Phrase Segmentation Model using Collocation and Translational Entropy

Hyoung-Gyu Lee[1], **Joo-Young Lee**[1], **Min-Jeong Kim**[1], **Hae-Chang Rim**[1],
Joong-Hwi Shin[2], and **Young-Sook Hwang**[3]
[1]Department of Computer and Radio Communications Engineering
Korea University, Seoul, Korea
[2]Technology Innovation Center, NHN Corporation, Gyeonggi-do, Korea
[3]Platform R&D Center, SK Telecom, Seoul, Korea
{hglee,jylee,mjkim,rim}@nlp.korea.ac.kr
joonghwi.shin@gmail.com
youngsook.hwang@sk.com

Abstract

In this paper, we propose a phrase segmentation model for the phrase-based statistical machine translation. We observed that good translation candidates generated by a conventional phrase-based SMT decoder have lexical cohesion and show more uniform translation for each phrase segment. Based on the observation, we propose a novel phrase segmentation model using collocation between two adjacent words and translation entropy of phrase segments. Experimental results show that the proposed model significantly improves the translation quality in both English-to-Korean and English-to-Chinese translation tasks.

1 Introduction

Phrase segmentation is to split a sentence into a sequence of multiple phrases. The phrase-based statistical machine translation (PBSMT) includes the phrase segmentation process (Koehn et al., 2003; Och and Ney, 2004; Zens and Ney, 2004). An input sentence is segmented into phrases at first, (here, a phrase is not necessarily linguistically motivated) and then translated by way of phrase-to-phrase.

Conventional PBSMT assumes that all of the possible phrase segmentation results are uniformly distributed.[1] One sentence can have multiple ways of segmentation, because there is no assumption or constraint for a phrase. For example, a sentence including three words $w_1w_2w_3$ possibly has four kinds of segmentation: $w_1/w_2/w_3$, w_1/w_2w_3, w_1w_2/w_3, and $w_1w_2w_3$. However, they do not consider any characteristics of the cohesive power between words consisting of source phrase segments.

Different phrase segmentation results usually generate different translation results. Therefore, it is obvious that there is a phrase segmentation which is appropriate for generating a correct translation result. It means that we should differentiate the probabilities of phrase segmentation candidates.

There are two previous works closely related to the phrase segmentation model for PBSMT. The first is the phrase segmentation model of Blackwood et al. (2008). They proposed a simple phrase bi-gram model which can be integrated into the translation model. They verified that estimating a phrase segmentation probability using a very large monolingual corpus is helpful for improving the translation quality.

The second is the study for the discriminative translation boundary classifier for PBSMT (Xiong et al., 2010). Their model classifies each word of an input sentence as beginning or ending. The predicted phrase boundary produced promising results in phrase-based translation. They did not, however, devise a probabilistic model which can be integrated into the translation model.

There also have been some strategies which score each source phrase in a phrase table from various viewpoints. While the first type is utilizing statistical collocation information (Ren et al., 2009; Liu et al., 2010), the second is based on the multiword expressions translation (Lambert and Banchs, 2005; Ren et al., 2009; Carpuat and Diab, 2010). These studies

[1] All imaginable phrase segmentation results do not actually have same probabilities because one segment is constrained to use only a phrase contained in a phrase table.

usually append additional features to a phrase table. It can be said that their methods indirectly give differential probabilities to possible segmentation results. However, they did not try to explicitly model the generation process of a phrase segmentation result.

Most previous works did not describe the condition of good segmentation, i.e. segmentation to generate a high quality translation result. The good segmentation has some characteristics which are different from the bad segmentation in terms of translation.

In this paper, we address the following questions.

- What characteristics of a source word sequence, i.e. a candidate segment help the model find the appropriate phrase segmentation in terms of the translation quality?
- How can we design a probabilistic segmentation model considering such characteristics?

