Machine Translation Evaluation

Tatiana Gornostay
NGSLT Machine Translation course
Reading assignment
October 23rd, 2008

Introduction

Nobody will argue that machine translation evaluation is an extremely important but challenging task. So, what is to be evaluated?

Machine translation (MT) is the translation of one natural language to another with the help of computer. Here lies the main difference between MT and human translation (HT). HT itself is not the point to be discussed herein. Although, some similarities should be mentioned against the background of evaluation problem. Firstly, either MT, or HT can be regarded both as a process and the result of this process. However, MT process is much more easier to examine with reference to HT process since it was developed by human. Secondly, any entity, as a rule, can be viewed critically. Hence, both MT and HT can be evaluated. MT evaluation methodology differs from HT evaluation for some reasons. On the one hand, the output of MT and HT is not the same (rough quality of MT versus high quality of HT). On the other hand, in contrast to HT, MT can be evaluated both as the process and the result along with the MT system itself.

MT evaluation concept

What is MT evaluation *per se*? One might say that MT evaluation is merely the statement whether this or that MT system is good enough or rather bad with or without any scale of defining to be rather "good" than "bad". The other will argue and insist on developing a hierarchical criteria classification. In other words, from the surface to the core of the question.

MT evaluation exists alongside with MT itself as a separate branch of computational linguistics in the field of MT systems development and has its own history and achievements, though is closely related to MT theory. A lot of researchers and developers discussed this issue already and have been discussing so far and MT evaluation has been an area of significant research in itself over the years (Banerjee, Lavie 2005). Moreover, as it was stated by Y. Wilks, "machine translation evaluation is better understood than machine translation" (Carbonell, Wilks 1991) and "machine translation evaluation is a better founded subject than machine translation" (Wilks 1994).

Evaluation and meta-evaluation

There are miscellaneous approaches to MT evaluation with a range of proposed methodologies and metrics, both automated and human, which can also be evaluated. Let us play upon words, the evaluation of evaluation, the process that is traditionally called meta-evaluation, represents an independent research area in the field of MT evaluation. There has been carried out several works on MT evaluation methodology review. Let us mention some of them.

In 1978 an international seminar on the problems of evaluation of MT was arranged and it was settled to carry out a critical review on the methods of evaluating MT. That review to be based on the presentations made at the seminar and on the studies on evaluation of MT already published appeared in 1979 in the final report "Critical Study of Methods for Evaluating the Quality of Machine Translation" (Slype 1979) summarizing all the technologies of MT evaluation as updated in that moment. That critical study met two requirements:

- to establish the state of the methodology of evaluation of machine translation;
- to make to the Commission a series of recommendations concerning both the methodology to be used to evaluate MT systems and research intended to improve in the long term the efficiency of those evaluations.

Later on the Machine Translation Market and Technology Study Committee of JEIDA (Japan Electronic Industry Development Association) had been working to develop criteria for MT evaluation for several years and published the first version in 1992. The paper presented the JEIDA discussion on the methodology and then lists of the criteria that permit both the overall evaluation of the system and the evaluation of technical components which have been incorporated into the system. Two sets of criteria for MT evaluation were developed: one by which users evaluate the possibility of introducing a MT system, and the other by which researchers and developers undertake technical evaluations of their systems for future research and development. However, this set of criteria does not evaluate the quality of MT system explicitly (Nomura, Isahara 1992).

Since the middle of the 90ies the Expert Advisory Group on Language Engineering Standards has been contributing to the issue. Two reviews were published in 1996 and 1999 (EAGLES 1996, 1999). EAGLES has been also developing the ISLE standards (International Standards for Language Engineering) on the basis of ISO/IES-9126-1 software engineering product quality standards of ISO (International Standard Organization) (EAGLES 1996-2008). As a result, a Framework for Machine Translation Evaluation (FMTE) in ISLE was developed[1] - the EAGLES guideline for NLP software evaluation (Hovy et al. 2002). It was an attempt to organize the various methods that were used to evaluate MT systems, and to relate them to the purpose and context of the systems.

