علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۱۲- ۲۲
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

بهبود ترجمه ماشینی مبتنی بر قاعده با استفاده از قواعد نحوی آماری

حکیمه فدائی هشام فیلی فرناز قاسمی تودشکی

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

چکیده

ترجمه ماشینی مبتنی بر قاعده[1] از مجموعه‌ای از قواعد که دربردارنده اطلاعات زبانی هستند در فرایند ترجمه استفاده می‌کند. نتایج تولید شده توسط این مترجم‌ها معمولاً از نظر دستور زبان و ترتیب کلمات بهتر از نتایج مترجم‌های آماری هستند. ولی تحقیقات نشان داده است که این ترجمه‌ها از نظر روانی و انتخاب کلمات مناسب، ضعیف‌تر از مترجم‌های آماری هستند. در این مقاله هدف، بهبود انتخاب لغات در مترجم مبتنی بر قاعده است. این کار با استفاده از مجموعه‌ای از قواعد نحوی- لغوی مبتنی بر گرامر درخت- پیوندی[2] (TAG) انجام می‌شود. این قواعد احتمالاتی به‌صورت آماری از یک پیکره موازی با اندازه بزرگ استخراج شده‌اند. در سیستم ارائه شده، کلمات با ترتیب پیشنهادی مترجم مبتنی بر قاعده در زبان مقصد قرار می‌گیرند و به همین دلیل در ترجمه جملات از یک رمزگشای یکنواخت مبتنی بر برنامه‌ریزی پویا[3] استفاده شده است. در این سیستم بهترین ترجمه با استناد به احتمال قواعد استفاده شده و امتیاز مدل زبانی انتخاب می‌شود. آزمایش‌ها روی ترجمه انگلیسی به فارسی نشان داد که کیفیت نتایج به‌دست آمده از روش پیشنهادی حدود ۱/۳+ واحد بلو از کیفیت ترجمه به‌دست آمده توسط مبتنی بر قاعده پایه بالاتر است.

کلمات کلیدی: ترجمه ماشینی ترکیبی، ترجمه ماشینی مبتنی بر قاعده، قواعد آماری، قواعد نحوی-لغوی، گرامر درخت-پیوندی.

۱- مقدمه

ترجمه ماشینی یکی از شاخه‌های پرکاربرد و پیچیده در پردازش زبان طبیعی است. با توجه به گسترش روزافزون اسناد تحت وب، درخواست استفاده از سرویس‌های ترجمه ماشینی بسیار بالاست و شرکت‌های بزرگ فعال در این حوزه به‌طور مداوم در حال تلاش برای بهبود محصولات خود هستند. در سال‌های گذشته رویکردهای متفاوتی برای این مسئله ارائه شده است که هر کدام نقاط قوت و ضعف خاص خود را دارند. از میان این رویکردها می‌توان به ترجمه مبتنی بر قاعده، ترجمه آماری، ترجمه مبتنی بر مثال و ترجمه بر پایه شبکه‌های عصبی اشاره کرد. همچنین با ترکیب رویکردهای مختلف روش‌های ترکیبی[4] ارائه داد. این روش‌ها سعی در هم‌افزایی نقاط مثبت رویکردهای ترکیب شده دارند. در روش‌های ترکیبی معمولاً یکی از رویکردهای عنوان شده به‌عنوان رویکرد اصلی انتخاب می‌شود و فرایند ترجمه با تکیه بیشتر بر آن روش انجام می‌شود و روش‌های استفاده شده دیگر برای بهبود و کم کردن خطاهای روش پایه به کار می‌آیند. از روش‌های ترکیبی متداول در ترجمه ماشینی می‌توان به اضافه کردن اطلاعات زبانی و نحوی به مترجم‌های آماری و یا استفاده از عبارات استخراج شده در مترجم آماری برای غنی‌سازی مترجم مبتنی بر قاعده اشاره کرد.

رویکرد مبتنی بر قاعده به‌عنوان قدیمی‌ترین رویکرد در حوزه ترجمه ماشینی شناخته می‌شود و بر پایه مجموعه‌ای از قواعد که معمولاً توسط انسان ایجاد شده است استوار است. این مجموعه قواعد نحوه انتقال نحوی و لغوی از یک زبان به زبان دیگر را مدل‌سازی می‌کنند. از آنجایی‌که این مجموعه قواعد با نظارت انسان تولید می‌شوند استفاده از این رویکرد هزینه بسیار زیادی به همراه دارد. در ازای این هزینه، نتایج تولید شده توسط مترجم‌های مبتنی بر قاعده از نظر دستور زبان صحیح‌تر هستند و ترتیب کلمات در آن‌ها بهتر رعایت شده است. ترجمه ارائه شده توسط این مترجم‌ها، به دلیل در نظر گرفتن اطلاعات زبان‌شناسی، معمولاً از نظر تطابق فعل و فاعل، زمان، شخص و شمار افعال و خصوصیات ساخت‌واژی دیگر، بهتر از نتایج مترجم‌های آماری است. این در حالی است که مترجم‌های آماری از نظر انتخاب معادل مناسب برای کلمات در زبان مقصد، معمولاً بهتر از مترجم‌های مبتنی بر قاعده عمل می‌کنند. مترجم‌های آماری در انجام جابجایی‌های نزدیک

قوی‌تر هستند و به دلیل استفاده از مدل زبانی معمولاً ترجمه‌های روان‌تری نسبت به ترجمه مبتنی بر قاعده ارائه می‌دهند [۱]. بدین‌ترتیب اگر بتوان مترجم مبتنی بر قاعده را با استفاده از اطلاعات مترجم آماری غنی کرد، امید است که نتایج بهتری نسبت به مترجم مبتنی بر قاعده پایه به دست آید.

در این مقاله هدف ما استفاده از اطلاعات آماری در بهبود نتایج مترجم مبتنی بر قاعده است. بدین‌ترتیب که می‌خواهیم برای ترتیب قرارگیری کلمات در زبان مقصد به مترجم مبتنی بر قاعده استناد کنیم. از آنجایی که این مترجم‌ها از قواعد نحوی استفاده می‌کنند معمولاً در تشخیص جابجایی‌های دور بین کلمات قوی‌تر از مترجم‌های آماری عمل می‌کنند و از این نظر برای ترجمه بین زوج زبان‌های دور مانند انگلیسی و فارسی مناسب‌تر هستند. از آنجایی که بافت نحوی کلمه در زبان مبدأ می‌تواند در انتخاب معادل مناسب در زبان مقصد بسیار مفید باشد، ما مجموعه‌ای از قواعد همگام نحوی-لغوی را به‌صورت آماری استخراج کرده‌ایم و برای هر کلمه یا مجموعه از کلمات در جمله مبدأ باتوجه به بافت نحوی این کلمات و با استناد به این قواعد معادل مناسب را پیشنهاد می‌دهیم.

قواعد استخراج شده مبتنی بر گرامر درخت- پیوندی [۲] هستند و با توجه به میزان رخدادشان در پیکره آموزشی به هریک از آن‌ها احتمالی نسبت داده شده است. سیستم ما در نهایت ترجمه کلمات را از بین ترجمه‌های پیشنهاد شده توسط مترجم مبتنی بر قاعده و ترجمه‌های ارائه شده توسط مدل آماری انتخاب می‌کند. همچنین از آنجایی که ترتیب کلمات توسط مترجم مبتنی بر قاعده تعیین می‌شود، ترجمه به‌صورت یکنواخت[۵] انجام می‌شود و می‌توان برای انتخاب ترجمه بهینه از برنامه‌ریزی پویا استفاده کرد که پیچیدگی زمانی بسیار کمتری نسبت به الگوریتم‌های جستجو در ترجمه آماری دارد. در این مقاله از مترجم مبتنی بر قاعده فرازین [۳] به‌عنوان مترجم پایه استفاده شده است.

ادامه مقاله به این ترتیب سامان‌دهی شده است: در بخش ۲ به مرور کارهای پیشین و مرتبط با این مقاله می‌پردازیم. در بخش ۳ گرامر درخت - پیوندی را به اختصار معرفی می‌کنیم. بخش ۴ به توضیح جزئیات سیستم پیشنهادی اختصاص داده شده است. در بخش ۵ نتایج به دست آمده از این سیستم را مورد بررسی قرار می‌دهیم و درنهایت در بخش ۶ به نتیجه‌گیری می‌پردازیم.

۲- کارهای پیشین

روش‌های ترکیبی در ترجمه ماشینی مورد توجه بسیاری از محققان بوده‌اند. این روش‌ها معمولاً یک رویکرد را به‌عنوان رویکرد پایه انتخاب کرده و با ترکیب رویکردهای دیگر سعی در بهبود نتایج رویکرد پایه را دارند. در [۴] روش‌های ترکیبی به دو دسته تقسیم شده‌اند؛ روش‌هایی که بر پایه مترجم‌های آماری هستند و سعی دارند نتایج این مترجم‌ها را با تزریق اطلاعات زبان‌شناسی تقویت کنند و دسته دیگر روش‌هایی که بر پایه مترجم‌های مبتنی بر قاعده هستند و با استفاده از داده‌های به دست آمده از روش‌های آماری سعی در بهبود کیفیت مترجم مبتنی بر قاعده دارند.

