NLPR Translation System for IWSLT 2006 Evaluation Campaign

Chunguang Chai, Jinhua Du, Wei Wei, Peng Liu
Keyan Zhou, Yanqing He, Chengqing Zong

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100080, China
cgchai@nlpr.ia.ac.cn

Abstract

In this paper we describe a hybrid approach to Chinese-to-English spoken language translation system used for the IWSLT 2006 evaluation campaign. In this system, the phrase-based statistical machine translation (SMT) engine is combined with the template-based machine translation (TBMT) engine and a simple way is proposed to select the best translation from the results generated by the two translation engines. The experiments prove that the combination can improve the performance of translation system. As the input sentences are speech recognition results and have no punctuation information, we restore the punctuation in source sentences in the processing and post-processing.

1. Introduction

In this paper, we describe a hybrid approach to Chinese-to-English spoken language translation system that combines the TBMT system with the phrase-based SMT system. We use a simple method to select the best translation from the results generated by the two translation system.

The template-based translation system is easy to realize and can generate correct translation result if the input sequence is matched properly with a template. So the template-based translator is often employed as one of the translation engines in machine translation systems. However, the template-based method has its weakness. For example, the extraction of templates is difficult and the coverage of the templates is small. In addition, the templates are generally fixed and the flexibility to match the inputs and to generate the target language is often limited. In our system, we use an improved template-based approach which is robust to some extent for the spoken Chinese language translation as described in [1].

In recent years, statistical machine translation method is becoming more and more popular. In the early 90s, IBM developed the candidate system. Since then, many statistical machine translation systems were proposed [2] [3]. These systems are based on the source channel model and apply a translation model to capture the relationship between the source language and the target language, and use a language model to drive the search process. The primary IBM model was purely word-based model and one improvement is phrase-based statistical translation model. The phrase-based SMT model is proposed to incorporate more complex structure and to get better lexical choice and more reliable local reordering. There are many researchers using the phrase-based translation method to improve their systems performance [4][5][6]. In our system, we apply a phrase-based translation model.

This paper is organized as follows: section 2 describes a hybrid approach to Chinese-to-English translation system, which combines the TBMT system and phrase-based SMT system. Section 3 presents the improvements of base system for the evaluation. In section 4, we present a series of experiments of Chinese-to-English translation and the results are analyzed. Section 5 gives the conclusion.

2. System Description

Our system combines two translation engines: the template-based translation engine and phrase-based statistical translation engine. We use a simple method to select the best translation from the results generated by the two translators. Figure 1 shows the architecture of our system. In this following, we will give an overview of the two translation models and introduce the selection method.

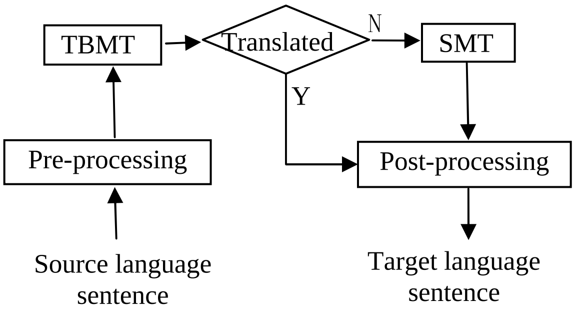

Figure 1: System architecture

2.1. Template-based translation method

The template-based translation use the templates defined beforehand to translate source language sentence into target language sentence. If the input sentence is matched properly with a template, the system can generate a correct translation.

The template is designed as:

$$C_1 C_2 \cdots C_n \Rightarrow T \qquad (1)$$

where n is an integer ($n \geq i \geq 1$), and C_i is a component source language has to meet. A component may be expressed as the following five types:

1) **Key words:** a fixed Chinese word such as “房间(room)”, “有没有(whether to have)”.

2) **Parts-of-speech and semantic features:** a part-of-speech, phrase name, a part-of-speech with semantic features or a phrase name with semantic features.

3) **Variable:** it means that any word or phrase may appear at the position, or nothing appears.

4) **Logical expression of candidate components:** for example, ‘C_1 | C_2’ means the condition is one of C_1 and C_2. The operator ‘|’ means logical OR.

5) **Dispensable components:** the components are written in square brackets and it means whether the components appear at the position or not.

T is the translation corresponding to the source sentence and it is similar to the format of left side. When an input sentence of source language meets the conditions expressed by the left side, it will be translated to the target language expressed by the right side *T*.

Example 1:

#我 想 预定 QP 个 N => I want to reserve !QP !N+s.

