

NLPR Machine Translation System for IWSLT 2006

Chengqing ZONG

National Laboratory of Pattern Recognition (NLPR)

Chinese Academy of Sciences (CAS)

cqzong@nlpr.ia.ac.cn



<http://www.ia.ac.cn>

2006-11-28

A decorative graphic on the left side of the slide, consisting of overlapping colored squares (yellow, red, blue) and a black crosshair.

Outline

- 1. Overview of NLPR MT System**
- 2. Adaptations to the Baseline System**
- 3. Experiments**
- 4. Future work**

A decorative graphic on the left side of the slide consists of overlapping yellow, red, and blue squares with a black crosshair.

1. Overview of NLPR MT System

A decorative graphic consisting of overlapping colored squares (yellow, red, blue) and a black crosshair.

1.1 NLPR MT System

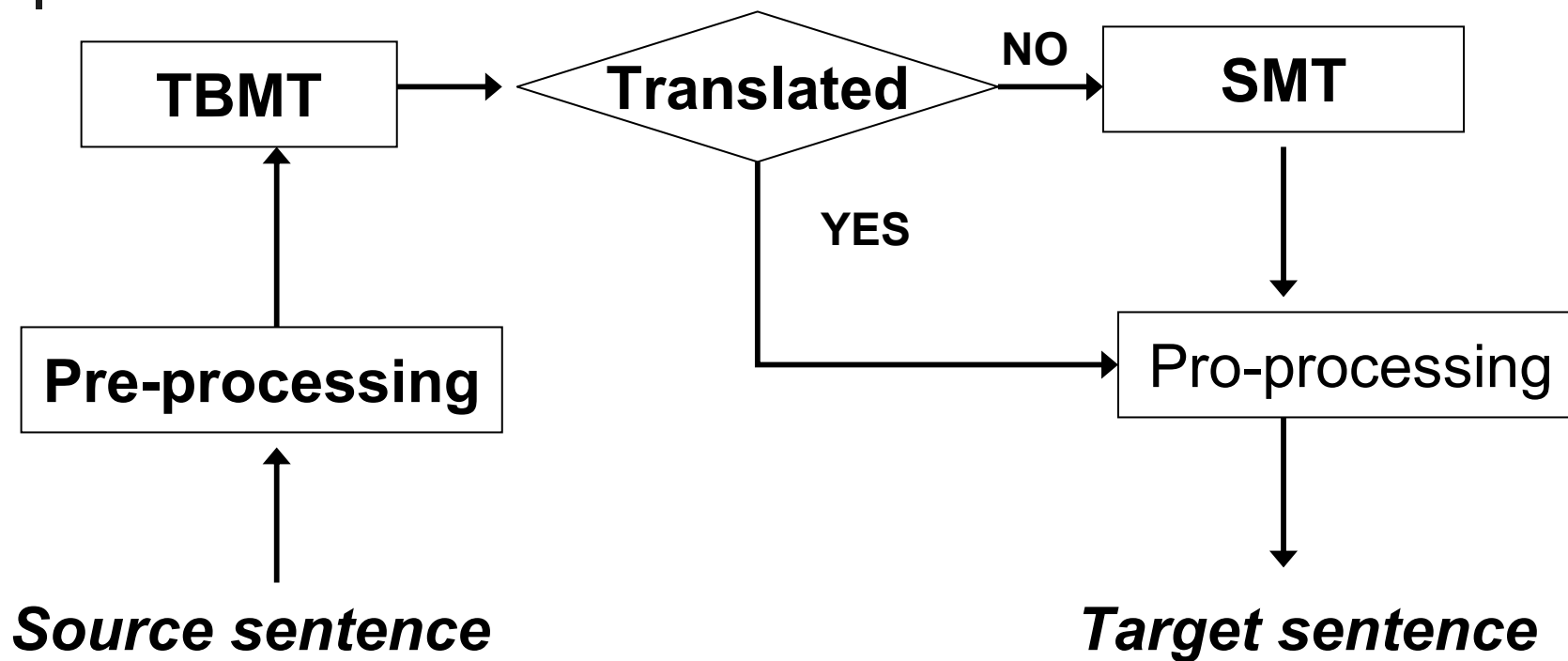
- Chinese-to-English
- A hybrid approach
 - Template based machine translation
 - Phrase based machine translation
- Some improvements
 - Restore punctuation information
 - Number pre-processing

A decorative graphic consisting of overlapping colored squares (yellow, red, blue) and a black crosshair.

