Automatic Evaluation of Chinese Translation Output: Word-Level or Character-Level?

Maoxi Li   Chengqing Zong
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190
{mxli, cqzong}@nlpr.ia.ac.cn

Hwee Tou Ng
Department of Computer Science
National University of Singapore
13 Computing Drive, Singapore 117417
nght@comp.nus.edu.sg

Abstract

Word is usually adopted as the smallest unit in most tasks of Chinese language processing. However, for automatic evaluation of the quality of Chinese translation output when translating from other languages, either a word-level approach or a character-level approach is possible. So far, there has been no detailed study to compare the correlations of these two approaches with human assessment. In this paper, we compare word-level metrics with character-level metrics on the submitted output of English-to-Chinese translation systems in the IWSLT'08 CT-EC and NIST'08 EC tasks. Our experimental results reveal that character-level metrics correlate with human assessment better than word-level metrics. Our analysis suggests several key reasons behind this finding.

1 Introduction

White space serves as the word delimiter in Latin alphabet-based languages. However, in written Chinese text, there is no word delimiter. Thus, in almost all tasks of Chinese natural language processing (NLP), the first step is to segment a Chinese sentence into a sequence of words. This is the task of Chinese word segmentation (CWS), an important and challenging task in Chinese NLP.

Some linguists believe that word (containing at least one character) is the appropriate unit for Chinese language processing. When treating CWS as a standalone NLP task, the goal is to segment a sentence into words so that the segmentation matches the human gold-standard segmentation with the highest F-measure, but without considering the performance of the end-to-end NLP application that uses the segmentation output. In statistical machine translation (SMT), it can happen that the most accurate word segmentation as judged by the human gold-standard segmentation may not produce the best translation output (Zhang et al., 2008). While state-of-the-art Chinese word segmenters achieve high accuracy, some errors still remain.

Instead of segmenting a Chinese sentence into words, an alternative is to split a Chinese sentence into characters, which can be readily done with perfect accuracy. However, it has been reported that a Chinese-English phrase-based SMT system (Xu et al., 2004) that relied on characters (without CWS) performed slightly worse than when it used segmented words. It has been recognized that varying segmentation granularities are needed for SMT (Chang et al., 2008).

To evaluate the quality of Chinese translation output, the International Workshop on Spoken Language Translation in 2005 (IWSLT'2005) used the word-level BLEU metric (Papineni et al., 2002). However, IWSLT'08 and NIST'08 adopted character-level evaluation metrics to rank the submitted systems. Although there is much work on automatic evaluation of machine translation (MT), whether word or character is more suitable for automatic evaluation of Chinese translation output has not been systematically investigated.

In this paper, we utilize various machine translation evaluation metrics to evaluate the quality of Chinese translation output, and compare their correlation with human assessment when the Chinese translation output is segmented into words versus characters. Since there are several CWS tools that can segment Chinese sentences into words and their segmentation results are different, we use four representative CWS tools in our experiments. Our experimental results reveal that character-level me-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 159–164,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

trics correlate with human assessment better than word-level metrics. That is, CWS is *not* essential for automatic evaluation of Chinese translation output. Our analysis suggests several key reasons behind this finding.

2 Chinese Translation Evaluation

Automatic MT evaluation aims at formulating automatic metrics to measure the quality of MT output. Compared with human assessment, automatic evaluation metrics can assess the quality of MT output quickly and objectively without much human labor.

Translation: 多_少_钱_的_伞_吗_？

Ref 1: 这_些_雨_伞_多_少_钱_？

……

Ref 7: 这_些_雨_伞_的_价_格_是_多_少_？

(a) Segmented into characters.

Translation: 多少_钱_的_伞_吗_？

Ref 1: 这些_雨伞_多少_钱_？

……

Ref 7: 这些_雨伞_的_价格_是_多少_？

(b) Segmented into words by Urheen.

Figure 1. An example to show an MT system translation and multiple reference translations being segmented into characters or words.

