Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):1628-1635

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Statistical machine translation based on translation rules

Hu Yulian

Foreign Language Department of Heze University, Heze, Shandong, China

ABSTRACT

Nowadays statistical machine translation shows its benefits and has received much attention. In this paper, phrase-based statistical machine translation was carefully studied. Improved Hidden Markov Model(HMM) was used to align words and solve the inconsistency between word alignment and phrase structures, and can serve word alignment better. Translation rules were extracted based on aligned phrases and English phrase trees. CYK+, an improved CYK algorithm, as adopted as the decoder to decode non-Chomsky translation rules; Two-round-decoding algorithm was proposed to integrate the language model during decoding. The experiment results showed the BLEU score of improved HMM was higher than the score of HMM, so it follows that the translation system based on translation rules has more stable translation effect on different data collection.

Keywords: statistical machine translation, improved Hidden Markov Model; translation rules, CYK+ algorithm, BLEU score

INTRODUCTION

Machine translation is to translate one natural language into another by computers. It can be viewed as a decision problem from the perspective of artificial intelligence, that is, every sentence of source language is translated according to translation knowledge[1]. Today, machine translation is still one of the most difficult decision problems because of the complexity of natural language. So the processing of natural language itself is quite a complicated question. It is even more difficult for machine to translate because the computer will process two or more languages at the same time. As is known to all, many languages have great differences in grammar rules and expressions. While people translate one language into another, they usually consider the grammar, semantics, and contexts of the language. Similarly, linguistic knowledge that machine translation uses can also be divided into different levels, which is showed in machine translation pyramid (see figure1), which reveals the complete process of machine translation through the machine system.

Statistical machine translation was proposed by Peter F. Brown et al in 1990[2-3], which has been a hot research topic in machine translation field today[4-5]. To some extent, statistical machine translation is to get the translation by calculation while adding probability to rules or examples rather than by direct judgment. This paper studies statistical machine translation based on translation rules which mainly involve phrase-based model. The research in this paper will be important for people to understand the feature of statistical machine translation because it tempts to explore some new ways on the basis of traditional translation method, whose result can be directly applied in machine translation practice, so its application value and practical significance are very clear.

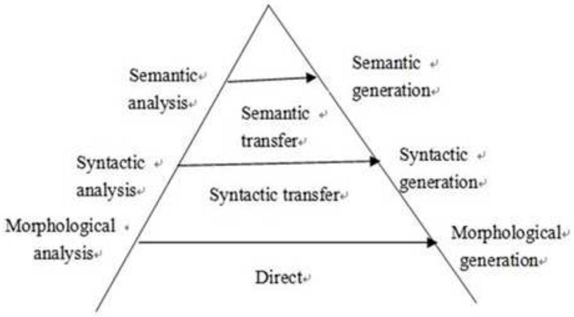

Figure1: Machine translation pyramid

lITERATURE REVIEW ON STATISTICAL MACHINE TRANSLATION

2.1 *The origin of statistical machine translation*

Statistical machine translation is also called data-driven translation in which leaning technology is automatically introduced. While the traditional rule-based machine translation uses rules to express translation knowledge, statistical machine translation does so with a set of model parameters. In 1949, W. Weaver suggested that statistical method and information theory should be used to study machine translation[6].In his memorandum he said, "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text". From his words, we can judge that it is W. Weaver who firstly put forward the idea of machine translation by decoding, which became the origin of noise channel theory. But statistical machine translation were not wide attentive until some researchers in the IBM T.J. Watson research center came up with the statistical machine translation based on source channel and successfully put it into translation practice.

2.2. *Framework of statistical machine translation*

Framework of statistical machine translation includes translation based on source channel, probabilistic parallel grammar, and maximum entropy models[7]. From the perspective of linguistic knowledge, the framework can also be classified into word-based, phrase-based, and syntax based models.

Brown from the IBM T.J. Watson research center first came up with the statistical machine translation based on source channel model (also called noisy channel model(see fig. 2). So such process of translation is divided into three sub-problems such as modeling of language model and translation model and decoding. The framework of translation based on source channel is showed in figure3.

