Mining the Correlation between Human and Automatic Evaluation at Sentence Level

Yanli Sun

School of Applied Language and Intercultural Studies, Dublin City University

yanli.sun2@mail.dcu.ie

Abstract

Automatic evaluation metrics are fast and cost-effective measurements of the quality of a Machine Translation (MT) system. However, as humans are the end-user of MT output, human judgement is the benchmark to assess the usefulness of automatic evaluation metrics. While most studies report the correlation between human evaluation and automatic evaluation at corpus level, our study examines their correlation at sentence level. In addition to the statistical correlation scores, such as Spearman's rank-order correlation coefficient, a finer-grained and detailed examination of the sensitivity of automatic metrics compared to human evaluation is also reported in this study. The results show that the threshold for human evaluators to agree with the judgements of automatic metrics varies with the automatic metrics at sentence level. While the automatic scores for two translations are greatly different, human evaluators may consider the translations to be qualitatively similar and vice versa. The detailed analysis of the correlation between automatic and human evaluation allows us determine with increased confidence whether an increase in the automatic scores will be agreed by human evaluators or not.

1. Introduction

It is widely recognized that evaluation plays an important role in the development of language technologies. In the area of Machine Translation (MT), there are two types of commonly used evaluation methods. While human evaluation is still the most important means of providing valuable feedback on the further development of an MT system, its cost, labour-intensive and highly subjective characteristics have led to the popularity of automatic evaluation metrics, such as BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2001), Precision and Recall (Turian et al. 2003), TER (Translation Error Rate) (Snover et al. 2006) etc. According to Coughlin (2001), automatic metrics have the advantages of high speed, convenience and comparatively lower-cost. However, as humans are the end-users of MT, human judgement is ultimately the benchmark to assess the usefulness of automatic metrics. How good an automatic metric is depends on its correlation with human evaluation. Two major forms of human evaluation in the area of MT are: scoring, which requires human evaluators to assign two scores (usually 1 to 5) representing the fluency and accuracy of a translation (LDC, 2005); and ranking, which asks human evaluators to compare the translations from different MT systems and assign rankings to them. The problem of scoring is that even with a clear guideline at hand, human evaluators still find it hard to assign appropriate scores to a translation. Ranking, on the other hand, is found to be quite intuitive and reliable (Vilar et al., 2007). Callison-Burch et al. (2008) concluded from their study that ranking was more reliable compared to scoring. Duh (2008) also pointed out that ranking could simplify the decision procedures for human evaluators compared to assigning scores.

Depending on the type of human evaluation used, the correlation between automatic and human evaluation is measured either by Pearson's correlation coefficient or Spearman's correlation coefficient. The correlation value ranges from -1 to 1 representing negative correlation to perfect positive correlation.

As automatic metrics are more effective at corpus level, more effort has been taken on finding out which automatic metric correlates better with human evaluation at corpus level. Nevertheless, increasing attention is being paid to correlation at sentence level. According to Lin and Och (2004), high sentence level correlation of automatic and human evaluation is crucial for machine translation researchers. Russo-Lassner et al. (2005) also pointed out that automatic metrics of high sentence level correlation could "provide a finer-grained assessment of translation quality" and could also "guide MT system development by offering feedback on sentences that are particularly challenging"(p3).

This paper extends the research on correlation at sentence level, aiming at finding out which automatic metric correlates better with human evaluation in terms of Chinese translation from English; and our second aim is to investigate how big a difference between two automatic scores has to be in order to reflect the qualitative changes of the translations. The remainder of the paper is organized as follows: Section two introduces the experiment setting; Section three reports the correlation level between automatic and human evaluation at sentence level; Section four examines the detailed difference between the judgement of automatic evaluation and human evaluation; and Section five summarizes the findings and points out future research questions.

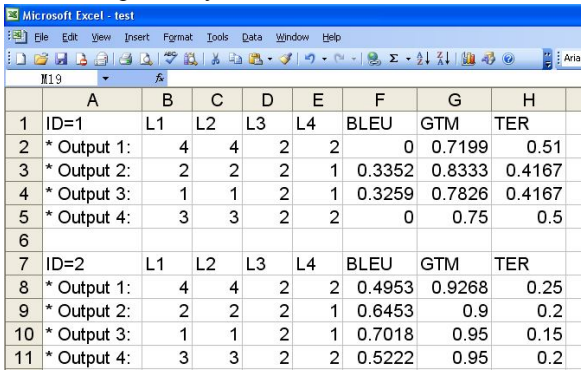
2. Experiment Setting

The automatic evaluation and human evaluation results reported in this paper were collected from an experiment comparing Chinese translations from different MT systems. However, the focus in this paper is to examine the correlation between human evaluation and automatic evaluation and not to discuss the translation quality per se. The corpus is an installation manual of an anti-virus software composed in English from Symantec (Ireland).

