره‌ی آزمایش به‌دست آورد. در روش‌های بدون نظارت پیکره‌ی آموزش و آزمایش دارای برچسب فعل مرکب نیستند. الگوریتم با پردازش پیکره و کسب دانش بسته به ویژگی‌های مورد استفاده، به شناسایی افعال می‌پردازد.

با توجه به اینکه افعال ممکن است چندجزئی باشند و این اجزاء با فاصله در جمله ظاهر شوند، تشخیص فعل یکی از مسائل چالش‌برانگیز در پردازش متون است. در زبان فارسی با توجه به ترتیب آزاد کلمات و همچنین فراوانی افعال مرکب و امکان وجود فاصله میان دو جزء فعل، امر شناسایی فعل با چالش‌های بیشتری روبروست که تاکنون حل نشده‌اند.

در ادامه‌ی مقاله ابتدا مروری بر کارهای انجام شده در زمینه شناسایی افعال در زبان فارسی و زبانهای دیگر انجام می‌شود. سپس روش پیشنهادی مبتنی بر رویکرد بدون نظارت شرح داده می‌شود. سپس روش پیشنهادی با نظارت شرح داده می‌شود و در ادامه نتایج و ارزیابی نشان داده می‌شود.

۲- کارهای مرتبط

از جمله روش‌های تشخیص افعال، الگوریتم مطرح‌شده برای زبان اُریا [۵] است. در این روش ابتدا هر کلمه‌ی فایل متنی ورودی، برچسب نحوی زده می‌شود. در ادامه کلمات کاندید برای هر فعل سبک مشخص می‌شود. طبق قواعدی از پیش تعریف شده‌ای، کلمات کاندیدی که صلاحیت ندارند حذف می‌شوند و در پایان گزاره‌های مرکب نهایی استخراج می‌شوند.

الگوریتم دیگر، یکی از الگوریتم‌های مطرح‌شده برای زبان هندی [۲] است. در این روش تعیین گزاره‌ی مرکب براساس تعیین عدم تطابق معنی فعل سبک هندی با معنی انگلیسی آن در جمله‌ی انگلیسی مطابقت داده شده با آن در یک پیکره موازی هندی- انگلیسی می‌باشد.

الگوریتم ارائه شده برای زبان بنگالی [۳] الگوهای لغوی افعال مرکب را به‌دست می‌آورد و اگر ترکیبی یکی از الگوها را دارا باشد به عنوان گزاره‌ی مرکب انتخاب می‌شود. گزاره‌های مرکب در زبان بنگالی از دو نوع افعال چندجزئی و افعال متصل تشکیل می‌شوند. اگر برچسب ریشه‌ی کلمه‌ی اول ([XXX]) اسم یا صفت (n/adj) و برچسب کلمه‌ی بعدی ([YYY]) فعل باشد (v)، الگوی لغوی مذکور به عنوان افعال از نوع متصل و اگر برچسب ریشه‌ی کلمه‌ی اول ([XXX]) اسم یا صفت (n/adj) و برچسب کلمه‌ی بعدی ([YYY]) فعل باشد (v)، الگوی لغوی مذکور به عنوان افعال از نوع افعال چندجزئی درنظر گرفته می‌شود. [XXX] و [XXX] می‌توانند نشان‌دهنده‌ی هر کلمه‌ای باشند. الگوریتم مذکور نیز از یک پیکره بدون برچسب نحوی استفاده می‌کند ولی با استفاده از یک تجزیه‌گر سطحی، اطلاعات نحوی مربوط به کلمات پیکره را به‌دست می‌آورد.

الگوریتم ارائه شده برای زبان چکسلواکی [۴] با استفاده از برچسب‌های نحوی پیکره و مشخص کردن گروه‌های فعلی، قواعدی را به طور خودکار از روی پیکره‌ی برچسب خورده یادگیری می‌کند و با توجه به قواعد یادگرفته شده، به تشخیص افعال در جملات می‌پردازد.

یکی از الگوریتم‌هایی که برای تشخیص افعال در زبان انگلیسی مطرح‌شده است [۶]، علاوه بر اینکه از یک پیکره‌ی برچسب خورده استفاده می‌کند، از بخش‌بند اسمی و بخش‌بند فعلی نیز بهره می‌برد. این الگوریتم به تشخیص افعال عبارتی از نوع مرکب پیوسته و مرکب گسسته می‌پردازد. همچنین برای تشخیص افعال مرکب از یک ماشین خودکار متناهی استفاده می‌کند. این ماشین برای هر نوع از افعال مرکب پیوسته و گسسته، توابع کنترلی خاص آن‌ها را دارد. F-score برای این روش، برابر با ۶۲.۷٪ است.

برای شناسایی فعل در زبان فارسی نیز در سالهای اخیر فعالیت هایی صورت گرفته است. یکی از روش‌های ارائه شده برای زبان فارسی روش صالحی و همکاران [۷] است. در این روش، ۷ ویژگی مطرح شده‌اند و بر اساس این ویژگی‌ها، افعال مرکب از افعال غیر مرکب مشخص می‌شوند. ویژگی‌های استنتاج شده به ۳ گروه تقسیم می‌شوند. گروه اول شامل ۳ ویژگی مبتنی بر بی‌نظمی (بی نظمی) هستند. گروه دوم شامل ۲ ویژگی جدید هستند که اتصالات کلمات تشکیل‌دهنده افعال مرکب در زبان مبدأ و زبان مقصد را مقایسه می‌کنند.

گروه سوم از ویژگی‌ها شامل ۲ ویژگی بوده و تعداد رخداد یا بسامد نگاشت‌هایی را که از مبدأ به مقصد و از مقصد به مبدأ صورت گرفته را بررسی می‌کنند. این ویژگی‌ها مختص پیکره‌ی دو زبانه‌ی TEP است. در این روش از دو رده‌بند استفاده شده است؛ رده‌بند MLP و 3NN. مقدار F-Score برای این الگوریتم با استفاده از رده‌بند 3NN برای افعال مرکب برابر با ۷۵.۸۸٪ و برای افعال غیر مرکب برابر با ۷۹.۵۸٪ است. با استفاده از رده‌بند MLP مقدار F-Score برای افعال مرکب برابر با ۷۵.۳۶٪ و برای افعال غیر مرکب برابر با ۸۲.۴۷٪ است.

روش ارائه شده‌ی دیگر برای زبان فارسی روش رسولی و همکاران [۸] است. در این روش دو الگوریتم مطرح‌شده است؛ یکی الگوریتم خودراه‌انداز و دیگری الگوریتم Kmeans. در الگوریتم خودراه‌انداز براساس معیار وابستگی اول تعریف شده ، تعدادی فعل مرکب (K تا) که مقدار معیار وابستگی آن‌ها از حد آستانه بیشتر باشد به عنوان فعل مرکب درست تشخیص داده می‌شوند و در هر مرحله با درنظر گرفتن افعال مرکبی که در مراحل قبل مشخص شدند، افعال مرکب باقی مانده در پیکره معین می‌شوند. سپس پس از چند مرحله که تعداد آن نیز از قبل مشخص می‌شود به جای محاسبه معیار وابستگی اول، از معیار وابستگی دوم تعریف شده استفاده می‌شود تا افعال مرکب گسسته نیز تشخیص داده شوند.

معیار وابستگی اول، نسخه تغییریافته معیار PMI[۱] [۹] است که در آن فقط کلمات کاندیدی که نسبت به جزء فعلی، بلافصل هستند را در نظر می‌گیرد. معیار وابستگی دوم نیز نسخه تغییریافته معیار PMI است. در این معیار احتمال فعل مرکب بودن هر دو کلمه که بدون فاصله در کنار یکدیگر ظاهر شده‌اند عددی بین صفر و یک است (برخلاف معیار PMI که این احتمال را یک درنظر می‌گیرد). مقدار F-score برای این روش بر اساس معیار وابستگی اول برابر با ۷۸.۱۱٪ و بر اساس معیار وابستگی اول و دوم برابر با ۷۸.۷۰٪ است.

