Phrase Table Combination Deficiency Analyses in Pivot-Based SMT

Yiming Cui, Conghui Zhu, Xiaoning Zhu, Tiejun Zhao, and Dequan Zheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{ymcui,chzhu,xnzhu,tjzhao,dqzheng}@mtlab.hit.edu.cn

Abstract. As the parallel corpus is not available all the time, pivot language was introduced to solve the parallel corpus sparseness in statistical machine translation. In this paper, we carried out several phrase-based SMT experiments, and analyzed the detailed reasons that caused the decline in translation performance. Experimental results indicated that both covering rate of phrase pairs and translation probability accuracy affect the quality of translation.

Keywords: Machine translation, Pivot method, Phrase table combination.

1 Introduction

In order to solve the parallel language data limitations, the pivot language method is introduced [1-3]. Pivot language becomes a bridge method between source and target languages, whose textual data are not largely available. When we choose a language as pivot language, it should provide a relatively large parallel corpus either in source-pivot direction, or in pivot-target direction.

In this paper, we focus on the phrase tables generated by two directions (*source-pivot, pivot-target*), that is *triangulation* method. This method multiplies corresponding translation probabilities and lexical weights in *source-pivot* and *pivot-target* phrase table to induce a new *source-target* phrase table.

2 Related Work

Utiyama and Isahara [3] investigate in the performance of three pivot methods. Cohn and Lapata [4] use multi-parallel corpora to alleviate the poor performance when using small training sets, but do not reveal the weak points of current phrase-based system when using a pivot method. What affects the pivot-based machine translation quality is discussed in general aspects by Michael Paul and Eiichiro Sumita [5], but not detailed explained in a certain aspect.

3 Pivot Method In SMT

When combining the two phrase tables generated by *source-pivot* and *pivot-target* corpora, we should take two elements into account.

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 355–358, 2013.

The first element is phrase translation probability. We assume that source phrases are independent with target phrases. In this way, we can induce the phrase translation probability $\varphi(s \mid t)$ when given the pivot phrases as Eq.1.

$$\varphi(s \mid t) = \sum_p \varphi(s \mid p) \cdot \varphi(p \mid t) \tag{1}$$

Where *s, p* and *t* denotes the phrases in the source, pivot and target respectively.

The second element is lexical weight, that is word alignment information *a* and in a phrase pair *(s,t)* and lexical translation probability *w(s|t)*[6].

We assume a_1 and a_2 be the word alignment inside phrase pairs (*s,p*) and (*p,t*) respectively, and the word alignment *a* of phrase pair (*s,t*) can be got by Eq.2.

$$a = \{(s,t) \mid \exists p : (s,p) \in a_1 \ \& \ (p,t) \in a_2\} \tag{2}$$

Then we can estimate the lexical translation probability by induced word alignment information, as shown in Eq.3. In this way, we can use source-pivot and pivot-target phrase table to generate a new source-target phrase table.

$$w(s \mid t) = \frac{count(s,t)}{\sum_{s'} count(s',t)} \tag{3}$$

4 Experiments

In our experiments, the pivot language is chosen as English, because of its large availability of bilingual corpus. Our goal is to build a Chinese-Japanese machine translation system. The corpus is selected as HIT trilingual parallel corpus [7]. There are two ways to divide the corpus. The first is *parallel* one, which indicates that both directions share the same training sets; the second is *non-parallel* one, which means the training sets of two directions are independent with each other. The Statistics are shown in Table 1.

Table 1. Zh-en-jp parallel corpus

	Train	Dev	Test
Parallel	59733	2000	1000
Non-parallel	29800*2	2000	1000

4.1 Coverage of Phrase Pairs

The coverage of phrase pairs shows how many phrases appear in the phrase table, and it can be an indicator that reveals the difference between standard and pivot model.

The scales of each phrase tables are shown in Table 2. Then we extracted phrases separately from standard and pivot phrase table, and deleted all repeated phrases respectively. We can calculate the number of phrases. The results are shown in Table 3.