Altogether 570 sentences were randomly selected as the test sample. The Chinese reference of the test sample was extracted from the company's Translation Memory. Four MT systems (one Rule-Based system and three Statistical-Based systems) were employed to translate the test sample into Chinese for comparison. Both human and automatic evaluations were applied in order to rank the quality of the output from the four systems. Four professional translators were employed to rank the outputs from 1 to 4 (1 being the best, 4 being the worst) sentence by sentence. BLEU, TER and GTM (General Text Matcher, an implementation of precision and recall) were used to get the automatic scores of each translation at both corpus level and sentence level. The reasons for using these three metrics are: first, they can be used (and have been used) to evaluate Asian language outputs (in this paper, Chinese); second, they are among the most widely used metrics in the area; third, they are relatively easy and cost-effective to use. There are also many other automatic metrics, such as Meteor (Banerjee & Lavie, 2005), TERp (Snover et al., 2009), etc. However, additional conditions are needed to get the best advantage from these metrics. For example, Meteor functions better with a database of synonyms, such as the WordNet for English; TERp requires paraphrases which also function as "synonyms" of phrases. Since these resources for Chinese were not available in our pilot project, these metrics were not employed in this paper. The next section compares the scores from the automatic metrics with the rankings from human evaluators to check how consistent the two evaluation methods are at sentence level with detailed analysis followed in section four.

3. Correlation Check

The correlation between automatic evaluation and human evaluation at sentence level was obtained following the practice of Callison-Burch et al. (2008). As mentioned earlier, we have 570 source English sentences to be translated by four MT systems into Chinese. Therefore, for each source English sentence, four translations can be produced which are ranked by four professional translators and scored by three automatic evaluation metrics. In other words, there are 570 groups (with four items per group) each of which contains four columns of rankings from the four human evaluators and three columns of scores from the three automatic metrics. Figure 1 below shows a sample of the final results sheet. L1, L2, L3, L4 in Figure 1 refer to the four human evaluators respectively.


Microsoft Excel - test

File Edit View Insert Format Tools Data Window Help

M19

	A	B	C	D	E	F	G	H
1	ID=1	L1	L2	L3	L4	BLEU	GTM	TER
2	* Output 1:	4	4	2	2	0	0.7199	0.51
3	* Output 2:	2	2	2	1	0.3352	0.8333	0.4167
4	* Output 3:	1	1	2	1	0.3259	0.7826	0.4167
5	* Output 4:	3	3	2	2	0	0.75	0.5
6								
7	ID=2	L1	L2	L3	L4	BLEU	GTM	TER
8	* Output 1:	4	4	2	2	0.4953	0.9268	0.25
9	* Output 2:	2	2	2	1	0.6453	0.9	0.2
10	* Output 3:	1	1	2	1	0.7018	0.95	0.15
11	* Output 4:	3	3	2	2	0.5222	0.95	0.2

Figure 1: Sample of the Final Results Sheet

One approach to computing the correlation is Spearman's ranking correlation coefficient (ρ). The process of getting Spearman's ranking correlation is as follows: first, the scores assigned by the automatic metrics should be converted into rankings as well; second, for each of the 570 groups, calculate the p value between each automatic metric and each human evaluator using the four items; third, average all the p values to get the mean p value between each metric and each human. Table 1 below reports the correlation values using this method.

	L1	L2	L3	L4	Average
GTM	0.32	0.50	0.14	0.26	**0.30**
TER	0.33	0.48	0.12	0.24	0.29
BLEU	0.34	0.44	0.13	0.26	0.29

Table 1: Spearman's Correlation between Automatic and Human Evaluation

However, the validity of this approach was questioned by Callison-Burch et al. (2008) who claimed that getting the general correlation value by averaging the p values from a limited number of (here only four) items is not appropriate. Instead, in their study, they conducted pair-wise comparison of any two outputs, examining whether the automatic scores were consistent with human rankings given any two outputs (that is the higher-ranked system received a higher score). Following this approach, the 570 groups were expanded into 3420 pairs (each of the 570 groups can be expanded into 6 pairs). For each automatic metric, the total number of consistent evaluations was divided by the total number of comparisons to get a percentage. Table 2 reports the consistency.

