Decoder-based Discriminative Training of Phrase Segmentation for Statistical Machine Translation

Hyoung − Gyu Lee   *Hae − Chang Rim*
Department of Computer and Radio Communications Engineering
Korea University, Seoul, Korea
`{hglee,rim}@nlp.korea.ac.kr`

Abstract

In this paper, we propose a new method of training phrase segmentation model for phrase-based statistical machine translation(SMT). We define a good segmentation as the segmentation producing a good translation. According to this definition, we propose a method that can discriminate between a good segmentation and a bad segmentation based on the translation quality. The proposed approach constructs the phrase labeled data by using the SMT decoder, so that the phrase segmentations supporting good translations can be acquired. Furthermore, our iterative training algorithm of the segmentation model can gradually improve the performance of the SMT decoder. Experimental results show that the proposed method is effective in improving the translation quality of the phrase-based SMT system.

Title and Abstract in another language (Korean)

통계 기계번역을 위한 디코더 기반 구 분할 차별 학습 방법

본 논문은 구 기반 통계 기계번역을 위한 구 분할 모델의 새로운 학습 방법을 제안한다. 우리는 좋은 번역(good translation)을 생성하는 구 분할을 좋은 분할(good segmentation)이라고 정의한다. 우리는 이 정의에 따라 번역 품질에 기반하여 좋은 분할과 좋지 않은 분할을 차별할 수 있는 방법을 제안한다. 제안하는 접근방법은 통계 기계번역(SMT) 디코더을 이용하여 구 부착 데이터를 구축함으로써, 좋은 번역을 만드는 구 분할을 얻을 수 있다. 또한 SMT 디코더의 성능을 점진적으로 개선시킬 수 있는 반복적인 학습 알고리즘을 제안한다. 실험을 통해, 제안 방법이 구 기반 SMT 시스템의 번역 품질 향상에 효과적이었음을 보인다.

Keywords: phrase-based SMT, phrase segmentation model, decoder-based approach.

Keywords in Korean: 구 기반 통계 기계번역, 구 분할 모델, 디코더 기반 접근방법.

Proceedings of COLING 2012: Posters, pages 611–620,
COLING 2012, Mumbai, December 2012.

1 Introduction

Phrase segmentation model for phrase-based statistical machine translation (SMT) has been studied by several researchers in recent years (Blackwood et al., 2008; Xiong et al., 2010; Lee et al., 2011; Xiong et al., 2011). They have emphasized the necessity of the phrase segmentation model for the following reasons. First, it is required to properly group adjacent words in a sentence so that the system can consider collocation or inter-phrase context (Blackwood et al., 2008; Lee et al., 2011; Xiong et al., 2011). Second, it is also required to properly segment the input sentence to keep the translation fluency despite of the phrase reordering process (Blackwood et al., 2008). Furthermore, there are some observable differences between the segmentations producing high quality translations and low quality translations (Lee et al., 2011).

The existing phrase segmentation models for phrase-based SMT are trained by different methods. Blackwood et al. (2008)'s phrase-level n-gram model is trained through the maximum likelihood estimation from a large monolingual corpus. Lee et al. (2011)'s segmentation model has been designed as multiple scoring functions, whose parameters are obtained from a parallel corpus or a monolingual corpus.

On the other hand, the maximum entropy based segmentation model (Xiong et al., 2011) requires a training data labeled with a segment boundary, because it uses two discriminative probabilistic classifiers. Their approach automatically identifies each phrase boundary of a source sentence by using the shift reduce algorithm (SRA) (Zhang et al., 2008). This study defines good segmentation in terms of cohesiveness of translation and focuses on learning cohesive segments from word aligned training corpus.

In aspects of the training of the phrase segmentation model, any previous studies did not differentiate between a good segmentation and a bad segmentation based on the translation quality. This paper defines a good segmentation as the segmentation producing a good translation. This definition has the goal for improving the performance of the end-to-end SMT system, and thus good segmentations according to this definition may be inconsistent from human translators' point of view.

In this paper, we develop a new decoder-based segmenter for automatically labeling segment boundaries on the training data of phrase segmentation model. This labeler uses the base SMT decoder including the conventional translation model without a phrase segmentation model. In this approach, we assume that there exists a good translation among translation candidates produced by the base decoder.

