Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates

نکته : تمامی عناوین و کلمات قرمز رنگ به صورت لینک های قابل کلیک هستند .

چکیده

واحدهای زیرواژه‌ای (Subword units) روشی مؤثر برای کاهش مشکلات واژگان باز در ترجمه ماشینی عصبی (NMT) هستند. در حالی‌که جملات معمولاً به دنباله‌های منحصربه‌فردی از زیرواژه‌ها تبدیل می‌شوند، تقسیم‌بندی زیرواژه‌ای ذاتاً مبهم است و حتی با استفاده از همان واژگان نیز چندین نوع تقسیم‌بندی ممکن است. سؤال اصلی که در این مقاله بررسی می‌شود این است که آیا می‌توان از این ابهام در تقسیم‌بندی به‌عنوان نوعی نویز برای افزایش مقاومت مدل NMT استفاده کرد یا خیر. در این پژوهش، ما یک روش ساده منظم‌سازی به نام «منظم‌سازی زیرواژه‌ای» (Subword Regularization) معرفی می‌کنیم که در آن مدل در طی آموزش با چندین تقسیم‌بندی زیرواژه‌ای که به‌صورت احتمالی نمونه‌برداری شده‌اند، آموزش داده می‌شود. علاوه بر این، برای بهبود فرآیند نمونه‌برداری زیرواژه‌ها، یک الگوریتم جدید تقسیم‌بندی زیرواژه‌ای مبتنی بر مدل زبانی تک‌واژه‌ای (Unigram Language Model) پیشنهاد می‌کنیم. ما این روش را بر روی چندین پیکره متنی آزمایش کرده‌ایم و نتایج بهبودهای پیوسته‌ای را نشان می‌دهند، به‌ویژه در شرایطی که منابع داده محدود هستند یا متون خارج از حوزه‌ی آموزشی مدل قرار دارند.

۱ مقدمه

مدل‌های ترجمه ماشینی عصبی (Neural Machine Translation, NMT)
(Bahdanau et al., 2014; Luong et al., 2015; Wu et al., 2016; Vaswani et al., 2017)
معمولاً با واژگان ثابتی از کلمات کار می‌کنند، زیرا فرآیند آموزش و استنتاج آن‌ها به شدت به اندازه‌ی واژگان وابسته است. با این حال، محدود کردن اندازه‌ی واژگان باعث افزایش تعداد کلمات ناشناخته می‌شود که این امر، به‌ویژه در شرایطی که واژگان باز هستند، دقت ترجمه را کاهش می‌دهد. یک روش رایج برای حل مسئله‌ی واژگان باز، شکستن کلمات نادر به واحدهای زیرواژه‌ای است
(Schuster and Nakajima, 2012; Chitnis and DeNero, 2015; Sennrich et al., 2016; Wu et al., 2016).
الگوریتم Byte-Pair-Encoding (BPE) (Sennrich et al., 2016) به عنوان استاندارد واقعی تقسیم‌بندی زیرواژه‌ای در بسیاری از سیستم‌های NMT به کار گرفته شده و در چندین وظیفه‌ی مشترک ترجمه، کیفیت بالایی را به دست آورده است
(Denkowski and Neubig, 2017; Nakazawa et al., 2017). تقسیم‌بندی با استفاده از BPE توازن مناسبی میان اندازه‌ی واژگان و کارایی رمزگشایی ایجاد می‌کند و همچنین نیاز به پردازش خاصی برای کلمات ناشناخته را از بین می‌برد. الگوریتم BPE یک جمله را به دنباله‌ای منحصربه‌فرد از زیرواژه‌ها رمزگذاری می‌کند. با این حال، حتی با استفاده از همان واژگان، یک جمله می‌تواند به چندین شکل مختلف به دنباله‌های زیرواژه‌ای تقسیم شود. جدول ۱ مثالی از این موضوع را نشان می‌دهد. در حالی‌که تمام این دنباله‌ها ورودی یکسانی یعنی «Hello World» را نشان می‌دهند، سیستم NMT آن‌ها را به عنوان ورودی‌هایی کاملاً متفاوت در نظر می‌گیرد. این موضوع زمانی آشکارتر می‌شود که دنباله‌های زیرواژه‌ای به دنباله‌های شناسه (id sequences) تبدیل می‌شوند (ستون سمت راست در جدول ۱). این تنوع‌ها می‌توانند نوعی ابهام کاذب در نظر گرفته شوند که ممکن است همیشه در فرآیند رمزگشایی برطرف نشود. در زمان آموزش مدل NMT، وجود چندین گزینه‌ی تقسیم‌بندی باعث می‌شود مدل در برابر نویز و خطاهای تقسیم‌بندی مقاوم‌تر شود، زیرا این تنوع‌ها به‌طور غیرمستقیم به مدل کمک می‌کنند تا قابلیت ترکیب‌پذیری واژه‌ها را یاد بگیرد؛ برای مثال، واژه‌ی «books» می‌تواند به صورت «book» + «s» تجزیه شود. در این پژوهش، ما یک روش جدید منظم‌سازی برای NMT با واژگان باز پیشنهاد می‌کنیم که «منظم‌سازی زیرواژه‌ای» (Subword Regularization) نام دارد و از چندین تقسیم‌بندی زیرواژه‌ای برای افزایش دقت و پایداری مدل NMT استفاده می‌کند. منظم‌سازی زیرواژه‌ای شامل دو زیرمشارکت (زیرنوآوری) زیر است:

- ما یک الگوریتم ساده برای آموزش مدل‌های ترجمه ماشینی عصبی (NMT) پیشنهاد می‌کنیم که امکان ادغام چندین گزینه‌ی تقسیم‌بندی را فراهم می‌سازد. رویکرد ما بر پایه‌ی نمونه‌برداری آنی از داده‌ها (on-the-fly data sampling) پیاده‌سازی شده است و به معماری خاصی از NMT وابسته نیست. روش منظم‌سازی زیرواژه‌ای (Subword Regularization) را می‌توان بدون نیاز به تغییر در ساختار مدل، در هر سیستم NMT مورد استفاده قرار داد.

- ما همچنین یک الگوریتم جدید برای تقسیم‌بندی زیرواژه‌ای مبتنی بر مدل زبانی (Language Model) پیشنهاد می‌کنیم که چندین نوع تقسیم‌بندی را همراه با احتمال مربوط به هر یک ارائه می‌دهد. این مدل زبانی امکان شبیه‌سازی نویزی را فراهم می‌کند که در طی فرآیند تقسیم‌بندی داده‌های واقعی به‌طور طبیعی ایجاد می‌شود.

آزمایش‌های تجربی انجام‌شده بر روی چندین پیکره‌ی متنی با اندازه‌ها و زبان‌های مختلف نشان می‌دهد که منظم‌سازی زیرواژه‌ای (Subword Regularization) نسبت به روشی که تنها از یک دنباله‌ی زیرواژه‌ای استفاده می‌کند، بهبودهای قابل‌توجهی ایجاد می‌کند. علاوه بر این، از طریق آزمایش‌هایی با پیکره‌های خارج از حوزه (out-of-domain corpora) نشان می‌دهیم که منظم‌سازی زیرواژه‌ای باعث افزایش پایداری و مقاومت مدل ترجمه ماشینی عصبی (NMT) می‌شود.

زیرواژه‌ها (علامت _ به معنی فاصله است)	دنباله شناسه واژگان
_Hell/o/_world	۱۳۵۸۶ ۱۳۷ ۲۵۵
_H/ello/_world	۳۲۰ ۷۳۶۳ ۲۵۵
_He/llo/_world	۵۷۹ ۱۰۱۱۵ ۲۵۵
_He/l/l/o/_world	۷ ۱۸۰۸۵ ۳۵۶ ۳۵۶ ۱۳۷ ۲۵۵
_H/el/l/o/_world	۳۲۰ ۵۸۵ ۳۵۶ ۱۳۷ ۷ ۱۲۲۹۵

جدول ۱: چندین دنباله زیرواژه‌ای که جمله یکسان «Hello World» را رمزگذاری می‌کنند.

