revista.tradumàtica
tecnologies de la traducció

464-470

Automated MT evaluation metrics and their limitations

Bogdan Babych
School of Computing
Faculty of Engineering
University of Leeds

ABSTRACT

This paper gives a general overview of the main classes of methods for automatic evaluation of Machine Translation (MT) quality, their limitations and their value for professional translators and MT developers. Automated evaluation of MT characterizes performance of MT systems on specific text or a corpus. Automated scores are expected to correlate with certain parameters of MT quality scored by human evaluators, such as adequacy of fluency of translation. Automated evaluation is now part of MT development cycle, but it also contributes to fundamental research on MT and improving MT technology..

Keywords: Machine translation evaluation, automated methods, future perspectives

RESUM *(Mètriques d'avaluació automatitzada de TA i les seves limitacions)*

Aquest article ofereix una visió general de les principals classes de mètodes d'avaluació automàtica de la qualitat de la Traducció Automàtica (TA), les seves limitacions i el seu valor tant per a traductors professionals com per a desenvolupadors de TA. L'avaluació automàtica de TA es caracteritza per l'actuació dels sistemes de TA amb textos o corpus específics. És d'esperar que els índexs automàtics es correlacionen amb aquells paràmetres que estableixen els avaluadors humans sobre la qualitat de la TA, com ara l'adequació o fluïdesa de la traducció. L'avaluació automàtica actualment és part del cicle de desenvolupament de la TA, i a més també permet fer avançar la investigació fonamental sobre TA i millorar la seva tecnologia.

Paraules clau: traducció automàtica, avaluació, mètodes automatitzats, perspectives futures.

RESUMEN *(Métricas de evaluación automatizada de TA y sus limitaciones)*

Este artículo ofrece una perspectiva general de las principales clases de métodos de evaluación automática de la calidad de la Traducción Automática (TA), sus limitaciones y su valor tanto para traductores profesionales como para desarrolladores de TA. La evaluación automática de TA se caracteriza por la actuación de los sistemas de TA con textos o córpora específicos. Es de esperar que los índices automáticos se correlacionen con aquellos parámetros que establecen los evaluadores humanos sobre la calidad de la TA, como por ejemplo la adecuación o fluidez de la traducción. La evaluación automática actualmente es parte del ciclo de desarrollo de la TA, y además también permite hacer avanzar la investigación fundamental sobre TA y mejorar su tecnología.

Palabras clave: traducción automática, evaluación, métodos automatizados, perspectivas futuras.

1. Introduction

This paper discussed classes of methods for automatic evaluation of the quality of Machine Translation, their limitations and value for professional translators and MT

developers. The main objective of methods and tools for automated evaluation of MT is to compute numerical scores, which characterize the 'quality', or the level of performance of specific Machine Translation systems. Automated MT evaluation scores are expected to agree (or correlate) with human intuitive judgments about certain aspects of translation quality, or with certain characteristics of usage scenarios for translated texts.

The development of automated evaluation techniques in 1990-ies shaped the research and development efforts of modern Machine Translation systems, so MT developers are now able to quickly monitor the progress and compare different system and assess effects of any changes, such as introduction of new data sources and processing algorithms. This shifted the focus to the large-scale wide-coverage systems, and to methods and tools that work not only for a few individual handpicked examples, but also generate improvements for the most frequent linguistic phenomena in a corpus, or a specific subject domain.

Automated MT evaluation works now in a wider context of translation industry, collaborative translation workflow that integrates Translation Memories, MT, terminology management systems and electronic dictionaries. As with any of such tools, it is important to understand optimal usage scenarios and limitations for various automated MT evaluation methods – specifically which user needs can be addressed by specific techniques, and which questions cannot be addressed.

