Self-supervised Dynamic Programming Encoding for Neural Machine Translation

Haiyue Song[†‡] Raj Dabre[‡] Chenhui Chu[†] Sadao Kurohashi[†] Eiichiro Sumita[‡]
[†]Kyoto University, Kyoto, Japan
[‡]National Institute of Information and Communications Technology, Kyoto, Japan
{song, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
{raj.dabre, eiichiro.sumita}@nict.go.jp

1 Introduction

Neural machine translation (NMT) [1] is known to give state-of-the-art translations for a variety of language pairs. Sub-word segmentation [2, 3] is one of the key reasons behind it as it enables the NMT model to simulate an infinite vocabulary, thereby eliminating the out-of-vocabulary problem. Among all proposed sub-word segmentation methods, byte pair encoding (BPE) [2] is the most commonly used approach which is used to learn sub-word segementation rules from a monolingual corpus. However, BPE uses a greedy search method to segment words and often ignores the sequence-to-sequence nature of NMT. To solve this, researchers have recently proposed a dynamic programming encoding (DPE) method [4] which uses parallel corpora and a neural network to perform sub-word segmentation. The advantage of DPE is that it employs a non-greedy search method in addition to being aware of the translation process by maximizing the probability of generating the target sentence. Although DPE is slower than BPE, it consistently outperforms BPE and related methods [5]. Compared to BPE, DPE tends to give linguistically intuitive segmentations which can potentially aid in understanding the working of NMT models. Thus far, DPE has been studied in resource-rich situations but its performance and working in low-resource settings is rather unknown.

The advantage of DPE is also its weakness, especially in a low-resource setting where parallel corpora sizes are too small to reliably train neural models. To this end, we decided to study how DPE works in a low-resource setting where monolingual corpora and self-supervision based training objectives are known to be extremely helpful [6, 7]. This in turn motivates our novel Self-Supervised DPE (SSDPE) method that simulates a parallel corpus by duplicating a monolingual corpus and then training a DPE model using a self-supervised objective. We use denoising methods involving token masking to prevent the encoder-decoder from learning trivial sequence-to-sequence mappings. The resultant SSDPE model needs only monolingual corpora, in addition to being source language agnostic. By analyzing the segmentation of words in different contexts, we found each word is consistently segmented regardless of its contexts, e.g. DPE has the context-free property. Motivated by this, we propose a novel Word-level DPE (WDPE) that uses only monolingual word-frequency table to train a word segmenter. In the decoding phase, each unique word only needs to be decoded once, which lowers the decoding time complexity.

To verify the utility of SSDPE and WDPE, we experiment with 11 Asian languages to English translation using the the low-resource Asian Language Treebank (ALT) corpus [8] belonging to the Wikinews domain. We observe that they are comparable, if not better than DPE and BPE by up to maximum 0.3 BLEU average improvement for 11 language pairs. Our analyses of the compared methods shows that DPE is actually source language agnostic and context-free which explains why SSDPE and WDPE works as well as DPE even without a parallel corpus and context.

2 Related Work

Our work focuses on sub-word tokenization for NMT where methods such as BPE [2], WPM [9] and Sentencepiece Model (SPM) [3] are some of the most popular ones. These methods use greedy search to learn sub-word merge rules, a fundamental component of sub-word methods, in a deterministic way. Stochastic variants of BPE called

BPE-drop [5] and of SPM called SPM-regularization [10] enable multiple sub-word segmentations of the same word thereby making the NMT model more robust.

BPE, SPM and the simple character level [11] methods, despite being effective, are greedy search based methods that tend to give non-intuitive sub-word segmentations. DPE [4] addresses this but its reliance on parallel corpora makes it unsuitable for low-resource settings. Our work fills in this gap by enabling DPE segmentation using only monolingual corpora that are easy to obtain and leverage through self-supervision methods for NLP applications [6, 12].