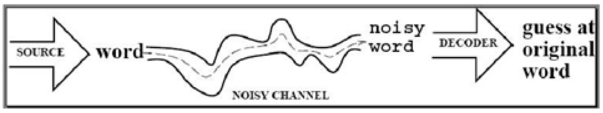

Figure2: Noisy channel

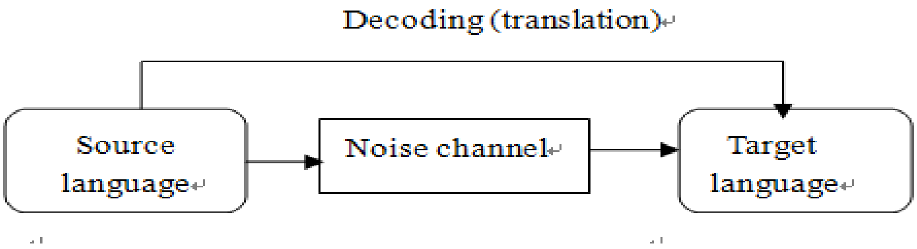

Figure 3: Source channel model for statistical machine translation

2.3 Statistical translation models

2.3.1Word-based statistical translation model

Based on word alignment, Brown et al (1993) built up 5 translation models which were called IBM Models 1-5[8]， whose complexity increases in turn. These models showed different plans in translation probability calculation when a source language sentence was turned into a target language sentence.

Similar to IBM2, Vogel improved and put forward 2 models [9] one of which is HMM-based alignment model. Compared with IBM2, alignment from HMM is more smooth and better. It is word-based statistical translation that opens up the study of statistical machine translation. It promotes the research on decoding, especially the application of every general algorithm. In addition, it also can be used in phrase-based and syntax-based translation models.

2.3.2 Phrase-based statistical translation model

It can be divided into the word-based model, which extracts phrase translation pairs based on word- alignment, and the non-word-based model, which extracts phrases based on other information.

By systematically comparing IBM model with HMM model, Och realized word alignment Giza++[10]. Heidi J Fox said that in Giza++, word alignment and sentence structure have more possibility to conflict, and large amount of wrong word alignment exists [11]. Och et al adopted alignment template technology to solve the problem of data sparseness[12] and used maximum entropy model to integrate various language characteristics and statistical information into statistical machine translation[13]. Based on monotonous phrase model, Koehn came up with phrase translation model based on word alignment[14-15]. This model also uses linear logarithmic model to be its framework, and has become the baseline of comparative trials in today's statistical machine translation. Al-Onaizan used word alignment and BLEU score to measure the similarity of word orders between the two languages[16]. He defined the outbound distortion, inbound distortion and pairwise distortion.

2.3.3 Syntax-based statistical translation model

The study of this model can be traced back to Inversion Transduction Grammar (ITG). It was firstly pointed out by Wu Dekai[17-18], and then was applied to construct Chinese-English translation system.

Yamada et al came up with syntax-based statistical translation model[19]. Different from the above direct translation model, this one is in fact a tree-to-string model, in which the input of the source channel is a syntax tree, and the output is a sentence. Yamada and Knight put forward the tree-to-string translation model based on syntax in the true sense[20].

Chiang put forward the Hierarchical Phrase-based Model[21]which can process discontinuous phrases. This model borrowed the structure of formal grammar and used CYK syntactic parser of beam search. Chiang realized complete compatibility of bilingual phrases in 2005. Recently, syntax-based translation research has further improvement, mainly including directly introduction of the syntax information in phrase approach and extracting phrases according to syntax structure.

2.4 Model training and decoding algorithm research

Liang came up with a judgment leaning method based on sensors, whose advantage was it could process characteristic set in a large scale and be used in every phase of decoding[22]. Tillman raised a global discrimination learning approach which regarded decoding as a dark box and could be used in any decoding algorithm[23]. Fraser brought forward the semi-supervised learning approaches which aimed at IBM 4. Its core was EMD algorithm[24]. Decoding algorithm is one of the important parts of statistical translation. The function of decoding directly affects the quality and effectiveness of the translation.