همانطور که ملاحظه شد، تعدادی از روش‌ها بر روی پیکره‌ی دو زبانه و تعدادی دیگر بر روی پیکره‌ی یک زبانه (با یا بدون برچسب نحوی) اعمال

می‌شوند. همچنین روش‌های کلی که برای شناسایی افعال تاکنون به کار رفته‌اند شامل موارد زیر است:

۱- استفاده از ماشین خودکار متناهی
۲- یادگیری قواعد حاکم بر پیکره
۳- یافتن الگوهای لغوی افعال
۴- وضع ویژگی‌های مختص پیکره و بهره‌گیری از آن‌ها
۵- استفاده از معیارهای آماری مانند PMI

برخی از مشکلاتی که در زبان فارسی در زمینه تشخیص افعال وجود دارد و تاکنون حل نشده‌اند عبارتند از ۱- درنظر نگرفتن فعل کمکی افعال اصلی ۲- پوشش ندادن تمامی اجزاء غیرفعلی ۳- بالا بودن خطای تشخیص افعال مرکب گسسته نسبت به افعال مرکب پیوسته و افعال ساده به دلیل وجود کلمه یا کلماتی که می‌تواند بین دو جزء فعل مرکب گسسته واقع شود. ۴- دقت پایین نتایج.

در ادامه با پیشنهاد دو روش بی نظارت و با نظارت سعی داریم تا حد ممکن مشکلات فوق را برطرف نماییم.

۳- روش پیشنهادی بدون نظارت برای تشخیص افعال

الگوریتم پیشنهادی ترکیبی از دو روش آماری و مبتنی بر قاعده است. روش آماری مبتنی بر محاسبات و فرمول‌های آماری و نتیجه گیری بر اساس اطلاعات آماری حاصل شده از آن‌هاست و روش مبتنی بر قواعد مبتنی بر تعدادی قاعده وضع شده و استنتاجی است و نتیجه گیری بر اساس تطابق یا عدم تطابق داده‌ها با این قواعد است و افعال مرکب توسط این قواعد پالایش می‌شوند. بنابراین به دلیل اینکه هر زبانی ویژگی‌های زبانی خاص خود را داراست، با بررسی و استخراج ویژگی‌های مختص زبان فارسی که بعضاً ممکن است در زبان‌های دیگر نیز صدق کند، قواعدی نتیجه گیری می‌شود و در الگوریتم پیشنهادی به کار می‌رود.

همچنین در این الگوریتم از هیچ گونه ساده سازی استفاده نمی‌شود فقط به منظور اختصار ، یک فعل را در پیکره‌ی مورد استفاده درنظر گرفته و الگوریتم پیشنهادی بر روی آن اعمال می‌شود و نتایج نشان داده می‌شود و تمام حالات از جمله تشخیص فعل کمکی را نیز مدنظر قرار می‌دهد. اعمال این روش بر روی هر فعل مرکب دیگر بدون صرف هزینه زیاد براحتی میسر است.

ساختار کلی الگوریتم از ۳ بخش اصلی تشکیل شده است: ۱- پردازش پیکره‌ی ورودی ۲- تشخیص کلماتی که به تنهایی و یا به همراه سایر کلمات می‌توانند به عنوان کاندیدا انتخاب شوند. ۳- امتیازدهی به کلمات منتخب و تشخیص فعل ساده از فعل مرکب با توجه به حد آستانه و در صورت مرکب بودن فعل، انتخاب کاندیدای دارای بیش‌ترین امتیاز.

۳-۱- پردازش پیکره ورودی

ابتدا کل پیکره بر اساس نقطه‌هایی که دارای برچسب "PUNC" هستند پردازش می‌شود و خروجی پردازش، جملات تشکیل‌دهنده پیکره است. به عبارت دیگر برای خرد کردن پیکره و تهیه‌ی ورودی برنامه از نقطه‌ای که برچسب "PUNC" دارد استفاده می‌شود.

لازم به ذکر است که در اینجا فرض بر این است که مرز هر جمله با نقطه معین می‌شود و سایر علائم نگارشی مانند "؟" یا "!" ملاک برای انتهای جمله درنظر گرفته نمی‌شوند.

۳-۲- تعیین افعال اصلی (ساده یا سبک) و کلمات کاندید افعال مرکب

در هر جمله، ابتدا با استفاده از برچسب کلمات و همچنین به کمک ریشه‌یاب [۱۰]، افعال تک واحدی که می‌توانند افعال ساده و یا افعال سبک باشند، تشخیص داده می‌شوند. افعال تک واحدی می‌توانند هم تک‌جزئی باشند مانند "خورد" و یا چندجزئی باشند مانند "خواهد گفت"، "زده شده باشد" و "گفته شده بود". از ریشه‌یاب برای تشخیص افعال چندجزئی استفاده می‌شود. برچسب کلماتی که فعل هستند در پیکره با حرف "V" شروع می‌شوند؛ برچسب افعال کمکی با "V AUX" شروع می‌شوند و برچسب افعال کمکی که برای نشان دادن زمان آینده به کار می‌روند مانند "خواهد" با "V AUX FUT" شروع می‌شود و صفت مفعولی (مانند خورده یا گفته) با برچسب "V PASTP" شروع می‌شود. روند تشخیص افعال تک واحدی با استفاده از ریشه‌یاب بدین شرح است:

۱. ابتدا جمله از آخر به اول پیمایش شده و افعالی که با برچسب "V" شروع می‌شوند (به جز افعال کمکی) تعیین می‌شوند و در لیست افعال اصلی ذخیره می‌شوند و همچنین افعال کمکی به استثناء افعال کمکی که برای نشان دادن زمان آینده به کار می‌روند در لیست افعال کمکی ذخیره می‌گردند.

۲. برای هر فعل در لیست افعال اصلی

الف. اگر کلمه‌ی قبل از آن صفت مفعولی (مانند "خورده") یا فعل کمکی برای زمان آینده (مانند "خواهم") باشد آن را نیز جزء فعل درنظر می‌گیریم.

ب. اگر دو کلمه‌ی ماقبل آن صفت مفعولی باشد و خروجی ریشه‌یاب که ورودی آن یک کلمه ۳ جزئی (کلمه آخر + کلمه ماقبل آخر + دو کلمه ماقبل آخر) است، مخالف تهی باشد (یعنی ریشه‌یاب این سه کلمه را یک فعل تصریف شده تشخیص دهد) آن کلمه نیز جزء فعل درنظر گرفته می‌شود.

پس از اینکه افعال ساده و سبک در جمله مشخص شدند برای هر فعل در جمله کلمات ماقبل آن را درنظر گرفته و آن‌هایی را که می‌توانند به عنوان جزء غیرفعلی فعل تک واحدی درنظر گرفته شوند را به عنوان اجزاء غیرفعلی کاندید برای فعل تک واحدی انتخاب می‌کنیم (کلمه‌ی کاندید).

برای انتخاب کلمات کاندید روند کار به این صورت است که کلمات از کلمه‌ی ماقبل فعل تا ابتدای جمله با اسامی مرتبط با فعل (اسامی که می‌توانند با فعل مذکور تشکیل یک فعل مرکب را دهند) مقایسه می‌شوند و در صورت برابری به عنوان کاندید غیرفعلی فعل مورد پردازش درنظر گرفته می‌شوند. برای تشخیص اجزاء غیرفعلی که بیش از یک جزء دارند مانند در+ جا (زدن)، نفس+ نفس (زدن) و ... که حداکثر طول این اجزاء غیرفعلی چندجزئی می‌تواند ۴ جزء باشد (مانند "به+ آب+ و+ آتش (زدن)") ۴ حالت (۴گرم، ۳گرم، ۲گرم و ۱گرم) از انتهای جمله درنظر می‌گیریم. به عنوان نمونه جمله‌ی "دوستم قبلاً در کارش در جا می‌زد " را درنظر می‌گیریم. ۴ حالت از انتهای جمله برای فعل "می‌زد" عبارتند از:

۱. درکارش‌درجا
۲. کارش‌درجا
۳. درجا
۴. جا

ابتدا برای هر کدام از حالت‌ها با استفاده از ریشه‌یاب شکل اصلی هر جزء آن‌ها به‌دست می‌آید (مانند "بیماری‌های لاعلاجی" که پس از ریشه یابی به "بیماری لاعلاج" تبدیل می‌شود). بنابراین داریم:

۱. درکاردرجا
۲. کاردرجا
۳. درجا
۴. جا

به ترتیب اولویت (اولویت با تعداد اجزاء بیشتر است) ابتدا حالت اول (شماره ۱) با مداخل موجود در پایگاه داده مقایسه می‌شود و در صورت عدم برابری حالت بعدی با مداخل موجود در جدول عبارات مورد مقایسه قرار می‌گیرد. اگر هر کدام از حالات با یکی از اسامی موجود در جدول عبارات مساوی بودند، از مقایسه‌ی بقیه‌ی حالات صرف‌نظر می‌شود و در مرحله‌ی بعد فرایند مقایسه از کلمه‌ی ماقبل کلمه‌ای که در مرحله‌ی قبل، تساوی در آن رخ داده انجام می‌گیرد. در این مثال "درکاردرجا" با هیچکدام از مداخل موجود در پایگاه داده برابر نیست. همچنین در حالت دوم یعنی "کاردرجا" نیز مانند حالت قبل تساوی رخ نمی‌دهد. ولی در حالت سوم یعنی "درجا" تساوی روی می‌دهد و بنابراین عبارت "درجا" به عنوان کلمه‌ی کاندید درنظر گرفته می‌شود. همچنین اگر هیچکدام از ۴ حالت با اسامی ذخیره شده در جدول عبارات برابر نبودند یک کلمه به عقب حرکت کرده و فرایند مقایسه از آنجا شروع می‌شود. همچنین در این مثال اگر "درجا" با مداخل موجود در جدول عبارات برابر نمی‌شد هیچ کلمه‌ی کاندیدی برای فعل "می‌زد" وجود نداشت و "می‌زد" فعل ساده درنظر گرفته می‌شد.

لازم به ذکر است که پایگاه داده بزرگی از افعال ساده و مرکب زبان فارسی در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی دسترس است که در این مرحله از آن استفاده می شود و نیاز به یادگیری انواع فعل مرکب را از بین می برد. این پایگاه مرتب در حال بروز رسانی است.

۳-۳- امتیازدهی

ابتدا توسط قواعدی از پیش تعریف شده، تعدادی از کلمات کاندید حذف می‌شوند. این قواعد شامل قواعد کلی و قواعد وابسته به فعل می‌شوند.

- ✓ قواعد کلی قواعدی هستند که در مورد تمام افعال صدق می‌کند.
- ✓ قواعد وابسته به فعل قواعدی هستند که برای تمام افعال صدق نمی‌کند و برای افعال خاصی صادق است.

سپس با استفاده از روش امتیازدهی به کلماتی که می‌توانند به عنوان جزء غیرفعلی در ترکیب فعل مرکب شرکت کنند، امتیاز داده می‌شود و بر اساس دو معیارِ امتیازهای به‌دست آمده و حد آستانه‌ی تشخیص فعل ساده از مرکب، افعال ساده و مرکب مشخص می‌شوند. برای ویژگی‌ها عدم قطعیت وجود دارد. چون ممکن است حالات استثنایی وجود داشته باشند که ویژگی‌ها را نقض کنند ولی در مورد قواعد عدم قطعیت وجود ندارد.

ویژگی‌هایی که برای امتیازدهی کلمات کاندید به‌کار می‌روند به دو دسته تقسیم می‌شوند:

- ✓ ویژگی‌های منفی. ویژگی‌هایی هستند که تاثیر منفی در امتیاز یک کلمه‌ی کاندید دارند.
- ✓ ویژگی‌های مثبت. برعکس ویژگی‌های منفی، تاثیر مثبت در امتیاز کلمه‌ی کاندید دارند.

۳-۳-۱- قواعد کلی

- ✓ اگر کلمه‌ی کاندیدی که به عنوان جزء غیرفعلی یکی از افعال مرکب موجود در یک جمله مشخص شود، در مراحل بعدی (تشخیص کلمات کاندید دیگر در همان جمله) نمی‌تواند به عنوان کلمه‌ی کاندید برای فعل مرکب دیگری درنظر گرفته شود.
- ✓ اگر قبل از کلمه‌ی کاندید حرف اضافه آمده باشد، کلمه‌ی کاندید نمی‌تواند به عنوان جزء غیرفعلی درنظرگرفته شود مگر اینکه حرف اضافه جزء کلمه‌ی کاندید باشد.
- ✓ اگر کلمه‌ی کاندید یک‌جزئی باشد مانند "پس" (مانند فعل مرکب "پس زدن") و یا "در" (مانند فعل مرکب "در زدن") و نقش آن حرف اضافه و یا حرف ربط باشد، نمی‌تواند به عنوان جزء غیرفعلی درنظر گرفته شود.

در اینجا هم یک‌جزئی بودن و هم حرف اضافه بودن مهم است، زیرا اگر کلمه‌ی کاندید دو جزئی باشد، برچسب دو جزء را نداریم. به عنوان مثال در(برچسب نحوی= حرف اضافه)+ جا(برچسب نحوی= اسم)= درجا، در حالی که برچسب "درجا" وجود ندارد. اگر برچسب عبارات چندجزئی را از قبل داشته باشیم، می‌توانیم قواعد جدیدی را نیز وضع نمائیم.

- ✓ اگر اسم کاندید اسم مصدر باشد یعنی دارای برچسب "N COM SING INFI" یا "N COM PL INFI" باشد (مانند گرفتن، شدن، کردن و ...) نمی‌تواند به عنوان جزء غیرفعلی درنظر گرفته شود. این قانون برای رفع خطای ریشه‌یاب وضع شده است.
- ✓ اگر کلمه‌ی پس از کلمه‌ی کاندید اسم مصدر باشد، کلمه‌ی کاندید نمی‌تواند به عنوان جزء غیرفعلی درنظرگرفته شود.
- ✓ اگر قبل از کلمه‌ی کاندید کسره اضافه آمده باشد و برچسب کلمه‌ی ماقبل کلمه‌ی کاندید اسم باشد و به علاوه پس از کلمه‌ی کاندید حرف ربط "که" نباشد، نمی‌تواند به عنوان جزء غیرفعلی درنظر گرفته شود.

۳-۳-۲- قواعد وابسته به فعل

در اینجا قواعد وابسته به فعل "زدن" شرح داده می‌شود.

- ✓ اگر بعد از کلمه‌ی "در"، اسم مکان مانند "کشور" و یا اسم زمان مانند "سال" آمده باشد و به علاوه پس از "در" کسره اضافه نباشد (معمولاً در چنین مواقعی "در" نقش حرف اضافه دارد)، اسم "در" نمی‌تواند به عنوان کلمه‌ی کاندید درنظر گرفته شود. بررسی وجود کسره اضافه پس از "در" مهم است؛ چون اگر پس از "در" کسره اضافه باشد مانند جمله‌ی "درِ خانه‌ی دوستم را زدم"، "در" جزء غیرفعلی فعل مرکب "در زدم" است.
- ✓ کلماتی هستند که اگر در جمله به عنوان جزء غیرفعلی ظاهر شوند، آنگاه مابین آن‌ها و افعال سبک آن‌ها نمی‌تواند کلمه یا کلماتی قرار گیرد. مانند "کلید" در جمله‌ی "او این پروژه را کلید زد". بنابراین اگر مابین "کلید" و فعل سبک کلمه‌ای آمده باشد، "کلید" نمی‌تواند جزء غیرفعلی باشد.
- ✓ کلماتی هستند که اگر در جمله به عنوان جزء غیرفعلی ظاهر شوند و مابین آن‌ها و افعال سبک، کلمه‌ای آمده باشد، آن کلمه به تعدادی کلمه خاص محدود می‌شود. به عنوان مثال اگر کلمه‌ی کاندید، "دست" باشد و بعد از آن حرف "به" آمده باشد، کلمه‌ی کاندید "دست" جزء غیرفعلی خواهد بود.