3 Proposed Method

3.1 Background: DPE

DPE segmentation is performed using an encoder-decoder model where $\boldsymbol{x}$ is the BPE segmented source sentence input to the encoder, $\boldsymbol{y}_c$ is the character segmented target sentence input to the decoder. The decoder produces $\boldsymbol{y}$, the DPE segmented target sentence where $z = (z_1, ..., z_{M+1})$ denotes some segmentation indices of $\boldsymbol{y}$ and $\mathcal{Z}_y$ denotes all possible segmentations of $\boldsymbol{y}$. Note that the decoder's output sub-word vocabulary is obtained using BPE. [4] optimized the exact log marginal likelihood as follows to learn DPE:

$$\begin{aligned} &log\ p(\boldsymbol{y}) = \\ &log \sum_{z \in \mathcal{Z}_y} exp \sum_{i=1}^{|z|} log\ p(\boldsymbol{y}_{z_i, z_{i+1}} | ..., y_{z_{i-1}, z_i}, \boldsymbol{x}) \end{aligned} \quad (1)$$

3.2 SSDPE: Self-Supervised DPE

We propose SSDPE that relies on replacing $\boldsymbol{x}$ with $\boldsymbol{y}_M$ in Equation 1, which is the BPE segmented target sentence $\boldsymbol{y}$ with certain tokens being masked thereby making it a monolingual approach. Consequently, SSDPE training resembles monolingual training tasks such as MASS [6] and Masked Language Model (MLM) in BERT [12]. In this paper, we consider the following:

1. SSDPE-LM: A trivial case where $\boldsymbol{y}_M = \boldsymbol{y}$.

2. SSDPE-MASS: Similar to [6], half the tokens in $\boldsymbol{y}$ are masked to give $\boldsymbol{y}_M$. The masked tokens are consecutive so as to enforce the model to learn sequence level information.

3. SSDPE-Mask: Motivated by [12], we keep 15% of the training sentences unmasked to bridge the gap between train phase and inference phase. Of the remaining 85% training sentences, $\boldsymbol{y}$, each token has a 15% chance to be masked thereby yielding $\boldsymbol{y}_M$.

3.3 WDPE: Word-level DPE

Based on SSDPE, we propose WDPE where the input $\boldsymbol{y}_M$ and target $\boldsymbol{y}$ are single words rather than sentences. We split sentences into words in the training corpus to keep the frequencies of words the same. We consider three variant methods as in SSDPE: WDPE-LM, WDPE-MASS and WDPE-Mask.

4 Experimental Settings

4.1 Datasets and Preprocessing

We used the ALT multi-way parallel dataset [8] consisting of 18,088, 1,000 and 1,018 sentences in the train, dev and test set respectively for 12 languages: Bengali (bg), English (en), Filipino (fil), Bahasa Indonesia (id), Japanese (ja), Khmer (km), Lao (lo), Malay (ms), Myanmar (my), Thai (th), Vietnamese (vi) and simplified Chinese (zh). We chose this corpus as it represents a realistic extremely low-resource setting and its multi-way nature enables us to effectively analyze the properties of our proposed method. As SSDPE needs only monolingual data, we experimented with 50,000 randomly selected English sentences from news commentary corpus[1)] belonging to the same news domain as ALT, to train SSDPE models. For WDPE, we split the same news commentary data into one word one line format.

We used Moses tokenizer [13] to tokenize Vietnamese, Malay, Indonesian, Filipino and English, Indic NLP to tokenize Bengali, deepcut [14] for Thai, LaoNLP[2)] for Lao, Juman++ [15] for Japanese, Stanford-tokenizer [16] for Chinese, and tokenized data from WAT for Khmer and Myanmar [17].

4.2 NMT Model Settings

We used the fairseq framework [18] with the Transformer architecture. We did hyper-parameter tuning to determine optimal vocabulary sizes, number of encoder-decoder layers and attention heads, as low-resource settings are quite sensitive to these parameters. As DPE, SSDPE