۲-۱- روش‌های ترکیبی بر پایه مترجم آماری

ترجمه مبتنی بر قاعده می‌تواند در مرحله رمزگشایی و یا پیش و پس پردازش با ترجمه آماری ترکیب شود [۵]. معمولاً هدف سیستم‌هایی که از روش مبتنی بر قاعده در فاز پیش‌پردازش استفاده می‌کنند، نزدیک کردن ساختار جمله ورودی به ساختار زبان مقصد است [۶] [۷]. در این روش‌ها ترتیب کلمات در جمله ورودی باتوجه‌به قواعد نحوی تغییر پیدا می‌کند و جمله تغییریافته توسط مدل مبتنی بر عبارت ترجمه می‌شود. این تغییر در ترتیب کلمات هم در زمان آموزش مترجم و هم در زمان ترجمه انجام می‌شود. یعنی هر جمله سمت مبدأ پیکره موازی آموزش، توسط این قواعد بازآرایی می‌شود و پیکره جدید در فرایند آموزش مورد استفاده قرار می‌گیرد. این تغییرات روی مجموعه داده‌های توسعه و آزمون نیز انجام می‌شود. در برخی روش‌ها به جای قواعد تولید شده توسط خبره، از قواعد تولید شده با روش‌های آماری استفاده می‌شود [۸] [۹].

اطلاعات به‌دست آمده از ترجمه مبتنی بر قاعده ممکن است در فاز رمزگشایی مورد استفاده قرار بگیرد. در [۱۰] از ترجمه‌های به‌دست آمده توسط چند سیستم مبتنی بر قاعده، مجموعه‌ای از عبارات دوزبانه استخراج شده است. این عبارات به جدول عبارات یک مترجم مبتنی بر عبارت اضافه شده‌اند و در فاز رمزگشایی مورد استفاده قرار گرفته‌اند. احسن و همکارانش [۱۱] به روش‌های مختلف از ترجمه مبتنی بر قاعده برای بهبود ترجمه مبتنی بر عبارت استفاده کرده‌اند. از جمله این روش‌ها می‌توان به جابجایی کلمات در زبان مبدأ توسط مترجم مبتنی بر قاعده و همچنین غنی‌سازی جدول عبارات اشاره کرد. در [۱۲] نیز، عبارات دوزبانه منطبق با قواعد و همچنین واژه‌نامه استفاده شده در ترجمه مبتنی بر قاعده را به جدول عبارات روش مبتنی بر عبارت اضافه کرده‌اند.

از قواعد می‌توان در فاز پس‌پردازش نیز استفاده کرد. در [۱۳] روشی ارائه شده است که برخی خطاهای دستور زبان را در خروجی مترجم آماری با استفاده از مجموعه‌ای از قواعد تشخیص می‌دهد و تصحیح می‌کند. این قواعد بر پایه گرامر درخت - پیوندی هستند که ویژگی‌هایی به درختان آنها نسبت داده شده است. با تطبیق این ویژگی‌ها بر خروجی تولید شده، خطاها تشخیص داده می‌شوند.

۲-۲- روش‌های ترکیبی بر پایه مترجم مبتنی بر قاعده

در [۱] مقایسه‌ای بین رویکرد مبتنی بر قاعده و رویکرد آماری ارائه شده است. براساس این تحقیق مترجم‌های مبتنی بر قاعده در رعایت کردن ترتیب صحیح کلمات و تولید صورت صحیح ساخت‌واژی کلمات زبان مقصد مشکلات کمتری نسبت به مترجم‌های آماری دارند. از این‌رو این رویکرد برای ترجمه بین زوج زبان‌هایی که از نظر ساختاری از هم فاصله زیادی دارند مناسب به نظر می‌رسد. همچنین در ترجمه به زبان‌هایی که از نظر ساخت‌واژی غنی هستند، مترجم‌های مبتنی بر قاعده می‌توانند بهتر عمل کنند. به همین دلیل ما در سیستم ترکیبی پیشنهادی خود، یک مترجم مبتنی بر قاعده را به‌عنوان مترجم پایه در نظر گرفتیم و با افزودن اطلاعات آماری در جهت بهبود کیفیت نتایج این مترجم قدم برداشتیم.

مترجم مبتنی بر قاعده نیز می‌تواند در مراحل مختلف ترجمه از ترکیب شدن روش آماری و یا اطلاعات آن بهره ببرد. به عنوان مثال در روش ارائه شده در [۱۴] نتایج تولید شده توسط مترجم مبتنی بر قاعده با استفاده از یک مدل آماری پس‌ویرایش می‌شوند. این مدل آماری با گرفتن ترجمه مترجم مبتنی بر قاعده به‌عنوان متن مبدأ و ترجمه انسانی به‌عنوان متن مقصد، آموزش داده شده است. پس از آموزش مدل، نتایج مترجم مبتنی بر قاعده بر روی مجموعه داده تست به مترجم آماری مبتنی بر عبارت جدید داده می‌شوند تا پس‌ویرایش‌های لازم روی آن‌ها اعمال شود.

برخی از سیستم‌های ترکیبی سعی بر غنی کردن دادگان مترجم‌های مبتنی بر قاعده دارند. در [۱۵] واژه‌نامه مترجم مبتنی بر قاعده با استفاده از واژه‌های استخراج شده از منابع تحت وب تقویت شده است.همچنین در [۱۶] از عبارات استخراج شده به روش آماری از پیکره موازی برای تقویت واژه‌نامه مترجم مبتنی بر قاعده استفاده شده است. ما نیز به نحوی در روش پیشنهادی خود این کار را انجام می‌دهیم، با این تفاوت که ما ترجمه کلمات و عبارات را به همراه بافت نحوی‌شان به مجموعه اضافه می‌کنیم.

۲-۳- گرامرهای مورد استفاده در مدل کردن اطلاعات نحوی

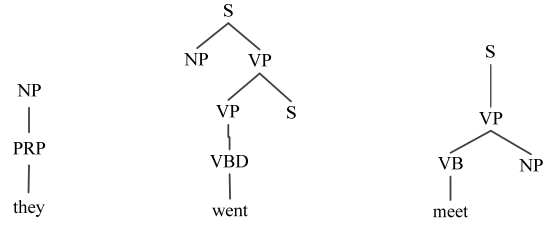
استفاده از اطلاعات نحوی در ترجمه ماشینی را از منظر نوع گرامر مورد استفاده نیز، می‌توان بررسی کرد. برخی مترجم‌ها از ساختارهای وابستگی در ترجمه استفاده می‌کنند [۱۷] در حالی که دیگر مترجم‌ها از گرامرهای مبتنی بر سازه استفاده می‌کنند. قدرت گرامرهای مختلف استفاده شده متفاوت است. یک فرمالیسم ممکن است قادر به مدل کردن پدیده‌هایی در زبان باشد که گرامر دیگر توانایی نمایش آن‌ها را ندارد. در بین گرامرهای مبتنی بر سازه، ابتدا گرامرهای مستقل از متن (CFG) در ترجمه ماشینی مورد استفاده قرار گرفتند [۸]. ناتوانی این گرامرها در مدل کردن برخی جابجایی‌ها باعث شد محققین این حوزه به سراغ گرامرهایی با دامنه محلی بزرگ‌تر[۶] مانند TSG[۷] بروند [۱۸] [۱۹]. در TSG تنها یک عملیات "جایگزینی[۸]" وجود دارد و در هر مرحله از اشتقاق یک غیرپایانه با یک درخت جایگزین می‌شود. در CFG قواعد به‌صورت درخت‌هایی دو سطحی هستند و به همین دلیل تنها می‌توانند جابجایی‌های در سطح همزادها[۹] را مدل کنند ولی در TSG این محدودیت وجود ندارد و قواعد می‌توانند از درختانی با بیش از دو سطح تشکیل شده باشند.

