نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

استفاده از بازآرایی نحوی جهت بهبود ترجمه ماشینی آماری انگلیسی به فارسی

رضا سعیدی[۱]، محسن کاهانی[۲]، سیداحمد جکیان طوسی[۳] و بهداد بهمدی مقدس[۴]
reza.saeedi@stu-mail.um.ac.ir[۱]
mohesn.kahani@um.ac.ir [۲]
ahmad.toosi@stu-mail.um.ac.ir [۳]
behdad.behmadi@stu-mail.um.ac.ir [۴]
آزمایشگاه فناوری وب دانشگاه فردوسی مشهد

چکیده

ترجمه ماشینی آماری به عنوان یکی از بهترین روش ها برای ترجمه از یک زبان به زبان دیگر شناخته می‌شود. برای زبان‌هایی که از لحاظ ساختار دارای شباهت زیادی به یکدیگر هستند خروجی این مترجم بسیار مناسب می‌باشد. تفاوت‌های ساختاری میان زبان انگلیسی و فارسی و همچنین عدم وجود پیکره دوزبانه بزرگ باعث شده است که این روش برای زبان ترجمه انگلیسی به فارسی ترجمه های مطلوبی را تولید نکند. ما در این مقاله سعی کرده‌ایم با استفاده از رهیافت بازآرایی کلمات، تا حد ممکن شباهت ساختاری میان عبارات انگلیسی و فارسی را افزایش دهیم. در ادامه تاثیر این عمل را بر روی بهبود نتایج خروجی مورد بررسی قرار داده‌ایم. به همین منظور ابتدا با کمك درخت تجزیه، مجموعه‌ای از قوانین بازآرایی استخراج شده است. سپس این قوانین به عنوان یك عمل پیش پردازشی برروی عبارات انگلیسی اعمال گردیده است. نتایج بررسی ها نشان می‌دهد که خروجی مترجم پس از اعمال این روش منجر به بهبود کیفیت ترجمه در معیار BLEU شده است.

کلمات کلیدی

ترجمه ماشینی آماری، بازآرایی نحوی، مدل زبانی، مدل ترجمه، درخت تجزیه

۱- مقدمه

ترجمه ماشینی به معنای استفاده از کامپیوتر برای ترجمه از یک زبان به زبان دیگر به صورت خودکار می‌باشد[۱]. تفاوت ذاتی میان برخی از زبانها کار ترجمه را بسیار مشکل می‌سازد. روشهای سنتی برای ترجمه ماشینی، برروی اطلاعات زبان شناختی بشر، بصورت قوانین تبدیل متن از زبانی به زبان دیگر تکیه داشت. با توجه به وسعت دامنه اطلاعات موجود در زبان ها، استفاده از این روشها، بسیار پیچیده و نیازمند دانش بسیار می‌باشد. ترجمه ماشینی آماری، یک رویکرد متفاوت است که سعی دارد بطور خودکار از روی حجم زیادی از داده های آموزشی، به دانش گفته شده در ترجمه، دست پیدا کند. این دانش که عموما بصورت احتمالاتی مبتنی بر قابلیتهای گوناگون زبانی می‌باشد و به منظور راهنمایی در فرایند ترجمه مورد استفاده قرار می‌گیرد.

از این رو ترجمه ماشینی آماری می‌تواند به عنوان یک رویه به منظور یافتن جمله ای با بالاترین احتمال در زبان مقصد تعریف شود. بر اساس رابطه (۱) با ورود یک جمله در زبان مبدا (e) شبیه ترین ترجمه آن در زبان مقصد(f) یافت می‌شود[۲].

$$e' = \arg\max_e P(e|f) \qquad (۱)$$

با توجه به رابطه (۲) $P(e|f)$ مبتنی بر دو فاکتور است. میزان احتمال بودن یک جمله در زبان e، که این فاکتور بعنوان مدل زبانی($P(e)$) شناخته می‌شود و روشی که جملات موجود در زبان e به جملات موجود در زبان f تبدیل می‌شوند که این فاکتور نیز بعنوان مدل ترجمه ($P(f|e)$) شناخته می‌شود.

$$e' = \arg\max_e \frac{P(e)P(f|e)}{P(f)} \qquad (۲)$$

همانطورکه در رابطه (۳) نشان داده شده است، این دو فاکتور از اعمال قانون بیز بر روی معادله اولیه حاصل می‌گردند. از آنجایی که مقدار $P(f)$ همواره ثابت است، آن را از بیشینه‌سازی حذف می‌کنیم. لذا داریم:

$$e' = \arg\max_e P(e)P(f|e) \qquad (۳)$$