From the above analysis, we know that the translation effect of word-based model is not so well as phrase-based model because purely word-based translation model cannot make full use of the contexts. So it is with syntax-based translation model without introducing the phrase template. Nowadays it is still phrase-based model that gains the highest scores in evaluating which is the mainstream method in statistical machine translation because of its simple and good translation quality.

Still, the quality of phrase alignment depends on word alignment. IBM 1 to IBM5 and HMM may lead to the conflict between word alignment and syntactic constraints. In this paper, the distance of phrase structure tree is used to overcome the conflict based on HMM to improve the word alignment quality. String-tree model has a strong ability to process syntactic feature, but a weak ability to deal with lexical feature; phrase model and hierarchical phrase model have a strong ability to process lexical feature, but a weak ability to process syntactic feature. Based on hierarchical phrase model, this paper extracts translation rules by using phrase structure tree of the target language, making full use of lexical feature and syntactic feature. CYK algorithm cannot decode because translation rules may be non-Chomsky. This paper adopts CYK+ (the improved CYK algorithm) as decoding.

PROCESS OF STATISTICAL MACHINE TRANSLATION SYSTEM

3.1 Preprocessing

It mainly includes the automatic words segmentation of Chinese sentences, case conversion of English sentences, filtering of long bilingual sentences (Chinese and English). The adopted software of automatic words segmentation is Stanford segmentation tool in 2008 version, and the part-of-speech tagging set used here is Beijing University version.

3.2 Word alignment

Word alignment is adopted on the basis of GIZA++ and improved HMM model whose input is Bilingual parallel corpora and output is the Viterbi alignment of a bilingual sentence pair. GIZA++ algorithm will follow the training sequence in the experiment done by Och. That is, first uses IBM model 1 and 2 to have training, then uses HMM and IBM model 4.

3.3 Alignment phrase extraction

The extended word alignment result is used to have alignment phrase extraction, whose input is the alignment result of bilingual and two-way words and the output is the lists of phrase inter-translation and the seat of a phrase in a sentence. Moses system2 is used here to align phrases[25].

3.4 Translation rules extraction

The input is the phrase alignment list, and the output is translation rules with syntactic tags and the responding characteristic function value. The parser is Stanford Parser in version 2007, and the tagging set is Penn tree-bank tagging set.

3.5 Translation rules filtering

Translation rules are filtered according to constraint condition so as to accelerate the decoding process. The input is translation rules and testing set, and the output is the translation rules that only aim at the testing set.

3.6 Minimum error rate training

Chinese-English reference set is used to adjust the weight of every characteristic function in translation models so as to have maximum entropy model training. The minimum error rate training adopts the algorithm studied by Och. The input of this module is the reference set, the language model in English, the translation rule set, and the decoder’s first K-Best translation result list. The output is the weight of every characteristic function.

3.7 Decoding

CYK+ is adopted as decoding algorithm, and meanwhile two-round decoding algorithm is used to integrate language model in the course of decoding. The input of this model is Chinese sentences for translation, and the output is 1-best or K-best English translations of these Chinese sentences. The language model tool used in the following experiment is Srilm version 1.5.5[26]. In addition, Web-1TB trigram language corpora (LDC for free) is used.

3.8 Automatic evaluating

The automatic evaluating tool of machine translation used here is mt-evaluation1.1 version of NIST3 and BLEU-4 meta-language model evaluation .

IMPROVED HIDDEN MARKOV MODEL

Improved HMM can improve the alignment of basic HMM, which makes alignment probability connect to the alignment between two source language strings. It happens at the string distance and phrase structure tree distance between two alignment positions in the target language string formula.

$$p(a_j \mid a_{j-1}, I) = \lambda_1 \frac{c(i-k)}{\sum_{l=1}^{I} c(l-k)} + \lambda_2 \frac{t(i,k)}{\sum_{M=1}^{I} t(m,k)} \qquad (1)$$

In (1), i=aj represents the alignment of the j source language word and the i target language word. Similarly, k= aj-1 means the alignment of the j-1 source language word and the k target language word. c(i- k) means the string distance of 2 target language words of the alignment of two source language words. The definition and operation of *c*(*i*-*k*) is similar to basic HMM. $\lambda_1 + \lambda_2 = 1$. *t*(*i, k*) stands for the distance of the target language words in the phrase structure tree of target language. The distance is based on the alignment of the two source language words *j*-1 and *j*. The denominators are all normalized factors.

