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

ABSTRACT

Recently, confusion network based system combination has applied successfully to various machine translation tasks. However, to construct the confusion network when combining the Chinese translation outputs from multiple machine translation systems, it is possible to either take a Chinese word as the atomic unit (word-level) or take a Chinese character as the atomic unit (character-level). In this paper, we compare word-level approach with character-level approach for combining Chinese translation outputs on the NIST'08 EC tasks and IWSLT'08 EC CRR challenge tasks. Our experimental results reveal that character-level combination system significantly outperforms word-level combination system.

KEYWORDS : machine translation; system combination; confusion network; Chinese translation output

1 Introduction

In recent years, the confusion network based system combination seems to be an expedient powerful means to improve the translation quality in many machine translation tasks empirically, which aims at combining the multiple outputs of various translation systems into a consensus translation (Chen et al., 2009; Feng et al., 2009; He et al., 2008; Rosti et al., 2007; Watanabe & Sumita, 2011). 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. However, for Asian Language, such as Chinese, Japanese, and Korean etc., words are not demarcated originally in the translation output. Thus, in those languages processing, the first step is to segment the translation output into a sequence of words. Instead of segmenting the translation output into words, an alternative is to split the translation output into characters, which can be readily done with perfect accuracy. It is possible that take either a word or a character as the smallest unit to construct the confusion network for system combination. So far, there has been no detailed study to compare the translation performance of these two combination approaches (word-level vs. character-level).

In this paper, we compare the translation performance of confusion network based system combination when the Chinese translation output is segmented into words versus characters. Since there are several Chinese word segmentation (CWS) tools that can segment Chinese sentences into words and their segmentation results are different, we use three representative CWS tools in our experiments. Our experimental results on the NIST'08 EC tasks and IWSLT'08 EC CRR challenge tasks reveal that character-level combination approach significantly outperforms word-level combination approach. That is, the Chinese translation outputs to be combined are not needed to be segment into words.

2 Related work

It is a long debating issue that which one, word or character, is the appropriate unit for Chinese natural language processing. J. Xu, et al. investigated CWS for Chinese-English phrase-based statistical machine translation (SMT), and found that a system which relied on characters performed slightly worse than when it used segmented words (Xu et al., 2004). R. Zhang, et al. reported that the most accurate word segmentation is not the best word segmentation for SMT (Zhang et al., 2008). P-C Chang, et al. optimized CWS granularity with respect to the SMT task (Chang et al., 2008). M. Li, et al. compared word-level metrics with character-level metrics, and demonstrated that word segmentation is not essential for automatic evaluation of Chinese translation output (Li et al., 2011). J. Du utilized a character-level system combination strategy to improve translation quality for English-Chinese spoken language translation (Du, 2011).

3 Confusion network based system combination for Chinese translation output

One of the crucial steps in confusion network based system combination is to align different hypotheses to each other. A variety of monolingual hypothesis alignment strategies have been proposed in recent years, such as GIZA++-like approach (Matusov et al., 2006; Och & Ney, 2003), TER (Snover et al., 2006), IHMM (He et al., 2008), and IncIHMM (Li et al., 2009) etc. It had been reported that IHMM is the most stable among the first three approaches (Chen et al., 2009). To get higher quality hypothesis alignment, we utilize the IHMM approach to align translation output.

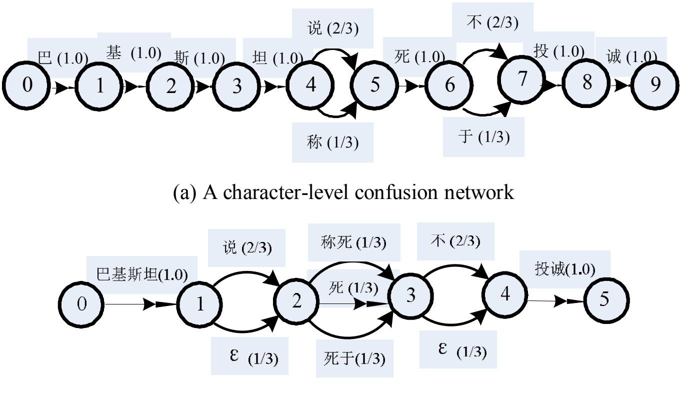
IHMM approach uses a similarity model and a distortion model to calculate the conditional probability that the hypothesis is generated by the skeleton. The similarity model, which models the similarity between a word in the skeleton and a word in the hypothesis, is a linear interpolation of the semantic similarity and surface similarity.

