Train the Machine with What It Can Learn

− Corpus Selection for SMT

Xiwu Han
School of Computer Science and Technology,
Heilongjiang University,
Harbin City 150080 China
hxw@hlju.edu.cn

Hanzhang Li
School of Computer Science and Technology,
Heilongjiang University,
Harbin City 150080 China
lhj@hlju.edu.cn

Tiejun Zhao
School of Computer Science and Technology,
Harbin Institute of Technology,
Harbin City 150001 China
tjzhao@mtlab.hit.edu.cn

Abstract

Statistical machine translation relies heavily on available parallel corpora, but SMT may not have the ability or intelligence to make full use of the training set. Instead of collecting more and more parallel training corpora, this paper aims to improve SMT performance by exploiting the full potential of existing parallel corpora. We first identify literally translated sentence pairs via lexical and grammatical compatibility, and then use these data to train SMT models. One experiment indicates that larger training corpora do not always lead to higher decoding performance when the added data are not literal translations. And another experiment shows that properly enlarging the contribution of literal translation can improve SMT performance significantly.

1 Introduction*

Parallel corpora are generally considered indispensable for the training of a translation model in statistical machine translation (SMT). And most researchers tend to agree on the opinion that the more data is used to estimate the parameters of the translation model, the better it can approximate the true translation probabilities, and in turn this will lead to a better translation performance. However, even if large corpora are easily available, does an SMT system have the ability or intelligence to make full use of a training set?

Another aspect is that larger amounts of training data also require larger computational resources. With increasing quantities of training data, the improvement of translation quality will become smaller and smaller. Therefore, while continuing to collect more and more parallel corpora, it is also important to seek effective ways of making better use of available parallel training data.

Literal translation and free translation are two basic skills of human translation. A literal translation is a translation that follows closely the form of the source language, also known as word-for-word translation (Larson 1984).

According to Mona Baker (1992) translation needs to maintain equivalence at different levels across languages. In bottom-up sequence, these levels are: the word level, the above word level, the grammatical level, the textual level and the pragmatic level. Lower levels of equivalence are often embedded in literal translation and easily maintained, whereas higher levels are very important for free translation and very difficult to be achieved even for experienced translators because this kind of equivalence more often than not calls for thorough analysis and understanding of the source language, which is obviously what an SMT system cannot be capable of. So from this perspective SMT may be regarded as a beginner in learning how to translate.

The training of statistical machine translation mainly depends on the alignment probabilities estimated from certain frequencies observed in a parallel corpus. Thus, we may say that SMT translates according to its bilingual scanning experiences, and there is actually no deep comprehension during the coding and decoding process.

Since human learners of translation generally begin with the comparatively simpler techniques of literal translation, our efforts described in this paper are intended to discover whether a corpus

* This research is jointly supported by the National Natural Science Foundation of China under Grant No.60773069 and 60873169.

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 27–33,
Suntec, Singapore, 6 August 2009.

of literal translations better suits the training of statistical machine translation.

In the following, section 2 introduces our corpus and proposes a combined method to recognize sentence pairs of literal translation. Section 3 describes our experiments with the acquired corpus on SMT training from two points of view. Section 4 analyzes the results from a linguistic point of view. And the conclusion is given in Section 5 with some suggestion for further work.

2 Literal Translation Recognition

Early machine translations were notorious for bad literal translations especially of idioms. However, good literal translation means to translate a sentence originally, and to keep the original message form, including the construction of the sentence, the meaning of the original words, use of metaphors and so on. Such a translation would be fluent and easy to comprehend by target language readers. If we suppose that the training corpus for SMT is mainly constituted of good translations, our first task is to identify those literally translated sentence pairs.

2.1 Our Corpus

The corpus used for our experiment consists of 650,000 bilingual sentence pairs of English and Chinese, which were gathered either from public and free Internet resources or from our own translation works. The sentences are either translated from Chinese to English or vice versa.

To facilitate the process of recognition, before the SMT experiment we preprocessed the corpus for the word and POS information, with English sentences parsed by (Collins 1999)'s head-driven parser and Chinese sentences by the head-driven parser of MI&TLAB at Harbin Institute of Technology (Cao 2006).

We define the literally translated sentence pairs as those that either embed enough word pairs which can be looked up in a bilingual dictionary, or share enough common grammatical categories. Hence, we invented two cross-lingual measures for the recognition of literal translation, i.e. lexical compatibility and grammatical compatibility.