FEMTI contains:

- a classification of the main features defining the context of use, that is: the type of user of the MT system, the type of task the system is used for, and the nature of the input to the system;
- a classification of the MT software quality characteristics;
- a mapping from the first classification to the second.

FEMTI helps:

- people who wants to use an MT system: they can select the quality characteristics that are most important to them and thereby choose the MT system that best suits these characteristics;
- people who want to compare several MT systems: they can brose and select the characteristics that best reflect their circumstances, and tereby find associated evaluation measures and tests;

[1] http://www.isi.edu/natural-language/mteval/

- people who want to design a new MT system or to upgrade an old one: they can learn about the needs of users and find niche applications for their system.

FEMTI is still under work: some parts need completion, while others are updated based on feedback from the community.

Some experiments were carried out with reference to the most commonly used methodologies of automated metrics of MT evaluation (Callison-Burch et al. 2007). The correlation of automated MT evaluation metrics with human judgments was measured and that meta-evaluation revealed surprising facts which will be described short while later in connection with automated metrics.

The experiment with meta-evaluation of human MT evaluation metrics was also carried out along with automated metrics. The researchers wanted to examine three types of manual evaluation and assess which was the best (Callison-Burch et al. 2007).

MT evaluation strategy

MT is evaluated for a number of different reasons and different types of evaluation are best suited to measure different aspects of an MT system. Comparison with human-quality translation, decision to use or buy a particular MT system, comparison of multiple MT systems, tracking technological process, improvement of a particular system are the reasons why MT systems may be evaluated (Nyberg et al. 1994).

Hence, MT evaluation strategy is task dependant. We can evaluate MT as a process and as a product (Vasconcellos 1992) or we can evaluate an overall MT system (Albisser 1991). MT evaluation techniques can be also classified according to evaluator (user and researcher / developer) (Nagao 1992; Nomura, Isahara 1992; Somers, Wild 2000). Evaluating MT is crucial for everyone involved: researches need to know if their theories make a difference, commercial developers want to impress customers, and users have to decide which system to employ (Nyberg et al. 1994; Hovy et al. 2002). Nevertheless, all researchers and developers agree that the main criterion of MT evaluation is the degree of the user's satisfaction with a MT system (Nomura, Isahara 1992; Popescu-Belis 2001, Yuste-Rodrigo 2001, Allen 2003).

Evaluating MT system as a process has two main advantages. Firstly, it enables us to set up typologies for comparing different MT systems cutting across glass box and black box. We may look inside the system, or we may look at it from the outside without getting into its inner workings, but the main point is that we are looking at a functional process, and we are going to see that not all systems have the same purpose. Secondly, with a functional approach it is also easier to make predictions about system's potential since unless we know the potential of an MT system, there is not much point in spending time to evaluate the formal output (Vasconcellos 1992).

The evaluation of generated output is an important issue for MT as well. Evaluating MT system as a product (MT output) depends on three main tasks performed by MT: dissemination, assimilation and communication (Gaspari, Hutchins 2007). It has become a fact so far that high-quality translation is needed for assimilation only. In all other cases rough raw MT is sufficient.

Automated vs. human metrics

Output quality analysis methodology can be performed both automatically and manually. Each approach has its advantages and disadvantages.

Advantages of automated metrics:

- automated metrics are fast, cheap and re-usable;
- they allow to compare multiple systems;
- automated metrics allow to perform system monitoring in day-to-day development of MT system, since they can be applied on a frequent and ongoing basis, and guide the development of the system based on concrete performance improvements;
- the results of automated metrics can be of use pointing out to work content for post-editing (Elliott et al. 2004; Banerjee, Lavie 2005; Callison-Burch et al. 2006, 2007; Lavie, Agarwal 2007).

Disadvantages of automated metrics:

- automated metrics require text preparing for testing both source text to be analyzed and one or more reference texts to compare with;
- all the automated metrics are not necessary reliable and rather subjective since they are based on human evaluation (source text is compared with already adjusted and evaluated texts);
- automated metrics do not produce very reliable sentence-level scores;
- these metrics do not give any details about the nature of translation errors (Turian 2003; Elliott et al. 2004; Popovic, Ney 2007; Popovic et al. 2006).