	L1	L2	L3	L4	Average
GTM	0.61	0.68	0.71	0.66	**0.66**
TER	0.58	0.64	0.70	0.64	0.64
BLEU	0.51	0.55	0.65	0.59	0.56

Table 2: Consistency of Automatic Evaluation with Human Evaluation

Table 2 indicates that these automatic metrics could correctly predict the human rankings of any pair of translations more than half the time. GTM correlates better with human evaluation than BLEU and TER at sentence level in Chinese output evaluation. Similar findings have been reported by Cahill (2009) in German evaluation which compared 6 metrics including the three metrics used in this paper. Besides, Agarwal and Lavie (2008) also mentioned that GTM and TER could produce more reliable sentence level scores than BLEU.

4. Further Analysis

As shown in Table 2, even for the best correlated metric GTM, there is only 66% consistency, indicating a large amount of discrepancy between humans and automatic evaluation metrics in ranking the quality of different translations. In order to further investigate the consistency and inconsistency at sentence level, we conducted a micro-analysis on the cases where humans and automatic metrics agree/disagree on the rankings of two translations. Given two translations of a source sentence, each of

which is associated with an automatic score, these two scores can suggest a difference in terms of the quality of these two translations. However, humans may or may not agree with the difference registered by the automatic metrics. Nevertheless, intuitively, the greater the differences between two automatic scores of two translations, the more likely that these scores predict the judgements of humans about the quality of the two translations. Based on such consideration, for any pairs of translations of a source sentence, the differences between the two corresponding automatic evaluation scores can be divided into different groups of scales. For example, if the GTM scores for two translations are 0.64 and 0.53 respectively, the difference between these GTM scores (0.11) falls into the difference scale (0.1-0.2). As mentioned in section 3, altogether there are 3420 pairs for comparison. For each automatic metric, the difference of scores within each pair were collected and categorized into different scales. Table 2 reports the number of pairs distributed in the difference scales of each automatic metric.

Difference Scale	GTM #pairs	TER #pairs	BLEU #pairs
0.9-1.0	/	/	18
0.8-0.9	/	/	7
0.7-0.8	/	/	28
0.6-0.7	/	4	35
0.5-0.6	4	11	58
0.4-0.5	12	52	137
0.3-0.4	73	127	201
0.2-0.3	232	278	261
0.1-0.2	627	659	364
0.0-0.1	1484	1026	776

Table 2: Number of Pairs Distributed in each Difference Scale of each Automatic Metric

Table 2 shows that the difference between the automatic scores of two different translations is mostly quite small. For example, 61.02% of the pairs have a difference below 0.1 in terms of GTM score, and this amounts to 47.57% in terms of TER and 41.17% in terms of BLEU.

It is worth pointing out that the scales refer to the difference between two scores for a pair of outputs, not the scale of the scores. The purpose of setting up these difference scales is to see whether the greater the difference between two scores, the more likely that humans agree with automatic metrics. For each of the three automatic evaluation metrics, we consider the following three scenarios: 1) the number of pairs for which human rankings are consistent with the scores assigned to the translations by the automatic metric ("Humans Agree"); 2) the number of pairs for which human rankings are contrary to the scores assigned by the automatic metric ("Humans Disagree"); 3) although the two translations in a pair are different and received two different automatic scores, humans do not think they are qualitatively different and rank the pair as ties ("Humans Assign Ties") (see Figures 2, 3 and 4).