[1)] http://data.statmt.org/news-commentary/v14/

[2)] https://github.com/wannaphong/LaoNLP

#	Method	bg	fil	id	ja	km	lo	ms	my	th	vi	zh	avg
	baseline												
1	SPM	10.33	23.71	25.49	9.94	16.59	*11.13	27.72	12.77	12.35	18.58	*11.74	16.40
2	BPE	9.90	23.09	25.70	9.42	17.12	10.56	28.19	12.11	13.52	19.94	*12.21	16.52
3	DPE	10.09	24.04	26.66	9.93	17.36	10.56	27.89	12.00	14.05	20.06	10.72	16.67
	proposed: SSDPE												
4	SSDPE-LM	10.10	23.49	25.15	10.27	17.19	10.49	28.35	11.53	13.70	21.37	*11.79	16.68
5	SSDPE-MASS	*10.50	24.28	25.37	10.74	17.00	*11.03	28.25	11.50	14.13	*21.36	*12.11	*__16.93__
6	SSDPE-Mask	9.35	24.01	25.78	10.07	*17.58	10.88	28.56	11.90	14.01	*21.45	*12.27	16.90
	proposed: WDPE												
7	WDPE-LM	10.07	24.11	26.45	10.23	*17.52	10.57	28.12	12.20	14.24	21.20	*11.85	*__16.96__
8	WDPE-MASS	*10.42	24.51	26.42	*11.17	*18.01	*11.09	28.60	10.53	13.94	21.15	10.97	*__16.98__
9	WDPE-Mask	*10.38	23.79	25.35	10.26	17.21	10.55	28.34	12.31	14.17	21.37	*12.00	*16.88
	analysis												
10	DPE-ms	9.03	24.26	26.26	10.35	16.92	10.57	27.89	12.11	13.44	*21.50	*12.26	16.78

Table 1: Asian languages to English MT using various subword segmentation methods. Rows 1-3 show baseline results. Rows 4-6 show proposed SSDPE results. Rows 7-9 shows results using words as training data. Row 10 uses English segmentation from Ms-En DPE model. *indicates statistically significant difference ($p < 0.05$) from DPE method.

Method	ja	id
SPM	12.58	26.01
BPE	12.69	28.08
DPE	13.46	**29.29**
SSDPE-Mask	**14.04**	28.46

Table 2: Results of English to Asian language MT. Japanese and Indonesian SSDPE segmenters used 100k news commentary monolingual sentences.

	# of lines	# of tokens	time (sec)
DPE	18k	615k	341.9
SSDPE-MASS	18k	615k	347.7
WDPE-MASS	30k	81k	58.3

Table 3: Decoding speed of DPE, SSDPE and WDPE methods of ALT-train set, averaged from 10 runs.

and WDPE use BPE vocabulary, the optimal vocabulary size is determined based on NMT performance only using BPE. Consequently, the optimal combination was: vocabulary size of 8,000, 6 layer encoder, 6 layer decoder and 1 attention head (except for Bengali, Filipino, Japanese and Lao, where 4 encoder layers were sufficient). We keep the same vocabulary size for SPM. We used one GPU with batch-size of 1,024 tokens. We used the ADAM optimizer [19] with betas (0.9, 0.98), warm up of 4,000 steps followed by decay, and perform early stopping based on the development set BLEU. Dropout of 0.1 and label smoothing of 0.1 is used. We used layer normalization for both the encoder and decoder. Decoding is done with a beam size of 12 and length penalty of 1.4. We reported sacreBLEU and performed statistical significance test.

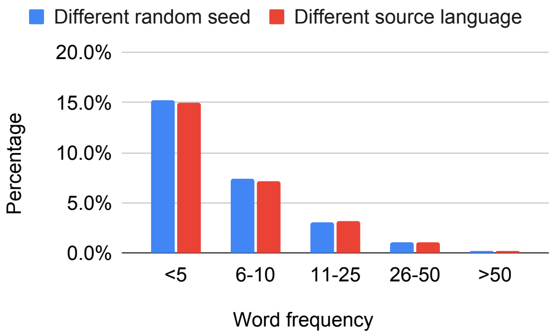

Figure 1: WSDR between DPE segmenters trained on Ja-En data with different random seeds (red), DPE segmenters trained on Zh-En and Ja-En (blue).

5 Results and Analyses

5.1 Main Results: SSDPE and WDPE

First, as shown in rows 1-6 of Table 1, for Asian language to English NMT, DPE gives results that are as good as if not better than BPE and SPM. However, our SSDPE and WDPE methods can improve over DPE by 0.26 and 0.31 BLEU respectively (significant at $p < 0.05$), averaged over 11 translation directions. For SSDPE methods, SSDPE-MASS and SSDPE-Mask give segmentations that yield better translations as compared to SSDPE-LM, which shows the advantage of masking based training approaches over trivial copy based methods. WDPE methods show better performance than SSDPE, which suggests that sentence contextual information is unnecessary for subword segmentation. For WDPE, a trivial LM method can give almost the best result. Next, Table 2 shows that SSDPE-Mask is better than DPE for English to Japanese NMT but worse than DPE for English to Indonesian NMT.

