AN IMPROVEMENT OF TRANSLATION QUALITY WITH ADDING KEY-WORDS IN PARALLEL CORPUS

LIANG TIAN, FAI WONG, SAM CHAO

Department of Computer and Information Science, University of Macau, Macau S.A.R., China
E-MAIL: tianliang0115@yahoo.com.cn, derekfw@umac.mo, lidiasc@umac.mo

Abstract:

In this paper, we propose a new approach to improve the translation quality by adding the Key-Words of a sentence to the parallel corpus. The main idea of the approach is to find the key-words of sentences that cannot be properly translated by the model, and then put it or them in the training corpus in a separated line as a sentence. During our experiment, we use two statistical machine translation (SMT) systems, word-based SMT (ISI-rewrite) and phrase-based SMT (Moses), and a small parallel corpus (4,000 sentences) to check our assumption. To our glad, we get a better BLEU score than the original parallel text. It can improve about 6% in word-based SMT (isi-rewrite) and 4% in phrased-based SMT (Moses). At last we build a 120,000 English-Chinese parallel corpus in this way.

Keywords:

Key-Words; Parallel corpus; Statistical machine translation

1. Introduction

In some way, we can say progress in natural language research is driven by the availability of data. This is particularly true for the field of statistical machine translation, which thrives on the emergence of large quantities of parallel text: text paired with its translation into a second language [1]. In this paper, we just want to propose a new way to improve the translation quality by adding some key-words in a parallel text.

Many researchers have improved the quality of the statistical machine translation system with the use of word-based, phrase-based or tree-based translation models and so on. The groundbreaking IBM models [2], [3] represent the first generation of SMT models and it gives many common models. After that, researchers, for example Marcu and Wong [4], Koehn et al. [5] and Och [6], give the phrase-based models, which produce better translations than word-based models, which are wildly used models in the community of SMT. To add the syntactic information to the translation model, researchers [7], [8], [9], [10] propose tree-based transfer models [11].

In this paper, we do not care the models mentioned above. What we want to do is to deal with the parallel corpus that can finally give a better translation quality. The main idea is to organize the parallel corpus in a reasonable way, and which as a result can improve the translation performance without depending on the translation models.

The motivation to have the idea to build a parallel corpus like that is from our human's learning process. Suppose that there is a long sentence in front of us. Maybe we do not know the correct meaning of the sentence. But when somebody tells us one or more words of the sentence, most of the time, we can better understand it. It is similar to that there are some loopholes in our memory. We should learn the words or phrases to close the loophole. So, we also call our method to build the corpus as Close Loophole Method.

We know that the statistical machine translation system treats translation as a machine leaning problem. This means that we can apply a learning algorithm to a large body of previously translated text, known as parallel corpus, parallel text and so on [12]. In this paper we will not develop a new learning algorithm, what we will focus on is the content of the parallel corpus for constructing a translation model. We can add some key-words as closing the loophole of the parallel corpus to get a higher probability and give a better translation.

In our experiment, we choose two statistical machine translation systems: word-based SMT (isi-rewrite decoder, we call *System 1*) and phrase-based SMT (Moses, we call *System 2*) for investigation. To our glad, we achieve an improvement about 6% BLEU score higher in *System 1* and 4% higher in *System 2.* Both results are better than the outcomes of only using original parallel text.

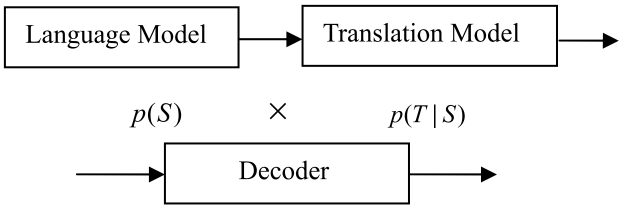
The remainder of this paper is organized as follows: principles of statistical machine translation are simply reviewed in section 2; Close Loophole Method to build the key-words parallel corpus is described in section 3. In section 4 we present experimental results and end this paper

978-1-4244-6525-5/10/$26.00 ©2010 IEEE

with conclusions and future work in section 5.

