中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

CASIA SMT System for IWSLT'09

Maoxi Li, Jiajun Zhang, Yu Zhou, and Chengqing Zong

Institute of Automation
Chinese Academy of Sciences (CASIA)

Outlines

Introduction

System Architecture

System Implementation

Experimental Results

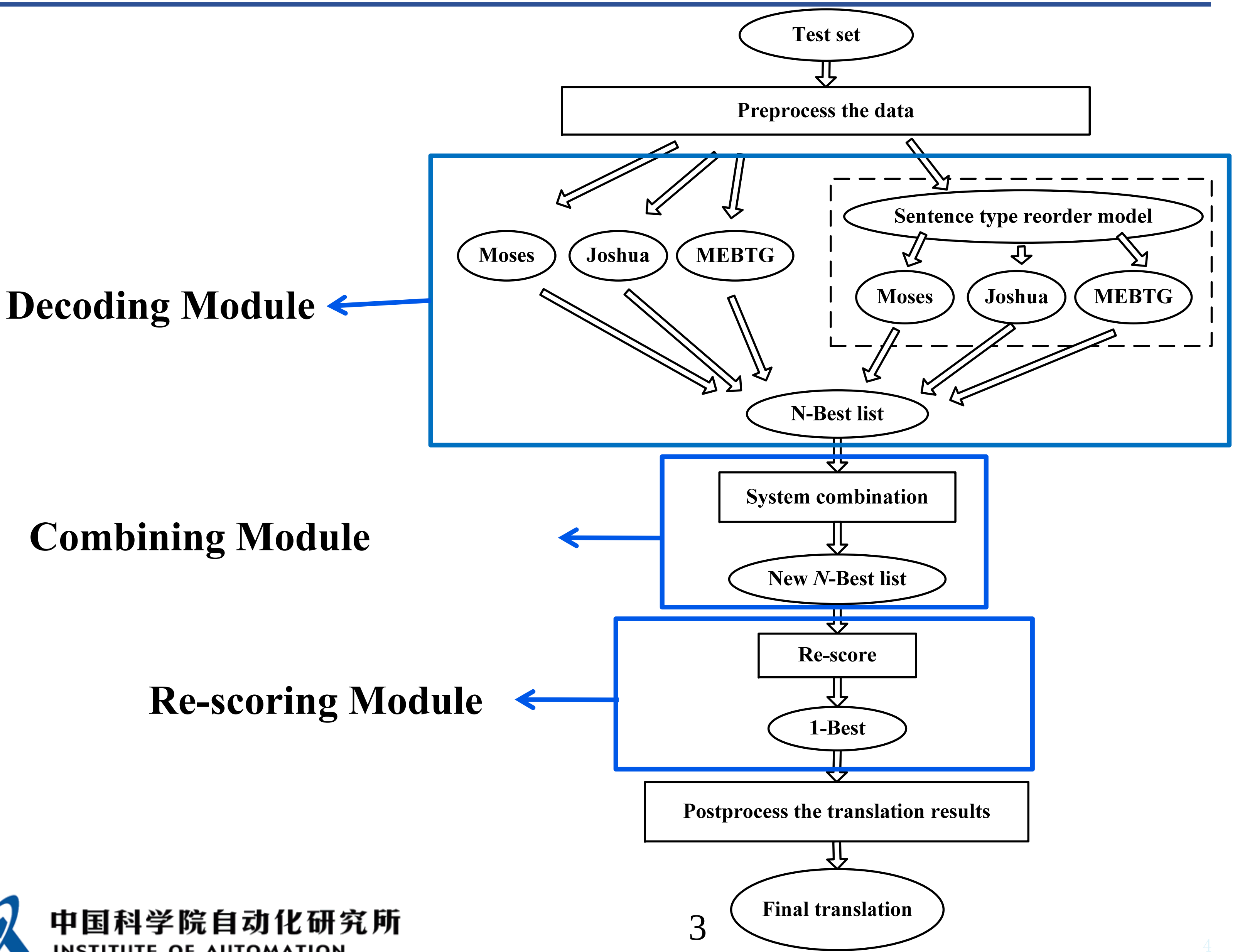
Conclusion

Introduction

1. In IWSLT 2009 Evaluation Campaign, the tasks that we participated in include:

- Challenge translation tasks:
  - ✓Chinese-to-English: CT_CE (CRR and ASR)
  - ✓English-to-Chinese: CT_EC (CRR and ASR)
- BTEC translation tasks:
  - ✓Chinese-to-English: BTEC_CE

System Architecture



中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

System Implementation

--Prepare the data 1

1 An important difference

Training corpus are limited to the released corpus for each translation task.