۲ ترجمه ماشینی عصبی با چندین تقسیم‌بندی زیرواژه‌ای

۱.۲ آموزش **NMT** با نمونه‌برداری آنی زیرواژه‌ها

فرض کنید جمله‌ی مبدأ X و جمله‌ی مقصد Y داده شده‌اند. اجازه دهید $x = (x_1, \cdots, x_M)$ و $y = (y_1, \cdots, y_N)$ دنباله‌های متناظر زیرواژه‌ای باشند که با استفاده از یک تقسیم‌کننده‌ی زیرواژه‌ای پایه‌ای (برای مثال BPE) به‌دست آمده‌اند. مدل ترجمه ماشینی عصبی (NMT)، احتمال ترجمه را به صورت $P(Y|X) = P(y|x)$ مدل می‌کند؛ به‌گونه‌ای که مدل زبان مقصد به‌صورت دنباله‌ای از زیرواژه‌های هدف y_n را با شرط‌گذاری بر تاریخچه‌ی هدف $y_{<n}$ و دنباله‌ی ورودی مبدأ x تولید میکند.

$$P(\mathbf{y} \mid \mathbf{x}; \theta) = \prod_{n=1}^{N} P(y_n \mid \mathbf{x}, y_{<n}; \theta) \quad (۱)$$

در این رابطه، θ مجموعه‌ای از پارامترهای مدل است. یک انتخاب رایج برای پیش‌بینی زیرواژه‌ی y_n ، استفاده از معماری شبکه عصبی بازگشتی (Recurrent Neural Network, RNN) است. با این حال باید توجه داشت که منظم‌سازی زیرواژه‌ای (Subword Regularization) به این معماری خاص محدود نیست و می‌تواند در سایر معماری‌های NMT بدون استفاده از RNN نیز به‌کار گرفته شود؛ برای مثال در مدل‌های (Vaswani et al., 2017; Gehring et al., 2017). مدل NMT با استفاده از روش برآورد بیشینه درست‌نمایی (Maximum Likelihood Estimation) آموزش داده می‌شود، به این معنا که لگاریتم درست‌نمایی $\mathcal{L}(\theta)$ مربوط به پیکره‌ی موازی داده‌شده $D =$ بیشینه می‌شود.

$$
\begin{aligned}
\{\langle X^{(s)}, Y^{(s)}\rangle\}_{s=1}^{|D|} &= \{\langle x^{(s)}, y^{(s)}\rangle\}_{s=1}^{|D|} \\
\theta_{MLE} &= argmax_\theta \mathcal{L}(\theta) \\
where, \mathcal{L}(\theta) &= \sum_{s=1}^{|D|} logP(y^{(s)}|x^{(s)};\theta)
\end{aligned}
\tag{۲}
$$

در اینجا فرض می‌کنیم که جملات مبدأ X و مقصد Y می‌توانند به چندین دنباله‌ی زیرواژه‌ای تقسیم شوند، به‌طوری‌که احتمال تقسیم‌بندی آن‌ها به‌ترتیب با P(x|X) و P(y|Y) نمایش داده می‌شود. در روش منظم‌سازی زیرواژه‌ای (Subword Regularization)، مجموعه پارامترها یعنی θ با استفاده از درست‌نمایی حاشیه‌ای‌شده به صورت فرمول (۳) بهینه‌سازی می‌شود.

$$
\mathcal{L}_{\text{marginal}}(\theta) = \sum_{s=1}^{|D|} \mathbb{E}_{\substack{\mathbf{x}\sim P(\mathbf{x}|X^{(s)}) \\ \mathbf{y}\sim P(\mathbf{y}|Y^{(s)})}} \left[\log P(\mathbf{y} \mid \mathbf{x};\theta)\right] \tag{۳}
$$

بهینه‌سازی دقیق رابطه‌ی (۳) امکان‌پذیر نیست، زیرا تعداد تقسیم‌بندی‌های ممکن با افزایش طول جمله به‌صورت نمایی رشد می‌کند. ما برای تقریب رابطه‌ی (۳)، از تعداد محدودی k دنباله‌ی نمونه‌برداری‌شده از توزیع‌های P(x|X) و P(y|Y) به‌ترتیب استفاده می‌کنیم.

$$
\begin{aligned}
\mathcal{L}_{\text{marginal}}(\theta) &\cong \frac{1}{k^2} \sum_{s=1}^{|D|} \sum_{i=1}^{k} \sum_{j=1}^{k} \log P(\mathbf{y}_j \mid \mathbf{x}_i;\theta), \\
\mathbf{x}_i &\sim P(\mathbf{x} \mid X^{(s)}), \quad \mathbf{y}_j \sim P(\mathbf{y} \mid Y^{(s)}).
\end{aligned}
\tag{۴}
$$

برای سادگی، مقدار $k = 1$ در نظر گرفته می‌شود. فرآیند آموزش در مدل‌های ترجمه ماشینی عصبی (NMT) معمولاً به‌صورت آموزش برخط (online training) انجام می‌شود تا کارایی افزایش یابد. در این روش، پارامتر θ به‌صورت تکراری با توجه به زیرمجموعه‌های کوچک‌تری از مجموعه داده D (که به آن مینی‌بتچ یا mini-batch گفته می‌شود) بهینه‌سازی می‌شود. زمانی که تعداد تکرارها کافی باشد، نمونه‌برداری زیرواژه‌ای از طریق همان نمونه‌برداری داده‌ها در فرآیند آموزش برخط انجام می‌شود، که حتی در حالتی که $k = 1$ باشد، تقریب مناسبی از رابطه‌ی (۳) به‌دست می‌دهد. با این حال، باید توجه داشت که دنباله‌ی زیرواژه‌ها در هر بار به‌روزرسانی پارامترها، به‌صورت آنی (on-the-fly) نمونه‌برداری می‌شود.

۲.۲ رمزگشایی

در مرحله‌ی رمزگشایی (Decoding) در مدل ترجمه ماشینی عصبی (NMT)، تنها جمله‌ی خام مبدأ X در اختیار داریم. یک رویکرد ساده برای رمزگشایی این است که ترجمه از بهترین تقسیم‌بندی x^* انجام شود ؛ به گونه ای که x^* بیشترین احتمال $P(x|X)$ را داشته باشد ، یعنی : $x^* = argmax_* P(x|X)$. علاوه بر این، می‌توان از n-best تقسیم‌بندی‌های احتمالی $P(x|X)$ برای در نظر گرفتن چندین گزینه‌ی تقسیم‌بندی استفاده کرد. به‌طور دقیق‌تر، اگر n تقسیم‌بندی برتر $(x_1, \cdots, x_n)$ داده شود، بهترین ترجمه‌ی y^* به‌گونه‌ای انتخاب می‌شود که امتیاز زیر را بیشینه کند.

$$score(\mathbf{x}, \mathbf{y}) = \log P(\mathbf{y} \mid \mathbf{x}) / |\mathbf{y}|^{\lambda}, \tag{۵}$$

که در آن $|y|$ تعداد زیرواژه‌ها در y است و $\lambda \in \mathbb{R}^+$ پارامتری است که برای تنبیه کردن جملات کوتاه‌تر به کار می‌رود. مقدار λ با استفاده از داده‌های توسعه (development data) بهینه‌سازی می‌شود. در این مقاله، ما این دو الگوریتم را به ترتیب one-best decoding و n-best decoding می‌نامیم.

