مدل ترجمهٔ عبارت-مرزی با استفاده از برچسب‌های کم‌عمق نحوی

شهرام سلامی* **و مهرنوش شمس‌فرد**

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

چکیده

مدل عبارت-مرزی برای ترجمه ماشینی آماری، قواعد را با طبقهٔ کلمات مرزی عبارات پیکره مقصد برچسب می‌زند. در این مقاله مدل عبارت-مرزی را با استفاده از برچسب‌های کم‌عمق نحوی شامل برچسب POS و برچسب قطعات توسعه می‌دهیم. با اولویت برچسب قطعات، مدل پیشنهادی، غیرپایانه‌ها را با برچسب‌های کم‌عمق نحوی در مرز عبارات مقصد نام‌گذاری می‌کند. در قیاس با مدل SAMT که قواعد را با درخت تجزیه نحوی جملات مقصد برچسب می‌زند، مدل پیشنهادی به تجزیه عمیق نحوی نیاز ندارد. همچنین، هرچه تفاوت ترتیب کلمات زبان مبداء و مقصد ترجمه بیشتر باشد، عبارات ترازشده قابل انطباق با درخت تجزیه نحوی، کمتر خواهد بود. تعدادی آزمایش در ترجمه از فارسی و آلمانی به انگلیسی به‌عنوان جفت‌زبان‌هایی با تفاوت زیاد در ترتیب کلمات انجام شد. در این آزمایش‌ها، مدل عبارت-مرزی پیشنهادی نسبت به مدل SAMT در حدود ۰/۵ واحد BLEU کیفیت ترجمه بهتری به‌دست آورد.

واژگان کلیدی: ترجمه ماشینی آماری، مدل سلسله‌مراتبی، برچسب کلمه، برچسب قطعه

Phrase-Boundary Translation Model Using Shallow Syntactic Labels

Shahram Salami* & Mehrnoush Shamsfard

Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

Abstract

Phrase-boundary model for statistical machine translation labels the rules with classes of boundary words on the target side phrases of training corpus. In this paper, we extend the phrase-boundary model using shallow syntactic labels including POS tags and chunk labels. With the priority of chunk labels, the proposed model names non-terminals with shallow syntactic labels on the boundaries of the target side phrases. In comparison to the base phrase-boundary model, our variant uses phrase labels in addition to word classes. In other words, if there is no chunk label in one boundary, the labeler uses the word POS tag. The boundary labels are concatenated where there is no label for the whole target span. Using chunks as phrase labels, the proposed model generalizes the rules to decrease the model sparseness. The sparseness has more importance in the language pairs with a lot of differences in the word order because they have less number of aligned phrase pairs for extraction of rules. Compared with Syntax Augmented Machine Translation (SAMT) that labels rules with the syntax trees of the target side sentences, the proposed model does not need deep syntactic parsing. Thus, it is applicable even for low-resource languages having no syntactic parser. Some translation experiments are performed from Persian and German to English as the source and target languages with different word orders. In the experiments, our model achieved improvements of about 0.5 point of BLEU over a variant of SAMT.

Keywords: Statistical machine translation, Hierarchical models, Word tag, Chunk label

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

۱- مقدمه

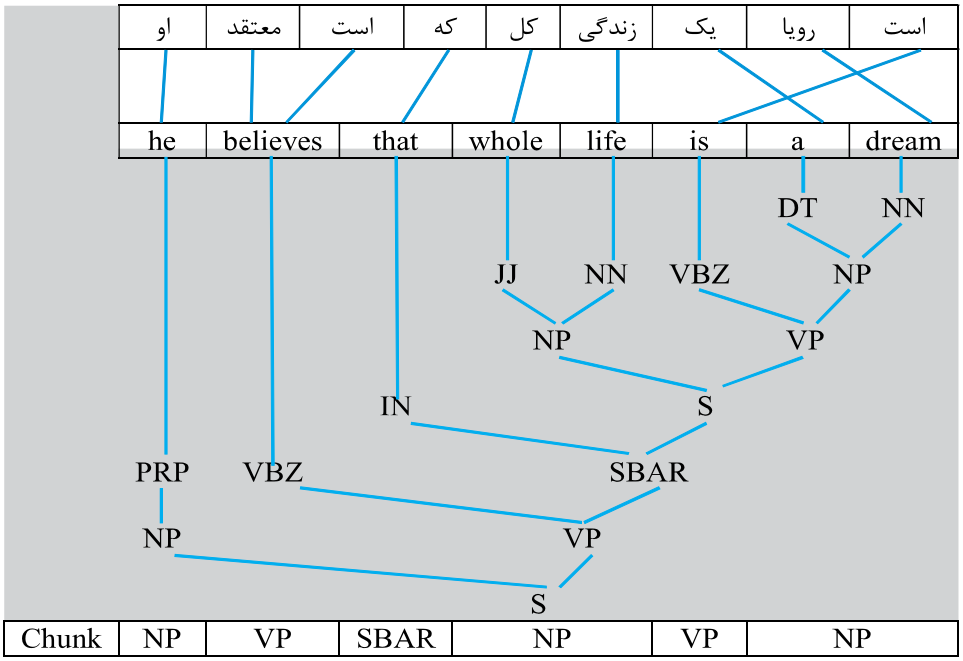
مدل‌های سلسله‌مراتبی با استفاده از عبارات تودرتو از بازترتیب‌های غیر محلی کلمات حمایت می‌کنند. برمبنای نوع برچسب‌های استفاده شده برای عبارات، مدل‌های سلسله‌مراتبی مختلفی پیشنهاد شده است. مدل مبتنی بر عبارت سلسله‌مراتبی [2] از یک برچسب عمومی برای تمام غیر پایانه‌ها استفاده می‌کند. SAMT[1]، یک مدل به‌خوبی شناخته شده است [3] که از طبقه‌های نحوی زبان مقصد برای برچسب قواعد استفاده می‌کند. با توجه به این که رمزگشایی به‌وسیلهٔ ورودی ترجمه هدایت می‌شود، استفاده از طبقه‌های نحوی عبارات مقصد، امکان ساخت نحوی خروجی ترجمه را فراهم می‌کند. در این مدل برچسب از تطابق سمت مقصد عبارات ترازشده با زیردرخت‌های درخت تجزیهٔ جمله در پیکره مقصد به‌دست می‌آید. عباراتی که با بازه‌ای در درخت تجزیه تطابق نداشته باشند با یک غیر پایانه پیش‌فرض برچسب‌گذاری می‌شوند. هرچه تفاوت ترتیب کلمات زبان مبداء و مقصد ترجمه بیشتر باشد، تعداد عبارات ترازشده منطبق با درخت تجزیهٔ نحوی جملات پیکره مقصد کمتر خواهد بود. اگر چه برخی تطابق‌های نسبی با درخت تجزیه نحوی در مدل SAMT پیش‌بینی شده است، کارایی این مدل در ترجمه جفت‌زبان‌هایی با تفاوت زیاد در ترتیب کلمات کاسته می‌شود. در زیر مثالی از قواعد SAMT برای تراز جملات فارسی-انگلیسی شکل (۱) آورده شده است (برای سادگی نمایش، کلمات فارسی از چپ به راست نمایش داده شده است.):

X → <کل که است معتقد او ; he believes that whole> (۱)

S/VP → <زندگی کل که است معتقد او ; he believes that whole life> (۲)

سمت چپ قاعده ۱ با غیر پایانهٔ پیش‌فرض X تعریف شده است؛ زیرا این عبارت در درخت تجزیه نحوی متناظری ندارد. سمت چپ قاعده ۲ یک تطابق نسبی در درخت تجزیه را نشان می‌دهد. برچسب سمت چپ قاعده با یک S فاقد VP (جمله فاقد عبارت فعلی) بیان شده است. این مثال نشان می‌دهد که در عباراتی با تراز یک‌نوای کلمات نیز امکان عدم انطباق با یک بازهٔ نحوی در درخت تجزیه وجود دارد.

