ترجمه ماشینی آماری با استفاده از برچسب‌های کم عمق نحوی

شهرام سلامی[۱]، مهرنوش شمس فرد [۲]

دانشگاه شهید بهشتی، دانشکده مهندسی و علوم کامپیوتر

sh_salami@sbu.ac.ir [۱], m-shams@sbu.ac.ir [۲]

چکیده

این مقاله مدل سلسله مراتبی جدیدی را برای ترجمه ماشینی آماری پیشنهاد می‌دهد که غیرپایانه‌ها را با تطابق مرز عبارات مقصد با برچسب-های کم عمق نحوی در سمت مقصد پیکره آموزش، نام‌گذاری می‌کند. در جایی که برچسبی برای کل عبارت موجود نباشد، نام غیرپایانه از اتصال برچسب‌های مرزی تعریف می‌شود. برچسب‌گذاری با کلاس کلمات مرزی عبارات قبلا معرفی گردیده است که می‌تواند شکل مبنای مدل پیشنهادی درنظر گرفته شود. ما این شکل مبنا را در مقاله حاضر با استفاده از برچسب قطعات توسعه می‌دهیم. در این توسعه، اگر برچسب قطعه در مرز عبارت وجود نداشته باشد، از برچسب POS کلمه مرزی استفاده می‌شود. با استفاده از برچسب عبارات به جای کلاس کلمات، قواعد مدل پیشنهادی تعمیم داده می‌شود. تعدادی آزمایش در ترجمه فارسی به انگلیسی انجام شد. با استفاده از معیار BLEU در قیاس با مدل SAMT که از درخت تجزیه نحوی برای برچسب گذاری استفاده می‌کند، مدل پیشنهادی بهبود قابل توجهی به دست آورد.

واژه‌های کلیدی

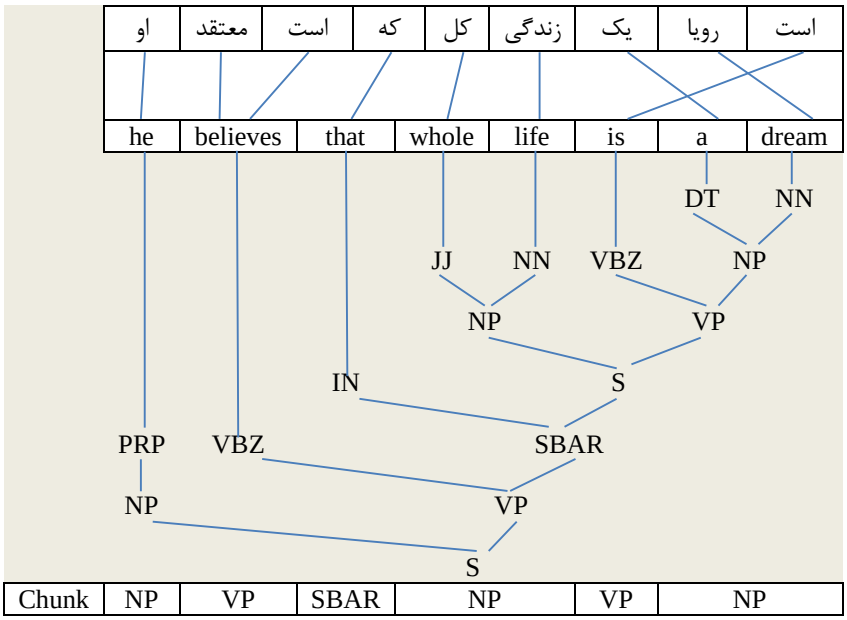
ترجمه ماشینی آماری، مدل سلسله مراتبی، برچسب کلمه، برچسب قطعه

