Confusion Network Based System Combination for Chinese Translation Output: Word-Level or Character-Level?

LI Maoxi[1] *WANG Mingwen*[1]

(1) School of Computer Information Engineering, Jiangxi Normal University,
Nanchang, China, 330022

mosesli@yeah.net, mwwang@jxnu.edu.cn

Introduction

- Recently, confusion network based system combination has applied successfully to various machine translation tasks.
- Confusion network based system combination picks one hypothesis as the skeleton and aligns the other hypotheses against the skeleton to form a confusion network.
- The path with the highest score represents the consensus translation.
- Previous work on system combination most focus on combining translation outputs in Latin alphabet-based languages, in which sentences are already segmented into words sequences with white space before constructing the confusion network.

Introduction

- When combining Chinese translation outputs
  - The first step is to segment the translation output into a sequence of words,
  - An alternative is to split the translation output into characters,
  - Both approach is possible.
- In this woks, we compare the translation performance of confusion network based system combination when the Chinese translation output is segmented into words versus characters.

Related work

- It is a long debating issue that which one, word or character, is the appropriate unit for Chinese NLP.
  - J. Xu, et al. investigated CWS for Chinese-English phrase-based SMT,
  - R. Zhang, et al. reported that the most accurate word segmentation is not the best word segmentation for SMT,
  - P-C Chang, et al. optimized CWS granularity with respect to the SMT task,
  - M. Li, et al. compared word-level metrics with character-level metrics,
  - J. Du utilized a character-level strategy to improve translation quality for spoken language translation.

Confusion network based system combination for Chinese translation output

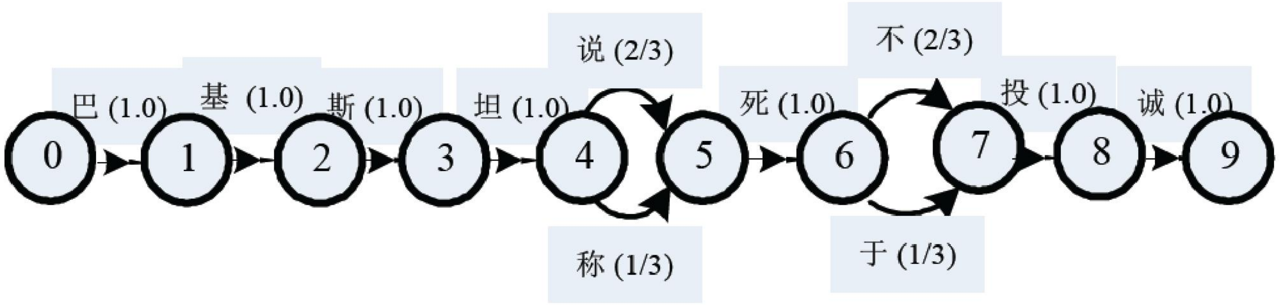
- IHMM monolingual hypothesis alignment approach is utilized to align the hypothesis to the skeleton.
- IHMM approach uses a similarity model and a distortion model to calculate the conditional probability that the hypothesis is generated by the skeleton.

$$p(e_j^{'} | e_i) = a \cdot p_{sem}(e_j^{'} | e_i) + (1-a) \cdot p_{sur}(e_j^{'} | e_i)$$

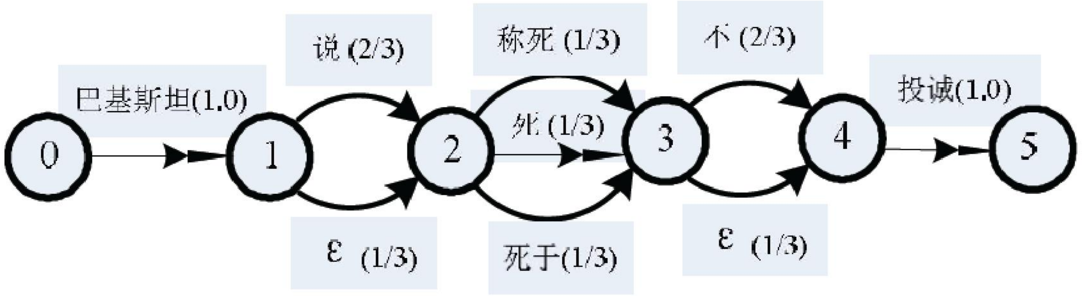
- Given a source sentence:
  - ➢ *Pakistan cleric says would rather die than surrender*
- And three translation hypotheses:
  - ➢ *巴 基 斯 坦 称 死 不 投 诚*
  - ➢ *巴 基 斯 坦 说 死 不 投 诚*
  - ➢ *巴 基 斯 坦 说 死 于 投 诚*