To evaluate English translation output, automatic MT evaluation metrics take an English word as the smallest unit when matching a system translation and a reference translation. On the other hand, to evaluate Chinese translation output, the smallest unit to use in matching can be a Chinese word or a Chinese character. As shown in Figure 1, given an English sentence "*how much are the umbrellas?*" a Chinese system translation (or a reference translation) can be segmented into characters (Figure 1(a)) or words (Figure 1(b)).

A variety of automatic MT evaluation metrics have been developed over the years, including BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (exact) (Banerjee and Lavie, 2005), GTM (Melamed et al., 2003), and TER (Snover et al., 2006). Some automatic MT evaluation metrics perform deeper linguistic analysis, such as part-of-speech tagging, synonym matching, semantic role labeling, etc. Since part-of-speech tags are only defined for Chinese words and not for Chinese characters, we restrict the automatic MT evaluation metrics explored in this paper to those metrics listed above which do not require part-of-speech tagging.

3 CWS Tools

Since there are a number of CWS tools and they give different segmentation results in general, we experimented with four different CWS tools in this paper.

ICTCLAS: ICTCLAS has been successfully used in a commercial product (Zhang et al., 2003). The version we adopt in this paper is ICTCLAS2009.

NUS Chinese word segmenter (NUS): The NUS Chinese word segmenter uses a maximum entropy approach to Chinese word segmentation, which achieved the highest F-measure on three of the four corpora in the open track of the Second International Chinese Word Segmentation Bakeoff (Ng and Low, 2004; Low et al., 2005). The segmentation standard adopted in this paper is CTB (Chinese Treebank).

Stanford Chinese word segmenter (STANFORD): The Stanford Chinese word segmenter is another well-known CWS tool (Tseng et al., 2005). The version we used was released on 2008-05-21 and the standard adopted is CTB.

Urheen: Urheen is a CWS tool developed by (Wang et al., 2010a; Wang et al., 2010b), and it outperformed most of the state-of-the-art CWS systems in the CIPS-SIGHAN'2010 evaluation. This tool is trained on Chinese Treebank 6.0.

4 Experimental Results

4.1 Data

To compare the word-level automatic MT evaluation metrics with the character-level metrics, we conducted experiments on two datasets, in the spoken language translation domain and the newswire translation domain.

The IWSLT'08 English-to-Chinese ASR challenge task evaluated the translation quality of 7 machine translation systems (Paul, 2008). The test set contained 300 segments with human assessment of system translation quality. Each segment came with 7 human reference translations. Human assessment of translation quality was carried out on the fluency and adequacy of the translations, as well as assigning a rank to the output of each system. For the rank judgment, human graders were asked to "*rank each whole sentence translation from best to worst relative to the other choices*" (Paul, 2008). Due to the high manual cost, the fluency and adequacy assessment was limited to the output of 4 submitted systems, while the human rank assessment was applied to all 7 systems. Evaluation based on ranking is reported in this paper. Experimental results on fluency and adequacy judgment also agree with the results on human rank assessment, but are not included in this paper due to length constraint.

The NIST'08 English-to-Chinese translation task evaluated 127 documents with 1,830 segments. Each segment has 4 reference translations and the system translations of 11 MT systems, released in the corpus LDC2010T01. We asked native speakers of Chinese to perform fluency and adequacy judgment on a five-point scale. Human assessment was done on the first 30 documents (355 segments) (document id “AFP_ENG_20070701.0026” to “AFP_ENG_20070731.0115”). The method of manually scoring the 11 submitted Chinese system translations of each segment is the same as that used in (Callison-Burch et al., 2007). The adequacy score indicates the overlap of the meaning expressed in the reference translations with a system translation, while the fluency score indicates how fluent a system translation is.

4.2 Segment-Level Consistency or Correlation

For human fluency and adequacy judgments, the Pearson correlation coefficient is used to compute the segment-level correlation between human judgments and automatic metrics. Human rank judgment is not an absolute score and thus Pearson correlation coefficient cannot be used. We calculate segment-level consistency as follows:

$$\rho = \frac{\text{The consistent number of pair-wise comparisons}}{\text{The total number of pair-wise comparisons}}$$

Ties are excluded in pair-wise comparison.