۳-۳-۳- ویژگی‌های منفی

۱. تعداد کلمات قرار گرفته مابین دو جزء فعل مرکب. مثلاً در جمله‌ی " آبی به سر و صورتش زد" تعداد کلمات قرار گرفته مابین فعل مرکب "آبی زد" (تعداد کلمات عبارت "به سر و صورتش") برابر با ۴ می‌باشد. (در معادله‌ی امتیازدهی با حرف A نشان داده شده است).

۲. تعداد افعال قرار گرفته مابین دو جزء فعل مرکب (تعداد کلماتی که مابین دو جزء قرار می‌گیرند و برچسب فعل دارند). به عنوان مثال در جمله‌ی "لطمه به مالتان می‌توانند بزنند" یک فعل (می‌توانند) مابین کلمه‌ی کاندید "لطمه" و فعل سبک "بزنند" قرار گرفته است. (تمام افعال اعم از ساده یا سبک صرفنظر از نوع آن‌ها درنظر گرفته می‌شوند). (در معادله‌ی امتیازدهی با حرف B نشان داده شده است).

۳. تعداد علائم نقطه‌گذاری مانند ":" ، "،" و "؟" قرار گرفته مابین دو جزء فعل مرکب. به عنوان مثال در جمله‌ی "پس از دو دهه انقلاب، دفاع مقدس و سرمایه گذاری مادی و معنوی مردم این مرز و بوم آیا در تعریف هزینه اصلاحات برای مردم تنها باید مشارکت سیاسی آن‌ها را مطلوب این پروسه بدانیم؟ یا تا کسی حرفی از لزوم نگاه اقتصادی برای مردم در روند اصلاحات می‌زند ..." مابین کلمه‌ی کاندید "پس" یا کلمه‌ی کاندید "در" و فعل سبک "می‌زند" یک علامت نگارشی "؟" واقع شده است (در معادله‌ی امتیازدهی با حرف C نشان داده شده است).

۴. تعداد حرف "را" قرار گرفته مابین دو جزء فعل مرکب. مثلاً در جمله‌ی "سیلی محکمی را به صورتش زد" مابین دو جزء "سیلی" و "زد" یک حرف "را" واقع شده است (در معادله‌ی امتیازدهی با حرف D نشان داده شده است).

۵. وجود یکی از اعداد شمارشی یا ترتیبی مانند ۲، یک، هفدهم، آخر و ... قبل از کلمه‌ی کاندید. مانند جمله‌ی "آن را با هزار ترفند و تردستی وارد ادبیات سیاسی نمائیم" که در آن "هزار" عدد شمارشی و "ترفند" کلمه‌ی کاندید برای فعل "زد" است (در معادله‌ی امتیازدهی با حرف E نشان داده شده است).

۳-۳-۴- ویژگی‌های مثبت

۱. نسخه‌ی تغییریافته معیار وابستگی PMI. یکی از معیارهای معروف وابستگی آماری است که مقدار آن از رابطه‌ی (۱) به‌دست می‌آید (در معادله‌ی امتیازدهی با حرف F نشان داده شده است):

$$PMI = Log_2(\frac{n\times f(x,y)}{f(x)\times f(y)}) \qquad (۱)$$

که در آن:

n: تعداد کل کلمات

$f(x,y)$: بسامد کلمه‌ی کاندید هنگامی که می‌تواند با فعل سبک همراه شده و تشکیل فعل مرکب بدهد

$f(x)$: بسامد کلمه‌ی کاندید در کل پیکره

$f(y)$: بسامد فعل سبک در کل پیکره

۲. درنظر گرفتن امتیاز مثبت برای کلمه‌ی کاندید بلافصل فعل اصلی. از آنجا که در متون فارسی در بیشتر مواقع مابین دو جزء فعل مرکب کلمه یا کلماتی نمی‌آیند، بنابراین با درنظر گرفتن امتیاز مثبت برای چنین کلمات کاندیدی، احتمال درنظر گرفتن آن‌ها به عنوان جزء غیرفعلی را افزایش می‌دهیم. به عنوان مثال در جمله‌ی "به سر و صورتش آبی زد"، "آبی" کلمه‌ی بلافصل فعل "زد" است (در معادله‌ی امتیازدهی با حرف G نشان داده شده است).

۳. مساوی بودن کلمه‌ی کاندید با شکل اصلی آن. مانند جمله‌ی "دستم را به پشتش زدم" که کلمه‌ی کاندید "دستم" برای فعل "زدم" به شکل شکل اصلی خود نیامده است (در معادله‌ی امتیازدهی با حرف H نشان داده شده است). همچنین در صورت نامساوی بودن کلمه با شکل اصلی خود، امتیاز منفی (۰.۵-) برای آن منظور می‌شود.

۴. درنظر گرفتن امتیاز مثبت برای دنباله‌ی کلمات کاندید (در معادله‌ی امتیازدهی با حرف I نشان داده شده است). چون روال معمول الگوریتم به این صورت است که فقط کلمه‌ی کاندید با بیش‌ترین امتیاز درنظر گرفته می‌شود و از مابقی کلمات کاندید که امتیاز کمتری دارند صرفنظر می‌شود، بنابراین در جملاتی نظیر جمله‌ی "او هر چه تلاش و کوشش می‌کند تا ..." فقط کوشش کلمه‌ی کاندید نهایی می‌شود. از اینرو در مورد عباراتی مانند "تلاش و کوشش کردن" که دو یا چند کلمه‌ی کاندید به وسیله‌ی حرف ربط "و" و یا "،" به دنبال هم می‌آیند (تلاش و کوشش) فقط یک کلمه‌ی کاندید به عنوان کلمه‌ی کاندید نهایی برگردانده نشود (کوشش) و تمام کلمات کاندید در یک دنباله برگردانده شوند (تلاش و کوشش) به صورت زیر عمل می‌شود.

ابتدا کل کلمات کاندید که به دنبال هم آمده‌اند جایگزین کلمات کاندید ماقبل آخر می‌شوند. در عبارت مذکور "تلاش و کوشش" جایگزین "تلاش" می‌شود چون "تلاش" کلمه‌ی کاندید آخر در دنباله‌ی "تلاش و کوشش" است. سپس یک امتیاز مثبت برای "تلاش و کوشش" لحاظ می‌شود. بنابراین کلمات "تلاش و کوشش" و "کوشش" به عنوان کلمات کاندید شناخته می‌شوند. (امتیاز حاصل از ترکیب دو کلمه‌ی کاندید متوالی در معادله‌ی امتیازدهی با حرف J نشان داده شده است.)

۳-۳-۵- برآورد امتیازها

در اینجا با ایجاد یک معادله که متغیرهای آن ضرایب ویژگی‌های استخراج شده که پیش‌تر ذکر آن‌ها گردید، می‌باشد به محاسبه امتیاز هر کلمه‌ی کاندید پرداخته می‌شود. این معادله به صورت رابطه (۲) است ($co_{\text{نام ویژگی}}$ = ضریب ویژگی):

$$score = -(\frac{co_A\times A + co_B\times B + co_c\times C + co_D\times D}{length_{sentence}} + co_E\times E) \qquad (۲)$$
$$+co_F\times F + \frac{co_G\times G}{length_{sentence}} + co_H\times H + co_I\times I$$

با پیدا کردن ضرایب بهینه و اعمال آن‌ها در معادله‌ی امتیازدهی، امتیاز هر کلمه‌ی کاندید مشخص می‌شود و با انتخاب بیش‌ترین امتیاز، در صورتی که این امتیاز از حد آستانه (حد آستانه به صورت تجربی به‌دست می‌آید) بیشتر باشد، کلمه‌ی کاندید با بیش‌ترین امتیاز به عنوان جزء غیرفعلی فعل مرکب انتخاب می‌شود و اگر از حد آستانه کمتر باشد، فعل سبک به عنوان فعل ساده درنظر گرفته می‌شود.