Confusion network based system combination for Chinese translation output

- We can construction a word-level and a character-level confusion network given the example.



(a) A character-level confusion network



(b) A word-level confusion network

Experimental Data

- We conducted experiments on two datasets
  - The NIST'08 English-to-Chinese translation task.
    - Contains 127 documents with 1,830 segments;
    - 4 human reference translations;
    - The best 7 submitted system outputs are chose to participate in system combination;
    - 3-fold cross-validation.
  - The IWSLT'08 English-to-Chinese CRR challenge task.
    - The development set contained 757 segments and the test set contained 300 segments;
    - 4 human reference translations;

Experimental Setting

- It has been reported that character-level automatic metrics correlate with human judgment better than word-level automatic metrics for Chinese translation evaluation.
- The system performance of Chinese translation output are measured with character-level metrics.
  - Character-level BLEU,
  - Character-level NIST,
  - Character-level METEOR,
  - Character-level GTM,
  - Character-level TER

Experimental Setting

- Because better automatic evaluation metrics leading to better translation performance for parameters optimization.
- The feature weights of confusion network based combination system are tuned based on character-level BLEU score.
- We experimented with three different CWS tools
  - ICTCLAS,
  - Stanford Chinese word segmenter (STANFORD),
  - Urheen.

Results on NIST’08 EC Tasks

- The submitted outputs of 7 systems are combined:
  - System 01, system 03, system 17, system 18, system 24, system 28, and system 31.
  - Words are not demarcated in the system outputs, we divide the output into words by different CWS tools or characters to facilitate hypothesis alignment before combining the outputs.

Results on NIST'08 EC Tasks

- The "Character" row shows the translation performance after the system outputs are split into characters.
- The "ICTCLAS", "STANFORD", and "Urheen" rows show the scores when the system outputs are segmented into words by the respective CWS tools.
- Experimental results given in Table 1 show that the character-level combination system significantly improves the translation performance ($p < 0.01$).

