

استفاده از تجزیه‌گرهای احتمالاتی زبان طبیعی

جهت بهبود ترجمه افعال گروهی

انگلیسی به فارسی

هشام فیلی

دانشکده مهندسی برق و کامپیوتر دانشکده فنی، دانشگاه تهران

چکیده:

یکی از بزرگ‌ترین مشکلات در ترجمه ماشینی زبان انگلیسی به فارسی، ترجمه افعال گروهی^۱ زبان انگلیسی است که به‌وفور در این زبان یافت می‌شود. افعال گروهی از متداول‌ترین عباراتی هستند که از ترکیب یک فعل با یک حرف اضافه یا قید (ادات) تشکیل شده است. تشخیص این که ادات به فعل مرتبط است که در آن صورت فعل گروهی تشکیل می‌دهد یا این که به گروه اسمی ما بعد آن مرتبط است، از جمله فعالیت‌های تا حدودی پیچیده و مبهم در تجزیه نحوی زبان انگلیسی به شمار می‌آید. در این مقاله با استفاده از تجزیه‌گر احتمالاتی زبان انگلیسی در مرحله تجزیه از یک سیستم مترجم ماشینی مبتنی بر قاعده، تشخیص افعال گروهی ابهام‌زدایی می‌شود. هم‌چنین با استفاده از تعدادی قواعد زبان‌شناسی که به‌صورت مکاشفه‌ای به‌دست آمده‌اند، خروجی‌های حاصل از تجزیه‌گر احتمالاتی بررسی شده و در صورت تشخیص ناسازگاری ساختاری، تجزیه نحوی بهبود داده می‌شود. آزمایش‌ها بر روی ۵۲۰ جمله حاوی افعال گروهی، نشان از کیفیت تشخیص افعال گروهی با استفاده از تجزیه‌گر احتمالاتی و قواعد مکاشفه‌ای تا حدود ۸۷٪ است.

واژگان کلیدی: ترجمه ماشینی، زبان فارسی، گرامر اتصال، درختی، تجزیه‌گر احتمالاتی

۱- مقدمه

اصطلاح چندکلمه‌ای^۲ از دیرباز یکی از بزرگ‌ترین مسائل در پردازش زبان طبیعی به‌شمار می‌آمده است؛ که پیچیدگی‌های زیادی در پردازش‌های مختلف زبانی، به‌خصوص مترجم ماشینی ایجاد کرده است. گفته می‌شود که تعداد اصطلاح چندکلمه‌ای در زبان انگلیسی به‌طور تقریبی با تعداد کل کلمات برابری می‌کند (Sag, et al., 2002). به‌عنوان مثال در سیستم WordNet 1.7، حدود ۴۱٪ از ذرایه‌ها را اصطلاح چندکلمه‌ای تشکیل داده‌اند (Fellbaum, 1998). یکی از متداول‌ترین انواع اصطلاحات چندکلمه‌ای، افعال گروهی هستند که به‌خصوص

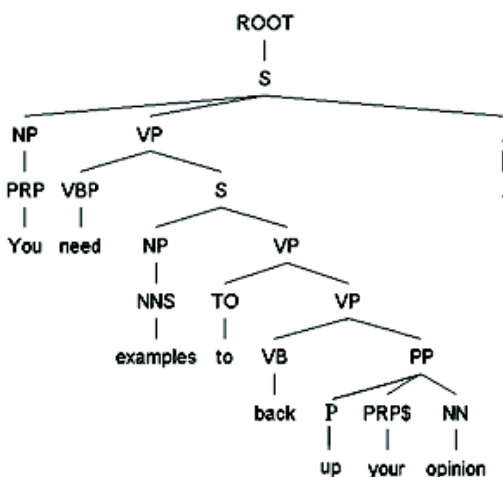
در زبان انگلیسی به‌وفور مورد استفاده قرار می‌گیرد. در زبان انگلیسی یک فعل گروهی از ترکیب یک فعل و یک حرف اضافه یا قید تشکیل شده است. به حرف اضافه (یا قید) که به یک فعل گروهی مرتبط است، ادات^۳ گفته می‌شود. از ترکیب فعل با ادات، افعال جدیدی در زبان ایجاد می‌گردد که معنای آن به هیچ وجه از معنای سازه‌های تشکیل دهنده آن استنتاج نمی‌گردد. از افعال گروهی متداول در زبان انگلیسی می‌توان به افعال "come across"، "come on"، "put on" و ... اشاره کرد.

هرقدر که ترکیب فعل و ادات سبب افزایش قدرت ارائه زبان و توانایی ایجاد مفاهیم جدید می‌شود، تشخیص این وابستگی خود به یکی از بزرگ‌ترین مشکلات در تجزیه‌گرهای

³ Particle

¹ Phrasal verb

² Multi-word Expression



(شکل ۲) تجزیه جمله شماره ۱ که در آن حرف اضافه on تشکیل یک گروه حرف اضافه کرده است.

تشخیص ساختار نحوی درست یک جمله که دارای افعال گروهی باشد، در پردازش‌های مختلف زبانی مانند ترجمه ماشینی مبتنی بر قاعده (Monti, et al., 2011) و آماری (Peng, et al., 2009) کاربرد فراوانی دارد. در صورت تشخیص صحیح افعال گروهی، علاوه بر آن ساختار نحوی جمله به درستی قابل تشخیص خواهد بود، امکان تعیین لغت مقصد نیز به خوبی وجود خواهد داشت.

در این مقاله از روش‌های احتمالاتی جهت تشخیص افعال گروهی به منظور استفاده در یک سیستم مترجم ماشینی مبتنی بر قاعده استفاده شده است. در سیستم مترجم ماشینی که در این مقاله بدان اشاره شده است، مدل ساختاری به خصوصی با نام دستور اتصال-درختی^۲ به عنوان مدل رسمی زبان منظور شده است.

گرامر اتصال-درختی به دلیل ویژگی‌های خاص خود، جهت استفاده در کاربردهای ترجمه ماشینی مناسب هستند (Abeille and Scabes, 1989). به این نوع از گرامرها که در طبقه‌بندی سلسله‌مراتبی چامسکی بین گرامرهای مستقل از متن و حساس به متن قرار می‌گیرد، گرامر حساس به متن ملایم^۳ گفته می‌شود (de la, et al., 1998). درواقع این نوع از گرامرها را می‌توان نسخه گسترش یافته گرامر مستقل از متن در نظر گرفت که قواعد تولید آن، خود درخت‌هایی هستند که ساختار نحوی جملات را نشان می‌دهند.

نحوی تبدیل شده است (Mudraya, 2008). تشخیص این که ادات، به فعل مرتبط است یا این که خود تشکیل یک گروه حرف اضافه مستقل بدهد، یکی از ابهامات موجود در عمل تجزیه نحوی به شمار می‌آید. به عنوان مثال جمله ۱، که در واقع دارای فعل گروهی “back up” است، می‌تواند به دو روش مختلف (با فعل گروهی و بدون فعل گروهی) تجزیه گردد.