The advantage of the decoder-based method is that it allows the segmentation model to learn more practically helpful segmentation boundaries. Phrase segmentation boundaries produced by the decoder are obviously helpful in terms of the translation quality, because they have been used in real decoding situations and have been selected by considering the reference translations. In other words, this decoder-based approach can effectively filter the bad phrase segments to train the segmentation model.

In addition to the segmentation labeling method, we design an iterative training algorithm, in which the phrase segmentation model and the decoder are iteratively trained. Through the algorithm, the performance of the phrase segmentation model and the decoder can be gradually improved.

2 System Overview

The proposed system is based on the phrase-based log-linear translation model (Och and Ney, 2004). The decision rule of the model has the following form:

$$
\begin{aligned}
\hat{e}_1^I &= \arg\max_{e_1^I}\{Pr(e_1^I|f_1^J)\} && (1)\\
&= \arg\max_{e_1^I}\{\sum_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)\} && (2)
\end{aligned}
$$

where e_1^I denotes a target sentence containing I words, f_1^J denotes a source sentence containing J words, h_m denotes a feature function, and λ_m denotes a weight of a feature function. Conventional phrase-based SMT employs components such as the language model, the phrase translation model, the phrase reordering model, the word penalty, the phrase penalty, and so on, as its feature functions.

Like the previous works for phrase segmentation model (Lee et al., 2011; Xiong et al., 2011), we integrate the phrase segmentation model into the log-linear model as an additional feature function.

The proposed system architecture is shown in Figure 1. In this system, two different sets of parallel sentences are used to train the phrase table and the phrase segmentation model, respectively. Our phrase segmenter using the SMT decoder automatically annotates source phrase boundaries on the training corpus. The phrase segmentation model learns the phrase segments from this labeled data. The SMT decoder employs this learned segmentation model. Our architecture allows a gradual improvement of both the segmentation model and the decoder through an iterative procedure. We describe the detailed training method in section 4.

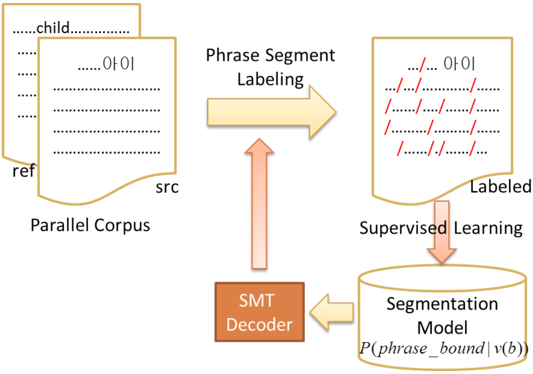

Figure 1: System architecture

3 Phrase Segmentation Model

The phrase segmentation model gives a score to the source segmentation of a given hypothesis. The proposed model outputs a probability that the source segmentation is good given both word boundaries and phrase boundaries of a given hypothesis. This model is simplified as a discriminative probabilistic classifier, which can judge whether each word boundary of a source sentence is a phrase segment boundary or not. The proposed model is described as the following equation:

$$\begin{aligned} h(x) &= P(good\_seg|PB(x), WB(x)) && (3)\\ &\propto \prod_{\forall b \in WB(x)} P(phrase\_bound|v(b))^{I_{PB}(b)} \times \{1 - P(phrase\_bound|v(b))\}^{1-I_{PB}(b)} && (4)\end{aligned}$$

where $h(x)$ denotes a feature function of a hypothesis x. $PB(x)$ and $WB(x)$ denote a set of source phrase segment boundaries and a set of source word boundaries of a hypothesis x, respectively. The label, $phrase\_bound$ indicates that a given word boundary is a phrase segment boundary, $v(b)$ denotes a function that outputs the feature vector of a word boundary, and $I_{PB}(b)$ denotes an indicator function of the existence of a word boundary b in $PB(x)$.

We simply and intuitively model the phrase segmentation, because this work is more interested in the effective training of the model than in the segmentation modeling. Now, according to this model, we have to train only one classifier, $P(phrase\_bound|v(b))$.

We adopt the maximum entropy log-linear model as a learning model. We propose lexical contexts, part-of-speech contexts, and the collocation score of two adjacent words as the feature set. We use the log likelihood ratio, which is widely used to measure the association of random variables, as the collocation measure. These features known as useful clues for phrase segmentation are used in previous works (Lee et al., 2011; Xiong et al., 2011).

