SAS STATISTICAL MACHINE TRANSLATION SYSTEM FOR WAT 2014

Rui Wang, Xu Yang and Yan Gao
SAS Institute Inc, Beijing, China

Ssas | THE POWER TO KNOW.

Introduction

This paper describes the machine translation system employed by SAS Institute Inc in the 1st Workshop on Asian Translation. We participate in two subtasks in this year's WAT:

Chinese to Japanese;

English to Japanese.

The sentence structure of Japanese is different with that of English/Chinese. Japanese is typically a Subject-Object-Verb (SOV) language while Chinese and English are Subject-Verb-Object (SVO) languages, as illustrated in the following Figure.
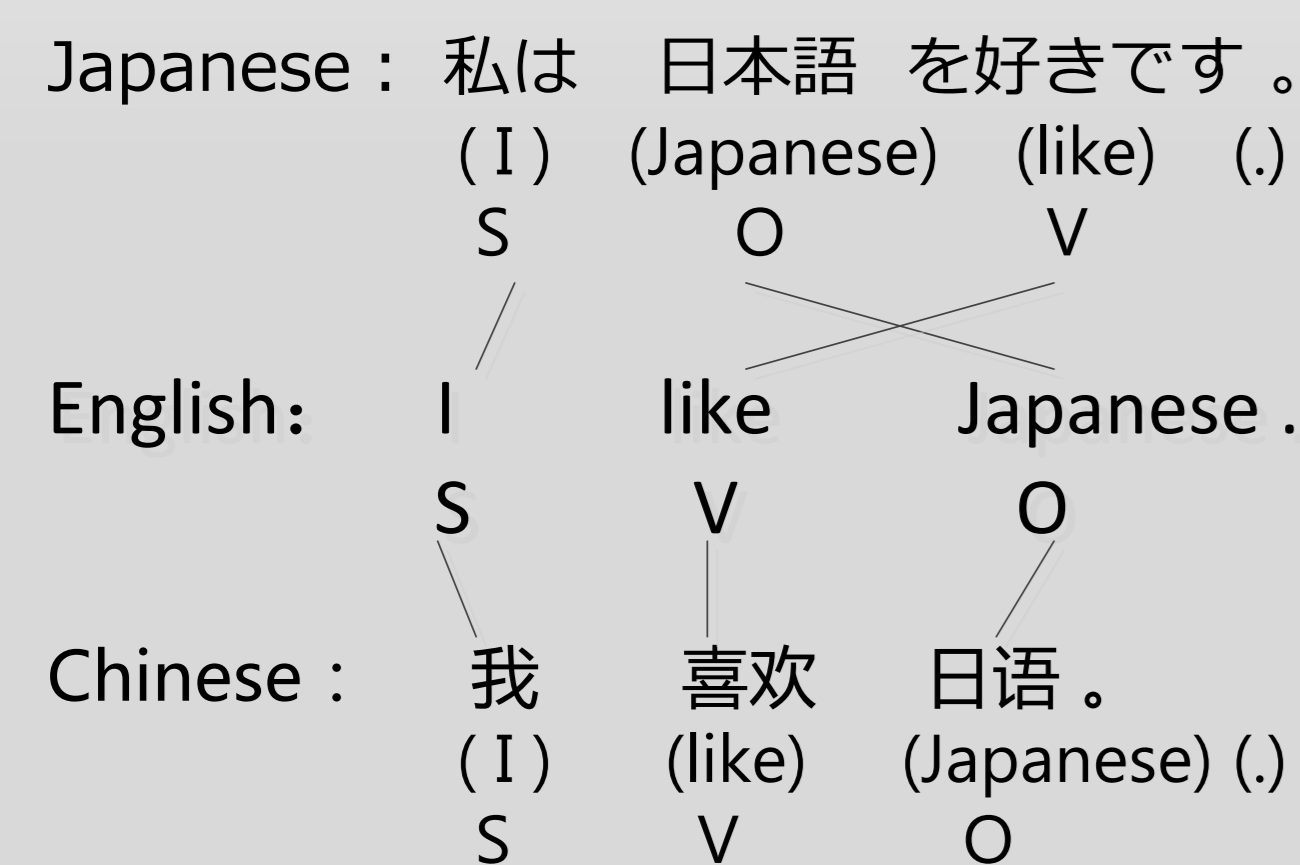
The statistic machine translation between Japanese and the SVO language is particularly difficult because of the long distance difference of word orders. We propose a simple syntactic reordering approach to transform Chinese/English into SVO languages.

In addition, we apply the segmentation tool in SAS® Text Miner to the corpus and obtain improvement of the translation.

