http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 4895-4900, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4895.4900

Phrase-table Filtering for Phrase-based Machine Translation

Li Bin, Ma Ning and Liang Wuqi
Anhui Radio and Television University, Hefei, Anhui, 230022, China

Abstract: Phrase-based machine translation models have shown better translations than Word-based models, but many phrase pairs do not encode any relevant context. In order to decrease the number of phrase pairs occurring in phrases table for phrase-based machine translation, we described and compared several approaches for filtering phrase pairs. While the phrase pairs extracted and tested in our Machine Translation (MT) system, they all performed satisfactorily. Comparing to each other, the method of Model-best got a very good result. Based on the Model-best, two methods CPS and MB and LLR and MB we proposed by combining Model-best with the method of "Composition" and Log Likelihood Ratio. Both of their translation performances are better than the method of Model-best.

Key words: Machine translation, phrase pairs filtering, combination of methods, phrase-table, model-best

INTRODUCTION

In the summer of early 1990s while Brown *et al.* (1993) had proposed a statistical machine translation model based on source-channel, statistical machine translation got a rapid development. Currently, the phrase-based statistical machine translation has become a mainstream statistical machine translation method. In phrase-based statistical machine translation, translation model used to reflect the correspondence relationship between source language and target language. The important step of this approach is regarded any continuous strings as phrases which extracted from the word-aligned bilingual corpus automatically and then translated these phrases to target language ones. Thus, the first step of the whole process for translation is segmenting the sentences into phrases and then translating each phrase into a corresponding target phrase; the sequence of phrases should be reordered according some rules at last. But there are some shortages of this approach as follow:

Firstly, the translation models are generated by the stages of the word alignment, phrase alignment, phrase grading and so on. There are so many phrases which are consisted of some simple words; the accuracy of results for translation will be bad if the models consist of those inappropriate phrase pairs.

Secondly, in generally speaking, phrase translation tables are the cores of phrase-based statistical machine translation systems. But it will take a lot of memory space and time to extracting so many phrase pairs by the method of word alignment.

Consequently, it is very significance to find a way to filtering the incorrect phrases and reducing the number of the phrases in phrase table to decrease the memory space meanwhile the translate effect become well or not bad at least.

The study is organized as follow. In section 2, related works are first introduced for filtering phrase pairs. Many approaches are proposed for filtering phrase pairs in different ways in section 3. Meanwhile, we have tried the methods by combining two of them. Section 4 is an analysis of the results for the experiments which aims to reveal the good and bad aspects of the approach. A conclusion is given and further study is proposed in section 5.

RELATED WORKS

Koehn (2004) proposed a method to extracting effective phrase pairs by the probability thresholds and the translation type thresholds. Johnson *et al.* (2007) removed the inappropriate phrase pairs of translation models in the phrase tables based on the method of p-value. In this case, the phrase pairs whose p-value are higher, are more likely to be pruned. Zettlemoyer and Moore (2007) also proposed a method based scoring function and redundancy limitations to remove phrase pairs. Wu and Wang (2007) obtained the better phrase pairs using the thresholds of calculated results of phrase pairs logarithmic likelihood ratio and the translation probability. Eck *et al.* (2007) got the relevant phrase pairs on the base of the frequency of sources phrase as well the role of the actual value (or score value) of phrases. And then they proposed two approaches based on Model-best and Metric-best to filter phrase pairs (Eck *et al.*, 2007). The research is followed by Tomeh *et al.* (2009) which

Corresponding Author: Li Bin, Anhui Radio and Television University, Hefei, Anhui, 230022, China

pairs' characters are measured by complexity. At present, the effective of translation based on Model-best is the best. Ling *et al.* (2012) proposed a relative entropy model for translation models, that measures how likely a phrase pair encodes a translation event that is derivable using smaller translation events with similar probabilities.

In our study, we compared the principles and results of such various methods and then proposed a new better approach by combining some of them.

