Computational Linguistics and Chinese Language Processing
Vol. 11, No. 3, September 2006, pp. 297-314

Performance Analysis and Visualization of Machine Translation Evaluation

Jianmin Yao[*+], Yunqian Qu[+], Qiang Lv[+], Qiaoming Zhu[+], and Jing Zhang[**]

Abstract

Automatic translation evaluation is popular in development of MT systems, but further research is necessary for better evaluation methods and selection of an appropriate evaluation suite. This paper is an attempt for an in-depth analysis of the performance of MT evaluation methods. Difficulty, discriminability and reliability characteristics are proposed and tested in experiments. Visualization of the evaluation scores, which is more intuitional, is proposed to see the translation quality and is shown as a natural way to assemble different evaluation methods.

Keywords: Machine Translation, Performance, Analysis, Visualization, Clustering, Natural Language Processing

1. Introduction

Machine translation (MT) evaluation activities have accompanied MT research and system development. The ALPAC report [ALPAC 1966], which has greatly influenced machine translation research activities, is the first historical MT evaluation activity. With new developments in natural language processing technology coming in the 1990s, the black-box evaluation has been instantiated by the methodology of DARPA [Doyon *et al*. 1998], 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 [ISLE 2000].

Since manual evaluation is labor-intensive and time-consuming, many researchers are making efforts towards reliable automatic MT evaluation methods. A problem is that the methods cannot be characterized by precision and recall as in other natural language

[*] Southeast University, Nanjing, China 210096 Tel: +86-512-68880263
E-mail: jyao@suda.edu.cn
[+] School of Computer Science and Technology, Soochow University, Suzhou, China, 2150006
[**] South China University of Technology, Guangzhou 510641, China

[Received March 21, 2006; Revised October 11, 2006; Accepted October 16, 2006]

processing activities such as POS tagging or phrase identification. A new quality system is necessary.

This paper aims for performance analysis and better illustration of machine translation evaluation, which can help developers know about the improvement in the quality of their system, and help users easily distinguish between MT systems. Section 2 reviews related research in the MT field and its evaluation. Section 3 studies the metrics and experiments for comparison of MT evaluation methods. Section 4 proposes an algorithm for visualizing the MT system quality, and draws a dendrogram for the systems by clustering. A conclusion is given in the last section.

2. Related Work

MT evaluation had not been a very powerful aid in machine translation research until automatic evaluation methods were broadly studied. Now, different heuristics are employed for automatic MT evaluation. This section gives a brief review of the main automatic MT evaluation methods and studies on the performance of these methods.

2.1 Automatic Evaluation Methods

Some automatic methods focus on specific syntactic features for translation evaluation. [Jones and Galliers 1993] utilizes linguistic information such as the balance of parse trees, N-grams, semantic co-occurrence, and other information as indicators of translation quality. A balanced tree was a negative indicator of Englishness, probably because English is right-branching. Other factors are also utilized in translation evaluation for their indication of the language quality. [Brew and Thompson. 1994], whose criteria involve word frequency, POS tagging distribution and other text features, compares human rankings and automatic measures to decide the translation quality. These linguistic features are extracted as a reflection of the overall translation quality.

Another type of evaluation method involves comparison of the translation result with human translations. [Keiji *et al*. 2001] evaluates the translation output by measuring the similarity between the translation output and translation answer candidates from a parallel corpus. [Yasuhiro *et al.* 2001] uses multiple edit distances to automatically rank machine translation output by translation examples. While the IBM BLEU method [Papineni *et al.* 2001] and the NIST MT evaluation [NIST 2002] compare MT output with expert reference translations in terms of the statistics of word N-grams. [Melamed *et al.* 2003] adopts the maximum matching size of the translation and the reference as the similarity measure for the score. [Niben and Och 2000] scores a sentence on the basis of scores of translations in a database with the smallest edit distance. [Yokoyama *et al.* 1999] proposes a two-way MT based evaluation method, which compares output Japanese sentences with the original

Japanese sentence for word identification, the correctness of the modification, the syntactic dependency and the parataxis.

Another path of MT evaluation 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. [Guessoum and Zantout 2001] proposes a semi-automatic evaluation method of the grammatical coverage machine translation systems via a database of unfolded grammatical structures. [Koh *et al.* 2001] describes their test suite constructed on the basis of fine-grained classification of linguistic phenomena.

