A Visualization method for machine translation evaluation results

Jian-Min YAO[1,2], Yun-Qian qu[1], Qiao-Ming Zhu[1],Jing Zhang[3]

[1]School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[3] School of Computer Science, South China University of Technology, Guangzhou 510641, China
E-mail: { jyao, quyq, qmzhu}@suda.edu.cn, zhjing@scut.edu.cn

Abstract. To make it easier to understand the machine translation evaluation results, a curve is utilized to stand for the performance of a machine translation system. The position of the curve in the graph depicts the quality of the system. The upper left curve stands for higher translation quality. System clustering is made and its dendrogram illustrates the quality difference between systems. These two methods visualize the machine translation evaluation results.
Keywords: machine translation, visualization, system clustering

1 Introduction

Machine translation evaluation activities have accompanied the MT research and system development. The ALPAC report is the first historical MT evaluation activity. In 1990s, because of the prosperity of machine translation research, great amount of evaluation activities are carried out according to the intelligibility and fidelity metrics followed DARPA methodology. The ISLE takes a software engineering point of view, which focuses on how an MT system serves the follow-on human processing rather than on what it is unlikely to do well.

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 can not be characteristic by its precision and recall as in other natural language processing activities such as POS tagging or phrase identification[4]. A new quality system is necessary. This paper aims for the better illustration of machine translation evaluation results and ensemble of different evaluation methods.

2 Related work

Evaluation has not been a very powerful aid in machine translation research until it is automated. Now different heuristics are employed for automatic MT evaluation. This section gives a brief review of main automatic MT evaluation methods and study on performance of these 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[1]. 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[2].

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[3]. Akiba (2001) uses multiple edit distances to automatically rank machine translation output by translation examples[4]. While the IBM BLEU method[5] and the NIST method[6] compare MT output with expert reference translations in terms of the statistics of word N-grams. The GTM method [7] adopts the maximum matching size of the translation and reference as similarity measure for score.

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[8]. Guessoum (2001) proposes a semi-automatic evaluation method of the grammatical coverage machine translation systems via a database of unfolded grammatical structures[9].

3 Visualization of MT system scores

The BLEU method scores MT quality in terms of a weighted sum of the counts of matching N-grams, including a penalty for translations whose length differs significantly from that of the gold standard translation, while the NIST method is a variation of BLEU. We make MT evaluation experiments using these methods and for a better understanding of the result, visualize the data in a graph as shown in Figure 2. The graph is produced with the algorithm in Figure 1. Figure 2 exhibits the MT evaluation 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.

```
INPUT: T←{Ti: t∈Ti, a translation by MT system MTSi}
// Process the MT translation and get the BLEU scores
FOR each machine translation system MTSi DO
FOR each translation t∈Ti by MT systems MTSi DO
Score{t}←{sti| the BLEU score of the translation ti}
END FOR
//Plot a line of the BLEU scores for each MT system
Score{t}←Score{t}{scores sorted in ascending order }
FOR i=1 to |T| {number of items in T} DO
Plot a point (i, sti) in the diagram
END FOR
END FOR
Output:a diagram where a system is presented as a curve
```

Fig. 1. Algorithm 1: Visualization of system scores by plotting lines in a graph

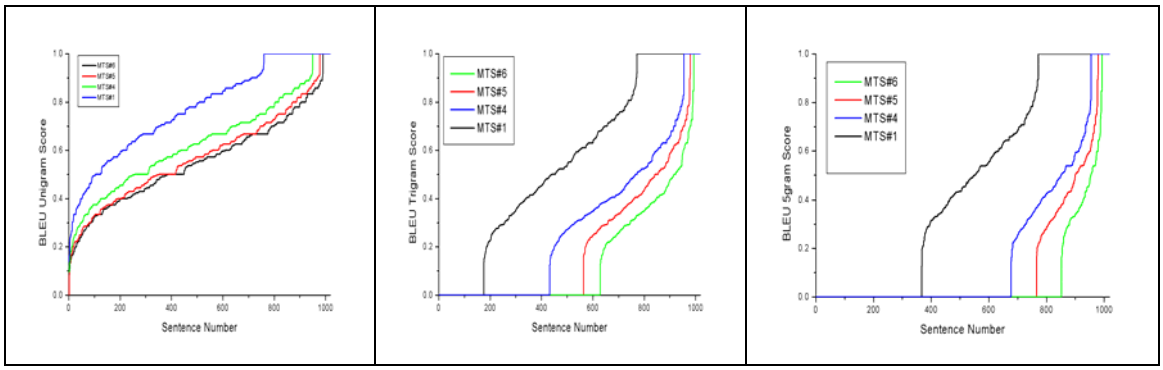

Fig. 2. MT evaluation scores of 4 MT systems with (a) 1-gram, (b) 3-gram and (c) 5-gram BLEU method