We propose a new phrase segmentation model which considers the lexical cohesion between adjacent words and the translational diversity of a word sequence as the characteristics of good segmentation. In order to reflect such characteristics in the model, we use statistical collocation and translational entropy. Our approach can reflect multiple characteristics and can be integrated into the phrase-based translation model.

The rest of the paper is organized as follows. Section 2 analyzes the phrase segmentation result of the conventional translation model and catches characteristics of good segmentation results. Section 3 proposes the phrase segmentation model for PBSMT. Section 4 presents the experimental setup and results on two translation tasks. Finally, section 5 concludes the paper.

2 Analyzing Phrase Segmentation of Conventional PBSMT

2.1 Phrase-based Translation Model

We describe the conventional phrase-based translation model before analyzing the phrase segmentation results produced by the model.

The traditional phrase-based translation model introduces the phrase segmentation model (Zens and Ney, 2004; Blackwood et al., 2008) assuming a one-to-one phrase alignment.

$$P(\mathbf{f}|\mathbf{e}) \approx \max_{S} \left\{ P(S|\mathbf{e}) P(\bar{f}_1^I|\bar{e}_1^I) \right\} \qquad (1)$$

where $\mathbf{f}$ and $\mathbf{e}$ indicate the source and target sentences respectively, $\bar{f}_1^I$ and $\bar{e}_1^I$ mean the segment sequences of source and target sentences respectively, and S denotes a source phrase segmentation.

The translation model is decomposed into the phrase segmentation probability and the phrase-level translation probability. The segmentation probability $P(S|\mathbf{e})$ is regarded as a constant as shown in the following equation.

$$P(S|\mathbf{e}) = Constant \qquad (2)$$

The phrase-level translation probability is modeled by using the independence assumption as follows.

$$P(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(start_i - end_{i-1} - 1) \qquad (3)$$

where ϕ means phrase translation probability and d means distance-based phrase distortion probability.

2.2 Analyzing Phrase Segmentation Results

In this section, we analyze n-best translation candidates produced by the PBSMT decoder with the uniform segmentation model.

We have performed the English-to-Korean translation task using the PBSMT engine, which was built using Moses (Koehn et al., 2007) toolkit and the parallel corpus described in section 4 for training, tuning and evaluation.[2]

Table 1 and 2 show the comparison of the segmentation statistics between high quality translation and low quality translation among 200-best candidate translation results of the input sentence by the conventional translation model. We classified each candidate translation as 'high quality' class or 'low quality' class. If the BLEU score of a translation candidate is higher than that of the 1-best translation result, it is classified as 'high quality', otherwise

[2] We have used *-n-best-list* and *-include-alignment-in-n-best* as additional options to obtain n-best outputs and their phrase segmentation result.

		68 High Quality Translation		131 Low Quality Translation	
i	Input Sentence	Not split	Split	Not split	Split
1	*most*	0%	100%	0%	100%
2	*bankers*	0%	100%	0%	100%
3	*required*	0%	100%	0%	100%
4	*that*	0%	100%	0%	100%
5	*customers*	0%	100%	0%	100%
6	*pay*	0%	100%	0%	100%
7	*5*	100%	0%	100%	0%
8	*percent*	**67.3%**	**32.7%**	**27.4%**	**72.6%**
9	*of*	**32.7%**	**67.3%**	**72.6%**	**27.4%**
10	*their*	**32.7%**	**67.3%**	**72.6%**	**27.4%**
11	*credit*	**67.3%**	**32.7%**	**27.4%**	**72.6%**
12	*card*	0%	100%	0%	100%
13	*balance*	0%	100%	0%	100%
14	*every*	5.8%	94.2%	12.6%	87.4%
15	*month*	94.2%	5.8%	87.4%	12.6%
16	.	0%	100%	0%	100%

Table 1: Statistics of split between adjacent words in source segmentation for translation candidates by PBSMT (Input: *Most bankers required that customers pay 5 percent of their credit card balance every month.*)