Automated evaluation tools are typically calibrated using more expensive, slow and intuitive process of human evaluation of MT output, where a group of human judges is asked to read either both the original text and the MT output, or just the MT-translated text, and to give their evaluation scores for certain quality parameters. These parameters usually include adequacy, measured by how much information from the original is preserved in MT output, fluency, measured by how naturally the MT output sounds in the target language, informativeness (or comprehension), measured by a multiple-choice questionnaire on the content of the evaluated text (White et al., 1994), or usability, measured by how useful would be the MT-translated text for getting a certain job done (completing a business transaction, following an instruction from a user manual, etc.). If a strong and consistent agreement is found between any of the human quality parameters and scores generated by a certain automated MT evaluation tool the automated tool can be used on its own to measure that specific human quality parameter.

Automated methods function in the context of a broader MT evaluation paradigm, as they usually address only one specific aspect – text quality evaluation. The text-external aspects that are relevant for the industrial settings are now not covered by automated MT evaluation tools, such as the quality of MT systems' user interface, user-friendliness, support of collaborative translation projects, customization for specific subject domains, extendibility and integration with other computer-assisted translation tools, suitability for large volume translation, client privacy considerations, systems' footprint and suitability of specific MT systems and architectures for different platforms – such as mobile devices, desktop applications for freelance translators, company networks or data centers, global web services and cloud computing services, systems' effectiveness, the dynamics of time and cost savings for translation teams of different sizes, and for specific subject domains, language pairs and translation directions. Other aspects, which are normally not covered by automated methods, are evaluation of individual systems' components or data sources, of feasibility evaluation for development of new systems or MT architectures.

Even though historically the area of MT evaluation started as a sub-field within MT development and primarily aimed at measuring improvements of systems and their features during the development cycle, it has now become a separate field with a wider set of goals and stakeholders, who include not only MT development community, but also translators and translator teams, localization project managers, investors, end users of the translated texts, who have to make decisions whether MT systems can address their specific needs (e.g., from the point of view of usability, text quality, post-editing effort, saving of time and cost in collaborative translation workflow, or for using unedited MT output for comprehension, etc.).

The FEMTI project (a Framework for Evaluation of Machine Translation in ISLE – the International Standards for Language Engineering) has given a comprehensive overview of such purposes and scenarios for MT evaluation (King et al., 2003), which point out this field has nowadays become much wider than even the area of MT development itself.

Apart from its practical use, MT evaluation also contributes to fundamental research on MT: it allows developers to go beyond an engineering perspective, and not only to implement known features and models, but also – to discover new facts about language and translation that can be further be used for improving MT technology.

2. Automated MT evaluation methods

There are two main types of methods for automatically evaluating translation text quality: reference proximity and performance-based methods.

Reference proximity techniques replicate a scenario of comparing the target text to the original or to the gold-standard human reference, so better MT output is considered to be closer to the reference. In these methods the distance between an MT output and a human professional translation is computed automatically, for example as a “word error rate” (WER), which is a Levenshtein edit distance (the minimal number of insertions, deletions and substitutions needed to transform corresponding sentences into each other) However, a standard edit distance (which was developed for the area of Automated Speech Recognition) is considered to be too simplistic for Machine Translation. The reason is that legitimate translation variants often involve differences in the order of words and phrases without major changes in meaning, and WER penalizes such re-orderings at the same level as using wrong words in one place and inserting redundant, or spurious words at another place.

A modification of the edit distance measure, which takes into account possible positional variation of continuous word sequences, was proposed in the Translation Error Rate (TER) metric (Snover et al., 2006), which is calculated as the number of “edits” divided by the average number of words in a reference. “Edits” cover insertions and deletions, as well as “shifts” – movement of continuous word sequences. TER can take into account several reference translations: the distance is calculated to one of such translations.

An alternative way of measuring the distance between MT output and a reference translation is to calculate their overlap in terms of N-grams (individual words and continuous word sequences of different length, the length is usually between 1 and 4. This method became the basis of the most widely used family of metrics, such as BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) and its modifications: NIST (National Institute of STandard's improved version of BLEU) (NIST 2005), METEOR (a metric which integrates additional linguistic features, such synonyms and stems, or dictionary forms of the inflected words found in the evaluated texts) (Banerjee and Lavie, 2005), WNM (Weighted N-gram model, which takes into account statistical salience scores, and assignes more weight to topic-specific terms and named entities) (Babych and Hartley, 2004), etc.