2 Chinese word segmentation

- ORI: Original Chinese word segmentation
- ICT: the free software toolkit ICTCLAS3.0 *(http://www.nlp.org.cn)*

Approaches	ORI	ICT	ORI+ICT
BLEU	35.31	36.24	36.63

Table1. The performances with different Chinese word segmentation approaches

System Implementation

--Prepare the data 2

3. English word Lowercased and tokenized

- A word in different positions of a sentence may have different morphology.
- To avoid data sparse, we use the lowercased and tokenized scripts of the open source toolkit Moses to do this job.

System Implementation

--Prepare the data 3

4. Named Entities process

- A hybrid named entity recognizer to identify Chinese NEs.
- Person names and location names are translated word by word.
- Organization names are translated by a structure-based translation model.
- A rule-based approach to translate the temporal and numerical NEs.

System Implementation

--Decoding module 1, three original SMT decoders

- Moses
- Joshua
- MEBTG:
  - ✓ An in-home maximum entropy-based reordering model decoder.
  - ✓ The prediction of relative orders of any two adjacent blocks is considered as a problem of classification.
  - ✓ A MaxEnt classifier is trained according to the training data.
  - ✓ A CKY algorithm is exploited to decode the test set.
  - ✓ We limits the phrase table within 40 and the partial hypotheses within 200.

System Implementation

--Decoding module 2, three deformed SMT decoders

- For Chinese to English translation, a preprocessing module, namely **Bandore**, reorders the Chinese sentences before decoding.
  - ✓ An SVM is used to classify Chinese sentences into three types exploiting all the words occurring in the sentence as features.
  - ✓ Corresponding reordering model is developed for specific sentence types.
  - ✓ Reordering the Chinese sentences of training set and test set.
  - ✓ Pass the reordered sentences into the original SMT decoders.
- We called them: Moses-Reorder, Joshua-Reorder and MEBTG-Reorder.

System Implementation

--Decoding module 3, SMT decoders setting

1 Decoder version selecting:

- ✓ Joshua 1.1 – the only version at that time.
- ✓ Moses 2009-04-13.

Tasks	Version 2009-04-13	Version 2008-07-11
CT_CE	35.64	35.37
CT_EC	33.70	33.53

Table 2. The performance on the Challenge CRR tasks with different Moses version

2 Decoder option: closest or shortest ?

Tasks	closest	shortest
CT_CE	36.56	35.64

Table 3. The performance on the development set with different Moses tuning option

Tasks	closest	shortest
CT_CE	38.00	36.58
CT_EC	31.96	31.03
BTEC	47.05	45.66

Table 4. The performance on the development set with different Joshua tuning option

System Implementation

--Combining module, a word-level system combination approach

- Our approach is similar to A.-V. I. Rosti etc. presented in ACL 2007.

- We improve the system combination performance by substituting a word reordering alignment (WRA) for alignment produced by TER.

- The 10-Best lists are used for system combination.

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

System Implementation

--Re-scoring Module

- The re-scoring method that we used this year is the same as last year.

- We merge the 100-Best hypotheses produced by the combining module and all the original 10-Best hypotheses generated by each single decoder.
  - ✓ Note that the 100-Best hypotheses produced by the combining module might include some original hypotheses, so we delete the repeated ones.

System Implementation

--Post-processing

- For Chinese to English translation tasks:
  - ✓ Re-case: train a re-caser with Moses and re-case the outputs.
  - ✓ De-tokenize: done by the de-tokenizer scripts of Moses package.
  - ✓ Tokenize the final submitted translation with the official tool: "ppEnglish.case+punc.pl" script.

- For English to Chinese translation task, evaluation by Chinese Character.
  - ✓ Transform segmentation into characters by the official tool: "splitUTF8characters.pl" script.

System Implementation

--For CT-CE ASR 1, replace Chinese character with Chinese Pinyin for CT-CE ASR task

- The mistakes made by ASR often focused on:
  - ✓ Homophone Chinese character. Such as "玲(ling2)" or "铃(ling2)"
  - ✓ Different Chinese character pronunciation tone. Such as "**直(zhi2)**" or "**智(zhi4)**" or "**知(zhi1)**".

