УДК 81'322.4:004.912

V. Yakovyna, V. Masyukevych
Lviv Polytechnic National University,
Software Department

REVIEW AND ANALYSIS OF MACHINE TRANSLATION QUALITY EVALUATION METRICS

This paper is devoted to the overview of main approaches of machine translation, to analysis of existing metrics for machine translation quality estimation. The advantages and disadvantages of these metrics have been revealed.

Key words – machine translation, machine translation quality estimation, reference translation

Introduction. The expansion of the Internet creates the conditions for the international communication. The only obstacle is the language barrier which is difficult to overcome using traditional translation because the amount of translations is growing, and even an increase in number of translators can not fully meet the need for translation. Computerization of the world improvements and new versions of software , attempts to make the software as compatible with a number of devices and applications , and other new developments of industry and technological advances bring to market new products, which need be translated into many languages to help deliver these products to the end user and to increase sales. Machine translation is one of means to overcome the language barrier. It plays an important role in the exchange of research and dissemination of information.

Machine translation systems. There are three main approaches to machine translation: rule-based machine translation, statistical and hybrid machine translation [1].

Rule-based machine translation is characterized by the use and manual creation of linguistic rules. The efficiency of rule-based machine translation systems is defined by the quality of bilingual dictionaries, and by precision of the rules, but their creation requires a long term work. [1].

The first machine translation system were created for specific pairs of languages and based on the complex processes of modeling languages, which were based on methods of analysis, transfer, and synthesis and interlingua. Machine translation of the first generation by direct substitution of words of the source language to the target language. Systems of second generation analyzed the structure of the source language, and then based on the transfer synthesized them to equivalent structures in the source language [1]. The third generation of machine translation systems are based on formal intermediary languages - interlingua. This concept included the transformation of words into the intermediary language, which is a universal language created for the system, independent of the process involved in the translation of languages. This approach uses two methods: analysis and synthesis [2 , 3].

Rule based machine translation systems are based on different levels of linguistic processing of language pair [1]:

1. morphological : lematization of lexical items, search of lexical items in the dictionary, morpheme analysis, recognition of grammar context class of lexical items etc.;

2. syntax : recognition of types of syntactic structures , relations between individual elements of syntactic structure etc.;

3. semantic : separation of lexical meaning meaningful lexical items and affixes , identifying their semantic function , synthesis of their syntactic uniqueness based on semantic analysis [1].

Rule based machine translation systems do not require access to databases of parallel texts, it can be set up, which increases the quality of the translation of specialized texts [4]. Rule-based systems can deal with many linguistic phenomena and are useful in the accompaniment. However, exceptions in grammar add difficulties [2], which requires the development of new and improvements previously developed algorithms [4].

The main disadvantage of rule based machine translation systems is that to improve the quality of these systems there is a need to improve previously created and develop new algorithms, which is very demanding, so modern machine translation systems are based mostly on statistical or hybrid methods. Statistical machine translation systems are based on the automatic extraction of segments of similar language pairs from bilingual full-text of cases that account for billions of tokens. Hybrid machine translation systems [5] are based on the existing rule based machine translation systems with the addition of these statistical methods. Thus, the study of statistical and hybrid machine translation systems is based on the bilingual corpus of texts and requires a deep and complex contrastive linguistic analysis . The possibility to reduce the cost of machine translation systems led to their rapid spread [1].

Statistical technology does not require special linguistic algorithms. The disadvantage of statistical translation is that it does not take into account the rules of grammar. Sometimes using statistical translation system can result in separate words instead of a coherent text. Another problem is that the use of this system requires a very large number of parallel texts. The larger the database of parallel texts is, the more likely matching is to find [4].

Linguistic peculiarities, ambiguities, lack of universal grammar and vocabulary are some of the reasons that the machine translation systems do not achieve 100% accuracy of translation [2]. There are a number of typical errors that occur during machine translation. The grammar translation problems include incorrect recognition of relationships types between sentence parts, which in turn disturbs the word order of the sentence, replacing one part of the sentence with another. Lexical problems, first of all, include wrong the principle of selection of the word correspondences, full or partial untranslatability, wrong translation of terminology [6].

Evaluation of translation quality. Evaluation of machine translation is an important area of research to determine efficiency and to optimize existing machine translation.