Background

Phrase-based model [Koehn, et.al 2003]
- translate phrases as units; "Standard Model" used by Google Translate
- Limited capacity for long distance reordering.

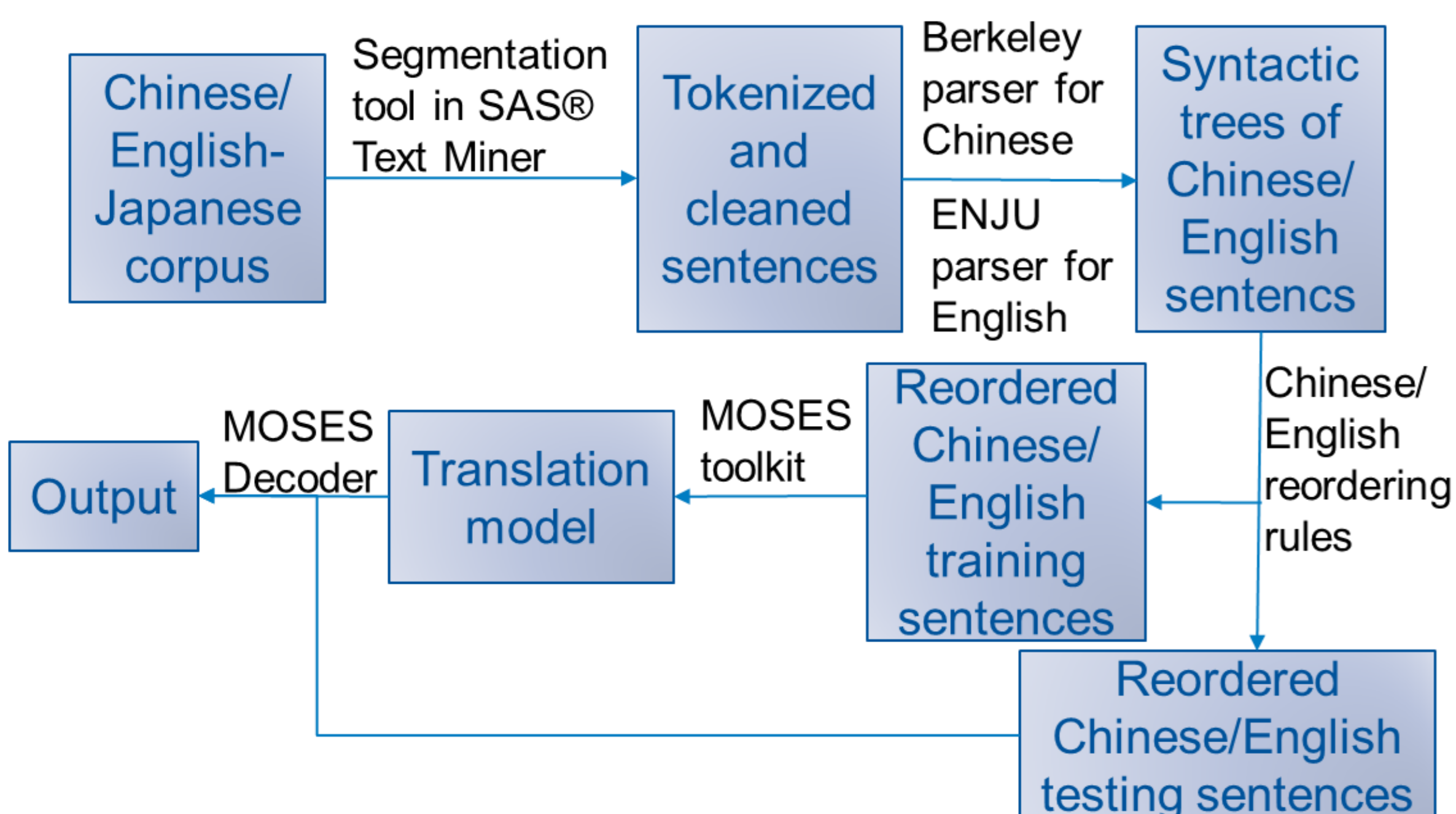Syntax-based models [Liu et al., 2006]

Forest-based models [Mi et al., 2008]
- Improve the translation
- System is complex and time consuming during decoding