TRANSLATION RULES EXTRACTION

5.1 Using alignment phrases to construct basic translation rules

The basic translation rules consist of syntactic markers, source language word strings, and target language word strings. For the alignment phrases we input, we search the syntactic markers of target language that is corresponding to the alignment phrases. If successfully searching these markers, the alignment phrases can be extended to be basic translation rules. If there are no corresponding syntactic markers on the target side of this alignment phrase, the following strategies can be used to seek extended syntactic markers for it.

Firstly, judge whether the target phrase corresponds with two or more syntactic markers. If so, all syntactic markers of sub-phrases can be merged as the syntactic markers of this phrase.

Secondly, try to use C1/C2 marker to align phrases. C1/C2 is an incomplete syntactic component on the right side, and its corresponding complete one is C1 which can only be formed by combining C1/C2 with C2 on the right side.

Thirdly, try to use C2\C1 marker to align phrases. C2\C1 is an incomplete syntactic component on the left side, and its complete syntactic component is C1 which can be formed by combining C2 \ C1with C2 on the left side.

5.2 Using basic translation rules to build combinational translation rules

Here we use derivation principle by Chiang[8] to obtain the combinational translation rules.

N→f1⋯fm/e1⋯en (2)

M→ f i⋯f u/ej⋯e v (3)

Two basic translation rules are showed in (2) and (3). If i⩾1,u⩽m, j⩾1, v⩽n, and the 4 equations do not equal at the same time, then the combinational translation rules can be deduced by (2) and (3): N→f1 ⋯fi-1 M k fu+1 ⋯fm /e1⋯ej-1 Mk ev+1 ⋯en..

A rectangle can be defined by using the initial position and ending position of source language and target language, indicating the span of bilingual strings in the translation rules.

CYK+ ALGORITHM

6.1 CYK and CYK+

Chiang's hierarchical phrase model uses the CYK(Cocke-Younger-Kasami)-based decoder. CYK algorithm demands that rewriting rules belong to Chomsky model, but the translation rules produced in this paper don't completely belong to Chomsky model. The general solution is to convert translation rules into Chomsky model, but this conversion will lead to the increasing number of nonterminal characters in translation rule sets. Such increasing amount of nonterminal characters has seriously cut down the computing speed of the decoder. To solve this problem, we use CYK+, the improved CYK algorithm.

Based on CYK algorithm, CYK+ builds the present two-dimensional matrix Chart [i][j] and make them contain the non-terminal character's collections of all the phrases that are likely formed. Meanwhile, we also make Chart [i][j] contain all the incomplete hypothesis(dotted rule) according to Earley algorithm. This improvement can make CYK+ algorithm process non-Chomsky translation rules. The data structure of CYK+ algorithm is two-dimensional matrix {Chart [i][j]}. Without considering the language model, we suppose the input string of the decoder is w^n_1, every Chart [i] [j] in the Chart corresponds to the nonterminal character's collections of all the possibly formed phrases in some span of the input sentence and all incomplete hypothesis, in which *i* indicates the position of first word on the left of the span, j indicates the number of the words contained in the span.

6.1.1 The process of initialization

As for all the position *i* and span *j* that are input in the source language, if the translation rule X→wi⋯wi+j-1 exists in the source language, then add X to the first list of Chart [i] [j].

6.1.2 The process of matching

For the given Chart [i] [j], carry out the following two operations:

Firstly, for all possible Chart[i][k] and Chart [i+k][j- k], if an incomplete hypothesis α exists in Chart[i][k] and a complete hypothesis β in Chart [i+k][j-k], together with α & β satisfying the translation rules A→ α β γ, if γ is empty, then add nonterminal character A to the first list of Chart [i] [j]; if γ is not empty, add nonterminal characters αβin the second list of Chart [i] [j]. Secondly, For Chart [i][j], for each nonterminal character in the first list, if A→ β γ exists and γ is empty, then add nonterminal character β in the first list of Chart [i] [j]; if γ is not empty, add nonterminal character β in the second list of Chart [i] [j]. Thirdly, Repeat step (2), until Chart[1][n] is finished structuring.

