

CASIA SMT System for the 2006 NIST MT Evaluation

Jinhua Du, Wei Wei, Yufeng Chen, Keyan Zhou, Wei Pang, Zhendong Yang,
Chengqing Zong, Bo Xu

Institute of Automation, Chinese Academy of Sciences
No.95 Zhongguancun East Road, Beijing, China

Abstract

This paper describes the first NIST MT Evaluation system of Institute of Automation, Chinese Academy of Sciences (CASIA). 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 the system performance, especially in named entity identification and translation. Our efforts resulted in an improved translation performance on 2005 NIST set. Participating in NIST MT Evaluation was also a valuable learning experience in large scale system building.

1. Introduction

This paper describes the first NIST MT Evaluation system of Institute of Automation, Chinese Academy of Sciences. Our system applies a phrase-based translation model to capture the corresponding relationship between two languages. We learn the phrase alignments from a corpus that the words are aligned by a training toolkit for a word-based translation model: the Giza++ toolkit (Och and Ney, 2000) 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 (Wei, 2006).

In this paper, we emphasize on our improvements in named entity identification and experiments designed for building our final system.

This paper is organized as follows: Section 2 describes the baseline system and some improvements. Section 3 describes the data we used to participate in this campaign. 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, and the next research to solve current problem in our system is also introduced.

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 the Chinese and English. In this section, we will give an overview of the system.

2.1 Baseline System

The phrase translation model is based on the noisy channel model. We employed a log-linear approach (Och and Ney, 2002) in our translation system. We searched for the most probable English sentence e given a foreign sentence f by maximizing the sum over a set of feature functions $h_m(e, f)$:

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e p(e/f) \\ &= \operatorname{argmax}_e \sum_{m=1}^M \lambda_m h_m(e, f)\end{aligned}\quad (1)$$

In our translation system, seven feature functions were employed when scoring candidate translation:

- Phrase mutual information translation probability (both directions)
- Lexical translation probability (both directions)
- Language model
- Future score
- Distortion model

The language model was a smoothed 4-gram model created and trained using the SRILM toolkit (Stolcke, 2002).

The phrase mutual information translation probability is different from the common phrase translation probability, which is defined as

$$p(\tilde{f}/\tilde{e}) = \frac{\operatorname{count}(\tilde{f}, \tilde{e})}{\sum_j \operatorname{count}(\tilde{f}, \tilde{e})}\quad (2)$$

Where, $\operatorname{count}(\tilde{f}, \tilde{e})$ represents the total numbers of co-occurrence in parallel training corpus, C is the monolingual foreign training corpus, and the denominator gives the total numbers of f occurred in C . We call this kind of phrase translation probability as mutual information based probability, which can absolutely reflect the real probability of phrases occurred in training corpus.

2.2 Improvements of NE identification and translation

In our previous system, the Chinese named entities (NE) were one of the most important problems in our research.

The accuracy of NE identification have the significantly impact on translation performance. In this evaluation, we concentrated on improve the accuracy of organization name identification and translation, which was ignored in our previous work.

We substitute the organization name with a variable ORG in the development set and the test set, which we called the variable template.

The Chinese NE identification model is defined as (Wu, 2005):

$$\begin{aligned}
(WC^*, TC^*) &= \arg \max_{(WC, TC)} P(WC, TC | W, T) \\
&= \arg \max_{(WC, TC)} P(WC, TC, W, T) / P(W, T) \\
&= \arg \max_{(WC, TC)} P(WC, TC, W, T) \\
&= \arg \max_{(WC, TC)} P(WC, W) (TC, T)^\lambda \\
&= \arg \max_{(WC, TC)} P(W | WC) P(WC) [P(T | TC) P(WC)]^\lambda
\end{aligned} \tag{4}$$

Where, WC^* represents the feature of word class, TC^* is the feature of word property, λ is the weight balancing WC and TC . $P(WC)$ is the context model of word class, $P(TC)$ is the context model of word property, $P(W/WC)$ is the entity model of word class, and $P(T/TC)$ is the entity model of word property.

2.3 Discriminative training

Minimum Error Rate (MER) training (Och, 2003) is a powerful method to train the parameters of log-linear model. We use Venugopal's trainer (Venugopal and Vogel, 2005) to tune the weights of feature functions, which runs in MATLAB version.

In our experiments, we use 108 sentences of 2005 NIST set as the development set to train parameters, and the remains are test set. In our log-linear model, we use seven feature functions (Section 2.1). The feature weights λ are

optimized in the following steps:

1. Initialize $\lambda_m = 1$ ($m = 1, 2, \dots, M$);
2. Compute an n -best list by the decoder, where $n = 200$;
3. Use the n -best list to perform MER training, and get the new parameters;
4. Use the new parameters to decoder, and get a new n -best list, and combine it with previous n -best list;
5. Repeat 3 till convergence or fixed times.

3. Training Data

We have made two experiments based on the development data set. One is oriented to the Large Data Track, and other one is for the Unlimited Data Track. As to the evaluation of Large Data Track, we designed and performed a series of experiments on development set, unfortunately, in 4 days' evaluation, we couldn't run a Large Data Track result because of the problem of our final system. The training data for Large Data Track are listed in Table 1.

Our final submitted evaluation condition is "unlimited". The training corpus used for translation model is given in Table 2 in detail.

The following Table 3 is statistics about the training data.