PHRASE PAIRS FILTERING

There are lots of approaches for filtering the phrase pairs from phrase tables. And some of them can obtain highly satisfied effect. In this section, we will introduce and analyze the internals of them.

Target language phrase length (TLPL): Extract phrases according by target language phrase length from phrase tables is a simple, but effective method in the previous research. In phrase tables, the lengths of target language phrases are unfixed; some of phrases are too long and some of them are short. In many cases, the longer lengths of target language phrases, the lower probability being used when we tested them, especially the words of phrase are too many. So that we can remove the long length phrases from the phrase tables by some relevant threshold values. On the one hand, it take few influence on the effect of translation, on the other hand, the phrase tables have been optimized.

Log likelihood ratio (LLR): There are some over-estimation problems for unusual phrases in the research of translation based on phrase pairs. To solve the problems of over-estimation, Dunning (1993) had proved that the log likelihood ratio method achieved better effect. The parameters of log-likelihood ratio can be defined in Table 1.

The formula of the log likelihood ratio as follow:

$$G^2(\bar{f},\bar{e}) = -2\log\lambda = \sum_{i,j} n_{ij}\log\frac{n_{ij}N}{R_iC_j} \quad (1)$$

We calculated the log likelihood ratio for each phrase pair and then filtered ones whose value are lower than threshold we defined.

p-value: In statistical significance testing the P-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. It represents a measure of correlation just like the mutual information. The formula as follow:

Table 1: Contingency table for phrase pairs

	Target language phrase	Non target language phrase	Total
Source language phrase	n_{11}	n_{12}	R_1
Non source language phrase	n_{21}	n_{22}	R_2
Total	C_1	C_2	N

$$p\_value(k) = \sum_{k=n_{11}}^{\infty} p_h(k) \quad (2)$$

$$p_h(k) = \binom{R_1}{n_{11}}\binom{R_2}{n_{21}} / \binom{N}{C_1} \quad (3)$$

where, p-value represents the value of correlation between the source language phrase and the target language phrase. In most cases, the larger of P-value, the greater relevance for source language phrase and target language phrase, thus it will be remained. Here, we can remove the phrase pairs whose p-value value is relatively small by some thresholds.

Perplexity (PPL): Perplexity is usually considered as the metrics of the quality for evaluating the language model. Here, it is used as the metrics of the probability of the source language phrases or target language phrases which existing in corpus. The formula of the perplexity containing k phrases as follow:

$$PPl(k) = 2^{-\log 2(k\sqrt{p(w_1)\cdot p(w_2|w_1)\cdots p(w_k|w_{k-1}\cdots w_1)})} \quad (4)$$

where, $p(w_1)$ represent the probability of the word w1 occurred in corpus. $p(w_2|w_1)$ represent the probability of the word w_2 occurred when w_1 have occurred in corpus. In order to describe conveniently, we have modified the formula as follow:

$$PPL(k) = \log_2(\sqrt[k]{p(w_1)\cdot p(w_2|w_1)\cdot\ldots\cdot p(w_k|w_kw_{k-1}\ldots w_1)} \quad (5)$$

From the formula five defined, we concluded that the value of perplexity is greater; the reusability of phrase is higher, so it can be taken for a meaningful phrase. On the contrary, the value of perplexity is lower; the phrase's reusability is smaller. So, we can select an appropriate threshold to filter those phrases whose perplexities are lower.

Entropy: In information theory, entropy is a concept used to measure the amount of information. A system is more orderly, the information entropy is lower. Conversely, a system is more confusion, the entropy is higher. Therefore, it can be said entropy is a measure system of the degree of ordering. In this experiment, we used to determine the value of the entropy of a phrase boundary.

