

## CASIA SMT System For TC-STAR Evaluation Campaign 2006

Wei Wei, Wei Pang, Zhendong Yang, JinHua Du, Zhenbiao Chen, Chengqing Zong, Bo Xu

Institute of Automation, Chinese Academy of Sciences  
No.95 Zhongguancun East Road, Beijing, China  
weiwei@hitic.ia.ac.cn

### Abstract

This paper presents our statistical machine translation (SMT) system for TC-STAR Evaluation Campaign 2006. In our Chinese-to-English SMT system, we use the phrase-based translation model. Considering the characteristics of Chinese-to-English translation, we propose some approaches to improve different tracing back and hypothesis expansion. For remediation of the intrinsic flaws in statistical framework, we make use of the existing language resources to improve the preprocessing model, word alignment and the estimation of phrase probability. The preliminary experimental results show that integrating kinds of knowledge resources into the statistical system has the potential to improve the performance of Chinese-to-English SMT system.

### 1 Introduction

Statistical machine translation is a popular approach in the current world for the large vocabulary text translation. In the early 90s, IBM developed Candide system. Since then, many statistical machine translation systems were proposed (Wang & Waibel, 1998; Och & Ney, 2000). These systems apply a translation model to capture the relationship between the source language and target language, and use a language model to drive the search process. The primary IBM model was purely word-based (Brown, 1993), and the phrase-based statistical translation models are proposed to incorporate more complex structure and to get better lexical choice and more reliable local reordering. Yamada and Knight (2001) used the phrase translation in a syntax-based SMT system. March and Wong (2002) introduced a joint-probability model for phrase translation. CMU and IBM also improved their systems with phrase translation capability (Vogel *et al.*, 2003).

Our system applies a phrase-based translation model to capture the corresponding relationship between two languages. We learn phrase alignments from a corpus that the words are aligned by a training toolkit for a word-based translation model: the Giza++ (Och and Ney, 2000) toolkit for the IBM models (Brown *et al.* 1993). The extraction heuristic is similar to the one used in the alignment template work by Och *et al.* (1999). The phrase-based decoder we developed employs a beam search algorithm, similar to the one in (Koehn *et al.*, 2003), but it applies the words with fertility probability of zero in the target language. A different tracing back algorithm to find the best path is proposed. In addition, we use some adaptive methods to the baseline system.

This paper is organized as follows: Section 2 describes the baseline system. Section 3 explains the method of some adaptations to the original system. In Section 4, we present a series of experiments in which the Chinese sentences are translated into English, and the results of these experiments are analyzed. We make a conclusion in Section 5.

### 2. System Description

The system uses the phrase-based statistical machine translation model (Koehn *et al.*, 2003). We developed a phrase-based decoder which is based on beam search, but improved the approaches of hypothesis expansion and tracing back while considering the big difference between Chinese and English. In this section, we will give an overview of the system.

#### 2.1 Phrase-based translation model

The phrase translation model is based on the noisy channel model. We use Bayes rule to reformulate the translation probability for translating a foreign sentence  $f$  into English sentence  $e$  as

$$\arg \max_e P(e/f) = \arg \max_e P(f/e)P(e) \quad (1)$$

This allows for a language model  $P(e)$  and a separated translation model  $P(f/e)$ . Recall that due to the Bayes rule, the translation direction is inverted from a modeling standpoint.

During decoding, the foreign input sentence  $f$  is segmented into a sequence of  $K$  phrases  $\bar{f}_1^K$ . We assume a uniform probability distribution over all possible segmentations. Each foreign phrase  $f_k$  in  $\bar{f}_1^K$  is translated into an English phrase  $e_k$ . The English phrases may be reordered. Reordering of the English output phrases is modeled by a relative distortion probability distribution  $P_{d(a_k-b_{k-1})}$ , where  $a_k$  denotes the start position of the foreign phrase that was translated into the  $k$ th English phrase, and  $b_{k-1}$  denotes the end position of the foreign phrase translated into the  $(k-1)$ th English phrase. In all our experiments, we use a simpler distortion model

$$P_{d(a_k-b_{k-1})} = \lambda |a_k - b_{k-1} - 1| \quad (2)$$

In summary, the best English output sentence  $e_{best}$  given a foreign input sentence  $f$  according to the phrase-based model is

$$\begin{aligned}
e_{best} &= \arg \max_e P(e|f) \\
&= \arg \max_e P_{LM}(e) \prod_{k=1}^K P_T(\bar{f}_k | \bar{e}_k) P_{d(a_k-b_{k-1})} \quad (3)
\end{aligned}$$