		99 High Quality Translation		100 Low Quality Translation	
i	Input Sentence	Not split	Split	Not split	Split
1	*need*	9.1%	90.9%	0%	100%
2	*a*	87.9%	12.1%	99%	1%
3	*place*	12.1%	87.9%	0%	100%
4	*to*	100%	0%	90%	10%
5	*live*	0%	100%	0%	100%
6	*while*	0%	100%	0%	100%
7	*taking*	**71.7%**	**28.3%**	**0%**	**100%**
8	*courses*	92.9%	7.1%	100%	0%
9	*at*	7.1%	92.9%	0%	100%
10	*our*	99%	1%	99%	1%
11	*english*	96%	4%	97%	3%
12	*school*	4%	96%	0%	100%
13	*?*	0%	100%	0%	100%

Table 2: Statistics of split between adjacent words (ith and i+1th) in source segmentation for translation candidates by PBSMT (Input: *Need a place to live while taking courses at our English school?*)

'low quality.' Each row of table 1 and 2 shows both the ratio that two adjacent words in the source sentence are segmented in a phrase segmentation and the ratio that they are not segmented.

The rows shown in bold have very different ratio between high and low quality. It means that these words are important segmentation points for translation. For example, when *percent* and *of* are not split, the system usually generates high quality translation. On the other hand, it is better to split *of* and *their* for good translation. Therefore, if we know that the segmentation *percent of* / *their* is better than *percent* / *of their* then we can improve the translation quality by choosing the better one during translation process. As a consequence of this observation, we need a good method of distinguishing between appropriate segmentation and inappropriate segmentation.

Through the additional analysis, we have found out the following two tendencies. First, when some collocations such as *credit card* are allocated to the same segment, the translation quality tends to be high. Second, it also does when a phrase whose translational diversity is low is not segmented. In other words, though individual words in a phrase may be diversely translated, the phrase may be usually translated as a few expressions. For example, the candidate segment *taking courses* in table 2, might be always translated as the Korean expressions, 수강하다(su-gang-ha-da) or 연수를 받다(yeon-su-reul bat-da), of which meanings are similar. Such word sequences are usually treated as idiomatic expressions, named entities or some expressions which are likely to be consistently translated by human translators.

3 Phrase Segmentation Model for PBSMT

We assume that the probabilities of all possible phrase segmentation results of a source sentence are not uniformly distributed. Here, we propose a translation model which distinguishes between a phrase segmentation candidate that is likely to generate high quality translation and other segmentation candidates. The probabilistic segmentation model outputs a probability of the generation of I segments as follows.

$$P(S|\mathbf{e}) = P(\bar{e}_1 \cdots \bar{e}_I | e_1 \cdots e_n), 1 \leq I \leq n \quad (4)$$

where e is a word, $\bar{e}$ is a segment, n is the number of words, and I is the number of segments. This segmentation model[3] replaces the constant of equation (2). This modeling is similar to the approach of Blackwood et al. (2008).

3.1 Phrase Segmentation Model using Collocation and Translational Entropy

In this section, we propose three different kinds of phrase segmentation models. The models basically output the segmentation probability of a given sentence by normalizing the product of the score of each candidate segment as follows:

$$\begin{aligned} & P(\bar{e}_1 \cdots \bar{e}_I | e_1 \cdots e_n) \\ & = \frac{1}{Z(e_1 \cdots e_n)} \left(\prod_{i=1}^{I} Score(\bar{e}_i) \right)^{1/I} \end{aligned} \quad (5)$$

where Z is a normalization factor and $Score(\bar{e})$ denotes the scoring function of a candidate segment.

