LETTER

Estimating Translation Probabilities Considering Semantic Recoverability of Phrase Retranslation

Hyoung-Gyu LEE[†], **Min-Jeong KIM**[†], **YingXiu QUAN**[†], ***Nonmembers***, **Hae-Chang RIM**[†*a)], ***and*** **So-Young PARK**[††], ***Members***

SUMMARY The general method for estimating phrase translation probabilities consists of sequential processes: word alignment, phrase pair extraction, and phrase translation probability calculation. However, during this sequential process, errors may propagate from the word alignment step through the translation probability calculation step. In this paper, we propose a new method for estimating phrase translation probabilities that reduce the effects of error propagation. By considering the semantic recoverability of phrase retranslation, our method identifies incorrect phrase pairs that have propagated from alignment errors. Furthermore, we define *retranslation similarity* which represents the semantic recoverability of phrase retranslation, and use this when computing translation probabilities. Experimental results show that the proposed phrase translation estimation method effectively prevents a PBSMT system from selecting incorrect phrase pairs, and consistently improves the translation quality in various language pairs.

key words: *statistical machine translation, phrase translation probability, semantic recoverability, phrase retranslation*

1. Introduction

Phrase-based statistical machine translation (PBSMT) is a translation model that has been extensively studied. One of the main problems with PBSMT, is the estimation of translation probabilities, because the system selects adequate target phrases based on the probabilities.

Phrase translation probabilities are generally estimated by the following sequential process: 1) Source words are first aligned with target words for each sentence pair, 2) Phrase pairs are extracted from each sentence pair, and 3) Translation probabilities are calculated by counting relative frequencies of phrase pairs [1], [2] or computing lexical weights [2], [3]. The general method is an interpolation of these two calculation methods. This approach is known to be relatively efficient and effective.

In order to improve phrase translation probability calculation, many researchers have tried to improve word alignment methods [4]–[6]. Nevertheless, there are still errors that exist in word alignment results, because of the difficulty in acquiring sufficient parallel sentences and bridging the structural gap for the language pairs. These errors may be propagated to the next step in the sequential process, and cause imprecise calculation of translation probabilities, because the calculation method does not discriminate phrase pairs. For example, if one correct pair and one incorrect pair, 'per month'-'매달(mae-dal)' and 'per month'-'연간(yeon gan)' were extracted from a parallel corpus, the translation probabilities by the general method should be p(매달|*per month*) = 0.5, p(연간|*per month*) = 0.5. From this example, we can find that the effect of error propagation is quite strong.

Therefore, we need a method to reduce the effect of error propagation from the word alignment step to achieve better translation quality. However, there have only been a few attempts to reduce the effect of alignment errors for achieving good probability estimation.

In this paper, we focus on reducing the effect of alignment errors when calculating phrase translation probabilities. We propose an improved method for estimating phrase translation probabilities. In order to discriminate phrase pairs, we consider the semantic recoverability of phrase retranslation and define the *retranslation similarity* of a phrase pair. We measure the *retranslation similarity* as the semantic similarity between one side of the phrase pair and every phrase paired with its other side. This approach effectively assigns some penalty to incorrect phrase pairs.

2. Related Work

The sequential process for the estimation of phrase translation probabilities from the training corpus consists of three steps: word alignment, phrase pair extraction, and translation probability calculation [1], [2]. For the third step, two typical approaches are used. The first approach is to calculate the translation probability as a relative frequency of target phrases, paired with a source phrase [1], [2].

$$\Phi(\bar{e}|\bar{f}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{e}'} count(\bar{f}, \bar{e}')} \quad (1)$$

where $\bar{f}$ is a source phrase, $\bar{e}$ is a target phrase, and $count(\bar{f}, \bar{e})$ represents the number of sentence pairs for which a particular phrase pair is extracted.

The second approach is to estimate translation probabilities by using word-to-word translation probabilities of word pairs occurred in phrase pairs [2], [3]. The word-level probabilities are obtained in the word alignment step. This approach was originally proposed in order to solve the

Manuscript received August 16, 2011.
Manuscript revised November 23, 2011.