In the decoding process, we use the conventional decoding algorithm of the phrase-based SMT to consider the additional segmentation model feature. The log-probability of good segmentation of current source phrase is added to the total score in every evaluation of translation options.

4 Training

In this section, we describe the proposed labeling method for acquiring the phrase segment boundaries that are likely to generate good translations. And then, we introduce a recursive procedure of training the segmentation model.

4.1 Phrase segment labeling and learning

We use the base decoder to label each word boundary with the phrase segment boundary. Most SMT decoders generate a lot of translation candidates and search the best translation according to their statistical models. There may be a relatively better translation among the translation candidates. We regard the source segmentation producing a better translation as a better segmentation. We also assume that a better translation can be found in the search space of the decoder by using reference translations and an evaluation metric.

Therefore, we try to find the best segmentation among the segmentations producing translation candidates generated by the base decoder. For this, we select the segmentation producing the best translation, whose BLEU score is the highest among the candidates, from the n-best list of translation candidates.[1] Our system set n to 200.

As described in the previous section, the maximum entropy log-linear model is used to estimate the probabilities of the boundaries. This discriminative model requires both the gold-labeled data and the parameter search algorithm. For this requirements, we use the automatically constructed data explained earlier, and the LBFGS algorithm[2].

Our decoder-based approach is similar to the tuning method of the translation model using the algorithms such as MERT (Och, 2003), MIRA (Chiang et al., 2008), or pairwise ranked optimization (Hopkins and May, 2011). Both approaches use a small set of bilingual sentences translated by the base decoder, reference translations and an evaluation metric.

4.2 Iterative training of segmentation model

Once the trained segmentation model is integrated into the base decoder, the improved decoder can be available to train the segmentation model again. In other words, our method utilizes a dependent relationship between the decoder and the phrase segmentation model. Therefore, we propose a recursive training algorithm that can iteratively train the segmentation model. In this algorithm, we assume that if a decoder is improved, the segment-labeled data obtained by the decoder will also be improved. The better segmentations, which were not included in the old n-best list of hypotheses, may be included in the new list.

Figure 2 shows the formal representation of the iterative training algorithm. It uses a decoder D including the pre-constructed phrase table and two equal-sized training sets, B_1, B_2 as inputs. Consequentially, it outputs an improved decoder. $Choose(B_1, B_2)$ alternately selects one training set between two sets. $Label(C, D)$ is a function in which the decoder D annotates segment boundary labels at the source side of the set C, using the method described in section 4.1. $DiffRatio(C_{labled}, C_{old-labled})$ returns the number of different segment boundary labels between two sets. $TrainSM(D, C)$ is a function of training the segmentation model of the decoder D by using the labeled data C. $TuneWeights(D)$ performs the weight optimization of log-linear translation model, by using algorithms such as MERT (Och, 2003), MIRA (Chiang et al., 2008) or pairwise ranked optimization (Hopkins and May, 2011).

The reason of dividing the training set into two sets, B_1 and B_2, is for preventing the decoder from being immediately applied again to the same data that is used for training the segmentation model of the decoder. This algorithm outputs a SMT decoder containing a trained segmentation model for each iteration of the training procedure. This algorithm is terminated when the ratio of the changed labels of the labeled result reaches the threshold θ, compared with the previous labeled result. We empirically determine the threshold.

5 Experimental Results

We have experimented with our method for Korean-to-English (K-E) and Chinese-to-English (C-E) translation tasks. We have used about 1.1M Korean-English parallel sentences[3] to build

[1] We use the Moses toolkit (Koehn et al., 2007) to implement the base decoder, and *-n-best-list* and *-include-alignment-in-n-best* as additional options to obtain n-best outputs and their phrase segmentation results.

[2] Our classifier and its trainer are implemented by using Zhang's MaxEnt Toolkit ($http : //homepages.inf.ed.ac.uk/lzhang10/maxent\_toolkit.html$).