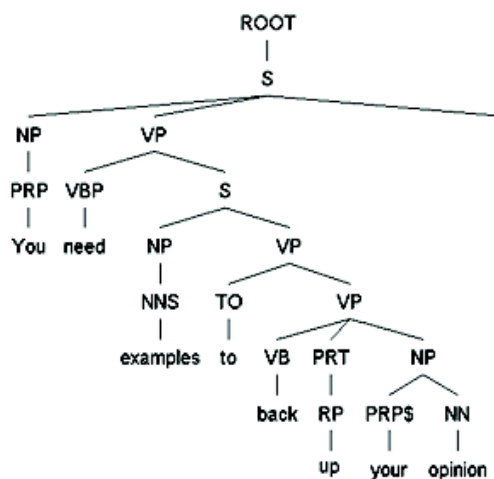
(1) You need examples to back up your opinion.

شکل‌های ۱ و ۲ به ترتیب درخت‌های تجزیه متناظر با هریک از حالت‌های ذکر شده را نشان می‌دهند.

بسته به این که “back up” به عنوان فعل گروهی تشخیص داده شود (شکل ۱) یا اینکه فعل back و حرف اضافه up داشته باشد (شکل ۲)، معنای آن در زبان فارسی تغییر می‌کند. در عمل تشخیص این موضوع به رفع ابهام معنای کلمه^۱ نیز کمک می‌کند. تشخیص این موضوع نیز با معنای جمله و احتمال حضور فعل مذکور با مفعول مربوطه مرتبط است. در زیر معانی این جمله در هر دو حالت ذکر شده است که البته معنی آن در حالت اول، درست است:

حالت ۱: شما برای پشتیبانی از ایده‌ی تان به مثال‌هایی نیاز دارید.

حالت ۲: شما برای پشت رفتن بالای ایده‌ی تان، به مثال‌هایی نیاز دارید.



(شکل ۱) تجزیه جمله شماره ۱ که در آن “back up” به عنوان فعل گروهی تشخیص داده شده است.

² Tree-Adjoining Grammar

³ Mildly context-sensitive grammar

¹ Word-Sense Disambiguation

با هدف استفاده از گرامر اتصال-درختی لغوی در ترجمه ماشینی، مفهوم جدیدی به نام گرامر اتصال-درختی هم‌زمان^۱ (S-TAG) ابداع شده است. اولین استفاده از S-TAG در ترجمه ماشینی به سال ۱۹۸۹ میلادی بر می‌گردد که در آن با استفاده از این مفهوم، روشی جهت ترجمه اصطلاحات ارایه شد (Abeille and Scabes, 1989). پس از آن، آزمایش‌ها و تحقیقات متعددی در این زمینه انجام شده است که اغلب آن‌ها بر اساس پروژه گرامر وسیع زبان انگلیسی که XTAG نام دارد، تعریف شده‌اند (Kallmeyer, 2010). S-TAG در واقع از دو گرامر TAG موازی، که متناظر با زبان‌های مبدأ و مقصد هستند، تشکیل شده است. هر جمله‌ای که در زبان مبدأ توسط گرامر TAG زبان مبدأ تجزیه می‌شود، به ساختار معادل خود در زبان مقصد تبدیل می‌گردد.

در مدل S-TAG به‌ازای هر کلمه در زبان مبدأ و ساختار نحوی که آن کلمه می‌تواند داشته باشد، یک ساختار به همراه معانی لغوی آن در زبان مقصد طراحی شده است. بنابراین اگر یک فعل به همراه ادات، به‌عنوان فعل گروهی توسط تجزیه‌گر تشخیص داده شود، آن‌گاه ابهام‌زدایی معنایی کلمه بسیار ساده‌تر خواهد بود و در زبان مقصد معنای دقیق آن مشخص خواهد شد. بنابراین، این‌که بتوان به‌درستی تشخیص داد که یک فعل به همراه ادات، فعل گروهی است یا سازه مجزا، در عمل در کیفیت ترجمه و ابهام‌زدایی معنایی آن فعل کمک شایانی خواهد کرد.

یکی از مشکلات اصلی استفاده از مدل مذکور، عدم توانایی ابهام‌زدایی در تشخیص افعال گروهی است که این موضوع به دلیل عدم استفاده از مدل‌های احتمالاتی در این ساختار است. بدین منظور از یک مدل احتمالاتی جهت بهبود و ابهام‌زدایی عمل تجزیه استفاده شده است. ساختارهای احتمالاتی متعددی در زمینه مدل‌سازی زبان انگلیسی وجود دارد که مدل مورد استفاده در این پروژه از ساختار گرامر مستقل از متن احتمالاتی لغوی^۲ استفاده شده است. گرامر مستقل از متن احتمالاتی لغوی، یک گرامر مستقل از متن است که در آن هر یک از قواعد دارای یک مقدار احتمالی هستند و هم‌چنین به هر یک از قواعد تعدادی لغت متناظر شده است. به عبارت دیگر، یک گرامر

مستقل از متن احتمالاتی لغوی، گرامری است که علاوه بر تولید جملات زبان، احتمال حضور آن جمله در زبان را نیز محاسبه می‌کند. تجزیه‌گرهای احتمالاتی، سیستم‌هایی هستند که جهت مدل‌سازی ساختاری زبان از یک مدل احتمالاتی (مانند مستقل از متن احتمالاتی لغوی) استفاده می‌کنند.

در این مقاله از تجزیه‌گر احتمالاتی استنفورد (Klein and Manning, 2003) جهت تجزیه جملات انگلیسی و تشخیص افعال گروهی استفاده شده است. این تجزیه‌گر دارای کیفیتی حدود ۸۸٪ دقت/صحت است. سپس یک مدل ترکیبی^۳ از طریق ادغام تجزیه‌گر مذکور با سیستم مترجم ماشینی مبتنی بر قاعده انگلیسی به فارسی (Faili and Ghassem-Sani, 2004)، ارایه می‌گردد. هدف اصلی از این ترکیب بهبود ترجمه افعال گروهی زبان انگلیسی است که به‌عنوان یکی از ضعف‌های مشهود در مدل TAG به‌شمار می‌آید. علاوه بر آن ارایه مدل ترکیبی تجزیه‌گر احتمالاتی و سیستم مترجم ماشینی مبتنی بر قاعده، استفاده دیگری نیز دارد که در پروژه مذکور در ابهام‌زدایی اتصال گروه حرف‌اضافه (PP Attachment)، ابهام‌زدایی اتصال گروه وصلی و ابهام‌زدایی اتصال گروه عطفی استفاده شده است.

در زمینه تشخیص افعال گروهی انگلیسی، یکروش بامری با استفاده از WordNet توسط (Bannard, 2002) ارایه شده است که از فهرست دست‌ساز برای تشخیص معنای حروف اضافه استفاده شده است و نتیجه‌گیری شد که پراکندگی داده‌ها باعث گرفتن نتایج نامناسب شده است. هم‌چنین با استفاده از روش Latent Semantic Analysis (LSA)، یک روش ترکیبی برای تشخیص افعال گروهی ارایه شده است (Baldwin, et al., 2003) که در آن با کمک LSA شباهت معنایی بین فعل گروهی یا ادات آن سنجیده شده است.