[†]The authors are with the Dept. of Computer and Radio Communications Engineering, Korea University, Seoul, Korea.
[††]The author is with Division of Digital Media Technology, SangMyung University, Seoul, Korea.
[*]Corresponding author
a) E-mail: rim@nlp.korea.ac.kr

DOI: 10.1587/transinf.E95.D.897

sparse data problem from the first approach. The equation is as follows:

$$p_w(\bar{e}|\bar{f},\bar{a}) = \prod_{i=1}^{len(\bar{e})} \frac{1}{|\{j|(i,j)\in\bar{a}\}|} \sum_{\forall(i,j)\in\bar{a}} w(e_i|f_j) \quad (2)$$

where $\bar{a}$ is the set of word alignment links of the phrase pair $\bar{f}$ and $\bar{e}$, $len(\bar{e})$ is the length of $\bar{e}$, and $w(e_i|f_j)$ indicates the word translation probability. The maximum value is chosen as a final score among scores for their all possible alignments as follows:

$$p_w(\bar{e}|\bar{f}) = max_{\bar{a}}\{p_w(\bar{e}|\bar{f},\bar{a})\} \quad (3)$$

The first approach is the phrase-level estimation and the second approach is the word-level estimation for phrase translation probabilities. A linear interpolation of these two estimation methods [2] is regarded as the state-of-the-art estimation method.

However, these approaches do not consider the errors propagated from the word alignment step. In order to reduce the effects of error propagation, we propose a method that can identify incorrect phrase pairs produced from alignment errors. Through the proposed method, we can obtain a better phrase table for machine translation.

One of the techniques used for mitigating the impact of mis-paired phrases is to use both bidirectional translation probabilities, $p(\bar{e}|\bar{f})$ and $p(\bar{f}|\bar{e})$ as features of the log-linear translation model [7]. Similarly, our approach considers phrase retranslation to reduce the impact of mis-paired phrases. Unlike the earlier work, our approach examines all possible retranslated phrases in the training set, and further considers the sense of each word occurred in a phrase.

A similar idea to the proposed approach is also found in the works using "back-translation" [8], [9]. They verified that exploiting the back-translation, which is obtained by decoding backward, is helpful to check correctness of the translation. Unlike their approaches, our approach does not require extra decoding, and applies the back-translation idea to the training of translation model.

3. Estimating Phrase Translation Probabilities

Let us assume that we translate a source sentence A into a target sentence B without any translation errors, and we retranslate B into the source sentence A'. Then, the retranslated sentence A' may be the original sentence A or a paraphrased sentence of A. If A is translated into a target sentence C with some translation errors, and we retranslate C into the source sentence, we may obtain few sentences that have the same meaning with the original sentence A.

We observed that most phrases paired with an incorrectly translated phrase do not obtain its original meaning from the retranslation. Figure 1 shows an example. For the English phrase 'per month', the upper part represents the correct translation, '매달(mae-dal)'. The bottom is an incorrect phrase translation, '연간(yeon gan)'. While the

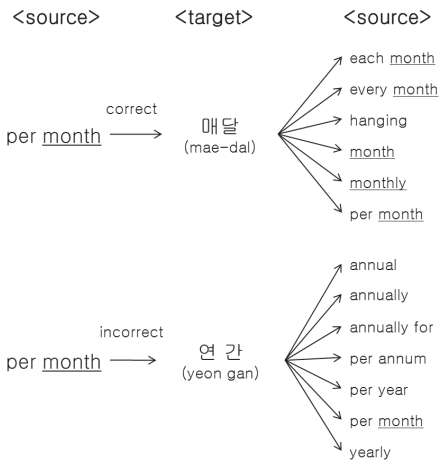

Fig. 1 Retranslation of correct and incorrect phrase pairs.

retranslation of '매달(mae-dal)' into English generates semantically similar expressions to 'per month', the retranslation of '연간(yeon gan)' generates a semantically different expression.

Based on this observation, we assume that if one side of a phrase pair is not semantically similar to the phrases that are generated by retranslation, the pair is incorrect. In other words, the assumption we made here is that the semantic recoverability of retranslation of a phrase pair reflects confidence of the pair.

3.1 Estimating Phrase Translation Probabilities Using Retranslation Similarity

We propose a method to estimate target-to-source (or source-to target) phrase translation probabilities with the assumption described above. The basic form for our estimation method is a multiplication of the relative frequency and the *retranslation similarity*. We define the *retranslation similarity* of a phrase pair as how well retranslated phrases of its one side are able to recover the meaning of its other side. We measure *retranslation similarity* as the semantic similarity between the target phrase e and every phrase paired with the source phrase f.

$$p(\bar{f}|\bar{e}) = \frac{count(\bar{e},\bar{f})}{\sum_{\bar{f}'} count(\bar{e},\bar{f}')} \times RS(\bar{f}|\bar{e}) \times \frac{1}{Z(\bar{e})} \quad (4)$$

where $count(e,f)$ indicates a count of the sentence pairs in which a particular phrase pair extracted, $RS(\bar{f}|\bar{e})$ represents a retranslation similarity of the phrase pair, and Z is a normalize factor. The final probability is obtained by normalizing the value calculated from the relative frequency and the retranslation similarity.