Model using Collocation Measure (CO)

As the first model, we present a segmentation model with the collocation concept. In this model, when mutually cohesive words belong to one segment, we give a high score to the segment. Segmentation candidates of a source sentence are differentiated by the score of each segment. The score is obtained by calculating a collocation score for each pair of adjacent words contained in the segment. Here, if only one word is contained in the candidate segment, the function gives a constant K to the segment according to the equation 6, and K is optimized by experiments on the development set.

$$\begin{aligned} & Score(\bar{e}_i) \\ & = \begin{cases} K & if\,|\bar{e}_i| = 1 \\ \left(\prod_{\forall e_j \in \bar{e}_i} Col(e_j, e_{j+1}) \right)^{1/(|\bar{e}_i|-1)} & otherwise \end{cases} \end{aligned} \quad (6)$$

where $|\bar{e}|$ is the number of words contained in a segment and the function $Col()$ means the collocation score of two words. The measurement of the score is presented in section 3.2.

This model may produce similar effects to the approach using the collocation probability for improving the phrase table, which is proposed in (Liu et al., 2010).

Model using Translational Entropy 1 (TE)

Based on the analysis of section 2, this model assumes that the segmentation in which a word sequence with low translational diversity is not split will generate a high quality translation. In order to reflect this idea, we use the translational entropy (Melamed, 1997) of each candidate segment. The following function gives a high score to the segment whose translations are not diverse.

$$Score(\bar{e}_i) = \begin{cases} K & if\,|\bar{e}_i| = 1 \\ \frac{1}{TE(\bar{e}_i)} & otherwise \end{cases} \quad (7)$$

where $TE(\bar{e})$ means the translational entropy of a segment. It has a high value when possible translation candidates of $\bar{e}$ are diverse and the translational entropy is computed as follows.

$$\begin{aligned} TE(\bar{e}) & = H(T_{\bar{e}}|\bar{e}) \\ & = -\sum_{\forall t \in T_{\bar{e}}} P(t|\bar{e}) log P(t|\bar{e}) \end{aligned} \quad (8)$$

where $T_{\bar{e}}$ is a set of all possible translation candidates of a segment $\bar{e}$.

Model using Translational Entropy 2 (GT)

The third model GT is a variation of the second model. The GT model measures the Gap between the word-level and the phrase-level Translational entropy.

The TE model focuses on the translational diversity of a phrase segment. To enhance the TE model, we develop the third model which uses the translational diversity of a phrase and that of a word together. Suppose that there are two English words w_1 and w_2. While each of them can be translated into various Korean words, the phrase "w_1 w_2" is translated into only one Korean phrase. In such a case, it is natural to group w_1 and w_2 into one phrase. The model reflects this intuition and therefore scores a segment candidate by the following equation.

[3] We note that the direction of source and target was reversed by the assumption of the noisy channel. The segmentation model is actually applied to the source side rather than to the target side.

$$
\begin{aligned}
&Score(\bar{e}_i)\\
&= \begin{cases} K & if\,|\bar{e}_i| = 1 \\ \frac{1}{|\bar{e}_i|} \sum_{\forall e \in \bar{e}_i} TE(e) - TE(\bar{e}_i) & otherwise \end{cases}
\end{aligned} \tag{9}
$$

This model is similar to a scoring function for identifying idiomatic expressions proposed in (Lee et al., 2010). According to the work, a lot of idiomatic expressions are highly ranked by the scoring function. It is obvious that when these expressions are not segmented, the system can generate better translation results.

Both TE and GT models differ from CO model. While CO model considers monolingual characteristics of candidate segments, both TE and GT models consider bilingual ones. We expect that TE and GT models will be more specialized models for machine translation.

3.2 Parameter Estimation

The parameter of CO model is estimated from monolingual corpus. In order to calculate the statistical collocation score between adjacent words, we adopt the log likelihood ratio which has been widely used to measure the association of two random variables. The measure is like the following equations.

$$
\begin{aligned}
&Col(e_j, e_{j+1})\\
&= log\frac{L(c_{12},c_1,p)L(c_2-c_{12},N-c_1,p)}{L(c_{12},c_1,p_1)L(c_2-c_{12},N-c_1,p_2)}
\end{aligned} \tag{10}
$$

$$
L(k,n,x) = x^k(1-x)^{n-k}, p = \frac{C(e_{j+1})}{N} \tag{11}
$$

$$
p_1 = \frac{C(e_j, e_{j+1})}{C(e_j)}, p_2 = \frac{C(e_{j+1}) - C(e_j, e_{j+1})}{N - C(e_j)} \tag{12}
$$

where $L()$ means a likelihood, $C()$ means the number of occurrences of a word or a word sequence in the corpus, and N is the number of word tokens in the corpus.