TSG نیز دارای محدودیت‌هایی است و نمی‌تواند ساختارهای حساس به متن[۱۰] را مدل کند. همچنین برخی از قواعد تولید شده توسط مدل مبتنی بر TSG ساختارهای مسطح و بزرگی را مورد پوشش قرار می‌دهند که باعث می‌شوند یک کلمه با تعداد زیادی از وابسته‌هایش در قالب یک قاعده دیده شود. بدین ترتیب با دیدن ترکیب جدیدی از وابسته‌ها در زمان ترجمه، سیستم قاعده‌ای برای پوشش آن‌ها پیدا نمی‌کند. برای رفع این مشکل، و استخراج قواعد عمومی‌تر از دودویی‌سازی درختان استفاده می‌شود. در این روش هر سازه با بیش از دو فرزند به تعدادی زیرسازه شکسته می‌شود و درخت مربوطه به شکل دودویی در می‌آید. این کار باعث شکسته شدن ساختارهای بزرگ و مسطحی که راجع به آن‌ها صحبت شد، به ساختارهای کوچک‌تر می‌شود. در [۲۰] عنوان می‌شود که دودویی‌سازی درختان که از آن به‌عنوان روشی برای کم کردن محدودیت‌های نحوی و رسیدن به قواعدی عمومی‌تر مطرح شد، ممکن است به آزادی بیش‌ازحد در مدل بینجامد که باعث تولید نتایج نادرست از نظر نحوی شود. همچنین در دودویی‌سازی تفاوتی بین وابسته‌های اجباری و اختیاری هسته[۱۱] گذاشته نمی‌شود و به همین دلیل در صورتی که بین یک هسته و یکی از وابسته‌های اجباری آن یک وابسته اختیاری قرار بگیرد، وابسته اجباری با هسته در قالب یک قاعده استخراج نمی‌شوند. برای کم کردن این مشکل و همچنین بالا بردن قدرت مدل در پذیرفتن ساختارهایی که TSG قادر به پذیرفتن آن‌ها نیست، مترجم‌هایی به سمت استفاده از گرامر درخت-پیوندی رفتند که در دسته‌بندی چامسکی جزء زبان‌های حساس به متن ملایم[۱۲] دسته‌بندی می‌شود. در این گرامرها علاوه بر عملیات جایگزینی که در TSG نیز وجود دارد، دارای عملیات الحاق[۱۳] نیز هست. این عملیات قادر است زیر درخت‌هایی را به جای یک گره در یک درخت موجود اضافه کند و برای مدل کردن اتصال سازه‌های اختیاری[۱۴] و ساختارهای بازگشتی استفاده می‌شود. درختان TAG اطلاعات ساختار وابستگی را نیز در خود دارند و به همین دلیل قدرتمندتر از گرامرهای وابستگی هستند. در سیستم پیشنهادی در این مقاله از گرامر درخت-پیوندی در مدل کردن اطلاعات نحوی استفاده شده است. در بخش بعد در مورد گرامر درخت- پیوندی توضیحاتی ارائه خواهد شد.

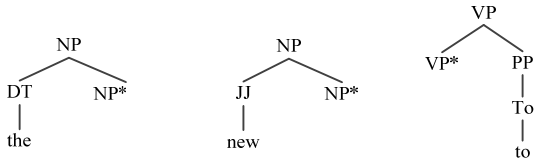
۳- گرامر درخت- پیوندی

گرامر درخت - پیوندی یکی از فرمالیسم‌های قوی در مدل کردن پدیده‌های زبانی است. در این گرامر واحدهای سازنده، درختانی هستند که به آن‌ها درختان اولیه[۱۵] گفته می‌شود. درختان اولیه خود به دو دسته درختان ابتدایی و درختان کمکی تقسیم می‌شوند. درختان ابتدایی ساختارهای نحوی اصلی در جمله را تشکیل می‌دهند و در آن‌ها تمام گره‌های داخلی، غیرپایانه و برگ‌ها پایانه یا غیرپایانه هستند. در گرامر درخت- پیوندی لغوی[۱۶] هر درخت اولیه حتماً دارای یک برگ لغوی است که به آن لنگر[۱۷] گفته می‌شود. یک اشتقاق TAG حتماً با یک درخت ابتدایی شروع می‌شود. نمونه‌هایی از درختان ابتدایی در شکل ۱ نمایش داده شده‌اند.

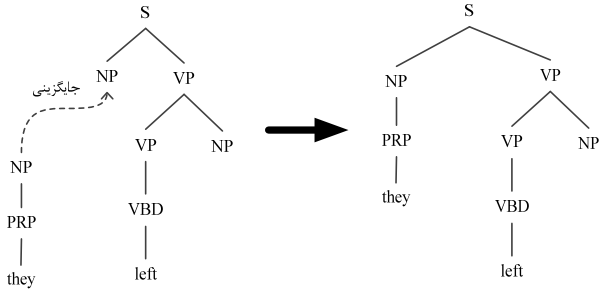
دسته دوم از درختان، که وجه تمایز TAG با TSG به‌حساب می‌آیند، درختان کمکی هستند. این درختان برای وارد کردن سازه‌ها در یک درخت دیگر استفاده می‌شوند و معمولاً نشان‌دهنده سازه‌های اختیاری و ساختارهای بازگشتی هستند. در هر درخت کمکی یک برگ با علامت ستاره (*) مشخص می‌شود که به آن گره انتهایی[۱۸] گفته می‌شود. غیرپایانه مربوط به این گره حتماً با غیرپایانه‌ای که در ریشه درخت قرار دارد یکسان است. چند نمونه از این درختان در شکل ۲ نمایش داده شده‌اند. قاعده سمت چپ نشان می‌دهد که حرف تعریف "the" می‌تواند قبل از یک عبارت اسمی اضافه شود و یک عبارت اسمی جدید تولید کند. قاعده وسط نشان می‌دهد که صفت "new" می‌تواند پیش از یک گروه اسمی اضافه شود و یک گروه اسمی جدید بسازد. و در نهایت قاعده سمت راست نشان دهنده اضافه شدن حرف اضافه"to" به انتهای یک عبارت فعلی و تولید یک عبارت فعلی جدید است.



شکل ۱- نمونه‌هایی از درختان ابتدایی TAG



شکل ۲- نمونه‌هایی از درختان کمکی TAG



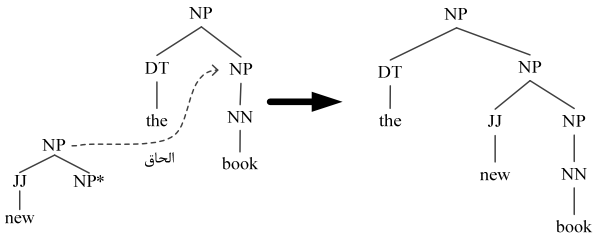
شکل ۳- نمونه‌ای از عملیات جایگزینی

در TAG دو عملیات روی درختان اولیه قابل اعمال است که باعث اتصال درختان اولیه به یکدیگر می‌شوند: عملیات جایگزینی و الحاق. جایگزینی نوعی اتصال محسوب می‌شود و روی درختان ابتدایی قابل اعمال است. در این عملیات یک درخت ابتدایی با ریشه X به یک درخت دیگر با برگ X متصل می‌شود. این

عملیات در TSG نیز وجود دارد و به‌طور مشابه انجام می‌شود. نمونه‌ای از این عملیات که اضافه شدن فاعل به جمله را نشان می‌دهد در شکل ۳ نمایش داده شده است.

الحاق عملیاتی است که روی درختان کمکی انجام می‌شود. در این عملیات یک درخت کمکی در میان یک درخت اولیه اضافه می‌شود. گره‌ای در درخت پدر که درخت کمکی در آن اضافه می‌شود دارای یک "جایگاه الحاق[۱۹]" است. در این عملیات فرزندان گره محل الحاق جدا شده و به زیر گره انتهایی متصل می‌شوند. سپس درخت کمکی تغییریافته در گره محل الحاق جایگزین می‌شود. نمونه‌ای از این عملیات که اضافه شدن یک صفت به یک عبارت اسمی را نشان می‌دهد در شکل ۴ نمایش داده شده است. قدرت گرامر درخت - پیوندی در عملگر الحاق نهفته است. این عملگر به ما این امکان را می‌دهد که یک درخت اولیه را در میان یک درخت اولیه دیگر وارد کنیم و این کار می‌تواند به دفعات انجام شود. در درخت اشتقاق درخت - پیوندی یک جمله، سازه‌های اختیاری معمولاً در قالب عملیات الحاق از هسته خود جدا می‌شوند. این جداسازی باعث می‌شود در استخراج قواعد به قواعد کلی‌تری دست پیدا کنیم که کمتر با خطر تنکی[۲۰] روبه‌رو هستند [۲۱].

درخت تجزیه هر جمله در زبان، از ترکیب درختان ابتدایی و کمکی با عملیات جایگزینی و الحاق ساخته می‌شود و به آن درخت استنتاج[۲۱] گفته می‌شود و از نظر ظاهری تفاوتی با درخت تجزیه CFG ندارد. سابقه درختان اولیه سازنده و عملیات انجام شده روی آن‌ها برای ساخت جمله، در درخت اشتقاق[۲۲] ذخیره می‌شود.



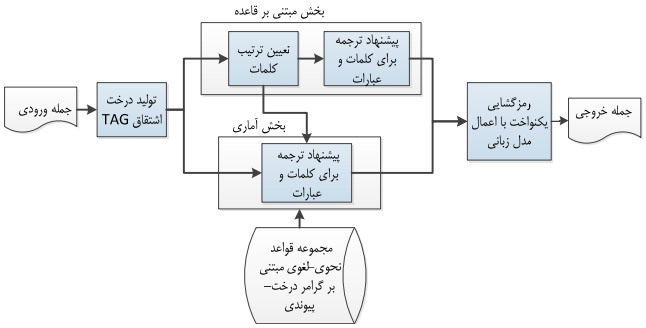
شکل ۴- نمونه‌ای از عملیات الحاق

۴- سیستم پیشنهادی

هدف کلی سیستم پیشنهادی، بهبود ترجمه ارائه شده توسط یک سیستم مبتنی بر قاعده با استفاده از مجموعه از قواعدی نحوی است که به‌صورت آماری از یک پیکره موازی استخراج‌شده‌اند. این قواعد مبتنی بر گرامر درخت - پیوندی هستند و نحوه استخراج آن‌ها در بخش ۴-۱ شرح داده شده است.