Syntactic reordering approaches
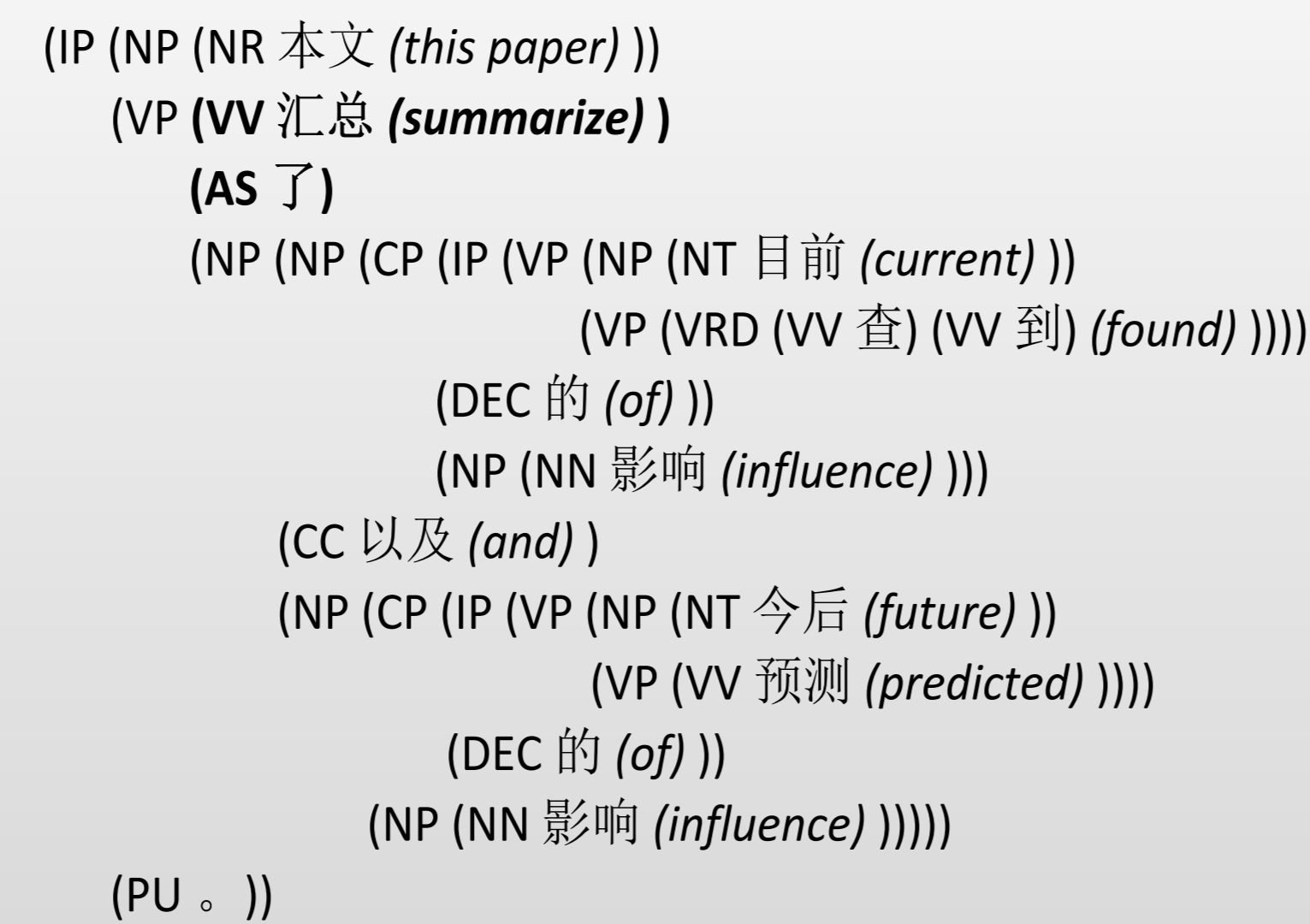- Effectively improve the translation results

System Architecture

Syntactic Reordering Approaches

1. Chinese to Japanese reordering rules

VP-rule: VP (VV AS (XXX)) →VP ((XXX) VV AS)

to move the verb (VV) and the auxiliary word (AS) behind VV to the end of the verb phrase (VP)

(a) Original   (b) Reordered

PP-rule: PP (P (XXX)) → PP ((XXX) P)