N-gram metrics are usually based on calculating three main parameters of the lexical overlap between the MT output and the reference: (A) the number of common word sequences; (B) the length of MT-translated text and (C) the length of the reference. The fraction (A/B) is called Precision: it is the score, which penalizes over-generation of spurious words that are not in the reference translation. The fraction (A/C) is called Recall: this is the score for avoiding “under-generation”: the words which are in the reference, but not in MT output. BLEU and NIST are calculated as a Precision score, while METEOR, WNM and other modifications of BLEU use some type of combination between the Precision and Recall scores. Importantly, these scores can be computed for several reference translations at the same time. Precision is calculated on a union of N-grams in all available reference translations, where the intuition is that MT is penalized for generating lexical items that are not used in any of human translations of the source text, which should be truly outside what could

be expected in a given sentence. Recall is calculated on an intersection of N-grams in human references, so MT is penalized for not generating really important words that have been produced by all human translators. The use of multiple reference translations of the same text addresses the issue of legitimate translation variation: with a single human reference unmatched N-grams can be interpreted either as translation errors, or as legitimate alternative translations. Several human reference translations (if produced independently) cover several possible translation alternatives for words and phrases where the variation can be expected. This lowers the number of genuine alternatives in MT output that do not match in N-gram based metrics. For practical evaluation purposes, however, it is often difficult to generate multiple independent translations of the same text, and a single longer human reference is used, which normally doesn't affect the accuracy of N-gram metrics (their correlation with human scores), e.g., if the size of the evaluated text is more than 7,000 words (Estrella et al., 2007).

Interestingly, reference proximity scores can be computed without using an actual reference translation. They can be calculated directly for the original text, e.g., he D-score proposed in (Rajman and Hartley, 2001). This score is computed on the basis of lexical similarity of the original and MT output to each of the documents in a parallel corpus of professional human translations – respectively: on the source and the target side, as a distance between the similarity matrices generated for MT and for the original text, which estimates the semantic distance between the text in two different languages.

The second group is performance-based methods of automated MT evaluation, which also have parallels in human evaluation: the idea is to measure how well someone can carry out a task on the basis of a degraded MT output. Different quantitative measures of performance for the task are taken to characterize the quality, or usability of MT output. In the context of human evaluation an early example of this idea is (Sinaiko, 1979), where an MT system was used to translate flying instruction manuals, and its performance was measured as the number of pilots who were successfully flying flight simulators.

In automated MT evaluation the performance-based methods are computed as evaluation scores for an automatic system on some well-defined tasks, such as text annotation or information extraction. These systems are normally designed for applications outside the filed of MT and are benchmarked on original texts, which are authored by human writers, usually – native speakers. The intuition is that these systems are expected to perform worse on imperfect MT output, and the amount of degradation should be proportional to human intuitions about MT quality. Examples of performance-based metrics that measure the performance of syntactic parsers on degraded MT output include C-score, which is computed as an average syntactic bracketing coverage of a sentence, and X-score, calculated as a combination of certain types of long- and short-distance dependencies identified in the translated text (Rajman and Hartley, 2001). Another metric that uses raw counts and the Recall score calculated for the Named Entity recognition task from MT output was presented in (Babych and Hartley, 2004b). Other proposals include measuring the performance Information Retrieval of MT translated texts, or the success of terminology extraction or of template filling in Information Extraction. Performance-based metrics can use human reference only for benchmarking the "upper limit" of the expected score, but reference texts are not strictly necessary for producing performance figures for MT systems. Important underlying assumptions behind performance-based metrics are that:

1. MT errors frequently destroy conditions in texts which trigger rules or statistical algorithms designed or trained on human texts; they rarely create spurious conditions: this assumption parallels the second law of thermodynamics applied to Machine Translation, since because of the redundancy of the natural language it is much easier to destroy certain highly-specialized conditions in a text than to construct such conditions by chance or by error.
2. The amount of degradation in performance is relative, so automated systems do not have to be 100% accurate on human texts, as again, this accuracy gives only the

upper limit for the scores. In this case automated systems' output can be characterized as a "silver standard" annotation (in contrast with the "gold-standard" human annotation often used in evaluation).