We considered that w_1, w_2, ..., w_{n-1} is an phrase containing n-1 words, now we can calculated the entropy of a word which behind the word w_{n-1} and then determine word w_{n-1} is the boundary of the phrase or not. The formula as follow:

$$\text{Entropy} = -\sum_{i=1}^{m} p(v_i \mid w_{n-1}w_{n-2}...w_1)\log p(v_i \mid w_{n-1}w_{n-2}...w_1) \quad (6)$$

where, the vocabulary table is (v_1, v_2 ..., v_m), the probability of phrase which consist of $w_1,w_2,\ldots,w_{n-1},w_n$ is low if the value of entropy is small. And we take word w_{n-1} as the boundary of the phrase and the phrase consisted words w_1, w_2, ... , w_{n-1} is a useful and meaningful. Otherwise, we take the word w_n as the boundary and considered that phrase with n words is a good one. Because there are two boundaries for each phrase, one is front of it, other one is behind of it. In the study, we chose the average of those two entropies.

Model-best: In the processing of translation, we always want to extract relevant phrase pairs from the best translation hypothesis. Then we calculated the scores of phrases based on the numbers that occurred in the best translation candidates. However, in many cases, only a few phrases that appeared in the best translation candidates and this will inevitably lead to a phenomenon that we can't get the suitable phrases when testing. To solve this problem, we consider the top 10 best translation candidates for scoring. The formula as follow:

$$\text{score}(pp) = \sum_{i=1}^{10} \frac{\#pp - in - i - best}{i^2} \quad (7)$$

In formula, #pp-in-i-best is the number of phrase 'pp' occurred in the ith candidates. Model-best 10 represented the top 10 best translation candidates as we all know. Considering the experiment, we scored the phrase pairs which have extracted from top 10 best translation candidates, then sorted them according the scores and selected the phrases with higher scores as the candidates of the phrases table.

Combination of methods: While training, we usually find a long phrase pair could be composed by other two or more existed phrase pairs. Thus, the long phrase pairs existing in phrase table are redundant. We can delete the long ones from the phrases table so that the memory of phrase table becomes smaller. Thus, the phrases in table are short units. The method can be called "Composition". In our study, we chose the phrase pairs by combining Model-best with those shorter ones. We called this method as CPS and MB. For study, we deleted the longer phrase pairs in phrases table at first. And then, we extracted the phrase pairs using the method of Model-best. The number of the phrase pairs we selected is as same as the experiment in Model-best.

Meanwhile, we have proposed another approach which is combined by LLR and Model-best. The method extracted the phrase pairs by two methods simultaneously. Firstly, we got about 60 thousands phrase pairs whose scores are in top of table calculated by LLR. Secondly, we added the 50 thousands phrase pairs exacted by the method of Model-best. Thirdly, we removed the same phrase pairs selected by both of them. At last, we added 1 million, 1.5 millions, 2 millions and 2.5 millions phrase pairs from Model-best method with LLR. We called the method as LLR and MB.

RESULTS AND ANALYSIS

Primary results: We took a generic field bilingual corpus as training and testing. Finally, we obtained total 5,591,086 phrase pairs which are extracted by GIZA++. There are 1167 sentences for statistics machine translation testing. The BLEU score was taken to measure the preformance of the translation (Johnson *et al*., 2007; Koehn, 2004; Koehn *et al*., 2007). The value of BLEU score for translation result was 17.14 when we took all phrase pairs for training. In order to find an effective and significant method for the experiment, we discarded at least a half of phrase pairs from phrase table. The max number of the phrase pairs we selected was 3 million.

Discussion and analysis: We tried the methods of Model-best and other single ones. The performance of machine translation of those single method can be observed in Fig. 1.

It can be seen in the Fig. 1, every method received a good BLEU score (about 16.8) when the number of phrase pairs was about a half of the total phrase pairs (near 3,000,000). The best score was 16.84 obtaining by Model-best at the number of phrase pairs was about 2.5 million. The BLEU score usually become smaller when the number of the phrase pairs decrease for training.

Although the method of TLPL was so simple, it received a better BLEU score yet. It performed stably and availably at the same time. The BLEU scores received by methods of LLR and P-value were quite different. Their scores performed suddenly because the number of phrase pairs remained was changed sharply by different thresholds. But the scores they got are better than TLPL's and PPL's.