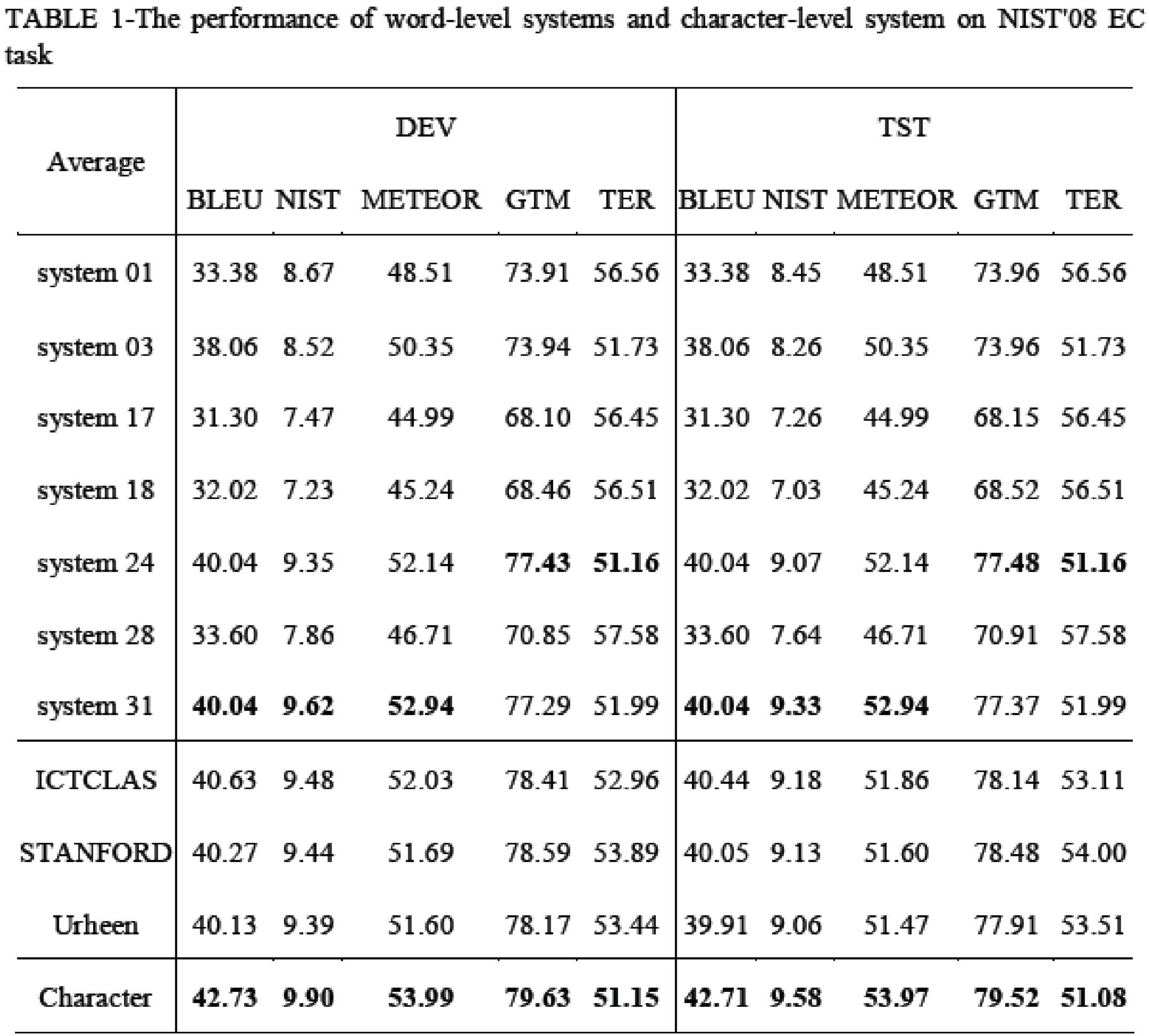


TABLE 1-The performance of word-level systems and character-level system on NIST'08 EC task

Average	DEV					TST				
	BLEU	NIST	METEOR	GTM	TER	BLEU	NIST	METEOR	GTM	TER
system 01	33.38	8.67	48.51	73.91	56.56	33.38	8.45	48.51	73.96	56.56
system 03	38.06	8.52	50.35	73.94	51.73	38.06	8.26	50.35	73.96	51.73
system 17	31.30	7.47	44.99	68.10	56.45	31.30	7.26	44.99	68.15	56.45
system 18	32.02	7.23	45.24	68.46	56.51	32.02	7.03	45.24	68.52	56.51
system 24	40.04	9.35	52.14	**77.43**	**51.16**	40.04	9.07	52.14	**77.48**	**51.16**
system 28	33.60	7.86	46.71	70.85	57.58	33.60	7.64	46.71	70.91	57.58
system 31	**40.04**	**9.62**	**52.94**	77.29	51.99	**40.04**	**9.33**	**52.94**	77.37	51.99
ICTCLAS	40.63	9.48	52.03	78.41	52.96	40.44	9.18	51.86	78.14	53.11
STANFORD	40.27	9.44	51.69	78.59	53.89	40.05	9.13	51.60	78.48	54.00
Urheen	40.13	9.39	51.60	78.17	53.44	39.91	9.06	51.47	77.91	53.51
Character	**42.73**	**9.90**	**53.99**	**79.63**	**51.15**	**42.71**	**9.58**	**53.97**	**79.52**	**51.08**