Table 2. The scale of two models

	Standard	Pivot
Parallel	1088394	252389200
Non-Parallel	1088394	92063889

Table 3. Number of phrases

	Parallel		Non-Parallel	
	zh	jp	zh	jp
Standard	521709	558819	521709	558819
Pivot	320409	380929	97860	131682

In general, we can see some problems revealed in figures above. Firstly, though pivot phrase table may be larger than the standard one in size (230 times bigger), the actual phrases are less than the standard one (about 60%). This reminds us that during the phrase table combination, some phrases would be lost. That is to say, the pivot language cannot bridge the phrase pairs in *source-pivot* and *pivot-target* directions. Secondly, due to a larger scale in phrase table and lower useful phrases, pivot phrase table brings so much noise during the combination. This would be a barrier, because the noise would affect both the quality and the efficiency in the translation process.

Then we carried out the following experiments to show what caused low phrase coverage. We extracted the phrase pairs (s,t) that exist in standard model but not in pivot model. When given phrase s, we searched the Chinese-English phrase table to get its translation e, and use corresponding phrase t to search the English-Japanese phrase table to get its origin e'. Then we compared output e and e', and see what reasons that caused the failure in connecting phrases in two models. We calculated the number of phrase pairs that was successfully connected by pivot in the Table 5.

Table 4. Connected phrase pairs in pivot model

	Parallel	Non-parallel
Connected phrase pairs	310439(34.75%)	73044(9.84%)

As we can see above, in parallel models there are only 34.75% phrase pairs connected, and in non-parallel situation, the rate goes down to 9.84%. So we examined the output file, and noticed some phenomenon which accounts for low number of connected phrase pairs. Firstly, Arabic numerals can be converted into English (e.g. 100 -> one hundred); secondly, the word with similar meanings can be converted (e.g. 8.76% -> 8.76 percent); thirdly, punctuations can be removed or added (e.g. over -> over.).

4.2 Translation Probability Accuracy

We also investigated whether translation probability accuracy affects the translation result a lot. We found the intersection phrase pairs in standard and pivot phrase tables, and generated two new phrase tables, using the common phrase pairs of standard and pivot phrase tables, and the probabilities of each. In this way, we can see in the condition of the same phrase pairs, how results differ when using different translation probability. The results are shown in Table 6, which the parameters were not tuned.

Table 5. BLEU scores of old and new generated models(with *parallel* data)

	Standard	Pivot
Old	26.88	17.56
New	24.99	21.44

We can see that, in new models, the variety of the probability brings a 3.55 BLEU score gap. We found a quite unusual phenomenon that, though new pivot model reduce to 0.85% of its original size, the BLEU score rise up to 21.44. This can also be a proof that there are too much noise in pivot phrase table. The noise affected the translation quality, and translation effectiveness is also impacted due to its large size.

5 Conclusion

The experiments showed that the translation result may decrease along with the change of coverage of phrase pairs and translation probability accuracy. We still need to improve the covering rate of phrase pairs, and we also should improve our translation probability accuracy, not merely using a multiplication of each probabilities.

Acknowledgements. This paper is funded by the project of National Natural Science Foundation of China (No.61100093) and the project of National High Technology Research and Development Program of China(863Program) (No. 2011AA01A207).

References

1. de Gispert, A., Marino, J.B.: Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In: Proceedings of 5th International Conference on Language Resources and Evaluation, pp. 65–68 (2006)
2. Wu, H., Wang, H.: Pivot Language Approach for Phrase-Based Statistical Machine Translation. In: Proceedings of 45th Annual Meeting of ACL, pp. 856–863 (2007)
3. Utiyama, M., Isahara, H.: A comparison of pivot methods for phrase-based sta-tistical machine translation. In: Proceedings of HLT, pp. 484–491 (2007)
4. Cohn, T., Lapata, M.: Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In: Proceedings of the 45th ACL, pp. 348–355 (2007)
5. Paul, M., Sumita, E.: Translation Quality Indicators for Pivot-based Statistical MT. In: Proceedings of 5th IJCNLP, pp. 811–818 (2011)
6. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 127–133 (2003)
7. Yang, M., Jiang, H., Zhao, T., Li, S.: Construct Trilingual Parallel Corpus on Demand. In: Huo, Q., Ma, B., Chng, E.-S., Li, H. (eds.) ISCSLP 2006. LNCS (LNAI), vol. 4274, pp. 760–767. Springer, Heidelberg (2006)