Advantages of human metrics:

- it is ultimately what we are interested in;
- human metrics allow to perform high-level analysis of the evaluation process (Callison-Burch et al. 2007).

Disadvantages of human metrics:

- time and labor consuming and expensive;
- not re-usable;
- trained bilingual evaluators are required;
- not suitable for comparisons of multiple systems;
- using "fluency and adequacy" metric people have a hard time separating two these aspects of translation, and the high correlation between people's fluency and adequacy scores indicate that the distinction might be false; and there are no clear guidelines on how to assign values to translations, no instructions are given to evaluators in terms of how to quantify meaning, or how many grammatical errors (or what sort) separates the different levels of fluency, thus, many judges either develop their own rules of thumb, or use the scales as relative rather than absolute (Banerjee, Lavie 2005, Callison-Burch et al. 2006, 2007; Popovic, Ney 2007).

Human metrics

There is a range of possibilities for how human evaluation of MT can be performed. For instance, it can be evaluated with reading comprehension test, or by

assigning subjective scores to the translations of individual sentences (Callison-Burch et al. 2007).

The most widely used methodology when manually evaluating MT is to assign values from five point scales representing fluency and adequacy. These scales were developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistics Data Consortium (LCD 2005). The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation (output translation). The second five point scale indicates how fluent the translation is. Separate scales for fluency and adequacy were developed under the assumption that a translation might be diffluent but contain all the information from the source.

In case with separate evaluation (ranking translated sentences relative to each other) people are simply asked to rank translations (from best to worst relative to other choices). Rather than having to assign each translation a value, people simply have to compare different translations of a single sentence and rank them.

In addition, the pilot study of a new type of MT evaluation methodology was conducted – constituent-based evaluation (Callison-Burch et al. 2007). The logics of this metric is to parse the source language sentence, to select constituents from the tree, and have people judge the translations of these syntactic phrases. The corresponding phrases in the translations are located via automatic word alignments.

Automated metrics

Automated metrics of MT evaluation have been receiving significant attention in recent years (Lavie, Agarwal 2007).

The way that automated evaluation metrics work is to compare the output of a MT system against reference HT (Callison-Burch et al. 2006) and they correlate with human judgments. Human judgments come in the form of "adequacy" and "fluency" quantitative scores (Lavie, Agarwal 2007).

BLEU (Papineni et al. 2002) is an IBM-developed metric and is probably the best known and most used in the MT community, and currently the de facto standard in MT evaluation (Callison-Burch et al. 2007). It is based on calculating n-gram precision between the system output and reference translation and a brevity penalty if the output differs from the reference.

TER (Translation Error Rate) measures the amount of editing required to change an output translation into a reference translation. The possible edits in TER include insertion, deletion, and substitution of single words (Snover et al. 2006).

METEOR (Metric for Evaluation if Translation with Explicit ORdering) was initially proposed and released in 2004, and was designed to explicitly address several observed weaknesses in IBM's BLEU metric (Lavie et al. 2004). This automated metric is based on an explicit word-to-word matching between the MT output being evaluated and one or more reference translations (Banerjee, Lavie 2005). METEOR's matching supports not only matching between words that are identical in the two strings being compared, but can also match words that are simple morphological variants of each other (i.e. they have an identical stem), and words that are synonyms of each other.

WER (Word Error Rate) is based on the minimum number of substitutions, deletions, and insertions that have to be performed to convert the generated text into the reference text (Popovic, Ney 2007).

NIST, a National Institute of Standards and Technology developed metric (Doddington 2002), is closely related to BLEU and aims in upgrading BLEU metric.

The F-measure was proposed as a comprehensible alternative for MT evaluation, and it can be defined as a simple composite of unigram precision and recall (Melamed et al. 2003).