Results on IWSLT'08 EC CRR challenge Tasks

- We segment the Chinese sentences in bilingual training data into word sequences, and train several English-to-Chinese SMT systems to decode the development set and test set.
  - Joshua$_{\text{ICTCLAS}}$ represent the Joshua system that Chinese sentences in the training data have been segmented into words by ICTCLAS tools, thus the outputs to be combined can be seemed to have been segmented into words by ICTCLAS tools.
  - Joshua$_{\text{STANFORD}}$ represent the Joshua system that Chinese sentences in the training data have been segmented into words by STANFORD tool.

Results on IWSLT'08 EC CRR challenge Tasks

- Because the outputs to be combined have been segmented into words with different granularity, we must consistently re-segment the outputs into words or characters before system combination.
  - The "ICTCLAS", and "STANFORD" rows show the scores when the system outputs are re-segmented into words by the respective Chinese word segmenters.
- The experimental results in Table 2 show when translation outputs to be combined are with different word granularity:
  - The character-level combination system significantly improves the translation performance.

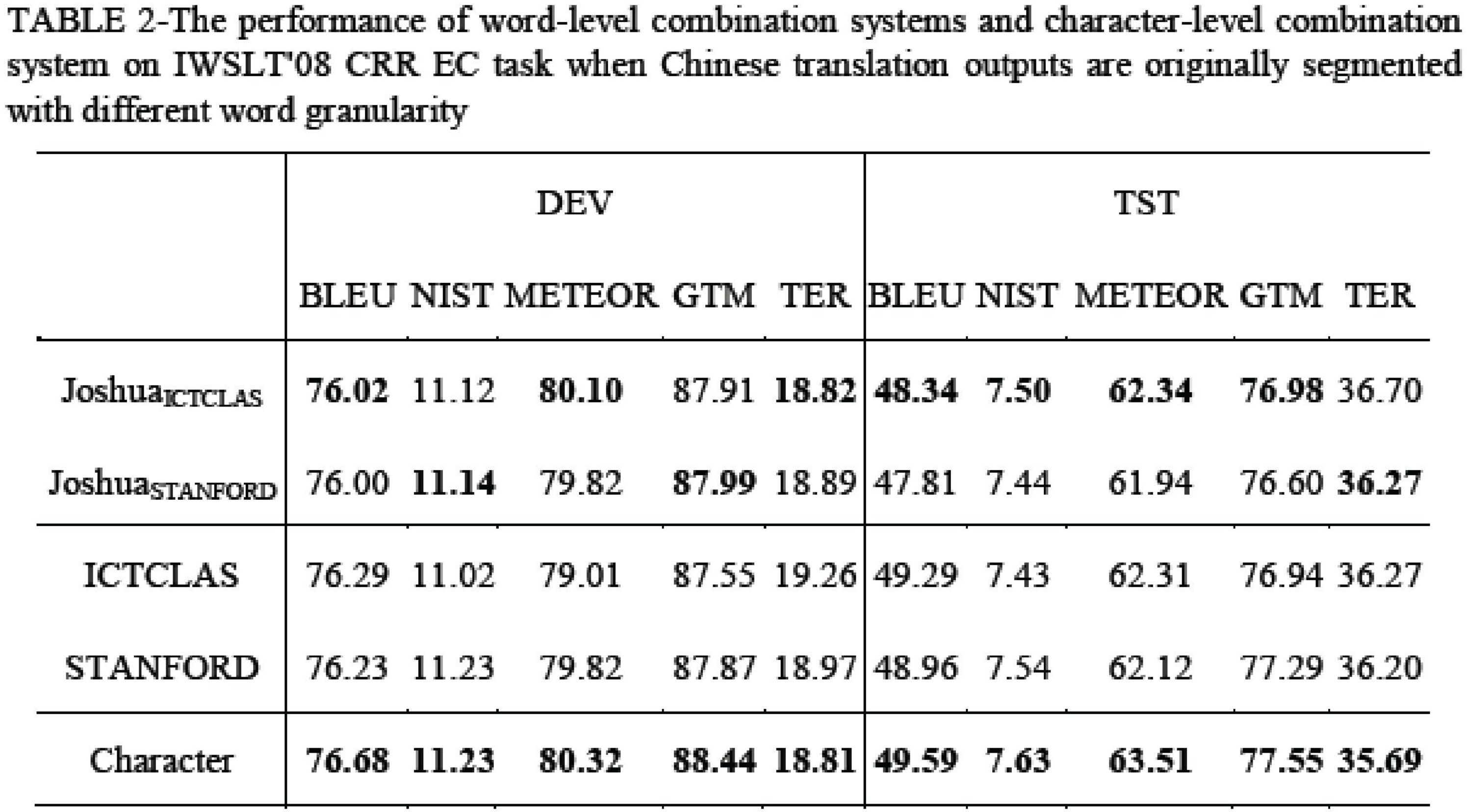


TABLE 2-The performance of word-level combination systems and character-level combination system on IWSLT'08 CRR EC task when Chinese translation outputs are originally segmented with different word granularity

	DEV					TST				
	BLEU	NIST	METEOR	GTM	TER	BLEU	NIST	METEOR	GTM	TER
Joshua$_{ICTCLAS}$	**76.02**	11.12	**80.10**	87.91	**18.82**	**48.34**	**7.50**	**62.34**	**76.98**	36.70
Joshua$_{STANFORD}$	76.00	**11.14**	79.82	**87.99**	18.89	47.81	7.44	61.94	76.60	**36.27**
ICTCLAS	76.29	11.02	79.01	87.55	19.26	49.29	7.43	62.31	76.94	36.27
STANFORD	76.23	11.23	79.82	87.87	18.97	48.96	7.54	62.12	77.29	36.20
Character	**76.68**	**11.23**	**80.32**	**88.44**	**18.81**	**49.59**	**7.63**	**63.51**	**77.55**	**35.69**

Results on IWSLT'08 EC CRR challenge Tasks

- When the outputs to be combined have been segmented into words by the same CWS tool ICTCLAS, we combined the output generated by two SMT systems:
  - $\text{Moses}_{\text{ICTCLAS}}$,
  - $\text{Joshua}_{\text{ICTCLAS}}$.
- Table 3 shows the character-level combination system still consistently outperforms the word-level combination system, "ICTCLAS", even though the translation outputs to be combined are with the same word granularity.