In the formula (3),  $P_T(\bar{f}_k | \bar{e}_k)$  is the probability of the  $k$ th phrase pairs  $f_k, e_k$ . For all our experiments we use the standard word-based trigram language model generated with smoothing Kneser-Ney methods by using SRILM (Stolcke, 1999).

## 2.2 Acquisition of phrase translation

Word alignments are first obtained by using the GIZA++ toolkit in both translation directions and then summarizing the two alignments. Since the IBM models implemented in GIZA++ are not able to map one target (English) word to the multiple source (Chinese) words, we improve this alignment with a number of heuristics, which are called refined method (Och and Ney, 2003).

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 to each other, and not to words outside (Och *et al.*, 1999).

Given the collected phrase pairs, we estimate the phrase translation probability distribution by the lexical probability of IBM model4:

$$P_T(\bar{f}_k | \bar{e}_k) = \prod_i \sum_j p(f_k^i / e_k^j) \quad (4)$$

## 2.3 Decoding strategy

The phrase-based decoder we developed employs a beam search algorithm, similar to Pharaoh decoder (Koehn, 2004b). But considering the different expression habits between Chinese and English, some words must be complemented when translating Chinese sentences into English. For example, some articles “a, an, the” and some prepositions “as, of”. All these words that appear frequently are difficult to extract, because they are words that have zero fertility and correspond to NULL in IBM Model 4. We call them F-zero-words. After each new hypothesis is expanded, the F-zero-words can be applied. That is to say, a NULL is added after the source phrase translated. In addition, considering there are many auxiliary words and mood words in the Chinese sentences, and there are no corresponding words in the English sentences, another improvement of our decoding algorithm is that we select the final hypothesis of the best translation in the last several stacks instead of the one covered all the source words, because perhaps not all words of the input sentence are necessary to be translated. This is different from that in Pharaoh. We will describe it in detail in Section 2.3.2.

### 2.3.1 Improvement of integrating F-zero-words bins

The decoder starts with an initial hypothesis. The initial hypothesis has two kinds: one is an empty hypothesis where no source phrase is translated and no target phrase is generated, and the other one is that generating F-zero-words

and corresponding to a NULL we supposed at the beginning of the input text.

New hypotheses are expanded from the current existing hypotheses as follows: if the target phrase of the existing hypothesis is an F-zero-words, the new hypothesis of a source phrase that has not been translated and one of its translation options are selected. If the target phrase is not the F-zero-words, there are two choices: one is expanding a hypothesis which is achieved as that described justly; the other is expanding a hypothesis by selecting one of the F-zero-words as output. This corresponds to a NULL which is added into the input text after the source phrase of the existing hypothesis. An example of hypothesis expansion is illustrated in Figure 1. The expansion with Cross is unallowable because F-zero-words must be expanded from Non F-zero-words

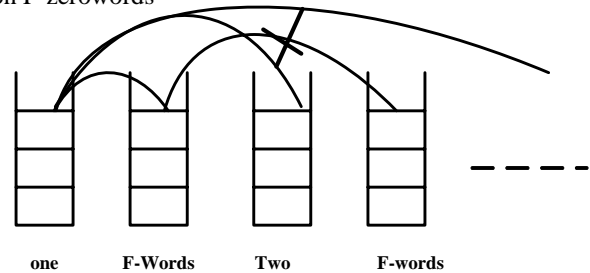


Figure 1: Different hypothesis expansion approach

The hypotheses are stored in different stacks and each of them has a sequence number. The odd stack  $S_{2p-1}$  ( $p=1, 2, \dots$ ) contains all hypotheses whose target phrases are not F-zero-words and in which  $p$  source words have been translated so far. The even stack  $S_{2p}$  contains all hypotheses whose target phrases are F-zero-words and in which  $p$  source words have been translated accumulatively. We recombine search hypotheses as done in (Och & Ney, 2001), and prune out weak hypotheses based on the probability they incurred so far and a future score estimated like that in Pharaoh. All these reduce the number of hypothesis stored in stacks to speed up the decoder. The different hypothesis expansion approaches are shown in Figure 1.

