Multi-Engine Based Chinese-to-English Translation System

Chengqing ZONG

National Laboratory of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences

cqzong@nlpr.ia.ac.cn

No.95, Zhongguancun East Road
Beijing 100080, China

http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

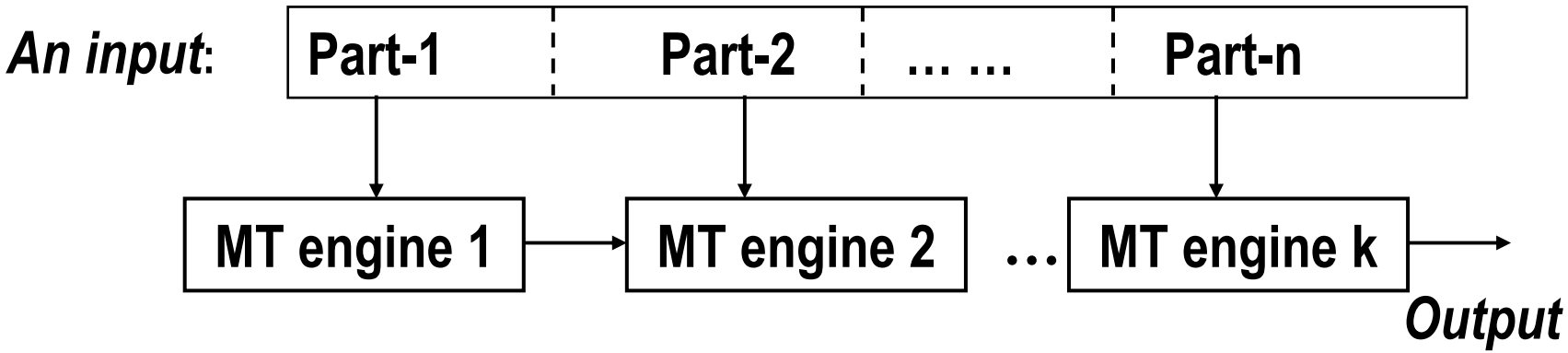
Outline

1. Introduction
2. Overview of our system
3. Experiments
4. Conclusion

1. Introduction

❑ Multi-engine based translation approach has been proposed and practiced.