The retranslation similarity is calculated by the following equation:

$$RS(\bar{f}|\bar{e}) \equiv sim(\bar{e},E_{\bar{f}}) = \sum_{\bar{e}'\in E_{\bar{f}}} sim(\bar{e},\bar{e}')p(\bar{e}'|\bar{f}) \quad (5)$$

where $E_{\bar{f}}$ denotes a set of phrases linked with $\bar{f}$; $sim(\bar{e},\bar{e}')$

indicates the semantic similarity between two phrases. When the similarities are summed up, a target-to-source translation probability is used as the weight of each phrase $\bar{e}'$.

A residual problem is to measure the similarity between two phrases, $sim(\bar{e}, \bar{e}')$.

3.2 Similarity Measure

We employed several simple methods to measure the similarity between two word sequences, because our aim was to verify the effect of our approach considering semantic recoverability of phrase retranslation. More elaborate similarity measures may be able to produce better translation probabilities for PBSMT.

The easiest way to measure the similarity is to use only the surface information of each phrase as in the following equation called the Dice's coefficient:

$$sim_{LEX}(\bar{p}_1, \bar{p}_2) = \frac{2\,|\{w|w \in \bar{p}_1 \text{ and } w \in \bar{p}_2\}|}{len(\bar{p}_1) + len(\bar{p}_2)} \tag{6}$$

where w denotes a word and $len(\bar{p})$ indicates the length of phrase $\bar{p}$.

We propose a method in which a similarity is measured as a ratio of the synonym matches between two phrases. We ignore stop words such as 'the', and 'is', and punctuations. The similarity formula is as follows:

$$sim_{SYN}(\bar{p}_1, \bar{p}_2) = \frac{\sum_{\forall a \in \bar{p}_1} is\_syn(a, \bar{p}_2) + \sum_{\forall b \in \bar{p}_2} is\_syn(b, \bar{p}_1)}{len(\bar{p}_1) + len(\bar{p}_2)} \tag{7}$$

$$is\_syn(a, \bar{p}) = \begin{cases} 1, & if\ \exists b \in \bar{p}\ is\ same\ as\ a\ or\ synonym\ of\ a \\ 0, & otherwise \end{cases} \tag{8}$$

where a and b are words occurring in phrase $\bar{p}_1$ and $\bar{p}_2$, respectively. For example, given two phrases, 'per month' and 'monthly', the words 'month' and 'monthly' are a synonym relation. Thus, the similarity is calculated as $(1+1)/(2+1) = 0.667$. To identify whether a word is a synonym of another word, we use the WordNet [10] synset information for English, the KorLex [11] for Korean, and a privately-constructed dictionary for Chinese.

We also use the part-of-speech similarity between two phrases. The part-of-speech information is expected to supplement the lexical information. This similarity is linearly interpolated with those two measures, as follows:

$$(1-\lambda) sim_{LEX}(\bar{p}_1, \bar{p}_2) + \lambda sim_{POS}(POS(\bar{p}_1), POS(\bar{p}_2)) \tag{9}$$

$$(1-\lambda) sim_{SYN}(\bar{p}_1, \bar{p}_2) + \lambda sim_{POS}(POS(\bar{p}_1), POS(\bar{p}_2)) \tag{10}$$

where $POS(\bar{p})$ represents the part-of-speech tag sequence of a phrase $\bar{p}$, and sim_{POS} is the same as Eq. (6) except that a part-of-speech is used instead of a word.

3.3 Gradual Update

The proposed estimation method makes use of translation probabilities for the opposite direction when estimating translation probabilities for the normal direction. Therefore, these bidirectional probabilities are complementary to each other, and can be gradually updated. We expect that this iterative method can produce a positive effect in the SMT system, because the conventional PBSMT system employs both bidirectional translation probabilities. The following equations describe the initial step and the iteration step in the update process.