2. Principle of translation model

Statistical machine translation is just using the statistical methods to automatically translate one natural language to another using computers [3]. As Brown et al. illustrated in their statistical approach to machine translation [2], there are three problems in statistical machine translation. Figure 1 shows the three problems: 1) Computing language model probability $p(S)$; 2) computing translation model probability $p(T \mid S)$; and 3) searching possible source sentences S for the one that gives the best probability of $p(S)p(T \mid S)$. Here S is Source language and T is Target language.



$$S' = \arg\max_{S} p(S \mid T) = \arg\max_{S} p(S)p(T \mid S)$$

Figure 1. A statistical machine translation system

Actually speaking, the decoding process is the actual translation. Given a sentence T in the target language, the decoder chooses a viable translation by selecting the sentence S' in the source language for which the probability $p(S \mid T)$ is the maximum. Here we do not care how to compute the probability of $p(S)$ or $p(T \mid S)$. What we really care is how to get a higher probability of them from some change to the corpus. Later we will discuss this problem. To better show the translation process, you can refer to the translation process as shown in Figure 2 [13].During the translation, we should get the language and translation model first and then using the decoding algorithm to get the final translation result.

Now let's go to our model. In our model, we want to translate English to Chinese, and we choose the Chinese to be the language model S as talked above. We know that both language model and translation model probability are estimated from a large amount of data. For the language model probability, we need only Chinese text in our experiment, which can be gotten from many ways from the Internet. This time we use the same language model (140,000 Chinese corpus), so we do not pay much attention to the language model probability of $p(S)$. For the translation model probability, we need pairs of sentences, which we call parallel corpus. Actually speaking, during the computing process, we want to get a better alignment. In other words, we want to get a higher translation model probability $p(T \mid S)$. Of course, with the help of GIZA++ or Moses toolkit (or else), we can get the probability. According to our experiment, we think it is possible to get a better translation probability without adding too many parallel texts to the original parallel corpus.

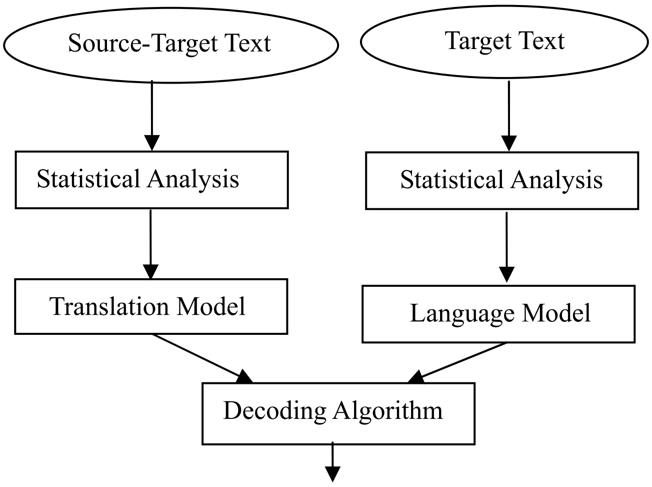


Figure 2. Translation process

As in section 3, we show the process of building the parallel corpus with key-words in our corpus.

3. Close Loophole Method to build Key-words corpus

The idea to use the Close Loophole Method to build the parallel corpus is from our human being's experience. Sometimes we could not understand a sentence very well. But after when someone tell us the meaning of some words or phrases, maybe we can understand the whole meaning of the sentence, at least better. There seems like that something lost in our mind. That's to say that our "brain database" has a loophole. We should learn or memory the lost words or phrases to close the loophole. This is the reason why we call the adding key-words to the corpus as a Close Loophole Method.

Consider the learning process. We first learn words and then the sentences and maybe we learn some grammar after that or at the same time. We should learn first and then understand. Most of the time, we can not understand a word or a sentence if we have not met before. Sometimes we may be not sure what the meaning of the word or a sentence is even we have met before. For example, "*The work shows* *the workman*.", all the words we can know , but we still cannot get the right Chinese translation. (A right translation of the sentence should be "什么匠人出什么活。".）

We want the machine to simulate this learning method. The difference is that we want the translation system to "learn" by short sentences. According to our experiment, translation system like Moses has a good translation quality when translate the short sentences (as shown in table 1). So we decide to have some short sentences as baseline data. In this way, we want the system to "learn" most of the words or phrases. We use the Close Loophole Method only to the long sentences that words more than 20 in a sentence or that we think it needed to mark (Proper names or idioms). Here you see that we just add some key-words to some sentences in the parallel corpus, which reduces the amount of work that building very large parallel corpus. According to our experiment, about 1000 parallel text with 100 key-words can exert some changes to the translation.