۳ بخش‌بندی زیرواژه‌ای با مدل زبانی

۱.۳ الگوریتم Byte-Pair-Encoding (BPE)

الگوریتم Byte-Pair-Encoding (BPE) (Sennrich et al., 2016; Schuster and Nakajima, 2012) یک روش بخش‌بندی زیرواژه‌ای است که به‌طور گسترده در بسیاری از سامانه‌های ترجمه ماشینی عصبی (NMT) استفاده می‌شود. در این روش، ابتدا کل جمله به کاراکترهای منفرد تقسیم می‌شود. سپس پرتکرارترین جفت‌های متوالی از کاراکترها به صورت پی‌درپی با هم ادغام می‌شوند تا زمانی که اندازه‌ی واژگان مطلوب حاصل شود. فرآیند بخش‌بندی زیرواژه‌ای برای جمله‌ی آزمون نیز از طریق اعمال همان عملیات ادغام انجام می‌گیرد. مزیت الگوریتم بخش‌بندی BPE در این است که می‌تواند به‌طور مؤثری میان اندازه‌ی واژگان و اندازه‌ی گام (یعنی تعداد توکن‌هایی که برای رمزگذاری جمله مورد نیاز است) تعادل برقرار کند. الگوریتم BPE عملیات ادغام را صرفاً بر اساس فراوانی کاراکترها آموزش می‌دهد. زیررشته‌های پرتکرار در مراحل اولیه با هم ترکیب می‌شوند و در نتیجه، کلمات متداول به‌صورت یک نماد یکتا باقی می‌مانند. در مقابل، کلماتی که شامل ترکیب‌های نادر از کاراکترها هستند، به واحدهای کوچک‌تری مانند زیررشته‌ها یا کاراکترها تقسیم می‌شوند. بنابراین، حتی با یک اندازه‌ی واژگان ثابت و کوچک (معمولاً بین ۱۶هزار تا ۳۲هزار)، تعداد نمادهای مورد نیاز برای رمزگذاری یک جمله افزایش چشمگیری نخواهد داشت، که این ویژگی مهمی برای فرایند رمزگشایی کارآمد به شمار می‌رود. با این حال، یکی از معایب الگوریتم BPE این است که بر پایه‌ی جایگزینی حریصانه و قطعی نمادها بنا شده است و در نتیجه، نمی‌تواند چندین بخش‌بندی مختلف را همراه با احتمالات آن‌ها ارائه دهد. اعمال الگوریتم BPE بر روی تنظیم‌سازی زیرواژه‌ای (subword regularization) که به احتمالات بخش‌بندی $P(x|X)$ وابسته است، کار ساده‌ای نیست.

۲.۳ مدل زبانی تک‌واژه‌ای (Unigram Language Model)

در این مقاله، ما یک الگوریتم جدید برای بخش‌بندی زیرواژه‌ای بر اساس مدل زبانی تک‌واژه‌ای (unigram language model) پیشنهاد می‌کنیم که قادر است چندین بخش‌بندی زیرواژه‌ای را همراه با احتمالات آن‌ها تولید کند. مدل زبانی تک‌واژه‌ای فرض می‌کند که هر زیرواژه به صورت مستقل رخ می‌دهد؛ در نتیجه، احتمال توالی زیرواژه‌ها $x = (x_1, \ldots, x_M)$ به‌صورت حاصل‌ضرب احتمالات وقوع هر زیرواژه $p(x_i)$ تعریف می‌شود.

$$
\begin{aligned}
P(x) &= \prod_{i=1}^{M} p(x_i), \\
\forall i \quad x_i \in \mathcal{V}, \quad & \sum_{x \in \mathcal{V}} p(x) = 1
\end{aligned}
\tag{۶}
$$

که در آن $\mathcal{V}$ مجموعه‌ی واژگان از پیش تعیین‌شده است. بخش‌بندی با بیشترین احتمال x^* برای جمله‌ی ورودی X به‌صورت زیر تعریف می‌شود:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{S}(X)} P(\mathbf{x}) \tag{۷}$$

که در آن $\mathcal{S}(X)$ مجموعه‌ای از گزینه‌های بخش‌بندی است که از جمله‌ی ورودی X ساخته شده‌اند. بخش‌بندی بهینه x^* با استفاده از الگوریتم Viterbi (Viterbi, 1967) به‌دست می‌آید. اگر مجموعه‌ی واژگان $\mathcal{V}$ داده شده باشد، احتمال وقوع زیرواژه‌ها $p(x_i)$ از طریق الگوریتم EM برآورد می‌شود؛ الگوریتمی که با بیشینه‌سازی درست‌نمایی نهایی $\mathcal{L}$ (با فرض اینکه $p(x_i)$ متغیرهای پنهان هستند) کار می‌کند.

$$\mathcal{L} = \sum_{s=1}^{|D|} \log\left(P(X^{(s)})\right) = \sum_{s=1}^{|D|} \log\left(\sum_{\mathbf{x} \in \mathcal{S}(X^{(s)})} P(\mathbf{x})\right) \tag{۸}$$

با این حال، در شرایط واقعی، مجموعه‌ی واژگان $\mathcal{V}$ نیز ناشناخته است. از آنجا که بهینه‌سازی هم‌زمان مجموعه‌ی واژگان و احتمال وقوع آن‌ها مسئله‌ای غیرقابل حل (intractable) است، در اینجا ما به‌دنبال یافتن آن‌ها از طریق الگوریتم تکراری زیر هستیم.

۱. به‌صورت ابتکاری (heuristically)، یک واژگان اولیه‌ی نسبتاً بزرگ از پیکره‌ی آموزشی (training corpus) ایجاد کنید.

۲. مراحل زیر را تکرار کنید تا زمانی که اندازه‌ی واژگان $|\mathcal{V}|$ به اندازه‌ی مطلوب برسد:

(آ) با ثابت نگه‌داشتن مجموعه‌ی واژگان، تابع $p(x)$ را با استفاده از الگوریتم EM بهینه کنید.

(ب) مقدار $loss_i$ را برای هر زیرواژه‌ی x_i محاسبه کنید، به‌طوری که $loss_i$ نشان‌دهنده‌ی میزان کاهش درست‌نمایی $\mathcal{L}$ در صورت حذف زیرواژه‌ی x_i از واژگان فعلی باشد.

(ج) نمادها را بر اساس $loss_i$ مرتب کنید و $\eta\%$ از زیرواژه‌ها را که مقدار $loss_i$ بالاتری دارند، حفظ کنید (برای مثال $\eta = 80$). توجه داشته باشید که همیشه زیرواژه‌هایی که از یک کاراکتر تشکیل شده‌اند نگه‌داشته می‌شوند تا از بروز حالت خارج از واژگان (out-of-vocabulary) جلوگیری شود.

روش‌های متعددی برای آماده‌سازی واژگان اولیه وجود دارد. انتخاب طبیعی، استفاده از اجتماع تمام کاراکترها و پرتکرارترین زیررشته‌ها در پیکره است. زیررشته‌های پرتکرار را می‌توان با استفاده از الگوریتم Enhanced Suffix Array (Nong et al., 2009) در زمان $O(T)$ و فضای $O(20T)$ شمارش کرد، که در آن T اندازه‌ی پیکره است. مشابه با (Sennrich et al., 2016)، زیرواژه‌هایی که از مرز میان کلمات عبور می‌کنند، در نظر گرفته نمی‌شوند. از آنجا که واژگان نهایی $\mathcal{V}$ شامل تمام کاراکترهای منفرد موجود در پیکره است، بخش‌بندی مبتنی بر کاراکتر نیز در مجموعه‌ی گزینه‌های بخش‌بندی $\mathcal{S}(X)$ گنجانده می‌شود. به عبارت دیگر، بخش‌بندی زیرواژه‌ای با مدل زبانی تک‌واژه‌ای (unigram language model) را می‌توان به‌عنوان ترکیب احتمالاتی از بخش‌بندی مبتنی بر کاراکتر، زیرواژه و کلمه در نظر گرفت.