نمودار بلوکی نمایش داده شده در شکل ۵ روند کار سیستم ترکیبی پیشنهاد شده در این مقاله را نمایش می‌دهد. مطابق این شکل برای هر جمله ورودی ابتدا درخت اشتقاق در فرمالیسم درخت - پیوندی تولید می‌شود. برای این کار ابتدا جمله توسط تجزیه‌گر استنفورد [۲۲] تجزیه نحوی شده و طی فرایندی مبتنی بر قاعده [۲۳] درخت اشتقاق آن تولید می‌شود. در مرحله بعد مترجم انگلیسی- فارسی فرازین با استناد به این درخت اشتقاق و قواعد داخلی خود، ترتیب صحیح ترجمه کلمات در زبان فارسی را تعیین می‌کند. همچنین برای هر کلمه یا هر دنباله از کلمات که معادلی در واژه‌نامه خود داشته باشد، واژه یا عبارت ترجمه را پیشنهاد می‌دهد. لازم به ذکر است که در روش ارائه شده استفاده از درخت اشتقاق در مترجم مبتنی بر قاعده ضروری نیست و هر مترجم مبتنی بر قاعده دیگری با ساختار دلخواه می‌تواند جایگزین فرازین شود.

بخش آماری سیستم با توجه به ترتیب کلمات ارائه شده توسط فرازین و با استفاده از قواعد خود برای هر کلمه یا هر دنباله متوالی از کلمات، ترجمه‌های پیشنهادی را از مجموعه قواعد آماری استخراج می‌کند. این ترجمه‌های پیشنهادی به همراه ترجمه‌های ارائه شده توسط فرازین در اختیار رمزگشای[۲۳] یکنواخت قرار داده می‌شود تا این رمزگشا با توجه به امتیاز هر ترجمه و با در نظر گرفتن مدل زبانی بهترین ترجمه را از بین ترجمه‌های ممکن انتخاب و به‌عنوان خروجی سیستم اعلام کند. جزئیات مربوط به رمزگشایی در بخش ۴-۲ شرح داده شده است.



شکل ۵- روال کار سیستم پیشنهادی

۴-۱- قواعد نحوی- لغوی

همان‌طور که گفته شد قواعد استفاده شده در این سیستم بر پایه گرامر درخت - پیوندی هستند. روال استخراج قواعد نحوی، که یک پیکره دوزبانه را به‌عنوان ورودی می‌گیرد، شامل مراحل زیر است [۲۰]:

۱. به دست آوردن هم‌ترازی در سطح کلمه بین جملات مبدأ و مقصد پیکره موازی
۲. تجزیه نحوی جملات سمت مبدأ
۳. تبدیل درختان تجزیه به درختان اشتقاق TAG
۴. استخراج قواعد کمینه از درختان اشتقاق
۵. استخراج قواعد مرکب
۶. تخمین احتمالات مربوط به قواعد

برای تولید هم‌ترازی در سطح کلمات از ابزاری مانند GIZA++ [۲۴] استفاده می‌شود. برای تجزیه نحوی جملات سمت مبدأ نیز از یکی از تجزیه‌گرهای موجود مانند تجزیه‌گر استنفورد [۲۲] می‌توان استفاده کرد. پس از تولید درخت تجزیه باید درخت اشتقاق TAG را از روی آن ساخت. برای این کار از روش‌های متفاوتی می‌توان استفاده کرد. روش ارائه شده در [۲۰] و [۲۳] تقریباً مشابه یکدیگر است. هر دو این روش‌ها از اطلاعات زبان‌شناسی برای تشخیص هسته سازه و وابسته‌های مربوط به آن استفاده می‌کنند و با توجه به این اطلاعات درخت اشتقاق TAG را تولید می‌کنند. این در حالی است که الگوریتم ارائه شده در [۲۵] تنها با توجه به ظاهر درخت و بدون استفاده از اطلاعات زبان‌شناسی دست به تولید درخت اشتقاق می‌زند. از این‌رو این روش برای یک درخت تجزیه واحد چند درخت اشتقاق محتمل تولید می‌کند در حالی که الگوریتم ارائه شده توسط چن [۲۳] تنها یک درخت اشتقاق به‌عنوان خروجی تولید می‌کند. به همین دلیل و همچنین به دلیل استفاده از اطلاعات زبانی در روش ارائه شده توسط چن، ما از این الگوریتم استفاده می‌کنیم.

۴-۱-۱- استخراج قواعد کمینه

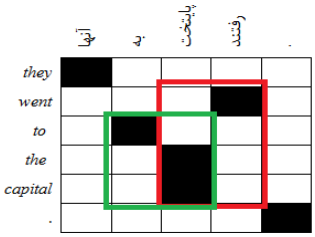
پس از تولید درخت اشتقاق برای جمله ورودی، با توجه به آن درخت و همچنین هم‌ترازی کلمات جمله با کلمات جمله زبان مقصد، قواعد کمینه را استخراج

می‌کنیم. برای این کار تعدادی تعریف را مرور کرده و همزمان روش استخراج قواعد را توضیح می‌دهیم.

تعریف ۱: زوج دنباله‌ای از کلمات در زبان مبدأ و مقصد را سازگار[۲۴] با هم‌ترازی (به‌اختصار "سازگار") گویند که هر کلمه در سمت مبدأ تنها به کلمات داخل دنباله سمت مقصد و یا Null هم‌تراز شده باشد و برعکس [۲۶].

این تعریف اساس استخراج عبارات در مترجم مبتنی بر عبارت است. شکل ۶ هم‌ترازی‌های بین کلمات دو جمله فارسی و انگلیسی را نمایش می‌دهد. با توجه به این شکل زوج عبارت <به پایتخت، to the capital> سازگار است. ولی زوج عبارت <پایتخت رفتند، went to the capital> به دلیل این‌که کلمه "to" به کلمه "به" که بیرون از عبارت فارسی است، هم‌تراز شده است، غیرسازگار است.

تعریف ۲: درخت اولیه‌ای را "کاندیدای استخراج قاعده[۲۵]" گویند که لنگر آن و لنگرهای تحت پوشش هر یک از درختان اولیه فرزندان آن در درخت تجزیه، تشکیل یک عبارت سازگار با سمت مقصد دهند [۲۰].



شکل ۶- مثال‌هایی از عبارات سازگار و غیرسازگار

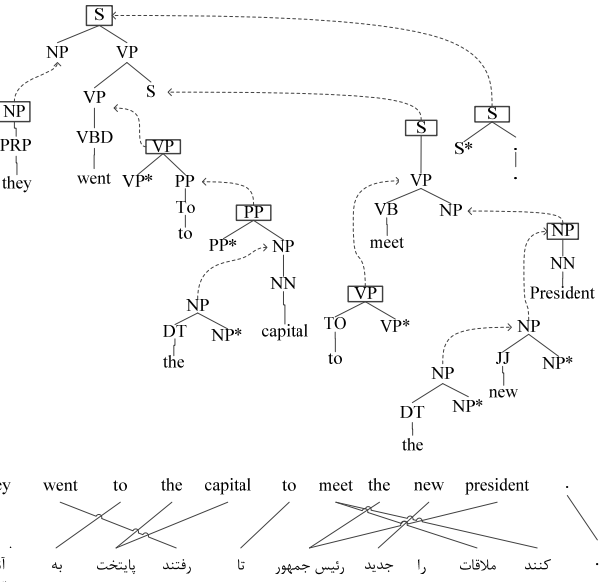
برای استخراج قواعد در مدل نحوی مبتنی بر TAG، درخت اشتقاق به‌صورت پایین به بالا مورد بررسی قرار می‌گیرد و هر درخت اولیه در صورت کاندیدای استخراج بودن، می‌تواند سمت چپ یک قاعده را تشکیل دهد. در غیر این صورت باید آن‌قدر با پدران خود ترکیب شود تا به یک پدر کاندیدای استخراج برسد. لازم به ذکر است که پدر مربوطه دیگر خود به‌تنهائی قادر به تشکیل یک قاعده جدید نخواهد بود. شکل ۷ درخت اشتقاق مربوط به جمله "They went to the capital to meet the new president." را نمایش می‌دهد که درختان اولیه کاندیدای استخراج در آن مشخص شده‌اند.

بدین ترتیب تعریف ۲ را می‌توان به شکل زیر بازنویسی کرد:

تعریف ۳: درخت اولیه و یا ترکیبی از درختان اولیه می‌تواند سمت چپ یک قاعده را تشکیل دهد که در آن درخت اولیه‌ی ریشه، "کاندیدای استخراج" باشد و تمامی فرزندان آن، خود ریشه یک درخت "کاندیدای استخراج" باشند.

