International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:7, 2008

Performance Analysis of MT Evaluation Measures and Test Suites

Yao Jian-Min, Lv Qiang, and Zhang Jing

***Abstract*—**Many measures have been proposed for machine translation evaluation (MTE) while little research has been done on the performance of MTE methods. This paper is an effort for MTE performance analysis. A general frame is proposed for the description of the MTE measure and the test suite, including whether the automatic measure is consistent with human evaluation, whether different results from various measures or test suites are consistent, whether the content of the test suite is suitable for performance evaluation, the degree of difficulty of the test suite and its influence on the MTE, the relationship of MTE result significance and the size of the test suite, etc. For a better clarification of the frame, several experiment results are analyzed relating human evaluation, BLEU evaluation, and typological MTE. A visualization method is introduced for better presentation of the results. The study aims for aid in construction of test suite and method selection in MTE practice.

***Keywords*—**Machine translation, natural language processing, visualization.

I. Introduction

MACHINE translation evaluation activities have accompanied the MT research and development. In 1960s, the US National Academy of Sciences set up the Automatic Language Processing Advisory Committee who proposed a framework for assessing machine translation and drew a negative conclusion to the MT research [1]. This ALPAC report is the first historical MT evaluation activity, which has greatly influenced the history of machine translation. With new development in natural language processing technology in 1990s, the black-box evaluation has been instantiated by the methodology of US Defense Advanced Research Projects Agency (DARPA) [2], which measures fluency, accuracy, and informativeness on a 5-point scale. The ISLE Project takes an approach that focuses on how an MT system serves the follow-on human processing rather than on what it is unlikely to do well [3].

Manuscript received March 29, 2006. This work was supported in part by the Jiangsu High-Tech Research Program under Grant GB2005020 and Natural Science Foundation of Guangdong Province, and partly by "211" Project Foundation of Soochow University in Computer Information Processing Branch.

Yao Jian-Min and LV Qiang are with the School of Computer Science and Technology, Soochow University, Suzhou 215006, China (e-mail: jyao@suda.edu.cn, and qiang@suda.edu.cn).

Zhang Jing is with the School of Automation, South China University of Technology, Guangzhou 510641, China (e-mail: zhj@scut.edu.cn).

Unfortunately, MTE has not been a very powerful aid in machine translation research because it requires human judgments and is thus expensive, time-consuming and not easily factored into the MT research agenda. That is why automatic evaluation methods are broadly studied and implemented using different heuristics to solve this problem. Followed is a more detailed review of automatic MT methods.

Some automatic methods focus on specific syntactic features for translation evaluation. Jones (2000) utilizes linguistic information such as balance of parse trees, N-grams, semantic co-occurrence and so on as indicators of translation quality [4]. Brew C (1994) compares human rankings and automatic measures to decide the translation quality, whose criteria involve word frequency, POS tagging distribution and other text features [5].

Another type of evaluation method involves comparison of the translation result with human translations. Yasuda (2001) evaluates the translation output by measuring the similarity between the translation output and translation answer candidates from a parallel corpus [6]. Akiba (2001) uses multiple edit distances to automatically rank machine translation output by translation examples [7]. While the IBM BLEU method [8] and the NIST MT evaluation [9] compare MT output with expert reference translations in terms of the statistics of word N-grams. Melamed (2003) adopted the maximum matching size of the translation and reference as similarity measure for score [10]. Niβen (2000) scores a sentence on basis of scores of translations in a database with the smallest edit distance [11]. Yokoyama (1999) proposed a two-way MT based evaluation method, which compares output Japanese sentences with the original Japanese sentence for the word identification, the correctness of the modification, the syntactic dependency and the parataxis [12].

Another path of MTE is based on test suites. A weighted average of the scores for separate grammatical points is taken as the score of the system. The typological test covers vocabulary size, lexical capacity, phrase, syntactic correctness, etc. Yu (1993) designs a test suite consisting of sentences with various test points [13]. Guessoum (2001) proposes a semi-automatic evaluation method of the grammatical coverage machine translation systems via a database of unfolded grammatical structures [14]. Koh (2001) describes their test suite constructed on the basis of fine-grained classification of linguistic phenomena [15].

For the complexity of the problem of MT evaluation, many researchers are making efforts towards taxonomy of automatic MTE measures. This paper aims for a specification of the performance of MTE measures. The second section is a list of definitions and mathematical formulae of the measures of

MTE performance. Section 3 gives some examples of the MTE practice. A conclusion is given in the last section.