Overall, our results show that in a low-resource setting, a parallel corpus dependent DPE model can simply be replaced with a monolingual corpus dependent SSDPE model and still improve translation quality. Further, a word-frequency table is enough to train the segmenter rather than sentence-level data.

5.2 Speed Analysis

Because our WDPE method only needs to decode each unique word once in a corpus, it shows a much faster speed than DPE as well as SSDPE. As shown in Table 3, the speed of WDPE is about 5.9x faster than DPE and SSDPE when decoding the ALT-train set.

5.3 Source Language Agnostic Analysis

We show the DPE method is source language agnostic by both the result of NMT and the analysis of word segmentations.

As the ALT corpus is multi-way, we can use the English segmentations obtained by the DPE model from one language pair for NMT task of another language pair. When we used English segmentations from the Malay-English DPE model and used it to train NMT models for Asian language to English translation, we saw (row 10 in Table 1) that the translation scores are barely affected.

To further investigate this issue, we propose word set difference rate (WSDR) between two segmenters to compute the probability that they generate different segmentations for one word. For a word W with frequency $nword$ which has n possible segmentations, let seg_i and $freq_i$ be the segmentation and frequency of the i^{th} segmentation. We define $S = \bigcup_{i=1 \ldots n}[(seg_i, freq_i)]$ to be the set of all possible segmentations where $\sum_{i=1}^{n} freq_i = nword$ holds true. Word difference rate (WDR) is defined as:

$$WDR = \frac{\sum_{i=1}^{|S_1|} \sum_{j=1}^{|S_2|} freq_i * freq_j * (1_{seg_i \neq seg_j})}{nword^2}$$

where S_1 and S_2 are segmentations of W generated by two segmenters. Then, word set difference rate (WSDR) for a set of words is defined as:

$$WSDR = \frac{\sum_{i=1}^{|words|} WDR(word_i) * nword_i}{\sum_{i=1}^{|words|} nword_i}$$

Figure 1 shows the WSDR of sets of words with different frequency ranges in the ALT corpus. Comparing the red bars (2 Japanese–English DPE segmenters with random seeds) against the blue bars (Japanese–English and Chinese–English DPE segmenters), it is clear that the WSDR rates are similar. This means that DPE model is actually source language agnostic. We believe this to be the main reason behind why the SSDPE, which uses masked language as input, works as well as DPE.

5.4 Context Agnostic Analysis

WDPE works due to the context agnostic property which means the segmentation of one word is consistent regardless of which sentence it is in.

We set S_1 equals to S_2 and calculated WSDR of sets of words with different frequency. We found for any set, the WSDR is less than 1%, which means the segmented result of one word is almost not affected by its context in the sentence. Furthermore, removing the context slightly improve the segmentation quality, shwon by the comparison of WDPE and SSDPE in Table 1.

6 Conclusion

We proposed novel SSDPE and WDPE methods for subword segmentation. Experimental results show NMT using proposed methods are either comparable to or significantly better than NMT using BPE and DPE. The WDPE method shows a faster decoding speed compared with the original DPE method. Our analyses show source language agnostic and context agnostic property of DPE. Our future work will focus on the performance of our methods in resource rich MT tasks as well as on developing a statistical SSDPE method which will be substantially faster.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, p. arXiv:1409.0473, September 2014.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[4] Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. Dynamic programming encoding for subword segmentation in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3042–3051, Online, July 2020. Association for Computational Linguistics.

[5] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1882–1892, Online, July 2020. Association for Computational Linguistics.

[6] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *arXiv e-prints*, p. arXiv:1905.02450, May 2019.

[7] Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, Haiyue Song, and Sadao Kurohashi. Jass: Japanese-specific sequence to sequence pre-training for neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 3683–3691, Marseille, France, May 2020. European Language Resources Association.

[8] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1574–1578, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[9] M. Schuster and K. Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152, 2012.

[10] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Jörg Tiedemann. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 141–151, Avignon, France, April 2012. Association for Computational Linguistics.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[14] Rakpong Kittinaradorn, Korakot Chaovavanich, Titipat Achakulvisut, Kittinan Srithaworn, Pattarawat Chormai, Chanwit Kaewkasi, Tulakan Ruangrong, and Krichkorn Oparad. DeepCut: A Thai word tokenization library using Deep Neural Network., September 2019.

[15] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 54–59, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[16] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[17] Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. Nova: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 18, No. 2, p. 17, 2018.

[18] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, p. arXiv:1412.6980, December 2014.