2.2 Performance of an Automatic Evaluation Method

The ISLE has made some efforts to develop a specification of performance for the MT evaluation methods [ISLE 2000]. A list of the desiderata demands that at least the measure: 1) must be easy to define, clear, and intuitive; 2) must correlate well with human judgments under all conditions, genres, domains, etc.; 3) must be 'tight', exhibiting as little variance as possible across evaluators, or for equivalent inputs; 4) must be cheap to prepare; 5) must be cheap to apply; 7) should be automated, if possible. These criteria give a broad coverage of the characteristics of the evaluation methods, but further work is needed to measure them in a consistent and objective way.

[Popescu-Belis 1999] argues that the MT evaluation metrics should have its upper limit, lower limit, and should be monotonic in quality measure. The above measures are qualitative attributives of MT evaluation methods. If it can further be automated, it will help the researchers find a much easier and consistent way to compare different systems.

Only recently, researchers began quantitative studies. Some recent works include [Forner and White 2001] on the correlation between intelligibility and fidelity and noun compound translation. [Papineni *et al.* 2001] and [Melamed *et al.* 2003] study the correlation between human scoring and automatic evaluation results. After DARPA took the BLEU method as the evaluation method for MT systems, the correlation between human and machine translation evaluation has become a standard criterion of MT quality scoring, though many researchers are arguing against its efficacy.

On the whole, methodological study of automatic evaluation methods has just started and needs to be further deepened. This paper is an attempt to refine the correlation measures and justify their usage in machine translation evaluation. The following section aims for a proposal of some criteria of the performance of MT evaluation measures, which will give linguists a better understanding of the MT evaluation task and its results.

3. MT Evaluation Performance Analysis

Up to now, the analysis of MT evaluation methods has remained a preliminary comparison of human and automatic scores. Further study is important to propose better evaluation measures and better understanding of the automatic evaluation results. This paper is an endeavor to provide more details of MT evaluation methods. A list of quantitative measures on basis of education measurement theory [Wang 2001] is proposed in section 3.1, and experimental study of the measures is made in section 3.2.

3.1 MT Evaluation Performance Metrics

3.1.1 Consistency and Reliability

Reliability is the most important issue in MT evaluation. Correlation is often utilized for description of the consistency between different score results as by various MT evaluation methods or test suites, as follows:

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

where X_a and X_b refer to scores of the two MT evaluation results; n is the number of test questions in the test suite; r_{tt} is the consistency between the two test results. If the scores are rank-based, reliability can be calculated by Spearman rank correlation as

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

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

The correlation coefficient between the automatic results and the human results shows the reliability of the automatic evaluation method. On the other hand, if the correlation is between two automatic results, it shows consistency between the two methods, thus, also showing whether they can compensate for each other.

3.1.2 Discriminability

The discriminability of an MT evaluation method reflects the ability to distinguish between minor differences in translation qualities. For a test with higher discriminability, a better system should be scored higher, and vice-versa. The MT evaluation result should be fine-grained so that even small changes in the translation quality could be correctly shown. The discriminability of a test can be calculated on the basis of the MT evaluation result, as

follows:

$$D = (X_H - X_L)/(H - L). \quad (3)$$

In the equation, X_H / X_L is the score for the best/worst system; H / L is highest/lowest possible score of the test.

3.1.3 Difficulty

The difficulty refers to the degree of the difficulty of the test, which has a great influence on the test result. The difficulty of the test changes the distribution, discriminability, and dispersion of the scores. For example, if the test is so difficult that none of the systems outputs the right answer, one cannot distinguish between systems via the MT evaluation result. This is also the case if the test is too easy. The difficulty of the test questions can be calculated as

$$P = (\overline{X} - L)/(H - L). \quad (4)$$

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 discriminability, efficacy, and other characteristics of the evaluation. According to education measurement theory, a difficulty of around 0.5 is helpful for discriminating the systems to be scored [Wang 2001].

In the section above, a proposal of performance metrics for MT evaluation measures and the proposal's test suite has been given. These metrics help in analyzing the efficacy of the evaluation methods. The next section gives some experimental examples of the evaluation performance, which verifies the metrics mentioned above.

3.2 Experiments on MT Evaluation Performance

3.2.1 Test of Consistency, Discriminability and Difficulty

Since the MT evaluation performance metrics proposed in section 3.1 are language-independent, they can be applied to evaluation results in any language. The open source of human evaluation results in [Darwin 2001] on eight English-to-Japanese MT systems is taken for analysis in this section. The authors of this paper do research on the open source evaluation results for two reasons: it is available to any researcher, and thus is easier to duplicate the experiment and analysis; also, the open source data is appropriate in data size and reliability and saves time for more manual work. In the experiment in [Darwin 2001], two evaluators score 8 systems on a 5-point scale showing intelligibility and accuracy. The experimental setup and details are listed in the appendix following this paper. Based on the measures proposed in the last section, this paper's authors make an analysis of the

characteristics of the MT evaluation results.