- The supplied corpus is very small, this often lead to:
  - ✓ A lot of OOV words emerge.
  - ✓ The data sparse is very severe.

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

System Implementation

--For CT-CE ASR 2

- Add Chinese Pinyin for Chinese character in the training data and the test data.

我 的 名字 是 铃木 直子
wo3 de5 ming2 zi4 shi4 **ling2** mu4 **zhi2** zi5

我 的 名字 是 铃木 智子
wo3 de5 ming2 zi4 shi4 ling2 mu4 **zhi4** zi5

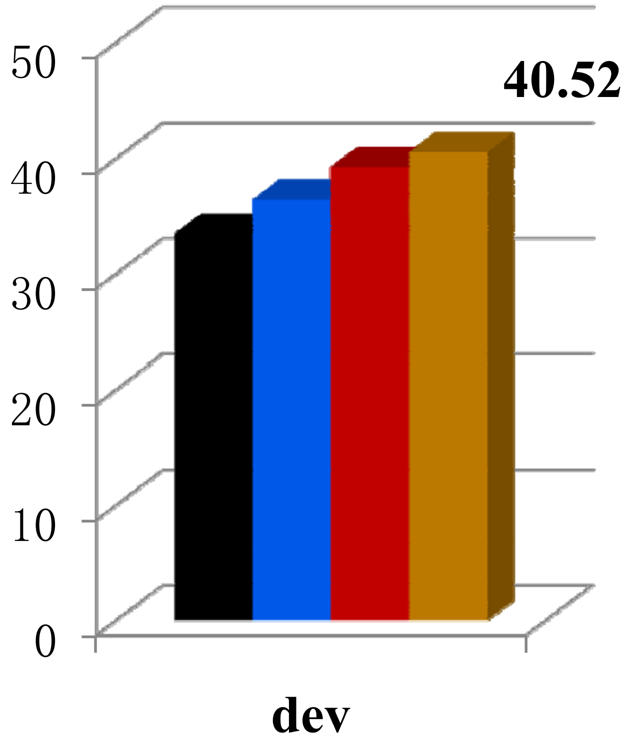
我 的 名字 是 铃木 知子
wo3 de5 ming2 zi4 shi4 ling2 mu4 **zhi1** zi5

我 的 名字 是 玲木 智子
wo3 de5 ming2 zi4 shi4 **ling2** mu4 **zhi4** zi5

- Use Chinese Pinyin to train the model.

System Implementation

--For CT-CE ASR 3, performance on the development set



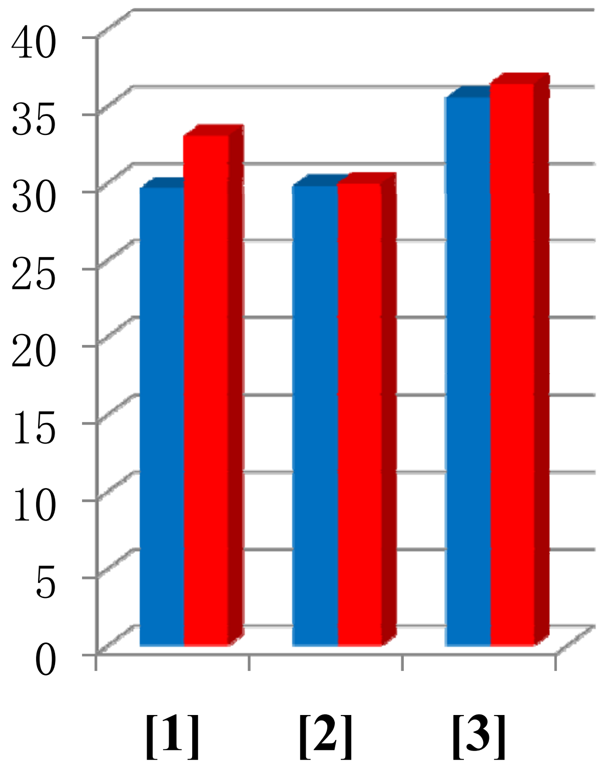
CT-CE-ASR	DEV	
ASR	33.48	
ASR-Pinyin	36.43	↑ 2.95

CT-CE-CRR	DEV	
CRR	39.24	
CRR-Pinyin	40.52	↑ 1.28

Table 5. The translation performance of substituting Chinese Pinyin for Chinese character on the DEV9 for CT-CE task

System Implementation

--For CT-CE ASR 4, performance on the test set



	ASR-Pinyin	CRR
no_case + no_punc [1]	29.65	33.04
case + punc [2]	29.81	29.95
Primary case + punc [3]	35.52	36.44

Table 6. The translation performance of substituting Chinese Pinyin for Chinese character on the test set for CT-CE task

Experimental Results

-- Corpus statistics 1

- When tuning the parameter:
  - ✓ We merge all individual development sets given a translation task.
  - ✓ We use the same pre-processing approach to deal with the source sentences and the reference translations.