We can dynamically program algorithm to build the above two-dimensional matrix Chart in the time of multiple items.

CYK+ and CYK have the identical time complexity: $O(n^3)$, n is the sentence length. CYK+ can process non-Chomsky translation rules, but CYK algorithm can't.

TWO-ROUND DECODING ALGORITHM

In the decoding algorithm, the state of every hypothesis in the course of decoding does not consider the state of the target language model corresponding to the current hypothesis. The reason is that the condition of equivalence hypothesis on which CYK+ algorithm depends is false under the premise of considering the language model. Even though the source language string of the 2 hypotheses and the syntactic markers of the target language are consistent, the two hypotheses that correspond to the target language model can be possibly inconsistent, so the score of the two hypotheses cannot be calculated. Though we can calculate without introducing language model, target language model plays the very important role in improving translation quality. So the integration of language model is a critical question for the decoder to solve in statistical machine translation based on syntax.

In this paper, two-round algorithm is used. In the first round, we add the approximate language mode score that can be quickly calculated so as to quickly makeup the corresponding Chart list and find the approximate optimal K-Best translation result; In the second round, we reversely search this chart list according to K-Best translation result in the first round. The considerations of such algorithm are that in the first round we can gain the approximately correct derivation sequence of K-Best based on translation rules and the approximate target language model. We find that mistakes in the translation result mainly appear in the target language model, so in the second round algorithm, the revision is done based on the language model score, and the final result should be similar to the result that completely uses target language model in the course of decoding.

EXPERIMENT AND RESULT ANALYSIS

The experiment adopts the bilingual parallel training corpora. The whole training set includes 500,000 parallel bilingual sentence pairs，in which the average length of Chinese sentence is 15.01 and the average length of English sentence is 13.84. The testing corpora used to calculate the BLEU score is another specially prepared 500 sentence pairs of Chinese-English translation; On the other hand, the testing corpora used to calculate word alignment quality is 500 sentence pairs of Chinese- English translation which is the alignment result of manually tagged words.

The main translation reference system here is phrase-based translation system Moses by Kohen and Chiang's hierarchical phrase-based translation system.

8.1 The word alignment quality of improved HMM

In the experiment, two kinds of word alignment modules are adopted, one is Giza++ , which realizes the alignment of IBM model 4 and basic HMM; the other is improved HMM word alignment module, which has the same input and output format as Giza++ module. The input is bilingual parallel corpora, the output is the two-way optimal word alignment result with Giza++ format. Table 1 shows the effects of the 3 word alignment results on the translation system. From table 1, we can see that BLEU score of basic HMM is higher than that of IBM model 4 in training set and testing set, which means that the word alignment quality of basic HMM is higher than that of IBM model 4. BLEU score of improved HMM is higher than that of IBM model 4 and basic HMM(the score can be increased by about 0.5-1 points), which means that the word alignment quality of improved HMM is higher than IBM model 4 and basic HMM.

Table1: Effects of 3 word alignment models on translation system

	BLEU score	Training set	Testing set	mean
1	IBM model 4	26.05	25.62	25.84
2	Basic HMM	26.44	25.89	26.17
3	Improved HMM	26.92	26.52	26.72

8.2 Effects of translation rules on statistical machine translation

The first aim of translation rules is to tackle the global phrase sorting, while phrase models can only solve the local sorting. Table 2 shows the comparison of machine translation quality based on phrase model, hierarchical phrase model and translation rules. As far as phrase-based translation mode is concerned, its greatest length of phrase sorting is 12(reo=12). The instruments of this experiment include the whole tri-meta model in Web1 T and the language models in the training corpora. From table2, we can see that BLEU score of hierarchical phrase model is higher than that of phrase model in training set, testing set and in means, which illustrates that the machine translation quality of hierarchical phrase model is higher than that of phrase model. BLEU score of improved hierarchical phrase model is higher than that of phrase model and hierarchical phrase model in training set, testing set, and in means, which shows that the machine translation quality of the improved hierarchical phrase model is higher than that of phrase model or hierarchical phrase model.