در بخش‌های بعدی ابتدا مختصری درمورد مدل ساخت‌یافته مترجم ماشینی و مدل احتمالاتی تجزیه‌گر زبان انگلیسی اشاره می‌گردد. سپس به معماری ترکیبی این دو اشاره می‌گردد. آزمایش‌ها نشان داده است که سیستم تجزیه‌گر احتمالاتی لغوی زبان انگلیسی، با وجود کیفیت حدود ۸۸٪ در تجزیه نحوی جملات، در تشخیص افعال

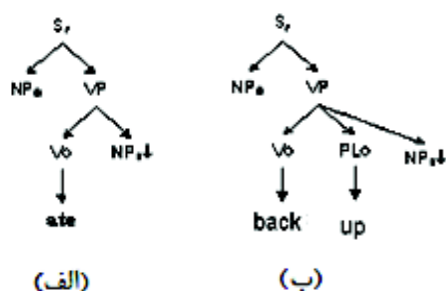
³ Hybrid

¹ Synchronous tree adjoining grammar

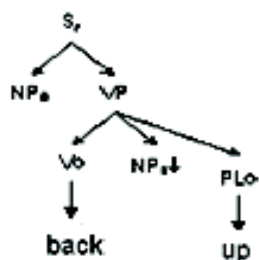
² Lexicalized Probabilistic Context-Free grammar (L-PCFG)

در سیستم مترجم ماشینی ساخت‌یافته به‌شمار می‌آید. آزمایش‌ها بر روی ۵۲۰ جمله حاوی افعال گروهی نشان داده است که تمامی جملات مذکور با هر دو نوع درخت بدوی قابل تجزیه هستند و در واقع تجزیه‌گر مبتنی بر TAG برای تمامی جملات مورد آزمایش هر دو نوع ساختار افعال گروهی و فعل به همراه گروه اسمی یا گروه حرف اضافه را تشخیص داده است.

علاوه بر ساختارهای ذکرشده، امکان ایجاد فاصله بین فعل و ادات آن (حرف اضافه یا قید) نیز وجود دارد. بدین‌صورت که ممکن است مفعول که خود یک گروه اسمی است، مابین این دو گره قرار گیرد. درخت نمایش داده شده در (شکل ۴)، یک نمونه از درخت‌های بدوی با در نظر گرفتن فاصله بین فعل و ادات را نشان می‌دهد. بدیهی است که گروه اسمی NP1 خود می‌تواند شامل چندین کلمه باشد.



(شکل ۳) نمونه‌ای از درخت بدوی یک لنگر برای فعل متعدی (الف) و چندلنگر فعل گروهی (ب)



(شکل ۴) نمونه‌ای از درخت بدوی چندلنگر فعل گروهی که بین فعل و ادات آن، یک سازه‌ی گروه اسمی وجود دارد.

معماری مترجم ساخت‌یافته انگلیسی به فارسی در (شکل ۵) نشان داده شده است. همان‌گونه که در شکل مشخص است، سیستم مترجم ماشینی ساخت‌یافته دارای سه مرحله با نام‌های مرحله تجزیه‌گر، مرحله انتقال و مرحله تحلیل ساخت‌وازی است. خروجی مرحله تجزیه‌گر، درخت

گروهی فقط حدود ۷۰/۳۸٪ کیفیت دارد. برای بهبود این وضع از یک روش مبتنی بر قاعده و پیاده‌سازی دو قاعده مکاشفه‌ای استفاده شده است که سبب افزایش کیفیت تشخیص افعال گروهی تا ۸۶/۹۲٪ شده است.

۲- ترجمه انگلیسی به فارسی مبتنی بر قاعده

همان‌طوری که در قبل ذکر شد، در سیستم مترجم ماشینی مبتنی بر قاعده از مدل گرامری اتصال-درختی استفاده شده است. گرامر اتصال-درختی از یک ۵-تایی مرتب (V_N, V_T, S, I, A) تشکیل شده است که در آن V_N مجموعه متناهی از غیرپایانه‌ها، V_T مجموعه متناهی از پایانه‌ها، S علامت شروع گرامر، I مجموعه متناهی از درخت‌های بدوی^۱ و A مجموعه متناهی از درخت‌های کمکی^۲ است.

جهت افزایش دقت و قدرت پردازشی TAG مفهوم لغوی نیز به آن اضافه می‌گردد. بدین صورت که هر درخت اولیه دست‌کم دارای یک گره است که به‌طور مستقیم به یک واژه متصل است و به آن گره لنگر^۳ گفته می‌شود. به این مدل از گرامر، اتصال-درختی لغوی (LTAG) اطلاق می‌گردد.

جهت مدیریت اصطلاحات چندکلمه‌ای، مانند افعال گروهی در گرامرهای اتصال-درختی لغوی درخت‌های بدوی چندلنگر تعریف شده است که در آن هر درخت بدوی به بیش از یک لغت مرتبط می‌گردد. در حالت افعال گروهی درختان متعددی تعریف شده‌اند که در آن گره‌های لنگر شامل دو گره فعل و حرف اضافه (یا قید) می‌باشد. (شکل ۳)، دو نمونه درخت بدوی در گرامر LTAG را نشان می‌دهد که قسمت الف نشان‌دهنده یک فعل متعدی مانند فعل "ate" می‌باشد. در قسمت ب یک فعل گروهی نشان داده می‌شود که دارای دو بخش فعل و حرف اضافه می‌باشد که به‌طور نمونه فعل گروهی "back up" نشان داده شده است. به‌طور تقریبی کلیه افعال گروهی با هر دو درخت مذکور قابل تجزیه هستند و درعمل در تجزیه تمامی جملاتی که دارای افعال گروهی هستند، ابهام ایجاد می‌گردد. تعیین نوع درختی که در تجزیه قابل استفاده است، از مشکلات موجود

¹ Initial trees

² Auxiliary trees

³ Anchor node

شده نیز می‌تواند بسیار زیاد باشد. برای این منظور از روش‌های مختلفی جهت فیلتر کردن تعداد درختان بدوی ممکن برای هر کلمه استفاده می‌شود (Faili and Ghassem-Sani, 2004). ایده اصلی این مقاله ارزیابی تعدادی فیلتر با استفاده از سیستم تجزیه‌گر احتمالاتی با مدل PCFG جهت کاهش تعداد حالت‌های تجزیه برای استفاده در یک سیستم مترجم ماشینی ساخت‌یافته است. درواقع در صورت استفاده از این فیلتر درعمل ابهامات موجود در فعل گروهی حذف می‌گردد.

۳- سیستم تجزیه‌گر احتمالاتی

همان‌طوری که در قبل مطرح گردید، جهت ابهام‌زدایی و انتخاب یک تجزیه واحد از بین کلیه تجزیه‌های ممکن، از یک سیستم تجزیه‌گر احتمالاتی^۳ زبان انگلیسی استفاده شده است. درواقع از این سیستم به‌عنوان یک فیلتر جهت کاهش تعداد درختان بدوی کلمات استفاده می‌گردد. این عمل علاوه بر آن که هزینه زمانی تجزیه‌گر مبتنی بر نمودار را کاهش می‌دهد، سبب کاهش ابهامات موجود در تجزیه و حذف ابهامات در حالت افعال گروهی می‌گردد.