مدل عبارت-مرزی [4] از طبقهٔ کلمات برای برچسب قواعد استفاده می‌کند. برچسب قواعد در این مدل با اتصال کلاس کلمات مرزی سمت مقصد عبارات تراز شده تعریف



(شکل-۱): تراز جملات فارسی-انگلیسی به همراه درخت تجزیه نحوی و برچسب قطعات کلمات انگلیسی

(جمله فارسی از چپ به راست نوشته شده است.)

(Figure-1): Alignment of Persian-English sentences along with syntax tree and chunk labels of English words

[1] Syntax Augmented Machine Translation

می‌شود. برای مثال، قواعد زیر (معادل قواعد ۱ و ۲) با استفاده از برچسب POS کلمات در مدل عبارت-مرزی برای تراز جملات فارسی-انگلیسی در شکل (۱) استخراج شده است:

PRP-JJ → <کل که است معتقد او ;
he believes that whole> (۳)

PRP-NN → <زندگی کل که است معتقد او ;
he believes that whole life> (۴)

همان‌طورکه در این مثال می‌بینید، مدل عبارت-مرزی در مقایسه با SAMT، بدون نیاز به برچسب پیش فرض، تمام عبارت ترازشده را پوشش می‌دهد؛ از سوی دیگر، با استفاده از برچسب POS به‌عنوان طبقهٔ کلمات، مدل عبارت-مرزی تنها از اطلاعات برگ‌ها در درخت تجزیه نحوی بهره می‌برد. با این حال، کارایی آن در آزمایش‌ها با مدل SAMT مشابه است.

در این مقاله، میزان دانش نحوی استفاده‌شده در مدل عبارت-مرزی را افزایش می‌دهیم. مدل عبارت-مرزی توسعه‌یافته، علاوه‌بر برچسب POS کلمات از برچسب قطعات[1] نیز برای نام‌گذاری غیرپایانه‌ها استفاده می‌کند. برچسب قطعات حاصل تجزیه کم‌عمق نحوی[2] (قطعه‌بندی) است که اجزای جمله (مانند عبارت اسمی و فعلی) را به‌صورت یکتا تعیین می‌کند. البته ساختار داخلی و نقش قطعات در جمله مشخص نمی‌شود. در مدل پیشنهادی، اولویت نام‌گذاری با برچسب قطعه است. اگر برچسب قطعه برای کل عبارت و یا در مرز عبارت وجود نداشته باشد، از برچسب POS کلمه مرزی استفاده می‌شود. نام‌گذاری غیرپایانه‌ها در این مدل سرراست است؛ زیرا برچسب POS در مرز همه عبارات وجود دارد. در قواعد زیر، دو قاعده ۳ و ۴ با استفاده از برچسب عبارات اسمی (NP) در مدل پیشنهادی تعمیم یافته است:

NP-JJ → <کل که است معتقد او ;
he believes that whole> (۵)

NP-NP → <زندگی کل که است معتقد او ;
he believes that whole life> (۶)

در مقایسه با شکل مبنای مدل عبارت-مرزی، قواعد مدل پیشنهادی با استفاده از برچسب عبارات به جای برچسب کلمات تعمیم‌یافته است. تعمیم قواعد تنکی مدل ترجمه را کاهش می‌دهد. در مقایسه با SAMT، مدل پیشنهادی ضمن استفاده از اطلاعات نحوی، تمام عبارات ترازشده را پوشش می‌دهد و به برچسب پیش‌فرض نیاز ندارد؛ به‌خصوص در جفت‌زبان‌هایی با تفاوت زیاد در ترتیب کلمات، عبارات ترازشده منطبق با درخت تجزیهٔ نحوی کمتر خواهد بود.

با استفاده از مدل مبتنی بر عبارت سلسله‌مراتبی، گونه‌هایی از مدل SAMT و مدل عبارت-مرزی آزمایش‌هایی در ترجمه از فارسی و آلمانی به انگلیسی به‌عنوان جفت‌زبان‌هایی با تفاوت زیاد در ترتیب کلمات انجام می‌دهیم. اگرچه، کیفیت ترجمه در مدل عبارت مرزی و SAMT نسبت به مدل مبتنی بر عبارت سلسله‌مراتبی بالاتر است، به‌دلیل تنوع برچسب‌ها، زمان ترجمه و اندازه این مدل‌ها بسیار بزرگتر است. راه حل شناخته‌شده برای رفع این مشکل، فیلتر قواعد است. فیلتر یک‌نوا[3] برای کاهش زمان ترجمه و اندازه شکل مبنای مدل عبارت-مرزی مطرح شده است. ما کارایی این فیلتر را برای مدل‌های سلسله‌مراتبی شامل مدل پیشنهادی بررسی و برای کاهش زمان ترجمه، مدل‌ها را با اعمال فیلتر نیز آزمایش می‌کنیم. همان‌طور که خواهیم دید، فیلتر یک‌نوا حافظه و زمان لازم برای آموزش و رمزگشایی مدل‌های سلسله‌مراتبی را به میزان قابل توجهی کاهش می‌دهد. دیده می‌شود که بر اساس معیار BLEU، کیفیت ترجمه مدل عبارت-مرزی پیشنهادی به‌خصوص در حالت فیلترشده از SAMT بالاتر است.

در ادامه، برخی کارهای مرتبط در بخش ۲ معرفی و در بخش ۳ مدل عبارت-مرزی توسعه یافته با استفاده از برچسب‌های کم‌عمق نحوی تعریف می‌شود. بخش ۴ فیلتر یک‌نوا را برای مدل‌های سلسله‌مراتبی پیشنهاد و بخش ۵ آزمایش‌های انجام‌شده را تشریح می‌کند. سرانجام، مقاله در بخش ۶ نتیجه گیری می‌شود.

۲- کارهای مرتبط

مدل مبتنی بر عبارت سلسله‌مراتبی [2] برای استخراج دستور[4] بدون نظارت بر مبنای تراز عبارات با یک غیرپایانه عمومی معرفی شد. تولید ناپیوسته کلمات مقصد، هرس فضای رمزگشایی با مدل زبانی مقصد را محدود می‌کند. با محدودکردن قواعد ترجمه به شکل GNF جمله مقصد از چپ به راست تولید شد [5]. در کاری دیگر [6]، با اجتناب از فرم بازگشتی قواعد سلسله‌مراتبی فضای رمزگشایی مدل مبتنی برعبارت سلسله‌مراتبی محدود شد. در این کار از دو غیرپایانه مختلف در سمت چپ و راست قواعد سلسله‌مراتبی استفاده و

[1] Chunk label
[2] Shallow syntactic parsing
[3] Monotonic filter
[4] Grammar

همچنین، بدون استفاده از منابع زبانی از الگوی تجزیه عبارات برای برچسب گذاری قواعد استفاده شد [7].

