Informative Manual Evaluation of Machine Translation Output

Maja Popović
ADAPT Centre, School of Computing
Dublin City University, Ireland
maja.popovic@adaptcentre.ie

Abstract

We present a new method for manual evaluation of Machine Translation (MT) output accepted at COLING 2020 (Popović, 2020). The method is based on marking actual issues in the translated text. The novelty is that the evaluators are not assigning any scores, nor classifying errors, but marking all problematic parts (words, phrases, sentences) of the translation.

The main advantage of this method is that the resulting annotations do not only provide overall scores by counting words with assigned tags, but can be further used for analysis of errors and challenging linguistic phenomena, as well as inter-annotator (dis)agreements. Detailed analysis and understanding of actual problems are not enabled by typical manual evaluations where the annotators are asked to assign overall scores or to rank two or more translations.

The proposed method is very general: it can be applied on any genre/domain and language pair, and it can be guided by various types of quality criteria. Also, it is not restricted to MT output, but can be used for other types of generated text guided by corresponding criteria.

1 Motivation

For manual evaluation of machine translation[1], the annotators are usually asked to assign an overall quality score for the given MT output, or to rank two or more competing outputs from best to worst. Still, neither of these two annotation methods provides any details about actual errors and problems.

This drawback is usually overcome by performing error classification, where the evaluators are asked to mark each translation error and assign an error tag from a set of predefined categories. However, this approach requires much more time and effort, both from annotators as well as from organisers (to define an appropriate error taxonomy which is not a trivial task, to prepare clear guidelines for each error class, and to train the annotators).

Our method can be seen as a "mid-way" between overall assessment and error classification, and its advantage is two-fold. First, it is more informative than assigning overall scores because the actual problematic words/phrases/sentences are marked and they can be further used for more detailed analysis. Second, the annotation process does not require any additional effort in comparison to assigning scores or ranking, which is much less effort than for error classification. Also, overall scores can be automatically extracted from annotated text.

2 Evaluation method

All existing methods for the human evaluation of MT output, such as (ALPAC, 1966; White et al., 1994; Koehn and Monz, 2006; Vilar et al., 2007; Graham et al., 2013; Forcada et al., 2018; Barrault et al., 2019), are essentially based on some of the following three quality criteria: adequacy (accuracy, fidelity), comprehensibility (intelligibility) and fluency (grammaticality). The choice of criteria often depends on the task and on the purpose of the translation.

Our evaluation process is guided by comprehensibility and adequacy. The annotators were asked to distinguish two levels of issues for each criterion: major issues (e.g. incomprehensible/not conveying the meaning of the source) and minor issues (e.g. grammar or stylistic errors/not an optimal translation choice for the source). The annotators were also asked to annotate omissions by adding "XXX" to the corresponding position in translation.[2]. It is arguable whether the distinction between major

[1] and other natural language generation tasks

[2] For comprehensibility, the omission tags were added if it seemed that some parts were missing

and minor errors is really necessary, but we did not want to let any errors unannotated. Furthermore, it is arguable whether the minor errors are actually representing fluency. An interesting direction for future work is to include fluency as criterion and annotate only major errors for the three criteria.

The evaluation is carried out on the review ("document") level, and not on the sentence level. In this way, it was ensured that the annotators were able to spot context-dependent issues. The translation outputs were given to the evaluators in the form of Google Doc, and they were asked to mark major issues with red colour and minor issues with blue colour. Each MT output was annotated by two annotators in order to obtain more reliable annotations and estimate inter-annotator agreement.

The described experiment was carried out on user reviews[3] (a case of "mid-way" genre between formal and informal written language) translated into Croatian and Serbian (a case of mid-size less-resourced morphologically rich European languages), but the method can be applied on any genre/domain and language pair. Two quality criteria were used in this work, comprehensibility (monolingual) and adequacy (bilingual), but the method can be guided by any other criterion (such as fluency). In addition, the method is not necessarily restricted to evaluation of MT output, it can be applied on any type of generated text. @inproceedingsevaluation20, Title = Informative Manual Evaluation of Machine Translation Output, Author = Maja Popović, BookTitle = Proceedings of the 28th International Conference on Computational Linguistics(COLING 2020), year = 2020, month = December, address = Online,

Inter-annotator agreement (IAA) In order to estimate IAA, we calculated two scores using the assigned word labels "Major", "Minor" and "None": F-score and normalised edit distance (also knows as WER – Word Error Rate):

- F-score: number of matched labels divided by the total number of words. Due to possible different lenghts of annotated sentences, the matches are defined as position-independent, which might introduce over-agreement.
- normalised edit distance, divided by the total number of words. It penalises differences in position, thus compensating the drawback of the position-independent F-score.

[3]IMDb movie reviews and Amazon product reviews

IAA (%)	compreh.	adequacy
F-score ↑	85.5	86.6
edit distance ↓	27.2	23.9

Table 1: Inter-annotator agreement (IAA) for comprehensibility and adequacy: F-score and normalised edit distance.

Table 1 shows that the agreements are high, which could be expected because no fine-grained classification was required.

It should be noted that we did not use Cohen's Kappa coefficient for several reasons. First reason is that it requires word-by-word comparisons, which is not possible for our annotations due to omission tags "XXX". Another reason is that it requires separate IAA for each pair of annotators, and in our annotations, there is a large number of different annotator pairs. Finally, the general property of the Kappa coefficient is debatable, namely the assumption that annotators will make random choices. This assumption heavily penalises a large number of agreements and understates the actual agreement.

3 Summary

We propose a method for manual evaluation of MT outputs where evaluators are marking actual problematic parts of the text (words, phrases, sentences) in the translation. The method thus does not provide only overall scores by counting the assigned tags, but also enables further detailed analysis of the annotated texts.

Apart from evaluation, the same method (marking errors) was used to improve an English-into-German NMT system by learning from marked errors (Kreutzer et al., 2020). It is also reported that marking errros lead to same improvements as post-editing while requiring much less time.

The method is not restricted to MT: it could be extended to other natural language generation tasks by defining the appropriate criteria and guidelines. For text simplification, for example, adequacy and fluency can be used directly ("meaning preservation" and "grammaticality"), while specific guidelines for "simplicity" could be defined such as "mark all words and/or parts of the text which are difficult to understand" or "which require more time to read".

References

ALPAC. 1966. Language and machines. Computers in translation and linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (WMT 2019)*, pages 1–61, Florence, Italy.

Mikel L. Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. 2018. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 192–203, Brussels, Belgium. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation (WMT 2006)*, pages 102–121, New York City. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct Me If You Can: Learning from Error Corrections and Markings. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020)*.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics(COLING 2020)*, Online.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT 2007)*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

John White, Theresa OConnell, and Francis OMara. 1994. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas (AMTA 1994)*, pages 193–205.