Table 1 and 2 show the segment-level consistency or correlation between human judgments and automatic metrics. The “Character” row shows the segment-level consistency or correlation between human judgments and automatic metrics after the system and reference translations are segmented into characters. The “ICTCLAS”, “NUS”, “STANFORD”, and “Urheen” rows show the scores when the system and reference translations are segmented into words by the respective Chinese word segmenters.

The character-level metrics outperform the best word-level metrics by 2−5% on the IWSLT’08 CT-EC task, and 4−13% on the NIST’08 EC task.

Method	BLEU	NIST	METEOR	GTM	1−TER
Character	**0.69**	**0.73**	**0.74**	**0.71**	**0.60**
ICTCLAS	0.64	0.70	0.69	0.66	0.57
NUS	0.64	0.71	0.70	0.65	0.55
STANFORD	0.64	0.69	0.69	0.64	0.54
Urheen	0.63	0.70	0.68	0.65	0.55

Table 1. Segment-level consistency on IWSLT’08 CT-EC.

Method	BLEU	NIST	METEOR	GTM	1−TER
Character	**0.63**	**0.61**	**0.65**	**0.61**	**0.60**
ICTCLAS	0.49	0.56	0.59	0.55	0.51
NUS	0.49	0.57	0.58	0.54	0.51
STANFORD	0.50	0.57	0.59	0.55	0.50
Urheen	0.49	0.56	0.58	0.54	0.51

Table 2. Average segment-level correlation on NIST’08 EC.

4.3 System-Level Correlation

We measure correlation at the system level using Spearman's rank correlation coefficient. The system-level correlations of word-level metrics and character-level metrics are summarized in Table 3 and 4.

Because there are only 7 systems that have human assessment in the IWSLT’08 CT-EC task, the gap between character-level metrics and word-level metrics is very small. However, it still shows that character-level metrics perform no worse than word-level metrics. For the NIST’08 EC task, the system translations of the 11 submitted MT systems were assessed manually. Except for the GTM metric, character-level metrics outperform word-level metrics. For BLEU and TER, character-level metrics yield up to 6−9% improvement over word-level metrics. This means the character-level metrics reduce about 2−3 erroneous system rankings. When the number of systems increases, the difference between the character-level metrics and word-level metrics will become larger.

Method	BLEU	NIST	METEOR	GTM	1−TER
Character	**0.96**	**0.93**	**0.96**	**0.93**	**0.96**
ICTCLAS	0.96	0.93	0.89	0.93	0.96
NUS	0.96	0.93	0.89	0.86	0.96
STANFORD	0.96	0.93	0.89	0.86	0.96
Urheen	0.96	0.93	0.89	0.86	0.96

Table 3. System-level correlation on IWSLT'08 CT-EC.

Method	BLEU	NIST	METEOR	GTM	1−TER
Character	**0.97**	**0.98**	**1.0**	**0.99**	**0.86**
ICTCLAS	0.91	0.96	0.99	0.99	0.81
NUS	0.91	0.96	0.99	0.99	0.79
STANFORD	0.89	0.97	0.99	0.99	0.77
Urheen	0.91	0.96	0.99	0.99	0.79

Table 4. System-level correlation on NIST'08 EC.

5 Analysis

We have analyzed the reasons why character-level metrics better correlate with human assessment than word-level metrics.

Compared to word-level metrics, character-level metrics can capture more synonym matches. For example, Figure 1 gives the system translation and a reference translation segmented into words:

Translation: 多少_钱_的_*伞*_吗_？

Reference: 这些_*雨伞*_多少_钱_？

The word "*伞*" is a synonym for the word "*雨伞*", and both words are translations of the English word "*umbrella*". If a word-level metric is used, the word "*伞*" in the system translation will not match the word "*雨伞*" in the reference translation. However, if the system and reference translation are segmented into characters, the word "*伞*" in the system translation shares the same character "*伞*" with the word "*雨伞*" in the reference. Thus character-level metrics can better capture synonym matches.

We can classify the semantic relationships of words that share some common characters into three types: exact match, partial match, and no match. The statistics on the output translations of an MT system are shown in Table 5. It shows that "exact match" accounts for 71% (29/41) and "no match" only accounts for 7% (3/41). This means that words that share some common characters are synonyms in most cases. Therefore, character-level metrics do a better job at matching Chinese translations.