1.1 NLPR MT System

- Motivations
 - How many complete correct translations from TBMT engine even the BLEU score is very low? And how many complete correct translations from the phrase-based MT engine even the BLEU score is higher than that of TBMT engine?

1.1 NLPR MT System



System architecture

1.2 Template-based MT

- Template format:

$$C_1 C_2 \dots C_n \Rightarrow T$$

- C_i ($n \geq i \geq 1$) is a component which expresses a condition that the input utterance has to meet.

T is the output result corresponding to the input.

It means if an input utterance of source language meets the conditions $C_1 C_2 \dots C_n$, the input will be translated into the target language expression T .



1.2 Template based MT

- The template is expressed by
 - (1) Keywords
 - (2) POS or POS with Semantic Features
Such as N, V, N(nso), QP,
 - (3) * Variable
It means any word or phrase may appear at the * position or nothing appears.
 - (4) Logical expression of candidate components
(See *Proc. ICSLP'2000*)

1.3 Phrase-based SMT model

- The log-linear model

$$p(e | c) = p_T(c | e)^{\lambda_t} \times p_L(e)^{\lambda_l} \times p_D(e, c)^{\lambda_d}$$

$P_T(c | e)$ is the translation model

$P_L(e)$ is the target language model

$P_D(e, c)$ is the distortion model and

$$P_D(e, c) = \lambda | a_i - b_{i-1} - 1 |$$

1.3 Phrase-based SMT model

◆ Phrase extraction

- Refined model from bi-directional word-based alignment (Och 2002, Koehn *et al.*, 2003)

◆ Phrase translation probability

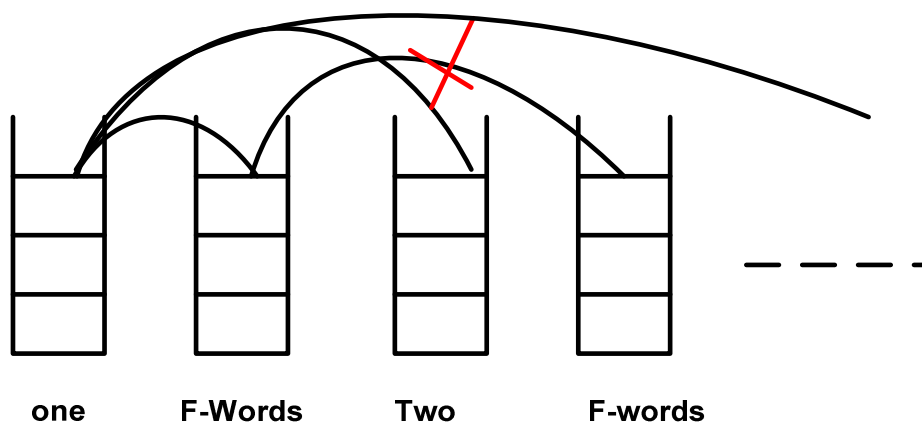
$$P_T(\bar{f}_k | \bar{e}_k) = \lambda_1 P_{lex}(\bar{f}_k | \bar{e}_k) + \lambda_2 P_{lex}(\bar{e}_k | \bar{f}_k) \\ + \lambda_3 P_{freq}(\bar{f}_k | \bar{e}_k) + \lambda_4 P_{lex}(\bar{e}_k | \bar{f}_k)$$

P_{lex} is the lexical probability based on IBM model-4

P_{freq} is the probability based on phrase frequency.

