Informative Manual Evaluation of Machine Translation Output

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

Abstract

This work proposes a new method for manual evaluation of Machine Translation (MT) output 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 disagreements. 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.

1 Introduction

While automatic evaluation metrics are very important and invaluable tools for rapid development of machine translation (MT) systems, they are only a substitution for human assessment of translation quality. Various methods have been proposed and used 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), and all of them are essentially based on some of the following three quality criteria: adequacy (accuracy, fidelity), comprehensibility (intelligibility) and fluency (grammaticality). Adequacy measures how well the meaning of the original text is conveyed to the translated text. Comprehensibility reflects how well a reader is able to understand the translated text without access to the original text. Fluency describes the grammar of the target language in the translated text. The choice of criteria often depends on the task and on the purpose of the translation: for example providing translations to users requires high comprehensibility and adequacy, comparing different MT systems is based mainly on adequacy, etc.

Regardless of the quality criterion, 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. One advantage of direct scoring over ranking is that the resulting annotations include not only the information that one MT output is better than another, but also the degree to which that output was better than the other. Also, it gives an estimation about the absolute quality of each MT output. Another advantage is that direct estimation of translation quality extends the usefulness of the annotated data to other tasks such as quality estimation.

Still, neither of these two annotation methods provides any details about actual errors and problems. The usual way to overcome this drawback is to perform error classification, where the evaluators are asked to mark each translation error and assign an error tag from a set of predefined categories. However,

Proceedings of the 28th International Conference on Computational Linguistics, pages 5059–5069
Barcelona, Spain (Online), December 8-13, 2020

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).

In the proposed approach, the evaluators are asked to mark problematic parts in translation without assigning any scores or error labels. The advantage of such approach 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. The results of one possible analysis, namely comparison of comprehensibility and adequacy issues, are presented in (Popović, 2020). 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 (in the form of percentage of problematic words) can be automatically extracted from annotated text.

Although the described experiment was carried out on user reviews (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), 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.

2 Related work

The first report about manual evaluation of machine translation (ALPAC, 1966) defines "intelligibility" (comprehensibility) and "fidelity" (adequacy) as the two major characteristic of a translation. For each of those two criteria, the evaluators assigned overall quality scores on 9 point scales. The ARPA MT Initiative (White et al., 1994) defines "adequacy", "fluency" and "comprehension" as a standard set of concepts for MT evaluation, and the evaluators are instructed to assign overall scores based on those three criteria.

Years later, the first WMT (Workshop/Conference on Machine Translation) translation shared task in 2006[1] as well as the subsequent task in 2007 adopted adequacy and fluency as official metrics (Koehn and Monz, 2006; Callison-Burch et al., 2007). The participants in the shared task are asked to assign adequacy and fluency scores on 5 point scales to each submitted MT output.

Later on, a binary ranking of two MT outputs was proposed (Vilar et al., 2007), and in a slightly changed form (comparing up to five outputs instead of two) became the official metric at WMT 2008 (Callison-Burch et al., 2008). It was reported that it required less effort and showed better inter-annotator agreement than adequacy and fluency scores, and it remained the official WMT metric until 2016. The quality critera for ranking were never explicitly defined, but it was implicitly based on a combination of adequacy and fluency.

Another method for assigning quality scores, continuous direct assessment (Graham et al., 2013), does not use discrete scales but assigns a continuous score between 0 and 100. Both adequacy and fluency were investigated as the guiding criterion, and it was concluded that the best option was to be guided by adequacy and to use fluency only as auxiliary criterion for adequacy ties (Bojar et al., 2016). Continuous direct assessment eventually replaced ranking at the WMT shared tasks in 2017 (Bojar et al., 2017) and is still used as the official WMT metric (Barrault et al., 2019).

Some publications dealt with evaluating user reviews: the annotators were asked to assign scores on 5 point scale for comprehensibility and scores on 2 point scale ("true" or "false") for fidelity (Roturier and Bensadoun, 2011). Those annotations were later used for quality estimation of these two evaluation criteria (Rubino et al., 2013). Other publications propose alternative methods for assessing comprehensibility, namely question answering (Scarton and Specia, 2016) as well as filling gaps as its cheaper version (Forcada et al., 2018).