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

The interpolation weight α is empirically set as 0.3.

For Chinese translation output, the semantic similarity between two Chinese words or two Chinese characters can also be estimated by using the source word sequence as a hidden layer. Because it is very hard to get the longest matched prefix or the longest common subsequence between two Chinese words or two Chinese characters, the surface similarity is based on exact match, that is, the surface similarity is set 1 if the word or character e' is the same as e, and is set 0 otherwise.

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*", we can use IHMM approach to align the hypotheses at character-level and word-level. The character-level and word-level confusion networks are built as shown in FIGURE 1. Finally, the consensus translation can be obtained by confusion network decoding.



(a) A character-level confusion network

(b) A word-level confusion network

FIGURE 1-Character-level and word-level confusion networks

4 Experimental results

4.1 Data

To compare the performance of word-level combination system with character-level combination system, we conduct experiments on two datasets, in the newswire translation domain and the spoken language translation domain.

The test set of NIST'08 English-to-Chinese translation task contains 127 documents with 1,830 segments. Each segment has 4 reference translations and the system translations of 11 machine translation systems, released in the corpus LDC2010T01. The best 7 submitted system outputs from the constrained training track are chose to participate in system combination, and a 4-gram language model is trained on the official released data LDC2005T14. A 3-fold cross-validation is used to compare the combination performance, the test set is randomly partitioned into three parts, two of them are utilized as development set and the rest is utilized as test set.

Experiments on spoken language translation domain are carried out on the IWSLT'08 English-to-Chinese CRR challenge task. We use the bilingual training data provided by IWSLT evaluation campaign (Paul, 2008). The development set contained 757 segments and the test set contained 300 segments, each segment with 7 human reference translations.

4.2 Automatic evaluation of Chinese translation output

It has been reported that character-level automatic metrics correlate with human judgment better than word-level automatic metrics for Chinese translation evaluation (Li et al., 2011). To measure the translation performance of word-level combination system and character-level combination system, several off-the-shelf automatic metrics, namely BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee & Lavie, 2005), GTM (Melamed et al., 2003), and TER (Snover et al., 2006), are used at character-level. Unless otherwise stated, the performance of Chinese translation is measured with character-level metrics scores. Because better automatic evaluation metrics leading to better translation performance for parameters optimization (Liu et al., 2011), the feature weights of confusion network based combination system are tuned based on character-level BLEU score.

4.3 Results

For NIST'08 EC task, the submitted outputs of 7 systems are combined: system 01, system 03, system 17, system 18, system 24, system 28, and system 31. Due to words are not demarcated in the system outputs, we must divide the output into words or characters to facilitate hypothesis alignment before combining the outputs. Since there are a number of CWS tools and they generally give different segmentation results. To consistently segment the Chinese outputs into word sequences, we experimented with three different CWS tools, namely ICTCLAS (Zhang et al., 2003), Stanford Chinese word segmenter (STANFORD) (Tseng et al., 2005), Urheen (Wang et al., 2010). TABLE 1 summary the performance for character-level combination system and word-level combination systems. 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. Compared to word-level combination systems, the character-level combination system improves the translation performance. This improvement is statistically significant ($p < 0.01$).

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**

Besides combining the submitted system outputs in which words are not delimited on NIST'08 EC task, we also conduct experiments on system outputs that have been segmented into word sequences on IWSLT'08 EC CRR challenge tasks. The state of the art SMT systems, Moses (Koehn et al., 2006) and Joshua (Li et al., 2009), are exploited to generate N-best lists for system combination. 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 of IWSLT'08 EC CRR challenge tasks. The N-best list hypotheses can be seemed to have been segmented into words by the same CWS tool that is used to segment the Chinese sentences in the training data.