II. Measures for MTE Performance

The ISLE (2000) has made some efforts to a specification of performance of the MTE methods [3]. A list of desiderata demands at least the measure must be easy to define, clear and intuitive; must correlate well with human judgments under all conditions, genres, domains, etc.; must be 'tight', exhibiting as little variance as possible across evaluators, or for equivalent inputs; must be cheap to prepare; must be cheap to apply; and should be automated if possible.

Popescu-Belis (1999) argues that the MTE metrics should have its upper limit, lower limit, and should be monotonic in quality measure [15]. Papineni (2001), Yao (2002) and Melamed (2003) make studies on the correlation between human scoring and automatic evaluation results. On the whole, the literature contains rather few methodological studies of this kind [16]. The present measures for MTE analysis is far from enough as a solid work on MTE and its application. On basis of general examination theory [17], this section aims for a proposal for some criteria of the performance of MTE measures, which will give us a better understanding of the MTE task and its results.

Reliability is the most important issue in an examination, which is also true in MTE. The consistency of the scores obtained in different cases of evaluation of the same system, especially the consistency with human evaluation, reflects to some extent the reliability of the MTE result. As for different purposes, reliability can be described by the following metrics:

A. Consistency between MTE Methods or Test Suites

The same MT system may be scored by different methods (including manual evaluation) on the same test suite, or on various test suites via the same MTE method. If the scores are in real numbers, the consistency can be calculated by

$$rtt = \frac{\sum X_a X_b - (\sum X_a)(\sum X_b)/n}{\sqrt{\sum X_a^2 - (\sum X_a)^2/n}\sqrt{\sum X_b^2 - (\sum X_b)^2/n}} \quad (1)$$

where X_a and X_b refers to scores of the two MTE results; n is the number of test questions in the test suite; r_{tt} is the consistency between the two test results.

If the evaluators just assign a rank to the translation results, we can use Kendall's coefficient of concordance (multiple ranking correlation) to calculate the consistency of the K evaluators on n examination questions. The formula is

$$r_{tt} = \frac{SS_R}{\frac{1}{12}K^2(n^3 - n)} \quad (2)$$

where $SS_R = \sum (R - \bar{R})^2$ is the square sum of the ranks.

B. Reliability of the Evaluators

Consistency between the different human evaluators reflects how reliable the evaluators are and how difficult the evaluation task is. Consistency between the automatic evaluation method and human scores is a sign of the reliability of the automatic method. Correlation between different methods reflects whether the methods complement each other in some aspects and whether they could be integrated into a better measure. If the scores are continuous real numbers, the consistency can be calculated by the equation (1). While if the scores are rank-based, it can be calculated by Spearman rank correlation as

$$r_{tt} = 1 - \frac{6\sum D^2}{n(n^2 - 1)} \quad (3)$$

Where D is the difference between ranks of the same test by different evaluators; n is the sample size.

From the above discussion, we know that the consistency of MTE results from different methods or test suites, and between evaluators is a reflection of the reliability of the MTE measure. The higher the reliability, the more confident we have that various cases of MTE give us similar evaluation results. Besides the reliability of MTE methods, we also should know that the consistent evaluation is not biased from the truth of the applicability of the systems. The efficacy issue is thus an important aspect of the evaluation measures. The efficacy of MTE refers to the power or capacity of the evaluation to produce a desired effect for a task, which reflects the degree to which the system is helpful to the end users. We can have two kinds of efficacy metrics for a MT evaluation.

C. Test Suite Content Efficacy

Different contents of the test suites lead to different MTE results, which greatly influences the evaluation result. One system may perform better in one domain while worse in another. An example of this phenomenon will be given in section three. By efficacy we refer to the extent to which the MTE result uncovers the quality of the translation for the user's application. The efficacy, which is the consistency between the MTE result and the user's experience of the MT systems' performance, can also be described by the consistency with the human result or a standard test result with a domain test suite calculated from equation (1) or (2). For example, the domain of the test corpus has much influence on the evaluation results. So for a specific purpose, the evaluator must choose carefully the contents of the test suite in both domain and difficulty.

D. Standard-Associated Efficacy

This refers to the consistency of the evaluation result and another independent test on a gold standard test suite or by a MTE as standard reference (e.g. the human evaluation result). This is often described by the correlation coefficient of the two results. In study of the performance of MTE methods, the standard-associated efficacy is a measure to estimate the virtue of the result. Of course, both the standard MTE measure and the test suite have to be carefully selected so as to avoid a bias of the analysis.

E. Discriminancy of a Measure

For two measures f, g on domain Ψ, we say f is more discriminating than g if there exist $a,b \in \Psi$ such that $f(a) > f(b)$ and $g(a) = g(b)$, and there exist no $a,b \in \Psi$ such that $g(a) > g(b)$ and $f(a) = f(b)$.