برای انتخاب بهتر قواعد در فرایند رمزگشایی مدل مبتنی برعبارت سلسله‌مراتبی، اطلاعات بافتار ورودی به شکل برچسب POS [8] و برچسب CCG [9] استفاده شده است. همچنین، با استفاده از دانش نحوی، اشتقاق‌ها در زمان رمزگشایی امتیازدهی شدند [10].

برای افزایش دقت قواعد، برچسب قواعد در مدل سلسله‌مراتبی با طبقه‌های درخت تجزیه نحوی با عنوان مدل SAMT [3] و طبقه‌های CCG [11] تعریف و برای کاهش تعداد قواعد در SAMT برچسب‌های نحوی خوشه‌بندی شدند [12]. همچنین، غیرپایانه‌های مدل سلسله‌مراتبی با برچسب POS کلمات سرآیند[1] نام‌گذاری شد [13].

طبقهٔ کلمات مرزی عبارات در برخی کارهای جدید استفاده شده است. برای بهبود بازترتیب کلمات در مدل مبتنی بر عبارت سلسله‌مراتبی [6] و مبتنی برعبارت [14] از طبقهٔ کلمات مرزی عبارات استفاده شد. با برچسب‌گذاری قواعد با استفاده از طبقهٔ کلمات مرزی عبارات، کیفیت ترجمه مشابه مدل SAMT به‌دست آمد [15]. برچسب‌گذاری قواعد با طبقهٔ کلمات مرزی عبارات همراه با استخراج فیلترشده قواعد برای کاهش اندازه مدل و زمان رمزگشایی در مدل عبارت–مرزی [4] پیشنهاد شد. مدل پیشنهادشده در این مقاله برچسب‌گذاری مرزی را به برچسب POS و قطعه برای کیفیت بهتر ترجمه تعمیم می‌دهد.

برای کاهش اندازه مدل و زمان ترجمه در مدل‌های سلسله‌مراتبی تعدادی روش برای فیلتر قواعد پیشنهاد شده است که در دو رویکرد قابل دسته‌بندی هستند. یک رویکرد قواعد غیر ضروری را از دستور استخراج‌شده حذف می‌کند. رویکرد دیگر روش استخراج قواعد را تغییر می‌دهد تا از استخراج قواعد غیر ضروری پیشگیری نماید. این رویکرد علاوه بر زمان و حافظه رمزگشایی، منابع لازم برای استخراج دستور را نیز کاهش می‌دهد.

در رویکرد حذف قواعد غیر ضروری از دستور استخراج‌شده، قواعدی حذف شدند که رخداد آنها از یک حد آستانه‌ای کمتر باشد [16]. افزایش حد آستانه کیفیت ترجمه را کاهش می‌دهد. برای فیلتر قواعد مدل مبتنی بر عبارت سلسله‌مراتبی، قواعدی که سمت مبداء آنها تنها در قواعدی با تراز یک‌نوا دیده شده باشد، حذف شدند [17]. لازم به ذکر است که قواعد چسب در دستور از اتصال یک‌نوای عبارات حمایت می‌کند. در روشی دیگر [18]، قواعد به الگوهای مختلف دسته‌بندی شدند و الگوهایی حذف شدند که حذف آنها تأثیر قابل توجهی بر کیفیت ترجمه نداشته باشد. همچنین، قواعد ترجمه برمبنای افزونگی اطلاعات موجود در آنها حذف شدند [19].

در رویکرد تغییر روش استخراج قواعد برای کاهش قواعد استخراج‌شده، استخراج قواعد به عباراتی محدود شد که احتمال تراز کلمات آنها بالاتر است [20]. در روش دیگر [21]، مجموعه کمینه‌ای از قواعد ترجمه استخراج شد که در این مجموعه برای هر جفت عبارت دست‌کم یک اشتقاق امکان‌پذیر است. فیلتر یک‌نوا [4] برای شکل مبنای مدل عبارت–مرزی بر مبنای الگوی تراز عبارات پیشنهاد شد. این فیلتر قواعد استخراج‌شده از عبارات تجزیه‌پذیر به زیررشته‌هایی با تراز یک‌نوا را کاهش می‌دهد. در این مقاله کاربرد فیلتر یک‌نوا را برای مدل پیشنهادی و سایر مدل‌های سلسله‌مراتبی تعمیم می‌دهیم.

۳- مدل

مدل پیشنهادی یک دستور همگام مستقل از متن را از عبارات ترازشده استخراج می‌کند. قواعد وزن‌دار به شکل واژگانی، سلسله‌مراتبی و چسب تعریف می‌شوند. قواعد واژگانی بیان‌گر عبارات ترازشده بدون غیرپایانه در سمت راست هستند. قواعد سلسله‌مراتبی با بیشینه دو جایگذاری زیرعبارات با غیرپایانه‌ها تعریف می‌شوند. قواعد چسب برای همه غیرپایانه‌های دستور جهت اتصال متوالی عبارات خروجی تعریف می‌شوند.

مدل عبارت–مرزی در شکل مبنا [4] به‌صورت یک‌نواخت، طبقهٔ کلمات مرزی عبارات مقصد را با یک خط پیوند برای نام‌گذاری غیرپایانه‌ها اتصال می‌دهد. با استفاده از برچسب POS به‌عنوان طبقهٔ کلمات، برچسب عبارت ترازشده $<f_i^j , e_m^n>$ (که f_i^j و e_m^n به‌ترتیب بیان‌گر زیررشته بسته از موقعیت i تا j و موقعیت m تا n است) به شکل زیر تعریف می‌شود:

$$X_{m,n} = \begin{cases} POS(m) & if\ m = n \\ POS(m) - POS(n) & else \end{cases} \qquad (۷)$$

برای مثال، قواعد زیر با استفاده از برچسب POS در شکل مبنای مدل عبارت–مرزی برای تراز جملات شکل (۲) تعریف شده است (برای سادگی، کلمات فارسی از چپ به راست نمایش داده شده است.):

[1] Head words

تعریف ۱

برچسب غیر پایانه مربوط به عبارت ترازشده $<f_i^j , e_m^n>$ که در آن عبارت مقصد با e_m شروع و با e_n خاتمه می‌یابد، به شکل زیر تعریف می‌شود:

$$X_{m,n} = \begin{cases} Chunk\ (e_m^n) & \text{اگر } e_m^n \text{ با یک برچسب قطعه منطبق باشد} \\ \text{POS}\ (e_m) & \text{در غیر این صورت اگر } m = n \text{ باشد} \\ Left\text{-}Label\ -\ Right\text{-}Label & \text{در غیر این صورت} \end{cases}$$

$$Left\text{-}Label = \begin{cases} Chunk\ (e_m^a) : m \le a < n & \text{اگر } e_m^a \text{ با یک برچسب قطعه منطبق باشد} \\ POS\ (e_m) & \text{در غیر این صورت} \end{cases}$$

$$Right\text{-}Label = \begin{cases} Chunk\ (e_b^n) : m < b \le n & \text{اگر } e_b^n \text{ با یک برچسب قطعه منطبق باشد} \\ POS\ (e_n) & \text{در غیر این صورت} \end{cases}$$

PRP → < من ; i > (۸)

DT-NN → < DT^{-1} آبی قلم; DT^{-1} blue pen> (۹)

DT-NN → <یک NN^{-1} آبی ; a blue NN^{-1}> (۱۰)

PRP-NN → < PRP^{-1} DT-NN^{-2} دادم پسر به ;
PRP^{-1} gave the boy DT-NN^{-2} > (۱۱)

مدل عبارت-مرزی پیشنهادی علاوه‌بر برچسب POS از برچسب قطعه در مرز عبارات مقصد نیز استفاده می‌کند. وقتی یک قطعه تمام عبارت ترازشده را پوشش دهد، به‌عنوان برچسب استفاده خواهد شد. در غیر این صورت قطعه واقع در عبارت مقصد که از سمت چپ شروع شود یا در سمت راست آن پایان یابد برای محاسبه برچسب استفاده می‌شود. درآخر، در حالتی که برچسب قطعه در مرز عبارت یا تمام عبارت به طول یک موجود نباشد، برچسب POS استفاده خواهد شد. تعریف ۱ برچسب عبارات را به‌صورت رسمی نشان می‌دهد.

