Efficient Statistical Machine Translation Algorithm based on IBM Model 4

Kai Li Yuejie Zhang
School of Computer Science, *Fudan University*, *ShangHai*, *China 200433*
062021173@fudan.edu.cn, *yjzhang@fudan.edu.cn*

Abstract

This paper describes our methodologies for NTCIR-7 Patent Translation Task, and reports the official results based on English and Japanese corpus. Our system was a novel combination pattern of machine translation algorithms including classical statistical method -- IBM model and highly efficient decoding algorithm. The result of this new method is relatively decent, and its speed is also fast. It can be considered as a candidate for such situations as people who want to get a kind of quick and simple grasp of the main idea of a text.

Keywords: *Patent Translation, Statistical Machine Translation, IBM Model 4, NTCIR.*

1. Introduction

The rapid development of statistical methods in machine translation requires larger parallel or multilingual corpus. The NTCIR-7 (NII Test Collection for IR Systems) Patent Translation Task provides a Japanese-English patent parallel corpus which is aligned in sentence level. This is a very valuable resource for the research communities.

Our experimental work is based on this precious corpus. We aim at combining effective and efficient methodologies in the field of Statistical Machine Translation (SMT) into a practical system fully utilizing this patent corpus. The resulted system will be able to provide people a quick glance at the main idea of the text that they concern.

Although the translated text is not perfect compared to human experts' work, its advantage lies in the speed to finish the translation process.

In the following sections, we will first discuss the overall system structure, the algorithm applied to implement each part of it and the data set that were used to test its efficiency. In the end, we will analyze the reason why it behave like what it did and propose what can be done to improve its performance.

2. System description

2.1. SMT algorithm overview

The formal representation of the basic theory of Statistical Machine Translation is to maximize the conditional probability in the following Formula (1).

$$\hat{e} = \arg\max_{e} \Pr(e \mid j) \tag{1}$$

Here, j is the sentence in the source language of Japanese and e is the sentence in the target language of English, and $\Pr(e \mid j)$ represents the probability that e is the translation of j.

Applying the Bayes' Law to the above Formula (1), we have the following Formula (2).

$$\hat{e} = \arg\max_{e} \frac{\Pr(j \mid e) \cdot \Pr(e)}{\Pr(j)} \tag{2}$$

As the probability of the sentence in the source language of Japanese is the same for all the sentences in the target language of English, it does not have to appear. Thus, Formula (2) can be converted into Formula (3) as follows.

$$\hat{e} = \arg\max_{e} \Pr(j \mid e) \cdot \Pr(e) \tag{3}$$

Take word alignments a into consideration, a more accurate model is constructed as the following Formula (4).

$$\hat{e} = \arg\max_{e} \sum_{a} \Pr(j, a \mid e) \Pr(e) \tag{4}$$

where $\Pr(e)$ is the language model computing the likelihood of e, while $\Pr(j \mid e)$ is the translation model computing the likelihood of e can be translated into j.

2.2. IBM Model 4

To be able to do the actual computation, we have to determine the specific way to convert an English sentence into a Japanese sentence. Here, we choose the classic IBM Model 4 [1][2]. It has four procedures as follows.

(1) ASSIGN the number of words translated from each word in the English sentence e, thus is the fertility model in the following Formula (5).

$$\prod_{i=1}^{l} n(\varphi_i \mid e_i) \tag{5}$$

Here, l is the number of words in e. For each of the l words, it is repeated φ_i times. Those words with zero fertility will not appear in the new sentence to be translated.

(2) INSERT the *NULL* words at the proper positions in *j*, thus is the *NULL* generation model shown in Formula (6) as follows.

$$\binom{m-\varphi_0}{\varphi_0} p_1^{\varphi_0} (1-p_1)^{m-2\varphi_0} \prod_{k=1}^{\varphi_0} t(j_{0k} \mid NULL) \quad (6)$$

Here, we assume the fertility of a *NULL* English word is φ_0 and the probability it do fertilize is p_1.

(3) DETERMINE the probability of a word in *e* being translated into a word in *j*, as described in the lexical model of Formula (7).

$$\prod_{i=1}^{l} \prod_{k=1}^{\varphi_i} t(j_{ik} \mid e_i) \quad (7)$$

Here, *l* is the number of words in *e*. According to this model, only one or many word-mappings are allowed.

(4) PERMUTE *j* to get a more fluent sentence using the distortion model. For words appear first in the word string generated by an English word, the model gives the specific value in the following Formula (8).

$$\prod_{i=1,\varphi_i>0}^{l} d_1(\pi_{i1} - c_{a_i} \mid class(e_{a_i}), class(j_{i1})) \quad (8)$$

For words appear in the other places, the model above gives the specific value in the following Formula (9).

$$\prod_{i=1}^{l} \prod_{k=2}^{\varphi_i} d_{>1}(\pi_{ik} - \pi_{i(k-1)} \mid class(j_k)) \quad (9)$$

2.3. Decoding algorithm

It is known that decoding in machine translation is a *NP*-Complete problem [3]. The amount of combinations will explode with the number of candidate words increasing. The problem here is that too many nonsense series of the irrelevant words are taken into consideration. One intuitive improvement is to limit the candidate sentences to those that are most possible to be meaningful. Such sentences can be those sentences that are most likely to be real and their variants with a limited set of mutations. To achieve fast decoding, we employ the Greedy Decoding Algorithm [4][5][6], which includes the following six steps.