System	#Sent. (parallel)	#Eng. Voc	#Chn. Voc
CASIA	2.4M	131K	118K

Table 3: The statistics about TM training

The monolingual resources to train the language model are as follows:

Source	Reference	Description	#Sent.(parallel)
LDC large data	LDC2004T08	Hong Kong Hansard Parallel Text	1,300,000
	LDC2004T08	Hong Kong News Parallel Text	700,000
	LDC2002E18	Xinhua Chinese English Parallel News Text Version 1 beta	270,000
	LDC2003E14	FBIS Multilanguage Texts	100,000
	LDC2005T10	Chinese English News Magazine Parallel Text	280,000
Total			2,650,000

Table 1: The training data used for Large Data Track in 2006 NIST MT Evaluation

Source	Reference	Description	#Sent.(parallel)
LDC large data	LDC2004T08	Hong Kong Hansard Parallel Text	1,300,000
	LDC2004T08	Hong Kong News Parallel Text	700,000
Chinese LDC (publicly available data)	CLDC-LAC-2003-004	Chinese-English Sentence aligned Bilingual Corpus	200,000
	CLDC-LAC-2003-006	Chinese-English/Chinese-Japanese parallel corpus	200,000
Total			2,400,000

Table 2: The training data for Unlimited Data Track in 2006 NIST MT Evaluation

- 1) English part of the above parallel corpora.
- 2) LDC2005T12
-English Gigaword Second Edition (only used “xin_eng”)

There are statistics about the LM training data in Table 4.

System	#Sent.	#Eng. vocabulary
CASIA	9.6M	131K

Table 4: The statistics about LM training

4. Experiments

In the days before evaluation, we designed and performed many experiments to tune the parameters of system. Finally, we got an improved result. In 4 days evaluation, we successfully ran an unlimited result by our improved system.

4.1 Model training

We use the above (Section 3) data to train IBM model 4 by GIZA++ (Och and Ney, 2000), then to extract the bilingual phrases and compute the phrase probability by equation (2).

We use SRI Language Modeling Toolkit (Stolcke, 2002) to train the language model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) by above monolingual corpus (Section 3, Table 4). In our experiments, we trained 2 language models, a trigram model and a 4-gram model. Since the corpora are very large, the 4-gram model we get is also large, and we reduce the size by pruning bi-gram and trigram.

4.2 Results and analysis

By the experiments we designed, we get the following comparable results in Table 5. All experiments results are evaluated on 2005 NIST set except 108 sentences which are used as development set.

ID	LM	NE(ORG)	BLEU	NIST
1	3-gram	No	0.1832	7.0944
2	4-gram	No	0.1889	7.2524
3	3-gram	Used	0.2186	7.7462
4	4-gram	Used	0.2207	7.7669

Table 5: The results of experiments

About 3.8M bilingual phrases are extracted from training data. In Table 5, the system with 4-gram and organization identification get the best performance, and the BLEU score rise 3% when introduced organization identification and translation. In our experiments, 4-gram language model only made a little improvement.

In 2006 NIST MT Evaluation, we select the 4th system as our final system, and the preliminary result is 0.1894 of BLEU score, which is not satisfied for some reasons:

- 1) It is the first time for us to participate in the NIST MT Evaluation, we prepared not enough in data preprocessing and system building.
- 2) Our system is not complete, and there are still short of some modules such as post-processing and so on.
- 3) We don't introduce the syntactic resources and the structure information into our system, which is proved

useful to translation result.

In our next research, we will complete our system in the following aspects:

- 1) Strengthen the preprocess and post-process modules;
- 2) We will build a syntax-based system and introduce structure information into our current system.
- 3) There are many compound and complex sentences in news translation, and we will perform deep research on long sentences and present some concepts and solutions.

References

- [Chen and Goodman, 1998] Stanley F. Chen and Joshua Goodman, 1998. An empirical study of smoothing techniques for language modeling, Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.
- [Koehn, 2003] Koehn, P, Franz Josef Och., 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*. Pages 127-133
- [Och et al., 1999] Franz Josef Och, Tillmann, C., and Hermann Ney.(1999). improved alignment models for statistical machine translation. In *proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.20-28.
- [Och, 2003] Franz Josef Och, 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41th Annual Meeting of the ACL*, pages 160-167.
- [Och and Ney, 2000] Franz Josef Och and Hermann Ney (2000). Improved Statistical Alignment Model. In *Proceeding of the 38th Annual Meeting of the ACL*, pages 440-447.
- [Och and Ney, 2002] Franz Josef Och and Hermann Ney, 2002. Discriminative training and maximum entropy models for statistical machine translation. In *proceedings of the 40th Annual Meeting of the ACL*, pages 295-302.
- [Stolcke, 2002] Andreas Stolcke, 2002. SRILM -- an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901-904.
- [Venugopal and Vogel,2005] Ashish Venugopal and Stephan Vogel. 2005. Considerations in maximum mutual information and minimum classification error training for statistical machine translation. In *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*.
- [Wei, 2005] Wei Wei, Wei Pang, Zhendong Yang, Zhenbiao Chen, Chengqing Zong, Bo Xu. 2006. CASIA SMT System For TC-STAR Evaluation Campaign 2006. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, pages 69-74, Barcelona, Spain, June.
- [Wu, 2005] Youzheng Wu, Jun Zhao, Bo Xu, Chinese Named Entity Recognition Model Based on Multiple Features. In *Proceedings of HLT/EMNLP 2005*, pp.427~434, October 6-8, Vancouver, B.C., Canada