[3] Part of this corpus is provided by SK Planet CO. only for research purpose. Part of this corpus is automatically constructed by using Hong et al. (2010)'s method. Part of this corpus is released by Kim et al. (2010), and the Sejong corpus (Kang and Kim, 2004) is also used.

```
$\textbf{\textit{ITA}}(D, B_1, B_2)$
    Input: a decoder $D$, a set of parallel sentences $B_1$, another set of parallel sentences $B_2$
    Output: a decoder containing a trained segmentation model
    $C = Choose(B_1, B_2)$
    $C_{labeled} = Label(C, D)$
    if $DiffRatio(C_{labeled}, C_{old-labeled}) < \theta$
        Return $D$
    else
        $D' = TrainSM(D, C_{labeled})$
        $D_{new} = TuneWeights(D')$
        $C_{old-labeled} = C_{labeled}$
        Return $\textbf{\textit{ITA}}(D_{new}, B_1, B_2)$
```

Figure 2: Iterative training algorithm (henceforth ITA)

the K-E SMT system. Among them, 1.1M, 10K, 1K and another 1K sentences were selected as the phrase table training set, the segmentation model training set, the tuning set, and our own test set, respectively. The official evaluation set of NIST OpenMT 2012 Evaluation (MT-12) has been used as another test set for K-E translation. We have also used 475K, 10K and 500 sentences from LDC Chinese-English corpora (LDC2005T10, LDC2005T06, and part of LDC2004T08) as the phrase table training set, the segmentation model training set, and the tuning set for the C-E SMT system, respectively. The official evaluation set of NIST OpenMT 2008 (MT-08) Evaluation has been used as the test set for C-E translation.

In this experiment, we use one half (5K) of the 10K segmentation training set and another one half (5K) as B_1 and B_2 for ITA.

		Korean-English		Chinese-English	
		Korean	English	Chinese	English
Train	Sentences	1,151K		485K	
	Words	27.7M	22.0M	10.5M	11.3M
Tune	Sentences	1,000		500	
	Words	27.6K	22.1K	10.7K	11.2K
Test A (Our own set)	Sentences	1,000		-	
	Words	26.5K	21.6K	-	-
Test B (MT-12/MT-08)	Sentences	3,074		1,357	
	Words	136.0K	90.7K*	32.5K	36.2K*

Table 1: Parallel corpus statistics (*Average of four references)

The SRILM toolkit[4] (Stolcke, 2002) has been used to train a 4-gram language model on 22.1M word tokens of English text. We have also used the morphological analyzer (Lee and Rim, 2009) for Korean tokenization and the Stanford Chinese word segmenter[5] (Tseng et al., 2005) for

[4] http://www.speech.sri.com/projects/srilm

[5] http://nlp.stanford.edu/software/segmenter.shtml

Chinese tokenization. We have used the open source SMT system, Moses[6] (Koehn et al., 2007) to implement the base decoder and the decoder that uses the proposed segmentation model. The minimum error rate training (MERT) (Och, 2003) was used to tune the feature weights. Both the BLEU score (Papineni et al., 2002) and the NIST score (NIST, 2001) are used as the evaluation metric.

We first verify the effectiveness of the proposed phrase segment boundary labeling in phrase-based SMT. We want to know how much the performance of the system can be improved, if the decoder is perfectly aware of the segment boundary information of input sentences. So, we labeled the test set by using the base decoder and the reference translation, and then used only the phrase segments in the labeled data when referring the phrase table during the decoding process. Through this setting, we can directly provide the acquired segmentation boundary information to the decoder. The experimental result is shown in Table 2. From these promising results, we found that if the translation system learns the segmentation boundaries labeled by using the base decoder well, the translation quality can be improved. In other words, these scores can be regarded as the upper bound of the system using the proposed decoder-based segmenter.

	K-E	C-E
System	(Our own)	(MT-08)
Baseline	16.81	17.86
Gold segmentation only	20.16	19.44

Table 2: Effectiveness of the proposed phrase segment boundary labeling (BLEU)

Next, we evaluate the segment boundary classifier by performing 10-fold cross validation on the labeled data. The accuracy was 78% when using the ME model in Korean. This result implies that the learning model and the features adopted in our model are effective enough in finding the phrase segment boundary.

Table 3 shows the performance of the proposed system. In this experiment, all scores of the proposed system were measured after the third iteration, in which the ITA reached the threshold θ, determined by experiments carried out on the development set. Our system outperformed the baseline in both K-E and C-E. From these results, we found that the proposed method can effectively train the segmentation model for the phrase-based translation, even though the system could not achieve the upper bound shown in Table 2. We also found that the performance gain in K-E task is larger than that in C-E task through both Table 2 and Table 3. It implies the relative importance of the phrase segmentation for K-E task, and encourages us to study the linguistically-motivated model of Korean phrase segmentation for Korean-to-X translation as the future work.