۱- مقدمه

مدل‌های سلسله مراتبی نسبت به مدل‌های مبتنی بر عبارت از بازترتیب کلمات بهتر حمایت می‌کنند. این مدل‌ها ترجمه بهتری را برای زبان‌هایی با اختلاف زیاد در ترتیب کلمات، نوید می‌دهند. برمبنای نوع برچسب‌های استفاده شده در قواعد، مدل‌های سلسله مراتبی مختلفی پیشنهاد شده است. مدل مبتنی بر عبارت سلسله مراتبی[۱] از یک برچسب عمومی برای تمام غیر پایانه‌ها استفاده می‌کند. مدل SAMT [۲] یک مدل به خوبی شناخته شده است که از کلاس‌های نحوی زبان مقصد برای برچسب قواعد استفاده می‌کند. در این مدل برچسب از تطابق سمت مقصد عبارات تراز شده با زیردرخت‌های درخت تجزیه نحوی جمله در پیکره مقصد به‌دست می‌آید. عباراتی که با بازه‌ای در درخت تجزیه تطابق نداشته باشند با غیر پایانه عمومی X برچسب گذاری می‌شوند. برای مثال، تراز جملات فارسی-انگلیسی شکل ۱ را در نظر بگیرید. قاعده زیر با یک بازه نحوی در درخت تجزیه جمله انگلیسی تناظر ندارد و با برچسب پیش فرض در مدل SAMT تعریف می‌شود (برای سهولت، کلمات فارسی از چپ به راست نمایش داده شده است.):

$$X \rightarrow \text{<کل که است معتقد او ; he believes that whole>} \quad (۱)$$

البته، برخی تطابق‌های نسبی با درخت تجزیه در این مدل پیش بینی شده است. برای مثال، قاعده زیر جمله‌ای فاقد عبارت فعلی در سمت راست را نشان می‌دهد:



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.)

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

در مدل عبارت-مرزی [۳] برچسب قواعد با کلاس کلمات مرزی سمت مقصد عبارات تراز شده تعریف می‌شود. برای مثال، با استفاده از برچسب POS به عنوان کلاس کلمات، قواعد زیر در این مدل، معادل قواعد ۱ و ۲ تعریف 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)

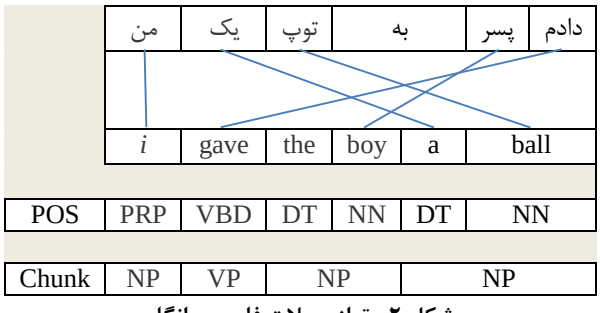
مدل عبارت-مرزی در مقایسه با SAMT همه عبارت تراز شده را پوشش می‌دهد و نیازی به برچسب پیش فرض X ندارد. در آزمایش‌ها، کیفیت ترجمه این مدل با مدل SAMT مشابه است. به عنوان مثالی دیگر، قواعد زیر در مدل عبارت-مرزی برای تراز جملات فارسی-انگلیسی در شکل ۲ استخراج شده است

PRP → < من ; i > (۵)

DT-NN → < $DT^{\sim 1}$ توپ ; $DT^{\sim 1}$ ball > (۶)

PRP-NN → < $PRP^{\sim 1}$ $DT\text{-}NN^{\sim 2}$ دادم پسر به ; $PRP^{\sim 1}$ gave the boy $DT\text{-}NN^{\sim 2}$ > (۷)

ما قصد داریم با اضافه کردن دانش زبانی بیشتر در مدل عبارت-مرزی کیفیت ترجمه بالاتری به‌دست آوریم. در این مقاله استفاده از برچسب قطعات همراه با برچسب POS کلمات برای نام‌گذاری غیرپایانه‌ها پیشنهاد می‌گردد. برچسب قطعات حاصل تجزیه کم عمق نحوی (قطعه بندی) هستند که اجزاء جمله (مانند عبارت اسمی و فعلی) را به صورت یکتا تعیین می‌کنند. 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 تعریف شده است.



شکل ۲- تراز جملات فارسی-انگلیسی

به همراه برچسب POS و قطعات کلمات انگلیسی

مدل پیشنهادی از برچسب مرزی عبارات مقصد برای بیان هم‌جواری عبارات خروجی استفاده می‌کند. نام‌گذاری غیرپایانه‌ها با استفاده از همه برچسب‌های واقع در عبارت مقصد، تنکی مدل را افزایش می‌دهد. در مقایسه با مدل عبارت-مرزی مبنا، قواعد مدل پیشنهادی با استفاده از برچسب عبارات به جای برچسب کلمات تعمیم می‌یابد. نام‌گذاری غیرپایانه‌ها با استفاده از برچسب‌های نحوی عبارات مقصد امکان ساخت نحوی خروجی ترجمه را فراهم می‌کند، در حالی‌که رمزگشایی به وسیله ورودی ترجمه هدایت می‌شود. از سوی دیگر، تناظر خوبی بین برچسب‌های کم عمق نحوی در عبارات مبداء و مقصد وجود دارد.