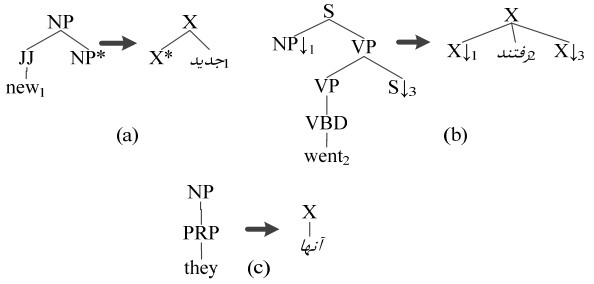
شکل ۸ نمونه‌هایی از قواعد کمینه استخراج شده برای جمله شکل ۷ را نمایش می‌دهد. در سمت راست قواعد تنها متغیر X و پایانه‌ها دیده می‌شوند که تشکیل یک درخت را می‌دهند. این درخت ماهیت نحوی ندارد و از تحلیل نحوی به دست نیامده است و معادل یک رشته در نظر گرفته می‌شود. در این قواعد، جایگاه‌های جایگزینی با ↓ مشخص شده‌اند. اعداد متصل به هر برگ برای مشخص کردن زوج گره‌های همگام در سمت مبدأ و مقصد هستند.

برای به دست آوردن سمت راست قواعد به هم‌ترازی بین عبارت زبان مبدأ و مقصد توجه می‌شود. به ازای هر جایگاه جایگزینی در سمت راست قاعده، یک X در سمت چپ قاعده قرار داده می‌شود تا جایگاه درختی را که در آینده در آن گره جایگزین می‌شود را معین کند. ترتیب قرارگیری این متغیرها به ترتیب قرارگیری لنگرهای وابسته به آن‌ها در زبان مقصد مرتبط است. این ترتیب جابجایی‌های لازم در ترجمه را ایجاد می‌کند و این مسئله قدرت اصلی سیستم‌های ترجمه مبتنی بر نحو است.



شکل ۷- درختان کاندیدای استخراج در جمله "They went to the capital to meet the new president."

در شکل ۸ قاعده اوّل نشان می‌دهد که اگر صفت "new" به یک گروه اسمی بپیوندد، در ترجمه به فارسی جای ترجمه صفت (جدید) با ترجمه گروه اسمی جابجا می‌شود. قاعده دوم ترجمه یک جمله متعدی با فعل "went" را نشان می‌دهد. در انگلیسی فعل بین فاعل و مفعول قرار دارد، در حالی که در ترجمه به فارسی فعل به انتهای جمله و بعد از مفعول انتقال پیدا می‌کند. قاعده سوم یک قاعده لغوی است که هیچ غیرپایانه‌ای در آن حضور ندارد و نشان می‌دهد که ضمیر "they" به "آنها" ترجمه می‌شود.



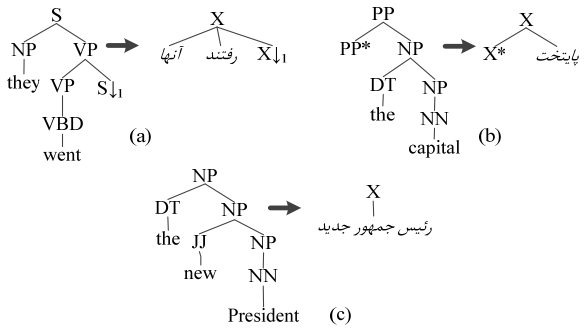
شکل ۸- نمونه‌هایی از قواعد کمینه استخراج شده از جمله مثال

۴-۱-۲- استخراج قواعد مرکب

قواعد کمینه مطابق آنچه در بخش ۴-۱-۱- توضیح داده شد، استخراج می‌شوند. این قواعد کوچک‌ترین قواعد قابل استخراج از داده‌های ورودی و غالباً شامل یک لنگر هستند و به همین دلیل بافت کمی را پوشش می‌دهند. به‌منظور نزدیک شدن به قدرت مترجم‌های مبتنی بر عبارت در ترجمه عبارات چند کلمه‌ای، محققین به سراغ قواعد پیچیده‌تر و بزرگ‌تر رفتند تا بتوانند بافت بزرگ‌تری را مورد پوشش قرار دهند [۲۷]. این قواعد که از ترکیب چندقاعده کمینه به دست می‌آیند قواعد مرکب[۲۶] نامیده می‌شوند. سمت چپ این قواعد ترکیبی از چند درخت اولیه‌ی همبند و سمت راست آنها رشته‌ای در زبان مقصد است که به‌طور مستقل با توجه به هم‌ترازی‌ها استخراج می‌شود. معمولاً برای تولید قواعد مرکب محدودیت‌هایی قائل می‌شوند که اندازه و تعداد قواعد از حدی بزرگ‌تر نشود. این محدودیت می‌تواند بر روی اندازه درخت تولید شده از نظر ارتفاع و پهنا و یا تعداد

غیرپایانه‌های میانی در آن باشد.

بنابراین در مدل پیشنهادی ما نیز از قواعد مرکب استفاده خواهد شد به طوری که هر زیر درخت از درخت اشتقاق که محدودیت‌های تعیین شده را ارضا کند، می‌تواند به‌عنوان سمت چپ یک قاعده مرکب در نظر گرفته شود. در انتخاب این زیردرخت نوع یال بین درختان اولیه و یا همان گره‌های درخت اشتقاق دارای اهمیت نیست. تعدادی از قواعد مرکب استخراج شده از روی این جمله در شکل ۹ نشان داده شده‌اند.



شکل ۹- نمونه‌هایی از قواعد مرکب استخراج شده برای جمله مثال

۴-۱-۳- تخمین احتمالات

پس از استخراج قواعد برای تمام جملات موجود در پیکره آموزش، احتمالات مربوطه به هر قاعده باید محاسبه شوند. برای این کار از روش تخمین بیشینه درست‌نمایی[۲۷] استفاده خواهد شد. برای هر قاعده احتمال P(f|e) محاسبه خواهد شدکه در آن‌ها e درخت نحوی سمت مبدأ قاعده به همراه لنگرهایش و f درخت ساختاری و یا رشته سمت مقصد است. این احتمال را می‌توان با استفاده از رابطه (۱) محاسبه کرد. در این رابطه تعداد رخداد درخت‌های سمت مبدأ و مقصد قاعده به همراه لنگرهای‌شان در کل پیکره آموزشی در صورت کسر قرار می‌گیرد. در مخرج کسر لنگرها در شمارش نادیده گرفته می‌شوند.

$$p(f|e) = \frac{\text{Count}(e,f)}{\sum_{f_i \in \text{Set of all target strings without anchor}} \text{Count}(e,f_i)} \qquad (۱)$$

از آنجایی که قواعد استفاده شده در این سیستم قواعدی نحوی - لغوی هستند، استفاده از آنها به ما کمک می‌کند که ترجمه کلمات را با توجه به ساختار نحوی جمله ورودی انتخاب کنیم. نمونه‌هایی از این قواعد در جدول ۱ نمایش داده شده‌اند.

در جدول ۱ قاعده اوّل نشان می‌دهد که کلمه "break" وقتی به‌عنوان اسم در جمله ظاهر شود به احتمال ۰/۱۴ به "استراحت" ترجمه می‌شود. قاعده دوم نشان می‌دهد عبارت "very different"هنگامی که در نقش صفت برای یک گروه اسمی باشد، در ۲۰٪ موارد به "بسیار متفاوتی" ترجمه شده است. قاعده آخر نشان می‌دهد اگر عبارت "I know" به یک جمله وابسته پیوند داده شود، به احتمال ۰/۱۴ به "من می‌دانم که" ترجمه می‌شود.

۴-۲- رمزگشایی

همان‌طور که گفته شد از آنجایی‌که ترتیب کلمات در زبان مقصد در ابتدای کار توسط مترجم مبتنی بر قاعده تعیین می‌شود، فضای جستجو در زمان رمزگشایی بسیار کوچک‌تر از مترجم‌های آماری است و همچنین می‌توان از برنامه‌ریزی پویا برای پیدا کردن بهترین ترجمه استفاده کرد.

جدول ۱- نمونه‌هایی از قواعد نحوی- لغوی مبتنی بر گرامر درخت- پیوندی

احتمال	ترجمه	درخت سمت مبدأ
۰/۱۴	استراحت	NP NN ***break***
۰/۲	بسیار متفاوتی	NP ADJP NP* RB ADJP ***very*** JJ ***different***
۰/۲۳	شکستن	S VP NP↓ VB ***break***
۰/۳۳	برآورد شده است	S NP↓ VP VP VP VBZ VBN S↓ ***is*** ***estimated***
۰/۱۴	من می‌دانم که	S NP VP PRP VBD SBAR↓ ***I*** ***know***

روش کار به این ترتیب است که یک جدول برنامه‌ریزی پویا با سایز $n \times n$ خواهیم داشت که در آن n تعداد کلمات جمله مبدأ است. و کلمات جمله مبدأ را با ترتیب تعیین شده توسط مترجم مبتنی بر قاعده در این جدول قرار می‌دهیم. بدین‌ترتیب خانه [i, j] در این جدول در بردارنده ترجمه‌های پیشنهادی برای کلمه iام تا jام (در ترتیب جدید) خواهد بود. و در نهایت ترجمه جمله در خانه [1, n] تولید خواهد شد. نمونه‌ای از این جدول برای جمله "I finally went to school yesterday." پس از جابجایی کلمات توسط فرازین در شکل ۱۰ نمایش داده شده است. همان‌طور که توضیح داده شد در این مرحله کلمات هنوز ترجمه

نشده‌اند و تنها ترتیب قرارگیری آنها مطابق ترتیب در زبان فارسی شده است.