Figure 3 shows BLEU scores measured for each iteration up to the fifth iteration of the ITA. From both graphs, we found that the ITA increases the BLEU score until the third iteration, and the score fluctuates in spite of the increase of iterations after the third iteration. We could learn that the proposed iterative training procedure gradually improves the system performance until a certain number of iterations as expected.

[6] http://www.statmt.org/moses

Language pair	K-E				C-E	
Test set	Our own		MT-12		MT-08	
System	BLEU	NIST	BLEU	NIST	BLEU	NIST
Baseline	16.81	5.8053	10.98	5.4596	17.86	6.0822
Proposed	18.04*	6.1020*	12.83*	6.0669*	18.25*	6.2007*

Table 3: Performance of the proposed system. All scores of the proposed system are measured after the third iteration of ITA. The scores marked with * are significantly better than the baseline ($p < 0.05$)

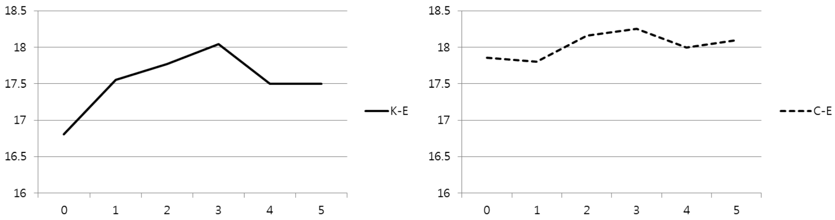

Figure 3: Iteration-BLEU graph

6 Conclusion

In this paper, we propose a new model of decoder based phrase segmentation and a new algorithm which can iteratively train the segmentation model. The main contribution of this paper can be summarized as follows. First, this paper is the first attempt to discriminate between a good segmentation and a bad segmentation based on the evaluation metric of the translation quality. Second, we have shown that the phrase segmentation model supporting the good translation quality can be trained by using the base SMT decoder. Finally, the proposed iterative training algorithm could gradually improve the translation quality of the phrase-based SMT, although the efficiency of the training may be reduced because of its iterative decoding.

For the future work, we try to integrate the decoder-based segmenter into other statistical translation models such as the hierarchical phrase-based model or the syntax-based model. This work is based on the hypothesis that our approach allows the system to select the practically useful boundaries of translation rules in the decoding process in the same way as the phrase-based model.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2011-0016878).

References

Blackwood, G., de Gispert, A., and Byrne, W. (2008). Phrasal segmentation models for statistical machine translation. In *Proceedings of Coling 2008*.

Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of EMNLP 2008*.

Hong, G., Li, C.-H., Zhou, M., and Rim, H.-C. (2010). An empirical study on web mining of parallel data. In *Proceedings of Coling 2010*, pages 474–482.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of EMNLP 2011*, pages 1352–1362.

Kang, B.-M. and Kim, H. (2004). Sejong korean corpora in the making. In *Proceedings of LREC 2004*, pages 1747–1750.

Kim, S., Jeong, M., Lee, J., and Lee, G. G. (2010). A cross-lingual annotation projection approach for relation detection. In *Proceedings of Coling 2010*, pages 564–571.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*.

Lee, D.-G. and Rim, H.-C. (2009). Probabilistic modeling of korean morphology. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):945–955.

Lee, H.-G., Lee, J.-Y., Kim, M.-J., Rim, H.-C., Shin, J.-H., and Hwang, Y.-S. (2011). Phrase segmentation model using collocation and translational entropy. In *Proceedings of MT Summit XIII*.

NIST (2001). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *http://www.nist.gov/speech/tests/mt/doc/ngramstudy.pdf*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A conditional random field word segmenter. In *Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*.

Xiong, D., Zhang, M., and Li, H. (2010). Learning translation boundaries for phrase-based decoding. In *Proceedings of HLT-NAACL 2010*.

Xiong, D., Zhang, M., and Li, H. (2011). A maximum-entropy segmentation model for statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2494–2505.

Zhang, H., Gildea, D., and Chiang, D. (2008). Extracting synchronous grammar rules from word-level alignments in linear time. In *Proceedings of Coling 2008*.