با استفاده از مدل مبتنی بر عبارت سلسله مراتبی، گونه‌ای از مدل SAMT و مدل عبارت-مرزی به عنوان مبنای مقایسه، مدل پیشنهادی در ترجمه فارسی به انگلیسی، بهبود قابل توجهی با معیار BLEU به‌دست آورد. در این مقاله برخی کارهای مرتبط در بخش ۲ معرفی می‌شود. در بخش ۳ نام‌گذاری غیر پایانه‌ها با استفاده از برچسب‌های کم عمق نحوی تعریف می‌شود. بخش ۴ آزمایش‌های انجام شده را تشریح می‌کند. سرانجام، مقاله در بخش ۵ نتیجه گیری می‌شود.

۲- کارهای مرتبط

مدل مبتنی برعبارت سلسله مراتبی [۱] برای استخراج گرامر بدون نظارت بر مبنای تراز عبارات با یک غیرپایانه عمومی معرفی گردید. تولید ناپیوسته کلمات مقصد هرس فضای رمزگشایی با مدل زبانی مقصد را محدود می‌کند. با محدود کردن قواعد ترجمه به فرم GNF [۵] جمله مقصد از چپ به راست تولید گردید. در کاری دیگر [۶]، با اجتناب از فرم بازگشتی قواعد سلسله مراتبی فضای رمزگشایی مدل مبتنی برعبارت سلسله مراتبی محدود گردید. در این کار از دوغیرپایانه مختلف در سمت چپ و راست قواعد سلسله مراتبی استفاده شد. بدون استفاده از منابع زبانی، الگوی تجزیه عبارات برای برچسب گذاری قواعد استفاده گردید [۷].

برای انتخاب بهتر قواعد در فرایند رمزگشایی مدل مبتنی برعبارت سلسله مراتبی، اطلاعات بافتار ورودی به شکل برچسب POS [۸] و برچسب CCG [۹] استفاده شده است. با استفاده از دانش نحوی، اشتقاق‌ها در زمان رمزگشایی امتیازدهی شدند [۱۰]. غیرپایانه‌ها با برچسب POS سرآیند کلمات نام‌گذاری شدند [۱۱]. دقت مدل مبتنی بر عبارت سلسله مراتبی با کلاس‌های نحوی به عنوان مدل SAMT [۲] و برچسب CCG [۱۲] بهبود یافت. برچسب‌های نحوی برای کاهش تعداد قواعد در SAMT خوشه بندی شدند [۱۳].

کلاس کلمات مرزی عبارات در برخی کارهای جدید استفاده شده است. کلاس کلمات مرزی عبارات برای بهبود بازترتیب کلمات در مدل مبتنی برعبارت سلسله مراتبی [۶] و مبتنی برعبارت [۱۴] به کار رفت. با برچسب گذاری قواعد با کلاس کلمات مرزی عبارات، کیفیت ترجمه مشابه مدل SAMT به دست آمد [۱۵]. برچسب گذاری قواعد با کلاس کلمات مرزی عبارات همراه با استخراج فیلتر شده قواعد برای کاهش اندازه مدل و زمان رمزگشایی در مدل عبارت-مرزی [۳] پیشنهاد شد. مدل پیشنهاد شده در این مقاله برچسب گذاری مرزی را به برچسب POS و قطعه برای کیفیت بهتر ترجمه تعمیم می‌دهد.

۳- مدل

مدل پیشنهادی یک گرامر همگام مستقل از متن را از عبارات تراز شده استخراج می‌کند. پیرو [۱]، قواعد وزن دار به شکل واژگانی، سلسله مراتبی و چسب تعریف می‌شوند. قواعد واژگانی بیانگر عبارات تراز شده بدون غیرپایانه در سمت راست هستند. قواعد سلسله مراتبی با حداکثر دو جایگذاری زیرعبارات با غیرپایانه‌ها تعریف می‌شوند. قواعد چسب برای همه غیرپایانه‌های گرامر جهت اتصال متوالی عبارات خروجی تعریف می‌شوند.