TABLE 1 SHORT SENTENCE TRANSLATION EXAMPLES IN MOSES

English Sentences
1. I 'm Kathy King. 2. It is January the 15th, 2010. 3. When I arrived at the station, the train had already left. 4. How about Mary? 5. I don't want to see any more of this TV show. 6. This is used for text GIZA++.
Translated Results in Moses
1.我是凯西·金。 2.今天是 2010 年 1 月 15 日。 3.我到车站时，火车已经开了 。 4.玛丽怎么样？ 5.我不想再看这个电视节目。 6.这是刚的检测在国际的 GIZA++。
Reference Translation Text
1.我是凯西·金。 2.今天是 2010 年 1 月 15 日。 3.我到车站时，火车已经开了。 4.玛丽怎么样？ 5.我不想再看这个电视节目了。 6.这是用来检测 GIZA++的。

In the following we will describe our idea. In table 2, there are three kinds of sentences that we have added some key-words to them. For the first Type, we want to deal with the words that polysemy phenomenon. In our corpus we have a short sentence that shows the common meaning of the "*inflection*" as "转折", in another sentence that shows another meaning as "音调变化". We add this meaning before the sentence. The best way to build this kind of parallel corpus is from the English-Chinese dictionary. We are seeking a good way to build the corpus using the *Oxford Advanced Learner's English-Chinese Dictionary, University English 4 and 6 dictionary and GRE vocabularies*. For the second type 2, we mainly want to deal with the proper names like school or corporation names and so on. We have collected many proper names in our corpus, which have mainly collected from the Internet. For the third type we want to deal with the idioms or phrases or proverbs. We know that these special sentences or phrases are not translated as the meaning of appearing on the surface, so we have to mark these sentences as key-words before the parallel sentences. In our corpus, we will not choose too many of this kind of sentences, since we think it will influence to the translation quality.

TABLE 2 SENTENCE TYPE TO USE CLOSE LOOPHOLE METHOD

Type	Example Sentences
1	We are in an inflection point. 我们正处于一个转折点。 voice inflection. 音调变化。 Correctly interprets non-verbal signs such as body language or voice inflection. 正确地解释身体语言或音调变化等非口语的信号。
2	Givenchy. "纪梵希"。 Elegance. "优雅"。 For several decades, Givenchy has kept its elegance style and become the pronoun of Elegance in the fashion world. 几十年来，"纪梵希"品牌一直保持着"优雅的风格"，在时装界几乎成了"优雅"的代名词。
3	so many people, so many minds. 仁者见人，智者见智。 just as the saying goes. 俗话说。 Just as the saying goes: "so many people, so many minds". 俗话说："仁智见人，智者见智"。

What we want to say is that we put the key-words before each sentence and lowercase the key-words at the same time. The reason why we do like that is that some word-based translation systems are sensible to case. They may know the meaning of “*Elegance*”, but without knowing the meaning of “*elegance*”. Although Philipp Koehn et al. [11] have proposed a new way to deal with this problem; we still think it is better to use the lowercase. Even in the phrase-based translation system as Moses, it suggests to use lowercase characters when in the training process [14], [15].

Another thing we want to point out is the tokenize problem of the key-words. For most key-words, we will not tokenize them or segmented. What we do is just treat the sentences or phrases as a whole. For example, we treat “*Givenchy*” as “纪梵希”, not “纪 梵 希” or something else; we should treat “*so many people, so many minds*.” as “仁者见人，智者见智。”，not“ 仁 者 见 人 ， 智 者见智 。” (Segmented in ICTCLAS). We can also say that we should align the Chinese as a whole to the English.

The following table (Table 3) is the corpus sentences we have built and a larger one will be finished soon about 2 million.