QP stands for the number phrase. By using the template, the input Chinese utterance ‘我想预定三个单人间’ will be translated into ‘I want to reserve 3 single rooms.’.

Example 2:

#QP 个 带|有 N1 的 N2 ＝> !QP !N2+s with !N1.

With the template, the input Chinese ‘一个带空调的单人间’ will be translated into ‘One single room with air condition.’, and the input Chinese ‘两个有电话的双人间’ will be translated into ‘Two double rooms with telephone’.

In our system, four hundred and sixty five templates are summarized from the training data. For more details of the template-based translation system, please refer to the reference [1].

2.2. Phrase-based translation model

2.2.1. Phrase-based translation model

In our system, the phrase-based translation model is based on the log-linear model [7] and the phrase we mention here is composed of a series of words that perhaps possess no syntax or semantic meanings. In the log-linear model, given the sentence f (source language), the translating process is searching the translation e (target language) with the highest probability:

$$e^* = \arg\max_e \sum_{m=1}^{M} \lambda_m h_m(e, f) \tag{2}$$

where $h_m(e,f)$ is a feature function and λ_m is the model parameter.

2.2.2. Phrase extraction

Word alignments are first obtained by using the GIZA++ toolkit in both translation directions and then summarizing the two alignments. We use a number of heuristics, which belongs to the refined method [3], to improve the alignment as the IBM model can not map one target (English) word to the multiple source (Chinese) words.

Based on the word alignment, we collect all aligned phrase pairs that are consistent with the word alignment: the words in a legal phrase pair are only aligned within the phrase pair and not to the words outside [3].

2.2.3. Feature functions

In the phrase-based system, we use the three feature functions: the target language model (LM), the translation model and the distortion model.

- **Target 3-gram LM:** standard language model with Kneser-Ney smoothing method by using the *ngram-count* tool in SRILM toolkit[1].
- **Translation model:** we use four different translation probabilities to calculate a hybrid translation probability.

$$P_T(\bar{f}_k \mid \bar{e}_k) = \lambda_1 P_{lex}(\bar{f}_k \mid \bar{e}_k) + \lambda_2 P_{lex}(\bar{e}_k \mid \bar{f}_k) + \lambda_3 P_{freq}(\bar{f}_k \mid \bar{e}_k) + \lambda_4 P_{freq}(\bar{e}_k \mid \bar{f}_k) \tag{3}$$

Where the $P_{lex}(\bar{f}_k \mid \bar{e}_k)$ is based on the lexical probability of IBM model 4 and $P_{lex}(\bar{e}_k \mid \bar{f}_k)$ is the inversed lexical probability. The $P_{freq}(\bar{f}_k \mid \bar{e}_k)$ and $P_{freq}(\bar{e}_k \mid \bar{f}_k)$ are the probabilities of phrase frequency. λ_i (*i*=1 to 4) is the parameter of probability.

- **Distortion model:** in our system, we use a simple distortion model:

$$P_{d(a_k - b_{k-1})} = \lambda \mid a_k - b_{k-1} - 1 \mid \tag{4}$$

Where a_k denotes the start position of the source phrase that was translated into the *k*th English phrase, and b_{k-1} denotes the end position of the source phrase translated into the (*k*-1)th English phrase.

2.2.4. Decoding strategy

In the phrase-based statistical machine translation system, the decoder employs a beam search algorithm that is similar to the Pharaoh decoder [9]. Considering the different expression habits between Chinese and English, some words must be complemented when translating Chinese sentences into English. For example, some frequent words, such as “a, an, of, the”, are difficult to extract because those words have zero fertility and correspond to NULL in IBM model 4. We call them F-zerowords. When decoding, the F-zerowords can be added after each new hypothesis, which means, a NULL is added after each phrase in the source sentences. At the same time, in Chinese sentence there are many auxiliary words and mood words which correspond to NULL in English. In our decoder, we select the final hypothesis of the best translation in the last several stacks instead of those cover all the source words, because not all the words in source language sentence are necessary to be translated. We will describe the two improvements in detail.

- **Expanding F-zerowords**

The decoder starts with an initial hypothesis. There are two kinds of initial hypothesis: one is an empty hypothesis that means no source phrase is translated and no target phrase is generated, and the other one is expanded from the empty hypothesis by adding F-zerowords.