- When decoding the test set:
  - ✓ We add the development sets to the training corpus to re-train the models.

Experimental Results

-- Corpus statistics 2

Corpus statistic for BTEC CE task

corpus	Size
Train corpus	19,972 sentence pairs
Development set	2,508 sentence with 16 references
Test set	469 sentence

Corpus statistic for CT-CE tasks

corpus	Size
Train corpus	30,033 sentence pairs
Development set	4,447 sentence with 16 references
Test set	405 sentence

Corpus statistic for CT-EC tasks

corpus	Size
Train corpus	30,033 sentence pairs
Development set	1,465 sentence with 7 references
Test set	393 sentence

Experimental Results

-- Word aligning 1, combine different word alignments

- We combine the word alignments produced by GIZA++ and BerkeleyAligner.
  - ✓ We use GIZA++ and BerkeleyAligner to generate different word alignment files.
  - ✓ We merge the two files into a big word alignment file by concatenating one alignment file to the other.
  - ✓ The big word alignment file is exploited by the decoders to generate the translation models.
- Improvement with combining word alignments

Challenge CT_CE CRR	Moses	Joshua
Baseline (GIZA++)	36.24	36.83
Combining word alignment	38.09	39.24

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

Experimental Results

-- Word aligning 2, a two-step word alignment approach

- In the first step:
  - ✓ We use GIZA++ to produce word alignment and phrase table;
  - ✓ We set a threshold value, such as 0.5, to filter the phrase table, and extract some phrase tables.
- In the second step:
  - ✓We add the reliable bilingual phrase tables into the training data and re-train the model.

Moses decoder	BLEU
Baseline	36.24
Two-step word alignment	36.83
Combining word alignment+ two-step word alignment	38.26