1.4 Beam-Search Decoding

- The basic algorithm is Beam-Search, same as Pharaoh. (Och & Ney, 2002; Zens *et al*, 2002)
- Improvement of integrating F-zerowords bins: the hypotheses are stored in different stacks; bin label: number of Chinese words covered; insert the Functional Words(of , the, for.....) bins.





1.4 Beam-Search Decoding

- *Considering there are some auxiliary words and mood words in the Chinese sentence, and these words sometimes don't have the corresponding words in the English sentence. So, the system does not require the all words in the Chinese sentence to be translated.*

The system selects the candidate sentences generated after $(L - a)$ Chinese words have been translated.

Where, L is the length of the input Chinese sentence; a is an integer.

1.4 Beam-Search Decoding

For example:

$L = 7$ words

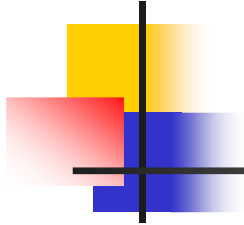
input: 我想预订一个单人间
(I want to reserve a single room)

if $a = 1$, *input* = 我想预订一个 → I would like to book one

if $a = 2$, *input* = 我想预订一 → I want to reserve one

... ..

I want to book one
I would like to reserve one
... ..



2. Adaptations to the Baseline System



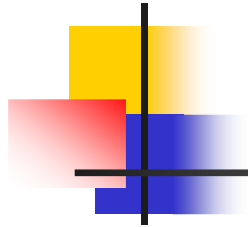
2.1 Restore Punctuations

- Restore punctuations
 - Using the *hidden n-gram* in the SRILM toolkit
 - In the pre-processing
 - to use the punctuation information in training data
 - In the post-processing
 - to restore the more punctuation information



2.2 Number pre-processing

- Considering the different characteristics of numerical expressions in the Chinese and English, we build the number pre-processing module based on rules that are summarized from the selected corpus covered with all domains (about 5,000 sentence pairs)
 - Numeral translation
 - Temporal translation



3. Experiments

3.1 Results of Improvements

- ◆ Training data: 40K sentences
- ◆ Test data: devset4_IWSLT06 (correct,489)

	BLEU	NIST
SMT (with punc in training set)	0.1486	5.4866
SMT+Deletion punctuation	0.1195	5.3029
SMT+Punctuation (test set)	0.1700	5.7994
SMT+Punctuation+TBMT	0.2238	6.1305



3.1 Results of Improvements

- 55 (in 489) sentences translated by TBMT
- An example:

Chinese sentence:

“我 喜欢 红色 或者 黑色 都 可以 。”

TBMT result:

“I would like either red or black .”

SMT result:

“I'd like a red , black be all right .”

3.2 Results of Test Set

		BLEU	NIST
Read speech	SMT	0.0972	3.5557
	SMT+TBMT(33)	0.1037	3.6384
Spontaneous speech	SMT	0.1001	3.4869
	SMT+TBMT(29)	0.1070	3.5755
Correct results	SMT	0.1167	3.9288
	SMT+TBMT(48)	0.1284	4.0658

The red number is the number of sentences translated by TBMT.

A decorative graphic on the left side of the slide consists of overlapping yellow, red, and blue squares with a black crosshair.

3.2 Results of Test Set

- The errors in ASR results
- Lack of punctuation information
- Only use the 1-Best of ASR results
- Set the parameters by hand



3.3 The Recent Experiments

Training the parameters of SMT by Minimum Error Rate criterion.

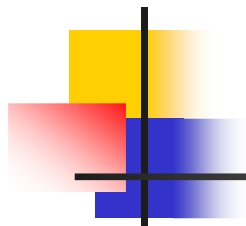
	BLEU	NIST
Manual	0.1167	3.9288
MER training	0.1252	4.1439

SMT results of correct test data

A decorative graphic consisting of overlapping yellow, red, and blue squares with a black crosshair.

4. The Future Work

- Collecting some ASR results
- Dealing with the noisy words in ASR results
- Approaches to word alignment and phrase extraction for SLT
- Try to use the phrasal structural information in TM



Thanks

谢谢!