3. An automated task that measures the performance of MT can turn out to be also the primary application for which MT has been generated, e.g., Information Extraction of terrorist activity event form MT output. In this case the performance-based metric removes the need for calibration of the scores with human intuitive judgments, about translation quality, because, e.g., the success of template filling here is measured directly, so procedure provides a technical definition of MT quality that does not need to rely on less tangible intuitive definitions of this concept.

However, performance-based methods have more limited application compared to reference proximity methods, mainly because they are more tightly linked to specific text types (instruction manuals, or news texts rich in Named Entities) or assume certain types of linguistic structures (e.g., an abundance of complex sentences with subordinate clauses needed to compute the X-score).

3. Limitations of automated MT evaluation techniques

Automated evaluation metrics were primarily designed to monitor the development progress of the same MT system. However, nowadays the range of their applications grew to other areas. Still, the use of these metrics for these new tasks is often based on a number of untested assumptions, such as the belief that some universal score (that characterizes translation accuracy independently of the purpose or a scenario for which MT is used) can capture translation quality.

One of the cases of improper use of automated MT evaluation was highlighted in (Callison-Burch et al., 2006): if N-gram-based metrics such as BLEU are used to compare statistical and rule-based MT systems, they consistently over-estimate statistical MT, so in the eyes of human judges the quality produced by rule-based systems is always higher than is reported by BLEU (in comparison with SMT). An extreme case of this type of inadequacy of BLEU for such comparisons is reported in (Babych et al., 2004): when a human translation was produced by a non-native speaker and was included into an evaluation set together with compared MT systems, human judges rated this translation well above the output of any of the compared MT systems. However, BLEU ranked this translation lower than the best performing MT system, because many lexical items were not natural and did not match the word sequences in the professional human reference produced by a native speaker; even though a non-native human translator successfully preserved most of the content (in contrast with MT output).

This points out to a fundamental limitation of automated scores: they cannot be reliably used to compare systems built within different architectures. Importantly, their values are meaningful only in comparison with the scores for similar systems or for the previous versions of the same system, so the numbers like BLEU=0.4 cannot be directly interpreted in terms of quality, or reliably mapped into human evaluation scores expected for a given MT system: this mapping will depend at least on the architecture used to build an MT system.

Secondly, it has been pointed out that mappings from automated to human scores also depends on subject domains and the target language (Babych et al., 2005): even though for all languages and domains BLEU the scores correlates well with human evaluation results, the regression parameters, such as the slope an the intercept of the fitted line (the parameters needed to calculate expected human scores on the basis on the correlated automated scores) were significantly different for specific combinations of domains and the target languages. So if human scores are associated with levels of MT usability for a certain project or task (e.g., it may be determined that for a post-editing scenario an MT system should score at least 4 out of 5 on human Adequacy), then BLEU needs to be re-calibrated

using a human evaluation experiment each for each new combination of the subject domain and the target language, as in each case the usability threshold expressed as a BLEU score will be different.

Thirdly, certain types of automated scores loose sensitivity on higher quality MT output (Babych and Hartley, 2008). For example, while BLEU-type scores have high correlation with human evaluation results for imperfect MT systems (those which require extensive poste-editing to achieve publishable quality), the correlation gradually drops when MT quality improves; so e.g., for high-quality MT between closely related languages (that needs little post-editing to produce a publishable text) BLEU may not be sensitive enough. It is interesting that performance-based automated metrics are more stable across the wider range of quality spectrum – there is no observable drop in sensitivity for higher quality MT output. The explanation can be that reference proximity methods rely on certain computable linguistic features at a certain language level, such as lexical features used by BLEU; but for higher quality MT human lexical issues are largely resolved and evaluators pay more attention to discourse-level features which are not prominent on the lexical level, e.g., textual connectors, co-reference chains, etc. Performance-based metrics, on the other hand, rely on external functional aspects of the text, which are not directly linked to features on a specific language level, so they are more stable across the quality spectrum.