Nevetheless, none of these methods provide information about actual parts of the translation which are problematic in terms of the given quality criterion. Our method overcomes this drawback by marking

[1] http://statmt.org/wmt06/

actual issues. Furthermore, our method is much less demanding than error classification, where the evaluators are asked to assign error classes according to a predefined typology, for example (Vilar et al., 2006), the MQM scheme[2] (Lommel et al., 2014; Klubička et al., 2018), etc. Although our method does not involve categorisation of marked issues, the annotated texts can be further used for many tasks with different error typologies (such as classifying translation errors, identifying linguistic phenomena causing the errors, concentrating on particular type(s) of erros and/or of phenomena, etc.) Furthermore, categorisation is then performed on already marked errors which reduces effort in comparison to finding and classifying errors from scratch.

3 Evaluation method

The novelty of the proposed method is that the evaluators are not assigning any scores, nor classifying errors, but marking all problematic parts of the text (words, phrases, sentences). The method can be seen as a "mid-way" between overall assessment and error classification. The advantage over overall assessment is that our annotation method provides much more information without requiring additional effort: assigning overall scores is guided by issues/errors in the translation, and our method requires direct marking of these issues instead of thinking about an overall score based on them. Our method also provides overall scores, by counting all marked words. The advantage over error classification is that it is much less demanding while still informative – the annotations can be further used for different kinds of analyses (with reduced effort in comparison to analysing from scratch), such as error classification, identification of problematic (linguistic) phenomena, etc.

The proposed evaluation process makes use of two well-known quality criteria, namely comprehensibility and adequacy which were already used for evaluating user reviews but in the form of assigning scores (Roturier and Bensadoun, 2011). Comprehensibility reflects the degree to which a translated review can be understood, and adequacy reflects the degree to which the translation conveys the meaning of the original text in the source language. 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). It should be stressed that comprehensibility is not fluency – a fluent text can be incomprehensible (for example "Colorless green ideas sleep furiously."), and vice versa (for example "All these experiment was carry out this year.").

The evaluation procedure consisted of two independent subsequent tasks:

1. Marking all issues related to comprehensibility – a monolingual evaluation task without access to the original source language text.

2. Marking all issues related to adequacy – a bilingual evaluation task with access to the original source language text.

The evaluation was performed by computational linguistics students and researchers, fluent in the source language and native speakers of the target language. In total, thirteen evaluators participated in the study, of which three have had experience with MT.

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. In addition to general definitions of comprehensibility and adequacy, the following detailed guidelines were given to the evaluators:

1. Comprehensibility:
    - mark as red all parts of the text (single words, small or long phrases, or entire sentences) which are not understandable (it does not make sense, it is not clear what it is about, etc.);
    - mark as blue all parts of the text (again: words, phrases or sentences) which might be understandable but contain grammatical or stylistic errors;

[2] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

(a) comprehensibility

MT output	Ako pravilno zavijate zglobove, vidjet ćete da su i preuska i prekratka, prekratka.
+ gloss	If properly you_are_wrapping wrists, see you_will that they_are both too_narrow and too_short, too_short.
	Ne shvaćajte ih ako udarite u tešku torbu.
	Don't understand them if you hit in heavy suitcase.
	Oni jednostavno neće zaštititi / podržavati vaše zglobove ili ručne zglobove.
	They simply won't protect / support your joints or hand joints.

(b) adequacy

source	If you wrap your wrists properly, you'll see these are both too narrow and too short, way too short.
	Do not get these if you are hitting the heavy bag.
	They just won't protect/support your wrists or knuckles.
MT output	Ako pravilno zavijate zglobove, vidjet ćete da su i preuska i prekratka, prekratka.
+ gloss	If properly you_are_wrapping wrists, see you_will that they_are both too_narrow and too_short, too_short.
	Ne shvaćajte ih ako udarite u tešku torbu.
	Don't understand them if you hit in heavy suitcase.
	Oni jednostavno neće zaštititi / podržavati vaše zglobove ili ručne zglobove.
	They simply won't protect / support your joints or hand joints.

Table 1: Example of a short review (three sentences) translated into Croatian and annotated with respect to: (a) comprehensibility (b) adequacy. Red colour denotes major issues whereas blue colour represents minor issues.

- if it seems that some parts are missing, add "XXX" in the corresponding color to the corresponding position.