The discriminancy of a method reflects the ability to distinguish between minor differences between translation qualities. For a test with higher discriminancy, a better system should be scored higher, and vice versa. And the MTE result should be fine grained so that even small change in the translation quality could be correctly reflected.

The discriminancy of a test can be calculated on basis of the MTE result, as follows:

$$D = (X_H - X_L) / N(H - L) \tag{4}$$

In the equation, XH/XL is the score for the best/worst system; N is the number of test sentences, while H/L is highest / lowest possible score of the test.

F. Difficulty of the Test Questions

This refers to the degree of the difficulty of the test suite, which has a great influence on the test result. The difficulty of the test changes the distribution, discriminancy and dispersion of the test results. For example, if the test suite is too difficult that none of the systems output the right answer, we cannot distinguish between system via the MTE result, also is the case if the test is too easy. The difficulty of the test questions can be calculated as

$$P = (\overline{X} - L)/(H - L) \tag{5}$$

In the equation $\overline{X}$ is the average score of the systems, while H/L is the highest/Lowest possible score for the test. The difficulty of the test question is closely interrelated with the discriminancy, efficacy, and other characteristics of the evaluation. Usually a difficulty around 0.5 is helpful discriminating the systems to be scored.

G. Relationship between Size and Efficacy of the MTE

The size of the MTE refers to the number of questions in the test suite. The larger the size of the test suite is, the higher is its reliability and efficacy. The relationship can be describe by

$$r_{nn} = \frac{n r_{ll}}{1 + (n-1) r_{ll}} \tag{6}$$

In the equation rnn denotes the correlation coefficient when the size of test suite is enlarged to n times, while rll is the original coefficient. Correspondingly, we can calculate the expected length of the test suite if we hope to improve the reliability and efficacy from rll to rnn, as follows

$$n = \frac{r_{nn}(1 - r_{ll})}{r_{ll}(1 - r_{nn})} \tag{7}$$

In this section, a preliminary proposal of metrics for the property of the MTE measures and its test suite has been given. This is an incomplete list while in our work for MTE, we need to use specific metrics based on the specific conditions. The next section, based on a MTE visualization method, gives some examples of the metrics proposed in this section. The examples will give us a better view of the characteristics of an MTE method and thoughtway of organizing a test suite.

III. Visualization and Experiment of MTE

A list of metrics has been proposed for the organization of MTE in both evaluation measures and construction of a test suite in the last section. Several instances are brought forth in this section for a demonstration of the metrics. The correlation of MTE results between different evaluators and various test suites is analyzed. The variance of MTE results shows that for a highly reliable MTE practice, both the MTE measure and the test suite should be well chosen.

A. Analysis of Human MTE Results

The human evaluation results in [18] on eight English-to-Japanese MT systems are listed in the appendix. Two evaluators score the systems on a 5 point scale with intelligibility and accuracy. Based on the measures proposed in the last section, we make an analysis of the characteristics of the human MTE results.

For consistency between MTE results from different measures (accuracy and intelligibility), different evaluators and different test suites, from equation (1) and (2), based on the data in Table A1 and A2 in the appendix, we get the correlation coefficients in table 1 which shows the reliability and efficacy of the MTE. For the different parts of the test suite, we have their discriminancy and difficulty on intelligibility calculated from equation (4) and (5), which can give us a hint of their influence on the MTE result.

TABLE I
Correlation Coefficients for MTE Results from Different Measures, Evaluators and Test Suites

Item1	Item2	Other conditions	Correlation option	Correlation coefficient
Intelligibility	Accuracy	Overall average scores	Pearson	0.998
Intelligibility	Accuracy	Overall average scores	Spearman	1.000
Evaluator A	Evaluator B	Intelligibility for all 300 sentences	Pearson	0.991
Evaluator A	Evaluator B	Accuracy for all 300 sentences	Pearson	0.998
Evaluator A	Evaluator B	Accuracy for all 300 sentences	Spearman	0.994
Sent#1-100	Sent#101-200	Intelligibility by evaluator A	Pearson	0.964
Sent#1-100	Sent#201-300	Intelligibility by evaluator A	Pearson	0.968
Sent#101-200	Sent#201-300	Intelligibility by evaluator A	Pearson	0.945

TABLE II
DISCRIMINANCY AND DIFFICULTY OF TEST SUITES WITH INTELLIGIBILITY WITH DIFFERENT EVALUATORS

Sentences	Evaluator	Discriminancy	Difficulty
1-100	A	0.23	0.50
1-100	B	0.31	0.44
101-200	A	0.23	0.56
101-200	B	0.31	0.62
201-300	A	0.24	0.43
201-300	B	0.34	0.53
All 300	A	0.23	0.50;
All 300	B	0.23	0.53

Above are some instances on consistency, discriminancy, and difficulty of an MTE task. The influence of test suite size is also studied.