برچسب‌های حاصل از *Left-Label* و *Right-Label* با یک خط پیوند متصل می‌شوند. تابع $Chunk\ (e_x^y)$ برچسب قطعه‌ای را که از موقعیت x شروع و در موقعیت y پایان می‌یابد، باز می‌گرداند. با توجه به عدم هم‌پوشانی قطعه‌ها، در این تعریف $a < b$ است.

برای مثال، قواعد زیر که از تراز جملات شکل (۲) استخراج شده‌اند، معادل قواعد ۸ تا ۱۱ هستند:

NP → < من ; i > (۱۲)

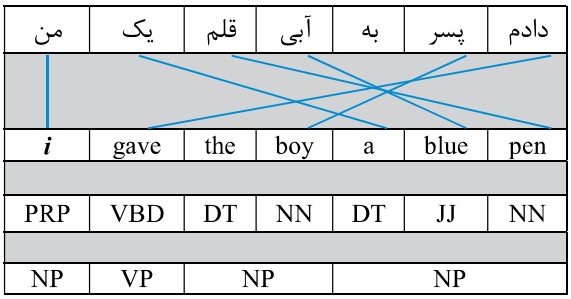
NP → < DT^{-1} آبی قلم ; DT^{-1} blue pen > (۱۳)

NP → < یک NN^{-1} آبی ; a blue NN^{-1}> (۱۴)

NP-NP → < NP^{-1} NP^{-2} دادم پسر به ;
NP^{-1} gave the boy NP^{-2} > (۱۵)

مقایسه این قواعد با قواعد ۸ تا ۱۱ نشان می‌دهد که استفاده از برچسب عبارات به جای برچسب کلمات تنکی مدل ترجمه را کاهش می‌دهد. برای مثال، فاعل در قاعده ۱۱ باید یک ضمیر باشد؛ ولی قاعده ۱۵ هر نوع عبارت اسمی (NP) را به‌عنوان فاعل می‌پذیرد. به‌طور خلاصه مدل پیشنهادی دارای مزایای زیر است:

- در مقایسه با SAMT، ضمن استفاده از برچسب نحوی عبارات، تمام عبارات ترازشده را پوشش می‌دهد.
- در مقایسه با شکل مبنای مدل عبارت-مرزی، استفاده از برچسب عبارات به جای برچسب کلمات، تنکی مدل ترجمه را کاهش می‌دهد.



(شکل-۲): تراز جملات فارسی-انگلیسی به همراه برچسب POS و قطعه کلمات انگلیسی

(جمله فارسی از چپ به راست نوشته شده است.)

(Figure-2): Alignment of Persian-English sentences along with POS and chunk labels of English words

اگر چه امکان تعریف قواعد با برچسب‌های مبداء پیکره موازی نیز وجود دارد، مدل پیشنهادی مانند شکل مبنای مدل عبارت-مرزی با استفاده از برچسب‌های سمت مقصد عبارات ترازشده تعریف شد. با توجه به این که رمزگشایی به وسیله ورودی ترجمه هدایت می‌شود، استفاده از برچسب‌های نحوی عبارات مقصد، امکان ساخت نحوی خروجی ترجمه را فراهم می‌کند. در کارهای دیگری [3]،[15] نیز از برچسب‌های سمت مقصد پیکره موازی استفاده شده و برتری این رویکرد نسبت به استفاده از برچسب‌های مبداء نشان داده شده است.

۴- فیلتر یک‌نوا

با استفاده از حداکثر دو غیرپایانه در سمت راست قواعد سلسله‌مراتبی، بخش مبداء یا مقصد قواعد با یکی از الگوهای زیر قابل تعریف است (w_i و x_i به‌ترتیب یک رشته از کلمات و یک غیرپایانه هستند.):

(۱۶)

$$\textit{all-patterns} = \{w_1, x_1 x_2, x_1 w_1, w_1 x_1, w_1 x_1 w_2, x_1 w_1 x_2, x_1 x_2 w_1, w_1 x_1 x_2, w_1 x_1 x_2 w_2, x_1 w_1 x_2 w_2, w_1 x_1 w_2 x_2, w_1 x_1 w_2 x_2 w_3\}$$

از آن جا که قواعد سلسله‌مراتبی با جایگزینی زیرعبارات ترازشده با غیرپایانه‌ها تعریف می‌شوند، تراز کلمات، الگوهای بالا را برای یک جفت عبارت ترازشده محدود می‌کند.

جفت‌عباراتی با تراز یک‌نوای کلمات دارای زیرعبارات ترازشده زیادی (دارای هم‌پوشانی) به‌عنوان مکان غیرپایانه‌ها در الگوهای مختلف هستند. پس، قواعد سلسله‌مراتبی زیادی از این جفت عبارات قابل استخراج است. از سوی دیگر، الگوهای ساده (از قبیل $x_1 w_1$, $w_1 x_1$) برای عبارات ترازشده یک‌نوا قابل تعریف هستند. برای مثال، سمت مبداء جفت عبارت (a) در شکل (۳) با <قلم آبی x_1> یا<x_1 یک> می‌تواند نشان داده شود. بر اساس این واقعیت، فیلتر یک‌نوا برای استخراج فیلترشده قواعد با مدل عبارت-مرزی برمبنای الگوی تراز عبارات پیشنهاد شد [4]. این فیلتر برای عبارات قابل تجزیه به دو عبارت با تراز یک‌نوا، الگوی قواعد را به الگوهای ساده محدود می‌کند. در ادامه، فیلتر یک‌نوا را به‌اختصار توضیح می‌دهیم و برای تعمیم این فیلتر به مدل‌های سلسله‌مراتبی، پوشش قواعد دستور فیلترشده را در حالت کلی بحث می‌کنیم:

فیلتر یک‌نوا قواعد نامزد برای استخراج از جفت عبارت $<f_i^j, e_m^n>$ را با شرایط زیر می‌پذیرد:

۱- قواعدی که مبداء آنها با یکی از الگوهای زیر سازگار باشد:

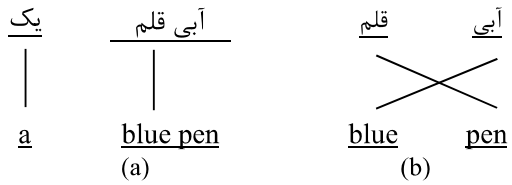
(۱۷)

$$\textit{Boundary}_2 = \{w_1, x_1 w_1, w_1 x_1, x_1 w_1 x_2\}$$

۲- سایر قواعد اگر $<f_i^j, e_m^n>$ *تجزیه‌پذیر به دو یک‌نوا* نباشد.