➢ Integration with cross processing

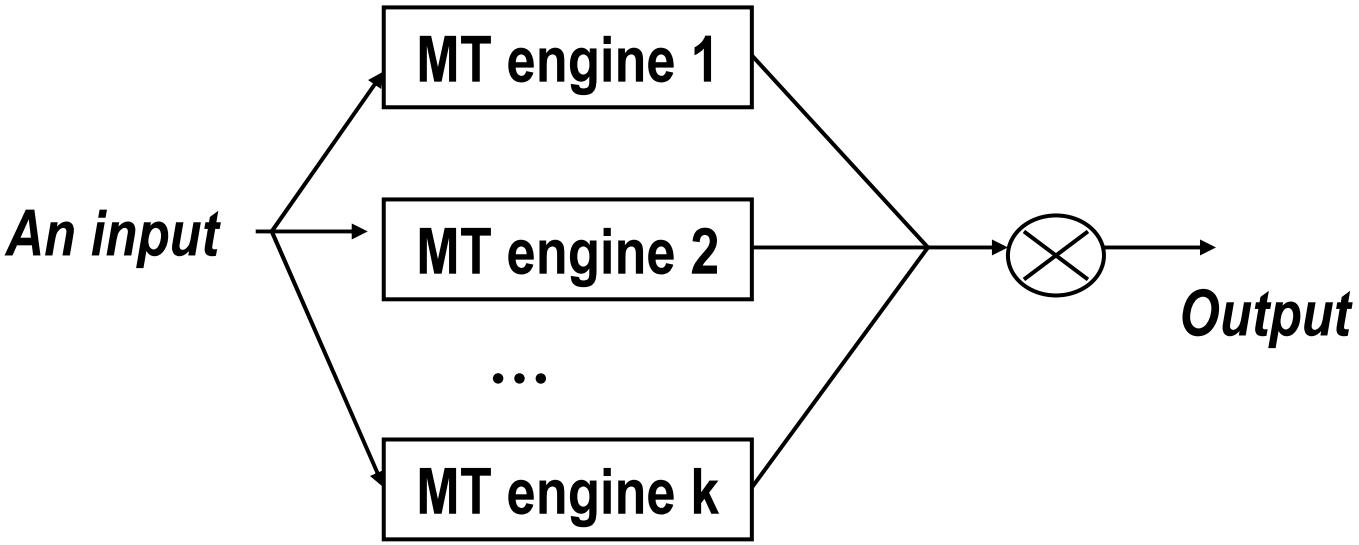


1. Introduction

- The different engine is closely dependent with each other.
- It is sometimes difficult to determine what translation engine works for a specific part of the input.

1. Introduction

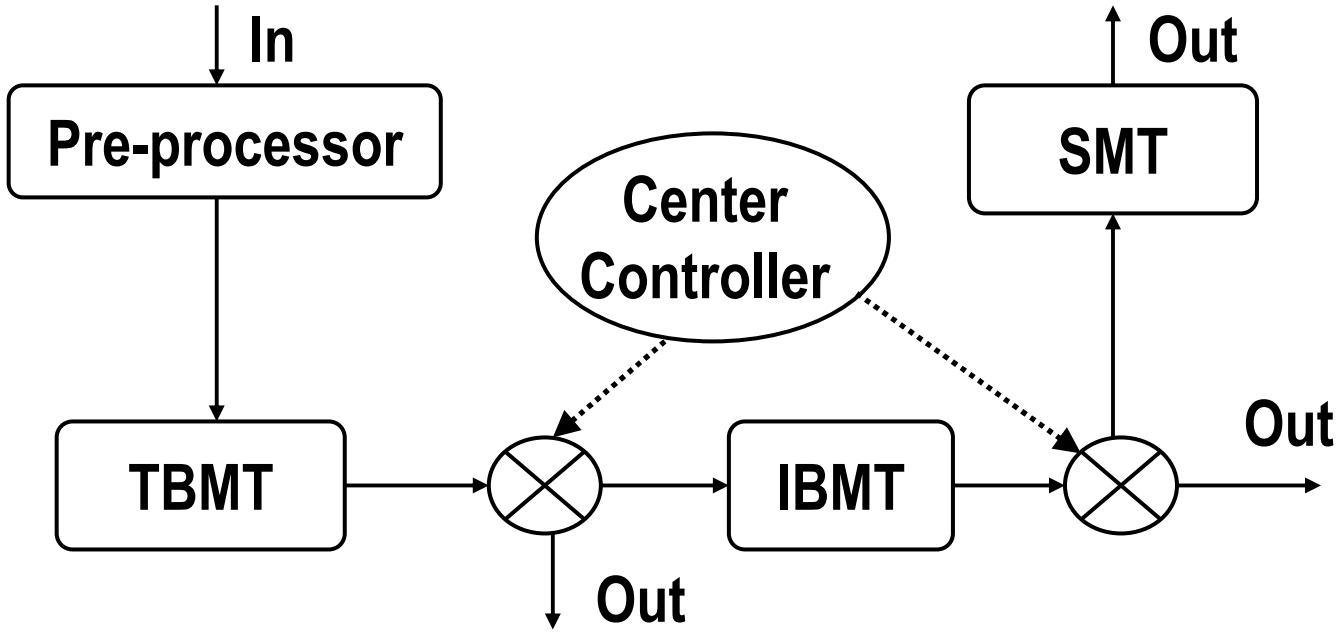
- Multiple engines work in parallel competitive mode



1. Introduction

- The different engine is independent with each other.
- The selector needs the effective function to select the real best result from all translations of different MT engines.

2. Overview of Our System



TBMT: Template Based Machine Translator

IBMT: Inter-lingua Based Machine Translator

SMT: Statistical Machine Translator

2. Overview of Our System

The center controller controls the flow of translation:

1) If the input sentence is translated by TBMT, the system outputs the result and ends the translation of the input sentence; Otherwise, the system performs 2);

2) If the input sentence is translated by IBMT, the system outputs the result of IBMT and ends the translation work; Otherwise, performs 3);

3) The translation of SMT is sent out as the system result, and system is ended.

