The POSTECH's Statistical Machine Translation System for the IWSLT 2010

Hwidong Na, Jong-Hyeok Lee

Division of Electrical and Computer Engineering
Pohang University of Science and Technology (POSTECH)
San 31 Hyoja Dong, Pohang, 790-784, Republic of Korea
{leona,jhlee}@postech.ac.kr

Abstract

In this paper, we utilize segmentation alternatives. Our research contribution is a novel estimation method of the translation probabilities used in phrase-based statistical machine translation in order to reflect the trustworthiness of the segmentation. Our system, however, underperforms the baseline.

1. Introduction

Natural language sentences often require natural language analyses prior to performing machine translation. For example, a Chinese sentence requires word segmentation since it is originally not segmented. However, a source sentence can be segmented at various levels of granularity when it is ambiguous. This is analogous to the recognization of an acoustic signal. In real world situations, ambiguities of natural language sentences are inevitable. Hence the results of natural language analyses such as word segmentation do not necessarily provide the correct information.

A distribution of the segmentation probabilities of our training corpus (segmented by the Stanford Chinese Segmenter [1]) confirms this problem. About 80 percent of sentences are segmented with relatively high probabilities (more than 0.8). However, the rest show lower probabilities, which indicate the segmentation results may be incorrect. Therefore we need a clever solution to cope with the analysis erorr at the word segmentation stage.

One of the popular research approaches has explored alternative segmentations for the following reasons [2, 3, 4]. Even if the best segmentation is incorrect, less strong alternatives would be correct. It is also possible to consider a combination of the segmentation results from different segmentation methods. Moreover, in a small-sized parallel corpus, utilizing different segmentations is useful to increase the size of the corpus. Generally, the translation quality in SMT becomes better as the size of the parallel corpus increases.

In this paper, we utilize segmentation alternatives. Our research contribution is a novel estimation method of the translation probabilities used in phrase-based statistical machine translation (PBSMT) in order to reflect the trustworthiness of the segmentation. Our system, however, does not outperform the baseline method.

2. Prior Work

Previous studies have proposed merging the alternative analyses to deal with analysis errors for two reasons: 1) the strongest alternative is not necessarily the correct analysis and 2) most alternatives contain similar element such as common sub trees. For segmentation alternatives, [2] proposed a word lattice that represents exponentially large numbers of segmentations of a source sentence. [4] further integrated reordering information into the lattice. For parsing alternatives, [5] suggested a packed forest that encodes alternative derivations from a node. [3] combined the two approaches to benefit from both. They treated the alternative segmentations equally in the training corpus. However, if the segmentation is not very accurate, it may also harm the estimation of translation probabilities as well as fail to decode the test corpus.

Meanwhile, previous participants in IWSLT translation tasks have also combined the different PBSMT systems based on different word segmenations [6, 7]. They have reported great improvements from the combination. Therefore, it is worthwhile to replicate the improvement from combination by merging the alternatives, or vice versa. Unfortunately, we do not compare both and leave it as futher work.

An Arabic-to-English system has been proposed for the same purpose [8]. The system accepts variation of input segmentation from two different segmentors, regarding the segmentation result equally probable. Although the system is quite similar to ours, it is different that we discriminate low probable segmentations in the probability estimation, and the translation is Chinese-to-English.

3. Method

3.1. Estimating translation probabilities

We assume that the extracted phrase pairs from a sentence having a lower segmentation probability are also less important to estimate the translation probabilities. Therefore we build a standard format phrase table used in PBSMT whereas the probabilities reflect the segmentation probability of the sentence.

PBSMT regards a sentence pair $\langle F, E \rangle$ consisting of pairs of phrases $(\bar{f}, \bar{e})$, where $\bar{f}$ and $\bar{e}$ are continuous word sequences. For each sentence pair, we duplicate E for K dif-

Table 1: Summary of the used resources

ID	Usage	File name
train	Training	IWSLT12_DIALOG.train.??.with_interpreter.txt
	Training	IWSLT10_BTEC.train.??.txt
tune	Parameter tuning (source)	IWSLT10_DIALOG.devset.case+punc.src.??.sgm
	Parameter tuning (reference)	IWSLT10_DIALOG.devset.case+punc.mref.??.sgm
dev[1-9]	Evaluation (source)	IWSLT10.devset[1-9]_*.case+punc.src.??.sgm
	Evaluation (reference)	IWSLT10.devset[1-9]_*.case+punc.mref.??.sgm