TABLE 2 shows the translation performance when translation outputs to be combined are with different word granularity. Two SMT systems are combined: Joshua$_{\text{ICTCLAS}}$, and Joshua$_{\text{STANFORD}}$. 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. While Joshua$_{\text{STANFORD}}$ represent the Joshua system that Chinese sentences in the training data have been segmented into words by STANFORD tool. 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. Compared to word-level combination systems, "ICTCLAS", and "STANFORD", the character-level combination system, "Character", 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$_{\text{ICTCLAS}}$	**76.02**	11.12	**80.10**	87.91	**18.82**	**48.34**	**7.50**	**62.34**	**76.98**	36.70
Joshua$_{\text{STANFORD}}$	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**

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$_{\text{ICTCLAS}}$	75.43	11.02	79.38	87.33	19.46	46.24	7.26	61.56	76.33	37.10
Joshua$_{\text{ICTCLAS}}$	**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**

When the outputs to be combined are generated by the SMT systems, Moses$_{\text{ICTCLAS}}$, and Joshua$_{\text{ICTCLAS}}$, in which the Chinese sentences in the training data have been segmented into words by the same CWS tool ICTCLAS, TABLE 3 shows the character-level combination system still consistently outperforms the word-level combination system even though the translation outputs to be combined are with the same word granularity.

Conclusion and discussion

In this paper, we conducted a detailed study of character-level versus word-level confusion network based system combination for Chinese translation output. The experimental results on NIST'08 EC tasks and IWSLT'08 EC CRR challenge tasks show that character-level combination system significantly outperforms word-level combination systems.

There are two possible reasons for character-level combination system better than word-level combination systems. First, 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. Secondarily, Chinese character is a smaller unit than Chinese word (containing at least one character) for constructing confusion network. Thus, character-level confusion network based system combination has more choice to produce better consensus translation.

Acknowledgments

This research has been funded by the Natural Science Foundation of China under Grant No. 6120 3313, 6127 2212, and 6116 3006, and also supported by the Natural Science Foundation of Jiangxi Educational Committee of China under Grant No. GJJ12212.

References

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chang, P.-C., Galley, M., & Manning, C. D. (2008). Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*.

Chen, B., Zhang, M., Li, H., & Aw, A. (2009). A Comparative Study of Hypothesis Alignment and its Improvement for Machine Translation System Combination. In *Proceedings of ACL 2009.*

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of HLT 02*.

Du, J. (2011). Character-Level System Combination: An Empirical Study for English-to-Chinese Spoken Language Translation. In *International Conference on Asian Language Processing*.

Feng, Y., Liu, Y., Mi, H., Liu, Q., & Lv, Y. (2009). Lattice-based System Combination for Statistical Machine Translation. In *Proceedings of EMNLP 2009*.

He, X., Yang, M., Gao, J., Nguyen, P., & Moore, R. (2008). Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems. In *Proceedings of EMNLP 2008*.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O. r., Callison-Burch, C., et al. (2006). Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding. In *John Hopkins University Summer Workshop*.

Li, C.-H., He, X., Liu, Y., & Xi, N. (2009). Incremental HMM Alignment for MT System Combination. In *Processing of ACL 2009*.

Li, M., Zong, C., & Ng, H. T. (2011). Automatic Evaluation of Chinese Translation Output: Word-Level or Character-Level? In *Processing of ACL 2011*.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., et al. (2009). Joshua: An Open Source Toolkit for Parsing-based Machine Translation. In *Proceedings of WMT 2009*.

Liu, C., Dahlmeier, D., & Ng, H. T. (2011). Better Evaluation Metrics Lead to Better Machine Translation. In *Proceedings of EMNLP 2011*.

Matusov, E., Ueffing, N., & Ney, H. (2006). Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of EACL*.

Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and Recall of Machine Translation. In *Proceedings of HLT-NAACL 2003*.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*.

Paul, M. (2008). Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of IWSLT 2008*.

Rosti, A.-V. I., Matsoukas, S., & Schwartz, R. (2007). Improved Word-Level System Combination for Machine Translation. In *Proceedings of ACL 2007*.

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L., & Makhoul, R. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Wang, K., Zong, C., & Su, K.-Y. (2010). A Character-Based Joint Model for Chinese Word Segmentation. In *Proceedings of Coling 2010*.

Watanabe, T., & Sumita, E. (2011). Machine Translation System Combination by Confusion Forest. In *Proceedings of ACL 2011*.

Xu, J., Zens, R., & Ney, H. (2004). Do We Need Chinese Word Segmentation for Statistical Machine Translation? In *Proceedings of ACL-SIGHAN Workshop 2004*.

Zhang, H.-P., Liu, Q., Cheng, X.-Q., Zhang, H., & Yu, H.-K. (2003). Chinese Lexical Analysis Using Hierarchical Hidden Markov Model. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*.

Zhang, R., Yasuda, K., & Sumita, E. (2008). Chinese Word Segmentation and Statistical Machine Translation. ACM Transactions on Speech and Language Processing, 5(2), 1-19.