BLEU and NIST metrics are the most popular to be applied to MT evaluation process among researchers and developers. Moreover, BLEU and NIST along with other automated metrics were accepted by MT translation community as panacea. MT system researchers and developers has been exploiting these metrics in an effort to claim improvements in translation quality by reporting improved BLEU / NIST scores, while neglecting to show any actual example translations. However, such automated metrics cannot answer the question: "Does minimizing the error rate indeed guarantee genuine translation improvements?" (Callison-Burch et al. 2006). So, Chris Callison-Burch, Miles Osborne and Philip Koehn showed that an improvement in BLEU is not sufficient to reflect a genuine improvement in translation quality and is not necessary to improve BLEU in order to achieve a noticeable improvement in translation quality. Moreover, the results of 2006 year's ACL workshop further suggested that BLEU systematically underestimated the quality of rule-based machine translation systems (Koehn, Monz 2006).

To eliminate the expense of producing human translations some experiments on designing an automated MT evaluation system, which does not require human reference translations, were carried out (Elliott et al. 2004).

Output quality analysis (linguistic analysis)

While both automated and human MT evaluation metrics are extremely important, they do not give a finer grained analysis of mistakes and should be augmented with detailed error analyses in order to identify the main problems, focus the research efforts, and therefore improve the system (Nyberg at al. 1994; Lange, Gerber 1992; Vilar et al. 2006; Kirchhoff et al. 2007; Popovic, Ney 2007). Some experiments have been carried out to develop output quality error analysis, or linguistic analysis, classifications.

Conclusion

MT evaluation task is a very important but complicated issue. There are miscellaneous ways to evaluate MT as a process and as a product (output) along with the MT system itself. MT evaluation is significant both for users and researches / developers. MT evaluation can be performed automatically and manually. Each approach has its own advantages and disadvantages. In every certain case it is necessary to choose an appropriate MT evaluation strategy in order to achieve optimal results. In addition, a considerable work on resource (corpus, test suite) preparation must be done. The latter is the subject of our practical paper.

References

1. *Albisser* D. (1991) Evaluation of MT Systems at Union Bank of Switzerland // Evaluator's Forum, Les Rasses, Vaud, Switzerland.

2. *Allen* J. (2003) Post-Editing // Computers and Translation: a transl. guide, p. 297–317.
3. *Banerjee* S., Lavie A. (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments // The 43rd 40th Annual Meeting of the ACL, Ann Arbor, Michigan, pp. 65-72.
4. *Callison*-Burch C., Cameron F., Koehn P., Monz C., Schroeder J. (2007) (Meta-) Evaluation of Machine Translation // The Second Workshop on Statistical Machine Translation, Prague, p. 136-158.
5. *Callison-Burch* C., Osborne M., Koehn P. (2006) Re-evaluating the Role of BLEU in Machine Translation Research // The 11th Conference of the EACL, pp. 249—256.
6. *Carbonell* J., Wilks Y. (1991) Machine Translation: An In-Depth Tutorial // The 29th Annual Meeting of the ACL, Berkeley, California.
7. *Doddington* G. (2002) Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics // Second Human Language Technologies Conference (HLT-02), San Diego, CA. pp. 128-132.
8. *EAGLES* Expert Advisory Group on Language Engineering Standards (1996–2008) Electronic resource: http://www.ilc.cnr.it/EAGLES96/home.html (21.09.08)
9. *EAGLES* MT Evaluation Working Group. (1996) EAGLES Evaluation of Natural Language Processing Systems: final report, Copenhagen.
10. *EAGLES* MT Evaluation Working Group. (1999) EAGLES Evaluation of Natural Language Processing Systems: final report, Copenhagen.
11. *Elliott* D., Hartley A., Atwell E. (2004) A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation // AMTA, pp. 64–73.
12. *Gaspari* F., Hutchins J. (2007) Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects // The 9th MT Summit, Copenhagen.
13. *Hovy* E., King M., Popescu-Belis A. (2002) Principles of Contex-Based achine Translation Evaluation // Machine Translation, 17, pp. 43-75.
14. *Kirchhoff* K., Rambow O., Habash H., Diab M. (2007) Semi-Automatic Error Analysis for Large-Scale Statistical Machine Translation Systems // MTS.
15. *Koehn* P., Monz C. (2006) Manual and automatic evaluation of machine translation between European languages // NAACL Workshop on Statistical Machine Translation.
16. *Lange* E., Gerber L. (1992) Internal Evaluation: Quality Analysis, an Internal Evaluation Tool at SYSTRAN
17. *Lavie* A, Agarwal A. (2007) Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments // The Second Workshop on Statistical Machine Translation, Prague, p. 228-231.
18. *Lavie* A., Sagae K., Jayaraman S. (2004) The Significance of Recall in Automatic Metrics for MT Evaluation // The 6th Conference of the AMTA, Washington DC.
19. *LDC* (2005) Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.
20. *Melamed* I., Green R., Turian P. (2003) Precision and recall of machine translation // HLT, NAACL.
21. *Nagao* M. (1992) in Panel: apples, oranges, or kiwis? Criteria for the comparison of MT systems // MT evaluation: basis for future directions, San Diego, p. 41.
22. *Nomura* H., Isahara H. (1992) Evaluation Surveys: The JEIDA Methodology and Survey // MT evaluation: basis for future directions, San Diego, pp. 11-12.