2. Overview of Our System

Why we use this simple integration approach?

- We know that the performance of the competitive mode is better than the performance of the best MT engine among multiple engines. (*See Akiba et. al 2002; Stephan Vogel's work*)
- We just implemented the chunk-based Chinese-to-English statistical translator when we took part in the evaluation, and we already have the template-based translation engine and the interlingua-based translation engine. We did not have more time to do more about the system integration, and in some sense, we wanted to probe into a simple way to integrate the multi-engine based SLT system.

2. Overview of Our System

2.1 Pre-processor

Main tasks:

1) To delete all repeated words except some special adverbs like "非常（very）", "十分（very）" etc.

2) To recognize and analyze the numerals and numeral phrases (QP) in the input sentences and translate the Chinese numerals into Arabic numerals.

3) To recognize the time words and time phrases (TP) and translate them into English expression.

2. Overview of Our System

2.2 Template based Translator

❑ **Template format:**

$$C_1 \ C_2 \ \ldots C_n \Rightarrow T_l$$

- **C_i ($n \geq i \geq 1$) is a component which expresses a condition that the input utterance of source language has to meet.**
- **T_l is the output result corresponding to the input.**

It means if an input utterance of source language meets the conditions C_1 C_2 … C_n, the input will be translated into the target language expression T_l.

2. Overview of Our System

2.2 Template based Translator

- **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*)

2. Overview of Our System

2.3 Inter-lingua Based Translator

Components:

- ❑ **Interchange Format, developed by C-STAR**
- ❑ **Chinese Parser, parsing the Chinese sentence into IF expression using HMM based model in combination with rule-based approach**
- ❑ **English generator, generating the English sentence based on IF expression**

2. Overview of Our System

2.3 Inter-lingua Based Translator

- **Interchange Format**

Speaker: *Speech-act* (+ *Concept*)* [*Arguments*]

Speaker: a indicating the agent and *c* indicating the client.