یک جفت عبارت، *تجزیه‌پذیر به دو یک‌نوا* است اگر بتواند به دو زیررشته با تراز یک‌نوا تجزیه شود. (۱۸)

به عبارت دیگر، عبارات تجزیه‌پذیر به دو یک‌نوا با اتصال یک‌نوای زیرعبارات ترازشده قابل تولید هستند. برای مثال، عبارات ترازشده (a) در شکل (۳) نمونه‌ای از عبارات تجزیه‌پذیر به دو یک‌نوا است. این جفت عبارت از زیرعبارات تراز شده یک‌نوای <a , یک> و <blue pen ,قلم آبی> تشکیل شده است. مطابق شکل (۳)، زیرعبارات <blue pen ,قلم آبی> تجزیه‌پذیر به دو یک‌نوا نیست. باید توجه داشت که نوع تراز زیرعبارات در شرایط استخراج قواعد برای یک جفت عبارت تأثیری ندارد. الگوی فیلتر $\textit{Boundary}_2$ قواعدی را با حداکثر دو غیرپایانه در مرزهای بخش مبداء نشان می‌دهد. باید توجه داشت که این الگو تنها استخراج قواعد از عبارات تجزیه‌پذیر به دو یک‌نوا را کاهش می‌دهد. برای مثال، قواعد ۱۴ و ۱۵ (با الگوی $x_1 x_2 w_1$ و $w_1 x_1 w_2$) با الگوی فیلتر $\textit{Boundary}_2$ سازگار نیستند و عبارت متناظر آنها تجزیه‌پذیر به دو یک‌نوا است. درنتیجه از استخراج آنها با فیلتر یک‌نوا صرف نظر می‌شود.



(شکل-۳): مثال‌هایی از (a) تراز یک‌نوا و (b) تراز متقاطع زیررشته‌ها

(Figure-3): Examples of (a) monotonic alignment and (b) cross alignment of substrings

بررسی قواعد فیلترشده نشان می‌دهد که درصد کمی از قواعد استخراج‌شده با الگوی فیلتر $\textit{Boundary}_2$ سازگار نیستند. درنتیجه، بیشتر عبارات ترازشده تجزیه‌پذیر به دو یک‌نوا هستند. از سوی دیگر، برخی عبارات به‌دلیل داده‌های نوفه‌ای تجزیه‌ناپذیر هستند. به این دلایل در رمزگشایی به قواعد سازگار با الگوی فیلتر اولویت داده می‌شود. علاوه‌بر خصوصیات متداول در مدل‌های سلسله‌مراتبی، یک خصوصیت جریمه الگو برای جریمه‌کردن قواعد ناسازگار با الگوی فیلتر قواعد استفاده می‌شود:

- جریمه الگو دارای مقدار 0 است اگر مبداء قاعده با یکی از الگوهای $\textit{Boundary}_2$ سازگار باشد. در غیر این‌صورت دارای مقدار 1 برای جریمه سایر قواعد است. (۱۹)

در ادامه پوشش قواعد دستور فیلترشده را برای مدل‌های سلسله‌مراتبی بررسی می‌کنیم. انتظار می‌رود، دو زیررشته عبارات تجزیه‌پذیر به دو یک‌نوا از قواعد دو بخشی ($x_1 w_1$ و $w_1 x_1$) در الگوی فیلتر $\textit{Boundary}_2$ قابل اشتقاق باشند؛ ولی به‌طور معمول، طول سمت راست قواعد سلسله‌مراتبی به پنج واژه (شامل غیرپایانه‌ها) محدود می‌شود. بنابراین یک جفت عبارت بزرگ ممکن است به شکل $w_1 x_1$ یا

$x_1 w_1$ به‌دلیل طول رشته w_1 قابل نمایش نباشد. از سوی دیگر، اتصال پشت سرهم زیررشته‌های عبارات تجزیه‌پذیر به دو یک‌نوا توسط قواعد چسب مشابه قاعده زیر حمایت می‌شود (S واژه آغازین دستور است.):

$$S \rightarrow < S^{\sim 1} X^{\sim 2}, S^{\sim 1} X^{\sim 2} > \quad (۲۰)$$

لازم به ذکر است که در دستور مدل‌های سلسله‌مراتبی منتج از مدل مبتنی بر عبارت سلسله‌مراتبی، قواعد چسب برای تمام غیرپایانه‌ها تعریف می‌شود. به این دلیل، فیلتر یک‌نوا تنها برای عبارات تجزیه‌پذیر به دو یک‌نوا پیشنهاد می‌شود که به بازترتیب زیررشته‌ها نیاز ندارند. البته، وابستگی‌های زبانی توسط قواعد سلسله‌مراتبی بهتر از قواعد چسب حمایت می‌شود. الگوی $x_1\ w_1\ x_2$ در الگوی فیلتر $Boundary_2$ با دو غیرپایانه شانس اشتقاق عبارات بزرگ را افزایش می‌دهد.

۵- آزمایش‌ها

مجموعه‌ای از آزمایش‌ها در ترجمه از فارسی و آلمانی به انگلیسی انجام شد. ترتیب کلمات در فارسی و انگلیسی اختلاف زیادی دارد. اگر چه ترتیب کلمات در زبان فارسی به‌طورتقریبی آزاد است، ساختار رسمی جملات آن SOV است که با ساختار SVO در انگلیسی تفاوت دارد. ترتیب کلمات آلمانی و انگلیسی نیز در بسیاری از موارد متفاوت است. برای مثال، در آلمانی افعال مصدری بعد از مفعول مرتبط آن قرار می‌گیرد.

برای ترجمه فارسی به انگلیسی، پیکره‌های موازی محدودی وجود دارد. همچنین کارهایی برای توسعه این پیکره‌ها (مانند [1]) انجام شده است. در این پژوهش، ترجمه فارسی به انگلیسی بر روی پیکره میزان [22] با حدود یک میلیون جمله آموزش داده شد. تعداد هزار جمله برای مجموعه تنظیم و هزار جمله برای مجموعه آزمون کنار گذاشته شد. ترجمه آلمانی به انگلیسی روی یک میلیون جمله ابتدای پیکره Europarl-V7 [23] آموزش داده شد. تعداد پانصد جمله انتهای وظیفه ترجمه WMT 2012 برای مجموعه تنظیم و هزار جمله ابتدای وظیفه ترجمه WMT 2013 برای مجموعه آزمون استفاده شد. آمار پیکره‌های استفاده‌شده در جدول (۱) آمده است.

ما نتایج مدل پیشنهادی را با سه مدل مبنا مقایسه کردیم: مدل مبتنی بر عبارت سلسله‌مراتبی [24] با یک غیرپایانه عمومی، مدل SAMT [3] که از درخت تجزیه جملات مقصد در پیکره آموزش برای برچسب گذاری قواعد استفاده می‌کند و مدل عبارت-مرزی در شکل مبنا [4] که قواعد را با برچسب POS عبارات مقصد برچسب می‌زند.