The first experiment is to test the consistency between MT evaluation results from different measures (accuracy and intelligibility), different evaluators, and different test suites. According to equation (1) and (2), based on the data in Table A1 and A2 in the appendix, one gets the correlation coefficients in Table 1, which shows the correlation coefficients for the MT evaluation results.

In Table 1, rows 1 and 2 show a consistency between MT evaluation results by metrics of intelligibility and accuracy. Rows 3 to 5 show consistency between two human evaluators A and B. Rows 6 to 8 show consistency between MT evaluation results by the same evaluator A on different parts of the 300 hundred sentences.

Table1. The correlation coefficients for the MT evaluation results achieved from different evaluation measures of intelligibility and accuracy), different evaluators (named as A and B) and various test suites (3 parts of 300 sentences).

Item1	Item2	Other conditions	Correlation option	Correlation
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 evaluator A	Pearson	0.964
Sent#1-100	Sent#201-300	Intelligibility evaluator A	Pearson	0.968
Sent#101-200	Sent#201-300	Intelligibility evaluator A	Pearson	0.945

From the definition in section 3.1, one knows that correlation between different human evaluation results is an upper bound of automatic MT evaluation performance. Correlation with a human evaluation also reflects the reliability of the automatic evaluation result. As seen in Table 1, all correlation coefficients are higher than 0.9, which is a strong hint of consistency. First, the correlation coefficient between intelligibility and accuracy are 0.998 and 1.000, respectively. This reminds researchers that the two metrics have quite similar scores, and a researcher may just measure one and know the other by regression analysis. Second, the coefficient is also high for correlation between different evaluators and different parts of the test suite, which shows that scores from both evaluators and from different sentences agree with each other on the whole. This is also the case for automatic measures. From previous study, one knows that some automatic evaluation methods are highly correlated with human evaluation, for example, a correlation of around 0.99 for BLEU and NIST [NIST 2002]. GTM (General Text Matching) claims a 0.8 level which is better than BLEU on the same test suite

[Melamed *et al*. 2003]. The difference between [Melamed *et al.* 2003] and [NIST 2003] gives researchers a strong signal that consistency is a key factor, but not the only one, in MT evaluation performance.

Another key issue seen from Table 1 is that rows 6 to 8 have a lower correlation coefficient than the rows above. It reminds the researchers that different metrics, such as intelligibility and accuracy, different evaluator A and B, as in the experiments, have a higher correlation coefficient than the same evaluator on different test suites with the same MT evaluation measure of intelligibility. Thus, the difficulty and size of the test suite is another key factor in MT evaluation. The following is further analysis of the influence of test suites.

3.2.2 Influence of the Test Suite

For the different parts of the test suite, the researchers have the discriminability and difficulty of intelligibility calculated using equations (3) and (4), which can give one a hint of the reason for their influence on the MT evaluation results.

Table 2. Discriminability and difficulty of test suites with intelligibility by different evaluators. The 300 sentences in the test suite are divided into 3 parts and evaluated with intelligibility separately.

Sentences	Evaluator	Discriminability	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	0.32	0.53

From Table 2, one can see that different parts of a test set may have different difficulty and discriminability levels. Since all evaluation tasks need better discriminability capability, the evaluator needs to pick out proper test sentences for the evaluation task. Taking evaluator A as an example, the difficulty of different parts of the test suites are 0.50 for sentences 1-100, 0.56 for sentences 101-200, and 0.43 for sentences 201-300. The different difficulty levels led to different correlation coefficients between different parts of the test suites. For example, sentences 101-200 and 201-300 differ greatly in difficulty, and the difference in correlation coefficients is also lower in Table 1 (only 0.945). Another factor found in Table 2 is that the results of evaluators A and B have different discriminability, the former about 0.23, and the latter 0.32. That means their evaluation score has a different distribution style. In fact, this

phenomenon has a vital influence on the correlation coefficient of two evaluation results, which is highly related to the evaluation result.

The above study of the evaluation performance is made on a public-available Japanese test suite. One does have to notice that the evaluation performance measures are language-independent, which ensures the applicability of the method to the Chinese language, or other language pairs.

To study other performance measures, a test on a Chinese suite is made below.