(1) INITIALIZATION. Construct an English sentence by aligning each Japanese word j_i with its most likely English translation e_i.

(2) CHANGE the translation of a word. For the English word e_i aligned with j_i, replace it with a new word e_i' if its fertility is one, or insert e_i' in the position that maximizes the alignment probability if its fertility is more than one. The candidates of e_i' are selected from the top words in the inverse translation model $\prod_{k=1}^{\varphi_i} t(e \mid j_i)$.

(3) INSERT an English word with zero-fertility which means it is not aligned with any Japanese word.

(4) DELETE a *NULL* word.

(5) JOIN two words. That is to delete one English word and align the Japanese words which were aligned with it to another English word.

(6) SWAP any substring pairs that don't overlap with each other. The number of all such pairs is computed by Formula (10) shown as follows.

$$\sum_{a=1}^{n-1} \sum_{b=a}^{n-1} \frac{(n-b)(n-b+1)}{2} = \frac{n(n^2-1)(n+2)}{24} \quad (10)$$

The complexity of decoding is reduced to quite a small percent of all possible combinations. That is the key reason that we achieve experimental result with the fast speed.

3. Experiments

3.1. Data description

The data used for the experiment is the Patent Corpus for NTCIR-7 Patent Translation Task. It consists of unexamined Japanese Patent Applications published in 1993-2002 and USPTO Patents published in the same period [7]. Applications/Patents published in 1993-2000 are used to create training data and the rest are made into testing data.

USPTO and Japanese Patents with the same priority number comprise "*Patent Families*" and their sentences aligned to build two data sets, that is, Parallel Sentence Data (PSD) and Parallel Patent Data (PPD). In the training data set, there are 1,801,312 sentence pairs from PSD and about 47,847 patent pairs which do no have sentence-alignment information from PPD. In the testing data for the formal run, there are 1,381 sentence pairs in the Japanese-English Intrinsic Subtask and in the English-Japanese Intrinsic Subtask. There are 224 patent topics for the Japanese-English Extrinsic Subtask.

3.2. Japanese segmentation

As many Asian languages, Japanese writing form leaves no space between words. So we must segment a sentence into the valid words before building the language model and the translation model. This process is mainly done by the Public Japanese Morphological Analysis System, that is, *ChaSen* [8].

3.3. Evaluation criteria

The criteria used to evaluate the translation quality is cased as the *BLEU* score [9].

BLEU is the metric which measures the *n*-grams co-occurrence between the automatically produced translations and the reference translations produced by human. It is defined in the following Formulae (11) and (12).

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n) \quad (11)$$

$$BP = \exp(\min(1 - \frac{L_{ref}}{L_{sys}}, 0)) \quad (12)$$

where N is the maximum number of n-gram's size, L_{sys} is the length of the automatically translated sentence, L_{ref} is the length of the shortest reference sentence, w_n is the weights of n-grams which sum to one, and p_n can be computed according the following Formula (13).

$$p_n = \frac{\sum_i \sum_{n-gram \in sys} count_{ref}(n-gram)}{\sum_i \sum_{n-gram \in sys} count(n-gram)} \quad (13)$$

The *BLEU* score favors the translations which consist of longer n-grams in the reference translation. It is positively correlated with the human experts' judgements for the adequacy and fluency of the translations, so the larger score represents the better translation quality.

3.4. Results

Firstly, the official evaluation results for Japanese-English translation of the formal run will be presented. The evaluation consists of the automatically intrinsic evaluation which mainly refers to the *BLEU* scores using single-reference from the original corpus and multi-reference from both the human experts and the corpus, and the extrinsic evaluation performed by the human experts which scores in a different way from *BLEU*. The intrinsic results are shown in Table 1 below.

Table 1. The intrinsic evaluation results for Japanese-English Patent Translation.

GROUP-ID	RUN	Adequacy	Fluency	Average
Fudan-MCandWI	1	1.75	2.42	2.08

Here, the "*Adequacy*" score represents how precise do the translated texts reflect the original meaning of the source texts, the "*Fluency*" score represents to what extent do the translated sentences meet the grammatical requirement of the target language, and the "*Average*" score is a half of the sum of the "*Adequacy*" score and the "*Fluency*" score.

The intrinsic evaluation is mainly for the purpose of research on the correlation between the *BLEU* scores and the human judgements.

According to the combinations of references used, there are two types of multi-reference *BLEU* scores, which are *m*600 and *m*300. Six hundred sentences, the *m*600, are randomly chosen from the 1,381 testing sentences for the formal run. Three experts (*A*, *B*, *C*) then translated them into English and another three (*D*, *E*, *F*) translated a half of them, that is, the *m*300. At the same time, the "*single*" (*S*) English sentences correspond to them in the corpus are used. Thus, a number of combinations of references lead to various evaluation results. The *BLEU* scores are shown in Table 2 below.