به عنوان مثال جمله‌ی "پس از دو دهه انقلاب، دفاع مقدس و سرمایه گذاری مادی و معنوی مردم این مرز و بوم آیا در تعریف هزینه اصلاحات برای مردم تنها باید مشارکت سیاسی آنها را مطلوب این پروسه بدانیم؟ یا تا کسی حرفی از لزوم نگاه اقتصادی برای مردم در روند اصلاحات می‌زند ..." را درنظر می‌گیریم. کلمات کاندید برای فعل اصلی "می‌زند" عبارتند از:

۱. پس

۲. تا

۳. حرفی

۴. در (قبل از کلمه "تعریف")

۵. در (قبل از کلمه "روند")

در اینجا امتیاز هر کلمه‌ی کاندید با استفاده از معادله‌ی امتیازدهی محاسبه می‌شود و بیشترین امتیازها محاسبه می‌شود و اگر این بیشترین امتیاز از حد آستانه بیشتر بود، به عنوان جزء غیرفعلی فعل سبک "می‌زند" درنظر گرفته می‌شود و اگر کمتر از حد آستانه بود فعل "می‌زند" به عنوان فعل ساده مشخص می‌شود.

۳-۴- تعیین افعال کمکی

در مرحله‌ی تشخیص افعال سبک که با کمک برچسب‌ها یا گروه‌های نحوی و همچنین بهره‌گیری از ریشه‌یاب مشخص می‌شوند، افعال کمکی نیز مشخص می‌شوند (در لیستی جداگانه ذخیره می‌شوند). برچسب این گونه افعال با "V AUX" شروع می‌شود؛ مانند می‌توان، می‌توانم، می‌شود، باید، بایستی و غیره. بعضی از این افعال هم به صورت فعل ساده و هم به صورت فعل کمکی می‌آیند

مانند داشتم، می‌توانستم که در اینجا با توجه به برچسب آن‌ها، افعال کمکی از افعال ساده مشخص می‌شوند.

پس از مشخص شدن افعال کمکی نوبت به تخصیص آن‌ها به افعال ساده یا افعال مرکب می‌رسد. ابتدا قواعد و ویژگی‌هایی که برای تعیین افعال کمکی استفاده می‌شوند بیان می‌شود، سپس روند کلی تشخیص افعال کمکی شرح داده می‌شود.

۳-۴-۱- قواعد مختص افعال کمکی

در اینجا قواعدی را که مختص افعال کمکی است و از پیکره‌ی بیجن‌خان استنتاج شده است، آورده می‌شود.

- ✓ افعالی که با برچسب "V AUX NIN" شروع می‌شوند نمی‌توانند با فعلی بیایند که زمان آن حال یا مضارع است.
- ✓ اگر یک فعل کمکی برای یک فعل اصلی دیگر در جمله در مراحل قبلی انتخاب شده باشد، آن فعل کمکی شانس انتخاب شدن برای فعل مورد پردازش را از دست می‌دهد و نمی‌تواند در لیست افعال کمکی کاندید قرار گیرد.
- ✓ فعل "می‌شود" و "نمی‌شود" در نقش فعل کمکی و "می‌توان" و "نمی‌توان" فقط می‌توانند با فعل ماضی سوم شخص مفرد بیایند.

۳-۴-۲- ویژگی‌های مختص افعال کمکی

برای برآورد امتیاز هر فعل کمکی کاندید از ۴ ویژگی زیر استفاده می‌کنیم:

۱. تعداد افعال مابین فعل کمکی و فعل اصلی.

ممکن است بین فعل اصلی و فعل کمکی کاندید چندین فعل صرفنظر از نوع آن‌ها قرار گیرد مانند جمله "او می‌تواند با ارواحی که در این جهان نیستند ارتباط برقرار کند".

۲. تعداد علائم نقطه گذاری مانند "?" و ":" مابین فعل کمکی و فعل اصلی.

۳. تعداد حرف "را" مابین فعل کمکی و فعل اصلی.

۴. تعداد کلمات قرار گرفته مابین فعل کمکی و فعل اصلی.

هر کدام از ۴ ویژگی مذکور تاثیر منفی در محاسبه امتیاز هر فعل کمکی کاندید دارند. همچنین برای محاسبه‌ی امتیاز هر فعل کمکی کاندید، هر کدام از ۴ ویژگی بر طول جمله تقسیم می‌شوند. منظور از جمله، جمله‌ای است که فعل اصلی و فعل کمکی کاندید در آن قرار گرفته‌اند.

۳-۴-۳- روش تعیین افعال کمکی

روش تعیین افعال کمکی به صورت زیر است:

- ✓ مشخص کردن افعال کمکی کاندید برای هر یک از افعال اصلی موجود در جمله
- ✓ اعمال قواعد مختص افعال کمکی بر روی افعال کمکی کاندید
- ✓ محاسبه‌ی امتیاز هر یک از افعال کمکی کاندید حاصل از پالایش مرحله قبل
- ✓ انتخاب فعل کمکی با بیش‌ترین امتیاز، در صورت بزرگ‌تر بودن امتیاز آن از حد آستانه

در اینجا حد آستانه از قبل به صورت دستی تعیین می‌شود. اگر فعل کمکی با بیش‌ترین امتیاز، امتیازش از حد آستانه بیشتر نباشد، فعل کمکی متعلق به آن فعل اصلی مورد پردازش نیست و به فعل دیگری تعلق دارد. در اینجا فعل اصلی می‌تواند فعل ساده یا فعل سبک باشد. مثلاً در جمله‌ی "او می‌تواند به من کمک کند"، "کند" فعل سبک و "می‌تواند" فعل کمکی آن و فعل این جمله "می‌تواند کمک کند" است و در جمله‌ی "او باید به مدرسه برود"، "برود" فعل ساده و "باید" فعل کمکی آن و "باید برود" فعل این جمله است.

۴- روش پیشنهادی با نظارت برای تشخیص افعال

در این روش دو سوم کل جملات پیکره برای آموزش و یک سوم باقیمانده برای آزمون به کار می‌رود. جملات به‌کار رفته برای مرحله‌ی آموزش و مرحله‌ی آزمون به‌صورت تصادفی از پیکره انتخاب می‌شوند. جداول مورد استفاده در این روش به شرح زیر است:

۱. جدول افعال که حاوی تمامی افعال مرکب که شامل بن ماضی، بن مضارع و شکل مصدری افعال است، می‌باشد (هر فعل مرکب جدیدی که قرار است برنامه آن را مدنظر قرار دهد به این جدول اضافه می‌شود).

۲. جدول افعال سبک که حاوی افعال سبک استخراج شده از جدول افعال مرکب به همراه بن ماضی و بن مضارع می‌باشد.

۳. جدول عبارات که حاوی اجزاء غیرفعلی افعال مرکب موجود در جدول افعال مرکب است و بین این جدول و جدول افعال سبک یک رابطه وجود دارد. همچنین این جدول حاوی فیلدهایی است که برای ذخیره مقادیر ویژگی‌های استخراج شده از پیکره‌ی مورد استفاده قرار می‌گیرد.

۴. جدول دیگری به نام جدول برچسب نیز وجود دارد که در آن حالاتی که یک جزء غیرفعلی با فاصله (حداقل یک کلمه مابین دو جزء غیرفعلی و فعل سبک قرار گیرد) در کنار فعل سبک قرار می‌گیرد و همچنین تعداد رخداد آنها در پیکره، ذخیره می‌گردد. داده‌های این جدول با پردازش پیکره‌ی آموزش به‌دست می‌آید. روند به‌دست آوردن داده‌های این جدول نیز به این صورت است که پس از اینکه کلمات کاندید در یک جمله مشخص شدند اگر پس از کلمات کاندید تا فعل سبک کلمه یا کلمات دیگری قرار گرفته باشند، آن کلمات در این جدول در سطر مربوطه ذخیره می‌شوند. همچنین تعداد رخداد آنها نیز در این جدول ذخیره می‌شود.