2.2 Method of Lexical Compatibility

The seed version of our bilingual dictionary is made up of 63,483 entries drawn from the bilingual dictionary for the rule-based Chinese-English machine translation system of CEMT2K developed by MI&TLAB at Harbin Institute of Technology (Zhao 2001). We extended the seed with synonyms from English WordNet v. 1.2 and Chinese Extended Tongyicicilin v. 1.0. The extending algorithm is as follows.

Input: The seed version dictionary ***SD***, Chinese Extended Tongyicicilin ***CT***, and English WordNet ***EW***

Output: An extended Chinese English dictionary ***ED***

Do:

a. For each entry in ***SD***,
   a) extend the Chinese part with all its synonyms found in ***CT***;
   b) extend the English part with all its synonyms found in ***EW***;
   c) accept the extended entry into ***ED***.

b. For each entry in ***ED***,
   a) if its Chinese part is a subset of that of another entry, merge them;
   b) if its English part is a subset of that of another entry, merge them.

An entry in our final extended dictionary in turn is organized as bilingual synonym classes, and there are altogether 43,820 entries including 212,367 Chinese and English lexical terms.

By looking up Chinese-English word pairs in the extended dictionary, we defined the cross-lingual measure of lexical compatibility for a Chinese-English sentence pair as $\boldsymbol{C}_{\mathrm{L}}$.

$$C_L = \frac{the\ number\ of\ word\ \ pairs\ looked\ up}{the\ total\ number\ of\ all\ \ words}$$

For the recognition task, we employed a maximum likelihood estimation filtering method with an empirical threshold of 0.85 on the lexical compatibility. Sentence pairs would be accepted as literal translation if their lexical compatibility $\boldsymbol{C}_{\mathrm{L}} > 0.85$.

Manual analysis on 15,000 sentence pairs showed that for this method the precision is 94.65% and the recall is only 16.84%. The low recall is obviously due to the limitations of our bilingual dictionary.

2.3 Method of Grammatical Compatibility

Although the diversity of grammatical categories tends to be great, some common word classes, such as nouns, pronouns, verbs, adjectives, etc, mainly constitute the vocabularies of most natural languages. And our observations on English

and Chinese parallel corpora show that the more literal a translation is, the more equivalent grammatical categories the pair of sentences may share.

We thus define the cross-lingual measure of grammatical compatibility as C_G.

$$C_G = \sum_{i=1}^{n} \lambda_i \frac{Min(|GE_i|, |GC_i|) + 1}{Max(|GE_i|, |GC_i|) + 1}$$

GE_i is an English grammatical category, $|GE_i|$ is the number it occurs in the English sentence, and GC_i is the Chinese counterpart (see Table 1). n is the number of common grammatical categories that make differences in the special task of recognizing literal translated sentence pairs. λ_i is the weight for the respective category, which is trained by a simple gradient descent algorithm on a sample of 10,000 manually analysed sentence pairs.

i	Chinese	English
1	noun	noun
2	pronoun	pronoun
3	verb	verb
4	adjective	adjective and adverb

Table 1: Equivalent grammatical categories

For the recognition task, we also employed a maximum likelihood estimation filtering method with an empirical threshold of 0.82 on the grammatical compatibility. Sentence pairs would be accepted as literal translation if their grammatical compatibility $C_G > 0.82$.

Evaluation on the held-out sample of 5,000 sentence pairs shows a precision ratio of 89.5% and a recall ratio of 42.34%.

2.4 Combination of the Two Methods

We simply combined the results of the two methods mentioned above to obtain a larger useful corpus. It is very interesting that the intersection between the results of the two methods accounts only for a very small part, which is estimated to be 17.2% of all the identified sentence pairs. The combined recognition results achieved a precision of 92.33% and a recall of 54.78% on the testing sample of 15,000 sentence pairs. And on the total corpus, our combined method acquired 201,062 sentence pairs that were classified to be the results of literal translation.

Further analysis on the sampled corpus shows that the wrongly unrecalled literally translated sentence pairs and the wrongly recalled ones are mainly due to bad segmentation of Chinese words or bad POS tagging results of both the Chinese and English parsers. In contrast, those sentence pairs correctly unrecalled are usually free transcriptions or bad translations.

3 SMT Experiments

3.1 Our Corpus and SMT System

After excluding some too long sentence pairs, we got our final training corpus, which includes 200,000 Chinese-English sentence pairs of literal translation and 400,000 pairs of free translation[1]. Our evaluation corpus was drawn from the IWSLT Chinese-to-English MT test set of 2004, which includes 506 Chinese sentences and 16 English reference sentences for each Chinese one.