در حال حاضر، بهترین سیستم‌های تجزیه‌گر احتمالاتی بر مبنای توسعه‌یافته از مدل گرامر مستقل از متن احتمالاتی (PCFG)، تولید شده‌اند (Cahill, 2008). این توسعه در راستای حذف فرضیه استقلال قواعد (مانند مدل-۱ ارایه شده توسط Collins (1997) و مدل ارایه شده توسط Bod, 2006)، یا در راستای لغوی شدن گرامر (مانند مدل-۲ ارایه شده توسط Collins (1997) و تجزیه‌گر استنفورد)، یا بر مبنای مدل آموزشی خودکار (مانند مدل ارایه شده توسط McClosky, et al., 2006) است.

درخت تجزیه درختی است که ساختار نحوی یک رشته را بر اساس یک سری گرامر رسمی (احتمالاتی) ارایه می‌دهد. یک تجزیه‌گر زبان طبیعی برنامه‌ای است که ساختار گرامری جملات را تولید می‌کند. برای مثال تشخیص می‌دهد چه کلماتی با هم یک سازه را تشکیل می‌دهند و چه کلماتی فاعل یا مفعول یک فعل هستند. تجزیه‌گر احتمالاتی از دانش حاصل از تجزیه دستی جملات استفاده می‌کند و

اشتقاق^۱ می‌باشد که در واقع یک نوع ساختار وابستگی به همراه اطلاعات نوع وابستگی و نوع هر گره درخت است (XTAG research group, 2003). درواقع با داشتن این ساختار نه تنها درخت تجزیه به راحتی به دست می‌آید، بلکه ساختار اشتقاق جملات در زبان مقصد نیز قابل دستیابی است.^۲ مرحله انتقال نیز خود شامل سه بخش انتقال درختی (ساختار نحوی)، انتقال لغوی و انتقال ویژگی است. جهت اطلاعات بیشتر به مرجع (Faili and Ghassem-Sani, 2004) مراجعه گردد.



(شکل ۵) مراحل کامل ترجمه بر مبنای گرامر اتصال-درختی هم-زمان (Faili, H., Ghassem-Sani, G. 2004)

از بزرگ‌ترین مشکلات موجود در استفاده از مدل TAG در مترجم ماشینی ساخت‌یافته، عدم امکان ابهام‌زدایی در مراحل مختلف به‌خصوص مرحله تجزیه زبان مبدأ می‌باشد. به‌طور مثال در جمله شماره ۱ این مقاله، درعمل هر دو درخت تجزیه شکل‌های ۱ و ۲ به‌عنوان خروجی تجزیه‌گر ارایه می‌شود. جهت ابهام‌زدایی و انتخاب یک تجزیه از بین نامزدهای موجود از روش‌های احتمالاتی استفاده می‌شود که در بخش بعدی توضیح داده می‌شود.

در صورت استفاده از تجزیه‌گر پایه بر مبنای مدل TAG، تمامی تجزیه‌های ممکن برای آن جمله ارایه می‌گردد (Kallmeyer, 2010). با استفاده از روش‌های برنامه‌نویسی پویا، الگوریتم تجزیه مبتنی بر نمودار برای مدل گرامری TAG ارایه شده است که یک جمله با طول n کلمه را با هزینه $O(C^3 n^6)$ تجزیه می‌کند که در آن C متوسط تعداد درختان بدوی برای هر کلمه در جمله است. علاوه بر مشکل زمان تجزیه این الگوریتم، تعداد تجزیه‌های ایجاد

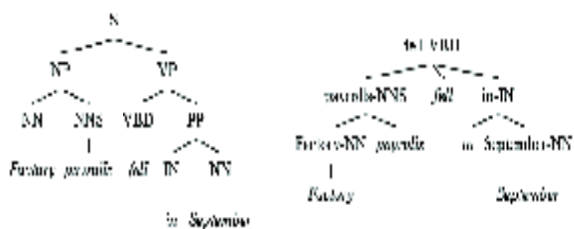
^۱ Derivation Tree

^۲ طبق مرجع (Kallmeyer, L. 2010) ساختار اشتقاق جملات هم معنی در زبان‌های مختلف از لحاظ شکل توپولوژی یکسان است و صرفاً برچسب‌های گره‌های این ساختار در این زبان‌ها متفاوت است.

^۳ Probabilistic parser

جمله به دست می‌آید و مقدار $P(D)$ از حاصل ضرب حضور هر یک از کلمات جمله محاسبه می‌گردد. به عنوان مثال، در (شکل ۶) درخت تجزیه و وابستگی جمله شماره ۲ نشان داده شده است.

(2) Factory payrolls fell in September



(شکل ۶) درخت تجزیه و وابستگی جمله شماره ۲

مقدار $P(T)$ بصورت زیر محاسبه می‌گردد:

$$P(T) = P(S \rightarrow NP \quad VP), \quad P(NP \rightarrow NN \quad NNS), \\ P(VP \rightarrow VBD \quad PP), \quad P(PP \rightarrow IN \quad NN)$$

درحالی که مقدار $P(D)$ از رابطه زیر به دست می‌آید:

$$P(D) = P(\text{fell} \mid \text{payrolls, left}) \times \\ P(\text{Factory} \mid \text{payrolls, right}) \times \\ P(\text{START_SYMBOL} \mid \text{Factory, left}) \times \\ P(\text{in} \mid \text{fell, right}) \times \\ P(\text{September} \mid \text{in, left})$$

که در آن $P(w_i \mid w_j, \text{dir})$ احتمال ظهور کلمه w_i در جهت dir (چپ یا راست) کلمه w_j است.

احتمال آن که یک فعل گروهی توسط تجزیه‌گر احتمالاتی تشخیص داده شود، به میزان وقوع وابستگی حرف اضافه (یا قید) به فعل و احتمال ایجاد ساختار نحوی فعل گروهی در جملات دادگان آزمون بستگی دارد. بنابراین در صورتی که یک فعل گروهی به کرات ظاهر شود، احتمال آن که در جملات مشابه نیز به درستی توسط سیستم تشخیص داده شود، بالا می‌رود. البته این که فعل و حرف اضافه تشکیل یک فعل گروهی دهند، ممکن است بدین صورت باشد که بین این دو کلمه، فاصله زیادی وجود داشته باشد و بین آن‌ها یک سازه گروه اسمی وجود داشته

مشابه‌ترین تحلیل را برای جملات جدید به دست می‌آورد. درواقع تجزیه‌گر احتمالاتی، یک سیستم مبتنی بر یادگیری است که در آن با استفاده از تعداد زیادی تجزیه که به صورت دستی یا نیمه دستی تهیه شده‌اند و به آن درخت-بانک^۱ اطلاق می‌گردد (Mitchell, et al., 1993)، آموزش یافته و امکان تجزیه جملات جدید را خواهد داشت. در این مقاله از تجزیه‌گر احتمالاتی استنفورد استفاده شده است که در آن میزان کیفیت تجزیه‌های ایجاد شده برابر با ۸۶/۳۶٪ است (Klein and Manning, 2003).