ابتدا دو سوم جملات پیکره به‌صورت دستی برچسب فعل مرکب زده می‌شوند. سپس با پردازش این جملات برچسب‌خورده، اطلاعات مورد نظر به‌دست می‌آیند. این اطلاعات شامل مقادیر ویژگی‌های مختص پیکره دارای برچسب فعل مرکب است. سپس اطلاعات به‌دست آمده در پایگاه داده ذخیره می‌شود. ویژگی‌های مختص پیکره دارای برچسب فعل مرکب به شرح زیر است:

ذخیره تعداد دفعاتی که یک کلمه‌ی کاندید به عنوان جزء غیرفعلی با فعل سبک ترکیب فعل مرکب را داده است به همراه تعداد کل دفعاتی که به عنوان کلمه‌ی کاندید درنظر گرفته شده است.

تعداد مواردی را که حرف "را" مابین دو جزء فعل مرکب قرار گرفته است، ذخیره می‌شود.

تعداد مواردی را که فعل یا فعل‌هایی (به همراه تعداد افعال) مابین دو جزء فعل مرکب قرار گرفته‌اند، ذخیره می‌شود (به عنوان مثال از ۱۰ مورد تکرار فعل مرکب "حرف زدن" ۳ حالت وجود دارد که فعل یا فعل‌هایی مابین "حرف" و "زدن" واقع شده است که در ۲ حالت تعداد افعال مابین "حرف" و "زدن" ۱ بوده و در یک حالت تعداد افعال، ۲ بوده است) تا بدین وسیله تعداد موارد استثناء (غالباً فعل مابین دو جزء فعل مرکب قرار نمی‌گیرد) در چنین مواردی برای یک جزء غیرفعلی می‌تواند رخ دهد، مشخص شود و در مرحله‌ی آزمون به عنوان امتیاز مثبت برای کلمه‌ی کاندید منظور شود.

تعداد مواردی را که علائم نگارشی (به همراه تعداد آن) مابین دو جزء فعل مرکب قرار گرفته است، ذخیره می‌شود (به عنوان مثال از ۱۰ مورد تکرار فعل

مرکب "حرف زدن" ۳ حالت علامت نگارشی مابین "حرف" و "زدن" واقع شده که در ۲ حالت تعداد علامت نگارشی مابین "حرف" و "زدن" ۱ بوده است و در یک حالت تعداد علامت نگارشی، ۲ بوده است) تا بدین وسیله تعداد موارد استثناء (غالباً فعل مابین دو جزء فعل مرکب قرار نمی‌گیرد) در چنین مواردی برای یک جزء غیرفعلی می‌تواند رخ دهد، مشخص شود و در مرحله‌ی آزمون به عنوان امتیاز مثبت برای کلمه‌ی کاندید نظیر آن منظور شود.

برچسب سطح اول یا عمومی کلمات واقع شده مابین دو جزء فعل مرکب ذخیره می‌شود. به عنوان مثال در جمله‌ی "مامان سیلی محکمی به صورتش زد"، برچسب سطح اول کلمات "محکمی به صورتش" به ترتیب عبارتند از"AJ"، "P" و "N" که این برچسب‌ها ذخیره می‌شوند.

تعداد کلمات به همراه تعداد رخداد آن به شکل دوتایی ذخیره می‌شود (مؤلفه اول تعداد کلمات و مؤلفه دوم تعداد تکرار آن). به عنوان مثال فرض شود مابین "حرف" و "زدن" در فعل مرکب "حرف زدن"، ۳ بار کلمات ۲ جزئی (عبارت ۲ جزئی مانند "بسیار خوبی" در جمله‌ی "حرف‌های بسیار خوبی زدم")، ۲ بار کلمات ۱ جزئی (کلمه ۱ جزئی مانند "خوبی" در جمله‌ی "حرف‌های خوبی زدم") در پیکره‌ی آموزش آمده باشد. بنابراین دو تایی‌های (۳،۲) و (۲،۱) در پایگاه داده ذخیره می‌گردند.

۴-۱- برآورد امتیازها

همانند روش بدون نظارت با ایجاد یک معادله که متغیرهای آن ضرایب ویژگی‌های استخراج شده که پیش‌تر ذکر آنها گردید، می‌باشد به محاسبه امتیاز هر کلمه‌ی کاندید پرداخته می‌شود این معادله به صورت رابطه (۳) است ($co_{\text{نام ویژگی}}$ = ضریب ویژگی):

$$score = co_A \times A + co_B + co_C + co_D + co_E + co_F \quad (۳)$$

- ✓ A: مقدار این متغیر برابر است با نسبت تعداد دفعاتی که کلمه‌ی کاندید در پیکره‌ی آموزش به عنوان جزء غیر فعلی فعل مرکب آمده به تعداد کل دفعاتی که در پیکره‌ی آموزش آمده است.
- ✓ B: در صورتی که تعداد افعال مابین دو جزء فعل مرکب حداقل یک باشد، برابر با نسبت تعداد افعال ذخیره شده به تعداد تکرار جزء غیرفعلی در پیکره‌ی آموزش است. اگر مقدار به‌دست آمده صفر باشد، مقدار ۱- در متغیرB ذخیره می‌شود.
- ✓ C: در صورتی که تعداد علائم نگارشی مابین دو جزء فعل مرکب حداقل یک باشد، برابر با نسبت تعداد علائم نگارشی ذخیره شده به تعداد تکرار جزء غیرفعلی در پیکره‌ی آموزش است. اگر مقدار به‌دست آمده صفر باشد، مقدار ۱- در متغیرC ذخیره می‌شود.
- ✓ D: در صورتی که تعداد حرف "را" مابین دو جزء فعل مرکب حداقل یک باشد، برابر با نسبت تعداد حرف "را" ذخیره شده به تعداد تکرار جزء غیرفعلی در پیکره‌ی آموزش است. اگر مقدار به‌دست آمده صفر باشد، مقدار ۱- در متغیر D ذخیره می‌شود.
- ✓ E: در صورتی که تعداد کلمات مابین دو جزء فعل مرکب حداقل یک باشد، برابر با مجموع امتیازات برچسب کلماتی که مابین دو جزء فعل مرکب واقع شده است و با برچسب‌های ذخیره شده در پایگاه داده در مکان‌های نظیرشان (یعنی برچسب کلمه اول از کلمات مابین دو جزء فعل مرکب با برچسب کلمه اول موجود در پایگاه داده) با هم برابر هستند.

به عنوان مثال فرض می‌کنیم برچسب‌های ذخیره شده در جدول عبارات پایگاه داده به‌صورت زیر است:

برچسب کلمه اول: می‌تواند حرف اضافه- ۳ (یعنی ۳ بار برچسب کلمه اول در پیکره‌ی آموزش برای جزء غیرفعلی مدنظر، حرف اضافه بوده است) یا اسم-۵ یا صفت- ۲ باشد.

برچسب کلمه دوم: اسم- ۲ یا حرف اضافه- ۶ می‌تواند باشد.

برچسب کلمه سوم: حرف اضافه- ۴ می‌تواند باشد.

به ازای هر تطابق با حفظ ترتیب برچسب‌ها، مقدار E برابر مجموع امتیاز هر کلمه می‌شود. فرض کنیم برچسب کلمات مابین دو جزء فعل مرکب به ترتیب از راست به چپ اسم – فعل – حرف اضافه - اسم باشد. چون برچسب کلمه‌ی اول می‌تواند اسم باشد۵/۱۰ یا ۰.۵ امتیاز که ۱۰ تعداد کل بسامد برچسب‌ها برای کلمه‌ی اول است (۳+۵+۲) و به این ترتیب تأثیر برچسبی که تکرار بیشتری در پیکره‌ی آموزش داشته است را افزایش می‌دهیم (در اینجا برچسب اسم بیشترین بسامد (۵ تکرار) را دارد). چون برچسب کلمه‌ی دوم و کلمه‌ی سوم نمی‌تواند فعل باشد جستجو خاتمه می‌یابد و برچسب کلمات بعدی ورودی (حرف اضافه - اسم) بررسی نمی‌شوند و امتیاز E برابر ۰.۵می‌شود.