### 2.3.2 Improvement of tracing back method

The hypotheses are generated continuously until all the words of the input sentence have been translated. Then, by searching not in the final stack which covers all the source words, but in the final several odd stacks, we find the final hypothesis of 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  $S$ , the tracing back method we used is denoted as back1, which is different from that in Pharaoh denoted as back2. Our experiments show the better performance of our method in Section 4.

### 3 Adaptations to the Baseline System

We focused on the pre-processing methods to deal with both the translation model and language model due to the diversity of language expressions between the Chinese and English, especially the named entity (NE) translation, such as proper name, number expressions. There are a lot of resources about proper name translations, especially for location name and person name. So we make use of the exiting tools, such as the NE dictionary, toolkit for recognizing the English NE, to deal with proper name. But for the number translation, we developed a in-house module to deal with the differences between the Chinese and English expressions. In addition, we use a mount of dictionary resources to improve the quality of pure statistical word alignment mentioned above. Finally, we use an efficient algorithm called suffix-array to calculate the frequency of phrase pairs in the whole training data, and integrate the result into the phrase translation probability.

#### 3.1 Building number pre-processing model

In our system, we gain better result by the use of existing resources, such as dictionary and character-based translation, to deal with proper name translation. However, dictionary lookup is particular difficult to translate numbers because of its limited coverage, and simply applying word-based or phrase-based translation without considering the inner rules, in most cases, can not achieve satisfactory result. Therefore, we pick out the numbers for special treatment to reduce the mistakes in translation.

Considering the characteristics of number expression in Chinese and English, we build the number pre-processing model based on rules that we summarized from selected corpus, which contains about five thousand sentences with different kinds of number expressions. And the corpus is covered with all kinds of domains, such as travel, news, sports and so on. According to different expression ways, we divide number translation into two sub-types: numeral translation and temporal translation. In section 3.1.1 and 3.1.2, we will describe the two modules separately.

##### 3.1.1 Numeral translation

Numeral is widely used in corpus. We summarized numeral translation as follow three types:

**Arabic number translation:** Both in Chinese and English, Arabic number is widely used. Taking the phone numbers and room numbers for example, they are only cardinal numbers. We can just translate them directly from the Chinese to English.

**Ordinal number translation:** The Chinese ordinal number can be easily recognized by the maker words, such as “第”, “头”. And there exist the corresponding ordinal numbers in English, thus we can translate this type of numeral by using summarized rules directly.

**Chinese number translation:** Chinese number, which is referred to the number written in Chinese, is the most difficult due to different ways of counting numbers. In

Chinese, we count numbers by four digits, such as “亿,万”. While in English, we count numbers by three digits, such as “billion, million”. So it’s inappropriate to translate them directly. We adopted 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”.

##### 3.1.2 Temporal translation

Temporal translation is a special kind of NE in Chinese-to-English translation, because there is no one-to-one corresponding relation in many expressions. Taking “四点十五” for example, we can find more than one corresponding translations as follows: 1) a quarter past four; 2) fifteen past four; 3)4:15. On the other hand, there exist three translation candidates corresponding English expression “Saturday”, such as “周三”, “星期三”, “礼拜三”. Thus we summarized temporal translation as follow two types:

**Duration translation:** For this type, there is no ambiguity between the two language pair, thus 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. When translating this type, we should use the corresponding rule.

With the number preprocessing model, we introduce the concept of phrase template. Before the training process, we replace the numbers appeared in the bilingual sentences with variables, and gain phrase template depository after phrase-extraction. In the process of translation, we translate the sentences by use of the template depository. The whole process can be seen in Figure 2.