این سیستم در واقع ترکیبی از یک تجزیه‌گر مبتنی بر گرامر مستقل از متن احتمالاتی و سیستم تجزیه‌گر مبتنی بر وابستگی^۲ است. هر یک از این دو تجزیه‌گر معیاری جهت یک تجزیه مناسب را ارائه می‌دهند که با استفاده از الگوریتمی که از هر دو معیار استفاده می‌کند و در فضای حالت‌های همه تجزیه‌های ممکن از الگوریتم A^* استفاده می‌کند، بهترین خروجی را ارائه می‌دهد. علاوه بر زبان انگلیسی، برای زبان‌های آلمانی، چینی و عربی نیز این سیستم پیاده سازی شده است.

در سیستم تجزیه‌گر مبتنی بر گرامر مستقل از متن احتمالاتی بامربی، احتمال حضور هر سازه^۳ و زیرسازه، محاسبه می‌شود که این موضوع از طریق شمارش کلیه سازه‌های موجود در درخت - بانک قابل محاسبه است^۴. در این مدل، از مقوله معنایی کلمات هیچ اثری دیده نمی‌شود. در حالی که در سیستم تجزیه‌گر مبتنی بر وابستگی، احتمال حضور کلمات مرتبط با یکدیگر محاسبه می‌گردد. درواقع مدل مورد استفاده در این تجزیه‌گر به صورت دوتایی $L = (T, D)$ است که در آن T مدل گرامر مستقل از متن احتمالاتی و D مدل وابستگی مورد نظر است. جهت ساده‌سازی مدل مذکور، استقلال احتمال وقوع هر دو مدل در یک جمله فرض می‌گردد که در آن $P(T, D) = P(T).P(D)$ خواهد شد. $P(T)$ احتمال آن است که ساختار نحوی مذکور ظاهر گردد درحالی که $P(D)$ احتمال حضور کلمات آن جمله در کنار یکدیگر است. مقدار $P(T)$ از حاصل ضرب احتمال هر یک از قواعد مورد استفاده در تجزیه

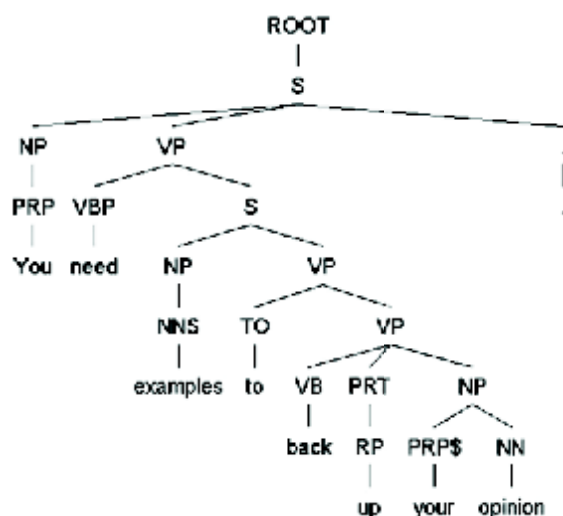
^۱ TreeBank

^۲ Dependency parser

^۳ Constituent

^۴ البته در تجزیه‌گرهای بی‌مربی مانند DOP (Bod, R. 2006) صرفاً از خود جملات یا برجسب کلام آن‌ها بدون تجزیه درختی استفاده می‌شود.

(3) You need examples to back up your opinion.



(شکل ۷) درخت تجزیه جمله ۳

در صورتی که حرف اضافه بخش دوم فعل گروهی تشخیص داده نشود، با یک اسم ترکیب می‌شود. در این صورت اولین ترکیبی که در آن شرکت می‌کند، ترکیب فعلی نمی‌باشد. برای مثال در درخت تجزیه جمله ۴ که در (شکل ۸) نشان داده شده است، حرف اضافه‌ی “on” با فعل جمله ترکیب نشده است. در نتیجه فعل جمله، به عنوان فعل گروهی تشخیص داده نشده است. در این حالت معنای جمله در زیر آن مشخص شده است.

(4) Professor Tanzer called on Tim to answer the question.

ترجمه فارسی: پروفسور Tanzer، Tim را فراخواند تا پاسخ سوالات را بدهد.

در صورت تشخیص گروه فعلی با استفاده از تجزیه‌گر احتمالاتی، می‌توان از این شناخت، در مرحله تجزیه‌گر مبتنی بر مدل TAG استفاده کرده و درخت بدوی چندلنگر برای فعل گروهی در نظر گرفت و در عمل با استفاده از این تجزیه‌گر امکان ابهام‌زدایی برای تجزیه‌گر گرامر مبتنی بر TAG ایجاد کرد. (شکل ۹)، الگوریتم کلی مرحله تجزیه در سیستم مترجم ماشینی را نشان می‌دهد. در این مرحله با استفاده از مدل TAG و فیلتر تجزیه‌گر احتمالاتی، امکان ایجاد ساختار اشتقاق فراهم می‌آید.

باشد. در این حالت وابستگی دور بین فعل و ادات وجود خواهد داشت.

آزمایش‌ها نشان داده است که تجزیه‌گر مذکور دارای مشکلاتی می‌باشد که به خصوص در کاربرد مترجم ماشینی در موضوع این مقاله حساسیت زیادی دارد. به عنوان مثال، تجزیه‌گر ممکن است که مقوله نحوی کلمه را در جمله درست تشخیص ندهد و درخت حاصل ساختار ناصحیح داشته باشد. این مشکل در خیلی از جملاتی که شامل افعال گروهی هستند به چشم می‌خورد.

تشخیص افعال گروهی یک زیرمجموعه از تشخیص اصطلاحات چندکلمه‌ای به شمار می‌آید که در این مجموعه روش‌های مختلف مبتنی بر دادگان و روش‌های قاعده‌مند وجود دارد (Seretan, 2011). ولی آن‌چه که در کار جاری مورد توجه است، بهبود میزان تشخیص افعال گروهی توسط یک تجزیه‌گر است که در مرحله تجزیه یک سیستم مترجم ماشینی اجرا شده است.

۴- بهبود تجزیه افعال گروهی با استفاده از تجزیه‌گر احتمالاتی

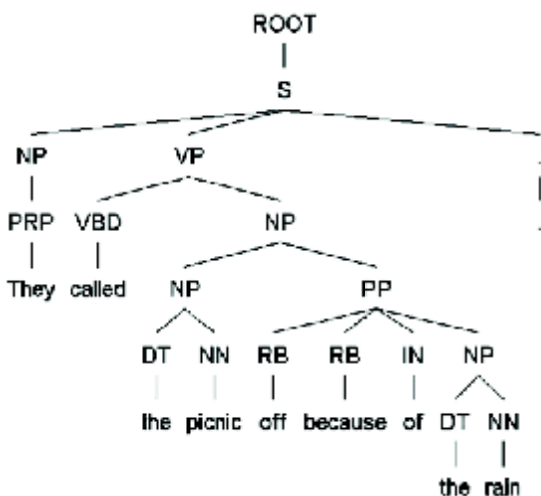
جهت بهبود کیفیت تشخیص افعال گروهی توسط تجزیه‌گر استنفورد، ابتدا باید مشخص شود که آیا تجزیه‌گر مذکور، جمله را دارای فعل گروهی تشخیص داده است یا نه؟ برای این منظور باید کلماتی که می‌توانند بخش دوم فعل گروهی باشند، بررسی شوند. این کلمات حروف اضافه یا قید می‌باشند. درخت تجزیه جمله‌ای که حاوی فعل گروهی تشخیص داده شده است به این صورت است که گروه فعلی، حداقل دو فرزند داشته باشد که یکی از آن‌ها فعل جمله و دیگری حرف اضافه یا قیدی باشد که در ترکیب دیگری شرکت نکرده است. برای این منظور حروف اضافه و قیده‌های جمله بررسی شده، در صورتی که یکی از آن‌ها سازه خواهر^۱ یک فعل باشد، آن جمله حاوی فعل گروهی تشخیص داده شده است. برای مثال در جمله ۳ که درخت تجزیه آن در (شکل ۷) نشان داده شده است، حرف اضافه “up” خواهر فعل “back” است و در اولین ترکیبی که شرکت کرده است، گروه فعلی است. در نتیجه جمله دارای فعل گروهی است.