For evaluation of machine translation human evaluation is used. Two of the most common metrics of human evaluation is fluency and adequacy [7, 8]. Fluency requires speaker fluent in the target language to judge whether the output of the system is fluent (the text is read as written by a native speaker), regardless of whether the content of the source text is an accurate translation of the input words.

Adequacy ignores the level of fluency of the initial data, while measures whether the information in the text can be obtained from the original text. The requirements for an annotator of adequacy are stricter than for fluency, as the annotator must be bilingual in both the source and target language in order to judge whether the information is preserved across translation.. In practice, annotator , who is fluent only in target language, can also annotate the adequacy using a set of source quality translations performed by human [9, 10].

Automatic machine translation evaluation metrics have been developed due to the high costs, lack of repeatability, subjectivity, and the slowness of the evaluation of machine translation output using human judgment, and the desire to apply the automatic configuration of the system parameters. Such metrics judge the quality of the machine translation output by comparing the output of machine translation systems with a set of reference translations [9]. The reference translation is a translation done by a qualified translator and is recognized to be of high quality.

Metric WER (Word error rate) [9] is defined as Levenshtein distance [11] between the words of the system output and the words of the reference translation divided by the length of the reference translation. Levenshtein distance is calculated using dynamic programming to find an optimal alignment between the output of machine translation and the reference translation. Each word machine translation output is aligned to 1 or 0 words in reference translation, and vice versa. The case where the reference word is aligned to 0 are called the deletion , whereas the alignment of word of machine translation to 0 is called insertion. If a reference word matches the MT output word it is aligned to, it is a substitution. WER is the sums of the number of substitutions (S), insertions (I), and deletions (D) divided by the number of words in the reference translation (N) [9]:

$$WER = \frac{S + I + D}{N}.$$

The disadvantage of the WER metric is that this metric does not adequately combine knowledge of several reference translations, and can not model the rearrangement of words or phrases in the translation.

Metric MWER (multi-reference WER) [12] - the application of WER metric to more than one reference translation – is determined by the minimum value of the WER between output of machine translation and each reference translation.

Metric PER (Position-independent error rate) [13] is an attempt to address the word-ordering limitation of WER. Reference and machine translation are treated as a set of words, so that the word of the reference translation can be related to the word of machine translation, regardless of position. Because of this the PER of an MT output is guaranteed to be lower than or equal to the WER of the MT output. The disadvantage is that the correct translation is not distinguished from that in which words have the wrong order.

The metric BLEU (Bilingual Evaluation Understudy) [14] is the current standard for automatic evaluation of machine translation. A key characteristic of the BLEU metric is the direct use of multiple reference translations. BLEU score of system output is determined by counting the number of n-grams or sequence of words in the output that occur in the set of reference translations. The BLEU score of a system output is calculated by counting the number of n-grams, or word sequences, in the system output that occur in the set of reference translations. BLEU is a precision-oriented metric as it shows how much of the system output is the correct translation, but does not measure whether a reference translation is fully reproduced in the system output. BLEU could be gamed by producing very short system outputs consisting only of highly confident n-grams, if it were not for the use of a brevity penalty which penalizes the BLEU score if the system output is shorter than the references.

$$p_n = \frac{\sum c \in \{can\} \sum n-gram \in c \; Cnt_{clip}\,(n-gram)}{\sum c' \in \{can\} \sum n-gram' \in c' \; Cnt_{clip}\,(n-gram')} \qquad (1)$$

$$BP = f(x) = \begin{cases} 1, & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)}, & if\ c < r \end{cases} \qquad (2)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (3)$$

Equation (1) shows the computation of BLEU precision (p_n) for *n*-grams of length *n*, where *Can* - sentences in the test-corpus, *Cnt (n-gram)* - number of occurrences of *n*-grams in machine translation and Cnt_{clip}*(n-gram)* - the minimum of the unclipped count and the maximum number of times it occurs in a reference translation. Equation (2) shows the calculation of the BLEU brevity penalty (*BP*), where *c* - length of machine translation and *r* - the length of reference translation. These terms are combined, as shown in equation (3), to calculate the total BLEU score, where *N* is usually equal to 4, and w_n is usually set *1/N.*

The disadvantage of BLEU metric is absence of recall in its formulations. BLEU was created and used for large test cases. BLEU score of individual sentences are considered unreliable. Also, the downside is the lack of synonym matching and inability to detect multiple correct word orders. That i, the semantic translation quality is not defined by BLEU metric [9].