		English	Korean
Train	Sentences	488K	
	Words	10.8M	13.4M
Dev	Sentences	1,000	
	Words	22.1K	27.6K
Test	Sentences	1,000	
	Words	21.6K	26.5K*

Table 3: English-Korean parallel corpus statistics. The asterisked number means an average of three references.

		English	Chinese
Train	Sentences	485K	
	Words	11.3M	16.4M
Dev	Sentences	500	
	Words	11.2K	16.9K
Test (MT-08 EtoC)	Sentences	1,859	
	Words	45.6K	76.5K*

Table 4: English-Chinese parallel corpus statistics. Chinese words are counted by the character segment. The asterisked number means an average of four references.

We have to calculate the translational entropy of a candidate segment for estimating parameters of both the TE model and the GT model. They need to estimate $P(t|\bar{e})$ in equation (8). It is obtained from phrase-aligned parallel corpus by relative frequency according to the following equation:[4]

$$
P(t|\bar{e}) = \frac{count(t, \bar{e})}{\sum_{t'} count(t', \bar{e})} \tag{13}
$$

where $count()$ means the frequency of phrase pairs in the corpus.

3.3 Decoding

We have explicitly integrated the phrase segmentation model into the phrase-based translation model. The segmentation model is regarded as one of feature functions in the log-linear model. We can use the conventional decoding algorithm of PBSMT to translate an input sentence using the translation model including the new segmentation model.

The decoder using our segmentation model is implemented to evaluate translation hypotheses with the total score in which the segmentation model score is added. In this decoding process, only the phrases in phrase table are considered as candidate segments. Therefore, the issue of the computational complexity is not critical because all possible segmentation candidates of an input sentence are not considered.

[4] $P(t|\bar{e})$ can be easily obtained from the phrase table which is constructed by training the translation model.

	E-to-K		E-to-C	
Model	BLEU	NIST	BLEU	NIST
Baseline	25.39	7.146	24.61	7.291
+CO	**25.85**	**7.206**	**26.26**	**7.605**
+TE	**25.94**	**7.371**	**25.98**	**7.343**
+GT	**25.92**	**7.343**	**25.75**	**7.388**

Table 5: Performance of proposed models

4 Experiments

4.1 Experimental Setup

We have experimented with our model in English-to-Korean and English-to-Chinese translation tasks.

Table 3 shows the statistics of English-Korean parallel corpus crawled from online newswires.[5] We have also used 485K English-Chinese sentence pairs and 500 sentence pairs from LDC corpora (LDC2005T10, LDC2005T06, and part of LDC2004T08) as training and development set, respectively. The official evaluation set of NIST OpenMT 2008 Evaluation (MT08) has been used as a test set for English-to-Chinese translation task. Their statistics are reported in table 4.

Both the BLEU score (Papineni et al., 2002) and the NIST score (NIST, 2001) are used as evaluation metrics of the translation quality. Performance on English-to-Korean task is measured with word-segmented translation sentences, while that on English-to-Chinese task is measured with character-segmented translation sentences.

We have used the open source SMT engine, Moses (Koehn et al., 2007) with default options as the baseline model which uses the uniform segmentation model. Our phrase segmentation models are trained by calculating scores for each source phrase of the phrase table in the model training step. We integrate our models to the baseline and evaluate their effects. The minimum error rate training (MERT) is used for the weight tuning of both the baseline and our proposed systems.