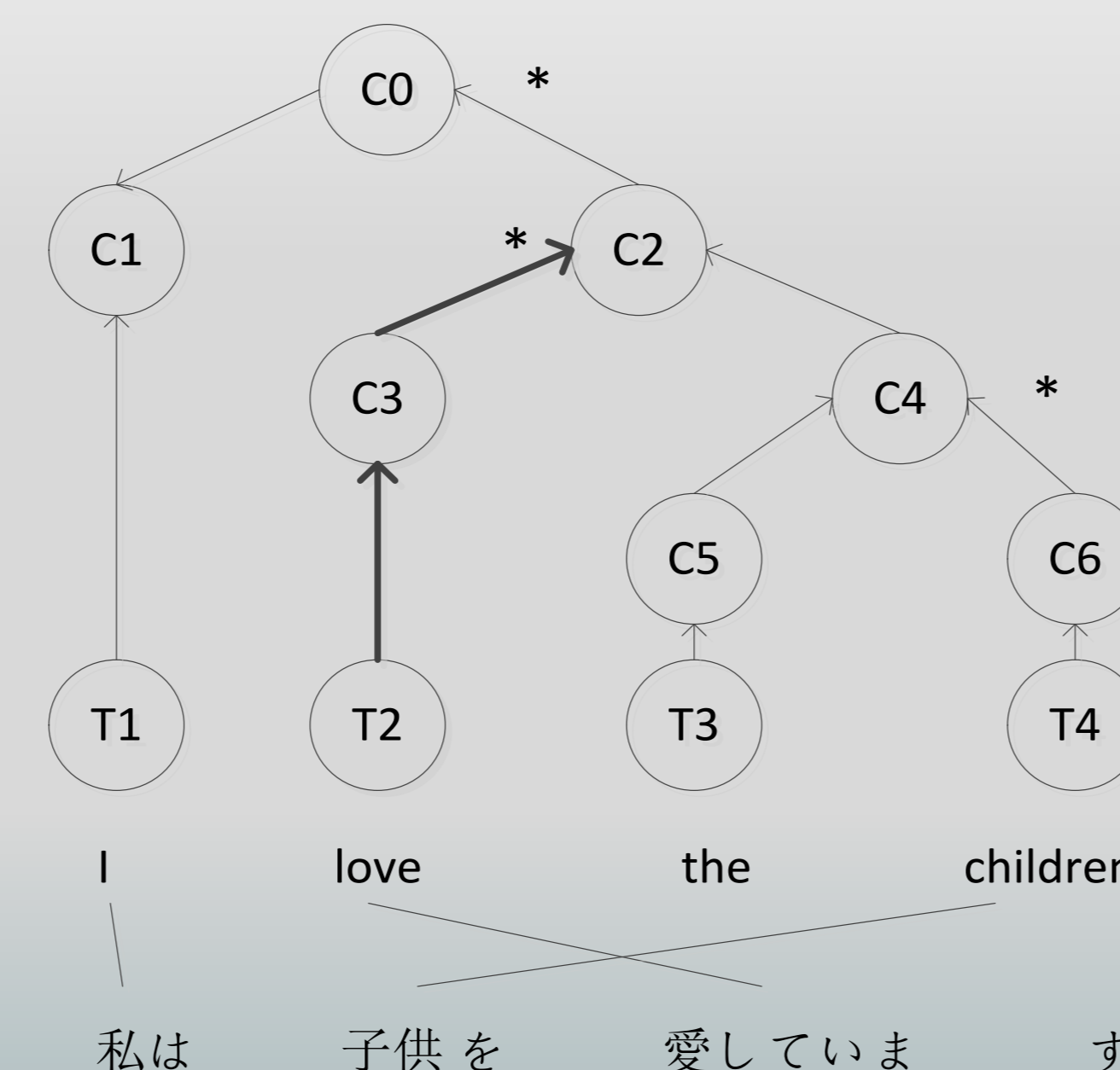to move P to the end of PP .