TABLE 3 PARALLEL CORPUS BUILDING

Sources	No.	Remarks
CET 4	15,000	*http://www.8844.43m.cn/index.asp* *http://dict.youdao.com*
CET 6	5,000	
GRE	20,000	
GMAT	6,000	
Proverb	6,000	*http://club.topsage.com/index.php*
Proper Names	30,000	*http://www.yesed.com/index-wordlist.htm* *http://www.fane.cn/*
New Concept English (1-4)	7,000	*http://club.topsage.com/index.php*
Others (News, art, Business…)	31,000	*http://www.i21st.cn/* *http://www.wwenglish.net/en/z/n/novel/* *http://www.cuyoo.com/*

Table 3 shows the situation of the parallel corpus (without the key-words). Here we use 46,000 example sentences to include four dictionaries (books) about 20,000 words. In this way we can see the most often used words. We also add some proper names, such as state or city names, corporation names, and academic vocabularies sentences. The other corpus is from *New Concept English* and *21 Century*，most of the sentences are news, we do not choose too many sentences from novels, here we just choose about 1000 sentences from novels(*Who Moved My Cheese, The Old Man and the Sea*), for words in a novel often consider context, which is not good for alignment.

4. Experiment Result

We present the result using parts of the corpus that we have built. This time we choose a small-scale corpus to conceptually illustrate its effectiveness. Here we choose 4,000 sentences as the training data and add 200, 400 and 600 key-words separately to check our idea. The translation text is 1130 parallel English-Chinese sentences, which 500 sentences are selected from New Concept English 3 and 4 and 400 short sentences are got from the training data we have used and the last 230 sentences are proverbs sentences. During the experiment, we first use the original parallel corpus and then the parallel corpus with adding key-words.

During the training process, table 4 shows the tools we used. After the decoding process, we get the translation result. To evaluate the result, we use an evaluation tool called: *multi-bleu.perl* in Moses packet. The result is shown in Table 5. From Table 4, we know that *System 1* is word-based translation model and *System 2* is the phrase-based model.

TABLE 4 DIFFERENT TRANSLATION TOOLS

Tools	LM	TM	Decoder
System 1	CMU-Cam	GIZA++, mkcls	Isi-rewrite
System 2	SRILM	GIZA++, mkcls	Moses

Note: LM is Language Model and TM is Translation Model

TABLE 5 THE BLEU [%] SCORE COMPARE

Test	BLEU 0	BLEU 1	BLEU 2	BLEU 3
System 1	5.95	11.97	11.93	11.49
System 2	55.60	59.63	60.93	59.93

From Table 5, we know that translation quality in phrase-based SMT is much better than the word-based one. After adding some key-words as we have shown above in section 3, the translation performance can be improved in a certain area. Here all the training data are tokenized or segmented sentences first. BLEU 0 is the score of the original sentences without key-words; BLEU 1 is the result of the sentences with 200 key-words; BLEU 2 is the score of the sentences with 400 key-words; BLEU 3 is the result of the sentences with 600 key-words. In *System 1*, when adding 200 key-words, there is a big improve (about 6%) to the translation quality , but after that, the BLEU score will be a little lower with 400 and 600 key-words, but still higher than the original one. Nearly the same case happens with the *System 2*. With 400 key-words added, the BLEU score improves about 4%. After adding 200 more key-words, the BLEU value begins to go down. This result shows that it is not the more key-words added to the corpus the better translation result will be. We should get the proper range. According to our experiment, we can add 1 key-word in 10(400/4000, here 400 is the key-words we added that can improve the translation quality, the 4000 is the actual sentences we deal with). Another enlightenment to us is that, maybe, there is no need to build very large corpus. We can find the balance point that gives the best translation result.

What we want to mention is that we should deal with the segmented Chinese text, because the segmentation is not always right; we should correct them manually. Even it is correctly segmented, we should change something. For example, sentence "这个孩子吃这么多是不正常的。" will be segmented into "这个 孩子 吃 这么 多 是 不 正常 的 。". Although the segmentation is all right, we do not want to segment the "不正常的" as "不 正常的", for the English sentence is "*It is abnormal for the boy to eat so much.*" ,we want the word "*abnormal*" to alignment to "不正常的". The word "不" and "正常的" should put together. Once we try to correct the segmentation by hand about 3,000 sentences, after that we use the *System 1* to get a better translation result, the BLEU score can be improved to about 15%. But we use the default segmentation by ICTCLAS, for we only want to check our idea this time.

5. Conclusion and future work

In this paper, we first analyze the principle of the statistical machine translation system, and then give our process of building a key-word parallel corpus. At last we use a small-scale parallel corpus to conceptually demonstrate our idea.