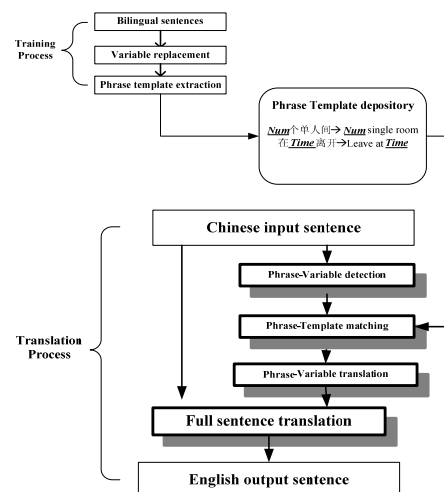


Figure 2 Process of Phrase-template translation

### 3.2 Optimizing word alignment

Though the refined word alignment has been applied in the most phrase-based statistical machine translation systems, the statistical techniques require almost no prior knowledge and based solely on the bilingual corpora. The pure statistical methods gain the better performance on the similar language pairs, such as alphabetic Indo-European languages. But for the Chinese-English word alignment task, there exist the phenomenon that can not be solved by simply calculating the co-occurrences of word pairs. Thus, we try to integrate the lexical clues based on the refined statistical model.

The dictionary we used contains both one-to-one and many-to-many translation entities. Based on the specific methods, we can correct one-to-one, many-to-one and one-to-many translation pairs which exist in the dictionary. The details of the approach are listed as follows:

**Lower case:** In order to avoid the mismatch between lower and upper case, we convert all the English words in the sentences into the lower case when finding the proper English words in the dictionary

**Stemmed English words:** In order to improve the morpho-syntactic word alignment, we set a constant  $\lambda$ , and extract  $\lambda * L$  characters of the word, where  $L$  is the number of characters in the word.

**Word disambiguation:** We integrate some restrictions of context when the word appears more than one time in a sentence, especially some functional words such as the preposition ‘of’, and so on.

### 3.3 Hybrid phrase translation probability

In the baseline model, we acquire the phrase translation probability only based on the results of IBM model 4, shown in formula (4). However, applying formula (4) into the Chinese-to-English translation has some drawbacks: if only one word of the source phrase has no appropriate corresponding word in the target phrase, the phrase translation probability will be small. Because there are many auxiliary words and mood words in Chinese. This problem became more serious. Furthermore, the dependence on the lexical probability is lack of regarding the phrase pairs as a whole part in the training data. Thus, we should integrate the frequency of phrase pairs in the training corpora into the estimation of phrase translation probability.

In order to efficiently calculate the frequency of extracted phrases in the large scale of corpora, we use the data structure of suffix array (Yamamoto & Church, 2001), which is designed to make it convenient to compute term frequencies for all substrings in the corpus. It is more compact and is amenable to storage in secondary memory. The data structure is simply an array containing all the pointers to the text suffixes sorted in lexicographical (alphabetical) order. Each suffix is a string starting at a certain position in the text and ending at the end of the text. Searching a text can be performed by binary search using the suffix array. The time complexity of the algorithm

is  $O(\log N)$ , where  $N$  is the number of running words in the whole corpus.

## 4 Experiments

We carried a number of experiments on the Chinese-to-English translation tasks. In order to perform the experiments effectively, we built two categories of the training data: 1) spoken language in small size (BETC corpus): 130,000 sentence-pairs, and the sentence length spans from 5 to 20 Chinese words. 2) General domain in large size (863<sup>1</sup> corpus): 870,000 sentence-pairs, and the sentence length spans from 10 to 50 Chinese words.

### 4.1 Results with baseline system

For the baseline system, the small bilingual corpus, are used as the training data for investigating the effect of F-zero words. The trace back method and numeral model we used, 1000 sentences of length 5-20 were reserved for testing. The results are shown in Table 1, where **WBT** means the word-based translation model, **PBT** means the phrase-based translation model, **F0** means F-zero words are applied, **back1** stands for our decoder, and **back2** stands for Pharaoh decoder.