New hypotheses are expanded from the current existing hypotheses as follow: if the last target phrase generated in the existing hypothesis is an F-zerowords, an un-translated source phrase and its translation options are selected to expand the hypothesis. If the last target phrase is not F-zerowords, the hypothesis can be expanded as described above or by selecting one of the F-zerowords. An example of hypotheses expansion is illustrated in Figure 2. The expansion with cross is unallowable because the F-zerowords can not be added after F-zerowords.

[1] http://www.speech.sri.com/projects/srilm

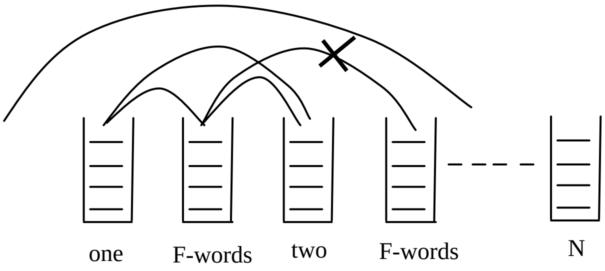

Figure 2: different hypothesis expansion approach

As is shown in Figure 2, the hypotheses are stored in different stacks and each of them has a sequence number. The hypothesis whose last target phrase is not F-zerowords and in which p source words have been translated accumulatively will be put into the odd stack S_{2p-1}(p=1,2……). In the same way, if the last target phrase is F-zerowords, the hypothesis will be in the even stack S_{2p}. We recombine the hypotheses and prune out the weak hypotheses that are similar to the Pharaoh decoder. Those operations will reduce the number of hypotheses and speed up the decoding.

- **New tracing back method**

When all the words of the source sentences have been translated, by searching not in the final stack which covers all the source words, but in the final several odd stacks, we find the best translation according to the accumulative score:

$$S_{best} = \arg\max\{P_s\} \quad (5)$$

where P_s is the accumulative probability of the hypothesis S.

2.3. Combine the two engines

The two translation systems have their own advantages. The TBMT system can generate more accurate results than the SMT system but the number of templates is limited while there are a lot of sentences that can not find proper templates The SMT system can translate more sentences than the TBMT system. So we combine the TBMT system with the phrase-based SMT system.

There are some different methods to select the best results generated by different translation system [10] [11]. The selectors in [10] [11] work after all the translation system separately have one or more results. In our system, the candidate selector is very simple.

Given a source sentence, the processes of translation are given as follows: First, the sentence is inputted into the TBMT system. If the TBMT can translate the sentence (there is suit template for the sentence), the translation is generated and the translation process is ended. Otherwise the sentence is sent to the SMT system and the SMT translate the sentence. The translation process is shown in Figure 1.

3. Improvements of the baseline system

3.1. Restore the punctuation information

The IWSLT 2006 evaluation campaign focuses on the translation of speech data and recognition results. The source language sentences do not contain any case information or punctuations, but the translation results must contain the information. The training data contain punctuations and our experiments show that the punctuation information is useful for translation. In order to use the punctuation information in training data, we insert the punctuations in source language sentences when preprocessing with the *hidden-ngram* tool in the SRILM toolkit fed with an *n*-gram punctuations language model. Because some punctuations are not restored, we purse this processing in the target language sentences after translation

3.2. Number pre-processing

In the phrase-based SMT system, it is difficult to gain the accurate translation of number because the training data can not contain all the ways to express number and the phrase extraction is not perfect. So we build a number pre-processing model based on rules to translate the number. According different expression ways, we summarize different rules.

1) Numeral translation

- **Arabic numeral translation:** We can just translate them directly from the Chinese to English.
- **Ordinal numeral translation:** The Chinese ordinal number can be easily recognized by the maker words, such as "头","第". And there exists the corresponding ordinal numbers in English, thus we can translate this type of numeral by using summarized rules directly.
- **Chinese numeral translation:** Considering the different expression ways between Chinese number and English number, we use the Arabic numerals as an intermediary in translation. For example, when translating "三 亿 四 千 万", we first transfer it as "3,4000,0000" according to Chinese expression; then we change the count unit (represented as comma) from four to three, which is "340,000,000" as a result; finally we gain English translation as "three hundred and forty million".

2) Temporal translation

- **Duration translation:** For this type, there is no ambiguity between the two languages pair, so we can translate them based on word translation directly.
- **Time translation:** This type is much different from duration translation because of its diversity. From the selected corpus, we summarized all the examples of time translation, and acquired about thirty pieces of rules. we use the corresponding rules for this type.

4. Experiment Result

We carried a number of experiments on the Chinese-to-English translation tasks. The training data supplied by IWSLT'06 is used for training the phrase-based SMT system and extracting the templates for template-based translation system. The correct recognition results in develop corpus are used for testing the translation system.