(جدول-۱): آمار پیکره‌های آموزش
(Table-1): Statistics of training corpus

پیکره آموزش	تعداد کلمه	کلمه یکتا	تراز کلمه
فارسی-انگلیسی	13M+15M	98K+135K	19M
آلمانی-انگلیسی	21M+20M	76K+222K	24M

کارایی مدل‌های ترجمه را با معیارهای کیفیت ترجمه و معیارهای عمومی کارایی (زمان و حافظه) ارزیابی می‌کنیم. در ارزیابی کارایی مدل‌ها، اولویت ارزیابی با کیفیت ترجمه است. زمان ترجمه را با معیار میانگین زمان مصرفی برای ترجمه یک جمله ارزیابی می‌کنیم. حافظه مصرفی را با معیار تعداد قواعد دستوری استخراج‌شده (اندازه مدل) ارزیابی می‌کنیم. از آنجا که اندازه مدل ثابت است و به طول ورودی بستگی ندارد، در ارزیابی زمان و حافظه، کاهش زمان ترجمه اولویت خواهد داشت.

برای ارزیابی کیفیت ترجمه از معیار ارزیابی خودکار BLEU[1] [25] استفاده می‌کنیم. این معیار برای ارزیابی کیفیت ترجمه ماشینی معمول است؛ زیرا همبستگی بالایی با ارزیابی انسان دارد [26]. معیار BLEU بر مبنای نسبت تعداد رشته‌های صحیح به طول n به کل تعداد رشته‌ها به همان طول در متن خروجی ترجمه (و نه یک جمله) محاسبه می‌شود. به‌طورمعمول (از جمله در این مقاله)، این معیار برای رشته‌های خروجی به طول یک تا چهار کلمه محاسبه و با نام BLEU-4 شناخته می‌شود.

مدل‌ها با مجموعه ابزار Joshua [27] آموزش دیده و ارزیابی شدند. تراز کلمات در دو جهت ترجمه با ابزار GIZA++ [28] انجام و نتایج متقارن شدند. مدل زبانی 3-gram بر روی سمت مقصد پیکره آموزش با ابزار Berkeley LM [29] ساخته و مقیاس پارامترهای مدل به روش کمینه نرخ خطا [30] آموزش داده شد.

نسخهٔ جدیدی از Thrax 2.0 [31] – ابزار استخراج دستور در Joshua - برای حمایت از مدل پیشنهادی توسعه داده شد. شکل (۴) نمونه‌ای از قواعد استخراج‌شده توسط این ابزار را برای مدل پیشنهادی نشان می‌دهد. استخراج دستور پیشنهادی به برچسب POS و برچسب قطعه نیاز دارد. این

[1] Bilingual Evaluation Understudy

برچسب‌ها با ابزار SENNA [32] روی سمت انگلیسی پیکره آموزش تعریف و درخت تجزیه جملات مقصد برای SAMT با ابزار تجزیه نحوی Stanford [33] تولید شد.

مدل‌های مبنا با تنظیمات پیش فرض پیکربندی شدند. این تنظیمات قواعد واژگانی را به طول ده کلمه و قواعد سلسله‌مراتبی را به طول پنج واژه شامل حداکثر دو غیرپایانه محدود می‌کند. استخراج قواعد سلسله‌مراتبی برای مدل مبتنی بر عبارت سلسله‌مراتبی به بازه ده کلمه و برای سایر مدل‌ها به بازه دوازده کلمه محدود شد.

از قواعد مجرد (فاقد کلمه) در قواعد صرف نظر شد. قواعد دستوری با خصوصیات پیش فرض زیر تعریف شدند:

- لگاریتم منفی احتمال عبارت (در دو جهت)
- لگاریتم منفی وزن واژگانی [34] (در دو جهت)
- جریمه کمیابی قواعد با مقدار exp (1-RuleFrequency)
- جریمه عبارت (با مقدار ثابت ۱) برای تشویق قواعدی با سمت راست بزرگ‌تر

طول قواعد و خصوصیات انتخاب‌شده برای مدل عبارت-مرزی پیشنهادی همانند مدل‌های مبنا انتخاب شد.

```
[PP-NP] ||| واقعیت‌های نهایی [PP,1] ||| [PP,1] ultimate facts ||| 0 6.68089 8.80287 0 0 1
[ADVP-VP] ||| او همسرش [ADVP,1] ||| [VBZ,2] [ADVP,1] she [VBZ,2] his wife ||| 0 2.24290 9.20967 0.69315 4.41884 1
```

(شکل-۴): نمونه‌ای از قواعد استخراج شده در ابزار Thrax
(Figure-4): Sample of extracted rules with Thrax tool
(هر سطر به ترتیب از چپ به راست، سمت چپ قاعده، بخش مبداء قاعده، بخش مقصد قاعده و مقدار خصوصیات مختلف را نشان می‌دهد که با نماد '|||' از هم جدا شده‌اند.)

۱-۵- نتایج بدون اعمال فیلتر

در این بخش نتایج ترجمه را بدون اعمال فیلتر روی مدل‌های مختلف بررسی می‌کنیم. گونه‌ای از SAMT که در آزمایش‌ها استفاده شد، عبارت مقصد را برای نام‌گذاری غیرپایانه‌ها با علایم زیر برچسب می‌زند:

- x: *عبارت* بدون تناظر با یک بازه در درخت تجزیه
- N_1: عبارت متناظر با طبقهٔ نحوی N_1
- $N_2\backslash N_1$ یا N_1/N_2 : عبارت متناظر با بخشی از طبقهٔ نحوی N_1 فاقد N_2 درسمت چپ یا راست
- N_1+N_2: عبارت متناظر با دو طبقهٔ نحوی همسایه

ابزار استخراج دستور - Thrax - دارای گزینه‌ای با نام Double-Plus برای برچسب‌گذاری قواعد در مدل SAMT است. با مقدار True برای این گزینه از نماد $N_1+N_2+N_3$ نیز در برچسب گذاری قواعد استفاده می‌شود. نتایج آزمایش‌ها در این حالت با نماد SAMT/double گزارش شده است.

نتایج آزمایش‌ها برای مدل مبتنی بر عبارت سلسله‌مراتبی (HPB)، SAMT، SAMT/double (همراه با گزینه Double-Plus) و گونه‌های زیر از مدل عبارت-مرزی در جدول (۲) مقایسه شده است:

- Boundary/base: شکل مبنای مدل عبارت-مرزی که از برچسب POS کلمات استفاده می‌کند.
- Boundary/CHK: مدل عبارت-مرزی پیشنهادی که با برچسب قطعات توسعه داده شده است.

جدول‌های (۲ و ۳) تعداد قواعد را به میلیون، امتیاز BLEU حاصل شده و میانگین زمان ترجمه هر جمله را به ثانیه نشان می‌دهد. بر اساس نتایج، بهترین کارایی مدل SAMT با فعال‌سازی گزینه Double-Plus حاصل می‌شود که از کیفیت ترجمه مدل پیشنهادی (Boundary/CHK) بهتر نیست.