Total count	Exact match	Partial match	No match
41	29	9	3

Table 5. Statistics of semantic relationships on words sharing some common characters.

Another reason why word-level metrics perform worse is that the segmented words in a system translation may be inconsistent with the segmented words in a reference translation, since a statistical word segmenter may segment the same sequence of characters differently depending on the context in a sentence. For example:

Translation: 你_在_*京都*_吗_？

Reference: 您_在_*京*_ *都*_做_什么_？

Here the word "*京都*" is the Chinese translation of the English word "*Kyoto*". However, it is segmented into two words, "*京*" and "*都*", in the reference translation by the same CWS tool. When this happens, a word-level metric will fail to match them in the system and reference translation. While the accuracy of state-of-the-art CWS tools is high, segmentation errors still exist and can cause such mismatches.

To summarize, character-level metrics can capture more synonym matches and the resulting segmentation into characters is guaranteed to be consistent, which makes character-level metrics more suitable for the automatic evaluation of Chinese translation output.

6 Conclusion

In this paper, we conducted a detailed study of the relative merits of word-level versus character-level metrics in the automatic evaluation of Chinese translation output. Our experimental results have shown that character-level metrics correlate better with human assessment than word-level metrics. Thus, CWS is *not* needed for automatic evaluation of Chinese translation output. Our study provides the needed justification for the use of character-level metrics in evaluating SMT systems in which Chinese is the target language.

Acknowledgments

This research was done for CSIDM Project No. CSIDM-200804 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore. This research has also been funded by the Natural Science Foundation of China under Grant No. 60975053, 61003160, and 60736014, and also supported by the External Cooperation Program of the Chinese Academy of Sciences. We thank Kun Wang, Daniel Dahlmeier, Matthew Snover, and Michael Denkowski for their kind assistance.

References

Satanjeev Banerjee and Alon Lavie, 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65-72, Ann Arbor, Michigan, USA.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder, 2007. (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136-158, Prague, Czech Republic.

Pi-Chuan Chang, Michel Galley and Christopher D. Manning, 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224-232, Columbus, Ohio, USA.

George Doddington, 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. *Proceedings of the Second International Conference on Human Language Technology Research (HLT'02)*, pages 138-145, San Diego, California, USA.

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo, 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161-164, Jeju Island, Korea.

I. Dan Melamed, Ryan Green and Joseph P. Turian, 2003. Precision and Recall of Machine Translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003) - short papers*, pages 61-63, Edmonton, Canada.

Hwee Tou Ng and Jin Kiat Low, 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 277-284, Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318, Philadelphia, Pennsylvania, USA.

Michael Paul, 2008. Overview of the IWSLT 2008 Evaluation Campaign. *Proceedings of IWSLT 2008*, pages 1-17, Hawaii, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla and Ralph Makhoul, 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the Association for Machine Translation in the Americas*, pages 223-231, Cambridge.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning, 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168-171, Jeju Island, Korea.

Kun Wang, Chengqing Zong and Keh-Yih Su, 2010a. A Character-Based Joint Model for Chinese Word Segmentation. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1173-1181, Beijing, China.

Kun Wang, Chengqing Zong and Keh-Yih Su, 2010b. A Character-Based Joint Model for CIPS-SIGHAN Word Segmentation Bakeoff 2010. *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010)*, pages 245-248, Beijing, China.

Jia Xu, Richard Zens and Hermann Ney, 2004. Do We Need Chinese Word Segmentation for Statistical Machine Translation? *Proceedings of the ACL SIGHAN Workshop 2004*, pages 122-128, Barcelona, Spain.

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang and Hong-Kui Yu, 2003. Chinese Lexical Analysis

Using Hierarchical Hidden Markov Model. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 63-70, Sapporo, Japan.

Ruiqiang Zhang, Keiji Yasuda and Eiichiro Sumita, 2008. Chinese Word Segmentation and Statistical Machine Translation. *ACM Transactions on Speech and Language Processing*, 5 (2). pages 1-19.