23. *Nyberg* E., Mitamura T., Carbonell J. (1994) Evaluation metrics for knowledge-based machine translation // 15th conference on Computational linguistics, pp. 95-99.
24. *Papineni* K., Roukos S., Ward T., Zhu W. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation // The 40th Annual Meeting of the ACL, Philadelphia, pp. 311-318.
25. *Popescu-Belis* A. (2001) Towards a Two-Stage Taxonomy for Machine Translation Evaluation // The 8th MT Summit Workshop on MT Evaluation, Santiago de Compostela, Spain. Electronic resource: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C2237D4782171FB08268D9A0B79FC586?doi=10.1.1.8.6802&rep=rep1&type=pdf (21.09.08)
26. *Popovic* M., Ney H. (2007) Word error rates: Decomposition over POS classes and applications for error analysis // ACL Workshop on Statistical Machine Translation.
27. *Popovic* M., Ney H., Gispert A., Marino J., Gupta D., Federico M., Lambert P., Banchs R. (2006) Morpho-syntactic Information for Automatic Error of Statistical Machine Translation Output // NAACL Workshop on Statistical Machine Translation.
28. *Slype* G. van. (1979) Critical Study of Methods for Evaluating the Quality of Machine Translation: final report, Bruxelles.
29. *Snover* M., Dorr B., Schwartz R., Micciulla L., Makhoul J. (2006) A study of translation edit rate with targeted human annotation // ASMTA.
30. *Somers* H., Wild E. (2000) Evaluating Machine Translation: the Cloze Procedure Revisited // Translating and the Computer, London.
31. *Turian* J. P. (2003) Evaluation of Machine Translation and its Evaluation // The 9th MT Summit, New Orleans, p. 386–393.
32. *Vasconcellos* M. (1992) in Panel: apples, oranges, or kiwis? Criteria for the comparison of MT systems // MT evaluation: basis for future directions, San Diego, p. 41.
33. *Vilar* D., Xu J., D'Haro L., Ney H. (2006) Error Analysis of Statistical Machine Translation Output Evaluation Metrics for Knowledge-Based // LREC.
34. *Wilks* Y. (1994) Keynote Traditions in the Evaluation of MT // MT Evaluation Basis for Future Directions, San Diego, California, pp. 1-3.
35. *Yuste-Rodrigo* E. (2001) Comparative Evaluation of the Linguistic Output of MT Systems for Translation and Information Purposes // The 8th MT Summit Workshop on MT Evaluation, Santiago de Compostela, Spain. Electronic resource: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=099EA3C100E1EEDB7E2CDFD7256D6D21?doi=10.1.1.9.4952&rep=rep1&type=pdf (21.09.08)