As described above, besides the difficulty and discriminability, another key factor for the test suite is the size. The larger the size of the test suite, the more stable and reliable the MT evaluation result becomes. Taking the popular automatic evaluation methods of BLEU and GTM as example, the influence of the size of the test suite, *i.e.* the number of sentences it contains, is tested using the 863 National High-tech Program MT evaluation corpus. This corpus is widely used for the evaluation of MT systems in mainland China. The corpus contains 1019 sentences. An experiment was carried out on the BLEU and GTM methods to test the influence of the size of a test suite for an English-to-Chinese translation system. The result is shown in Figure 1.

When the test suite is small, *i.e.* there is small number of sentences in the test suite, the MT evaluation score fluctuates violently. While when the test suite contains more than 80 sentences, the fluctuation becomes less violent and goes flat after 400 sentences. Figure 1 shows that the two methods have similar tendencies, which shows that they have similar demands of the test suite size.

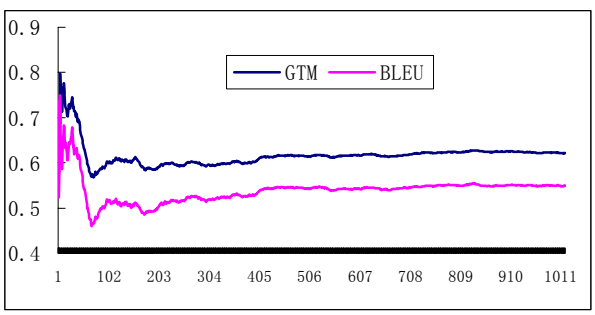
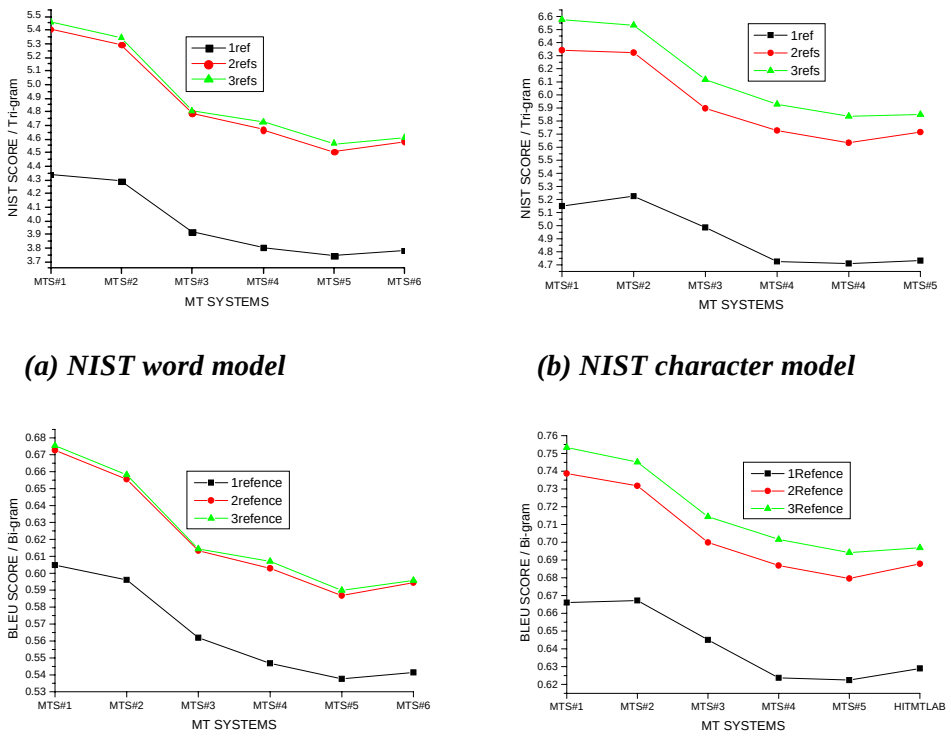


Figure 1. MT evaluation score changes with the increasing of sentence in the test corpus. The score stabilizes when the corpus contains more than 400 sentences. The experiment is made on an English-to-Chinese MT system.

Another aspect of the influence of the size of test suite can be revealed by the number of reference translations in NIST and BLEU evaluation. To get a higher quality of evaluation result, the BLEU and NIST methods can have multiple reference translations. Figure 2 shows the influence of the number of reference translations on BLEU and NIST evaluation results.



(a) NIST word model ***(b) NIST character model***

(c) BLEU word model ***(d) BLEU character model***

Figure 2. Evaluation results with different number of reference translations by (a) NIST word model, (b) NIST character model, (c) BLEU word model and (d) BLEU character model for 6 English-Chinese MT systems. The word model calculates the MT evaluation scores in terms of Chinese words, while the character model is in terms of Chinese characters.