Table 1 Results with Baseline system

Methods	Bleu (4-gram)
WBT+back2	0.1833
WBT+back1	0.1919
WBT+F0+back2	0.2372
WBT+F0+back1	0.2663
PBT+back2	0.2730
PBT+back1	0.2864
PBT+F0+back2	0.2763
PBT+F0+back1	0.2882
PBT+F0+back1+NUM	0.3177

#### 4.1.1 Comparison of different tracing back methods

We can see the result of the word-based system with no F-zero words and Pharaoh decoder is the lowest. When the tracing back method used in Pharaoh decoder is replaced by the method proposed by us, the score increases from 0.1833 to 0.1919. The score increases more obviously from 0.2372 to 0.2663 when F-zero words are added. When based on phrase translations, the score also goes up owing to using back1. All these show back1 is superior to back2 because some source language words are not necessary to be translated.

<sup>1</sup> 863 Programme: another name of the Evaluation on Chinese Information Processing and Intelligent Human-Machine Interface Technology, which is supported by China’s national High-Tech Research and Development Programme(HTRDP)

Table 3: The training data we used in TC-STAR 2006

	Reference	Description	#Sent.(parallel)
LDC large data	LDC2000E49	Hong Kong Hansards Parallel Text, aligned-sentence level	1,300,000
	LDC2003E25	Hong Kong News Parallel Text, sentence-aligned	700,000
Publicly available data (all from Chinese LDC)	CLDC-LAC-2003-004	Chinese-English Sentence aligned Bilingual Corpus	200,000
	CLDC-LAC-2003-006	Chinese-English/Chinese-Japanese parallel corpora	200,000
Total			2,400,000

#### 4.1.2 The role of F-zero words

From Table 1, when F-zero words are added through the decoding of word-based system, the score goes up sharply from 0.1919 to 0.2663 with back1, increasing by 0.0744, which denotes F-zero words play an important role. This is because some words, such as of, the and so on, are complemented by the language model, distortion model. The same conclusion can be drawn when translating based on phrase translations, but the increase is not so obvious (from 0.2864 to 0.2882), probably because with the phrase number rising, some F-zero words are extracted in the phrase, and the effect of F-zero words is decreased.

#### 4.1.3 Effect of phrase template

As mentioned in section 3.1, after replacing Named Entity (Location, Person, Number and time) with specific variables, the original phrase pair became phrase template, which has the ability of generalization. For example:

X 个单人间 → X single Room

In our experiment using the small bilingual corpus, about 5% extracted phrases contain variables. From Table 1, the performance of the whole system has been absolutely improved about 0.032 by using the phrase template.

## 4.2 Results with adaptive approaches to baseline

In order to test the different approaches of improvement on the baseline system, we use the large scale of the bilingual corpus as the training data and 500 sentences with 5 translations (863-2005 standard dialogue test set) as the test data.

Table 2: Results of adaptive approaches to the baseline

Hybrid probability	Bleu(4 gram)
$P_T(\bar{f}_k   \bar{e}_k)$ (Baseline)	0.1814
$P_T(\bar{f}_k   \bar{e}_k) + P_T(\bar{e}_k   \bar{f}_k)$	0.1816
$P_T(\bar{f}_k   \bar{e}_k) + P_T(\bar{e}_k   \bar{f}_k) + P_{freq}(\bar{f}_k   \bar{e}_k) + P_{freq}(\bar{e}_k   \bar{f}_k)$	0.2090
$0.55P_T(\bar{f}_k   \bar{e}_k) + 0.15P_T(\bar{e}_k   \bar{f}_k) + 0.15P_{freq}(\bar{f}_k   \bar{e}_k) + 0.15P_{freq}(\bar{e}_k   \bar{f}_k)$	0.2190
Best + improving word alignment	0.2211

#### 4.2.1 Results of hybrid phrase translation probability

The baseline system only use  $P_T(\bar{f}_k | \bar{e}_k)$ , which is based on the lexical probability of IBM model 4. In the experiments, we added the inverse probability  $P_T(\bar{e}_k | \bar{f}_k)$ , but the score almost remains the same, because it is only another form of the lexical probability and can not overcome the flaws in formula (3). Furthermore, we added the other two probabilities of phrase frequency:  $P_{freq}(\bar{f}_k | \bar{e}_k)$ ,  $P_{freq}(\bar{e}_k | \bar{f}_k)$ , which is described in Section 3.3. This method has an obvious increase with BLEU score, which is from 0.1814 to 0.2090. Finally, we adjusted the parameters with each probability, and the score increase by 0.01. Because of the intrinsic flaws of formula (3), we should integrate other knowledge resources to calculate the probabilities. Here, the parameter is acquired simply by hand. Recently, we have realized MER training process, and we hope to submit the optimizing system soon.