برای هر خانه [i, j]، m بهترین ترجمه آن نگه داشته می‌شود و بقیه ترجمه‌ها نادیده گرفته می‌شوند. ترجمه‌های کاندیدا برای خانه [i, j] از دو طریق ایجاد می‌شوند: ۱- مستقیماً توسط فرازین یا مجموعه قواعد ما ارائه می‌شوند. ۲- از ترکیب ترجمه عبارات تشکیل دهنده آن تولید می‌شوند.

برای هر دنباله i تا j از کلمات جمله مبدأ، ترجمه p ممکن است مستقیماً توسط فرازین و یا قواعد ما ارائه شود. امتیاز اختصاص داده شده به این ترجمه که با score(p) نمایش داده می‌شود مطابق رابطه (۲) محاسبه می‌شود. در این رابطه $lm(p)$ امتیاز مدل زبانی عبارت p به شرط کلمات قبل از آن است و $tm(p|w_{i-j})$ احتمال ترجمه کلمات i تا j به عبارت فارسی p است. بسته به این‌که P توسط کدام مرجع ترجمه ارائه شده است، $tm(p|w_{i-j})$ از فرازین یا مجموعه قواعد همگام نحوی گرفته می‌شود. w_1 و w_2 وزن‌هایی هستند که برای امتیاز مدل زبانی و امتیاز ترجمه در نظر گرفته می‌شوند.

$$Score(p) = w_1 \log(lm(p)) + w_2 \log\left(tm(p|w_{i-j})\right) \quad (۲)$$

در روش دوم برای پر کردن خانه [i, j] می‌توان از ترجمه‌های خانه‌های [i, k] و [k+1, j] استفاده کرد. بدین‌ترتیب که ترجمه‌های خانه [i, k] و خانه [k+1, j] را دوبه‌دو باهم در نظر می‌گیریم و به هم متصل می‌کنیم. بدین‌ترتیب برای هر k که بین i و j انتخاب شود حداکثر m^2 ترجمه تولید شده و جزء کاندیداهای خانه [i, j] محسوب می‌شود. امتیاز اختصاص داده شده به هر کاندیدا که از ترکیب ترجمه p از خانه [i, k] و ترجمه q از خانه [k+1, j] تولید شده است مطابق رابطه (۳) محاسبه می‌شود. Score(p) از خانه [i, k] و Score(q) از خانه [k+1, j] خوانده می‌شود. از آنجایی که امتیاز مدل زبانی نیز در محاسبه امتیاز هر خانه دخالت داده شده است لازم است که احتمال n-gram کلمات اولیه q به نسبت کلمات انتهایی p نیز در محاسبه امتیاز Score(pq) دخالت داده شوند.

$$Score(pq) = Score(p) + Score(q) + w_1 \log\left(lm(\text{کلمات مرزی } q_9 p)\right) \quad (۳)$$

در جدول برنامه‌ریزی پویا، قطر اصلی جدول شامل کلمات تکی است. این خانه‌ها با ترجمه‌های فرازین برای هر کلمه و همچنین با استفاده از ترجمه‌های ارائه شده توسط قواعد نحوی-لغوی ما پر می‌شوند. سپس خانه‌های دیگر جدول با ترتیب نمایش داده شده در شکل ۱۰ و با توجه به توضیحات ارائه شده، پر می‌شوند.

یکی از چالش‌هایی که در استفاده از این روش با آن روبه‌رو هستیم این است که خروجی‌های تولید شده توسط سیستم فرازین احتمالاتی نیستند و تنها ترجمه کلمات به‌عنوان خروجی داده می‌شود، یعنی مقداری برای $tm(p|w_{i-j})$ مستقیماً در اختیار ما قرار داده نمی‌شود. راه‌های متعددی برای برطرف کردن این مشکل وجود دارد که ما ساده‌ترین آن‌ها را انتخاب کرده‌ایم و به تمام ترجمه‌های ارائه شده توسط مترجم فرازین یک احتمال ثابت اختصاص دادیم. این احتمال با توجه به مجموعه داده توسعه انتخاب شده است.

۵- آزمایش‌ها و ارزیابی

برای انجام آزمایش‌ها نیاز به یک پیکره دوزبانه برای استخراج قواعد نحوی - لغوی داریم. آزمایش‌ها روی ترجمه انگلیسی- فارسی انجام شده است و ما از پیکره دوزبانه AFEC [۲۸] به‌عنوان پیکره آموزشی خود استفاده کردیم که اطلاعات مربوط به آن در جدول ۲ ارائه شده است.

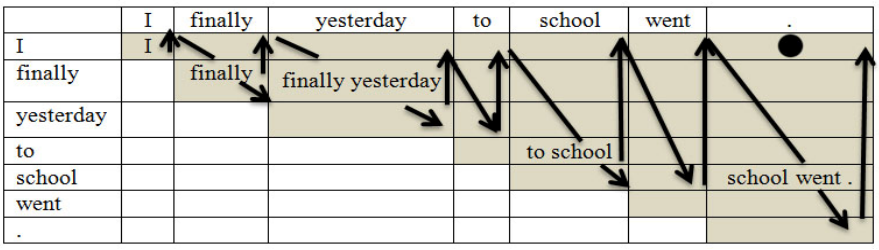
با استفاده از پیکره آموزشی و روش ارائه شده در بخش ۴-۱ مجموعه‌ای از قواعد نحوی- لغوی را استخراج کردیم. همان‌طور که گفته شد هر مجموعه از درختان اولیه که تشکیل یک زیردرخت متصل را بدهند و شرایط گفته شده در بخش ۴-۱ را داشته باشند می‌توانند یک قاعده نحوی را تشکیل دهد. ولی از آنجایی که به‌طور متوسط هر چه اندازه درخت بزرگ‌تر شود، تعداد رخداد آن در پیکره کمتر می‌شود، ما اندازه قواعد خود را محدود کردیم. در آزمایش‌های انجام شده از قواعدی استفاده شده است که سمت راست آن‌ها حداکثر از ترکیب سه درخت اولیه ایجاد شده باشد. همچنین برای بالا بردن کیفیت، قواعدی که تنها یک بار در کل پیکره آموزش دیده شده بودند را حذف کردیم. جدول ۳ اطلاعاتی در مورد مجموعه قواعد استخراج شده ارائه می‌دهد. نمونه‌های از مجموعه قواعد استخراج شده در جدول ۱ نمایش داده شده بود.

جدول ۲- اطلاعات مربوط به پیکره آموزش و مجموعه داده‌های تست و توسعه

		پیکره آموزشی	مجموعه توسعه	مجموعه تست
	تعداد جملات	۶۸۳ هزار	۴۶۰	۴۲۷
انگلیسی	تعداد کلمات	۱۴/۵ میلیون	۱۱،۶۰۰	۱۰،۸۰۰
	تعداد کلمات یکتا	۱۵۸ هزار	۳۰۰۰	۲،۸۰۰
	متوسط طول جمله	۲۱	۲۵	۲۵
فارسی	تعداد کلمات	۱۵/۴ میلیون	-	-
	تعداد کلمات یکتا	۲۰۲ هزار	-	-
	متوسط طول جمله	۲۲	-	-

جدول ۳- آمار مربوط به قواعد استخراج شده

تعداد قواعد کمینه (با ۱ لنگر)	۳۰۶ هزار
تعداد قواعد مرکب (با ۲-۳ لنگر)	۴۰۵ هزار
متوسط تعداد ترجمه‌های مختلف برای هر کلمه و بافت نحوی آن در قواعد کمینه	۳
متوسط تعداد ترجمه‌های مختلف برای هر کلمه و بافت نحوی آن در قواعد مرکب	۲/۶



شکل ۱۰- نمونه‌ای از جدول برنامه‌ریزی پویا

پس از اتمام مرحله استخراج قواعد می‌توانیم روش خود را روی مجموعه داده تست ارزیابی کنیم. اطلاعات مربوط به مجموعه داده‌های توسعه و تست در جدول ۲ ارائه شده است. این مجموعه‌ها از دامنه خبر هستند و در آن‌ها هر جمله انگلیسی چهار ترجمه مرجع دارد.