The BLEU and NIST evaluations are implemented with two different language models: The character model, which takes Chinese characters as unit of scoring, while the word model takes the Chinese word as the unit. The Chinese sentences are segmented into words by a Chinese segmentor (which was developed at Harbin Institute of Technology, http://ir.hit.edu.cn). In BLEU and NIST evaluation, one can see that the scores go up with the increasing number of reference translations. Compared to the character model, the word model scores saturate faster with an increasing number of references, which means it has a lower demand for references. This is also the case for the BLEU models. A possible reason for this phenomenon is that a word is not easy to be matched in extra-translation reference, while new characters come out even after a big number of references. This experiment gives researchers a hint that synonyms can improve the performance of similarity-based MT evaluation methods such as BLEU and NIST.

4. Visualization of MT evaluation scores and system clustering

MT evaluation has been extensively studied in recent years. However, the various MT evaluation methods just render a score for each system or translation sentence. The score scales also vary among methods. The BLEU and GTM score has a value between 0 and 1. NIST has a lower bound of 0 with no upper bound. The manual evaluation of fidelity and accuracy usually has discrete quality levels. This makes it quite ambiguous to understand the meanings of the scores. This section intends to make it easier to understand the MT evaluation scores by visualizing the scores of evaluation results.

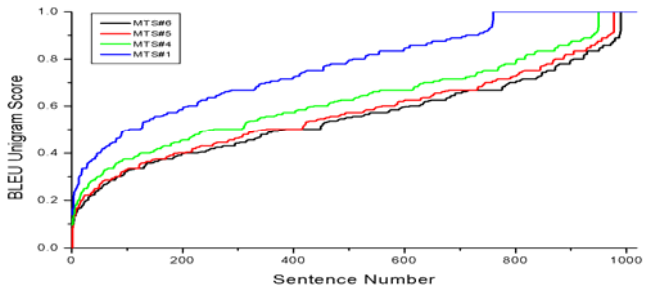
4.1 Visualization of MT Evaluation Scores

The BLEU and NIST evaluation methods have been popular in MT evaluation research. This research project makes MT evaluation experiments using these methods for a better understanding of the result. The MT evaluation data is visualized in the diagram as shown in Figure 4. Figure 4 exhibits the MT evaluation results with the test suite of 1019 sentences selected from the 863 National High-tech Program MT Evaluation corpus for machine translation, as introduced in section 3.2. Four systems are evaluated with the BLEU method. The diagram is produced with the algorithm in Figure 3.

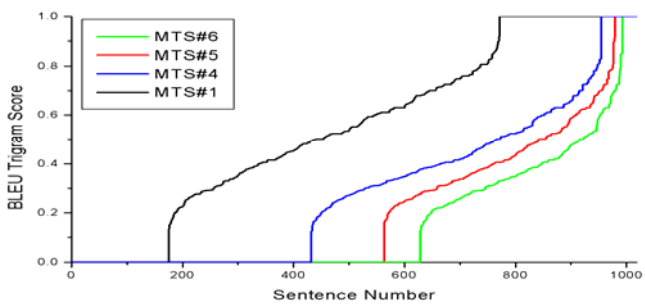
Algorithm: Visualization of system scores by plotting lines in a diagram

```
1:  INPUT: T←{Ti: t∈Ti, t is a translation by MT system MTSi}
2:  //Process the MT translation and get the BLEU scores
3:  For each machine translation system MTSi do
4:  For each translation t∈Ti by machine translation systems MTSi do
5:      Score{t}←{sti| sti is the BLEU score of the translation ti}
6:  End for
7:  //Plot a line of the BLEU scores for each MT system
8:  Score{t}←Score{t} {the BLEU scores sorted in ascending order}
9:  For i=1 to |T| {number of items in the translation set T} do
10:     Plot a point (i,sti) in the diagram
11: End for
12: End for
13: Output: a diagram in which every MT system is presented with a curve
```

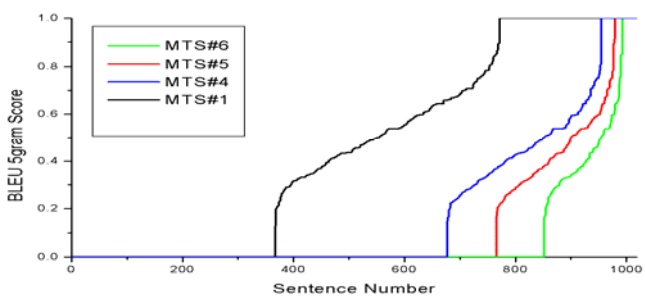
Figure 3. Algorithm: Visualization of system scores by plotting lines in a diagram



(a) BLEU 1-gram



(b) BLEU 3-gram



(c) BLEU 5-gram

Figure 4. Machine translation evaluation scores of 4 MT systems on 1019 sentences with (a) 1-gram, (b) 3-gram and (c) 5-gram BLEU method. Each line manifests the quality performance of a MT system. A line on the left and upper stands for a system with higher translation quality.