From the result we can know that it can improve the translation quality by using Close Loophole Method, especially to the word-based translation system. We are glad to the result as we expected. But we have to admit that we haven't found a way to add the key words automatically to the large corpus that we have built. So this time we just use a small corpus to prove our idea. The next step we want to do is to continue to expand the key-words parallel corpus to about 1 million parallel sentences. We will try our best to have an automatically way to add the key-words to the parallel corpus.

Another challenge to us is that we still think we can build a fit parallel corpus that can give us the best translation quality. Although we build very large parallel corpus wanting to get the best probability, sometimes it may be not. The highest probability may be not the best translation! In another experiment, we first used 1500 sentences to do the training process. Although the whole translation quality was not very good, some sentences especially those appeared in the training corpus were not bad. When we added the training data to 15,000, some translation result was not as good as with 1,500 sentences. In our translation sentences, there is a sentence "*I do not think the German Air Force has the numbers or quality to overpower our air defences.*". The following are the translation result from *System 2*, the result is shown on Table 6.

TABLE 6 TRANSLATION RESULTS WITH DIFFERENT NUMBER OF CORPUS

1,500 parallel texts: 我不认为德国空军在数量和质量上能击溃我们的空防。
15,000 parallel texts: 我不认为德国空军在数量和质量我们的空防。
Reference translation: 我不认为德国空军在数量和质量上能击溃我们的空防。

We can see that the system can not translate the "*overpower*" correctly, although the word appears more times in 15,000 sentences than in 1500 sentences. It seems that the system is "not sure" how to translate it. Of course it is because of the changes of the probability.

We can also have an intuitionistic to understand this idea. For example, "*right*" is often used in English, during the statistical process we can find that the world as the meaning of "右边" is much higher as the meaning of "彻底地，不折不扣地". When we give a sentence like "*Her son is a right little horror .*", we cannot translate as "她的儿子是个右小淘气 。". (In the word-based translation system *System 1*)

This is a hard job to find the balance point or how many corpus needed that can give the best translation result, but it is worth a try!

Acknowledgements

This work was partially supported by the Research Committee of University of Macau under grant CATIVO: 5868 and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2. I am grateful for some of my classmates before, who gave me a lot of help during the building of the key-words parallel corpus. Within about two weeks, they gave me about 70,000 parallel texts from different dictionaries or books. Without their help I can not finish the experiments.

References

[1] Philipp Koehn, Europarl : A Parallel Corpus For Statistical Machine Translation, MT Summit, 2005.

[2] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, A statistical approach to machine translation, Computational Linguistics, 16(2), pp. 79–85, Jun 1990.

[3] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R.L., The mathematics of statistical machine translation, Computational Linguistics, pp. 263–313, 1993.

[4] Daniel Marcu and William Wong, A phrase-based joint probability model for statistical machine translation, In Proc. of EMNLP, pp. 133–139, Jul 2002.

[5] Philipp Koehn, Franz Josef Och, and Daniel Marcu, Statistical phrase-based translation, In Proc. of HLT-NAACL, pp. 127–133, May. 2003.

[6] Franz Josef Och and Hermann Ney, The alignment template approach to machine translation, Computational Linguistics,30(4),pp. 417–449, Jun 2004.

[7] Wu D. , Stochastic inversion transduction grammars and bilingual parsing of parallel corpora, Computational Linguistics, 23(3), 1997.

[8] Alshawi, H., Bangalore, S., and Douglas, S.,Automatic acquisition of hierarchical transduction models for machine translation, In Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL), 1998.

[9] Yamada, K. and Knight, K., A syntax-based statistical translation model, In Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL), 2001.

[10] Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S.,Wang, W., and Thayer, I., Scalable inference and training of context-rich syntactic translation models, In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, pp. 961–968, 2006.

[11] Philipp Koehn, Hieu Hoang, Factored Translation Models, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pp. 868–876, June. 2007.

[12] Adam David Lopez, Doctor of Philosophy Paper, Machine Translation By Pattern Matching, University of Maryland, 2008.

[13] Philipp Koehn,Machine translation Lecture 9: Machine translation Decodeing , 2009.

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June. 2007.

[15] Philipp Koehn, A Beam-Search Decoder for Factored Phrased-based Statistical Machine Models User Manual and Code Guide, pp. 66-67, November 11, 2009.