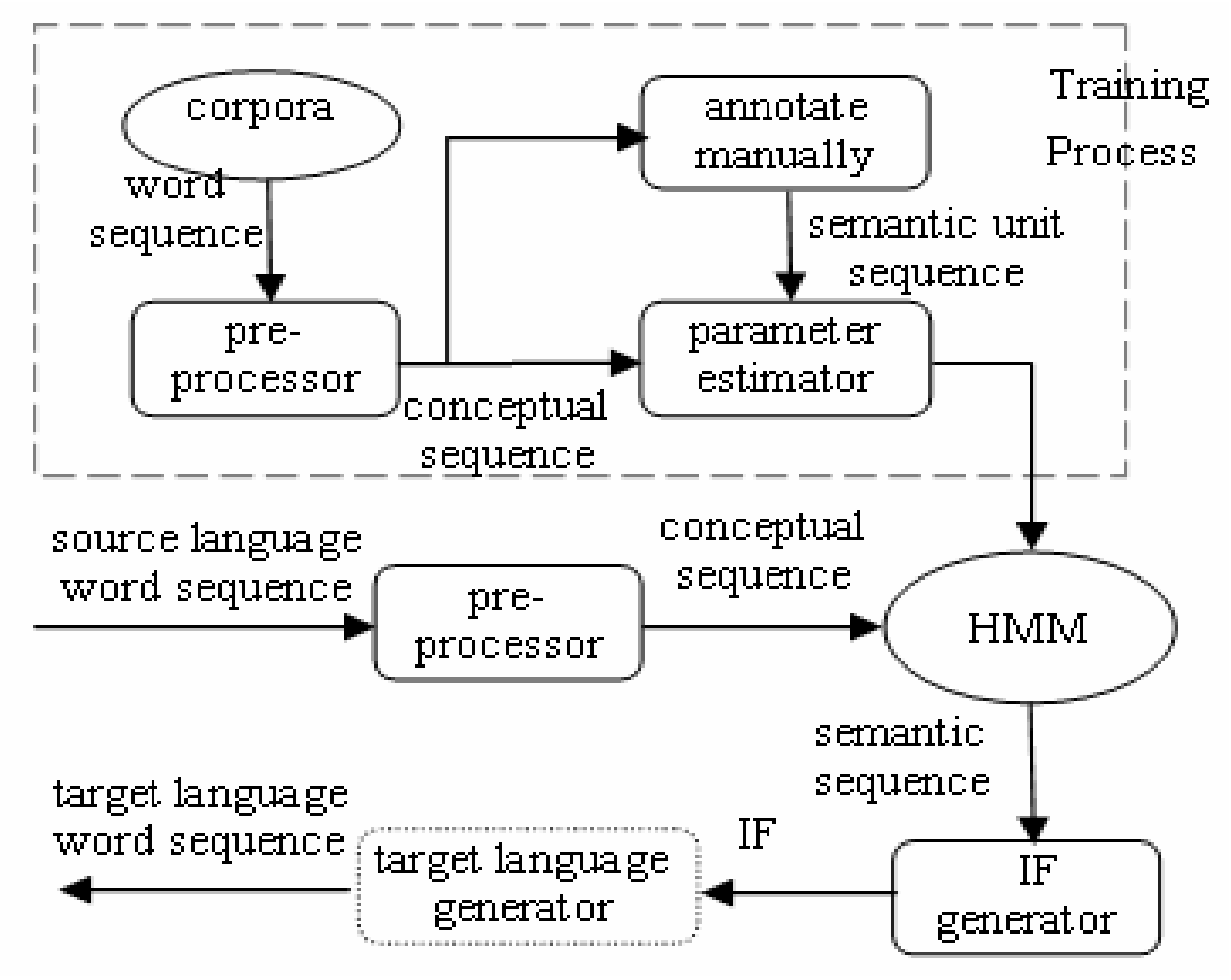
Speech-act: denotes the speaker's intentions that are expressed by a set of definitions.

Concept: is defined to restrict and describe the *Speech-act* in a specific domain (hotel reservation).

Argument is a list of n ($n \geq 1$) arguments that are expressed by pairs of parameter variables and their corresponding attribute values.