- Initial step

$$p_0(\bar{f}|\bar{e}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{f}'} count(\bar{e}, \bar{f}')}\ ,\ p_0(\bar{e}|\bar{f}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{e}'} count(\bar{e}', \bar{f})} \tag{11}$$

- Iteration step

$$p_{i+1}(\bar{f}|\bar{e}) = \sum_{\bar{e}' \in E_{\bar{f}}} sim(\bar{e}, \bar{e}') p_i(\bar{e}'|\bar{f}) \times p_i(\bar{f}|\bar{e}) \times \frac{1}{Z_i(\bar{e})} \tag{12}$$

$$p_{i+1}(\bar{e}|\bar{f}) = \sum_{\bar{f}' \in F_{\bar{e}}} sim(\bar{f}, \bar{f}') p_{i+1}(\bar{f}'|\bar{e}) \times p_i(\bar{e}|\bar{f}) \times \frac{1}{Z_i(\bar{f})} \tag{13}$$

where $p_i(\bar{f}|\bar{e})$ and $p_i(\bar{e}|\bar{f})$ represent the phrase translation probabilities, which are iteratively updated i times, and $F_{\bar{e}}$ denotes a set of phrases paired with $\bar{e}$. The iteration step is terminated when a total change of probabilities reaches a threshold, which is determined by measuring the changes of the BLEU scores of the development set.

4. Experiments

We have experimented with the proposed method for English-to-Korean (E2K), Korean-to-English (K2E), English-to-Chinese (E2C) and Chinese-to-English (C2E) translation tasks. We have used an English-Korean parallel corpus, consisting of 488 K for training, 1 K for tuning, and 1 K sentence pairs for testing.[†] We have also used 485 K and 500 English-Chinese sentence pairs from the LDC corpora (LDC2005T10, LDC2005T06, and part of LDC2004T08) as training and development sets, respectively. The official evaluation set of NIST OpenMT 2008 Evaluation has been used as the test set for E2C and C2E translation.

We have used the open source SMT system, Moses [12], with default options as the baseline translation system, and have also used the minimum error rate training (MERT) for weight tuning of the system. The proposed method has been implemented by modifying the phrase pair scoring step in the training process, and replacing original translation probabilities in the phrase table by the updated probabilities.

The BLEU score [13] is used as the evaluation metrics. Performance in E2K, K2E and C2E are measured with

[†]This corpus is provided by SK Telecom only for research purposes. The parallel sentences are crawled over various online newswires. We have constructed three references for E2K evaluation, and have used only one reference for K2E evaluation.

Table 1 Performance of proposed methods (BLEU).

	E2K	K2E	E2C	C2E
Baseline (i=0)	25.49	14.87	24.86	16.27
LEX (i=1)	**25.88***	14.79	25.10*	**16.85***
LEX (i=2)	25.69	14.92*	25.21	16.33
LEX (i=3)	25.41	14.64	**25.72**	16.34
SYN (i=1)	**25.97***	15.08*	**25.42***	16.52*
SYN (i=2)	**25.97**	14.96	**25.45**	16.53
SYN (i=3)	**25.96**	14.95	**25.68**	16.00
LEX+POS (i=1)	25.67	15.02	**25.32***	**16.57***
LEX+POS (i=2)	**25.96***	15.03*	**26.06**	**16.63**
LEX+POS (i=3)	25.72	15.08	25.05	16.44
SYN+POS (i=1)	**25.83**	15.01*	**25.42**	16.43*
SYN+POS (i=2)	**25.91**	15.05	**26.07***	16.47
SYN+POS (i=3)	**25.84***	14.96	25.27	16.44

word-segmented translation results, while the performance in E2C is measured with character-segmented translation results.

Table 1 shows the BLEU scores in various language pairs.[†] We have compared four methods that use different similarity measures. In these experiments, we have iteratively updated each method three times, regardless of its termination condition. The asterisked scores in Table 1, which indicate the end of the iteration optimized on the development set, are regarded as the practical performance. The proposed method increased the BLEU score at the end of the iteration in all language pairs, thus implicating that our method is effective in improving the performance of the PBSMT system and is independent of the language pair. The gradual update increased the score until i=2 and decreased it from i=3 in many cases. We found that the method using SYN outperforms the method using LEX in E2K and K2E translation while it does not outperform the methods in E2C and C2E translation. This is because the resource used for obtaining the Korean synonym information is of a higher quality than that used for Chinese.