Since our focus is not on a specific SMT architecture, we use the off-the-shelf phrase-based decoder Pharaoh (Koehn 2004). Pharaoh implements a beam search decoder for phrase-based statistical models, and has the advantages of being freely available and widely used. The phrase bilingual lexicon is derived from the intersection of bi-directional IBM Model 4 alignments, obtained with GIZA++ (Och and Ney 2003). For better comparison between experimental results, we kept all the system parameters as default, while only tuning our own parameters.

3.2 Experiment on Incremental Training Corpora

This experiment was designed to check whether it is true that larger training corpora always lead to better SMT decoding performance. We randomly segmented the 400,000 free translation sentence pairs into 4 subsets, with each of them including 100,000 pairs. A baseline SMT model was trained with the 200,000 literal translation sentence pairs, and then 4 other SMT models were trained on extended corpora, of which each later used corpus includes one more subset than the previous one.

The decoding performances in terms of BLEU and NIST scores of all 5 models are listed in the second and third column of Table 2, and the last column gives the numbers of out-of-vocabulary (OOV) words of each model on the test set. Curves in Figure 1 and 2, respectively, show the trajectories of BLEU and NIST scores in accordance with the sizes of extended training corpora.

[1] Note that "free translations" are identified statistically using our recognition method for literal translations.

Corpus Size	BLEU	NIST	OOV
200,000	0.3835	7.0982	47
300,000	0.3695	6.9096	45
400,000	0.4113	7.1242	32
500,000	0.4194	7.1824	21
600,000	0.4138	7.1566	18

Table 2: SMT performance with extended corpora

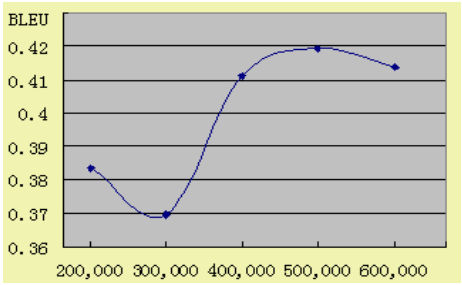

Figure 1: Trajectory of BLEU score

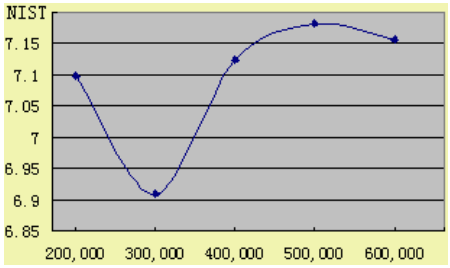

Figure 2: Trajectory of NIST score

A comparison between the different models' BLEU and NIST scores shows that a larger training data set does not necessarily lead to better SMT decoding performance. Based on the literal translation data, when more and more free translation data are added to the training set, the performance measures of the relevant SMT models fall at first, then rise, and at finally fall again. Furthermore, according to our manual analysis of the decoding results, free translation data have actually harmed the SMT model. It is just because the much smaller numbers of OOV words have made up for the impairment that the performance measures have risen for two times. They, however, will fall when the decrease in OOV words fails to make it up.

3.3 Experiment on Weighted Training Corpora

This experiment was designed to exploit both the contribution of literal translation and the advantage of a large vocabulary from a larger corpus. To achieve such a goal, minor modifications need to be made towards the training corpus and the module of GIZA++.

We start with an SMT training data set $\boldsymbol{X}$, which includes $\boldsymbol{n}$ bilingual sentence pairs, i.e. the input vector $\boldsymbol{X} = \{x_1, x_2, x_3, \ldots, x_i, \ldots, x_{n-1}, x_n\}$. During the original training process, every sentence pair x_i contributes in the same way to the estimation of parameters in the translation model since the corpus has not been weighted. Now we tried to adjust the contribution of x_i according to our previous decision whether it is literal translation or free translation. If we set the weight vector to be $\boldsymbol{W} = \{w_1, w_2, w_3, \ldots, w_i, \ldots, w_{n-1}, w_n\}^T$, the weighted corpus would become $\boldsymbol{X'} = \boldsymbol{WX} = \{w_1x_1, w_2 x_2, w_3x_3, \ldots, w_ix_i, \ldots, w_{n-1} x_{n-1}, w_n x_n\}$, where

$$w_i = \begin{cases} \lambda \text{ when } x_i \text{ is literal translation,} \\ 1-\lambda \text{ otherwise.} \end{cases}$$

Hereby λ is an empirical weighting parameter in the range of $0 <= \lambda <= 1$.