در روش ارائه شده چند پارامتر وجود دارد که مقادیر آن‌ها باید با توجه به مجموعه داده توسعه، تنظیم شود:

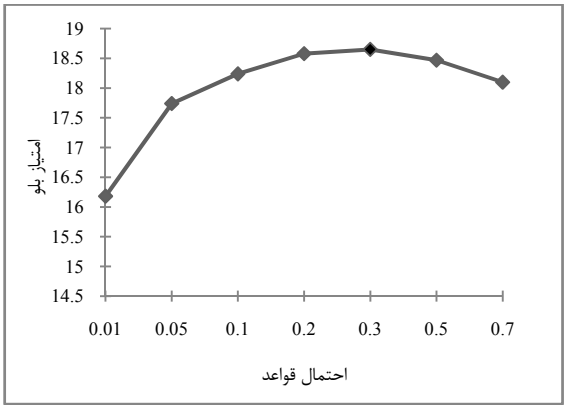
۱- احتمال ثابت اختصاص داده شده به ترجمه‌های فرازین

۲- وزن امتیاز مدل زبانی (w_1)

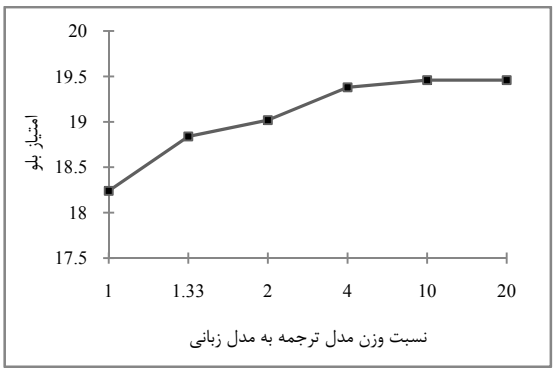
۳- وزن امتیاز ترجمه (w_2)

برای تنظیم مقادیر این پارامترها، در فرایندی تکراری مقادیر این پارامترها را تغییر دادیم و مجموعه داده توسعه را با توجه به آن مقادیر ترجمه کردیم و کیفیت ترجمه را با توجه به ترجمه‌های مرجع سنجیدیم. درنهایت مقادیری که بهترین نتیجه را روی مجموعه داده‌های توسعه داشتند به‌عنوان مقادیر نهایی انتخاب و در ارزیابی سیستم از آن‌ها استفاده کردیم. در ارزیابی‌ها از معیار بلو[۲۸] [۲۹] به‌عنوان معیار ارزیابی کیفیت ترجمه استفاده شده است. همچنین در آزمایش‌ها از مدل زبانی ۴-گرام استفاده شده است و در هر خانه از جدول برنامه‌ریزی پویا ۲۰ کاندیدای ترجمه بهتر را نگه داشتیم.

شکل ۱۱، امتیاز بلوی به‌دست آمده با تغییر احتمال اختصاص داده شده به ترجمه‌های فرازین نمایش داده شده است. در این آزمایش بقیه پارامترها ثابت نگه داشته شده‌اند. با توجه به این شکل اختصاص داده احتمال ۰/۳ به ترجمه‌های ارائه شده توسط فرازین، سیستم را به بهترین نتایج رسانده است.



شکل ۱۱- نمودار امتیاز بلو بر حسب احتمال ثابت نسبت داده شده به ترجمه‌های مترجم مبتنی بر قاعده



شکل ۱۲- نمودار بلو برحسب نسبت وزن مدل ترجمه به وزن مدل زبانی

همان‌طور که توضیح داده شد امتیاز نهایی هر ترجمه از ترکیب خطی لگاریتم امتیاز مدل ترجمه و لگاریتم امتیاز مدل زبانی محاسبه شده برای آن تعیین می‌شود. هریک از مدل‌های ترجمه و زبانی در این ترکیب خطی دارای وزنی است. شکل ۱۲ تأثیر تغییر این وزن‌ها روی کیفیت ترجمه نهایی برای مجموعه داده توسعه بررسی شده است. با توجه به این شکل سیستم پیشنهادی وقتی بهترین عملکرد را داشته است که وزن مدل ترجمه ۱۰ برابر وزن مدل زبانی در نظر گرفته شده است.

جدول ۴ امتیاز های بلوی دو سیستم مورد آزمایش را روی مجموعه داده‌های توسعه و آزمون را پس از پیدا کردن و اعمال وزن‌های بهینه نمایش می‌دهد. همان‌طور که در این جدول نمایش داده شده است، ترجمه‌های ارائه شده توسط سیستم پیشنهادی در این مقاله برای مجموعه داده تست، نسبت به ترجمه‌های مترجم فرازین حدود ۱/۳+ واحد بلو افزایش کیفیت داشته‌اند. برای بررسی معناداری نتایج به‌دست‌آمده از آزمون معناداری ارائه شده در [۳۰] ($P \leq 0.01$; 1000 iterations) استفاده شده است و اختلاف کیفیت‌های به‌دست آمده توسط روش پیشنهادی از نظر آماری معنادار هستند.

در جدول ۵ نیز دو نمونه از جملات مجموعه داده تست به همراه ترجمه‌های ارائه شده توسط فرازین و سیستم پیشنهادی، ارائه شده است. با توجه به این مثال‌ها می‌توان دید که انتخاب لغات در سیستم پیشنهادی بهتر از سیستم پایه است. اختلاف‌ها در این جدول با رنگ متفاوت مشخص شده‌اند.

جدول ۴- نتایج به دست آمده در آزمایش‌ها با استفاده از معیار بلو

مجموعه داده	فرازین	سیستم پیشنهادی
توسعه	۱۷/۸۹	۱۹/۵۹ (+۱/۷)
آزمون	۱۹/۶۵	۲۰/۹۳ (+۱/۲۸)

جدول ۵- نمونه‌هایی از نتایج ترجمه

جمله انگلیسی	mcguinness is expected to return to his role as northern ireland 's deputy first minister.
ترجمه فرازین	مکگاینس **احتمال داده می‌شوند** به نقش او به‌عنوان **جانشین اولین وزیر** ایرلند شمالی بر بگردند.
ترجمه سیستم پیشنهادی	مکگینس**انتظار می‌رود** به نقش او به‌عنوان **معاون نخست وزیر** ایرلند شمالی بر بگردند.
جمله انگلیسی	four afghans , including two students , were also killed , said hashmatstanikzai , spokesman for kabul 's police chief .
ترجمه فرازین	حشمت stanikzai، سخنگو برای رییس پلیس کابل گفت چهار افغان‌ها، از جمله **دو تن از دانشجویانی**، **همچنین** کشته شدند.
ترجمه سیستم پیشنهادی	حشمت stanikzai، سخنگوی رییس پلیس کابل گفت چهار افغان‌ها، از جمله **دو دانشجو**، **نیز** کشته شدند.

۶- نتیجه‌گیری و کارهای آتی

در این مقاله روشی برای غنی‌سازی ترجمه مبتنی بر قاعده پیشنهاد شد. این روش بر پایه مجموعه‌ای از قواعد نحوی- لغوی مبتنی بر گرامر درخت- پیوندی است که

با استفاده از روش‌های آماری از یک پیکره موازی استخراج شده‌اند. این قواعد احتمالاتی برای هر کلمه یا عبارت با توجه به بافت نحوی کلمه، ترجمه‌ای پیشنهاد می‌کنند. ترتیب قرارگیری کلمات در زبان مقصد در ابتدای کار توسط مترجم مبتنی بر قاعده تعیین می‌شود و ادامه کار با ثابت در نظر گرفتن این ترتیب دنبال می‌شود. این امر باعث می‌شود که بتوان عمل رمزگشایی را با استفاده از برنامه‌ریزی پویا انجام داد. نتایج به‌دست آمده از این روش به نسبت مترجم مبتنی بر قاعده پایه از کیفیت بالاتری برخوردار هستند و بهبود ۱/۳+ در واحد بلو در آزمایش‌ها ملاحظه شد. روش پیشنهادی مستقل از زبان است و در صورت وجود پیکره موازی و تجزیه‌گر نحوی مناسب قابل استفاده برای زوج زبان‌های دیگر نیز هست.

در روش ارائه شده، به ترجمه‌های پیشنهادی توسط مترجم مبتنی بر قاعده احتمال یکسانی داده شد. انتخاب روشی مناسب‌تر برای تعیین احتمال برای این ترجمه‌ها می‌تواند باعث بهبود نتایج شود. یکی از راهکارها در این زمینه می‌تواند استفاده از جدول عبارات مدل مبتنی بر عبارت برای اختصاص دادن احتمال به ترجمه‌های مترجم مبتنی بر قاعده باشد. مشکل دیگری که وجود دارد تنکی مجموعه قواعد استخراج شده است. برای کم کردن تأثیر این مشکل می‌توان در مواردی که قاعده‌ای برای کلمه مبدأ در بافت نحوی موردنظر پیدا نمی‌شود به مدلی بدون در نظر گرفتن بافت نحوی عقب‌گرد[۲۹] کرد.

قدردانی

پژوهشی که نتایج آن در این مقاله ارائه شده است در قالب یک طرح تحقیقاتی مصوب و با حمایت مالی صندوق حمایت از پژوهشگران و فناوران کشور انجام شده است.