- ✓ F: اگر تعداد کلماتی که مابین دو جزء فعل مرکب آمده‌اند با یکی از مؤلفه‌های اول دو تایی‌های (تعداد کلمات، تعداد تکرار آن) موجود در پایگاه داده برابر باشد مقدار F، برابر نسبت مؤلفه‌ی دوم دوتایی منطبق شده به کل تعداد تکرارها (جمع مؤلفه‌های دوم تمام دو تایی‌ها) و در غیراینصورت صفر است.

به عنوان مثال فرض کنیم دو تایی‌های (۶،۲)، (۵،۱) و (۳،۴) در جدول عبارات برای یک کلمه‌ی کاندید ذخیره شده باشد؛ اگر تعداد کلمات مابین دو جزء فعل مرکب در پیکره‌ی آزمایش با یکی از اعداد ۲ یا ۱ یا ۴ یکسان باشد (مثلاً ۴)، مقدار F برابر با ۳/(۳+۵+۶) می‌شود.

روش تعیین فعل ساده یا فعل مرکب و انتخاب کلمه‌ی کاندید به عنوان جزء غیرفعلی برای فعل مرکب همانند روش بدون نظارت انجام می‌گیرد.

۵-آزمون و ارزیابی

پیکره‌ی ورودی برای آزمون باید برچسب نحوی داشته باشد مانند پیکره‌ی بیجن‌خان. برای آزمون و اعمال روش پیشنهادی و به منظور ساده شدن و پرهیز از پیچیدگی نتایج، فقط فعل‌های مرکبی که با فعل سبک "زد" ساخته می‌شوند (که یکی از فعل‌های رایج مرکب در زبان فارسی می‌باشد) را از پیکره‌ی بیجن‌خان [۱۱] استخراج و الگوریتم پیشنهادی را بر روی آن اعمال می‌کنیم. اگرچه این الگوریتم برای هر فعل دیگری که در زبان فارسی وجود دارد قابل استفاده است مانند فعل سبک "کرد" و غیره.

بهترین مقدار برای حد آستانه به منظور تشخیص افعال شامل فعل کمکی، برابر با صفر، مقدار حد آستانه برای تشخیص فعل ساده از فعل مرکب برابر با ۲.۲۵- است. در روش بدون نظارت ضریب بهینه برای ویژگی B برابر با ۵، ویژگی D برابر با ۰.۸، ویژگی G برابر با ۲۹، ویژگی I برابر با ۴ و بقیه ویژگی‌ها برابر با ۱ است. در روش بانظارت ضرایب بهینه‌ی ویژگی‌ها برابر با ۱ منظور شده است. این مقادیر به صورت تجربی (روش آزمون و خطا) به‌دست آمده‌اند.

تعداد کل افعال با مصدر "زدن" در کل پیکره بیجن خان برابر با ۳۷۸۱ عدد است. از این تعداد، تعداد کل افعال ساده برابر با ۳۱۱، تعداد کل افعال مرکب پیوسته برابر با ۲۵۶۲، تعداد کل افعال مرکب گسسته برابر با ۹۰۸ و تعداد افعالی که شامل فعل کمکی هستند برابر با ۱۲۵ عدد می‌باشد. پیکره آزمون مورد استفاده

در روش بانظارت ۱/۳ کل پیکره حاوی افعال "زدن" و مابقی پیکره، پیکره‌ی آموزش است.

نتایج به‌دست آمده برای روش‌های بدون نظارت و با نظارت به تفکیک نوع افعال در جداول جدول) و جدول) و دقت کل در جداول جدول) و جدول) آمده است. اغماض از تمام ویژگی‌ها به عنوان Baseline درنظر گرفته شده است. همچنین یکی از ویژگی‌های مهم روش پیشنهادی (بدون نظارت و بانظارت) برابری مقدار معیارهای دقت، فراخوان و F-score است و بنابراین فقط مقدار دقت در نتایج نشان داده شده است.

جدول ۱- نتایج به تفکیک نوع افعال (روش بدون نظارت)

	افعال ساده (٪)	افعال مرکب پیوسته (٪)	افعال مرکب گسسته (٪)
اغماض از تمام ویژگی‌ها	۳۱.۵۱	۸۳.۱	۷۲.۱۴
اعمال تمام ویژگی‌ها	۹۱.۶۴	۹۹.۲۶	۹۲.۲۹

جدول ۲- نتایج به تفکیک نوع افعال (روش با نظارت)

	افعال ساده (٪)	افعال مرکب پیوسته (٪)	افعال مرکب گسسته(٪)
اغماض از تمام ویژگی‌ها	۷۶.۵۳	۹۴.۶۹	۹۰.۰۹
اعمال تمام ویژگی‌ها	۹۶.۴۶	۹۹.۶۵	۹۷.۱۴

جدول ۳- دقت کل (روش بدون نظارت)

	دقت کل (٪)
اغماض از تمام ویژگی‌ها	۷۶.۲
اعمال تمام ویژگی‌ها	۹۶.۹۶

جدول ۴- دقت کل (روش با نظارت)

	دقت کل (٪)
اغماض از تمام ویژگی‌ها	۷۶.۱
اعمال تمام ویژگی‌ها	۹۶.۲۵

تعداد افعال غلط تشخیص داده شده در روش بدون نظارت برابر با ۱۱۵ عدد می‌باشد. از این تعداد افعال غلط تشخیص داده شده ۴۸ عدد مربوط به ناتوانی تشخیص ریشه‌یاب در تشخیص شکل اصلی کلمات می‌باشد. مقدار خطا در جدول) نشان داده شده است.

جدول ۵- خطای روش پیشنهادی (بدون نظارت)

خطا (٪)	خطای ریشه‌یاب (٪)
۳.۰۴	۱.۲۷

از آنجا که با رفع مشکل ریشه‌یاب، افعال به درستی تشخیص داده می‌شوند، می‌توان خطای نهایی را بدون لحاظ کردن خطای ریشه‌یاب بیان نمود. به علاوه تعداد افعال غلط تشخیص داده شده ناشی از حذف جزء غیرفعلی به قرینه (مانند "حرفی را که باید در دانشگاه زد، در روستا می‌زند") برابر با ۴ است. بنابراین خطای ناشی از آن ۰.۱۶٪ است. به این ترتیب خطای کل روش بدون نظارت برابر با ۱.۶۱٪ است.

تعداد افعال غلط تشخیص داده شده در روش با نظارت برابر با ۴۷ عدد می‌باشد. از این تعداد افعال غلط تشخیص داده شده ۲۰ عدد مربوط به ناتوانی تشخیص ریشه‌یاب در تشخیص شکل اصلی کلمات می‌باشد. مقدار خطا در جدول) نشان داده شده است.

جدول ۶- خطای روش پیشنهادی (با نظارت)

خطا (٪)	خطای ریشه‌یاب (٪)
۳.۷۵	۱.۵۹

با رفع مشکل ریشه‌یاب خطای کل برای روش با نظارت برابر با ۲.۱۵٪ است.

اگر یکی از مداخل موجود در جدول عبارات، در یک جمله بیش از یک بار ظاهر شده باشد و حتی رخدادهای آن از نظر تصریف نیز مانند هم باشند، الگوریتم همه‌ی آنها را درنظر می‌گیرد و برای آنها تفاوت قائل است. مثلاً در جمله‌ی "پس از شنیدن حرف‌های ایشان، حرف‌های خود را دوباره زد." دو کلمه‌ی کاندید "حرف‌های" برای فعل "زد" وجود دارد که با توجه به امتیاز هر کدام از آنها، یکی انتخاب می‌شود. به علاوه در جمله‌ی "پس از شنیدن حرف دوستش، بالاخره حرف زد" نیز دو کلمه‌ی کاندید یکسان وجود دارد (کلمه "حرف"). در این جمله کلمه‌ی کاندید اول با فعل "زد" تشکیل فعل مرکب گسسته "حرف زد" را می‌دهد. همچنین کلمه‌ی کاندید دوم با فعل سبک "زد" تشکیل فعل مرکب پیوسته "حرف زد" را می‌دهد.