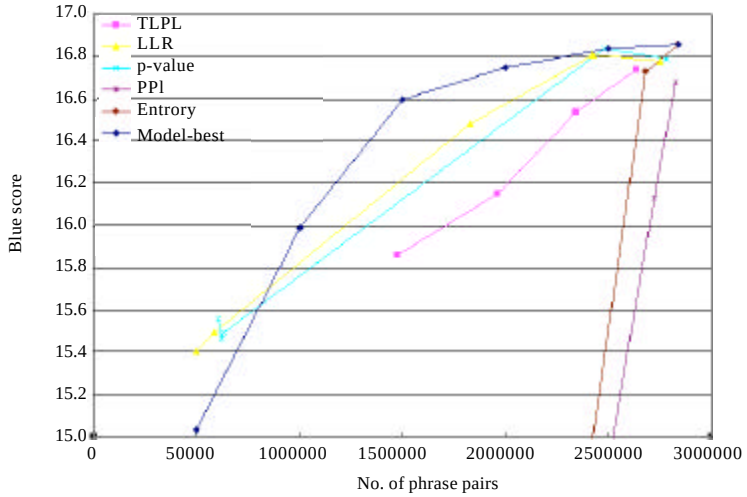

Fig. 1: Performance of each single method

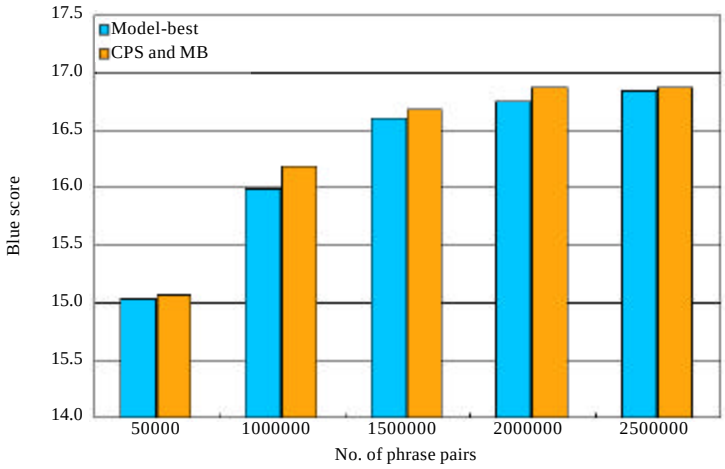

Fig. 2: Performance of CPS and MB and Model-best

The approach of Model-best was sure of the best if only just using the single method for punching. It performed the better result than others. The results giving by PPL and entropy were bad, because the number of the phrase pairs changed sharply through different thresholds. The characters of the phrase pairs were not obvious extracted by those two methods. When the number of phrase pairs was less, about 50 thousands, just only one-tenth of the total phrase pairs, the methods p-value and LLR reached the better performance.

What's more, we compared the result of two method Model-best and CPS and MB in Fig. 2. The performance got better when the number of the phrase pairs increased and then kept in a certain levels. However, even the numbers of the phrase pairs of two methods were the same, the BLEU scores they performed were different and the method of Model-best got a little lower score than CPS and MB's. The method of CPS and MB got the better scores at each point, because the phrase pairs extracted by CPS and MB were more effective for machine translation. The best BLEU score was 16.88 obtained by CPS and MB at the number of phrase pairs was 2,500,000. The performance of the combined methods showed in Fig. 3. Based on Model-Best, the methods combined others with it, CPS and MB and LLR and MB, got the best performance. The BLEU scores they performed were the similar. The former reached the best score 16.9 at the number of phrase pairs for training was 2,103,317, the later got the best score 16.88 at the number was 2,000,000. They were all better than 16.86 which received by the method of Model-best at the number of phrase pairs is 2,847,683. If only took about 2,000,000 phrase pairs for