METEOR (Metric for evaluation of translation with explicit ordering) [15] - is a tunable, alignment-oriented metric, whereas BLEU is generally a precision-oriented metric. Unlike BLEU, METEOR calculates both precision and recall, combining them to calculate the parameterized harmonic mean:

$$F_{mean} = \frac{P \cdot R}{\alpha P + (1-\alpha) R}.$$

METEOR metric uses several stages to establish correspondence between machine translation and the reference translation in order to compare the two strings [9]:

1. Exact matching. Strings which are identical in the reference and the hypothesis are aligned.

2. Stem matching. Stemming is performed, so that words with the same morphological root are aligned.

3. Synonymy matching. Words which are synonyms according to WordNet [16] are aligned.

In each of these stages, only words that were not matched in previous stages are allowed to be matched.

Unlike BLEU metric, METEOR metric does not penalize longer answers and incorporates a level of linguistic knowledge in the form of its stem and synonym matching allowing it to identify equivalences between the machine translation output and the reference translation. METEOR lacks one of BLEU's key features however: the direct exploitation of multiple references [9].

The GTM metric (General text matcher) [17] attempts to model the movement of phrases during translation by using the maximum matching size to compute the quality of a translation. It finds the longest sequences of words that match between the hypothesis and the reference. The size of the matches is defined by

$$size(M) = \sqrt{\sum_{r \in M} length(r)^2},$$

where M - set of matches found, r - the length of the reference translation. The precision and recall are computed as the size of the matches divided by the length of the system output or the reference, respectively. The score of the GTM is the harmonic mean of precision and recall.

As BLEU, GTM does not incorporate any linguistic knowledge, and only considers words in the MT output and the references as matching if they are identical. If one word in the translation is incorrect, the GTM value depends on the location error in the sentence. Error in the middle of a sentence has the greatest impact, and the error in the end of a sentence has a minimal impact [9].

The TER metric (Translation edit rate) [18] addresses the phrase reordering failing of WER by allowing block movement of words, also called shifts. TER uses a greedy search to select the words to be shifted, and as well as further constraints on the words to be shifted. These constraints are intended to simulate the way in which a human editor might choose the words to shift.

The shifting constraints used by TER serve to both reduce the computational complexity of the model and better model the quality of translation.TER metric uses the following constraints:

1. Shifts are selected by a greedy algorithm that selects the shift that most reduces the WER between the reference and the system output.

2. The sequence of words shifted in the system output must exactly match the sequence of words in the reference that it is being shifted to.

3. The words to be shifted must contain at least one error, according to the WER, before being shifted. This prevents the shifting of words that currently correctly matched.

4. The matching words in reference that are being shifted to must also contain at least one error. This prevents shifting to align to words that already correctly aligned.

If TER is used in case of multiple reference translations, the machine translation is estimated for each reference translation separately. TER is calculated by the formula [9]:

$$TER = \frac{SUB+INS+DEL+SHIFT}{\bar{N}},$$

where SUB, INS, DEL, SHIFT - number of substitutions, insertions, deletions, shifts respectively and $\overline{N}$ - the average number of words in the reference translation.

TER does not use simultaneously several reference translations as BLEU and ignores external linguistic knowledge as METEOR [9].

Metric CDER (Cover disjoint error rate) [19] exploits the fact that the number of blocks in a sentence is equal to the number of gaps among the blocks plus one. Movement blocks can be equivalently described as a long jump operation over the interval between the two blocks. The costs of a long jump are considered constant. Long jumps are combined with the classical Levenshtein edit operations. Long jump distance is the minimum number of operations required for the transformation of machine translation output into reference translation. Like Levenstein distance, long jump distance can be represented using alignment grid [19].

The metric HTER (Human-mediated translation error rate) [18] requires the use of mono-lingual human annotators who create references that are targeted to a particular system output. Target reference translations are created by changing the source data system with minimal changes, so that they fluent and preserves the meaning of the other reference translations. Because a minimal number of edits are used to correct the system output, creating a targeted reference can be thought of as selecting from the set of all possible references the one which is closest to the system output.

Target reference translations can be used for any quality evaluation metrics that use reference translations [9]. However, as the target reference translations are created for each system output , they can not be reused for different systems output and even output from different versions of machine translation systems .