2. Overview of Our System

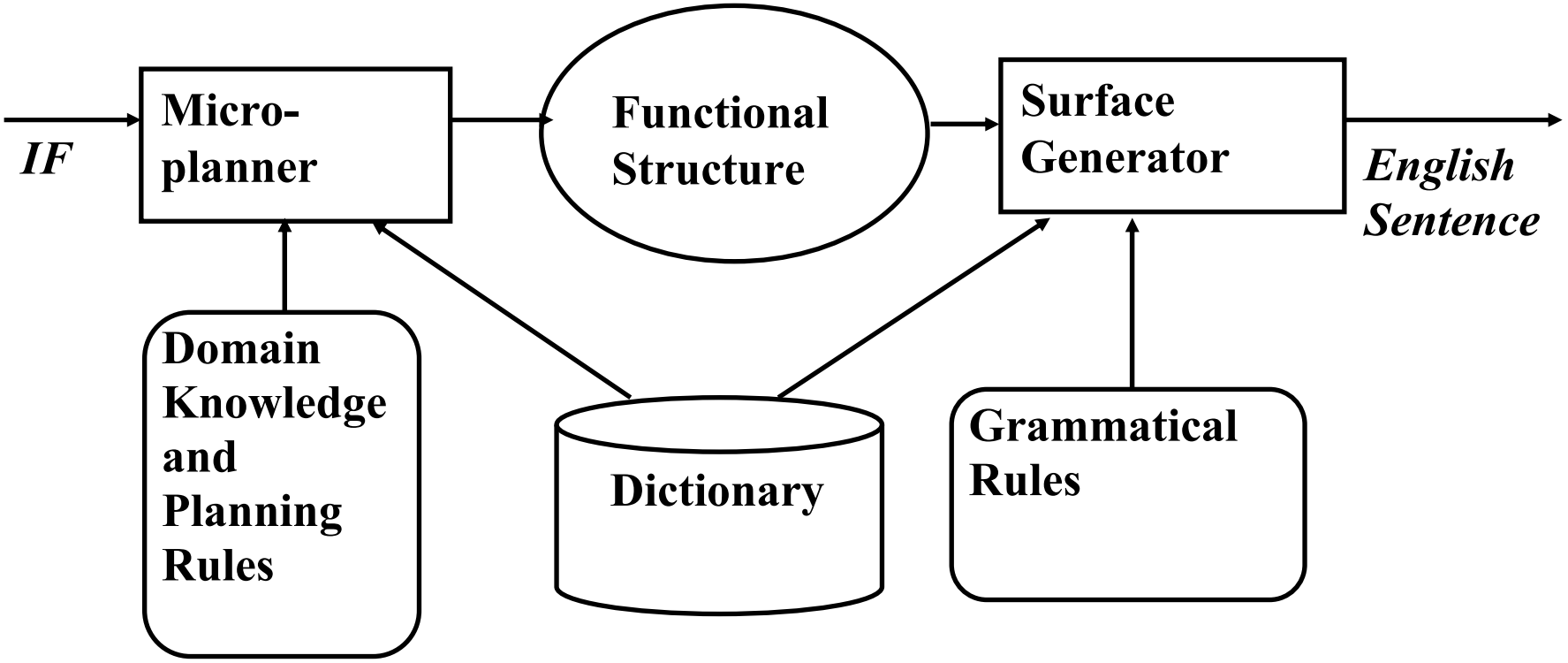
➢ **Chinese Parser**



See *Proc. ICSLP'2002*

2. Overview of Our System

- **English Generator**



See *Proc. ICSLP'2004*

2. Overview of Our System

The Chinese parser and generator have been integrated with the components developed by UKA and CMU, and the integration Chinese-English SLT system was successfully demonstrated in 2004 Beijing International Exhibition of Sci. & Tech., May 22 – 26, and in 2004 Barcelona International Culture Forum, July 16 –18.