^۱ Sibling

۵- مدل ترکیبی جهت بهبود تشخیص افعال گروهی

با توجه به آن که تجزیه‌گر احتمالاتی با استفاده از احتمال وابستگی و احتمال ساختار گرامر مستقل از متن احتمالاتی، جمله را تجزیه می‌کند، در مواقعی که فعل گروهی در دادگان آزمون تاحدودی کمیاب باشد، یا این که حرف اضافه (یا قید) همراه فعل گروهی در ساختارهای دیگر مانند گروه حرف اضافه به کرات یافت شود، احتمال آن که تجزیه‌گر آن را فعل گروهی تشخیص دهد، کم خواهد شد. به عنوان مثال جمله شماره ۵ توسط تجزیه‌گر احتمالاتی بدون فعل گروهی تشخیص داده شده است که در (شکل ۱۰)، تجزیه آن نشان داده شده است. همان‌طور که در شکل مشاهده می‌شود حرف اضافه "off" به اشتباه به عنوان بخشی از یک ترکیب اضافی^۱ در نظر گرفته شده است. البته این که بین فعل و حرف اضافه یک سازه اسمی وجود دارد و وابستگی دور در این جمله اتفاق افتاده است، در عدم تشخیص درست مقوله نحوی تأثیرگذار است.

(5) They called the picnic off because of the rain.

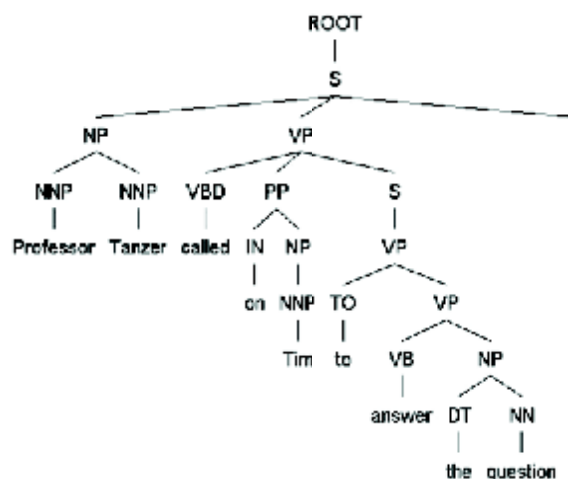


(شکل ۱۰) تجزیه نادرست جمله ۵ توسط تجزیه‌گر استنفورد

با بررسی تجزیه ایجاد شده، مشاهده می‌شود که برای کلیه کلمات "off because of the rain" یک گروه حرف اضافه در نظر گرفته است در حالی که چنین ساختاری برای گروه حرف اضافه در زبان انگلیسی وجود ندارد. یعنی با وجود آن که احتمال:

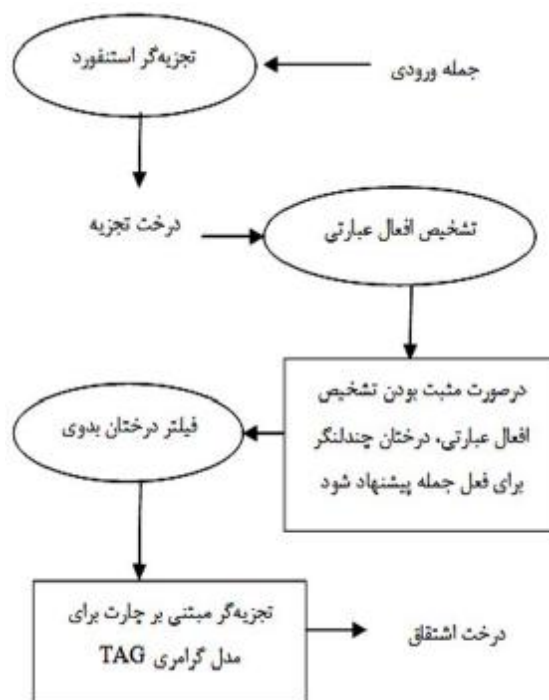
$$P(T) = P(PP \rightarrow RB RB IN NP)$$

^۱ Prepositional phrase



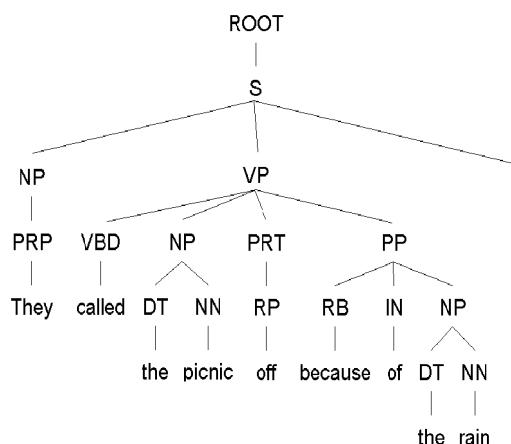
(شکل ۸) درخت تجزیه جمله ۴

پس از ترکیب تجزیه‌گر احتمالاتی با تجزیه‌گر قاعده‌مند مبتنی بر مدل TAG آزمایش‌ها بر روی ۵۲۰ جمله که همگی دارای افعال گروهی هستند دوباره اجرا گردید. این جملات نمونه‌هایی هستند که از مرجع شده‌اند. از بین ۵۲۰ جمله‌ای که دارای فعل گروهی هستند، تجزیه‌گر استنفورد توانست ۳۶۶ جمله را به درستی تشخیص دهد و در عمل درخت تجزیه چندلنگر را برای آن تشخیص دهد. میزان دقت ابهام‌زدایی در این مرحله به ۷۰/۳۸٪ رسیده است.