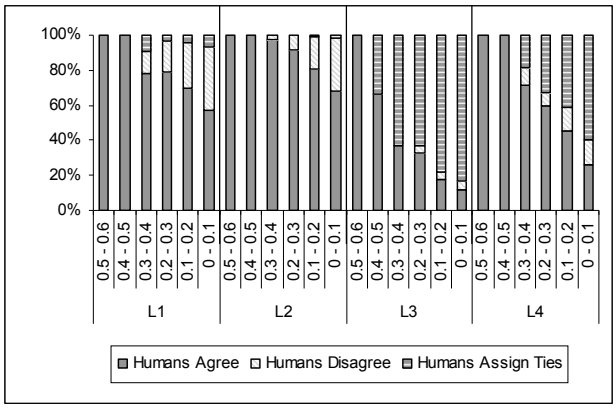


Figure 2: Distribution of Human Evaluation within GTM Difference Scales

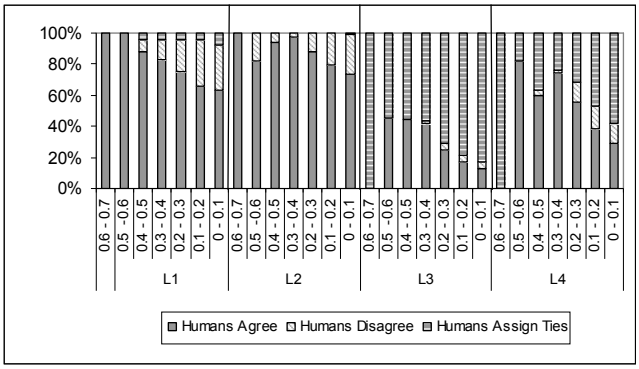


Figure 3: Distribution of Human Evaluation within TER Difference Scales

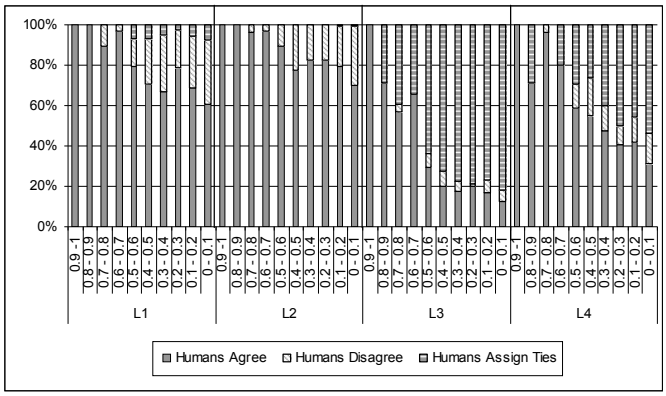


Figure 4: Distribution of Human Evaluation within BLEU Difference Scales

The height of the solid grey bars in Figures 2 to 4 show that for GTM (Figure 2), it is true that the greater the difference between two automatic scores, the more cases that humans **agree** with the judgements of GTM; the smaller the difference, the more cases that humans **disagree** with the judgements of GTM. On the contrary, even with very high TER or BLEU score differences, humans may still disagree with the judgement of TER (Figure 3) or BLEU (Figure 4). In this experiment, when the difference between two GTM scores is bigger than 0.11, the majority of the human evaluators agree with the judgement of the GTM score about which translation is better. The average difference between two TER scores and BLEU scores has to be bigger than 0.18 and 0.29

before the majority of the human evaluators agree with the judgement of these automatic metrics.

Figures 2 to 4 also reflect that different evaluators have different criteria in judging the quality of different translations. As can be seen from the Figures, L3 assigned many more ties in pair-wise comparison than other evaluators. The inter-evaluator correlation within the four human evaluators was measured using the Kappa coefficient (K), a measurement of the agreement between categorical data (Boslaugh & Watters, 2008). One widely accepted interpretation of Kappa was proposed by Landis and Koch (1977): 0-.2 is slight correlation, .2-.4 is fair correlation, .4-.6 is moderate correlation, .6-.8 is substantial correlation and .8-1 is almost perfect correlation. Using the Microsoft Kappa Calculator template (King, 2004), the inter-evaluator agreement score between the four human evaluators is (K=.273). Excluding human evaluator L3, the K value increases to .381.

Generally speaking, even if there are slight differences in two translations, automatic metrics could generate different scores for them. However, there are also cases where the automatic scores are the same for two different translations. In this experiment, we found that for some pairs of different translations for which the automatic metrics assigned the same scores, humans didn't consider them qualitatively different either. On the other hand, there are some other translations that were evaluated as qualitatively different by humans but not by automatic metrics. For each automatic metric, we summed the number of pairs that received the same scores by automatic evaluation but different rankings by human evaluators. As there are four human evaluators, only those pairs that were differentiated by the majority of human evaluators (i.e. three or more evaluators assigned different rankings to the translations in one pair) were taken into consideration. Table 3 contains the total number of pairs where no differentiation was made by the automatic metrics but where humans differentiated.

	GTM	TER	BLEU
#pairs	141	209	331

Table 3: No. Pairs of Translations Differentiated by Humans but not by Automatic Metrics

GTM appears to have the smallest number of pairs that were not differentiated demonstrating a stronger differentiation ability at sentence level more in line with the human evaluation while BLEU left a large number of pairs undifferentiated showing its weakness at sentence level evaluation in relation to the human evaluation. This finding shows that in some cases automatic evaluation cannot reflect the difference between two translations which are apparent according to the human assessments. Hence, if two scores show no sign of difference, it does not always indicate there is no qualitative difference between two translations.