در حالت کلی، تنوع ساخت‌واژی کلمات در زبان مبداء یا مقصد ترجمه، تعداد کلمات ترازشده را کاهش می‌دهد. تعداد تراز کلمات در پیکره‌های آموزش پس از ترازبندی در جدول (۱) گزارش شده است. این جدول نشان می‌دهد که تعداد تراز کلمات در پیکره فارسی-انگلیسی کمتر است. کاهش تراز کلمات با کاهش تراز عبارات همراه است. از سوی دیگر، تفاوت در ترتیب کلمات زبان مبداء و مقصد ترجمه نیز تراز عبارات را کاهش می‌دهد و کاهش تعداد عبارات ترازشده، تعداد قواعد استخراج‌شده را کاهش می‌دهد. جدول (۲) نشان می‌دهد که تعداد قواعد استخراج‌شده برای ترجمه فارسی-انگلیسی نسبت به آلمانی-انگلیسی بسیار کمتر است. با این وجود، زمان ترجمه فارسی به انگلیسی نیز زیاد است. در ادامه، برای کاهش زمان ترجمه مدل‌های مختلف را فیلتر می‌کنیم.

(جدول-۲): نتایج ترجمه با مدل‌های مختلف بدون اعمال فیلتر
(Table-2): Translation results of different models without filtering
(تعداد قواعد به میلیون و زمان ترجمه به ثانیه است.)

	Persian-English			German-English		
Model	Rules	BLEU	Time	Rules	BLEU	Time
HPB	11	11.75	0.29	101	18.53	0.13
SAMT	17	11.91	14.0	114	18.75	10.7
SAMT/double	21	12.41	16.1	186	19.16	13.7
Boundary/base	29	12.27	21.2	411	18.91	7.9
Boundary/CHK	28	12.63	18.2	399	19.15	13.5

(جدول-۳): نتایج ترجمه با مدل‌های فیلتر شده
(Table -3): Translation results of filtered models

		Persian-English			German-English		
Model	Filter	Rules	BLEU	Time	Rules	BLEU	Time
HPB	*Monotonic*	8	11.49	0.25	42	18.23	0.09
SAMT	*Monotonic*	12	11.73	4.5	56	18.50	2.5
SAMT/double	*Monotonic*	14	11.50	4.2	79	18.41	2.5
SAMT/double	MRC_2	2	09.03	4.7	18	18.96	4.7
Boundary/base	*monotonic*	16	11.95	7.3	128	18.86	2.2
Boundary/CHK	*monotonic*	16	**12.50**	8.0	125	**19.37**	4.4

۲-۵- نتایج با اعمال فیلتر

در این بخش برای کاهش زمان ترجمه، مدل‌های مورد آزمایش را فیلتر می‌کنیم. خصوصیت جریمه الگو (۱۹) به قواعد دستور، همراه سایر خصوصیات به مدل‌های فیلترشده با فیلتر یک‌نوا اضافه می‌شود. جدول (۳) نتایج اعمال فیلتر یک‌نوا (*monotonic*) را روی همه مدل‌ها نشان می‌دهد. با توجه به افت زیاد کیفیت ترجمه SAMT/double با فیلتر یک‌نوا، این مدل با یک پسا فیلتر شناخته‌شده نیز فیلتر شد. این فیلتر قواعد نادری را حذف می‌کند که کمتر از یک مقدار آستانه رخ داده باشند [16]. زیادکردن مقدار آستانه کیفیت ترجمه را کاهش می‌دهد. این فیلتر در ابزار Thrax با مقداردهی پارامتری با نام "Min-Rule-Count" فعال می‌شود. ما این فیلتر را با مقدار آستانهٔ دو فعال کردیم (MRC_2 در جدول (۳)). در مقایسه با فیلتر یک‌نوا، این پسا فیلتر با مقدار کمینه آن تنکی بسیار بیشتری در مدل ترجمه فارسی به انگلیسی ایجاد می‌کند.

در مقایسه با SAMT، استفاده از برچسب‌های بیشتر در مدل SAMT/double با تنکی بیشتر این مدل همراه است. فیلتر یک‌نوا اثر محسوسی روی سایر مدل‌ها نداشته و در ترجمه آلمانی به انگلیسی با مدل پیشنهادی تا حدی کیفیت ترجمه را نیز افزایش داده است.

اگر چه در ترجمه با مدل‌های فیلتر نشده، نتایج مدل پیشنهادی و SAMT (از نظر کیفیت و زمان ترجمه) مشابه است، براساس نتایج مدل‌های فیلترشده، مدل عبارت-مرزی پیشنهادی دارای بهترین کیفیت ترجمه برمبنای معیار BLEU است.

در ادامه، میزان اهمیت آماری بهبود نتایج را برای مدل ترجمه پیشنهادی بررسی می‌کنیم. برای این منظور، آزمون معناداری را با بازه اطمینان ۹۵٪ به روش نمونه‌گیری مجدد خودراه‌انداز[1] [35] با استفاده از مجموعه ابزار Moses [36] انجام می‌دهیم. جدول (۴) نتایج این آزمون را در مقایسه نتایج ترجمه مدل پیشنهادی و شکل مبنای مدل عبارت-مرزی نشان می‌دهد.

نتایج آزمون معناداری در جدول (۴) نشان می‌دهد که بهبود کیفیت ترجمه در مدل پیشنهادی نسبت به مدل عبارت-مرزی مبنا از نظر آماری با "p-value < 0.05" قابل توجه است.

(جدول-۴): نتایج آزمون معناداری با بازه اطمینان ۹۵٪
(Table-4): Results of statistical significance test with 95% confidence interval

	German-English		Persian-English	
Model	BLEU	p-value	BLEU	p-value
Boundary/base	12.11	-	18.70	-
Boundary/CHK	12.69	0.041	19.18	0.45

[1] Bootstrap resampling method

۶- نتیجه‌گیری

مدل عبارت-مرزی پیشنهادشده در این مقاله غیرپایانه‌ها را با برچسب‌های کم‌عمق نحوی در مرز عبارات مقصد نام‌گذاری می‌کند. این مدل در شکل مبنا تنها از برچسب POS به‌عنوان طبقهٔ کلمات مرزی استفاده می‌کند؛ ولی توسعه پیشنهادی برچسب قطعات را نیز به کار می‌برد. با درنظرگرفتن کیفیت و زمان ترجمه، مدل پیشنهادی در مقایسه با مدل مبتنی بر عبارت سلسله‌مراتبی، شکل مبنای مدل عبارت-مرزی و مدل SAMT، ترجمه زبان‌هایی با تفاوت زیاد را در ترتیب کلمات بهبود داد. اگرچه آزمایش‌های ما در ترجمه به زبان انگلیسی انجام شد، برای بیشتر زبان‌ها، تجزیه‌گر کم‌عمق نحوی (برای تولید برچسب قطعه) از تجزیه‌گر نحوی در دسترس‌تر است.

مدل SAMT برای نام‌گذاری عبارات فاقد بازه متناظر در درخت تجزیه نحوی از یک برچسب پیش‌فرض استفاده می‌کند. از سوی دیگر، بخش بیشتری از عبارات نحوی در زبان‌هایی با اختلاف زیاد در ترتیب کلمات تراز نمی‌شوند. درنتیجه استفاده از فیلتر برای کاهش زمان ترجمه به تنکی زیاد مدل SAMT در ترجمه فارسی به انگلیسی منجر شد. همچنین، نتایج آزمایش‌ها نشان داد که فیلتر یک‌نوا برای کاهش زمان ترجمه و اندازه مدل پیشنهادی و سایر مدل‌های سلسله‌مراتبی قابل استفاده است.