#### 4.2.2 Results of improving word alignment

In the experiment, first, we preserve the dictionary entities which exist in the training corpus. After filtering, there are three kinds of useful translation pairs in the dictionary (Chinese-to-English): 1) one-to-one: 19,000 pairs, 2) one-to-many: 12,100 pairs, and 3) many-to-one: 73,811 pairs. Then, we integrated the three methods mentioned in Section 3.2 to improve the word alignment. From Table 2, we can see the final BLEU score increases from 0.2190 to 0.2211, and the promising result shows the potential to improve the translation system through integrating the different kinds of language resources.

## 4.3 TC-STAR 2006 test result

For the TC-STAR 2006 evaluation, we only use a small part of LDC data, and some public data from Chinese LDC ([www.chineseldc.org](http://www.chineseldc.org)). So, our system is under the SECONDARY data condition. Table 3 shows the data we used in details

The test result of TC-STAR in 2006 is shown in Table 4. ASR is the result after speech recognize. We use the tool provided by RWTH to gain the broken of every sentence in ASR result.

Table 4: Result of TC-STAR 2006

	BLEU	NIST
VERBATIM	0.0849	4.6290
ASR	0.0744	4.1339

## 5 Conclusions

In summary, this paper presents a phrase-based statistical machine translation system including the improvement of tracing back and hypothesis expansion methods in decoding, and some adaptations to the baseline system. Our experiments show that phrase-based translation gets much better performance than the traditional word-based methods. The F-zerowords usually play an important role in decoding, and the tracing back method we used is superior to that used in Pharaoh decoder. Considering the dissimilarity between the Chinese and English, we optimized the baseline system in pre-processing model, word alignment and the methods of calculating phrase probability. The results of experiments shows that integrating much language resources can overcome the intrinsic flaws in statistical methods, and lead to the higher quality of the whole translation system. Focused on the characteristic of the Chinese-to-English translation, now we are trying to use the more approaches and language resources that can be integrated in the current statistical system.

## References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19, no. 2, pp. 263-311.

Koehn, P, Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. *In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics*. pp.127-133

Mikio Yamamoto, Kenneth W. Church.(2001).Using suffix arrays to compute term frequency and document

frequency for all substrings in a corpus. *Computational Linguistics*, Vol.27 (1):pp. 1-30.

Och, F. J., Tillmann, C., and Ney, H.(1999). improved alignment models for statistical machine translation. *In proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.20-28.

Och, F. J. and Ney, H (2000) .Improved Statistical Alignment Model. *Proceeding of ACL-00*, pp. 440-447.

A. Stolcke (2002), SRILM -- An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver.

S. Vogel, Y. Zhang, F. Huang, A. Venugopal, B. Zhao, A. T. and M. Eck, and A. Waibel, (2003) "The CMU statistical machine translation system," in *Proc. of the Machine Translation Summit IX*, New Orleans, LA. pp.110-117

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

Yeyi Wang and Alex Waibel.(1998).Fast Decoding for Statistical Machine Translation. *Proc. ICSLP 98*, Vol. 6, pp.2775-2778

Wei Pang ,Z.Yang, Z.Chen, W.Wei, B.Xu and C.Zong(2005) "The CASIA Phrase-Based Machine Translation System." *In Proceedings of the International Workshop on Spoken Language Translation*. Pittsburgh, USA. pp.114-121.

Z.Yang, Z.Chen, W.Pang, W.Wei and B.Xu.(2005).“The CASIA Phrase-Based Machine Translation System”. *In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*. October 30th - November 1st, 2005. Wuhan, China. pp.416-419