4.1. Experiments on the development set

The training data contains the punctuations, so a lot of phrases extracted from it contain punctuations. But the input sentences have no punctuations, so the phrases with punctuations are useless. We delete all the punctuations in the training data and then extract the phrase. As shown in line 2 of Table 1, this processing make the BLEU score decrease from 0.1486 to 0.1195. This means that the punctuations are useful for word alignment and phrase extraction because the punctuations appear frequently in the sentences.

Considering usage of the punctuations, we insert the punctuations into the source sentences before the translation process. The performance of the system has been absolutely improved about 0.0214.

From Table 1, when combining the SMT system with the TBMT system, the BLEU score increases more obviously from 0.1700 to 0.2238. The reason is the TBMT can generate more accurate translation result and there are 55 sentences are translated by the TBMT system. For example, the sentence ‘我 喜欢 红色 或者 黑色 都 可以’ is translated into ‘I would like either red or black .’ by the TBMT system and ‘I'd like a red , black be all right .’ by the phrase-based SMT system. The result generated by the TBMT is more fluency and accurate.

Table 1: Results of improvement

methods	BLEU	NIST
SMT	0.1486	5.4866
SMT+Delete punctuation	0.1195	5.3029
SMT+Punctuation	0.1700	5.7994
SMT+Punctuation+TBMT	0.2238	6.1305

4.2. IWSLT 2006 test results

The test results of IWSLT 2006 are shown in Table 2. In the task of translating the ASR outputs, our systems only translate the 1-best results. Because of the influence of ASR errors, the systems perform better on the correct recognition results than the ASR results.

Table 2: Results of IWSLT 2006 test data

		BLEU	NIST
read	SMT	0.0972	3.5557
	SMT+TBMP	0.1037	3.6384
spontaneous	SMT	0.1001	3.4869
	SMT+TBMP	0.1070	3.5755
correct	SMT	0.1167	3.9288
	SMT+TBMT	0.1284	4.0658

Table 3: The numbers of sentences translated by TBMT

	read	spontaneous	correct
number	33	29	48

5. Conclusions

In summary, this paper presents a hybrid approach to translation system. This system combines the template-based translation engine with the phrase-based statistical translation engine by using a simple method to select the best translation from the results generated by the two engines. Considering the dissimilarity between Chinese and English, we develop the decoder with new tracing back and hypotheses expansion methods. In order to make use of punctuation information, we restore the punctuations in the source sentences. The experiments show that the punctuation information can improve the performance of the translation system.

6. References

[1] Chengqing Zong, Taiyi Huang and Bo Xu, An improved Template-Based Approach to Spoken language Translation. In *Proc. ICSLP2000*, vol, IV, pp. 440-443, Beijing, 2000.

[2] Yeyi Wang and Alex Waibel, Fast Decoding for Statistical Machine Translation. In *Proc. of ICSLP 98*, Vol. 6, pp. 2775-2778, 1998.

[3] F. J. Och and H. Ney, Inproved Statistical Alignment Model. *Proceeding of ACL-00*, pp. 440-447, 2000.

[4] S. Vogel, Y. Zhang, et al, The CMU statistical machine translation system. In *Proc of the Machine Translation Summit IX*, pp. 110-117. 2003.

[5] Yamada, K. and Knight, A Syntax-based Statistical Translation Model. In *Proc of the 39 Annual Meeting of ACL*. pp. 6-11. 2001.

[6] Marcu Daniel, and William Wong. 2002. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia, PA, USA. July 2002.

[7] F.J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic*, pp.295-302, 2002.

[8] Wei Wei, Wei Pang, Zhendong Yang, JinHua Du, Zhenbiao Chen, Chengqing Zong and Bo Xu. CASIA SMT System for TC-STAR Evaluation Campaign 2006. *TC-STAR workshop on “Speech-to-Speech Translation”*,pp 69-74, 2006.

[9] Koehn Philipp. 2004. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, pages 115-124. (http://www.isi.edu/licensed-sw/pharaoh)

[10] Nomoto Tadashi, Predictive Models of Performance in Multi-Engine Machine Translation. *Proceedings of MT Summit IX*, pp 269-276, 2003.

[11] G. Foster, S. Gandrabur, P. Langlais, G. Russell, M. Simard. Statistical Machine Translation: Rapid Development with Limited Resources. *Proceedings of MT Summit IX*, 2003.