2. Adequacy:

   - mark as red all parts of the translation (single words, small or long phrases, or entire sentences) which have different meaning than the original English text;
   - mark as blue all parts of the translation (again: words, phrases or sentences) which do not essentally change the meaning of the source text, but contain grammar errors or sub-optimal lexical choices.
   - if some parts of the original English text are missing in the translation, add "XXX" to the corresponding position in the translation and mark it with the corresponding color; mark the English part as yellow;
   - if there are any errors in the source language[3] (spelling or grammar errors, etc.):
     - mark it with green colour;
     - mark its translation as red or blue if it does not correspond to the intented English word even though it is a correct translation of the erroneous English word.

An example of a short review annotated for comprehensibility (above) and adequacy (below) can be seen in Table 1. Table 2 presents examples of annotating omissions (above) and errors in the original text (below).

Each annotator performed first the monolingual comprehensibility task and afterwards the bilingual adequacy task. In total, each annotator evaluated about 5000 (non-tokenised) words for each of the two quality criteria. All annotators managed to complete both tasks in 3 to 7 (non-consecutive) hours. Each MT output was annotated by two annotators in order to obtain more reliable annotations and estimate inter-annotator agreement.

[3] Detailed instructions for errors in the source text are particularly relevant for translating user generated content.

source	I cannot recommend this grill cover.
MT output	Ne mogu vam preporučiti ovo XXX roštilj.
gloss	Not I_can to_you recommend this XXX grill.
source	Red and Blue where always off the shelfs.
MT output	Red i Blue gdje uvijek off police.
gloss	Red and Blue where always off shelf.

Table 2: Examples of annotating omissions (above) and errors in the source text (below). Red colour denotes major issues whereas blue colour represents minor issues. Omitted parts are marked yellow in the source text and as "XXX" in the translation. Errors in the source text are marked green.

It should be noted that in the view of recent findings and recommendations regarding manual MT evaluation (Läubli et al., 2018; Castilho et al., 2020), 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.

4 Data sets

We have been working with two types of publicly available user reviews:

- IMDb movie reviews[4] (Maas et al., 2011)

  IMDb movie reviews consist of 10.8 sentences and 230.1 words on average. Each review is labelled with a score: negative reviews have a score<4 out of 10, positive reviews have a score>7 out of 10, and the reviews with more neutral ratings are not included.

- Amazon product reviews[5] (McAuley et al., 2015)

  Amazon product reviews are generally shorter, consisting of 5.4 sentences and 93.2 words on average. Each review is labelled with a rating from 1 (worst) to 5 (best). The reviews are divided into 24 categories, such as ”Sports and Outdoors”, ”Musical Instruments”, etc. In this evaluation, we used the reviews from the following 14 categories: ”Beauty”, ”Books”, ”CDs and Vinyl”, ”Cell Phones and Accessories”, ”Grocery and Gourmet Food”, ”Health and Personal Care”, ”Home and Kitchen”, ”Movies and TV”, ”Musical Instruments”, ”Patio, Lawn and Garden”, ”Pet Supplies”, ”Sports and Outdoors”, ”Toys and Games”, and ”Video Games”.

In total, 28 IMDb and 122 Amazon reviews (16807 untokenised English source words) are covered in this evaluation. Average length of the selected IMDb reviews was 12.4 sentences and 214.3 words, whereas for Amazon reviews was 6.5 sentences and 87.7 words.

The selected English user reviews were then translated into Croatian and into Serbian. The goal of the experiment is not to evaluate or compare particular MT systems, but to explore a new evaluation strategy. For this purpose, MT outputs[6] of three on-line systems were used: Google Translate[7], Bing[8] and Amazon translate[9].

