Unsupervised Word Segmentation Using Minimum Description Length for Neural Machine Translation

Hao Wang Yves Lepage
Graduate School of Information, Production and Systems,
Waseda University
{oko_ips@ruri., yves.lepage}@waseda.jp

概 要

For Chinese and Japanese, there is no distinct word boundary. Word segmentation is widely applied as a pre-processing step in neural machine translation (NMT) pipelines. However, conventional segmenters probably produce massive rare words. NMT cannot make use of such rare words because it commonly limits the vocabulary to most high-frequency words. In this paper, we investigate unsupervised word segmentation using the principle of Minimum Description Length (MDL). We propose a novel two-phase MDL-based method for word segmentation. Experimental results show that our method improves over the strong baseline (monolingual segmenters, e.g., Juman/Stanford Segmenter) for the WAT Japanese–Chinese and Chinese–Japanese translation tasks by up to 1.5 and 2 BLEU points, respectively.

1 Introduction

For the languages without distinct word boundaries, e.g., Chinese and Japanese, word segmentation is widely applied as a pre-processing step in many tasks of natural language processing (NLP). Differing from phrase-based statistical machine translation (SMT) [8], in which makes use of phrase pairs for translation, neural machine translation [9] treats words as atomic units for processing. NMT pipeline for these languages requires word vectors as the input for the network to train the neural model. There are many publicly available segmenters, e.g., Juman, KyTea, Mecab for Japanese and ICTCLAS, Stanford Segmenter for Chinese. Translation results vary with different segmentation tools used, which mainly because segmentation consistency and granularity of those tools are different. [2] show that segmentation consistency and granularity will affect the final SMT results. For NMT, there is a similar conclusion can be drawn from the recent Workshop on Asian Translation (WAT) [10], a simple change in segmenters does not make any big influence on NMT systems.

However, conventional segmenters are prone to generate massive rare words, and most of the low-frequency words will be discarded during training. In this paper, we investigate and tackle the rare word/out-of-vocabulary (OOV) problem within the scope of word segmentation. We propose a novel unsupervised segmentation approach with fixed vocabulary for NMT. Firstly, based on the principle of minimum description length [5], the size of vocabulary will continue to grow in an iterative procedure. Secondly, the inferred codebook allows reuse for word segmentation.

2 OOV Problem in NMT

NMT achieves state-of-the-art performance in large-scale translation tasks. It trains single neural network with a large parallel corpus. We use Figure 1 to illustrate the NMT system used in this paper. Our baseline system is an NMT with attention mechanism [9] which follows the encoder-decoder architecture. Both encoder and decoder are recurrent neural networks (RNN) with a Long Short-Term Mem-

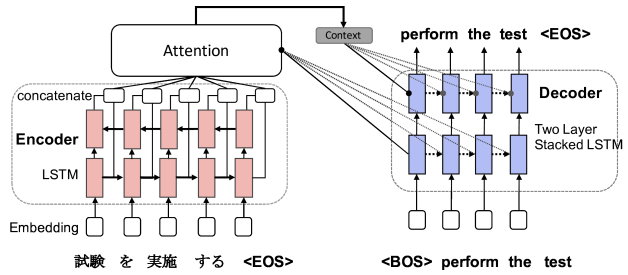

図 1: Bidirectional LSTM encoder-decoder architecture with attention mechanism for NMT.

ory (LSTM). Given the source sentence, the decoder generates one target word at a time, which tries to find the target word with the maximal conditional probability among all target words in the vocabulary. Hence, NMT models commonly limit the vocabulary to 50k∼80k most frequent words to control the translation quality. However, the translation of OOV word is simply handled by converting these words into a single $\langle unk \rangle$ symbol. Hence, there exists an obvious problem that NMT is unable to deal with OOV words. Previous work which explores the problem mainly divides into two categories: Some approaches aim to solve the OOV problem directly. Other approaches aim to is to pre-split the rare words into higher frequency subwords [12, 13]. These approaches provide a good balance between flexibility of single characters and the efficiency of full words, which have exhibited impressive results in morphologically rich languages.