Experimental Results

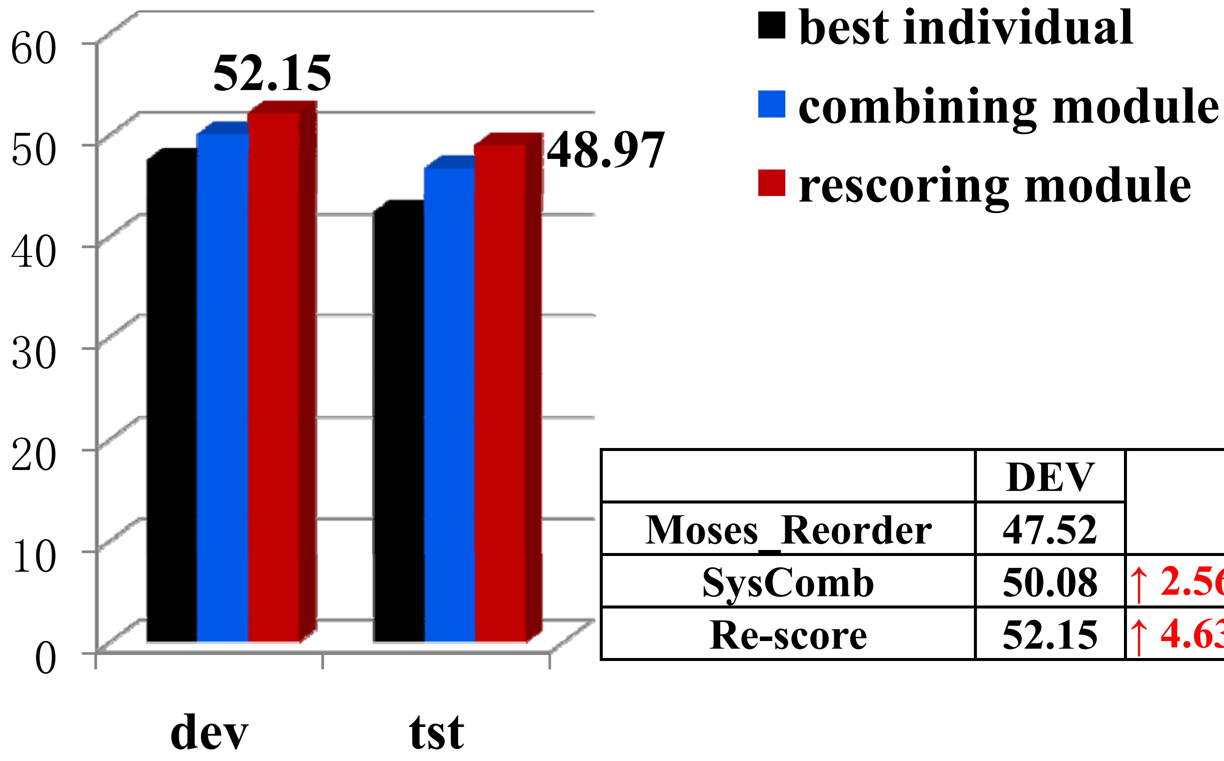
--Performance 1

- For each task, we submit three system running results:
  - ✓ Re-score result (primary).
  - ✓ System combination result (contrastive 1).
  - ✓ The results of the best individual system on the development set (**contrastive 2**).

- The performance on the development sets is case-insensitive, while on the test set is case-sensitive, which is released by the official.

Experimental Results

--Performance 2

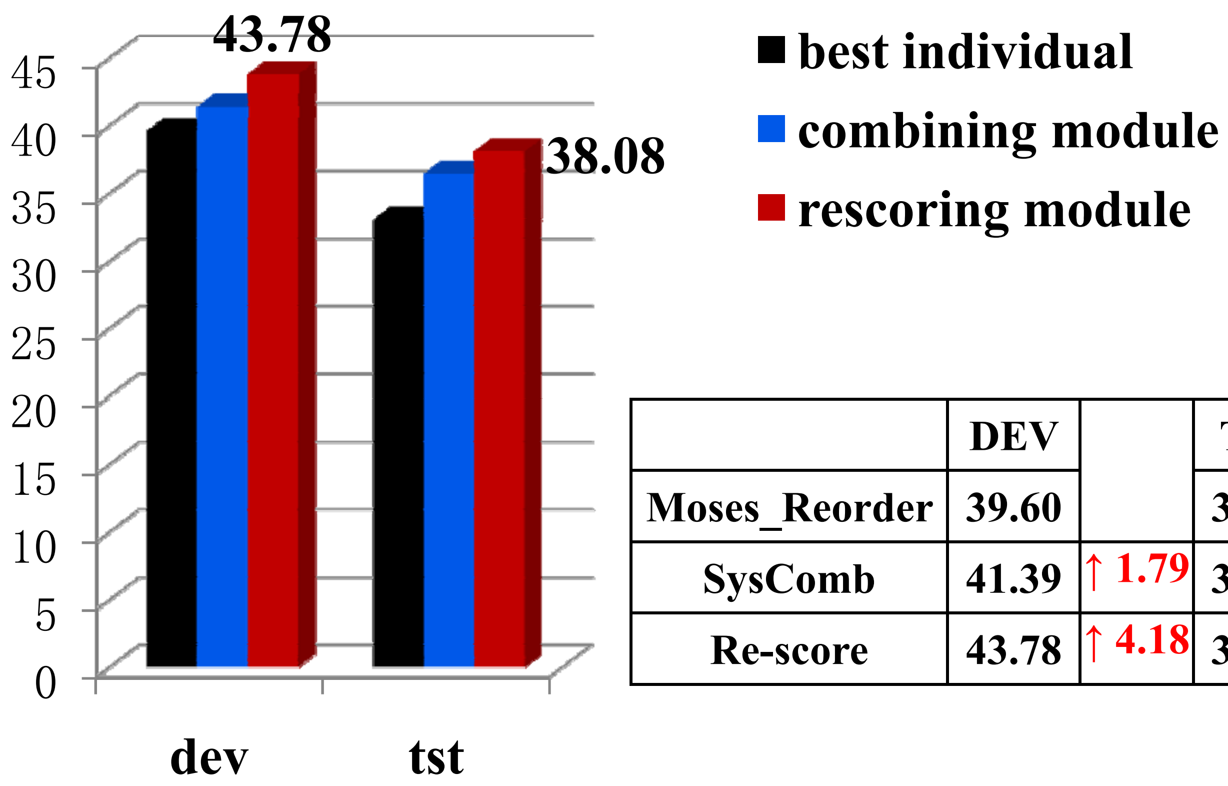


	DEV		TST	
Moses_Reorder	47.52		42.39	
SysComb	50.08	↑ 2.56	46.68	↑ 4.29
Re-score	52.15	↑ 4.63	48.97	↑ 6.58