From the selected English reviews, 900 MT outputs were generated (150 reviews were translated by three MT systems into two target languages, thus six translations for each of the 150 reviews = 900), and 270 of them were included in the manual evaluation. More details about the data can be seen in Table 3. The annotated data sets are publicly available under the Creative Commons CC-BY licence.[10]

[4] https://ai.stanford.edu/~amaas/data/sentiment/

[5] http://jmcauley.ucsd.edu/data/amazon/

[6] generated at the end of January 2020

[7] https://translate.google.com/

[8] https://www.bing.com/translator

[9] https://aws.amazon.com/translate/

[10] https://github.com/m-popovic/QRev-annotations

	number of source			number of translated reviews			
corpus	reviews	sentences	words	all	amazon	bing	google
IMDb	28	347	6111	50	21	12	17
Amazon	122	791	10707	220	86	55	79

Table 3: Statistics of the annotated data set: number of reviews, sentences and (untokenised) words in the original English text (left), and number of translated reviews included in the evaluation: all, translated by Amazon, translated by Bing and translated by Google (right).

	annotated texts F-score / edit distance
1)	Ako pravilno zavijate zglobove, vidjet ćete da su i preuska i prekratka, prekratka.
	Ako pravilno zavijate zglobove, vidjet ćete da su i preuska i prekratka, prekratka.
seg. IAA	92.3 / 7.0
2)	Ne shvaćajte ih ako udarite u tešku torbu.
	Ne shvaćajte ih ako udarite u tešku torbu.
seg. IAA	33.3 / 66.0
3)	Oni jednostavno neće zaštititi / podržavati vaše zglobove ili ručne zglobove.
	Oni jednostavno neće zaštititi / podržavati vaše zglobove ili ručne zglobove.
seg. IAA	63.6 / 36.0
4)	Nadmašio me na svakom koraku i stalno me iznenadila priča.
	Nadmašio me XXX na svakom koraku i stalno me XXX iznenadila priča.
seg. IAA	87.0 / 26.0
5)	Zabavno je gledati u prvih petnaestak minuta jer je loše, ali nakon toga ide sve XXX gore.
	Zabavno je gledati u prvih petnaestak minuta jer je loše, ali nakon toga ide sve gore.
seg. IAA	86.5 / 27.0
total IAA	76.2 / 30.2

Table 4: Examples of inter-annotator (dis)agreement and the scores (segment-level and total): F-score and edit distance.

5 Results

5.1 Inter-annotator agreement

As a part of the evaluation of the results of the annotation task, we assessed inter-annotator agreement (IAA), sometimes known as inter-rater reliability. One reason for human disagreement in the case of MT evaluation is the simple fact that there is no single correct translation for a given source text so that errors/issues can often be interpreted in multiple ways. 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.

Table 4 presents examples of annotations obtained by different evaluators together with the IAA scores, for each segment and overall. Example 1) has very high agreement (high F-score and low edit distance) because only one word is different (in terms of major vs. minor issue). On the other hand, examples 3) and 4) have much higher edit distance despite the high F-score.

Table 5 shows that the agreements are high, which could be expected because no fine-grained classification was required. Besides disagreements between the issue types (examples 1, 4, 5 in Table 4) or

IAA (%)	F-score ↑	edit distance ↓
comprehension	78.0	27.8
adequacy	81.8	23.9

Table 5: Inter-annotator agreement (IAA) for comprehensibility and adequacy: F-score and normalised edit distance.

presence/absence of minor issues (example 3), disagreements are often caused by different spans (example 2). It can be also noted that the agreement is higher for adequacy than for comprehension, probably because adequacy is guided by the source text while comprehension allows more subjective judgments. Further analysis of exact nature of disagreements can also provide insights into human perception of translation quality.

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.

5.2 Feedback from annotators

Before the evaluation started, all annotators obtained detailed description of comprehension and adequacy together with the guidelines. After that, each evaluator was given a small text of about 500 English words. The purpose of this preliminary round was to check how the evaluation will go in practice and to clarify all potential doubts which might seem clear in theory but arise only in practice. In this round, many annotators asked how exactly to treat translations of incorrect English words, which was not explained in the initial guidelines. Therefore, the adequacy guidelines were adjusted, and after completing both tasks on the introductory text, the evaluators were presented with the rest of the text and updated guidelines. They were encouraged to ask any question at any moment during the whole evaluation. The majority of evalutors did not ask any additional questions, some asked a few minor questions, whereas three evaluators were asking questions frequently. They were previously involved in error classification tasks (although not for translation), so they occasionally had doubts whether they should think about error categories, too.