مدل عبارت-مرزی در شکل مبنا [۳] به صورت یکنواخت، کلاس کلمات مرزی عبارات مقصد را با یک خط پیوند برای نام‌گذاری غیرپایانه‌ها اتصال می‌دهد. با استفاده از برچسب POS به عنوان کلاس کلمات، برچسب عبارت تراز شده $<f_i^j, e_m^n>$ (که f_i^j و e_m^n به ترتیب بیانگر زیر رشته بسته از موقعیت i تا j و موقعیت m تا n است) به شکل زیر تعریف می‌شود:

$$X_{m,n} = \begin{cases} POS(m) & if\ m = n \\ POS(m) - POS(n) & else \end{cases} \qquad (٤)$$

علاوه بر برچسب POS، مدل عبارت-مرزی توسعه داده شده از برچسب قطعه در مرز عبارات مقصد نیز استفاده می‌کند. وقتی یک قطعه تمام عبارت تراز شده را پوشش دهد، به عنوان برچسب استفاده خواهد شد. در غیر این صورت قطعه واقع در عبارت مقصد که از سمت چپ شروع شود یا در سمت راست 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 به صورت رسمی نشان می‌دهد.

تعریف ۱- برچسب غیر پایانه مربوط به عبارت تراز شده $<f_i^j, e_m^n>$ که در آن عبارت مقصد با e_m شروع و با e_n خاتمه می‌یابد، به شکل زیر تعریف می‌شود:

$$X_{m,n} = \begin{cases} Chunk\ (e_m^n) & if\ a\ Chunk\ Label\ matches \\ \text{POS}\ (e_m) & else\ if\ m = n \\ Left\text{-}Label\ -\ Right\text{-}Label & else \end{cases}$$

$$Left\text{-}Label = \begin{cases} Chunk\ (e_m^a) : m \le a < n & if\ a\ Chunk\ Label\ matches \\ POS\ (e_m) & else \end{cases}$$

$$Right\text{-}Label = \begin{cases} Chunk\ (e_b^n) : m < b \le n & if\ a\ Chunk\ Label\ matches \\ POS\ (e_n) & else \end{cases}$$

برچسب‌های حاصل از *Left-Label* و *Right-Label* با یک خط پیوند متصل می‌شوند. تابع $Chunk\ (e_x^y)$ برچسب قطعه‌ای که از موقعیت x شروع و در موقعیت y پایان می‌یابد را باز می‌گرداند. با توجه به عدم هم‌پوشانی قطعه‌ها، در این تعریف $a < b$ است. برای مثال، قواعد زیر که از تراز عبارات شکل ۲ استخراج شده‌اند، تعمیم قواعد ۵ تا ۷ با برچسب عبارات اسمی (NP) را نشان می‌دهد:

$$NP \rightarrow < \text{من} ; i > \quad (۸)$$

$$NP \rightarrow < DT^{\sim 1}\ \text{توپ} ; DT^{\sim 1}\ ball > \quad (۹)$$

$$NP\text{-}NP \rightarrow < NP^{\sim 1}\ \text{به}\ NP^{\sim 2}\ \text{دادم پسر} ; NP^{\sim 1}\ gave\ the\ boy\ NP^{\sim 2} > \quad (۱۰)$$

همچنین، قواعد زیر که از تراز شکل ۱ استخراج شده است، معادل قواعد ۱ و ۲ را در توسعه مدل عبارت-مرزی نشان می‌دهد:

$$NP\text{-}JJ \rightarrow <\text{کل که است معتقد او} ; he\ believes\ that\ whole> \quad (۱۱)$$

$$NP\text{-}NP \rightarrow <\text{زندگی کل که است معتقد او} ; he\ believes\ that\ whole\ life> \quad (۱۲)$$

همان‌طور که این قواعد نشان می‌دهد، توسعه مدل عبارت-مرزی ضمن پوشش همه عبارات تراز شده، درصورت امکان از برچسب قطعات در تعریف قواعد استفاده می‌کند.