2.4 Statistical Translator

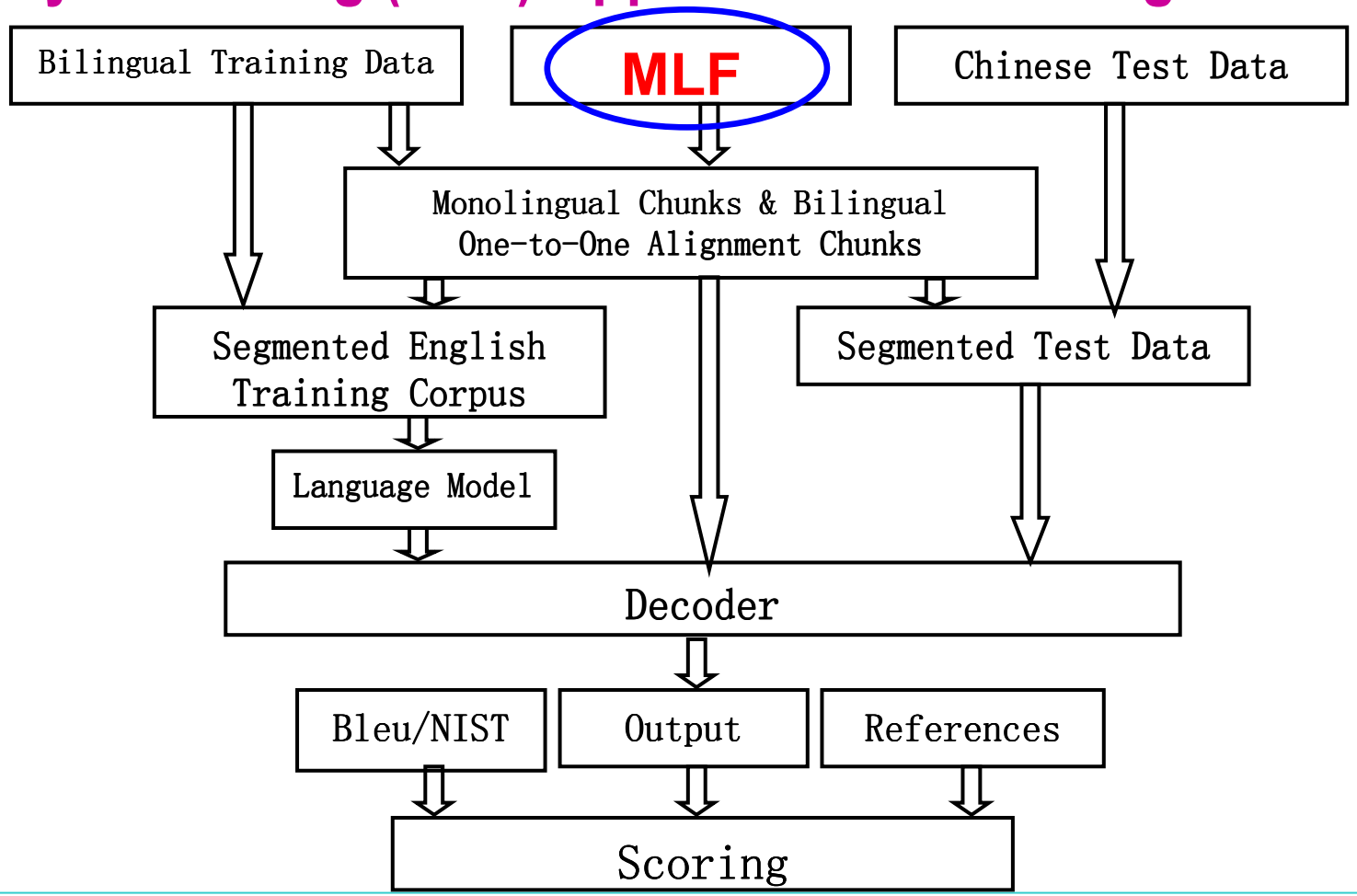
- **The basic SMT model has been proposed by IBM**
  - **Language Model (LM)** : ***P(S)***
  - **Translation Model (TM)** : ***P(T/S)***

$$S = \arg \max_{s} P(S)P(T \mid S)$$

- the LM is used to assign a probability to any English String (ES) (often seen as representing its fluency)
- the TM is used to assign a probability to any pair of English and Chinese string (CS) (often used to reflect the fidelity of ES to CS)
- Our system is based on chunk-to-chunk translations extracted from a bilingual corpus.

2.4 Statistical Translator

Multi-layer Filtering (MLF) Approach to chunk segmentation.



MLF: The First Layer
Filtering out the most frequent chunks

- **The most co-occurrent word lists may be a potential chunk. These word lists are first filtered out as initial monolingual chunks. e. g.,**

请给我来些**加糖的咖啡

*please give me some** coffee with sugar*

- **A lot of word lists like this, for example：**

I want to reserve {a, one, two…} {single, double, standard…} {room(s)}.

=>I want to reserve || 我想预定

MLF: The First Layer

Filtering out the most frequent chunks

$$D(w_1, w_2) = (1-\beta) \times MI(w_1, w_2) + \beta \times P(w_1, w_2)$$

$$n = \text{int}\{\frac{length\ of\ a\ sentence}{the\ maximum\ length\ of\ defined\ chunk}\}$$

β=0.5

- 0.17 0.076 0.87 1.27
- 你 || 想 || 预 || 定 || 什 || 么 || 样 || 的 || 房 || 间
- 0.69 2.39 7.80 0.30 4.52

- 1.31 0.063 0.61 0.077
- what || kind || of || room || do || you || want || to || reserve
- 1.36 0.046 10.07 2.11

MLF: The First Layer

Filtering out the most frequent chunks

Two principles: the maximum matching / no overlapping

$$\mu = D_k / D_{k-1} \qquad \nu = D_{k_i} / D_{k_{i+1}} \qquad D'_k = D_k \times \frac{Max(D_k)}{Max(D_2)}$$

Table 2 English initial chunks & their cohesion degree (10^{-3})