3 MDL-based Segmentation

3.1 Minimum description length

Minimum description length has been previously used in various NLP tasks. For example, grammar induction [4], word segmentation [1, 14], translation model compression [3]. Given a set of data $\mathcal{D}$, the MDL principle aims at finding the minimal model (i.e., codebook) Φ which can describe the $\mathcal{D}$. We formalize MDL inference as:

$$\hat{\Phi} = \underset{\Phi}{arg\,min}\ \mathrm{DL}(\mathcal{D}, \Phi) \tag{1}$$

$$= \underset{\Phi}{arg\,min}\ \mathrm{DL}(\mathcal{D}|\Phi) + \mathrm{DL}(\Phi) \tag{2}$$

The objective function divides into two components, the model description length $\mathrm{DL}(\Phi)$ and data description length $\mathrm{DL}(\Phi)$.

$$\mathrm{DL}(\Phi) = \sum_{w \in \Phi} len(w) \tag{3}$$

Given the codebook Φ, we can segment the data. The total data description length is calculated as:

$$\mathrm{DL}(\mathcal{D}|\Phi) = -\sum_{w \in \Phi} \#w(\log \#w - \log N) \tag{4}$$

where $\#w$ is the count number of word w (coding entry). N it the count number of all tokens in data. len is the length of characters of the lexicon.

3.2 Proposed method

For Chinese and Japanese, character vocabulary size is further smaller than the size of fixed vocabulary for NMT. Consider the initialization stage, and each character is an entry in the codebook. To reduce the description length, given two adjacent characters w_1, w_2, we try to insert a new entry into codebook which a bigram w_1w_2 by merge w_1 with w_2. A greedy method to minimizing DL is finding an new longer entry that has the maximal $\triangle$DL and updates codebook recursively.

$$\triangle \mathrm{DL} = \mathrm{DL}(\tilde{\mathcal{D}}, \tilde{\Phi}) - \mathrm{DL}(\mathcal{D}, \Phi) \tag{5}$$

$$= \triangle\, \mathrm{DL}(\tilde{\Phi}, \Phi) + \triangle \mathrm{DL}(\tilde{\mathcal{D}}, \mathcal{D}) \tag{6}$$

Each operation of updating codebook, i.e., inserting a longer entry and deleting shorter unused entries if necessary, should reduce the description length at each time. $\tilde{\Phi}$ and $\tilde{\mathcal{D}}$ is the new model and data after insertion respectively. We employ an iterative updating procedure for codebook inference. Figure 2 describes the details . The most difficult in Figure 2 is computing the description length changes. For the model description length difference between codebook $\tilde{\Phi}$ and Φ,

$$\triangle \mathrm{DL}(\tilde{\Phi}, \Phi) = \begin{cases} len(w_1w_2) & , \{w_1, w_2\} \subset \tilde{\Phi} \\ len(w_1w_2) - len(w_1) - len(w_2) & , \{w_1, w_2\} \not\subset \tilde{\Phi} \\ len(w_1w_2) - len(w_1) & , w_1 \notin \tilde{\Phi} \\ len(w_1w_2) - len(w_2) & , w_2 \notin \tilde{\Phi} \end{cases} \tag{7}$$

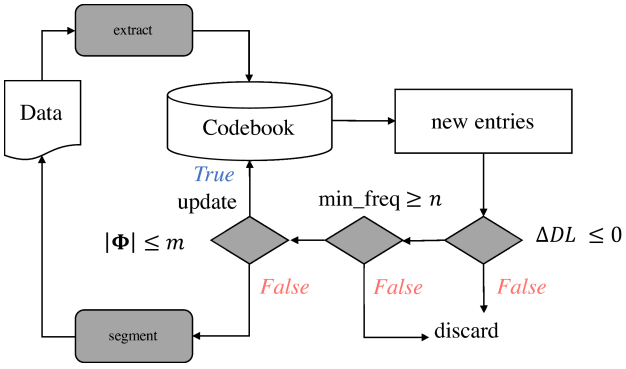

図 2: Iterative procedure for inference of codebook.