(شکل ۹) الگوریتم تجزیه مبتنی بر مدل TAG

فعل جمله، به‌عنوان فعل گروهی تشخیص داده می‌شود. در نتیجه حرف اضافه مذکور به‌عنوان ادات فعل و ترکیب اسمی “the picnic” مفعول جمله در نظر گرفته می‌شود. درخت تجزیه تصحیح شده برای این جمله در (شکل ۱۱) نشان داده شده است:



(شکل ۱۱) تجزیه صحیح جمله ۵

۵-۲- قاعده مکاشفه‌ای ۲

در صورتی که تجزیه‌گر استنفورد و قاعده مکاشفه‌ای، ۱ فعل جمله را به‌عنوان فعل گروهی تشخیص ندهند، وجود مفعول در جمله، بررسی می‌گردد^۱ و اگر مفعولی در جمله وجود نداشت، متعدی بودن ترکیب فعل جمله با تک‌تک حروف اضافه موجود در جمله را به کمک پایگاه داده افعال گروهی بررسی می‌شود. اگر یکی از این ترکیب‌ها متعدی باشد، می‌توان نتیجه گرفت که جمله اشتباه تجزیه شده است. تجزیه درست آن به این صورت است که فعل جمله به‌عنوان فعل گروهی است و حرف اضافه‌ای که ترکیبش با فعل متعدی بود، بخش دوم فعل گروهی است و اسم بعد از آن حرف اضافه مفعول جمله است. در (شکل ۱۲) تجزیه جمله ۶ توسط تجزیه‌گر استنفورد نشان داده شده است:

(6) Democracy brought about great change in the lives of the people.

در درخت تجزیه جمله ۶ مشاهده می‌شود که کلمه “about” به‌عنوان حرف اضافه یک گروه حرف اضافه تشخیص داده شده است، در حالی که بخش ادات فعل گروهی جمله است.

^۱ وجود مفعول در جمله به‌وسیله تجزیه وابستگی که توسط تجزیه‌گر استنفورد خروجی می‌شود، می‌توان تشخیص داد. در واقع با استفاده از مسند *obj(verb, noun)* در گرامر وابستگی می‌توان متوجه شد که کدام کلمه مفعول فعل اصلی جمله است.

بسیار کم است، ولی به‌دلیل این که احتمال وابستگی $P(D) = P(\text{off} | \text{called, right})$ بسیار کمتر است، ساختار فوق به اشتباه، خروجی داده شده است و با وجود آن که ساختار نحوی تشخیص داده شده، بسیار نادر است، ولی به‌دلیل آن که احتمال وابستگی ساختارهای دیگر کم‌تر است، چنین ساختاری به‌طور نادرست خروجی داده شده است.

در واقع چنین مشکلی به‌دلیل کمبود دادگان جهت آموزش وابستگی‌های واژگان به‌خصوص در وابستگی‌های مربوط به حرف اضافه (یا قید) با فعل مربوطه در افعال گروهی ایجاد شده است. هم‌چنین وابستگی دور موجود مابین ادات و فعل سبب می‌شود که مشکل دوچندان شود. البته در صورتی که دادگان آموزشی به اندازه کافی برای این حالت‌ها (به‌خصوص در حالت وابستگی دور) وجود داشته باشد، نتایج مناسب‌تر خواهند شد.

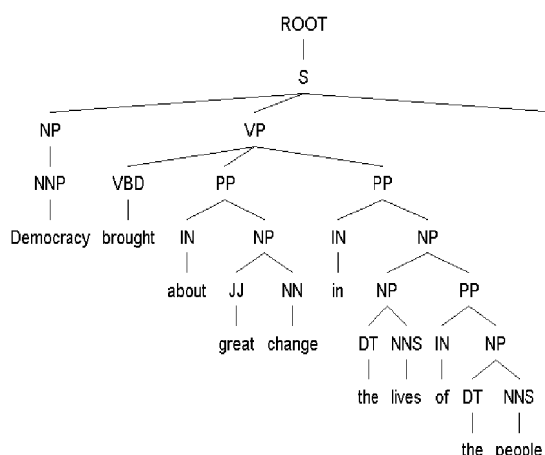
ایده اصلی مطرح شده در این پروژه، استفاده از یک سیستم مبتنی بر قاعده جهت تشخیص ناسازگاری‌های موجود در ساختارهای نحوی حاصل از تجزیه‌گر استنفورد می‌باشد. در واقع با استفاده از یک سیستم مبتنی بر قاعده، خروجی‌های تجزیه احتمالاتی بررسی مجدد شده و در صورت تشخیص ناسازگاری، مشکلات ساختاری آن رفع می‌گردد. در سیستم موجود تاکنون دو قاعده برای تشخیص ناسازگاری به‌صورت زیر ارایه شده است:

۵-۱- قاعده مکاشفه‌ای ۱

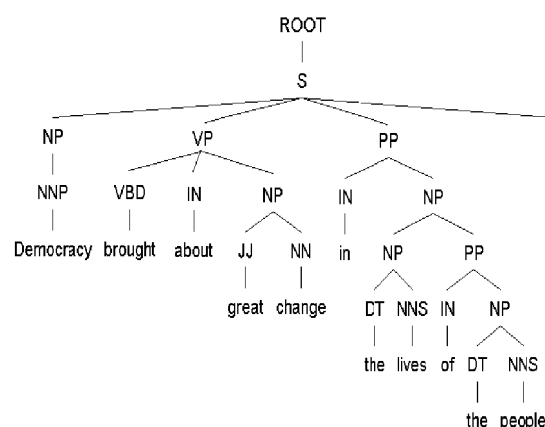
در قاعده مکاشفه‌ای ۱، با توجه به مقوله نحوی کلمه‌ای که بعد از حرف اضافه در جمله می‌آید، گروهی بودن فعل بعضی از جملات تشخیص داده می‌شود. حرف اضافه در جمله، دو نقش می‌تواند داشته باشد یا مربوط به کلمه بعد از آن و یا بخش دوم فعل گروهی باشد (Megerdooian, 2004). در صورتی که حرف اضافه، مربوط به کلمه بعدش باشد، نوع کلمه بعد باید اسم یا گروه اسمی باشد. در نتیجه اگر این کلمه، اسم نباشد به‌حتم بخش دوم فعل است و فعل ما گروهی است. در این تابع مکاشفه‌ای حرف اضافه‌هایی که به‌طور معمول بخش دوم فعل گروهی نیستند، بررسی نمی‌شوند.

با اعمال تابع مکاشفه‌ای ۱ بر روی مثال (شکل ۱۰)، با مشاهده حرف اضافه “off” نوع کلمه بعد از آن (because) بررسی می‌شود و چون اسم یا گروه اسمی نیست،

با توجه به وابستگی‌هایی که از تجزیه‌گر استنفورد به‌دست می‌آید، مشاهده می‌شود در این جمله مفعولی وجود ندارد. بنابراین متعدی بودن ترکیب فعل “brought” با حرف اضافه بعد از آن بررسی می‌شود. مشاهده می‌شود فعل “brought about” یک فعل متعدی است؛ ولی در جمله مذکور مفعولی وجود ندارد. در نتیجه تجزیه جمله نادرست بوده است و فعل گروهی “brought about” فعل جمله و ترکیب اسمی “great change” مفعول جمله است. (شکل ۱۳) درخت تجزیه صحیح را نشان می‌دهد.