5. Conclusion and Future Work

It is well known that precise automatic evaluation metrics at sentence level can help MT developers determine what sentence structures their MT system can or can not deal with appropriately. This study examines the correlation of automatic evaluation and human evaluation at sentence level in terms of Chinese translation evaluation. Several conclusions have been drawn from this study: first, for evaluation of Chinese translations of English technical document, GTM correlates better with human evaluation than TER and BLEU do at sentence level; second, only when the difference between two scores is greater than a certain value will the majority of human evaluators agree with the judgement of the automatic metrics; third, when two automatic scores of two translations are the same, it does not always mean there is no qualitative difference between the translations. There are also questions remained unanswered: first, the statistical significance of the correlation and consistency is not examined; second, we are aware that the correlation between human and automatic evaluation may vary depending on the MT system involved; however no such distinction was made in this study. Therefore, there is a lot of further work to be done in the future. In addition to these, we have shown that for a considerable number of paired, human judgements are inconsistent with automatic metrics. In the future, we plan to conduct a further analysis into the causes for such discrepancies in an attempt to provide some linguistically motivated patterns that may benefit the design of the automatic metrics. Finally, although human evaluation has been regarded as the golden standard in the process of MT evaluation, the results in this paper reflects some problems of human evaluation. How to standardize human evaluation is another question worthy of exploring in the future.

Acknowledgement

This work was financed by Enterprise Ireland and Symantec Corporation (Ireland). The author would like to thank Dr. Fred Hollowood for his inspiring ideas and suggestions, Dr. Sharon O'Brien, Dr. Minako O'Hagan and Dr. Johann Roturier for their precious corrections and comments. Thanks also to the anonymous reviewers for their insightful comments. However, the author is responsible for any errors in the paper.

Reference

Agarwal, A. & Lavie, A. (2008). 'Meteor, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output.' In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, June, pp. 115-118.

Banerjee, S. & Lavie, A. (2005). 'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments'. In *Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan, pp. 65-72.

Boslaugh, S. & Watters, P.A. (2008). '*Statistics in a Nutshell.*' O'Reilly Media, Inc., the United States of America.

Cahill, A. (2009). 'Correlating Human and Automatic Evaluation of a German Surface Realiser'. In *Proceedings of the ACL-IJNLP 2009 Conference Short Papers,* Suntec, Singapore, August, pp. 97-100.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2008). 'Further Meta-evaluation of Machine Translation'. In *Proceedings of the Third*

Workshop on Statistical Machine Translation, Columbus, Ohio, June, pp. 70-106.

Coughlin, D. (2001). 'Correlating Automated and Human Assessments of Machine Translation Quality'. In *Proceedings of MT Summit IX,* Santiago de Compostela, Spain, September, pp. 63-70.

Duh, K. (2008). 'Ranking vs. Regression in Machine Translation Evaluation'. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, June, pp.191–194.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A.& Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session,* Prague, June, pp.177-180.

King, J. E. (2004). 'Software Solutions for Obtaining a Kappa-type Statistic for Use with Multiple Raters'. *Presented at th*e *Annual Meeting of the Southwest Educational Research Association*, Dallas, TX.

Landis, J.R. & Koch, G.G. (1977). 'The Measurement of Observer Agreement for Categorical Data.' *Biometics*, 33:159-174.

LDC (2005). Linguistic Data Annotation Specification: Assessment of fluency and adequacy in translations. http://projects.ldc.upenn.edu/TIDES/tidesmt.html.

Lin, C. & Och, F.J. (2004). 'ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation'. In *Proceedings of the 20th International Conference on Computational Linguistics,* Geneva, Switzerland, August, pp. 501-508.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2001). 'BLEU: A Method for Automatic Evaluation of Machine Translation'. Research Report RC22176 (W0109-022), IBM T.J.Watson Research Center, September.

Russo-Lassner, G., Lin, J. & Resnik, P. (2005). '*A Paraphrase-Based Approach to Machine Translation Evaluation'. Technical report,* University of Maryland, College Park.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Weischedel, R. (2006). 'A Study of Translation Edit Rate with Targeted Human Annotation'. In *Proceedings of AMTA*, Cambridge, MA, August, pp.223-231.

Snover, M., Madnani, N., Dorr, B.J. & Schwartz, R. (2009). 'Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric'. In *Proceedings of the EACL-2009 Workshop on Statistical Machine Translation (WMT09), Athens,* pp. 259-268.

Turian, J.P., Shen, L., & Melamed, I.D. (2003). 'Evaluation of Machine Translation and its Evaluation'. In *Proceedings of the MT Summit IX,* New Orleans, LA, September, pp. 386-393.

Vilar, D., Leusch, G., Ney, H., & Bachs, R. (2007). 'Human Evaluation of Machine Translation Through Binary System Comparisons'. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, June, pp.96–103.