The following section will give an example of content efficacy, difficulty-discriminancy relationship based on visualization of the MTE result.

B. *Visualization of Performance of MT Evaluation Scores*

The BLEU and NIST evaluation methods have been popular in MT evaluation research. We make MTE experiments using these methods and for a better understanding of the result, visualize the data in a chart as shown in Fig. 1. Fig. 1 exhibits the MTE results with a test suite of 1019 sentences selected from the 863 National High-tech Program MTE corpuses for Chinese-to-English translation. Four systems are evaluated with the BLEU method. The chart is produced in the following steps:

STEP 1: Get the machine translations by the four systems;

STEP 2: With the human translations as reference set, calculate the BLEU score for each sentence;

STEP 3: Sort the scores for each system in ascending order;

STEP 4: Put the sentence number on the X-axis, and BLEU score on the Y-axis to get a line for each system.

From the Fig. 1, we can draw the following conclusions about the MTE performance:

1) The longer the N-gram, the more difficult the test is, and the lower the scores obtained by MT systems. The lines in the figure are shifting to the right side when the N-gram shifts from unigram to 5-gram. The most northwest line represents the performance of the best system.

2) The gap between the lines changes with the difficulty of the test. As seen in the first figure of unigram scores, the lines representing systems #2, #3, and #4 are very near to each other, while the gap become much larger between the trigram lines. This is because the difficulty of the test influences the discriminancy of the measure, which is also observed in a typological MTE experiment (Yu 1993).

For the same typological evaluation tool, we adopted two test suites, separately consisting of 124 sentences and 293 sentences. The sentences in the first suite are difficult to translate while the other is easier. Four machine translation systems are scored on the test suites, as shown in Table III.

For the difficult test suite, the systems are distributed in a two-peak shape, which is not suitable for distinguishing the systems. The test suite #1 is much better, because the systems are distributed more evenly in the whole range of score. This experiment shows us the influence of selection of test questions of appropriate difficulty.

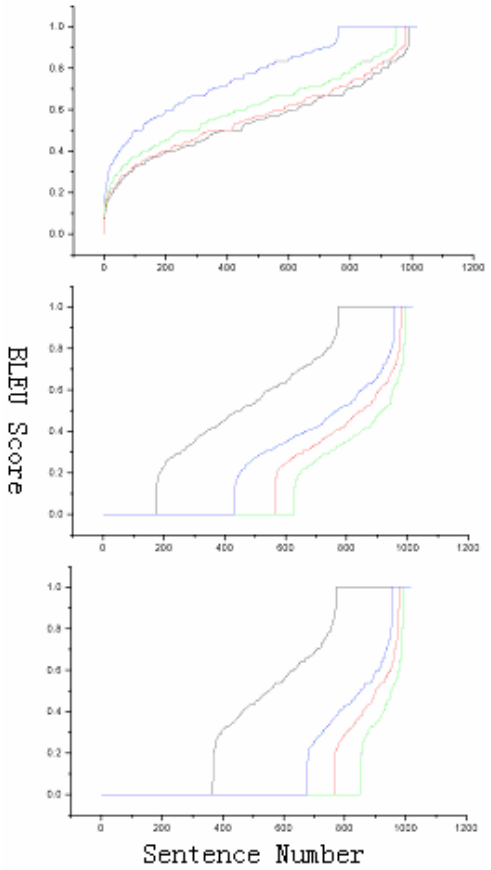

Fig. 1 BLEU scores on 1019 sentences with unigram, trigram and 5-gram

TABLE III
TYPOLOGICAL TEST RESULT OF FOUR MT SYSTEMS ON TWO TEST SUITES

	MTS#1	MTS#2	MTS#3	MTS#4
Suite#1	87	65	50	27
Suite#2	30	3.4	2.0	2.0

IV. CONCLUSION

Many measures have been proposed for machine translation evaluation (MTE) while little work has been done on the performance of MTE methods. This paper is an effort towards MTE performance analysis. After a general frame proposed for description of the MTE measure and the test suite, some instances are given including whether the automatic measure is consistent with human evaluation, whether MTE results from various measures or test suites are consistent, whether the content of the test suite is suitable for performance evaluation, the degree of difficulty of the test suite and its influence on the MTE, the relationship of MTE result significance and the size of the test suite, etc. For a better clarification of the frame, a visualization method is introduced for presenting the results. This paper aims for a framework for construction of MTE task in measure and test suite.