Although HTER is built on human judgments, its most obvious weakness is that it is a purely quantitative metric that weights all errors equally, when in fact some edits, some translation errors, are of trivial importance while others such as some instances of polarity errors can be devastating [9].

The STM metric (Syntax tree based metric) [20]. In order to give a clear and direct evaluation for the fluency of a sentence, syntax trees are used to generate metrics based on the similarity of the machine translation hypothesis's tree and those of the references of trees reference translation and machine translation. One can not hope that the whole syntax tree of the hypothesis can always be found in the references, however, this approach is to be based on the fractions of the subtrees which also appear in the reference syntax trees. For each hypothesis, the fractions of subtrees with different depths are calculated and their arithmetic mean is computed as the syntax tree based metric:

$$STM = \frac{1}{D}\sum_{n=1}^{D}\frac{\sum_{t\in subtrees\ _n(hyp)} count\ _{clip}\ (t)}{\sum_{t\in subtrees\ _n(hyp)} count\ (t)},$$

where D - maximum depth subtrees, *count (t)* - the number of occurrences of subtree t in the machine translation output tree, $count_{clip}$ *(t)* - clipped number of occurrences of t in the syntactic trees of reference translation [20].

Conclusions. To create an advanced machine translation systems there is a need in automated metrics for machine translation quality evaluation. Review of the existing automated metrics showed that:

• all examined the metrics require a set of reference translations, which makes it impossible to evaluate new texts that appear while the machine translation system performance;

• differences in syntactic structure of sentences requires the adaptation of STM metric to the specific language.

1.Mischenko AL Machine translation in the context of modern scientific and technical translation / / Bulletin of Kharkiv National University VN History. - A series of "Philology. Teaching of Foreign Languages. "- № 1051. - P.172-180, 2013. 2. Tripathi S., Sarkhel J. K. Approaches to Machine Translation \\ Annals of Library and Information Studies., Vol. 57, pp. 388-393, 2010. 3. Sanyal S. , Borgohain R. Machine Translation Systems in India \\ Computing Research Repository, 2013 4. Franchuk NP Machine translation / / Scientific Journal nous of MP Drahomanova. Series number 2. Computer-oriented learning systems: Proc. Science. works / Redrada. - K.: NEA Dragomanov, 2010. - № 8 (15). - S. 185-190. 5. Costa-jussà M. R., Banchs R., Rapp R., Lambert P., Eberle K., Babych B. Workshop on Hybrid Approaches to Translation: Overview and Developments //Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 1-6, Sofia, Bulgaria, 2013 6. Smirnova TV Advantages and disadvantages of computer translation / TV Smirnov / / Proceedings of the X International Scientific and Technical Conference "AVIA-2011". - V.4. - K.: NAU, 2011. - S. 37.36 - 37.39. 7. White, J.S., T. O'Connell and F. O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. Proceedings of the First Conference of the Association for Machine Translation, pages 193–205. 8. Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz and J. Schroeder. 2007. (Meta-)evaluation of machine translation. Proceedings of the Second Workshop on Statistical Machine Translation, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics. 9. Olive J. et al.

Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, Springer Science+ Business Media, 2011. 10. Sthamych JS Adequacy and equivalence translation in the context of computational linguistics / / Herald of Zhytomyr State University named after Ivan Franko, Philology. - № 66. - S.235-238, 2012. 11. Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, 10:707–710. 12. Nießen, S., F.J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000), pages 39–45. 13. Tillmann, C., S. Vogel, H. Ney, A. Zubiag and H. Sawaf. 1997. Accelerated DP Based Search For Statistical Translation. European Conference on Speech Communication and Technology, pages 2667–2670, Rhodes, Greece, September. 14. Papineni, K., S. Roukos, T. Ward and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, PA. 15. Banerjee, S. and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pages 65–72, Ann Arbor, Michigan, June. 16. Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. MIT Press. 17. Turian, J.P., L. Shen and D.I. Melamed. 2003. Evaluation of machine translation and its evaluation. Proc. MT Summit IX, pages 386–393, New Orleans, LA. 18. Snover, M., B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of Association for Machine Translation in the Americas (AMTA-2006), pages 223–231, Cambridge, Massachusetts. 19. Leusch, G., N. Ueffing and H. Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006). 20. Liu D., Gildea D. 2005. Syntactic Features for Evaluation of Machine Translation. Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization pages 25–32.