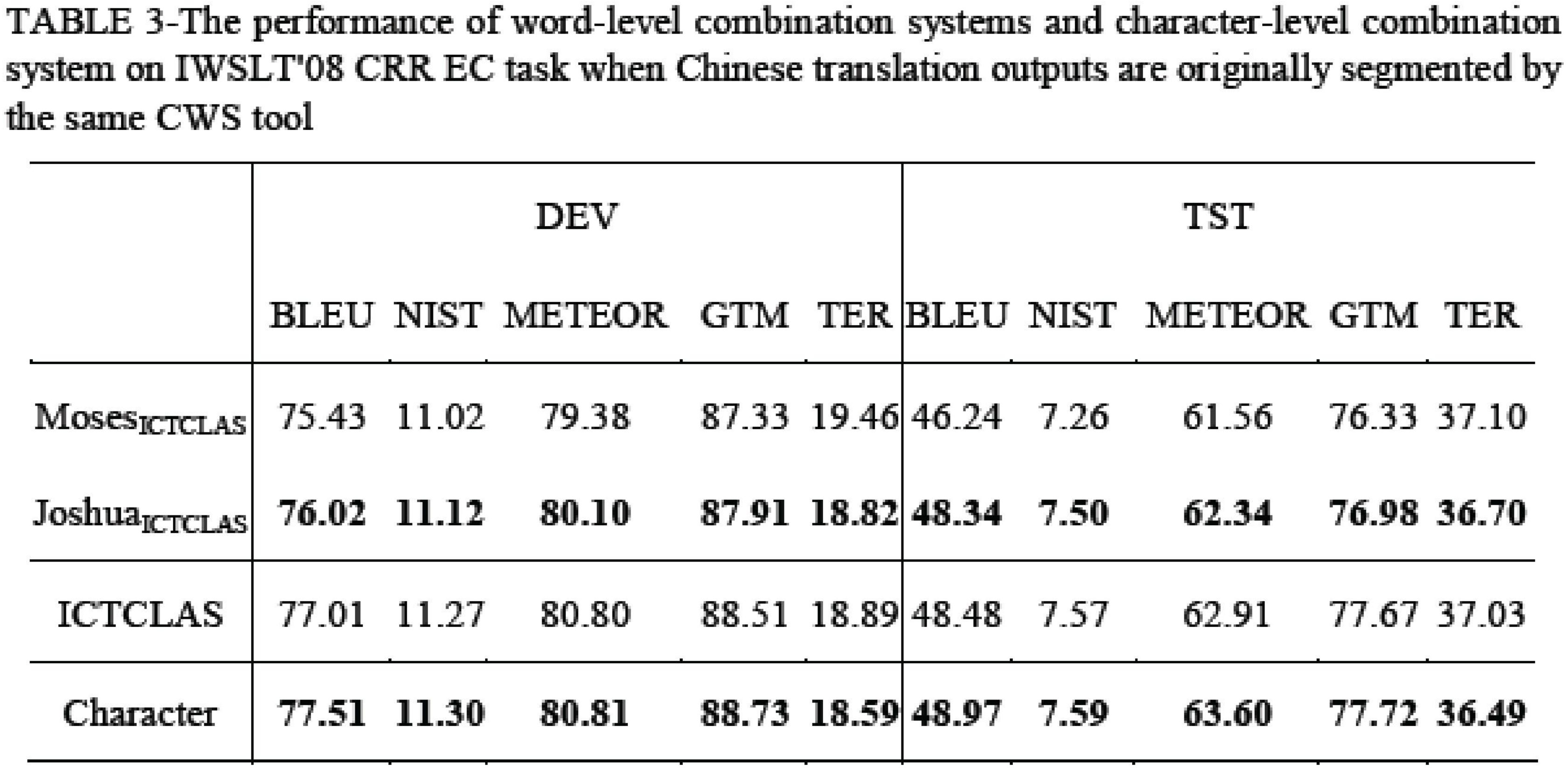


TABLE 3-The performance of word-level combination systems and character-level combination system on IWSLT'08 CRR EC task when Chinese translation outputs are originally segmented by the same CWS tool

	DEV					TST				
	BLEU	NIST	METEOR	GTM	TER	BLEU	NIST	METEOR	GTM	TER
Moses$_{ICTCLAS}$	75.43	11.02	79.38	87.33	19.46	46.24	7.26	61.56	76.33	37.10
Joshua$_{ICTCLAS}$	**76.02**	**11.12**	**80.10**	**87.91**	**18.82**	**48.34**	**7.50**	**62.34**	**76.98**	**36.70**
ICTCLAS	77.01	11.27	80.80	88.51	18.89	48.48	7.57	62.91	77.67	37.03
Character	**77.51**	**11.30**	**80.81**	**88.73**	**18.59**	**48.97**	**7.59**	**63.60**	**77.72**	**36.49**

Conclusion and discussion

- We conducted a study of character-level versus word-level confusion network based system combination for Chinese translation output.
- The experimental results show that character-level combination system significantly outperforms word-level combination systems.

Conclusion and discussion

- Reasons:
  - Chinese sentences can be split into characters with perfect accuracy; however, there is not a CWS tool to perform 100% yet. Therefore, outputs can be segmented into characters more consistently. which lead to generate high quality monolingual hypothesis alignment to help construct confusion network.
  - Chinese character is a smaller unit than Chinese word (containing at least one character) for constructing confusion network. Thus, character-level approach has more choice to produce better consensus translation.

Thanks!