Initial Chunks	D_k	D'_k	Initial Chunks	D_k	D'_k
do you want to	0.13	0.90	you want to reserve	0.086	0.60
what kind of	2.10	5.25	do you want	0.31	0.77
you want to	0.33	0.82	want to reserve	0.056	0.14
what kind	1.36	1.36	kind of	1.31	1.31
do you	10.07	10.07	you want	0.61	0.61
want to	2.11	2.11	to reserve	0.077	0.077

Table 1 Chinese initial chunks & their cohesion degree (10^{-3})

Initial Chunks	D_k	D'_k	Initial Chunks	D_k	D'_k
什么样的	0.58	2.44	么样的房	0.13	0.55
样的房间	0.21	0.88	什么样	0.44	1.00
么样的	0.37	0.84	样的房	0.13	0.30
的房间	2.45	5.88	你想	0.69	0.69
预定	2.39	2.39	什么	7.80	7.80
么样	0.87	0.87	样的	0.30	0.30
的房	1.27	1.27	房间	4.52	4.52

你 || 想 || 预定 || 什么样的 || 房间 What&kind&of || room || do&you || want&to || reserve

MLF: The Second Layer

Filtering out the most frequent structures

- Observing the corpus after the first filtering step, we have found that there are many frequent structures distributed throughout it which are similar on inspection but different in detail:

  at five o'clock => *在五点钟*

  at six o'clock => 在六点钟

- These structures may include word sequences with low frequency of occurrence (like "*five*" and "*six*"). Similar issues often arise in other structures, e. g., the potentially good chunks "*a single room*" has been broken into several fragments ("*a*", "*single*" and "*room*") after the first filtering process. This segmentation would bring problems for the subsequent alignment process.

MLF: The Second Layer

Filtering out the most frequent structures

- We cluster similar words together (like numbers) according to the position vectors of their behavior relative to the anchor words.

MLF: The Second Layer

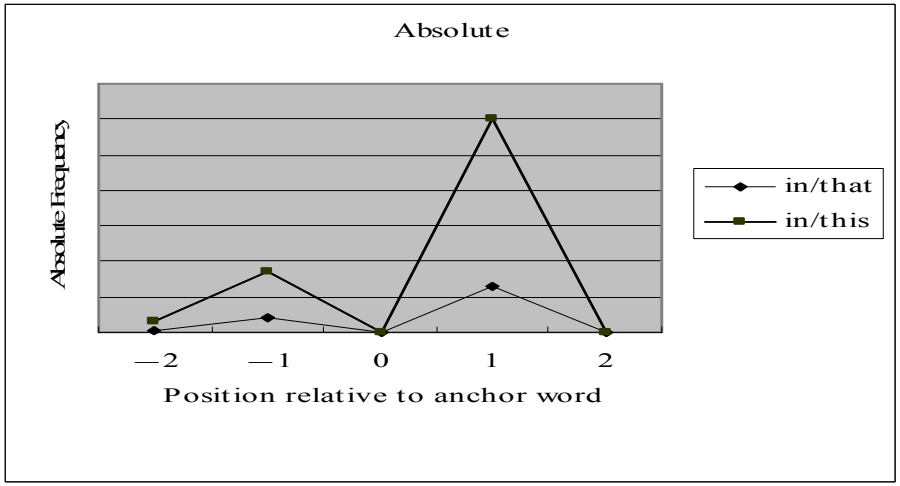
Filtering out the most frequent structures