The data change is computed as following:

$$\begin{aligned}\triangle \mathrm{DL}(\tilde{\mathcal{D}}, \mathcal{D}) = \mathrm{DL}(\tilde{\mathcal{D}}|w_1 w_2) \\ + \triangle \mathrm{DL}(w_1) + \triangle \mathrm{DL}(w_2) + O(1)\end{aligned} \tag{8}$$

The term of $\triangle \mathrm{DL}(w_1)$ can be rewritten as:

$$\begin{aligned}\triangle \mathrm{DL}(w_1) = \#w_1 \times \log(\frac{\#w_1}{N}) - (\#w_1 - \#w_1 w_2) \\ \times \log(\frac{(\#w_1 - \#w_1 w_2}{N - \#w_1 w_2})\end{aligned} \tag{9}$$

The estimation of the cost for w_2 is analogous to w_1. Additionally, the insertion operation also affects the frequency of other entries in codebook, the bias is computed as follow:

$$O(1) = (N - \#w_1 - \#w_2) \times [\log(\frac{N - \#w_1 w_2}{N})] \tag{10}$$

[14] point out that retrieving word indices in the corpus is a challenging work and replacing on the entire corpus is prohibitive. We use *suffix array* [7] for remembering the indices and updating the pair statistic on-the-fly.

Word segmentation is processed by looking up the codebook with longest common string matching. Although this forward maximum matching algorithm is straightforward, in the experiment, we found it can output good segmentations.

3.3 Translation Experiments

To measure the performance, we evaluate our method on Chinese-to-Japanese translation tasks. Translation quality is measured by the BLEU [11] and RIBES [6] metrics.

表 1: Chinese-Japanese and Japanese–Chinese translation results on ASPEC Corpus. Boldface indicates no significant difference with the best system. FMDL stands for MDL with fixed vocabulary.

	ja → zh		**zh → ja**	
	BLEU	**RIBES**	**BLEU**	**RIBES**
baseline	31.12	82.51	39.61	86.34
BPE	**32.60**	**84.33**	**41.62**	**86.70**
WPM	32.28	84.28	**41.87**	**86.84**
FMDL	**32.64**	**84.67**	**41.90**	**86.78**

Our baseline system is a NMT system armed with a bidirectional LSTM encoder (two layers RNN), a stacked-LSTM decoder (two layers RNN) and a global attention layer. It is similar to the system configuration proposed in [9] without any <*unk*> replacement. For all experiment, we have used the basic setting following the baseline system in Workshop on Asian Translation (WAT) translation champion. For word embedding, we limit both the source and target vocabulary to 50k with 500 dimensions for each word vector in our experiments. The size of hidden states is 500. We also compare our method with other segmentation methods, e.g., *byte pair encoding* (BPE) [12] and *wordpieces model* (WPM) [13]. We limit both the source and target vocabulary to 50k for all segmentation models.

4 Conclusion

We proposed a novel unsupervised MDL-based method for NMT. Differing from the previous MDL-based method, our approach limits the vocabulary to a fixed size. We also compared our method with other state-of-the-art unsupervised segmentation methods for NMT. Our MDL-based segmentation achieved the comparable results in the Chinese–Japanese and Japanese–Chinese end-to-end NMT experiments.

Remark

A similar paper has been submitted to an international conference.

参考文献

[1] Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. Efficient unsupervised recursive word segmentation using minimum description length. In *COLING*, p. 1058. Association for Computational Linguistics, 2004.

[2] Pi-Chuan Chang, Michel Galley, and Christopher D Manning. Optimizing Chinese word segmentation for machine translation performance. In *WMT*, pp. 224–232. Association for Computational Linguistics, 2008.

[3] Jesús González-Rubio and Francisco Casacuberta. Inference of phrase-based translation models via minimum description length. In *EACL, volume 2: Short Papers*, pp. 90–94, 2014.

[4] Peter Grünwald. A minimum description length approach to grammar inference. In *International Joint Conference on Artificial Intelligence*, pp. 203–216. Springer, 1995.

[5] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[6] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *EMNLP*, pp. 944–952. Association for Computational Linguistics, 2010.

[7] Juha Kärkkäinen and Peter Sanders. Simple linear work suffix array construction. In *ICALP*, Vol. 2719, pp. 943–955. Springer, 2003.

[8] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT:NAACL-Volume 1*, pp. 48–54. Association for Computational Linguistics, 2003.

[9] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[10] Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. Overview of the 3rd Workshop on Asian Translation. *Proc. WAT*, 2016.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318. Association for Computational Linguistics, 2002.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. pp. 1715–1725, August 2016.

[13] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[14] Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. An efficient algorithm for unsupervised word segmentation with branching entropy and mdl. *Information and Media Technologies*, Vol. 8, No. 2, pp. 514–527, 2013.