Table 2. The intrinsic evaluation results for Japanese-English Patent Translation by using Multi-Reference *BLEU* Scores.

GROUP-ID	RUN	Multi-Reference *BLEU* Scores	Low	High	All
Fudan-MCandWI	1	*single*	9	10.1	9.38
		*m*600-*ABC-S*	18.25	21.73	19.94
		*m*300-*DE*	19.01	21.4	20.27

Here, the "*Low*" score represents the smallest value among all the translated sentences' *BLEU* scores using different references set, while the "*High*" score represents the biggest value. The "*All*" score is the average score of all the *BLEU* scores of the sentences.

For English-Japanese Patent Translation, only references from the corpus are used because there are not human experts to provide additional references. It is the counterpart of the "*single*" *BLEU* score for Japanese-English Patent Translation. The related evaluation results are shown in Table 3 below.

Table 3. The intrinsic evaluation results for English-Japanese Patent Translation by using Multi-Reference *BLEU* Scores.

GROUP-ID	RUN	Low	High	All
Fudan-MCandWI	1	10.08	10.95	10.52

The above results can be combined in one figure, as shown in Figure 1. It can be seen that the *BLEU* scores of Japanese-English (JE) Patent Translation Task are correlated to each other and the score of English-Japanese (EJ) Patent Translation Task is a little better than that of JE task.

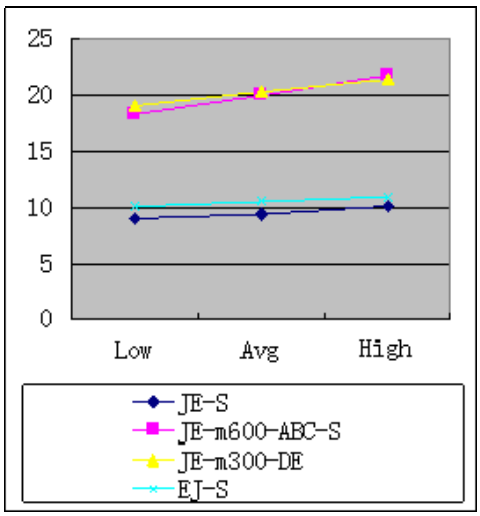

Figure 1. The evaluation results for the formal run, GROUP-ID = FDU-MCandMI, RUN = 1.

The time cost required to finish the translation process of the sentences in the testing data set of the formal run is about seventeen minutes. However, it is followed by a sacrifice in the result quality.

4. Conclusions

Compared to the scores of the other teams in NTCIR-7 Patent Translation Task [10], our results are not good enough in the aspects of the fluency and adequacy. But its speed is faster. We can conclude that the method used in our system to build the language model and translation model is obsolete. It does not take the phrasal language unit into account and leads to many mistakes in this aspect. Many idiomatic phrases are treated as the separate words, thus are translated into the meaningless word strings. Better system may take phrase as the basic units of the language model [11].

It is neither able to deal with the syntactic problem, the word is viewed as a word without linguistic attributes. So the complex relationship between the words, their attributes and the order of them is totally ignored. The future work must put much more emphasis on this problem. The Syntactic Analysis should be modeled into a more effective, efficient and satisfactory system.

The only advantage achieved is the speed. It resulted from the fast decoding pattern. Because a lot redundant manipulation can be deleted according to some statistically learned rules, the speed can be still further improved.

Acknowledgement

This paper is supported by National Natural Science Foundation of China (No. 60773124) and Shanghai Municipal R&D Foundation (No. 07dz15007). And the corresponding author of this paper is Yuejie Zhang from Fudan University.

References

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, Vol.19(2):263-311, 1993.

[2] Peter F. Brown, John Cocke, Stephen A. 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):79-85, 1990.

[3] Koehn Philipp. Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models. Machine Translation: from Real Users to Research, Association for Machine Translation in the Americas, Conference No.6, Washington DC, ETATS-UNIS, Vol.3265:115-124, 2004.

[4] Ulrich Germann. Greedy Decoding for Statistical Machine Translation in Almost Linear Time. In Proc. of HLT-NAACL 2003, Edmonton, Canada, 2003.

[5] Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. Fast Decoding and Optimal Decoding for Machine Translation. In Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01), pp.228-235, 2001.

[6] Franz Josef Och. An Efficient Method for Determining Bilingual Word Classes. In Proc. of EACL'99, Bergen, Norway, June 1999.

[7] Masao Utiyama, and Hitoshi Isahara. A Japanese-English Patent Parallel Corpus. MT summit XI, pp.475-482, 2007.

[8] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. Japanese Morphological Analysis System ChaSen Version 2.3.3 Manual. 2003.

[9] Papineni K., Roukos S., Ward T., and Zhu W. J. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp.311-318, 2002.

[10] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In Proc. of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. 2008.

[11] David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, pp.263-270, 2005.