۴- آزمایش‌ها

مجموعه‌ای از آزمایش‌ها در ترجمه فارسی به انگلیسی انجام شد. ترتیب کلمات در فارسی و انگلیسی اختلاف زیادی دارد. اگر چه فارسی تقریبا یک زبان با ترتیب آزاد کلمات است، ساختار رسمی جملات آن SOV است که با ساختار SVO در انگلیسی تفاوت دارد. ترجمه بر روی پیکره میزان [۱۶] با حدود یک میلیون جمله آموزش داده شد. تعداد هزار جمله برای مجموعه تنظیم و هزار جمله برای مجموعه تست کنار گذاشته شد.

با استفاده از معیار ارزیابی BLEU-4 [۱۷]، ما نتایج را با سه مدل مبنا مقایسه کردیم: مدل مبتنی بر عبارت سلسله مراتبی [۴] با یک غیرپایانه عمومی، SAMT [۲] که از درخت تجزیه جمله مقصد در پیکره آموزش برای برچسب گذاری قواعد استفاده می‌کند و مدل عبارت-مرزی در شکل مبنا [۳] که قواعد را با برچسب POS عبارات مقصد برچسب می‌زند. با توجه به عدم استفاده از دانش زبانی مبداء یا مقصد ترجمه در مدل مبتنی بر عبارت سلسله مراتبی، این مدل به عنوان یک مدل مبنای مستقل از زبان در آزمایش‌ها استفاده گردید.

مدل‌ها با مجموعه ابزار Joshua [۱۸] آموزش دیده و ارزیابی شدند. تراز کلمات در دو جهت ترجمه با ابزار GIZA++ [۱۹] انجام شد و نتایج متقارن شدند. مدل زبانی 3-gram بر روی سمت مقصد پیکره آموزش با ابزار Berkeley LM [۲۰] ساخته شد. مقیاس پارامترهای مدل به روش حداقل نرخ خطا [۲۱] آموزش داده شد.

نسخه جدیدی از Thrax 2.0 [۲۲] – ابزار استخراج گرامر در Joshua - برای حمایت از مدل پیشنهادی توسعه داده شد. استخراج گرامر پیشنهادی به برچسب POS و برچسب قطعه نیاز دارد. این برچسب‌ها با ابزار SENNA [۲۳] روی سمت انگلیسی پیکره آموزش تعریف گردید. درخت نحو جملات مقصد برای SAMT با ابزار تجزیه گر نحوی Stanford [۲۴] تولید شد.

مدل‌های مبنا با تنظیمات پیش فرض پیکربندی شدند. این تنظیمات قواعد واژگانی را به طول ۱۰ کلمه و قواعد سلسله مراتبی را به طول ۵ نشانه شامل حداکثر دو غیرپایانه محدود می‌کند. استخراج قواعد سلسله مراتبی برای مدل مبتنی بر عبارت سلسله مراتبی به بازه ۱۰ کلمه و برای سایر مدل-ها به بازه ۱۲ کلمه محدود شد. قواعد مجرد (فاقد کلمه) در گرامر تولید نشد. خصوصیات پیش فرض برای قواعد گرامر شامل لگاریتم منفی احتمال عبارت (در دو جهت)، لگاریتم منفی وزن واژگانی [۲۵] (در دو جهت)، جریمه کمیابی قواعد و جریمه عبارت (با مقدار ثابت ۱) اعمال شدند. طول قواعد و خصوصیات انتخاب شده برای مدل پیشنهادی همانند مدل‌های مبنا انتخاب گردید.

۴-۱- نتایج مدل‌های فیلتر نشده

در این بخش ابتدا نتایج مدل‌های فیلتر نشده را بررسی می‌کنیم و سپس در بخش بعد برای کاهش زمان ترجمه مدل‌های مختلف را فیلتر می‌کنیم. گونه‌ای از SAMT که در آزمایش‌ها استفاده شد، عبارت مقصد را برای نام-گذاری غیرپایانه‌ها با علائم زیر برچسب می‌زند:

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

ابزار استخراج گرامر - Thrax - دارای دو گزینه برای برچسب قواعد در مدل SAMT است. این ابزار با مقدار True برای گزینه Double-Plus از نماد $N_1+N_2+N_3$ نیز در برچسب گذاری استفاده می‌کند. گزینه دیگر آن Non-Lexical-X است که با مقدار False از قواعد سلسله مراتبی (غیر واژگانی) شامل غیرپایانه پیش فرض X صرف نظر می‌کند. جدول ۱ نتایج آزمایش‌ها با مقادیر مختلف این پارامترها را نشان می‌دهد. نتایج در این جدول (و جداول بعدی) تعداد قواعد به میلیون، امتیاز BLEU حاصل شده و میانگین زمان ترجمه هر جمله به ثانیه را نشان می‌دهد.