The module of GIZA++ was modified to ensure that the weights imposed on sentence pairs could be effectively transmitted to smaller translation units. GIZA++ builds word alignments by means of counting occurrences of word pairs in the training corpus. Given a possibly translatable Chinese-English word pair $\boldsymbol{D} = \langle c, e \rangle$, the number $\boldsymbol{N}$ of its occurrences in our original training corpus $\boldsymbol{X}$ can be calculated by summing up its occurrence number N_{xi} in each sentence pair, i.e.

$$N = \sum_{i=1}^{n} N_{xi}$$

Thus the weighted occurrence number $\boldsymbol{N'}$ of word pair $\boldsymbol{D}$ in the weighted training corpus can be calculated via the following equation.

$$N' = \sum_{i=1}^{n} N_{wi*xi} = \sum_{i=1}^{n} (w_i * N_{xi})$$

Finally, GIZA++ estimates word alignment parameters on the basis of $\boldsymbol{N'}$. Apart from this modification, all other parts of PHARAOH had been untouched to guarantee comparable experimental results.

We trained five SMT models of different weights on the previously mentioned corpora of free and literal translations. Table 3 lists both the training parameters and relevant decoding performances of the five models. Figures 3 and 4 show the trajectories of BLEU and NIST scores in accordance with the weight variable. We can see that the SMT model achieved the best performance when λ was set to be 0.67.

Corpus Size	λ	BLEU	NIST	OOV
400,000	0	0.4001	6.9082	23
600,000	0.5	0.4138	7.0796	18
600,000	**0.67**	**0.4259**	**7.2997**	**26**
600,000	0.8	0.4243	7.2706	39
200,000	1	0.3835	7.0982	47

Table 3: SMT performances with weighted corpora

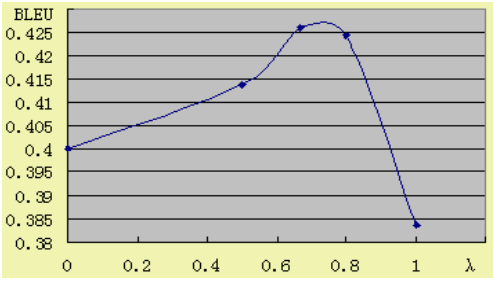

Figure 3: Trajectory of BLEU score

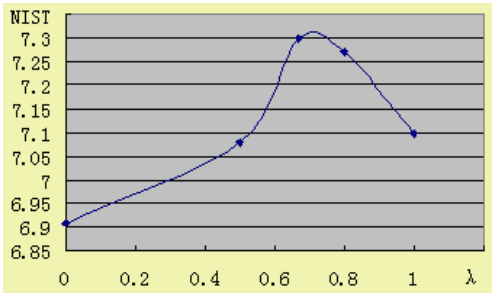

Figure 4: Trajectory of NIST score

Among the five models, that of $\lambda = 0.5$ is the baseline since here all sentence pairs contributed in the same way. Those of $\lambda = 0$ and 1 are two special cases designed to explore the isolated contribution of free and literal translation corpora in a contrastive way. Hereby the two models of $\lambda = 0.67$ and 0.8 are the central part of our experiment. According to the performance trajectories it seems that a reasonable increase in the contribution of the corpus of literal translations effectively improves the decoding performance of the SMT system since the BLEU scores with $\lambda = 0.67$ and 0.8 are higher than that of the baseline which are 0.0121 and 0.0105, and of the NIST scores which are 0.2201 and 0.191.

Our further analysis of the translation results and the related evaluation scores with different weight parameters showed that there exists some potential for literal translations to be used to improve SMT systems.

Our analysis indicates that two facts caused most of the out-of-vocabulary words (see Table 3). First, some OOV words never occurred in the training corpus; second, most others had been pruned off due to their much lower frequencies. Training corpora for $\lambda = 0.67$ and 0.8 have the same size of as that for $\lambda = 0.5$, but they resulted in much more OOV words than those for $\lambda = 0.5$ because the lower weight had decreased some related alignment probabilities very much. It seems that the large OOV increase must have counteracted the potential improvement to a certain degree although it did not have a devastating effects in these two cases. Therefore, a proper selection of a corpus of literal translations as training data would contribute more to the improvement of SMT models should some heuristic pruning methods be employed to avoid a possible OOV increase.

4 Related work

There have been a lot of studies on SMT training data. Most of them are focused on parallel data collections. Some work tried to acquire more parallel sentences from the web (Nie et al. 1999; Resnik and Smith 2003; Chen et al. 2004). Others extracted parallel sentences from comparable or non-parallel corpora (Munteanu and Marcu 2005, 2006). These works aim to collect more parallel training corpora, while our work aims to make better use of existing parallel corpora.