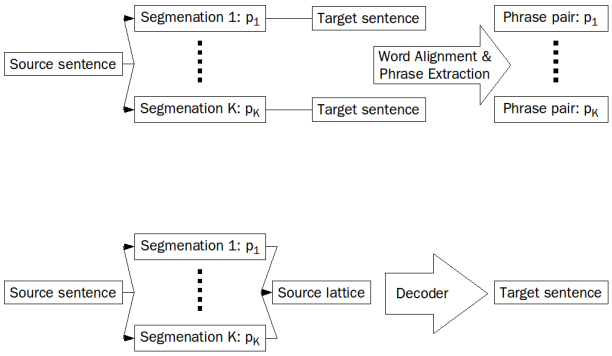

Figure 1: Overall architecture of our system. The upper diagram shows our propsed estimation, and the lower one shows lattice decoding.

ferenct segmentations of F. Hence, the size of the parallel corpus is multiplied by K. Then we obtain the word alignment of the increased parallel corpus.

Unlike traditional phrase pair extraction, we assign the segmentation probability p_{seg} of the source sentence for each phrase pair $(\bar{f}, \bar{e})$ extracted from a sentence pair. Then we define the fractional count for each unique phrase pair in the corpus. Finally, the maximum likelihood estimates for a phrase pair p_{phr} are computed based on the fractional count of extracted phrases. Lexical translation probabilities p_{lex} of a phrase pair are also computed using the fractional count.

$$count(\bar{f}, \bar{e}) = \sum p_{seg}(\bar{f}, \bar{e}) \tag{1}$$

$$p_{phr}(\bar{f}|\bar{e}) = \frac{p_{phr}(\bar{f}, \bar{e})}{\sum_{\bar{e}'} p_{phr}(\bar{f}, \bar{e}')} \tag{2}$$

$$p_{lex}(\bar{f}|\bar{e}) = \frac{p_{lex}(\bar{f}, \bar{e})}{\sum_{\bar{e}'} p_{lex}(\bar{f}, \bar{e}')} \tag{3}$$

, where

$$\begin{aligned} p_{phr}(\bar{f}, \bar{e}) &= \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}', \bar{e}'} count(\bar{f}', \bar{e}')} \\ p_{lex}(\bar{f}, \bar{e}) &= \sum_a p_{lex}(\bar{f}, \bar{e}, a) \\ &= \sum_a \prod_{(i,j) \in a} p_{lex}(f_j, e_i) \\ &= \sum_a \prod_{(i,j) \in a} \frac{count(f_j, e_i)}{\sum_{f'_j, e'_i} count(f'_j, e'_i)} \end{aligned}$$

Our proposed method is different from previous works that treat all segmentation alternatives the same or discard less probable segmentations under a thresold [2]. The upper diagrams in Figure 1 shows our proposed method.

3.2. Decoding lattice

Theoretically for n characters, 2^{n-1} segmentations are possible at most. In practice, however, we have K different segmentations with probabilities for a sentence. The K different segmentations can be effeciently compressed into a lattice as described in [2]. Then we decode the lattice for each source sentence in order to translate it into the target language. The lower diagram in Figure 1 shows lattice decoding.

4. Experiment

4.1. Resource

Thoughout our experiments, we use the supplied resources only. The usage of the resources is summarized in Table 1. We ignore the dialog annotations such as ID and speaker information. The provided Chinese sentences are already segmented, and we regard the provided segmentation as the "1-best" (the first and last row in Table 2). For the variation of segmenation, we use two segmentation alternatives (K=2) according to two different segmentation standards that follow the Peking University (PKU) and Penn Chinese Treebank (CTB), and generate a lattice as input (the second and third row in Table 2). The increased size of the training corpus is about twice as large as the original corpus. No other supplientary data such as a large monolingual corpus is used.