جدول ۱- نتایج آزمایش‌ها با پارامترهای مختلف مدل SAMT
(تعداد قواعد به میلیون زمان به ثانیه)

Double-Plus	Non-Lexical-X	Rules	BLEU	Time
False	False	17	11.91	14.0
True	False	21	**12.41**	16.1
False	True	29	11.83	32.2
True	True	30	12.15	31.4

براساس نتایج، بهترین کارایی مدل SAMT با فعال سازی گزینه Double-Plus حاصل می‌شود. بهترین کارایی مدل SAMT با مدل مبتنی بر عبارت سلسله مراتبی (HPB) 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.

جدول ۲- نتایج ترجمه فارسی به انگلیسی با مدل‌های مختلف

Model	Rules	BLEU	Time
HPB	11	11.75	0.29
SAMT	21	12.41	16.1
Boundary/Base	29	12.27	21.2
Boundary/CHK	28	**12.63**	18.2

بر اساس نتایج، کیفیت ترجمه مدل پیشنهادی (Boundary/CHK) از سایر مدل‌ها بیشتر است.

۴-۲- نتایج مدل‌های فیلتر شده

در این بخش برای کاهش زمان ترجمه، مدل‌های عبارت-مرزی و SAMT را فیلتر می‌کنیم. فیلتر یکنوا [۳] برای فیلتر قواعد در مرحله استخراج گرامر بر مبنای الگوی تراز عبارات پیشنهاد شده است. ما این فیلتر را در ادامه اعمال می‌کنیم که منابع آموزش و رمزگشایی را به میزان قابل توجه کاهش می‌دهد. در ادامه این فیلتر به اختصار توضیح داده می‌شود. فیلتر یکنوا قواعد نامزد برای استخراج را با شرایط زیر می‌پذیرد:

۱- تمام قواعد واژگانی (فاقد غیرپایانه در سمت راست)

۲- قواعد سلسله مراتبی که مبداء آنها با یکی از الگوهای زیر سازگار باشد (w_i و x_i به ترتیب یک رشته از کلمات و یک غیرپایانه هستند):

$$Boundary_2 = \{x_1\ w_1,\ w_1\ x_1,\ x_1\ w_1\ x_2\} \qquad (۱۳)$$

۳- قواعد سلسله مراتبی دیگر جایی که عبارات متناظر *تجزیه پذیر*-۲ *یکنوا* نباشد.

یک جفت عبارت تراز شده تجزیه پذیر-۲ یکنوا است اگر بتواند به زیر رشته‌هایی با تراز یکنوا تجزیه شود. برای مثال، معادله ۱۴ یک عبارت تجزیه پذیر-۲ یکنوا در شکل ۱ است:

< یک ; a > + < توپ ; ball > = < یک ; a ball > < توپ (۱۴)

به عبارت دیگر، با استخراج همه قواعد از جفت عبارات دیگر، استخراج قواعد سلسله مراتبی از عبارات تجزیه پذیر-۲ یکنوا به قواعدی با حداکثر دو غیرپایانه در مرزهای سمت مبداء قاعده (۱۳) محدود می‌شود. یک خصوصیت جریمه نیز به گرامر فیلتر شده با نام جریمه الگو اضافه می‌شود:

- جریمه الگو دارای مقدار 0 است اگر قاعده واژگانی باشد یا اگر یک قاعده سلسله مراتبی باشد که مبداء آن با یکی از الگوهای $Bondary_2$ سازگار باشد. در غیر این‌صورت دارای مقدار 1 برای جریمه قواعد سلسله مراتبی دیگر است.