۳.۳ نمونه‌برداری زیرواژه‌ای (Subword Sampling)

تنظیم‌سازی زیرواژه‌ای (Subword regularization) در هر مرحله‌ی به‌روزرسانی پارامتر، یک بخش‌بندی زیرواژه‌ای را از توزیع $P(x \mid X)$ نمونه‌برداری می‌کند. یک روش مستقیم برای نمونه‌برداری تقریبی، استفاده از بخش‌بندی‌های $l-best$ است. به‌طور دقیق‌تر، ابتدا l بخش‌بندی برتر بر اساس احتمال $P(x \mid X)$ به‌دست می‌آوریم. جست‌وجوی $l-best$ به‌صورت خطی و با استفاده از الگوریتم Forward-DP Backward-A* (Nagata, 1994) انجام می‌شود. سپس یک بخش‌بندی x_i از توزیع چندجمله‌ای زیر نمونه‌برداری می‌شود $P(x_i|X) \cong P(x_i)^\alpha / \sum_{i=1}^{l} P(x_i)^\alpha$ که در آن $\alpha \in \mathbb{R}^+$ ابرپارامتری است که میزان نرمی (smoothness) توزیع را کنترل می‌کند. مقدار کوچک‌تر α منجر به نمونه‌برداری از توزیعی یکنواخت‌تر می‌شود، در حالی که مقدار بزرگ‌تر α تمایل به انتخاب بخش‌بندی Viterbi دارد. به‌صورت نظری، با قرار دادن $l \to \infty$ می‌توان تمام بخش‌بندی‌های ممکن را در نظر گرفت. با این حال، افزایش صریح مقدار l عملی نیست، زیرا تعداد گزینه‌های ممکن به‌صورت نمایی نسبت به طول جمله افزایش می‌یابد. برای نمونه‌برداری دقیق از میان تمام بخش‌بندی‌های ممکن، از الگوریتم فیلترسازی رو به جلو و نمونه‌برداری رو به عقب (Forward-Filtering and Backward-Sampling, FFBS) (Scott, 2002) استفاده می‌کنیم؛ این الگوریتم نوعی از برنامه‌ریزی پویا است که در ابتدا برای آموزش مدل پنهان مارکوف بیزی (Bayesian Hidden Markov Model) معرفی شده بود. در الگوریتم FFBS، تمام گزینه‌های بخش‌بندی در یک ساختار شبکه‌ای فشرده (lattice) نمایش داده می‌شوند که در آن هر گره نشان‌دهنده‌ی یک زیرواژه است. در مرحله‌ی اول، الگوریتم FFBS مجموعه‌ای از احتمالات پیشرو (forward probabilities) را برای تمام زیرواژه‌ها در شبکه محاسبه می‌کند؛ این مقادیر، احتمال رسیدن به هر زیرواژه‌ی خاص w را نشان می‌دهند. در مرحله‌ی دوم ، با پیمایش گره‌های شبکه از انتهای جمله تا ابتدای آن، زیرواژه‌ها برای هر شاخه به‌صورت بازگشتی و بر اساس احتمالات پیشرو نمونه‌برداری می‌شوند.

۴ کارهای مرتبط

منظم‌سازی با نویز (Regularization by noise) یک تکنیک شناخته‌شده و پرمطالعه در شبکه‌های عصبی عمیق است. یکی از نمونه‌های معروف آن، dropout (Srivastava et al., 2014) است که طی آن، بخشی از واحدهای پنهان در حین آموزش به‌صورت تصادفی غیرفعال می‌شوند. Dropout به‌عنوان نوعی آموزش گروهی (ensemble training) تحلیل می‌شود، که در آن مدل‌های گوناگون بر روی زیرمجموعه‌های متفاوتی از داده‌ها آموزش می‌بینند. به‌طور مشابه، منظم‌سازی زیرواژه‌ای (Subword regularization) مدل را بر روی ورودی‌هایی آموزش می‌دهد که به‌صورت تصادفی از جملات ورودی اصلی نمونه‌برداری شده‌اند، و از این رو می‌توان آن را نوعی از آموزش گروهی در نظر گرفت. ایده‌ی تزریق نویز (noise injection) پیش‌تر در چارچوب خودرمزگذارهای حذف‌نویز (De-noising Auto-Encoders, DAEs) (Vincent et al., 2008) به‌کار رفته است، که در آن نویز به ورودی‌ها افزوده می‌شود و مدل برای بازسازی ورودی‌های اصلی آموزش می‌بینند. چندین پژوهش از خودرمزگذارهای حذف‌نویز در حوزه‌ی پردازش زبان طبیعی استفاده کرده‌اند. (Lample et al., 2017; Artetxe et al., 2017) به‌صورت مستقل، استفاده از DAEها را در چارچوب یادگیری دنباله‌به‌دنباله (sequence-to-sequence learning) پیشنهاد کردند؛ در این روش، ترتیب کلمات جمله‌ی ورودی به‌صورت تصادفی تغییر می‌کند و مدل آموزش می‌بیند تا جمله‌ی اصلی را بازسازی کند. تکنیک آن‌ها در ترجمه‌ی ماشینی بدون نظارت (unsupervised machine translation) به‌کار گرفته شده است تا رمزگذار بتواند ترکیب‌پذیری (compositionality) جملات ورودی را به‌درستی بیاموزد. روش Word Dropout (Iyyer et al., 2015) یک رویکرد ساده برای نمایش «کیسه‌ی کلمات» (bag-of-words) است. در این روش، بردار نهفتگی (embedding) یک دنباله از واژه‌ها با میانگین‌گیری از بردارهای نهفتگی واژه‌های آن دنباله محاسبه می‌شود. در فرآیند word dropout، برخی واژه‌ها به‌صورت تصادفی از مجموعه حذف می‌شوند، و سپس میانگین بردارهای باقیمانده گرفته می‌شود. به این ترتیب، مدل می‌تواند برای هر ورودی X، تا $2^{|X|}$ توالی توکن متفاوت را مشاهده کند. (Belinkov and Bisk, 2017) به بررسی آموزش ترجمه‌ی ماشینی عصبی مبتنی بر کاراکتر (character-based NMT) با استفاده از نویز مصنوعی می‌پردازند؛ در این روش، ترتیب کاراکترهای هر واژه به‌صورت تصادفی تغییر داده می‌شود. همچنین، (Xie et al., 2017) یک مدل زبانی مقاوم بر پایه‌ی شبکه‌های عصبی بازگشتی (RNN language model) ارائه می‌دهد که از ترکیب تصادفی با مدل زبانی تک‌واژه‌ای (unigram language model) برای بهبود پایداری مدل استفاده می‌کند. ایده‌ی اصلی و انگیزه‌ی پشت منظم‌سازی زیرواژه‌ای (Subword regularization) مشابه پژوهش‌های پیشین است. برای افزایش پایداری مدل (robustness)، در این روش‌ها با ایجاد تغییرات تصادفی در نمایش درونی جملات، نوعی نویز به ورودی‌ها تزریق می‌شود. با این حال، این رویکردهای پیشین معمولاً به روش‌های ابتکاری (heuristics) برای تولید نویز مصنوعی وابسته‌اند، که همواره بازتاب‌دهنده‌ی نویزهای واقعی در مرحله‌ی آموزش و استنتاج نیستند. علاوه بر این، این روش‌ها تنها بر جملات مبدأ (رمزگذار یا encoder) قابل اعمال هستند، زیرا سطح ظاهری جمله‌ها را به‌صورت غیرقابل بازگشت تغییر می‌دهند. در مقابل، منظم‌سازی زیرواژه‌ای با استفاده از مدل زبانی زیربنایی، توالی‌های زیرواژه‌ای مصنوعی تولید می‌کند تا نویزها و خطاهای بخش‌بندی را به‌صورت واقع‌گرایانه‌تر شبیه‌سازی کند. از آنجا که منظم‌سازی زیرواژه‌ای بر پایه‌ی تبدیل قابل بازگشت (invertible conversion) است، می‌توان آن را با اطمینان هم بر جملات مبدأ و هم بر جملات مقصد اعمال کرد. منظم‌سازی زیرواژه‌ای (Subword regularization) را همچنین می‌توان به‌عنوان نوعی افزایش داده (data augmentation) در نظر گرفت. در این روش، یک جمله‌ی ورودی به چندین توالی معادل و پایدار تبدیل می‌شود؛ فرآیندی که شباهت زیادی به روش‌های افزایش داده در وظایف دسته‌بندی تصویر دارد ، مانند وارونه‌سازی