مراجع

[1] M. R. Costa-jussàa, M. Farr´us, J. B. Mari˜no, and J. A.R. Fonollosa, "*Study And Comparison of Rule-Based And Statistical Catalan-Spanish Machine Translation Systems,"* Computing and Informatics, vol. 31, no. 2, pp. 245-270, 2012.

[2] A. K. Joshi, L. S. Levy, and M. Takahashi, *"Tree Adjunct Grammars,"* Journal of Computer and System Sciences, vol. 10, no. 1, pp. 136–163, 1975.

[3] فرازین: مترجم خودکار متون انگلیسی به فارسی, [Online]. Available: www.faraazin.ir.

[4] M. R. Costa-jussàa, and J. A. R. Fonollosa, *"Latest trends in hybrid machine translation and its applications,"* Computer Speech & Language, vol. 32, no. 1, pp. 3-10, July 2015.

[5] A. Bisazza, and M. Federico, *"A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena,"* Computational Linguistics, vol. 42, no. 2, pp. 163–205, 2016.

[6] M. Collins, P. Koehn, and I. Kucerova, *"Clause Restructuring for Statistical Machine Translation,"* In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 531–540, 2005.

[7] R. N. Patel, R. Gupta, P. B. Pimpale, and M. Sasikumar, *"Reordering rules for English-Hindi SMT,"* In Proceedings of the 2nd Workshop on Hybrid Approaches to Translation (HyTra), pp. 34-41, 2013.

[8] F. Xia, and M. McCord, *"Improving a Statistical MT System with Automatically Learned Rewrite Patterns,"* In Proceedings of the 20th international conference on Computational Linguistics, pp. 508, 2004.

[9] A. Mansouri, H. Fadaei, H. Faili, and M. Arabsorkhi, "*Using Synchronous TAG for Source-Side Reordering in SMT,"* International Journal of Information & Communication Technology Research, vol. 5, no. 4, pp. 47-58, Autumn 2013.

[10] A. Eisele, C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. "*Using Moses to integrate multiple rule-based machine translation engines into a hybrid system,"* In Proceedings of the 3rd Workshop on Statistical Machine Translation (WMT), pp. 179–182, 2008.

[11] A. Ahsan, P. Kolachina, S. Kolachina, D. Misra Sharma, and R. Sangal, *"Coupling statistical machine translation with rule-based transfer and generation,"* In Proceedings of the 9th Conference of the Association for Machine Translation in the Americas. 2010.

[12] V. M. S´anchez-Cartagena, J. A. P´erez-Ortiz, and F. S´anchez-Mart´ınez, *"Integrating Rules and Dictionaries from Shallow-Transfer Machine Translation into Phrase-Based Statistical Machine Translation,"* Journal of Artificial Intelligence Research, vol. 55, pp. 17-61, 2016.

[13] W. Ma, and K. McKeown, *"Detecting and Correcting Syntactic Errors in Machine Translation Using Feature-based Lexicalized Tree Adjoining Grammars,"* Computational Linguistics and Chinese Language Processing, vol. 17, no. 4, pp. 1-14, December 2012.

[14] A. L. Lagarda, V. Alabau, F. Casacuberta, R. Silva, and E. Diaz-de Liano, *"Statistical post-editing of a rule-based machine translation system,"* In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 217-220, January 2009.

[15] A. Göhring, *"Building a Spanish-German dictionary for hybrid MT,"* The 3rd Workshop on Hybrid Approaches to MachineTranslation (HyTra), pp. 30–35, April 2014.

[16] A. Antonova, and A. Misyurev, *"Improving the precision of automatically constructed human-oriented translation dictionaries,"* In Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra), pp. 58–66, April 2014.

[17] L. Shen, J. Xu, and R. Weischedel, *"A New String-to-dependency Machine Translation Algorithm with a Target Dependency Language Model,"* In Proceedings of The

Association for Computational Linguistics, pp. 577-585, 2008.

[18] M. Galley, M. Hopkins, K. Knight, and D. Marcu, *"What's in a Translation Rule,"* In Proceedings of The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL- HLT), Boston, Massachusetts, USA, pp. 273-280, 2004.

[19] L. Huang, K. Knight, and A. Joshi, *"Statistical Syntax-directed Translation with Extended Domain of Locality,"* In Proceedings of AMTA, pp. 66-73, 2006.

[20] S. DeNeefe, *"Tree-adjoining Machine Translation,"* PhD Thesis, Faculty of the USC graduate school University of Southern California, 2011.

[21] S. DeNeefe, K. Knight, W. Wang, and D. Marcu, "*What Can Syntax-based MT Learn from Phrase-based MT?,"* In Proceedings of EMNLP-CoNLL, pp. 755-763, 2007.

[22] D. Klein, and Ch. D. Manning, *"Accurate Unlexicalized Parsing,"* In Proceeding of the 40th Annual meeting of the Association for Computational Linguistics, vol.1, pp. 423-430, 2003.

[23] J. Chen, and K. Vijay-Shanker, *"Automated Extraction of TAGs from the Penn Treebank,"* In Proceedings of the Sixth International Workshop on Parsing Technologies, pp. 73-89, 2000.

[24] F. J. Och, and H. Ney, *"Improved Statistical Alignment Models,"* In Proceedings ofthe 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, 2000.

[25] Y. Liu, Q. Liu, and Y. Lu, *"Adjoining Tree-to-String Translation,"* In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 1278–1287, 2011.

[26] P. Koehn, *"Statistical Machine Translation,"* Cambridge University Press, 2010.

[27] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang and I. Thayer, *"Scalable Inference and Training of Context-Rich Syntactic Translation Models,"* In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 961-968, 2006.

[28] F. Jabbari, S. Bakhshaei, S. M. MohammadzadehZiabary, and S. Khadivi, *"Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus,"* In Proceedings of the fourth Workshop on Computational Approaches to Arabic Script-based Languages, pp. 17, 2012.

[29] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "*BLEU: a method for automatic evaluation of machine translation,"* In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics, pp.311–318, 2002.

[30] P. Koehn, *"Statistical Significance Tests for Machine Translation Evaluation,"* In Proceedings of EMNLP. pp. 388–395, 2004.

حکیمه فدایی دانشجوی مقطع دکتری رشته مهندسی نرم‌افزار در دانشگاه تهران است. همچنین وی مدرک کارشناسی و کارشناسی‌ارشد خود را در رشته مهندسی نرم‌افزار به ترتیب در سال‌های ۱۳۸۵ و ۱۳۸۸ از دانشگاه شهید بهشتی کسب کرده است. زمینه پژوهشی وی پردازش زبان طبیعی و به‌طور خاص ترجمه ماشینی است.
آدرس پست‌الکترونیکی ایشان عبارت است از:

h.fadaei@ut.ac.ir

هشام فیلی تحصیلات خود را در مقطع کارشناسی مهندسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شریف با رتبهٔ یک در سال ۱۳۷۶ به پایان رساند؛ سپس مقاطع کارشناسی‌ارشد نرم‌افزار و دکترای هوش مصنوعی را به ترتیب در سال‌های ۱۳۷۸ و ۱۳۸۵ در همان دانشکده تکمیل کرد. از سال ۱۳۸۷ تاکنون عضو هیأت علمی دانشکدهٔ مهندسی برق و کامپیوتر دانشکدهٔ فنی دانشگاه تهران است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش هوشمند متن و زبان طبیعی، ترجمهٔ ماشینی، داده‌کاوی، بازیابی اطلاعات و شبکه‌های اجتماعی هستند.
آدرس پست‌الکترونیکی ایشان عبارت است از:

hfaili@ut.ac.ir

فرناز قاسمی دوره کارشناسی خود را در سال ۱۳۹۵ در رشته مهندسی فناوری اطلاعات در دانشگاه تهران به پایان رساند. او در حال حاضر دانشجوی کارشناسی‌ارشد دانشگاه تهران در گرایش سامانه‌های شبکه‌ای است. زمینه‌ی پژوهشی مورد علاقه وی پردازش زبان طبیعی است.
آدرس پست‌الکترونیکی ایشان عبارت است از:

f.ghasemi.91@ut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۰/۲۷
تاریخ اصلاح: ۱۳۹۵/۱۱/۱۵
تاریخ قبول شدن: ۱۳۹۵/۱۱/۲۵
نویسنده مرتبط: دکتر هشام فیلی، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران.

[1] Rule-Based Machine Translation
[2] Tree Adjoining Grammar
[3] Dynamic Programing
[4] Hybrid Approaches
[5] Monotone
[6] Extended Domain of Locality
[7] Tree Substitution Grammar
[8] Substitution
[9] Siblings

[10]Context Sensitive
[11]Head
[12]Mildly Context Sensitive
[13]Adjunction
[14]Optional
[15]Elementary Tree
[16]Lexicalized Tree Adjoining Grammar
[17]Anchor
[18]Foot Node
[19]Adjunction Site
[20]Sparseness
[21]Derived Tree
[22]Derivation Tree
[23]Decoder
[24]Consistent
[25]Extractable
[26]Composed Rules
[27]Maximum Likelihood Estimation
[28]BLEU
[29]Back-Off