From the figure, 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 shift to the right side when the N-gram shifts from unigram to 5-gram. The upper 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 first figure of 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. This is because the difficulty of the test influences the discriminability of the measure.

The visualization method is different from BLEU/NIST in the following aspects: 1) the evaluation is not only presented for the whole system, but also each translation; 2) The tendency of the lines manifests the quality characteristics of MT systems, while the gap represents the difference.

4 System clustering for visualizing system quality difference

System clustering is utilized for visualizing the distances of MT systems in respect of translation quality. The process involves calculating the distances of system quality, as shown in figure 3.

INPUT: Score{MTSi}←{sco_mti| sco_mti is the BLEU score set of the translations by MT system 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

Fig. 3. Similarity histogram-based incremental MT system clustering

Table 1. Normalized scores of MT systems by various MT evaluation methods. F-score is F measure of human evaluation of fidelity and intelligibility; ET is human evaluation of error types and weighted score; SLP is statistical language model probability; Edist is edit distance-based similarity; DICE is DICE coefficient-based similarity

MTS	F-score	ET	SLP	BLEU	NIST	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

We evaluate the system by several manual and automatic evaluation methods. For they have different value scopes, we normalize the scores as in algorithm 2. After the normalization, the value of MT scores varies between 0~1. The normalized scores are shown in Table **1.** 1. The clustering dendrogram is shown in figure 4.

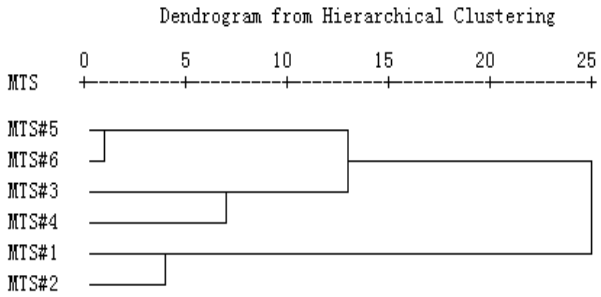

Fig. 4. Cluster chart and distance between clusters of 6 MT systems

5 Conclusions

This paper is an effort towards better rendering of machine translation evaluation results. A ROC-like curve is introduced to stand for a translation system, and a upper left curve represents a higher quality of the translation. The curves make it easy to tell the difference of translation of different systems.

After normalizing the scores from various evaluation methods, system clustering is made which manifests the quality gaps between translation systems. This clustering not only visualizes the quality difference but also integrates evaluation results from various methods.

Acknowledgements. The research project is supported by the Natural Science Foundation of Jiangsu Province (Contract No. BK2006539), Natural Science Foundation of Guangdong Province(Contract No. 108B6040600) and Jiangsu High-Tech Research Program (Contract No. GB2005020).

.

References

1. Douglas J, Rusk G. (2000). Toward a scoring function for quality-driven machine translation. Proceedings of the International Conference on Computational Linguistics. pp. 376-382.
2. Brew C, Thompson H. (1994). Automatic evaluation of computer generated text: a progress report on the TextEval project. Proceedings of the Human Language Technology Workshop. pp. 108-113.
3. Keiji Y, Sugaya F, 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.
4. Yasuhiro A, Imamura K, Sumita E. (2001). Using multiple edit distances to automatically rank machine translation output. MT Summit Conference, Santiago de Compostela. pp. 15-20.
5. 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/
6. 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
7. Melamed I.D., Green R, Turian J.P. (2003). Precision and recall of machine translation. NAACL/Human Language Technology 2003, Edmonton, Canada.
8. Yu SW (1993). Automatic Evaluation of Quality for Machine Translation Systems. Machine Translation, 8. pp. 117-126.
9. 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.