تصادفی (random flipping)، تغییر شکل (distorting) یا برش تصادفی (cropping) تصاویر. چندین پژوهش بر روی ابهام‌های مربوط به بخش‌بندی در مدل‌سازی زبانی تمرکز داشته‌اند. روش تجزیه دنباله‌های نهفته (Latent Sequence Decompositions, LSDs) (Chan et al., 2016)، نگاشت بین ورودی و خروجی را با حاشیه‌گیری (marginalizing) بر روی تمام بخش‌بندی‌های ممکن یاد می‌گیرد. هر دو روش یعنی LSDs و منظم‌سازی زیرواژه‌ای (Subword regularization) فرض نمی‌کنند که برای جمله، بخش‌بندی از پیش تعیین‌شده‌ای وجود دارد، بلکه با استفاده از تکنیک مشابه حاشیه‌گیری، چندین بخش‌بندی ممکن را در نظر می‌گیرند. تفاوت این دو در آن است که منظم‌سازی زیرواژه‌ای، چندین بخش‌بندی را از طریق مدل زبانی جداگانه و با نمونه برداری آنی زیرواژه‌ای (on-the-fly subword sampling) وارد می‌کند. این رویکرد باعث می‌شود مدل ساده‌تر شده و از معماری‌های ترجمه‌ی ماشینی عصبی (NMT) مستقل باشد. مدل‌های شبکه‌به‌دنباله (Lattice-to-Sequence Models) (Su et al., 2017; Sperber et al., 2017) گسترشی طبیعی از مدل‌های دنباله‌به‌دنباله (Sequence-to-Sequence Models) هستند که در آن‌ها عدم‌قطعیت موجود در ورودی‌ها از طریق ساختار شبکه‌ای (lattice) نمایش داده می‌شود. در این روش، شبکه با استفاده از گونه‌ای تغییریافته از TreeLSTM (Tai et al., 2015) رمزگذاری می‌شود، که مستلزم تغییر در معماری مدل است. علاوه بر این، در حالی که منظم‌سازی زیرواژه‌ای (Subword regularization) هم بر جملات مبدأ و هم بر جملات مقصد اعمال می‌شود، مدل‌های شبکه‌به‌دنباله قادر به مدیریت ابهام‌های موجود در سمت مقصد نیستند. مدل ترکیبی واژه/کاراکتر (Mixed Word/Character Model) (Wu et al., 2016) برای رفع مشکل واژه‌های خارج از واژگان (Out-of-Vocabulary Problem) با استفاده از یک واژگان ثابت طراحی شده است. در این مدل، واژه‌های خارج از واژگان در یک نماد واحد UNK ادغام نمی‌شوند، بلکه به توالی‌ای از کاراکترها تبدیل می‌شوند که هر کاراکتر دارای پیشوندی ویژه برای نشان دادن موقعیت آن در واژه است. مشابه با الگوریتم BPE، این مدل نیز جمله را به یک دنباله‌ی ثابت و یکتا رمزگذاری می‌کند؛ بنابراین، بخش‌بندی‌های متعدد در این روش در نظر گرفته نمی‌شوند.

۵ آزمایش ها

۱.۵ تنظیمات

ما مجموعه‌ای از آزمایش‌ها را با استفاده از چندین پیکره‌ی زبانی در اندازه‌ها و زبان‌های مختلف انجام دادیم. جدول ۲ داده‌های ارزیابی مورد استفاده در این پژوهش را خلاصه می‌کند IWSLT15، IWSLT17، KFTT، ASPEC، WMT14 .

پیکره‌های IWSLT15/17 و KFTT نسبتاً کوچک هستند، اما شامل طیف گسترده‌تری از زبان‌ها با ویژگی‌های زبانی متفاوت می‌باشند. این مجموعه‌ها برای ارزیابی ویژگی مستقل از زبان (language-agnostic) در منظم‌سازی زیرواژه‌ای مناسب هستند. پیکره‌های ASPEC و WMT14 (en↔de) در اندازه‌ی متوسط قرار دارند، در حالی‌که پیکره‌ی WMT14 (en↔cs) نسبتاً بزرگ است و شامل بیش از ۱۰ میلیون جمله‌ی موازی می‌باشد. ما از مدل GNMT (Wu et al., 2016) به‌عنوان پیاده‌سازی سیستم ترجمه‌ی ماشینی عصبی (NMT) برای تمامی آزمایش‌ها استفاده کردیم. به‌طور کلی تنظیمات و روش آموزش توضیح‌داده‌شده در (Wu et al., 2016) دنبال شد، با این تفاوت که تنظیمات بر اساس اندازه‌ی هر پیکره تغییر داده شدند. جدول ۲ ابرپارامترهای مورد استفاده در هر آزمایش را نشان می‌دهد. به‌عنوان تنظیم مشترک، احتمال dropout برابر با 0.2 در نظر گرفته شد. برای برآورد پارامترها، از ترکیب دو الگوریتم Adam (Kingma and Adam, 2014) و SGD استفاده کردیم. هر دو پارامتر مربوط به نرمال‌سازی طول (length normalization) و