These limitations indicate that the development of automated evaluation metrics needs to be complemented with studies of applicability and limitations of these metrics in different usage scenarios in the workflow of MT developers, professional translators, translation teams and companies.

From MT development perspective there is an important link between the automated metrics that accurately predict human evaluation scores, and new methods for improving MT: generally those methods and features which work for MT evaluation can be used for improving MT quality. For example, the success of Named Entity recognition from MT output characterizes systems' quality; but it can be also implemented on the pre-processing stage to improve sentence segmentation and prevent parts of the person and organization names from being translated as common names (e.g., Bill Fisher should not be translated as 'send a bill to a fisherman'). But once this pre-processing becomes integrated into MT systems, it can no longer be used for evaluation. Therefore, as the MT development catches up with automated MT evaluation methods, MT users will always need new "surprise" metrics that are not yet part of state-of-the-art MT architectures.

References

King, M., Popescu-Belis, A., & Hovy, E. (2003). FEMTI: creating and using a framework for MT evaluation. In Proceedings of MT Summit IX, New Orleans, LA (pp. 224-231).

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of Association for Machine Translation in the Americas.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318.

NIST (2005). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.itl.nist.gov/iad/mig//tests/mt/doc/ngram-study.pdf

Banerjee, S. and Lavie, A. (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005.

Babych, B., and Hartley, A. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. Paper presented at the 42nd International Conference of the Association for Computational Linguistics, ACL 2004: Barcelona, Spain.

Babych, B., Hartley A. (2004b). Comparative Evaluation of Automatic Named Entity Recognition from Machine Translation Output. Paper presented at the Workshop on Named Entity Recognition for Natural Language Processing Applications. In Conjunction with the First International Joint Conference on Natural Language Processing IJCNLP-04, Sanya.

Babych, B., D. Elliott, & A. Hartley. (2004): Calibrating resource-light automatic MT evaluation: a cheap approach to ranking MT systems by the usability of their output. LREC-2004: Fourth International Conference on Language Resources and Evaluation, Proceedings, Lisbon, Portugal, 26-28 May 2004; pp.2031-2034.

Babych, B., A. Hartley & D. Elliott (2005). Estimating the predictive power of n-gram MT evaluation metrics across language and text types. In: Proc of MT Summit X, Phuket, Thailand, September 13-15, 2005, Conference Proceedings: the tenth Machine Translation Summit; pp.412-418.

Rajman, M & A. Hartley (2001). Automatically predicting MT systems rankings compatible with fluency, adequacy and informativeness scores. MT Summit VIII, Santiago de Compostela, Spain, 18-22 September 2001. Workshop on MT Evaluation

Sinaiko, H. W. (1979). Measurement of usefulness by performance test. In Van Slype, G. In: Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR 19142. Bureau Marcel van Dijk, p.91

Estrella, P., O. Hamon, & A. Popescu-Belis (2007). How much data is needed for reliable MT evaluation? Using bootstrapping to study human and automatic metrics. MT Summit XI, 10-14 September 2007, Copenhagen, Denmark. Proceedings; pp.167-174.

White, J. S., T. A. O'Connell, F. E. O'Mara (1994). The ARPA MT evaluation methodologies: evolution, lessons,and future approaches. Technology partnerships for crossing the language barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas,5-8 October, Columbia, Maryland, USA

Callison-Burch, Chris, Miles Osborne, & Philipp Koehn. (2006). Re-evaluating the role of BLEU in machine translation research. EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April 3-7, 2006; pp.249-256.

Babych, B. & A. Hartley (2008). Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. LREC 2008: 6th Language Resources and Evaluation Conference, Marrakech, Morocco, 26-30 May 2008; 4pp.