Figure 4 is the diagram from the algorithm for visualization of system scores, in which each system is represented by a line drawn according to the scores of the translations. From the lines of MT systems, one can draw the following conclusions about the MT evaluation 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 diagram shift to the right side when the N-gram shifts from unigram to 5-gram. The leftmost 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 Figure 4(a) of the unigram scores, the lines representing systems #2, #3, and #4 are very near to each other, while the gap becomes much larger between the trigram lines in Figure 4(b). This is because the difficulty of the test influences the discriminability of the evaluation.

The visualization method is based on NIST, BLEU or a similar MT evaluation score, but is more intuitional and easier to understand. On the one hand, the evaluation is not only presented for the whole system, but also each translation; on the other hand, the tendency of the lines manifests the quality characteristics of MT systems, while the gap represents the difference. From the diagram, one can directly see the difficulty and discriminability of the MT evaluation. This has fully taken advantage of the diagrams over pure numbers.

4.2 System Clustering Based on Various MT Evaluation Scores

The above section presents a diagram presenting the evaluation scores of the MT systems, which shows the translation quality of several systems. To make the quality difference clearer, system clustering is utilized for visualizing the distances of MT systems in respect to translation quality in this section. This process involves calculating the distances of translation quality, as shown in the algorithm of Figure 5.

The MT systems are evaluated by several manual and automatic evaluation methods. The evaluation methods are: F-measure of intelligibility and accuracy, error typology scoring ET as in [Guessoum and Zanout 2001], separate linguistic points as in [Yu 1993], BLEU word model, NIST word model, language model probability, edit distance and DICE coefficient as in [Yao *et al.* 2002]. As different evaluation methods have different value scopes, the scores as in step 3 to step 9 of the algorithm have been normalized. After the normalization, the value of MT scores varies between 0~1. The normalized scores are shown in Table 3. The clustering dendrogram is shown in Figure 6.

The methods introduced in this experiment are as follows: 1) F-measure is the F1 measure, which integrates the manual metrics of intelligibility and fidelity. 2) ET is a weighted sum of scores from different Types of Errors. 3) SLP comes from the automatic scoring based on a Separate Language Points, which measures different linguistic phenomena based on a human-edited test suite. 4) BLEUW and NISTW is the BLEU/NIST score measured on Chinese word model, which takes words instead of characters as the unit of

comparison. 5) LM is the score from a language model, specifically a bi-gram model in this article. 6) EDist is a score from edit distance between the translation and the reference. 7) DICE is a score based on the DICE coefficient of the translation and the reference.

Algorithm: Similarity histogram-based incremental MT system clustering

INPUT: Score{MTSi}←{sco_mti| BLEU scores of translations by MTSi }
// Normalize the MT BLEU scores
For each machine translation system MTSi **do**
max{sco_mti}←sco_mti{the maximum BLEU score in Score{MTSi}
min{sco_mti}←sco_mti{the minimum BLEU score in Score{MTSi}
For each sco_mti **do**

$$sco\_mti = \frac{sco\_mti - \min\{sco\_mti\}}{\max\{sco\_mti\} - \min\{sco\_mti\}}$$

End for
End for
//Similarity histogram-based incremental MT system clustering
L←Empty list{Cluster list}
For each MT system mts **do**
    For each cluster c in L **do**
        HRold = HRc
        Simulate adding mts to c
        If (HRnew ≧ HRold) OR ((HRnew > HRmin) AND (HRold – Hrnew < ε)) **then**
            Add mts to c
        End if
    End for
    If mts was not added to any cluster **then**
        Create a new cluster c
        Add mts to c
        Add c to L
    End if
End for
Output: a histogram of MT systems

Figure 5. Similarity histogram-based incremental MT system clustering

Table 3. Normalized scores of MT systems by various MT evaluation methods. The scores are obtained with various MT evaluation methods that have different score scopes. The scores are normalized for system clustering.