جریمه‌ی همگرایی (converge penalty) روی مقدار 0.2 تنظیم شدند (بخش ۷ در (Wu et al., 2016) را ببینید). اندازه‌ی پرتو (beam size) در فرایند رمزگشایی برابر با 4 در نظر گرفته شد. داده‌ها پیش از آموزش مدل‌های زیرواژه‌ای، با استفاده از توکنایزر Moses پیش‌پردازش شدند. با این حال، باید توجه داشت که در زبان‌های چینی و ژاپنی مرزهای واژه‌ای به‌صورت صریح وجود ندارند و توکنایزر Moses نیز جملات را به واژه‌ها تقسیم نمی‌کند. بنابراین، در این زبان‌ها بخش‌بندی زیرواژه‌ای تقریباً از روی جملات خام و بدون بخش‌بندی آموزش داده شده است. برای ارزیابی، از معیار امتیاز BLEU حساس به حروف بزرگ و کوچک (Papineni et al., 2002) استفاده کردیم. از آنجا که جملات خروجی در زبان‌های چینی و ژاپنی به‌صورت واژه‌ای بخش‌بندی نشده‌اند، پیش از محاسبه‌ی امتیاز BLEU، آن‌ها را به‌ترتیب با استفاده از کاراکترها برای چینی و KyTea برای ژاپنی بخش‌بندی کردیم. بخش‌بندی BPE به‌عنوان سیستم پایه مورد استفاده قرار گرفت. ما سه سیستم آزمایشی را با استراتژی‌های نمونه‌برداری متفاوت ارزیابی کردیم: ۱. بخش‌بندی زیرواژه‌ای مبتنی بر مدل زبانی تک‌واژه‌ای (Unigram Language Model) بدون منظم‌سازی زیرواژه‌ای، با $l = 1$. ۲. بخش‌بندی زیرواژه‌ای با منظم‌سازی زیرواژه‌ای، با $l = 64$ و $\alpha = 0.1$. ۳. حالت سوم با $l = \infty$ و $\alpha = 0.2/0.5$ (0.2 برای IWSLT و 0.5 برای سایر پیکره‌ها). این پارامترهای نمونه‌برداری از طریق آزمایش‌های مقدماتی تعیین شدند. حالت $l = 1$ برای مقایسه‌ی مستقیم بین BPE و مدل زبانی تک‌واژه‌ای در نظر گرفته شد. علاوه بر این، ما دو روش رمزگشایی را نیز با یکدیگر مقایسه کردیم: رمزگشایی تک‌بهترین(one-best decoding) و رمزگشایی چندبهترین (n-best decoding) (بخش ۲.۲ را ببینید). از آنجا که BPE قادر به تولید چندین بخش‌بندی مختلف نیست، برای آن تنها رمزگشایی تک‌بهترین مورد ارزیابی قرار گرفت. در نتیجه، ما در مجموع ۷ سیستم را برای هر جفت‌زبان مقایسه کردیم: $(1+3\times 2)$.

۲.۵ نتایج اصلی

جدول ۳ نتایج آزمایش‌های ترجمه را نشان می‌دهد. در ابتدا، همان‌طور که در جدول مشاهده می‌شود، بخش‌بندی BPE و مدل زبانی تک‌واژه‌ای (unigram language model) بدون منظم‌سازی زیرواژه‌ای $l = 1$ امتیازهای BLEU تقریباً مشابهی دارند. این نتیجه قابل انتظار است، زیرا هر دو روش BPE و مدل زبانی تک‌واژه‌ای بر پایه‌ی الگوریتم‌های فشرده‌سازی داده بنا شده‌اند. مشاهده می‌شود که منظم‌سازی زیرواژه‌ای $l > 1$ باعث بهبود قابل توجهی در امتیازهای BLEU در تمامی جفت‌زبان‌ها (به میزان حدود ۱ تا ۲ امتیاز) شده است، به‌جز در مجموعه‌داده‌ی WMT14 (en→cs). این بهبودها به‌ویژه در شرایط با منابع محدود (پیکره‌های IWSLT و KFTT) چشمگیرتر هستند. می‌توان نتیجه گرفت که تأثیر مثبت افزایش داده از طریق منظم‌سازی زیرواژه‌ای در محیط‌های کم‌منبع مؤثرتر است، که این ویژگی در میان سایر روش‌های منظم‌سازی نیز رایج است. از نظر الگوریتم نمونه‌برداری، حالت $(l = \infty, \alpha = 0.2/0.5)$ کمی بهتر از حالت $(l = 64, \alpha = 0.1)$ در پیکره‌ی IWSLT عمل کرده است، اما در مجموعه‌داده‌های بزرگ‌تر نتایج تقریباً مشابهی ارائه می‌دهد. تحلیل دقیق‌تر این نتایج در بخش ۵.۵ ارائه شده است. علاوه بر بهبودهای حاصل از منظم‌سازی زیرواژه‌ای، استفاده از رمزگشایی چندبهترین (n-best decoding) نیز در بسیاری از جفت‌زبان‌ها باعث افزایش بیشتر امتیاز BLEU شده است. با این حال باید توجه داشت که منظم‌سازی زیرواژه‌ای برای رمزگشایی چندبهترین ضروری است؛ زیرا در غیاب آن $l = 1$، امتیاز BLEU در برخی از جفت‌زبان‌ها کاهش یافته است. این نتیجه نشان می‌دهد که در صورتی که چندین بخش‌بندی در مرحله‌ی آموزش مورد بررسی قرار نگیرند، رمزگشا در مواجهه با بخش‌بندی‌های متعدد دچار سردرگمی بیشتری می‌شود.

۳.۵ نتایج با پیکره‌ی خارج از دامنه

برای بررسی تأثیر منظم‌سازی زیرواژه‌ای در محیط‌های بازتر و متنوع‌تر از نظر دامنه، سیستم‌ها را با داده‌های داخلی خارج از دامنه (out-of-domain) که شامل چندین ژانر مختلف از جمله وب، پتنت‌ها و لاگ‌های جست‌وجو بودند، مورد ارزیابی قرار دادیم. لازم به ذکر است که مقایسه با پیکره‌های KFTT و ASPEC انجام نشد، زیرا دامنه‌ی این دو پیکره بسیار خاص است و ارزیابی‌های اولیه نشان داد که امتیاز BLEU بر روی داده‌های خارج از دامنه بسیار پایین (کمتر از ۵) است. نتایج در جدول ۴ ارائه شده است. در مقایسه با بهبودهایی که در ارزیابی‌های درون‌دامنه (جدول ۳) مشاهده شد، منظم‌سازی زیرواژه‌ای در تمام دامنه‌های پیکره، بهبود قابل‌توجه‌تری (حدود +۲ امتیاز) ایجاد کرده است. نکته‌ی جالب این است که حتی در مجموعه‌داده‌های بزرگ آموزشی مانند WMT14 نیز بهبود در همان سطح مشاهده می‌شود، در حالی‌که این مجموعه‌ها در داده‌های درون‌دامنه تنها افزایش اندکی نشان داده بودند. این نتیجه به‌طور قوی از این ادعا پشتیبانی می‌کند که منظم‌سازی زیرواژه‌ای در تنظیمات باز و متنوع از نظر دامنه بسیار مفیدتر است.

۵،۴ مقایسه با سایر الگوریتم‌های بخش‌بندی

جدول ۵ مقایسه‌ای میان الگوریتم‌های مختلف بخش‌بندی ارائه می‌دهد، از جمله: مدل‌های واژه‌ای (word-based)، کاراکتری (character-based)، ترکیبی واژه/کاراکتر (Wu et al., 2016)، BPE (Sennrich et al., 2016) و مدل تک‌واژه‌ای ما (unigram model) با یا بدون منظم‌سازی زیرواژه‌ای. امتیازهای BLEU برای مدل‌های واژه‌ای، کاراکتری و ترکیبی واژه/کاراکتر از پژوهش (Wu et al., 2016) نقل شده‌اند. از آنجا که زبان آلمانی زبانی با ساختار صرفی غنی است و برای مدل‌های واژه‌ای به واژگان بسیار بزرگی نیاز دارد، الگوریتم‌های مبتنی بر زیرواژه نسبت به مدل واژه‌ای بیش از ۱ امتیاز BLEU بهبود عملکرد نشان می‌دهند. در میان الگوریتم‌های مبتنی بر زیرواژه، مدل زبانی تک‌واژه‌ای همراه با منظم‌سازی زیرواژه‌ای بهترین امتیاز BLEU (۲۵.۰۴) را به‌دست آورد، که نشان‌دهنده‌ی کارایی و اثربخشی استفاده از چندین بخش‌بندی زیرواژه‌ای است.