Though the method considering part-of-speech significantly improved the translation quality in E2C, the performance gain cannot compensate for the time cost. A short phrase consisting of one or two words can have different part-of-speech tags according to its context. Nevertheless, we assign only one for each phrase, which is the most frequently occurring part-of-speech tag sequence in the training corpus, and thus possibly incorrect part-of-speech information may be unhelpful in estimating translation probabilities.

The proposed method is expected to reduce target phrase selection errors during the decoding process. In order to evaluate how our method actually affects the phrase selection, we have measured the reduction rate of inadequate phrase selection by adopting the proposed method. We have first sampled 100 sentences from the test set, and extracted only phrases changed by applying the proposed method, and then manually judged each phrase as correct or incorrect.

[†]The bold font in Table 1 indicates the result where the difference to the baseline result is statistically significant at the 95% confidence level. We have used Zhang's significance tester [14].

Table 2 Phrase-level adequacy of 100 sampled translation results.

E2K	Baseline	SYN (i=2)
Total phrases	1,437	1,435
Changed phrases	293	292
Adequate phrases	222	234 (+5.4%)
Inadequate phrases	71	58 (−18.3%)
E2C	**Baseline**	**SYN (i=3)**
Total phrases	1,848	1,842
Changed phrases	302	295
Adequate phrases	215	227 (+5.6%)
Inadequate phrases	87	68 (−21.8%)

This was done by a Korean native speaker and a Chinese native speaker for E2K and E2C, respectively.

Table 2 shows the evaluation results in E2K and E2C translations. As shown in Table 2, the number of adequate phrases is increased in both translations, and the number of inadequate phrases are decreased by 18% in E2K and by 21% in E2C. From these experimental results, we can claim that the proposed method, considering semantic recoverability of phrase retranslation, helps prevent the PBSMT system from selecting inadequate phrase pairs.

5. Conclusions

We proposed an improved method for estimating translation probabilities by considering the semantic recoverability of phrase retranslation. Experimental results showed that the proposed method effectively prevented the PBSMT system from selecting incorrect phrase pairs. We also found that the method is effective in improving the translation quality of various language pairs.

In our future work, we will develop a more accurate method of measuring retranslation similarity, and a method for considering multi-word expressions, such as idioms and compound nouns. Furthermore, we also plan to devise a strategy that will enable the better selection of source phrases from a phrase table during the decoding process, as well as the better selection of target phrases.

Acknowledgements

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), and 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

[1] F.J. Och and H. Ney, "The alignment template approach to statistical machine translation," Computational Linguistics, vol.30, no.4, pp.417–449, 2004.

[2] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation," Proc. HLT/NAACL'03, pp.48–54, 2003.

[3] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, and A. Waibel, "The cmu statistical machine translation system," Proc. MT Summit IX, 2003.
[4] Y. Deng and W. Byrne, "Hmm word and phrase alignment for statistical machine translation," Proc. HLT'05, pp.169–176, 2005.
[5] Y. Liu, T. Xia, X. Xiao, and Q. Liu, "Weighted alignment matrices for statistical machine translation," Proc. EMNLP'09, 2009.
[6] G. Hong, S.W. Lee, and H.C. Rim, "Bridging morpho-syntactic gap between source and target sentences for english-korean statistical machine translation," Proc. ACL-IJCNLP'09, 2009.
[7] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2010.
[8] M. Sammer, K. Reiter, S. Soderland, K. Kirchhoff, and O. Etzioni, "Ambiguity reduction for machine translation: Human-computer collaboration," Proc. AMTA'06, 2006.
[9] C.L. Goh, T. Watanabe, A. Finch, and E. Sumita, "Discriminative reranking for smt using various global features," Proc. IUCS'10, pp.8–14, 2010.
[10] G.A. Miller, "Wordnet: A lexical database for english," Commun. ACM, vol.38, no.11, pp.39–41, 1995.
[11] A. Yoon, S. Hwang, E. Lee, and H.C. Kwon, "Construction of korean wordnet korlex 1.5," J. Korean Institute of Information Scientists and Engineers: Software and Applications, vol.36, no.1, pp.92–108, 2009.
[12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," Proc. ACL'07, 2007.
[13] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "Bleu: a method for automatic evaluation of machine translation," Proc. ACL'02, 2002.
[14] Y. Zhang and S. Vogel, "Measuring confidence intervals for the machine translation evaluation metrics," Proc. TMI'04, 2004.