در توسعهٔ آینده مدل عبارت-مرزی، این مدل را با برچسب‌های نقش معنایی به‌عنوان برچسب‌های کم‌عمق معنایی توسعه خواهیم داد. برچسب‌های نقش معنایی بیان‌گر گزاره‌ها و آرگومان‌های مرتبط در جمله هستند. تعریف غیرپایانه‌ها با برچسب‌های نقش معنایی خطاهای معنایی ترجمه را می‌تواند کاهش دهد و بازترتیب آرگومان‌های معنایی را بهبود دهد.

۷- مراجع 7-References

[1] رحیمی زینب، ثمنی محمد حسین، خدیوی شهرام. استخراج پیکره موازی از اسناد قابل مقایسه برای بهبود کیفیت ترجمه در سامانه‌های ترجمه ماشینی. پردازش علائم و داده‌ها، ۱۲(۲)، ۷۲-۵۵، ۱۳۹۴

[1] Z. Rahimi, M. H. Samani, and S. Khadivi, "Extracting parallel corpus from comparable documents to improve the quality of translation in machine translation systems," *Signal data Process.*, vol. 12, no. 2, pp. 55–72, 2015.

[2] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 263–270.

[3] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing," in *Proceedings of the Workshop on Statistical Machine Translation*, 2006, pp. 138–141.

[4] S. Salami, M. Shamsfard, and S. Khadivi, "Phrase-boundary model for statistical machine translation," *Comput. Speech Lang.*, vol. 38, pp. 13–27, 2016.

[5] T. Watanabe, H. Tsukada, and H. Isozaki, "Left-to-right target generation for hierarchical phrase-based translation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 777–784.

[6] M. Huck, S. Peitz, M. Freitag, and H. Ney, "Discriminative reordering extensions for hierarchical phrase-based machine translation," in *Proc. of the 16th Annual Conf. of the European Assoc. for Machine Translation*, 2012, pp. 313–320.

[7] M. de B. Wenniger and K. Sima'an, "Labeling hierarchical phrase-based models without linguistic resources," *Mach. Transl.*, vol. 29, no. 3–4, pp. 225–265, 2015.

[8] Z. He, Q. Liu, and S. Lin, "Improving statistical machine translation using lexicalized rule selection," in *Proceedings of the 22nd In-ternational Conference on Computational Lin-guistics-Volume 1*, 2008, pp. 321–328.

[9] R. Haque, S. Kumar Naskar, A. Van Den Bosch, and A. Way, "Supertags as source language context in hierarchical phrase-based SMT," in *Association for Machine Translation in the Americas (AMTA 2010)*, 2010.

[10] B. Zhou, X. Zhu, B. Xiang, and Y. Gao, "Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels," in *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, 2008, pp. 19–27.

[11] H. Almaghout, J. Jiang, and A. Way, "CCG augmented hierarchical phrase based machine-translation," 2010.

[12] H. Mino, T. Watanabe, and E. Sumita, "Syntax-Augmented Machine Translation using Syntax-Label Clustering.," in *EMNLP*, 2014, pp. 165–171.

[13] J. Li, Z. Tu, G. Zhou, and J. van Genabith, "Using syntactic head information in hierarchical phrase-based translation," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 2012, pp. 232–242.

machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 311–318.

[26] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[27] Z. Li, C. Callison-Burch, C. Dyer, J. Ganitkevitch, S. Khudanpur, L. Schwartz, W. N. G. Thornton, J. Weese, and O. F. Zaidan, "Joshua: An open source toolkit for parsing-based machine translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009, pp. 135–139.

[28] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000, pp. 440–447.

[29] A. Pauls and D. Klein, "Faster and smaller n-gram language models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 258–267.

[30] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003, pp. 160–167.

[31] M. Post, J. Ganitkevitch, L. Orland, J. Weese, Y. Cao, and C. Callison-Burch, "Joshua 5.0: Sparser, better, faster, server," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 2013, pp. 206–212.

[32] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[33] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003, pp. 423–430.

[34] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 48–54.

[35] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation.," in *EMNLP*, 2004, pp. 388–395.

[36] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and others, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 2007, pp. 177–180.

Based Translation using Sparse Features.," in *HLT-NAACL*, 2013, pp. 22–31.

[15] A. Zollmann and S. Vogel, "A word-class approach to labeling pscfg rules for machine translation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 1–11.

[16] A. Zollmann, A. Venugopal, F. Och, and J. Ponte, "A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, 2008, pp. 1145–1152.

[17] Z. He, Y. Meng, and H. Yu, "Discarding monotone composed rule for hierarchical phrase-based statistical machine translation," in *Proceedings of the 3rd International Universal Communication Symposium*, 2009, pp. 25–29.

[18] G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne, "Rule filtering by pattern for efficient hierarchical translation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 380–388.

[19] S.-W. Lee, D. Zhang, M. Li, M. Zhou, and H.-C. Rim, "Translation model size reduction for hierarchical phrase-based statistical machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012, pp. 291–295.

[20] B. Sankaran, G. Haffari, and A. Sarkar, "Bayesian extraction of minimal scfg rules for hierarchical phrase-based translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 533–541.

[21] B. Sankaran, G. Haffari, and A. Sarkar, "Compact rule extraction for hierarchical phrase-based translation," in *The 10th biennial conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA. Association for Computational Linguistics*, 2012.

[22] S. C. of ICT, "Mizan English-Persian Parallel Corpus," 2013.

[23] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, 2005, vol. 5, pp. 79–86.

[24] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguist.*, vol. 33, no. 2, pp. 201–228, 2007.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of

شهرام سلامی مـــدارک کارشناســـی و کارشناسـی ارشد خـود را به‌ترتیـــب در رشته مهندسی کامپیوتر گرایش نـرمافزار و گرایـــش هـوش مصنوعی از دانشگـــاه شهید بهشتی و دانشگاه علوم و تحقیقات تـهران در سـال‌های ۱۳۷۶ و ۱۳۷۹ اخـــذ کـرده و در سـال مدرک دکترای خود را در رشته مهندسـی کامپیوتر گرایـــش نرم افـــزار از دانشکـده مهندسـی و علـوم کامپیوتر دانشگـــاه شهید بهشتی دریافت کردند. زمینه‌های پژوهشی مورد علاقـه ایـشان پردازش زبان، ترجمه ماشیـنی آماری و پایگاه داده‌هـا می‌باشند.

نشانی رایانامه ایشان عبارت است از:

sh_salami@sbu.ac.ir

مهرنوش شمس فرد مدارک کارشناسی و کارشناســـی ارشد خـــود را در رشتـــه مهندســـی کامپیوتر گرایـــش نرم‌افزار از دانشگاه صنعتی شریف و مـدرک دکترای خود را در رشتـــه مهندســـی کامپیوتـــر گرایش هوش مصنوعـی ازدانشگاه صـنعتی امیرکبـیر دریافت کردند. ایشان در حال حاضـر دانشیار دانشگاه شهیدبهشتـی و سرپرســـت آزمایشگـــاه پردازش زبـــان طبیعی در دانشکـــده مهندسی و علوم کامپیوتر هستند. زمینه‌های پژوهشی مـــورد علاقه ایشان هوش مصنوعی، پردازش زبان طبیعی با تأکید بر پردازش زبـان فارســـی، مهندسی دانـش و هستان‌شناســـی و وب‌معنایی است.

نشانی رایانامه ایشان عبارت است از:

m-shams@sbu.ac.ir