MTS	F-measure	ET	SLP	BLEUW	NISTW	LM	EDist	DICE
MTS#1	1.00	0.92	1.00	1.00	1.00	1.00	0.92	1.00
MTS#2	0.84	1.00	0.85	0.78	0.78	0.46	1.00	1.00
MTS#3	0.60	0.71	0.45	0.22	0.24	0.18	0.23	0.27
MTS#4	0.44	0.71	0.20	0.22	0.15	0.14	0.69	0.80
MTS#5	0.16	0.38	0.10	0.00	0.00	0.00	0.00	0.00
MTS#6	0.00	0.00	0.00	0.11	0.03	0.11	0.08	0.20

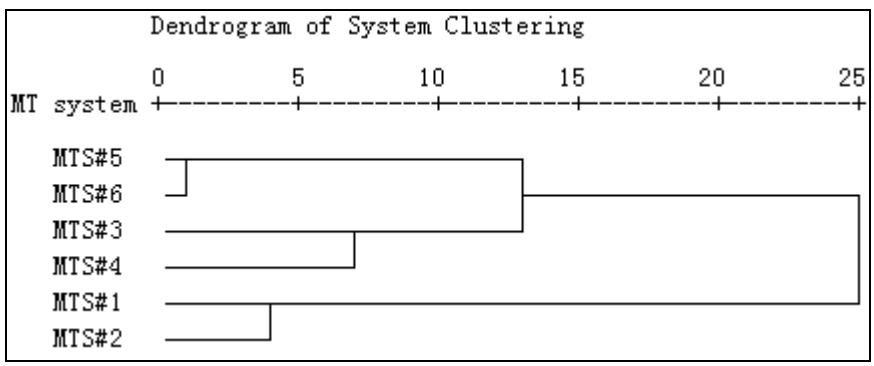


Figure 6. Cluster chart and distance between clusters of 6 MT systems. Systems are clustered according to their quality difference.

The cluster chart in the dendrogram in Figure 6 is a clear representation of the machine translation system quality. As seen from this dendrogram, the systems MTS#5 and MTS#6 are very similar to each other and are clustered first. The MTS#1 and MTS#2 have a second smallest difference. After MTS#3 and MTS#4 are clustered as one, the clustering goes on, and all the systems cluster into a binary tree. This clustering dendrogram is an easy way for a clear presentation of MT system quality based on ensemble of various evaluation scores.

5. Conclusion

This paper is an effort towards MT evaluation performance analysis and better rendering of MT evaluation results. After a general framework is proposed for the description of MT evaluation measure and the test suite, some instances are given including whether the automatic measure is consistent with human evaluation, whether MT evaluation 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 MT evaluation, the relationship of MT evaluation result significance and the size of the test suite, etc. For better clarification of the framework, a visualization method is introduced for presenting the results. The MT evaluation performance analysis can help a lot in designing test suites for different MT evaluation methods. The visualization method, on the one hand, gives an intuitive representation of the quality difference of MT systems; on the other hand, it is an easy way to assemble of the different evaluation results.

Acknowledgements

The research project is supported by the High-Tech Research and Development Program of Jiangsu Province China (Contract No. GB2005020, BK2006539), the Natural Science Foundation for Higher Education in Jiangsu Province (Contract No. 06KJB520095), Natural Science Foundation of Guangdong Province (Contract No. 108B6040600).

References

ALPAC, "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, 1966.

Brew, C., and H.S. Thompson, "Automatic evaluation of computer generated text: a progress report on the TextEval project," In *Proceedings of the Human Language Technology Workshop*, 1994, pp. 108-113.

Darwin, M., "Trial and Error: An Evaluation Project on Japanese English MT Output Quality," In *Proceedings of the MT Summit*,2001, Santiago de Compostela, Galicia, Spain, pp.57-63.

Doyon, J., K. Taylor, and J. White, "The DARPA Machine Translation Evaluation Methodology: Past and Present," In *Proceedings of the AMTA*, 1998, Philadelphia, PA.

Forner, M., and J. White, "Predicting MT fidelity from noun-compound handling," In *Proceedings of the Workshop MT Evaluation: Who Did What To Whom held in conjunction with Machine Translation Summit VIII*, 2001, Santiago de Compostela, Spain, pp.45-48.

Guessoum, A., and R. Zantout, "Semi-automatic evaluation of the grammatical coverage of machine translation systems," In *Proceedings of the MT Summit Conference*, 2001, Santiago de Compostela, pp.133-138.

ISLE, "The ISLE classification of machine translation evaluations, draft 1," A document by the International Standards for Language Engineering, http://www.isi.edu/natural-language/mteval/, 2000.

Jones, S., and J. Galliers, "Evaluating Natural Language Processing Systems," Technical Report 291, University of Cambridge Computer Laboratory,1993.