Table 2: Chinese-English CRR BLEU scores on various test corpora

Estimation	Input type	tune	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	Avg.
Moses	1-best	47.99	46.16	48.52	49.85	22.65	20.72	30.80	43.01	39.29	35.70	37.41
Moses	lattice	48.85	46.03	48.43	52.04	23.22	21.53	32.50	42.88	39.80	36.05	38.05
Proposed	lattice	47.35	44.86	45.63	49.96	22.52	19.85	28.08	40.03	35.23	35.18	35.93
Moses-chart	1-best	47.04	47.87	50.18	52.22	23.09	21.00	34.78	44.20	40.85	36.48	38.96

4.2. Setting

We carried out four different experiments with two experimental variables using Moses [9]. One variable is whether our proposed estimation method is used or not, and the other is whether Moses decodes the best segmentation or lattice of the segmentation alternatives. As a baseline system, we also utilize Moses-chart, which is a hierarchical PBSMT [10]. We use GIZA++ [11] to obtain bidirectional word alignments and refine them using "grow-diag-final-and" heuristics. The weights of the probabilities are tuned using the minimum error rate training (MERT) [12]. We use SRILM [13] for trainining a 5-gram language model with modified Kneser-Ney smoothing. All settings are assembled and executed using the experimental management system in Moses with a minor modification for tuning parameters on multiple references.

4.3. Result

We focused on Chinses-English CRR PBSMT using our proposed method. Table 2 shows the automatic evaluation results by BLEU [14] on the test corpora as well as the paramter tuned score. We report the best score among several results since we observed that MERT gives quite unstable translation results. Decoding lattice (the second row) is competitive with the original PBSMT (the first row). However, our proposed method underperforms the other settings (the third row). The tuned scores using lattice are relatively higher than the others. The best performance on average comes from Moses-chart, which is our primary run. Our contrastive run is the translation result applying our proposed method. We also submitted our primary run for Chinese-English ASR, and English-Chinese CRR and ASR using Moses-chart.

5. Discussion

On the test corpora, Our proposed method failed to achieve improvement over Moses and Moses-chart, which are very strong baselines. We suspected that the low quality segmentations may cause the word alignment error, and discarded the segmentations which have lower probabilities than a given threshold. However, it did not improve performance either. It is interesting that the lattice input for tuning gives higher BLEU score than 1-best. As we mentioned above, MERT gives very unstable results for this corpus, and this causes an overfitting problem in the lattice input case.

6. Conclusion

In this paper, we described our proposed method at the IWSLT 2010 DIALOG Task. Lattice decoding worked competitively, but tailoring translation probability for lattice decoding proposed in this paper did not work. We will further investigate better methods in the future IWSLT.

7. Acknowledgments

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST No. 2010-0016489), in part by the Electronics and Telecommunications Research Institue (ETRI), and in part by the BK 21 Project in 2010.

8. References

[1] P.-C. Chang, H. Tseng, D. Jurafsky, and C. D. Manning, "Discriminative reordering with Chinese grammatical relations features," in *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 51–59. [Online]. Available: http://www.aclweb.org/anthology/W09-2307

[2] C. Dyer, S. Muresan, and P. Resnik, "Generalizing word lattice translation," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 1012–1020. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1115

[3] H. Mi, L. Huang, and Q. Liu, "Machine translation with lattices and forests," in *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, August 2010, pp. 837–845. [Online]. Available: http://www.aclweb.org/anthology/C10-2096

[4] C. Dyer and P. Resnik, "Context-free reordering, finite-state translation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, June 2010, pp. 858–866. [Online]. Available: http://www.aclweb.org/anthology/N10-1128

[5] H. Mi, L. Huang, and Q. Liu, "Forest-based translation," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 192–199. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1023

[6] M. Li, J. Zhang, Y. Zhou, and C. Zong, "The CASIA Statistical Machine Translation System for IWSLT 2009," in *Proc. of the International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 83–90.

[7] P. Nakov, C. Liu, W. Lu, and H. T. Ng, "The NUS Statistical Machine Translation System for IWSLT 2009," in *Proc. of the International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 91–98.

[8] F. Bougares, L. Besacier, and H. Blanchon, "Lig approach for iwslt09: using multiple morphological segmenters for spoken language translation of arabic," in *Proc. of the International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 60–64.

[9] 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," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180. [Online]. Available: http://www.aclweb.org/anthology/P07-2045

[10] D. Chiang, "Hierarchical phrase-based translation," *computational linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[11] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[12] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, July 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021

[13] A. Stolcke, "Srilm – an extensible language modeling toolkit," 2002. [Online]. Available: http://citeseer.ist.psu.edu/stolcke02srilm.html

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. [Online]. Available: http://www.aclweb.org/anthology/P02-1040