4.2 Experimental Results

Table 5 compares the baseline model with the proposed models. The baseline model includes no phrase segmentation model. All of the proposed models have consistently increased both the BLEU score and the NIST score on two translation tasks. The bold font in the table represents the results that pass a significance test.[6] These results demonstrate that the proposed phrase segmentation models are effective in improving the translation quality.

	E-to-K		E-to-C	
Model	BLEU	NIST	BLEU	NIST
Baseline	25.39	7.146	24.61	7.291
+CO+TE	**25.90**	**7.422**	**25.04**	**7.377**
+CO+GT	**25.93**	**7.356**	**26.53**	**7.571**
+TE+GT	25.63	7.199	**25.99**	**7.369**
+CO+TE+GT	**25.91**	**7.395**	**25.51**	**7.446**

Table 6: Performance of various combinations of proposed models

An interesting result is that the performance improvement of CO model is largest in English-to-Chinese, while smallest in English-to-Korean. From this result, we can find that the helpfulness of each segmentation model depends on the language pair.

We have also obtained the results of the model combinations as shown in table 6. We expect that their combination might produce a complementary effect, because three proposed models reflect different characteristics of segmentation. We have implemented the combinations of our phrase segmentation models by adding multiple features into the log-linear model. Unlike our expectation, their BLEU scores are little different from those of the single models. However, one positive aspect is that CO+TE and CO+GT models give higher NIST scores than each single model of the combined model. It also can be said that the low performance of TE+GT is because of a high correlation between TE and GT model's feature values.

4.3 Discussion

In order to analyze the coverage and the practical effect of the proposed model, we first evaluate only sentences whose source phrase segmentation is changed by the proposed model in the test set. The results are shown in table 7. Here, "#Sent"

[5] This corpus is provided by SK Telecom only for research purpose. The parallel sentences are crawled over various online newswires.

[6] We have used Zhang's significance tester (Zhang and Vogel, 2004).

input reference		∼ the number of abductees amounts to 82,959 . ∼ 납북자 의 수가 82,959 명 에 이르ㄴ다 . (∼ DPRK abductee of number 82,959 amounts to .)
baseline	segmentation translation	∼ the number of / abductees / **amounts / to** / 82,959 / . / ∼ 피랍 자 들 / **은** / 82,959 / **하** / 았 다 . / (∼ abductees / / 82,959 / do / was . /)
CO+TE+GT	segmentation translation	∼ the number of / abductees / **amounts to** / 82,959 / . / ∼ 피랍 자 들 / 82,959 / **에 이르** / ㄴ다 . / (∼ abductees / 82,959 / amount to / is . /)
input reference		∼ workers with shabby clothes and dirty faces . ∼ 허름하ㄴ 옷에 더럽ㄴ 얼굴 을 가지ㄴ 노동자 들 (∼ shabby cloth dirty face have workers)
baseline	segmentation translation	∼ workers with / shabby / **clothes / and dirty / faces .** / ∼ 전환 / 하 / **고 더럽ㄴ / 옷 을 입 / 고 있 다 .** / (∼ change / do / dirty / wear a cloth / be ∼ing . /)
CO+TE+GT	segmentation translation	∼ workers / with / shabby / **clothes and / dirty faces** / . / ∼ 직원 들 의 / **옷 ,** / **더럽ㄴ 얼굴** / 로 / 하 / 았 다 . / (∼ workers' / cloth , / dirty face / with / do / was . /)

Table 9: Examples of English-to-Korean translation using the baseline and CO+TE+GT model

Task	Model	#Sent	Baseline	Proposed
E-to-K	+CO	726	24.88	25.46
	+TE	786	24.84	25.48
	+GT	751	24.82	25.48
E-to-C	+CO	1,786	24.56	26.23
	+TE	1,733	24.42	25.83
	+GT	1,698	24.47	25.65

Table 7: Performance in segmentation-changed sentences (BLEU)

Model	Top-10	Top-50	Top-100
Baseline	41.9%	54.2%	60.9%
+CO	50.0%	62.1%	68.6%
+TE	47.8%	61.3%	66.1%
+GT	44.9%	57.2%	63.3%

Table 8: Agreement rate between the segmentation result of 1-best translation and the most frequent segmentation result of high quality (BLEU top-N) translation candidates in test set

means the number of sentences whose segmentation is changed. The column shows that our method affects more than 70% of input sentences in English-to-Korean and more than 90% of input sentences in English-to-Chinese. More important result is that the performance gain of these sentences is larger than that of all sentences (see table 5). Thus, we believe that the proposed method has enough coverage and positive effect on translation.