بنابراین این دو کلمه‌ی کاندید یکسان، تشکیل افعال مرکب یکسانی را می‌دهند ولی نوع افعال مرکب ایجاد شده توسط آنها، متفاوت است (یکی فعل مرکب گسسته و یکی فعل مرکب پیوسته). الگوریتم پیشنهادی میان این دو نوع فعل مرکب یکسان تمییز قائل است و در محاسبه‌ی دقت نوع افعال نیز مدنظر قرار می‌گیرد. در مثال مذکور اگر کلمه‌ی کاندید اول به عنوان کلمه‌ی کاندید نهایی انتخاب شود (تشکیل فعل مرکب گسسته) با اینکه جزء غیرفعلی درست تشخیص داده می‌شود ولی برنامه آن را غلط درنظر می‌گیرد چون فعل مرکب در این جمله از نوع فعل مرکب پیوسته است. همچنین اگر در جمله‌ای فعل مرکب گسسته باشد و الگوریتم فعل مرکب پیوسته را به عنوان جواب برگرداند، برنامه جواب را غلط درنظر می‌گیرد.

ممکن است داخل جمله، چندین جمله قرار گرفته باشد. به عبارت دیگر جملات تودرتو وجود داشته باشد که این جملات می‌توانند از هر نوعی باشند مثلاً جملات سؤالی یا امری. این الگوریتم تمام افعال را در داخل یک جمله تشخیص می‌دهد که این افعال می‌توانند افعال جملات پرسشی، امری و غیره باشند.

۶- نتیجه‌گیری

اگرچه داده‌های آزمون در اینجا با داده‌های آزمون در کارهای اخیر انجام شده برای زبان فارسی یکسان نیست ولی به طور تقریبی به این نتیجه می‌رسیم که دقت روش پیشنهادی در مقایسه با حدود دقت سایر کارهای مشابه مخصوصاً در حوزه زبان فارسی و مخصوصاً در مقایسه با دو کار اخیر انجام شده [۷]، [۸]، برتری قابل ملاحظه‌ای دارد. زیرا در روش رسولی و همکاران [۸] با درنظر گرفتن فعل "کردن"، بیش‌ترین مقدار F-score برای فعل "کردن" برابر با ۷۸.۷۰٪ و در روش صالحی و همکاران [۷] با درنظر گرفتن افعال "زدن"، "خوردن"، "دادن"، "گذاشتن" و "گرفتن"، مقدار F-Score به طور متوسط با استفاده از رده‌بند 3NN

برای افعال مرکب برابر با ۷۵.۸۸٪ و برای افعال غیر مرکب برابر با ۷۹.۵۸٪ و با استفاده از رده‌بند MLP برای افعال مرکب برابر با ۷۵.۳۶٪ و برای افعال غیر مرکب برابر با ۸۲.۴۷٪ است. همچنین با بررسی و مشاهده‌ی نتایج روش پیشنهادی، مهم‌ترین عامل در بالا بردن دقت، اعمال قواعد استنتاجی است. علاوه بر آن می‌توان مهم‌ترین ویژگی‌های استنتاجی بر اساس بیش‌ترین تاثیر در افزایش دقت را به صورت زیر برشمرد.

۱- معیار PMI تغییریافته
۲- درنظر گرفتن فاصله‌ی بین دو جزء فعل مرکب
۳- درنظر گرفتن تاثیر جداگانه افعال، علائم و حرف "را" مابین دو جزء فعل مرکب

در اولویت‌های بعدی می‌توان سایر ویژگی‌های استنتاجی به همراه حد آستانه‌ی بهینه و ضرایب بهینه برای ویژگی‌ها را در افزایش دقت موثر دانست.

از جمله کارهایی که برای بهبود دقت الگوریتم می‌توان انجام داد، عبارتند از:

۱- بهره گیری از روش‌های تشخیص آرگومان‌های هر فعل در جمله.
۲- مشخص کردن اجزای غیرفعلی ناسازگار با زمان‌های مختلف یک فعل سبک.
۳- استفاده از هرزواژه برای کلماتی که نمی‌توانند مابین دو جزء فعل مرکب واقع شوند.
۴- استفاده از روش‌های پیشرو مانند CRF، Log-Linear و یا Structured Perceptron به جای معادله‌ی امتیازدهی.
۵- اعمال روش پیشنهادی بر روی بیش از یک فعل.

مراجع

[1] M. Dabir-Moghaddam, "Compound verbs in Persian, " *Studies in the Linguistic Sciences*, vol. 27, no. 2, pp. 25-59, 1997.

[2] R. M. K. Sinha, "Mining complex predicates in Hindi using a parallel Hindi-English corpus," *Proc, Int'l Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pp. 40-46, 2009.

[3] D. D. S. P. T. Mondal, T. Chakraborty, and S. Bandyopadhyay, "Automatic Extraction of Complex Predicates in Bengali," *Proc, Int'l Conf. Computational Linguistics*, pp. 37-45, 2010.

[4] E. Zackova, L. Popelinsky, and M. Nepil, "Recognition and tagging of compound verb groups in Czech," *Proc, Int'l Conf. Computational natural language learning*, pp. 219-225, 2000.

[5] R. C. Balabantaray, M. K. Jena, and S. Mohanty, "Shallow morphology based complex predicates extraction in Oriya," *Journal of Computer Applications*, vol. 16, no. 1, pp. 1-5, 2011.

[6] C. Xiao, and D. Rösner, "Detecting multiword verbs in the English sublanguage of MEDLINE abstracts," *Proc, Int'l Conf. Computational Linguistics*, pp. 23-27, 2004.

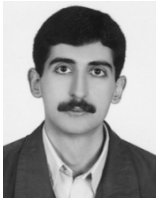
[7] B. Salehi, N. Askarian, and A. Fazly, "Automatic identification of persian light verb constructions," *Proc, Int'l Conf. Computational Linguistics and Intelligent Text Processing*, pp. 201-210, 2012.

[8] M. S. Rasooli, H. Faili, and B. Minaei-bidgoli, " Unsupervised Identification of Persian Compound Verbs," *Proc, Int'l Conf. Advances in Artificial Intelligence*, pp. 394-406, 2011.

[9] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proc, Int'l Conf. Biennial GSCL*, pp. 31-40, 2009.

[10] M. Shamsfard, H. S. Jafari, and M. Ilbeygi, "Step-1: A set of fundamental tools for persian text processing," *Proc, Int'l Conf. Language Resources and Evaluation*, pp. 124-130, 2010.

[11] H. Amiri, H. Hojjat, and F. Oroumchian, "Investigation on a feasible corpus for Persian POS tagging," *Proc, Int'l Conf. CSI Computer*, pp. 66-72, 2007.



آرش چاقری مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه خوارزمی (تربیت معلم) تهران و مدرک کارشناسی‌ارشد گرایش هوش مصنوعی را در سال ۱۳۹۱ از دانشگاه شهید بهشتی تهران اخذ نموده است. پردازش زبان طبیعی یکی از زمینه‌های مورد علاقه ایشان است و پایان نامه کارشناسی‌ارشدشان نیز در همین زمینه به تشخیص افعال در جملات فارسی اختصاص داشته است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

arashchaghari@gmail.com

مهرنوش شمس‌فرد مدرک کارشناسی و کارشناسی‌ارشد خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه صنعتی شریف و مدرک دکتری را در گرایش هوش مصنوعی از دانشگاه صنعتی امیرکبیر اخذ نموده است. وی در حال حاضر عضو هیئت علمی دانشکده مهندسی برق و کامپیوتر دانشگاه شهید بهشتی و سرپرست آزمایشگاه پردازش زبان طبیعی آن دانشکده است. عمده کار ایشان در حوزه تکنولوژی‌های معنایی قرار دارد و در زمینه‌های پردازش زبان‌های طبیعی، مهندسی دانش و هستان‌شناسی، متن کاوی، و وب معنایی فعالیت دارند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

m-shams@sbu.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۹۲/۱/۳۱
تاریخ اصلاح: ۹۲/۵/۲۸
تاریخ قبول شدن: ۹۲/۶/۷
نویسنده مرتبط: دکتر مهرنوش شمس‌فرد، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران.

[1] Pairwise Mutual Information