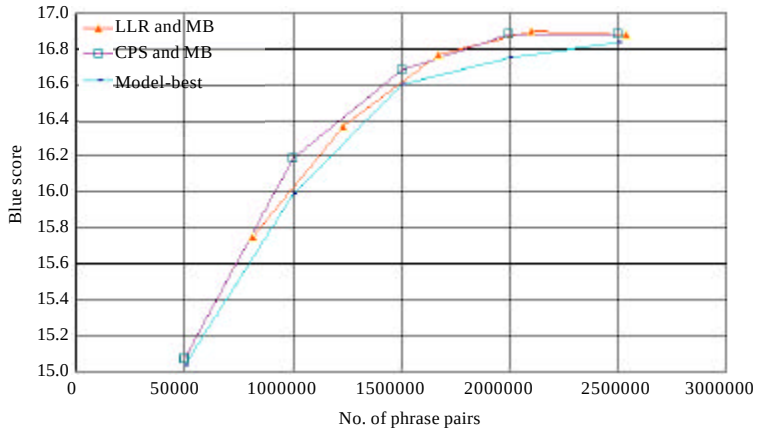

Fig. 3: Performance of combined methods

training, the BLEU score received by the method of Model-best was just 16.75. At same number of phrase pairs in every point, CPS and MB and LLR and MB received the better scores than only using the method of Model-best.

CONCLUSION AND FUTURE WORK

In the study, we have filtered the phrase pairs from the phrases table using different approaches. Taking those phrase pairs for machine translation test, we have got some good results. The method of Model-best performed particularly prominently if only taking just one method to our experiments. Based on Model-best, CPS and MB and LLR and MB were proposed in our experiments. They all achieved the best BLEU scores at the number of phrase pairs was about only one third of total phrases.

In the future, we will try more effective approaches for phrase pairs filtering and consider the techniques of combining and fusion. And we will do our best to reduce the number of the phrase pairs to decrease the memory store base on the premise that the BLEU score change slightly.

ACKNOWLEDGMENT

This study was supported by the Foundation for Young Talents in College of Anhui Province under Grant 2012SQRL230.

REFERENCES

Brown, P.F., V.J.D. Pietra, S.A.D. Pietra and R.L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. Comput. Ling., 19: 263-311.

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. Comput. Linguist., 19: 61-74.

Eck, M., S. Vogel and A. Waibel, 2007. Translation model pruning via usage statistics for statistical machine translation. Proceeding of NAACL HLT, April 22-27, 2007, Rochester, New York, pp: 21-24.

Eck, M., S. Vogel and A. Waibel, 2007. Estimating phrase pair relevance for translation model pruning. Proceedings of the 11th Machine Translation Summit, September 10-14, 2007, Copenhagen, Denmark, pp: 159-165.

Johnson, J.H., J. Martin, G. Foster and R. Kuhn, 2007. Improving translation quality by discarding most of the phrasetable. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Canada, pp: 965-967.

Koehn, P., 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, September 28-October 2, 2004, Washington, DC., USA., pp: 115-124.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch and M. Federico et al., 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, June 2007, Prague, pp: 177-180.

Ling, W., J. Graca, I. Trancoso and A. Black, 2012. Entropy-based pruning for Phrase-based machine translation. Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, July 12-14, 2012, Jeju Island, Korea, pp: 962-971.

Tomeh, N., N. Cancedda and M. Dymetman, 2009. Complexity-based phrase-table filtering for statistical machine translation. Proceedings of the 12th Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada, pp: 144-151.

Wu, H. and H. Wang, 2007. Comparative study of word alignment heuristics and phrase-based SMT. Proceedings of the 11th Machine Translation Summit, September 10-14, 2007, Copenhagen, pp: 507-514.

Zettlemoyer, L. and R. Moore, 2007. Selective phrase pair extraction for improved statistical machine translation. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York, April 22-27, Association for Computational Linguistics, Morristown, NJ, USA., pp: 209-212.