Now we want to examine whether our model actually generates appropriate segmentation in terms of the translation quality. Unfortunately, there exists no gold standard phrase segmentation of an input sentence. Furthermore, the view of good segmentation for MT system may be different from that for human. Therefore, we compare the agreement rate between the segmentation result of 1-best translation and the most frequent segmentation result of high quality translation candidates for each input sentence. In this analysis, we assume that the typical segmentation result of n translation candidates with high BLEU score approximates good segmentation in terms of the translation quality. We use 10, 50, and 100 as n.

Table 8 shows the segmentation agreements in English-to-Korean translation results. The segmentation of 1-best translation result generated by the proposed model agrees with the typical segmentation result at a higher rate than that by the baseline model. This statistics imply that the proposed model generates better phrase segmentation as expected.

Two examples described in table 9 show the effect of our method more clearly. In the first example, *amounts* and *to* are grouped together into a phrase, and thus the system successfully generates a target phrase 에 이르(e i-reu). Our model also helps the system adequately translate *clothes and dirty faces* as 옷 , 더럽ㄴ 얼굴(ot, deo-reo-un eol-gul) in the second example, while the baseline model generates incorrect answer.

5 Conclusions

In this paper, we propose a phrase segmentation model for the phrase-based statistical machine translation. We demonstrate that the characteristics of good segmentation can be found based on the analysis of translation candidates generated by the conventional PBSMT decoder. We observed that good translation candidates produced by the SMT decoder have lexical cohesion and show more uniform translation for each phrase segment. Based on the observation, we propose a novel phrase segmentation model using collocation between two adjacent words and translation entropy of phrase segments. Experimental results show that the proposed model significantly improves the translation quality in both English-to-Korean and English-to-Chinese translation tasks.

For the future work, we plan to incorporate the phrase segmentation model using collocation and translational entropy into other statistical translation models such as hierarchical model or syntax-based model. It is also required to study translational diversity in the phrase pair extraction step. We expect to obtain a more improved phrase table which can help source phrase segmentation through the work.

Acknowledgments

This work was supported by the Second Brain Korea 21 Project, the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2011-0016878). This research was also supported by the MKE(The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-C1810-1002-0025).

References

Graeme Blackwood, Adria de Gispert, and William Byrne. 2008. Phrasal segmentation models for statistical machine translation. In *Proceedings of COLING 2008*.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proceedings of HLT-NAACL 2010*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*.

Patrik Lambert and Rafael Banchs. 2005. Data inferred multiword expressions for statistical machine translation. In *Proceedings of MT Summit X*.

Hyoung-Gyu Lee, Min-Jeong Kim, Gumwon Hong, Sang-Bum Kim, Young-Sook Hwang, and Hae-Chang Rim. 2010. Identifying idiomatic expressions using phrase alignments in bilingual parallel corpus. In *Proceedings of PRICAI 2010*.

Zhanyi Liu, Haifeng Wang, Hua Wu, and Sheng Li. 2010. Improving statical machine translation with monolingual collocation. In *Proceedings of ACL 2010*.

I. Dan Melamed. 1997. Measuring semantic entropy. In *Proceedings of SIGLEX-97 Workshop*, pages 41–46.

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*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of MWE 2009 (ACL-IJCNLP)*.

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Learning translation boundaries for phrase-based decoding. In *Proceedings of HLT-NAACL 2010*.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of HLT-NAACL 2004*.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of TMI 2004*.