After completing both tasks on the entire assigned text, each evaluator was asked to fill up the questionnaire about the evaluation process. The questions and the summary of responses were:

1. Did you read the whole review first or were you annotating while reading?
    - five annotators first read the whole review and then marked issues
    - five were annotating while reading
    - one annotated paragraph-by-paragraph
    - two evaluated comprehensibility while reading, and adequacy sentence-by-sentence
2. Which task (quality criterion) was more difficult for you to evaluate, comprehension or adequacy?
    - seven evaluators found comprehensibility more difficult
    - five evaluators found adequacy more difficult
    - one found both tasks equal
3. What did you find difficult in the first task (comprehension)?
    - Five annotators were often trying to figure out what the incomprehensible parts of the translation are actually supposed to mean, and found it very hard. This was not at all required/expected by the definition of comprehensibility, and, of course, not even possible without the source text.

- Four evaluators found difficult to decide between minor vs. major issues. Two of them said that they had some influence of English knowlegde so that they might have marked some objectively major issues as minor.
- Eight annotators found difficult to decide about the span (whether to mark words, phrases, or entire sentence). One annotator mentioned span for minor word order errors.

4. What did you find difficult in the second task (adequacy)?
   - Four annotators found it difficult to follow and understand the text in two languages (first source and then target). Two of them said that it would be much easier to see a sequence of split sentences instead of a sequence of entire reviews as shown in Table 1 (English sentence above and translated sentence below instead of English review above and translated review below).
   - Five annotators found difficult to decide between minor vs. major issues. Two mentioned informal nature of the text, and two said that they were trying to avoid personal translation preferences.
   - Five annotators found difficult to decide about the span (whether to mark words or phrases). Two mentioned the span for word order errors.
   - One annotator found marking omissions very difficult.

Finally, although they were not asked for that, several annotators reported that certain linguistic phenomena were repeatedly appearing in translations as problematic: subject-verb agreement, ambiguous words, gender, named entities, as well as overuse of "it" and "this" in target languages.

5.2.1 Lessons learnt

Taking into account all evaluators' comments, questions and responses, the following lessons have been learnt from this first evaluation and will be considered in future.

Comprehensibility Considering that comprehensibility is a monolingual task and adequacy bilingual, it was kind of surprise that seven annotators found comprehensibility more difficult than adequacy. However, listing the difficulties in both tasks revealed that many annotators tried to figure out what the translation really means, which is irrelevant and out of scope of the comprehensibility by its definition, and also not at all possible without the original source language text. This should not happen in future evaluations and therefore it should be clearly explained in guidelines.

Adequacy The user reviews are short, so we did not perform any sentence alignment between the original text and the translation, but showed the whole review in English followed by the whole translated review. However, the responses indicated that the cognitive load of working on two sources of information required by the definition of adequacy can be reduced by replacing the review-level view with the sentence-level view.

Distinguishing major and minor issues For comprehensibility, one possibility is to find annotators without knowledge of the source language. Another possibility is to ask evaluators to mark only major issues, and add a task with fluency as the quality criterion. The exact concept might be dependent on the main goal of evaluation.

Span Many evaluators were not sure how to decide about the exact span of certain issues. This is, however, very subjective, and therefore cannot be precisely guided by instructions.

5.3 Overall scores: word issue rates

Table 6 shows the overall scores obtained in this evaluation, namely percentages of words marked as problematic. It can be seen that the annotations provide both overall translation performance of a particular MT system as well as comparison between different MT systems.

For example, the following can be observed: overall, there are less issues for translation into Croatian than into Serbian. For both target languages, Bing exhibits the largest number of issues, especially

(% of issues)		comprehension		adequacy	
	system	major	minor	major	minor
en→hr	all	9.1	12.4	8.0	12.4
	amazon	7.9	11.9	6.6	11.7
	bing	14.7	15.7	13.0	16.8
	google	7.2	10.9	6.8	10.6
en→sr	all	13.0	19.3	12.0	14.6
	amazon	13.1	20.0	10.8	15.5
	bing	17.7	19.5	17.3	14.5
	google	9.5	18.4	10.3	13.8

Table 6: Percentages of words problematic for comprehensibility and adequacy for each target language: overall and for each of the three on-line MT systems.

major issues. Google generates the most comprehensible translations for both target languages, whereas Amazon generates the most adequate translations in terms of major issues.

5.4 Ongoing analysis: categorisation of major issues

As mentioned in previous sections, the annotated data can be further used for different types of detailed analyses. Ongoing analysis aims to identify types and causes of all major issues.