(شکل ۱۲): تجزیه نادرست جمله ۶ توسط تجزیه‌گر استنفورد



(شکل ۱۳): تجزیه صحیح جمله ۶

۶- ارزیابی

با توجه به افزودن یک سیستم مبتنی بر قاعده جهت تشخیص ناسازگاری‌های تجزیه‌های حاصل از تجزیه‌گر احتمالاتی با قواعد زبان‌شناسی، آزمایش‌های قبلی دوباره اجرا شده و نتایج کلی در (جدول ۱) ذکر شده است. این آزمایش‌ها بر روی ۵۲۰ جمله که همگی دارای افعال

گروهی هستند، انجام شده است. این جملات از مرجع (List of English Phrasal Verbs, 2010) که شامل افعال گروهی انگلیسی است، استخراج شده است. همان‌گونه که در جدول مذکور نشان داده شده است، اعمال این قواعد، سبب افزایش کیفیت تشخیص افعال گروهی از ۷۰/۳۸٪ به ۸۶/۹۲٪ شده است و در مجموع حدود ۱۶/۵۴٪ بهبود یافته است.

همان‌گونه که در جدول نشان داده شده است، با افزایش طول جمله میزان تشخیص افعال گروهی کاهش می‌یابد ولی در هر حالت بهبود کیفیت حاصل از استفاده از قواعد مکاشفه‌ای قابل توجه است.

هم‌چنین برای بررسی مانعیت این قواعد، تعداد ۵۰۰۰ جمله بدون فعل گروهی به‌صورت تصادفی طوری انتخاب شده است که حروف اضافه یا قیدهایی وجود داشته باشد که می‌توانست به اشتباه با فعل تشکیل یک فعل گروهی دهد. آزمایش‌های فوق دوباره بر روی این جملات اجرا شده و میزان تشخیص افعال گروهی بررسی شده است. از بین ۵۰۰۰ جمله مذکور، تنها دو جمله به‌صورت نادرست، حاوی فعل گروهی تشخیص داده شده است، یعنی حدود ۰/۰۴٪ از جملات، نادرست تشخیص داده شده است که این موضوع دقت بالای قواعد تعریف شده را نشان می‌دهد.

(جدول ۱) مقایسه کیفیت تجزیه‌گر استنفورد و توابع مکاشفه‌ای

طول جمله	تعداد جملات	درصد تشخیص تجزیه‌گر استنفورد	درصد تشخیص توابع مکاشفه‌ای	درصد بهبود
کمتر از ۱۰ کلمه	۲۷۰	۶۹/۲۶٪	۸۷/۷۸٪	۱۸/۵۲٪
بین ۱۰ و ۲۰ کلمه	۲۴۰	۷۲/۰۸٪	۸۵/۸۳٪	۱۳/۷۵٪
بیشتر از ۲۰ کلمه	۱۰	۶۰٪	۸۰٪	۲۰٪
مجموع	۵۲۰	۷۰/۳۸٪	۸۶/۹۲٪	۱۶/۵۴٪

۷- تشکر و قدردانی

از آقای فرشاد کوتی و خانم معصومه بختیاری ضیابری که طی دوره کارآموزی در آزمایشگاه پردازش زبان طبیعی دانشگاه تهران، در اجرای این کار کمک شایانی کرده‌اند، تشکر و قدردانی می‌گردد.

- List of English Phrasal Verbs, 2010. <http://www.learn-english-today.com/phrasal-verbs/phrasal-verb-list.htm>, August.
- Megerdooian, K., 2004. A Semantic Template for Light Verb Constructions, In Proceeding of First Workshop on Persian Language and Computers, Tehran University, Iran. May 25-26.
- Mudraya, O., 2008. Automatic Extraction of Translation Equivalents of Phrasal and Light Verbs in English and Russian, In Phraseology: an interdisciplinary perspective, Benjamins, Amsterdam, pp. 293-309.
- Monti, J., *et al.*, 2011. Taking on new challenges in multi-word unit processing for Machine Translation, In Proceeding of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation.
- McClosky, D., Charniak, E., Johnson, M., 2006. Effective self-training for parsing, In Proceeding of the Human Language Technology Conference of the NAACL, Main Conference, pages 152-159, New York City, USA. Association for Computational Linguistics.
- Mitchell, P.M., Marcinkiewicz, M.A., Santorini, B., 1993. Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics - Special issue on using large corpora: II archive Volume 19 Issue 2, June 1993 MIT Press Cambridge, MA, USA.
- Peng, X., *et al.*, 2009. Using a dependency parser to improve SMT for subject-object-verb languages, In Proceeding of NAACL/HLT, pages 245-253, June.
- Sag, I.A., *et al.*, 2002. Multiword Expressions: A Pain in the Neck for NLP, In Proceeding of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City.
- Seretan, V., 2011. Syntax based Collocation Extraction, Springer.
- XTAG research group., 2003. A Lexicalized Tree Adjoining Grammar for English, Technical Report IRCS 01-03, Institute for Research in Cognitive Science, University of Pennsylvania.
- Abeille, A., Scabes, Y., 1989. Parsing idioms in tree adjoining grammars, In Proceeding of the Fourth conference of the European chapter of the association for computational linguistics, Manchester, England.
- Baldwin, T., *et al.*, 2003. An empirical model of multiword expression decomposability, In Proceeding of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment- Volume 18, pages 89--96. Association for Computational Linguistics.
- Bannard, C., 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions, Master's thesis, University of Edinburgh, Edinburgh, UK.
- Cahill, A., 2008. Treebank-Based Probabilistic Phrase Structure Parsing, Language and Linguistics Compass 2/1, 18-40.
- Bod, R., 2006. An All-Subtrees Approach to Unsupervised Parsing, In Proceeding of ACL.
- Collins, M., 1997. Three generative, lexicalized models for statistical parsing, In Proceeding of ACL.
- de la Clergerie, E., Alonso Pardo, M.A., Cabrero Souto, D., 1998. A Tabular Interpretation of Bottom-up Automata for TAG, In Proceeding of TAG+4, Fourth International Workshop on Tree-Adjoining Grammars and Related Frameworks, pp. 42-45, Philadelphia, PA, USA.
- Faili, H., Ghassem-Sani, G., 2004. An Application of Lexicalized Grammar in English-Persian Translation, In Proceeding of 16th European conference on Artificial Intelligence (ECAI 2004), Universidad Politecnica de Valencia, Valencia, Spain, pp. 596-600, 24-27.
- Fellbaum, C., 1998. WordNet: An Electronic Lexical Database, MIT Press.
- Kallmeyer, L., 2010. Parsing Beyond Context-Free Grammars, volume 0 of Cognitive Technologies. Springer.
- Klein, D., Manning, C., 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing, In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.



هشام فیلی تحصیلات خود را در مقطع

کارشناسی نرم افزار در دانشکده مهندسی

کامپیوتر، دانشگاه صنعتی شریف با رتبه

یک، در سال ۱۳۷۶ به پایان رساند. سپس

مقاطع کارشناسی ارشد نرم افزار و دکتری

هوش مصنوعی را به ترتیب در سال های ۱۳۷۸ و ۱۳۸۵ در

همان دانشکده تکمیل کرد. از سال ۱۳۸۷ تا کنون عضو

هیئت علمی دانشکده مهندسی برق و کامپیوتر دانشکده

فنی دانشگاه تهران است. زمینه های تحقیقاتی مورد علاقه

ایشان عبارتند از: پردازش هوشمند متن و گفتار، مترجم

ماشینی، داده کاوی، بازیابی اطلاعات و شبکه های اجتماعی.

نشانی رایانامک ایشان عبارتست از:

hfaili@ut.ac.ir