پیکره	جفت‌زبان	تعداد جملات				پارامترها	
		آموزش	توسعه	آزمون	واژگان	ابعاد LSTM / نهفتگی	لایه‌های (رمزگذار+رمزگشا) LSTM
IWSLT15	vi ↔ en	۱۳۳ k	۱۵۵۳	۱۲۶۸	۱۶ k	۵۱۲	۲+۲
IWSLT15	zh ↔ en	۲۰۹ k	۸۸۷	۱۲۶۱	۱۶ k	۵۱۲	۲+۲
IWSLT17	fr ↔ en	۲۳۲ k	۸۹۰	۱۲۱۰	۱۶ k	۵۱۲	۲+۲
IWSLT17	ar ↔ en	۲۳۱ k	۸۸۸	۱۲۰۵	۱۶ k	۵۱۲	۲+۲
KFTT	ja ↔ en	۴۴۰ k	۱۱۶۶	۱۱۶۰	۸ k	۵۱۲	۶+۶
ASPEC	ja ↔ en	۲ M	۱۷۹۰	۱۸۱۲	۱۶ k	۵۱۲	۶+۶
WMT14	de ↔ en	۴.۵ M	۳۰۰۰	۳۰۰۳	۳۲ k	۱۰۲۴	۸+۸
WMT14	cs ↔ en	۱۵ M	۳۰۰۰	۳۰۰۳	۳۲ k	۱۰۲۴	۸+۸

جدول ۲: جزئیات مجموعه‌داده‌های ارزیابی

پیکره	جفت‌زبان	پایه (BPE)	تک‌بهترین			چندبهترین (n=64)		
			$l=1$	$l=64$ $\alpha=0.1$	$l=\infty$ $\alpha=0.2/0.5$	$l=1$	$l=64$ $\alpha=0.1$	$l=\infty$ $\alpha=0.2/0.5$
IWSLT15	en→vi	۲۵.۶۱	۲۵.۴۹	۲۷.۶۸ *	۲۷.۷۱ *	۲۵.۳۳	۲۸.۱۸ *	۲۸.۴۸ *
	vi→en	۲۲.۴۸	۲۲.۳۲	۲۴.۷۳ *	۲۶.۱۵ *	۲۲.۰۴	۲۴.۶۶ *	۲۶.۳۱ *
	en→zh	۱۶.۷۰	۱۶.۹۰	۱۹.۳۶ *	۲۰.۳۳ *	۱۶.۷۳	۲۰.۱۴*	۲۱.۳۰ *
	zh→en	۱۵.۷۶	۱۵.۸۸	۱۷.۷۹ *	۱۶.۹۵ *	۱۶.۲۳	۱۷.۷۵ *	۱۷.۲۹ *
IWSLT17	en→fr	۳۵.۵۳	۳۵.۳۹	۳۶.۷۰ *	۳۶.۳۶*	۳۵.۱۶	۳۷.۶۰ *	۳۷.۰۱ *
	fr→en	۳۳.۸۱	۳۳.۷۴	۳۵.۵۷ *	۳۵.۵۴ *	۳۳.۶۹	۳۶.۰۷ *	۳۶.۰۶ *
	en→ar	۱۳.۰۱	۱۳.۰۴	۱۴.۹۲ *	۱۵.۵۵ *	۱۲.۲۹	۱۴.۹۰ *	۱۵.۳۶ *
	ar→en	۲۵.۹۸	۲۷.۰۹ *	۲۸.۴۷ *	۲۹.۲۲ *	۲۷.۰۸ *	۲۹.۰۵ *	۲۹.۲۹ *
KFTT	en→ja	۲۷.۸۵	۲۸.۹۲ *	۳۰.۳۷ *	۳۰.۰۱ *	۲۸.۵۵ *	۳۱.۴۶ *	۳۱.۴۳ *
	ja→en	۲۱.۳۷	۲۱.۴۶	۲۲.۳۳ *	۲۲.۰۴ *	۲۱.۳۷	۲۲.۴۷ *	۲۲.۶۴ *
ASPEC	en→ja	۴۰.۶۲	۴۰.۶۶	۴۱.۲۴ *	۴۱.۲۳ *	۴۰.۸۶	۴۱.۵۵ *	۴۱.۸۷ *
	ja→en	۲۶.۵۱	۲۶.۷۶	۲۷.۰۸ *	۲۷.۱۴ *	۲۷.۴۹ *	۲۷.۷۵ *	۲۷.۸۹ *
WMT14	en→de	۲۴.۵۳	۲۴.۵۰	۲۵.۰۴ *	۲۴.۷۴ *	۲۲.۷۳	۲۵.۰۰ *	۲۴.۵۷
	de→en	۲۸.۰۱	۲۸.۶۵ *	۲۸.۸۳ *	۲۹.۳۹ *	۲۸.۲۴	۲۹.۱۳ *	۲۹.۹۷ *
	en→cs	۲۵.۲۵	۲۵.۵۴	۲۵.۴۱	۲۵.۲۶	۲۴.۸۸	۲۵.۴۹	۲۵.۳۸
	cs→en	۲۸.۷۸	۲۸.۸۴	۲۹.۶۴ *	۲۹.۴۱ *	۲۵.۷۷	۲۹.۲۳ *	۲۹.۱۵ *

جدول ۳: نتایج اصلی (BLEU%) (l: اندازه نمونه‌گیری در SR، α: پارامتر هموارسازی). علامت * بیانگر تفاوت آماری معنادار با سطح معناداری $p < 0.05$ نسبت به مقادیر مبنا با استفاده از روش بازنمونه‌گیری بوت‌استرپ است (Koehn, 2004). همین نشان در جدول‌های ۴ و ۶ نیز به‌کار رفته است.

۵.۵ تأثیر ابرپارامترهای نمونه‌گیری

تنظیم‌سازی زیرواژه‌ای دارای دو ابرپارامتر است: l: اندازه مجموعه نمونه‌گیری، و α: ثابت هموارسازی. شکل ۱ امتیازهای BLEU را برای مقادیر مختلف این ابرپارامترها در مجموعه داده IWSLT15 (en→vi) نشان می‌دهد. در وهله نخست مشاهده می‌شود که نقطه اوج امتیازهای BLEU نسبت به پارامتر هموارسازی α به اندازه نمونه‌گیری l بستگی دارد. این نتیجه قابل انتظار است، زیرا در حالت $l = \infty$ فضای جست‌وجو گسترده‌تر از $l = 64$ است و برای نمونه‌گیری از توالی‌هایی نزدیک به توالی ویتربی $\mathbf{x}^*$، مقدار بزرگ‌تری از α باید تنظیم شود. مشاهده جالب دیگر این است که در مقدار $\alpha = 0.0$، کاهش کارایی به‌ویژه در حالت $l = \infty$ مشاهده می‌شود. هنگامی که $\alpha = 0.0$ باشد، احتمال بخش‌بندی $P(x|X)$ عملاً نادیده گرفته می‌شود و تنها یک بخش‌بندی به‌صورت یکنواخت نمونه‌گیری می‌گردد. این نتیجه نشان می‌دهد که نمونه‌گیریِ جهت‌دار بر پایه یک مدل زبانی، در بازنمایی نویز واقعی در فرایند ترجمه عملی مؤثر است. به‌طور کلی، مقادیر بزرگ‌تر l موجب منظم‌سازی (regularization) قوی‌تر و کارآمدتر می‌شوند و در محیط‌های با منابع محدود مانند IWSLT عملکرد بهتری دارند. با این حال، برآورد مقدار α در این حالت حساس‌تر است و زمانی که α بیش از حد کوچک باشد، عملکرد حتی از حالت پایه نیز پایین‌تر می‌آید. برای تضعیف اثر منظم‌سازی و جلوگیری از انتخاب پارامترهای نامعتبر، منطقی‌تر است که در زبان‌های با منابع بالا از مقدار $l = 64$ استفاده شود.