APPENDIX

Here presents the human evaluation results from [18] on eight English-to-Japanese MT systems. Two popular metrics are used in the human evaluation: intelligibility and accuracy. The evaluators score the systems on a 5 point scale. For more detailed data please refer to the original paper.

TABLE A-I
OVERALL ENGLISH-TO-JAPANESE AVERAGE SCORES
(POSSIBLE SCORE 5 POINTS)

MT System	EJsys-1	EJsys-2	EJsys-3	EJsys-4
Intelligibility	2.33	3.39	3.42	3.32
Accuracy	2.42	3.60	3.62	3.45
MT System	EJsys-5	EJsys-6	EJsys-7	EJsys-8
Intelligibility	3.00	3.01	3.11	2.87
Accuracy	3.13	3.15	3.27	2.99

ACKNOWLEDGEMENTS

We would like to show our deepest gratefulness to Dr. Darwin for our citation of his research results.

REFERENCES

[1] ALPAC (1966). Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, National Research Council. Washington, D.C. National Academy of Sciences.
[2] White J.S., T.A. O'Connell (1994). The ARPA MT evaluation methodologies: evolution, lessons, and further approaches. Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas, Columbia, MD. pp. 193-205.
[3] ISLE (2000). The ISLE classification of machine translation evaluations, draft 1. A document by the International Standards for Language Engineering. See http://www.isi.edu/natural-language/mteval/
[4] Jones Douglas A., Gregory M. Rusk (2000). Toward a scoring function for quality-driven machine translation. Proceedings of the International Conference on Computational Linguistics. pp. 376-382.
[5] Brew C, Thompson H.S. (1994). Automatic evaluation of computer generated text: a progress report on the TextEval project. Proceedings of the Human Language Technology Workshop. pp. 108-113.
[6] Yasuda Keiji, Fumiaki Sugaya, et al (2001). An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. MT Summit Conference, Santiago de Compostela. pp. 373-378.
[7] Akiba Yasuhiro, Kenji Imamura, Eiichiro Sumita (2001). Using multiple edit distances to automatically rank machine translation output. MT Summit Conference, Santiago de Compostela. pp. 15-20.
[8] Papineni K., S.Roukos, T.Ward, W.-J. Zhu (2001). BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176(W0109-022), IBM Research Division, T.J.Watson Research Cente. See http://domino.watson.ibm.com/library/
[9] NIST (2002). The NIST 2002 machine translation evaluation plan. A document by the National Institute of Standards and Technology. See http://www.nist.gov/speech/tests/mt/doc/2002- MT-EvalPlan- v1.3.pdf
[10] I.D.Melamed, R.Green, J.P.Turian (2003). Precision and recall of machine translation. NAACL/Human Language Technology 2003, Edmonton, Canada.
[11] S. Niβen, F. J. Och, et al (2000). An evaluation tool for machine translation: fast evaluation for MT research. 2nd International Conference on Language Resources and Evaluation. Athens, Greece. pp. 39-45.
[12] Yokoyama S. et al. (1999). Quantitative evaluation of machine translation using two-way MT. Proceeding of Machine Translation Summit VII. pp. 568--573.
[13] Yu Shiwen (1993). Automatic Evaluation of Quality for Machine Translation Systems. Machine Translation, 8. pp. 117-126.
[14] Guessoum A., R. Zantout (2001). Semi-automatic evaluation of the grammatical coverage of machine translation systems. MT Summit Conference , Santiago de Compostela. pp. 133-138.
[15] Popesc-Belis (1999). Evaluation of natural language processing systems: a model for coherence verification of quality measure. Marc Blasband and Patrick Paroubek, editors, A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment. ELSE Project LE4-8340 (Evaluation in Language and Speech Engineering).
[16] Yao Jianmin, Ming Zhou et al (2002). An automatic evaluation method for localization oriented lexicalised EBMT system. The 19th International Conference on Computational Linguistics, Taipei. pp. 1142-1148.
[17] Zhang Minqiang (2003). Education measurement. 2nd edition. People's Education Press Beijing. (in Chinese). pp. 98-132.
[18] Darwin, M. (2001). Trial and Error: An Evaluation Project on Japanese-English MT Output Quality. In Maegaard, B. (Ed.). MT Summit VIII, 77-82. Santiago de Compostela, Spain: European Association for Machine Translation (EAMT).