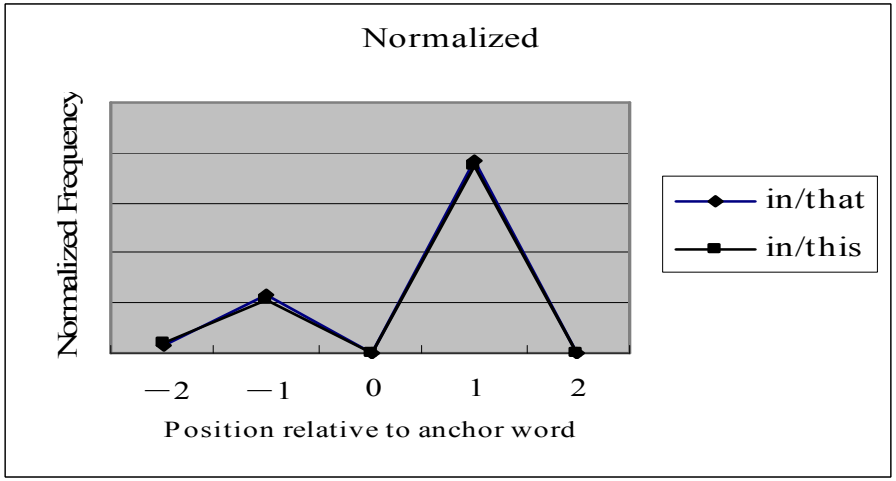
$$V_{ij} = \sum_{k=1}^{N} \delta(w_j, w)$$

$$\delta(w_j, w) = \begin{cases} 1 & w_j = w \\ 0 & w_j \neq w \end{cases}$$

$$V_{ij}' = V_{ij} / \sum_{j=1}^{m} V_{ij}$$

$$D(V_X, V_Y) = \sqrt{\sum_{j=1}^{K} (V_{xj} - V_{yj})^2}$$





MLF: The Second Layer

Filtering out the most frequent structures

- All of the words in the same class are substituted with a particular symbol, and the symbol is treated as an ordinary word. Then the most frequent structures are filtered according to the method used in the First Layer process.

MLF: The Third Layer

Filtering out the rest fragments

- The lengths of the sentences are quite short after the previous filtering steps. Maybe only one or two individual or sequential words remain. We simply combine such word sequences as a chunk, and treat each lone word as a chunk. Finally, these chunks are filtered too in the same way.

Extract the bilingual alignment chunks

$$\theta = \frac{2 \times Num\{Co - occurrence\ (C\_CHK, E\_CHK)\}}{Num(C\_CHK) + Num(E\_CHK)}$$

Table 3 Aligned bilingual chunks (10^{-3})

θ	你	想	预定	什么样的	房间
what kind of	**0.025**	**0.021**	**0.053**	**0.889**	**0.016**
room	**0.021**	**0.029**	**0.09**	**0.0140**	**0.888**
do you	**0.460**	**0.014**	**0.002**	**0.012**	**0.020**
want to	**0.007**	**0.069**	**0.013**	**0.002**	**0.023**
reserve	**0.002**	**0.001**	**0.083**	**0.034**	**0.047**

3. Experiments

- **Training Corpus**
  - 20,000 BTEC sentences
- **Test Corpus**
  - 500 sentences

3. Experiments

- **Translations**
  - 238 (47.6%) sentences are translated by TBMT
  - 237 (47.4%) sentences are translated by SMT
  - 25 (5%) sentences are translated by IBMT

3. Experiments

Table-4. Evaluation of System

Eval / Perfor.	ADEQ	FLUE	WER	PER	GTM
SMT 237	2.5316	3.6123	0.6123	0.5454	0.5546
TBMT 238	3.1343	3.2700	0.5338	0.5007	0.5999
System	2.8000	3.4000	0.5788	0.5310	0.5639

4. Conclusion

- The simple system integration approach is easy to implement.
- Each translation engine works independently. The system is easy to extend, translation substitution, or even transplant.
- The more important is that the new approach to chunk identification based on the multi-layer filtering works very well. This approach doesn't rely on the Chinese word segmention and any other information from tagging, parsing, or syntax analyzing.

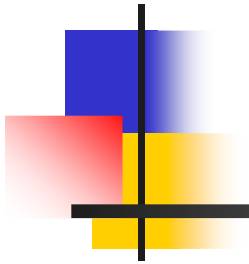
4. Conclusion

❑ Future work

➢ Automatic extracting templates from the training corpus

➢ Further make the definition of bilingual chunks clear to easy compute and operate

➢ Find out the effective method to select the best results from multiple translation engines and search for the more reasonable system integration approach



Thanks

谢谢!