The translation performance on the development set and the test set for BTEC CE task

Experimental Results

--Performance 3

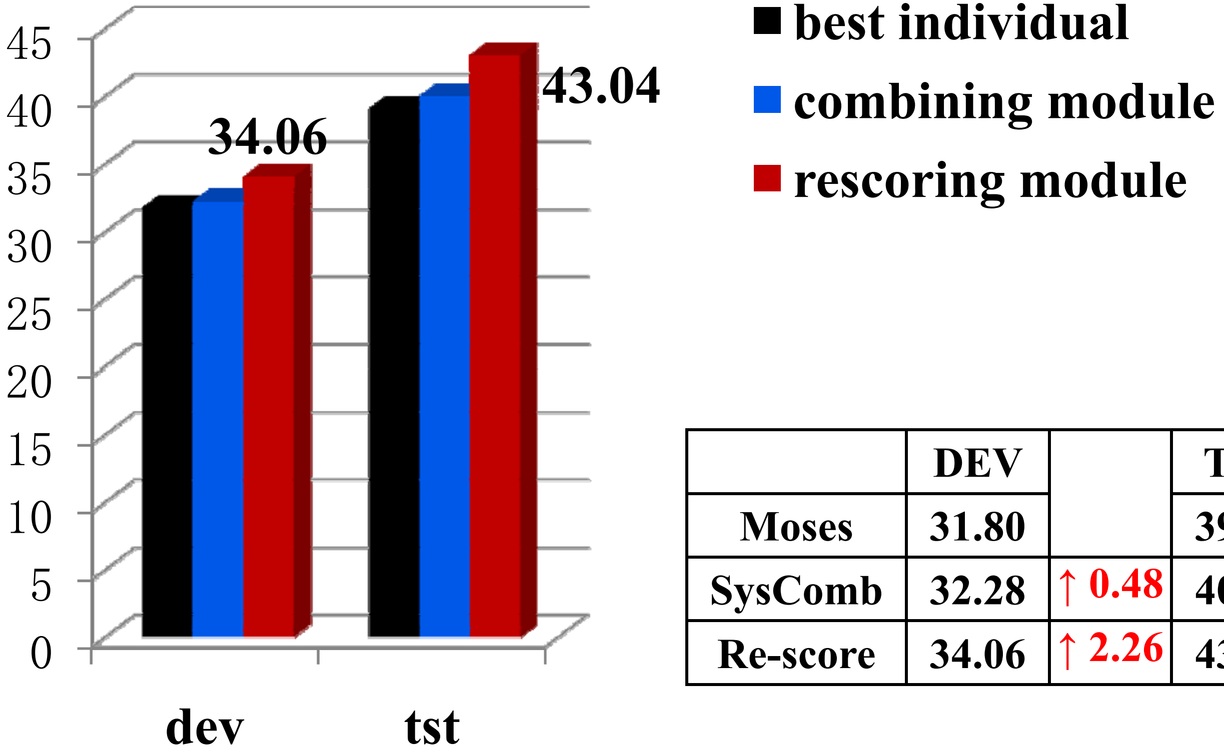


	DEV		TST	
Moses_Reorder	39.60		33.04	
SysComb	41.39	↑ 1.79	36.44	↑ 3.40
Re-score	43.78	↑ 4.18	38.08	↑ 5.04

The translation performance on the development set and the test set for CT-CE CRR task

Experimental Results

--Performance 4



	DEV		TST	
Moses	31.80		39.10	
SysComb	32.28	↑ 0.48	40.03	↑ 0.93
Re-score	34.06	↑ 2.26	43.04	↑ 3.94

The translation performance on the development set and the test set for CT-EC CRR task

Conclusion 1

- The combination module and rescoring module are effective, ↑ 3~6 Bleu points.
- Replace Chinese character with Chinese Pinyin are effective for CT-CE ASR, ↑ 3 Bleu points.
- Combine different word alignments are effective, ↑ 2 Bleu points.
- The two-step word alignment are effective, ↑ 0.6 Bleu points.
- Combine different Chinese word character improve the system performance, ↑ 1 Bleu points.
- Processing NE to the correct formats improves the translation quality.

Conclusion 2

Details determine success or failure!

Thanks for your attention!

Any questions?