Table2: Blue score comparison among 3 models

	BLEU score	Training set	Testing set	mean
1	Phrase model (reo=12)	26.39	25.96	26.18
2	Hierarchical phrase model	26.78	26.13	26.46
3	improved hierarchical phrase model	26.92	26.52	26.72

8.3 BLEU effects of two-round decoding algorithm on the improved hierarchical phrase model

Table 3 shows the comparison of BLEU score between cubic pruning algorithm and the two-round decoding algorithm. As for the improved hierarchical phrase model, BLEU score of cubic pruning algorithm parallels that of two-round decoding algorithm, which shows that the translation quality of the two algorithms is similar, but the decoding speed of two-round algorithm is faster than cubic pruning.

Table3: Comparison of the two algorithms

	BLEU score	Reference set	Testing set	mean
1	Cubic pruning algorithm	26.92	26.52	26.72
2	Two-round decoding algorithm	26.89	26.51	26.70

CONCLUSION

This paper realizes a more complete statistical machine translation system by integrating the advantages of string-tree model and hierarchical phrase model. In the practice we find that word alignment result and syntactic constraint conflict have influenced the quality of translation rules and the property of translation system. So firstly this paper uses HMM which is based on phrase structure tree distance in word alignment because it can effectively reduce the conflict between word alignment result and syntactic constraint so as to improve the property of syntax-based translation system. Secondly, this paper introduces the hierarchical phrase model based on phrase structure tree, which is using the thought of string-tree model to improve Chiang's hierarchical phrase model. Based on the bilingual alignment phrase, this paper extracts the translation rules by integrating English phrase structure tree, and puts forward the heuristic rules to get the improved syntactic markers of the translation rules. Compared with hierarchical phrase model, this improved model has more stable property in different data set and has better evaluation results. In addition, the paper provides not only CYK+ algorithm that can process the non-Chomsky translation rules in the course of decoding, but also two-round algorithm to integrate language model. The algorithm has the similar decoding quality as cubic pruning algorithm, but the former has a faster decoding speed. Last but not least, the proposals given in this paper can be more effective in obtaining translation knowledge and improving translation quality by machine, which will be very helpful for machine translation research and will bring many conveniences for people.

REFERENCES

[1] F. J. Och, Statistical Machine Translation: From Single-Word Models to Alignment Templates, Ph.D. Thesis, RWTH-Aachen, **2002**,1-122.

[2] Yang G, He Q, Deng X. *International Journal of Advancements in Computing Technology*, **2011**, 3(10).pp. 266−273.

[3] P. F. Brown; J. Cocke; S. A. Della Pietra; V. J. Della Pietra; F. Jelinek; J. D. Lafferty, A Statistical Approach To Machine Translation, Computational Linguistics, **1990**,16(2),79-85.

[4] Jinkui Hou, Lei Wang, JNW, Formal Description for Component-based Architecture Model Transformation, **2013**, 8(4),874-881.

[5] Fucheng You; Ying Zhang, JMM ,Research of an Improved Wavelet Threshold Denoising Method for Transformer Partial Discharge Signal, **2013**,8(1),56-63.

[6] YANG Gelan, Yue WU, and Huixia JIN. *Journal of Computational Information Systems*, **2012**, 8(10): 4315-4322.

[7] Liu Qun, Literature review on statistical machine translation, Journal of Chinese information, **2003**,17(4),1-12.

[8] Yang G., Le D., Jin Y., Cao S. Q. *International Journal of Distributed Sensor Networks*, **2014**, vol. 2014, Article ID 363584, doi:10.1155/2014/363584.

[9] Gelan Yang, Su-Qun Cao, Yue Wu. *Mathematical Problems in Engineering*, **2014**,vol. 2014, Article ID 549024, 4 pages, doi:10.1155/2014/549024.