دامنه (اندازه)	پیکره	جفت‌زبان	پایه (BPE)	پیشنهادی (SR)
وب (۵هزار)	IWSLT15	en→vi	۱۳.۸۶	۱۷.۳۶ *
		vi→en	۷.۸۳	۱۱.۶۹ *
		en→zh	۹.۷۱	۱۳.۸۵ *
		zh→en	۵.۹۳	۸.۱۳ *
	IWSLT17	en→fr	۱۶.۰۹	۲۰.۰۴ *
		fr→en	۱۴.۷۷	۱۹.۹۹ *
	WMT14	en→de	۲۲.۷۱	۲۶.۰۲ *
		de→en	۲۶.۴۲	۲۹.۶۳ *
		en→cs	۱۹.۵۳	۲۱.۴۱ *
		cs→en	۲۵.۹۴	۲۷.۸۶ *
پتنت (۲هزار)	WMT14	en→de	۱۵.۶۳	۲۵.۷۶ *
		de→en	۲۲.۷۴	۳۲.۶۶ *
		en→cs	۱۶.۷۰	۱۹.۳۸ *
		cs→en	۲۳.۲۰	۲۵.۳۰ *
پرس‌وجو (۲هزار)	IWSLT15	en→zh	۹.۳۰	۱۲.۴۷ *
		zh→en	۱۴.۹۴	۱۹.۹۹ *
	IWSLT17	en→fr	۱۰.۷۹	۱۰.۹۹
		fr→en	۱۹.۰۱	۲۳.۹۶ *
	WMT14	en→de	۲۵.۹۳	۲۹.۸۲ *
		de→en	۲۶.۲۴	۳۰.۹۰ *

جدول ۴: نتایج به‌دست‌آمده با پیکره‌های خارج از حوزه (out-of-domain corpus)
(l = ∞, $\alpha = 0.2$: IWSLT15/17, $l = 64$, $\alpha = 0.1$: others, one-best decoding)

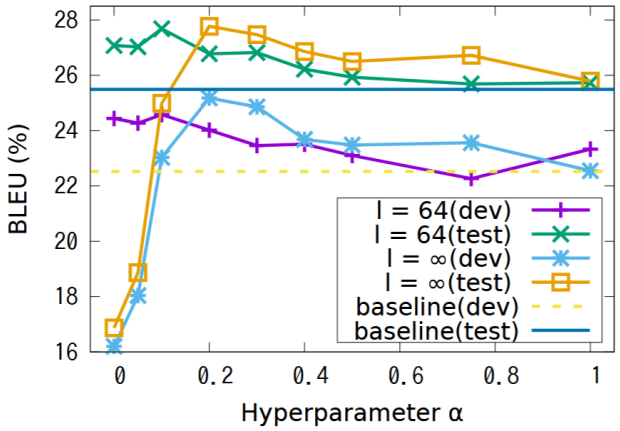
BLEU	مدل
۲۳.۱۲	Word
۲۲.۶۲	Character (512 nodes)
۲۴.۱۷	Mixed Word/Character
۲۴.۵۳	BPE
۲۴.۵۰	Unigram w/o SR ($l = 1$)
۲۵.۰۴	Unigram w/ SR ($l = 64$, $\alpha = 0.1$)

جدول ۵: مقایسه الگوریتم‌های مختلف بخش‌بندی (WMT14 en→de)

اگرچه به‌طور کلی مشاهده می‌شود که ابرپارامترهای بهینه به‌صورت تقریبی از طریق برآورد داده‌های کنارگذاشته‌شده (held-out estimation) قابل پیش‌بینی هستند، اما همچنان این پرسش باز باقی می‌ماند که چگونه باید اندازه بهینه l را در فرایند نمونه‌گیری زیرواژه‌ای انتخاب کرد.

۶.۵ نتایج مربوط به منظم‌سازی یک‌سویه

جدول ۶ امتیازهای BLEU را در حالتی خلاصه می‌کند که منظم‌سازی زیرواژه‌ای تنها بر روی جمله مبدأ یا جمله مقصد اعمال شده است تا مشخص شود کدام جزء (رمزگذار یا رمزگشا) تأثیر بیشتری می‌پذیرد. همان‌طور که انتظار می‌رود، مشاهده می‌شود که امتیازهای BLEU در منظم‌سازی یک‌سویه کمتر از حالت منظم‌سازی کامل است. با این حال باید توجه داشت که منظم‌سازی یک‌سویه همچنان اثرات مثبتی دارد. این نتیجه نشان می‌دهد که منظم‌سازی زیرواژه‌ای نه‌تنها برای معماری‌های رمزگذار□رمزگشا مفید است، بلکه در سایر وظایف پردازش زبان طبیعی (NLP) که تنها از یکی از این دو مؤلفه استفاده می‌کنند ـ مانند طبقه‌بندی متون (Iyyer et al., 2015) و تولید شرح تصاویر (Vinyals et al., 2015) نیز قابل به‌کارگیری است.



شکل ۱: تأثیر ابرپارامترهای نمونه‌گیری

نوع منظم‌سازی	en→vi	vi→en	en→ar	ar→en
بدون منظم سازی (خط پایه)	۲۵.۴۹	۲۲.۳۲	۱۳.۰۴	۲۷.۰۹
مبدأ فقط	۲۶.۰۰	*۲۳.۰۹	۱۳.۴۶	*۲۸.۱۶
مقصد فقط	۲۶.۱۰	*۲۳.۶۲	*۱۴.۳۴	*۲۷.۸۹
مقصد و مبدأ	*۲۷.۶۸	*۲۴.۷۳	*۱۴.۹۲	*۲۸.۴۷

جدول ۶: مقایسه راهبردهای مختلف منظم‌سازی (IWSLT15/17، $l = 64$ ،$\alpha = 0.1$)

۶ نتیجه گیری

در این پژوهش، روشی ساده برای منظم‌سازی، با عنوان منظم‌سازی زیرواژه‌ای (subword regularization) برای ترجمه ماشینی عصبی (NMT) ارائه شد، بدون آنکه نیازی به تغییر در معماری شبکه وجود داشته باشد. ایده اصلی آن، افزایش مجازی داده‌های آموزشی از طریق نمونه‌گیری زیرواژه‌ای در حین آموزش است، که به بهبود دقت و همچنین افزایش پایداری مدل‌های ترجمه ماشینی منجر می‌شود. علاوه بر این، برای نمونه‌گیری بهتر زیرواژه‌ها، یک الگوریتم جدید برای بخش‌بندی زیرواژه‌ها بر پایه مدل زبانی تک‌واژه‌ای (unigram language model) پیشنهاد شده است. آزمایش‌ها بر روی پیکره‌های مختلف با اندازه‌ها و زبان‌های گوناگون نشان داد که منظم‌سازی زیرواژه‌ای بهبودهای چشمگیری ایجاد می‌کند، به‌ویژه در محیط‌های با منابع محدود و حوزه‌های باز (open-domain). مسیرهای امیدبخش برای پژوهش‌های آینده شامل به‌کارگیری منظم‌سازی زیرواژه‌ای در سایر وظایف پردازش زبان طبیعی مبتنی بر معماری رمزگذار ، رمزگشا است؛ برای مثال، تولید گفت‌وگو (Vinyals and Le, 2015) و خلاصه‌سازی خودکار متون (Rush et al., 2015). در مقایسه با ترجمه ماشینی، این وظایف معمولاً داده‌های آموزشی کافی ندارند، بنابراین فضای زیادی برای بهبود عملکرد با استفاده از منظم‌سازی زیرواژه‌ای وجود دارد. همچنین، قصد داریم کاربرد منظم‌سازی زیرواژه‌ای را در سایر حوزه‌های یادگیری ماشین بررسی کنیم؛ از جمله در خودرمزگذارهای حذف‌نویز (Vincent et al., 2008) و یادگیری خصمانه (Goodfellow et al., 2015).