Adequacy issues in Table 2 can be categorised as follows: The major issue in the first sentence is the omitted part of the noun phrase "grill cover". In the second sentence, the first major issue is caused by spelling error in the source ("where" instead of "were"), so that, although the translation is correct for the misspelled word, it is not conveying the intented meaning of the source. Interestingly, the second misspelled Enlish word "shelfs" did not result in issues in the MT output. Second major issue is the untranslated word "off". Apart from major issues, there is a minor issue in the first sentence, namely incorrect gender of the determiner "this" for the object "grill cover".

First results show that the most frequent cause of major adequacy issues in both target languages are ambiguous words (translation of the word is in principle correct, but not in the given context). For translation into Serbian, the next frequent causes are mistranslation, noun phrases, translation of named entities which should not be translated, and word-by-word translation (phrases/word groups where for each English word, a potentially correct Serbian word is chosen randomly without taking the context into account). For translation into Croatian, the next frequent issue types are untranslated words (English words copied into translation), mistranslation, subject-verb agreement, noun phrases, and non-existing words.

6 Conclusions

This work presents a novel method for manual evaluation of MT output where the evaluators are not assigning any scores, nor classifying errors, but are marking actual problematic parts of the text (words, phrases, sentences) in the translation. The proposed method does not provide only overall scores by counting the assigned tags, but also enables further detailed analysis of the annotated texts.

The annotators were asked to distinguish only two levels of issues, major and minor. Since it does not include fine-grained error classification, inter-annotator agreement is high – annotators agree on about 70-80% of issues on the word level. Many disagreements are due to assigning different spans – some annotators preferred to mark whole phrases/sentences whereas others preferred to chose one or two words. A detailed analysis of disagreements which could represent an interesting direction towards better understanding the evaluation process is also possible on texts annotated in the described way.

The evaluation was carried out on English user reviews translated into Croatian and Serbian, and was based on two quality criteria: comprehensibility and adequacy. Hovever, the method can be applied on any genre/domain and language pair, and can be guided by any other quality criterion (such as fluency). It could also be used for comparing ("ranking") two[11] MT outputs by marking differences between them as "better" or "worse". Generally, the method can be applied to any type of generated text.

[11] not possible for more

Future evaluations should include more domains and language pairs, and take into account the most important lessons learnt from the annotators' feedback: show the original-translation pairs on the sentence level (instead of showing whole translated review ("document") below the whole original review) in order to reduce cognitive load, and clearly explain in the guidelines that comprehensibility does not include figuring out the "true" meaning of the original text. Apart of that, it might be worth to develop a more sofisticated annotation tool, or to extend an existing tool. Furthermore, it would be good to directly compare the proposed method with conventional types of manual evaluation, namely ranking, assigning overall scores, or even post-editing. Also, it might be worth to explore how to end up with a final quality score, because simple average overall error count might not be the optimal solution.

Acknowledgments

This research is being conducted with the financial support of the European Association for Machine Translation (EAMT) under its programme "2019 Sponsorship of Activities" at the ADAPT Research Centre at Dublin City University. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. We would like to thank all the evaluators for providing us with annotations and feedback.

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 (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, May.

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: Research Papers*, pages 192–203, Brussels, Belgium, October. 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, August. Association for Computational Linguistics.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative Fine-grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, 32(3):195–215, September.

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*, pages 102–121, New York City, June. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 4791–4796, Brussels, Belgium, October-November.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014)*, pages 165–172.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and Human Language Technologies (ACL-HLT 2011)*, pages 142–150, Portland, Oregon, USA, June.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 43–52, Santiago, Chile.

Maja Popović. 2020. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 2020)*, Online, November.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the MT Summit XIII*, Xiamen, China, September.

Raphael Rubino, Jennifer Foster, Rasoul Samad Zadeh Kaljahi, Johann Roturier, and Fred Hollowood. 2013. Estimating the Quality of Translated User-Generated Content. In *Proceedings of 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 1167–1173, Nagoya, Japan, October.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia, May. European Language Resources Association (ELRA).

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

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*, pages 96–103, Prague, Czech Republic, June. Association for Computational Linguistics.

John White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*, pages 193–205.