جدول ۳ نتایج اعمال فیلتر یکنوا را روی همه مدل‌ها نشان می‌دهد (*monotonic* در جدول ۳). خصوصیت جریمه الگو به قواعد گرامر همراه سایر خصوصیات به مدل‌های فیلتر شده یکنوا اضافه می‌شود. چون کیفیت ترجمه SAMT با فیلتر یکنوا افت زیادی دارد، این مدل با یک پسا فیلتر شناخته شده [۲۶] نیز فیلتر گردید. این فیلتر قواعد نادر را که کمتر از یک مقدار آستانه رخ داده باشند را حذف می‌کند. زیاد کردن مقدار آستانه کیفیت ترجمه را کاهش می‌دهد. این فیلتر در ابزار Thrax با مقداردهی پارامتر "Min-Rule-Count" فعال می‌شود. ما این فیلتر را با مقدار آستانه‌ای ۲ فعال کردیم (MRC_2 در جدول ۳).

جدول ۳- نتایج ترجمه با مدل‌های فیلتر شده

Model	Filter	Rules	BLEU	Time
SAMT	*monotonic*	14	11.50	4.2
SAMT	MRC_2	2	09.03	4.7
Boundary/Base	*monotonic*	16	11.95	7.3
Boundary/CHK	*monotonic*	16	**12.50**	8.0

همان‌طور که نتایج نشان می‌دهد، در مدل‌های فیلتر شده نیز بهترین نتیجه با مدل پیشنهادی حاصل شد. هر دو فیلتر تنکی زیادی در مدل SAMT ایجاد کرده و کیفیت ترجمه آن را به شدت کاهش می‌دهند. لازم به ذکر است که اختلاف زیاد در ترتیب کلمات فارسی و انگلیسی با تنکی بیشتر قواعد نحوی همراه است.

۵- نتیجه گیری

مدل عبارت-مرزی پیشنهاد شده در این مقاله غیرپایانه‌ها را با برچسب-های کم عمق نحوی در مرز عبارات مقصد نام‌گذاری می‌کند. این مدل در شکل مبنا تنها از برچسب POS به عنوان کلاس کلمات مرزی استفاده می-کند ولی توسعه پیشنهادی برچسب قطعات را نیز به کار می‌برد. با معیار BLEU، مدل پیشنهادی امتیاز بیشتری نسبت به مدل مبتنی بر عبارت سلسله مراتبی، شکل مبنای مدل عبارت-مرزی و مدل SAMT در ترجمه فارسی به انگلیسی (به خصوص در حالت فیلتر شده) به دست آورد. اگرچه سمت مقصد آزمایش ما زبان انگلیسی است، برای بیشتر زبان‌ها، تجزیه گر کم عمق نحوی (برای تولید برچسب قطعه) از تجزیه‌گر نحوی در دسترس‌تر است.

استفاده از فیلتر برای کاهش زمان ترجمه به تنکی زیاد مدل SAMT منجر شد. در حالت کلی، به دلیل ترتیب مختلف کلمات فارسی و انگلیسی و شکل ساختواژی غنی کلمات فارسی تعداد کلمات تراز شده و تعداد قواعد استخراج شده نسبت به جفت زبان‌های اروپایی بسیار کمتر است. از سوی دیگر، مدل SAMT عبارات فاقد بازه متناظر در درخت تجزیه نحوی را با غیرپایانه X برچسب می‌زند. در زبان‌هایی با اختلاف زیاد در ترتیب کلمات، بخش زیادی از عبارات نحوی تراز نمی‌شوند که به تنکی بیشتر مدل SAMT منجر می‌شود.

مراجع

[1] 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.

[2] A. Zollmann and A. Venugopal, "Syntax augmented machine

Volume 1, 2011, pp. 1–11.

[16] S. C. of ICT, “Mizan English-Persian Parallel Corpus,” 2013.

[17] 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 on association for computational linguistics*, 2002, pp. 311–318.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

translation via chart parsing,” in *Proceedings of the Workshop on Statistical Machine Translation*, 2006, pp. 138–141.

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

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

[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] G. M. de Buy Wenniger and K. Sima’an, “Labeling hierarchical phrase-based models without linguistic resources,” *Mach. Transl.*, pp. 1–41, 2016.

[8] Z. He, Q. Liu, and S. Lin, “Improving statistical machine translation using lexicalized rule selection,” in *Proceedings of the 22nd International Conference on Computational Linguistics-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,” 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] 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.

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

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

[14] C. Cherry, “Improved Reordering for Phrase-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-*