(a) Original   (b) Reordered

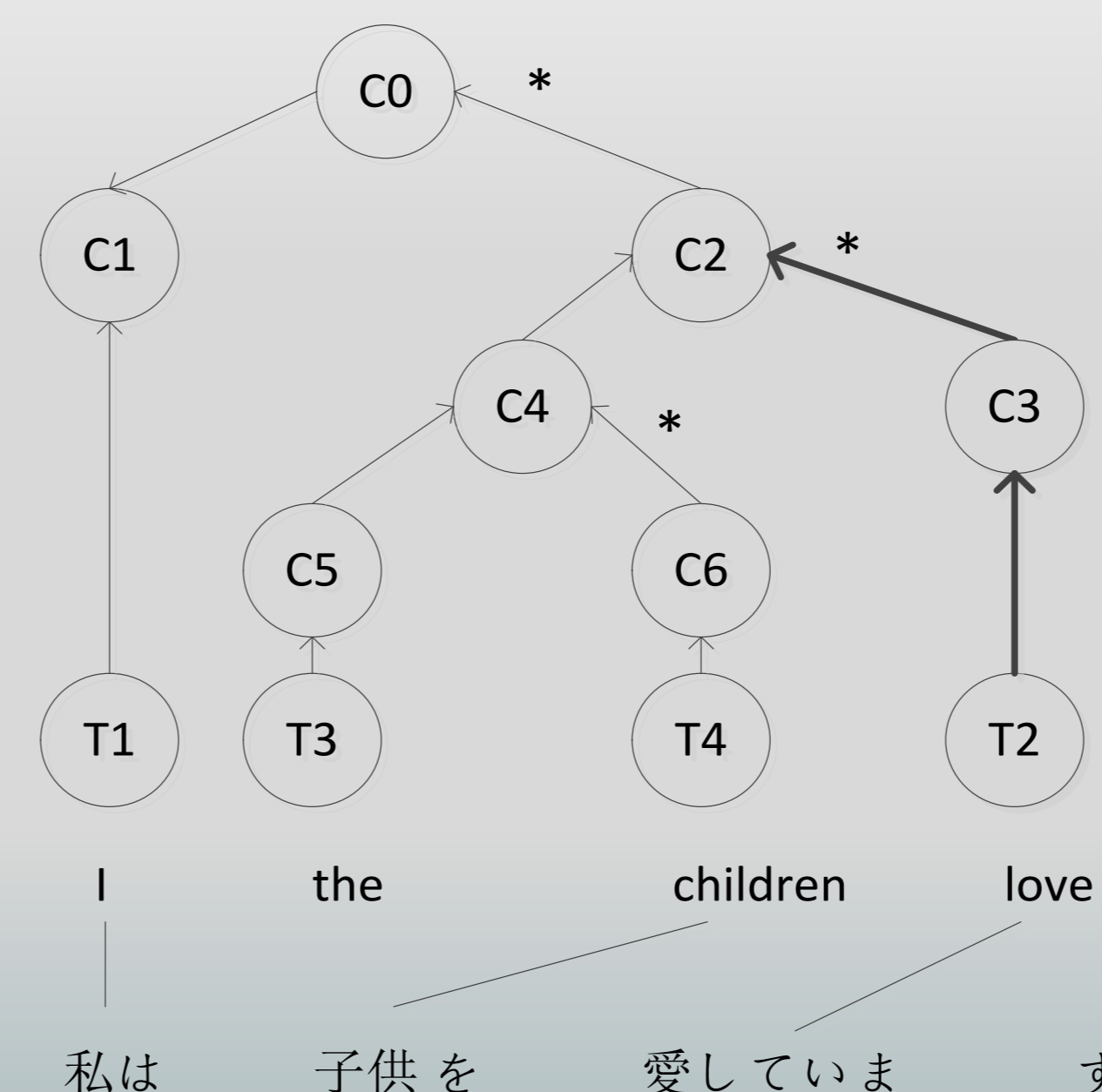### 2. English to Japanese reordering rules

Reordering rule: Head Finalization [Isozaki 2010]

Move syntactic heads to the end of the corresponding syntactic constituents.

Use dependency parser: ENJU Parser (developed by University of Tokyo)

(a) Original   (b) Reordered

Experiments

1. Effect of the segmentation tool of SAS® Text Miner

Baseline (phrase-based model provided by the organizer) :
   Japanese: Juman segmentation tool
   Chinese: Stanford Word Segmenter

SAS segmentation:
   SAS segmentation tool of SAS® Text Miner for Chinese and Japanese.

	BLEU	RIBES
Baseline	34.86	0.769962
SAS segmentation	35.31	0.809631

2. Chinese to Japanese translation

Baseline (phrase-based model provided by the organizer)

We gain 2.07 in BLEU scores compared with the baseline.

	BLEU	RIBES
Baseline	34.86	0.769962
Baseline+VP	36.19	0.826146
Baseline+PP	36.30	0.815694
Baseline+PP+VP	36.40	0.826015
Hierarchical	36.06	0.814207
Hierarchical+PP+VP	37.38	0.830909

3. English to Japanese translation

Baseline (phrase-based model provided by the organizer)

We gain 3.13 in BLEU scores compared with the baseline.

	BLEU	RIBES
Baseline	28.52	0.690350
Baseline+reorder	31.09	0.765005
Hierarchical	31.23	0.743135
Hierarchical+reorder	31.65	0.767323

Conclusion & Future Work

Conclusion:

Introduce the system architecture of SAS at WAT 2014;
- Describe the reordering approaches in detail;
- Show experiments results to illustrate the effect of our system.

Future work:
- Consider Japanese Case Marker in the translation;
- Add more reordering rules on Chinese to Japanese translation;
- Attend the work to English to Chinese translation.