Keiji, Y., F. Sugaya, T. Takezawa, S. Yamamoto,and M. Yanagida, "An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus," In *Proceedings of MT Summit Conference*, 2001, Santiago de Compostela, pp.373-378.

Koh, S., J. Maeng, J. Y. LEE, Y. S. CHAE, and K. S. Choi, "A test suite for evaluation of English-to-Korean machine translation systems," In *Proceedings of MT Summit Conference*, 2001, Santiago de Compostela.

Melamed, I.D., R.Green, and J.P.Turian, "Precision and recall of machine translation," In *Proceedings of the NAACL/Human Language Technology*, 2003, Edmonton, Canada.

Niβen, S., F. J. Och, G. Leusch, and H. Ney, "An evaluation tool for machine translation: fast evaluation for MT research," In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000, Athens, Greece, pp.39-45.

NIST, "The NIST 2002 machine translation evaluation plan," A document by the National Institute of Standards and Technology, http://www.nist.gov/speech/ tests/mt/doc/2002-MT-EvalPlan- v1.3.pdf. 2002.

Papineni, K., S.Roukos, T.Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of MT," Research Report, Computer Science RC22176(W0109-022), IBM Research Division, T.J.Watson Research Center, 2001.

Popescu-Belis, A., "Evaluation of natural language processing systems: a model for coherence verification of quality measure," In 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), 1999.

Wang, X., "Education Measurement," East China Normal University Press, 2001, pp. 129~161. (in Chinese)

Yao, J., M. Zhou, H. Yu, T. Zhao, and S. Li, "An Automatic Evaluation Method for Localization Oriented Lexicalised EBMT System," In *Proceedings of the 19th Conference on Computational Linguistics (COLING'2002)*, August 24, 2002, TaiPei, Taiwan, pp.1142-1148.

Yasuhiro, A., K. Imamura, and E. Sumita, "Using multiple edit distances to automatically rank machine translation output," In *Proceedings of the MT Summit Conference*, 2001, Santiago de Compostela, pp. 15-20.

Yokoyama, S., H. Kashioka, A. Kumano, M. Matsudaira, Y. Shirokizawa, M. Kawagoe, S. Kodama, H. Kashioka, T. Ehara, S. Miyazawa, and Y. Nakajima, "Quantitative evaluation of machine translation using two-way MT," In *Proceeding of Machine Translation Summit VII*, 1999, pp.568--573.

Yu, S., "Automatic Evaluation of Quality for Machine Translation Systems," *Machine Translation*, 8, 1993, pp.117-126.

Appendix

This section presents the human evaluation results from [Darwin 2001] 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.

Table A1. Overall English-to-Japanese Average Scores (Possible Score from 1 to 5 Points).

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

Table A2. E-to-J Average Scores by Evaluator A and B (phase by phase), the column "I" lists intelligibility scores, and A column lists accuracy scores.

Test Suite	EJsys-1		EJsys-2		EJsys-3		EJsys-4		EJsys-5		EJsys-6		EJsys-7		EJsys-8	
	I	A	I	A	I	A	I	A	I	A	I	A	I	A	I	A
Sent#1-100(A)	2.38	2.62	3.25	3.56	3.30	3.54	3.14	3.48	3.10	3.29	2.97	3.26	3.08	3.33	2.81	3.04
Sent#101-200(A)	2.67	2.83	3.53	3.87	3.58	3.91	3.32	3.65	3.17	3.45	3.17	3.53	3.33	3.69	3.14	3.43
Sent#201-300(A)	2.11	2.41	3.02	3.54	3.05	3.61	3.01	3.40	2.67	3.06	2.71	3.02	2.65	3.07	2.56	2.86
All 300(A)	2.39	2.62	3.27	3.66	3.31	3.69	3.16	3.51	2.98	3.27	2.95	3.27	3.02	3.36	2.84	3.11
Sent#1-100(B)	1.91	1.76	3.15	3.08	3.08	2.98	3.08	2.87	2.73	2.55	2.78	2.65	2.83	2.75	2.48	2.39
Sent#101-200(B)	2.65	2.60	3.86	3.86	3.89	3.90	3.74	3.60	3.32	3.29	3.42	3.35	3.59	3.53	3.31	3.22
Sent#201-300(B)	2.25	2.29	3.50	3.66	3.61	3.77	3.60	3.68	3.03	3.15	3.02	3.09	3.20	3.25	2.89	2.97
All 300(B)	2.27	2.22	3.50	3.53	3.53	3.55	3.47	3.38	3.03	3.00	3.07	3.03	3.21	3.18	2.89	2.86