Some studies have also been conducted on parallel data selection and adaptation. Eck et al. (2005) proposed a method to select more informative sentences based on n-gram coverage. They used n-grams to estimate the importance of a sentence. The more previously unseen n-grams exist in the sentence, the more important the sentence is regarded. A TF-IDF weighting scheme was also tried in their method, but did not show improvements over n-grams. Their goal was to decrease the amount of training data to make SMT systems adaptable to small devices.

Some other works select training data according to domain information of the test set. Hildebrand et al. (2005) used an information retrieval method for translation model adaptation. They selected sentences similar to the test set from available in-of-domain and out-of-domain training data to form an adapted translation model. Lü et al. (2007) further used smaller adapted data to optimize the distribution of the whole training data. They took advantage both of larger data and adapted data.

Unlike all the above-mentioned studies, our method selected the training corpus according to basic theories of literal and free translation. This is somewhat similar to Lü et al. (2007), however, our weighting scheme also tried to make use of

both larger and smaller data, which are free translations and literal translations in our case.

Besides, there have also been some studies on language model adaptation in recent years, motivated by the fact hat large-scale monolingual corpora are easier to obtain than parallel corpora.. Examples are Zhao et al. (2004), Eck et al. (2004), Zhang et al. (2006) and Mauser et al. (2006). Since a language model is built for the target language in SMT, a one pass translation is usually needed to generate the n-best translation candidates in language model adaptation. The principle in our research could also be used for translation re-ranking to further improve SMT performance.

5 Conclusions

This paper presents a new method to improve statistical machine translation performance by making better use of the available parallel training corpora. We at first identified literally translated sentence pairs by means of lexical and grammatical compatibility, and then used these data to train SMT models. Experimental results show that literal and free translation corpora contribute differently to the training of SMT models. It seems that literal translation training data better suit SMT system at its present level of intelligence. The weighted training data can further improve translation performance by enlarging the contribution of literal translations while maintaining a larger vocabulary from the larger corpus of free translations. Detailed analysis shows that a literal translation corpus would contribute more to the improvement of SMT models if some heuristic pruning methods would be employed to avoid possible OOV increase.

In future work, we will improve our methods in several aspects. Currently, the recognition method for literal translations and the weighting schemes are very simple. It might work better by trying some supervised recognition techniques or using more complicated methods to determine the weights of sentence pairs with variant literal degree. What's more, our present test corpus is an out-of-domain one, and this might have impacted the observations made in this work. Last, employing our method to the language model might also improve translation performance.

Acknowledgments

We are obliged to the authors of English WordNet version 1.2 and Chinese Extended Tongyicicilin version 1.0 for the free dictionary resources they provided. We also thank the two reviewers for their constructive advices that we referred to when preparing the last version of this paper.

References

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. *Proceedings of EAMT* 2005: 133-142.

Arne Mauser, Richard Zens, Evgeny Matusov, Sasa Hasan, Hermann Ney. 2006. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. *Proceedings of International Workshop on Spoken Language Translation*: 103-110.

Bing Zhao, Matthias Eck, Stephan Vogel. 2004. Language Model Adaptation for Statistical Machine Translation with structured query models. *COLING-2004.*

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Comparable Corpora. *Computational Linguistics*, 31 (4): 477-504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Comparable Corpora. *ACL-2006*: 81-88.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-52.

Hailong Cao. 2006. *Research on Chinese Syntactic Parsing Based on Lexicalized Statistical Model*, Dissertation for PhD, Harbin Institute of Technology, Harbin.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, Richard Durand. 1999. Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. *SIGIR-1999*: 74-81.

Jisong Chen, Rowena Chau, Chung-Hsing Yeh. 2004. Discovering Parallel Text from the World Wide Web. *ACSW Frontiers 2004*: 157-161.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Language Model Adaptation for Statistical Machine Translation Based on Information Retrieval. *Proceedings of Fourth International Conference on Language Resources and Evaluation*: 327-330.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD Dissertation, University of Pennsylvania.

Mona Baker. 2000. *In Other Words: A Coursebook on Translaton*, Foreign Language Teaching and Research Press, Beijing.

Mildred L. Larson. 1984. *Meaning-based translation: A guide to cross-language equivalence*. Lanham, MD: University Press of America.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *In 6th Conference of the Association for*

Machine Translation in the Americas (AMTA), Washington, DC.

Philip Resnik and Noah A. Smith. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3): 349-380.

Tiejun Zhao. 2001. *Technical Reports for CEMT2K*. MI&TLAB, Harbin Institute of Technology, Harbin.

Yajuan Lü, Jin Huang and Qun Liu. 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 343-350.

Ying Zhang, Almut Silja Hildebrand, Stephan Vogel. 2006. Distributed Language Modeling for N-best List Re-ranking. *EMNLP-2006*: 216-223.