Machine Translation Quality of User Reviews
– final report, January 2021 –

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

1 Introduction

This project was funded by the European Association for Machine Translation through its 2019 sponsorship of activities programme. This document contains the final report of the project describing the completed activities.

The aim of the project was to identify the important aspects of translation quality on the case of user reviews. User reviews were chosen as a case of "mid-way" genre between formal and informal, an abundant genre which has not been widely investigated yet in machine translation. A set of publicly available user reviews were translated into Croatian and Serbian, as a case of mid-size less-investigated morphologically rich European languages.

Altough the project covered only one genre and two language pairs including two similar target languages, the described methods can be applied on any genre/domain and language pair.

2 Data

We were working with two types of publicly available user reviews: IMDb movie reviews[1] and Amazon product reviews.[2]. In total, 28 IMDb and 122 Amazon English reviews (16807 untokenised English source words) are covered. In total, five translation systems were used during the project to produce translation outputs for quality analysis: three on-line systems (Google Translate[3], Amazon Translate[4] and Bing[5]) and two in-house systems (one trained on general domain and one tailored for user reviews). The on-line MT outputs were generated in January 2020, and the in-house MT outputs were generated in September 2020.

From the original English text, 1500 MT reviews were generated (150 reviews translated by five MT systems into two target languages, thus ten translations for each of the 150 reviews = 1500), and 428 of these translations were included in the manual annotation (3334 sentences and about 45000 words in total). Each of those 428 translated reviews is annotated by two annotators in order to obtain more reliable annotations and estimate inter-annotator agreement.

After the annotators completed their work, we have analysed the marked errors in order to determine their type and cause.

All data created during this project is publicly available under the Creative Commons CC-BY licence.[6]

The fully annotated texts consist of words, issue type for each word (if identified, otherwise *None*) and its error mark (*Major*, *Minor* or *None*). An example of a sentence can be seen below, with three minor errors: two of them were caused by incorrect mood of a verb and one by ambiguity of the source word.

[1] https://ai.stanford.edu/~amaas/data/sentiment/
[2] http://jmcauley.ucsd.edu/data/amazon/
[3] https://translate.google.com/
[4] https://aws.amazon.com/translate/
[5] https://www.bing.com/translator
[6] https://github.com/m-popovic/QRev-annotations

Da—None—None smo—None—None znali—None—None koliko—None—Minor
bi—MOOD—Minor bile—MOOD—Minor neprijatne—AMBIGUITY—Minor
ove—None—None stolice—None—None ...—None—None

Table 1: Example of a marked and analysed sentence.

3 Manual annotation (marking) of issues in MT outputs (funded by EAMT)

In order to be able to identify the important aspects of translation quality, the very first step is manual annotation of problematic parts of MT outputs. Two quality criteria were used in this project: comprehensibility (monolingual) and adequacy/fidelity (bilingual), but the procedure can also be guided by other quality criteria. The annotators were asked to concentrate on problematic parts of the text and to mark them, without assigning any scores or classifying errors. In order to not let any error unmarked, they were asked to distinguish major and minor errors.

The funding from EAMT was necessary for this annotation, in order to cover as much diverse MT outputs as possible and to annotate each of them by two annotators. The money was spent to pay 14 different annotators in total. Each of them evaluated about 2500-5000 target language words, first for comprehensibility, without seeing the English source text, and afterwards for adequacy. All the annotators were computational linguistic students and/or researchers, with backgrounds in translation studies, humanities, technical and/or computer science.

Since this type of manual annotation is new, a paper about it is published at COLING (Popović, 2020a), and also presented at the Workshop on Evaluating NLG Evaluation at INLG 2020.[7]

Timeline of the annotation

The initial plan (presented in the project proposal) was to annotate Google and Bing outputs and to try to develop an in-house system, so that if this system demonstrates acceptable performance to include its output in the second round. In total, outputs from two or three different MT systems were initially planned.

However, the plans have been slightly changed during the project. First, another on-line system became available at the end of 2019: Amazon Translate published their systems for the given two languages pairs, so we included these outputs, too.

Furthermore, the performance of the in-house system was much better than initially expected – the automatic scores after first steps published at the Worskhop on NLP for Similar Languages, Language Varieties and Dialects (Popović et al., 2020) were already promising, and later we managed to develop two systems which outperformed all three on-line systems in terms of automatic evaluation. Therefore, two in-house systems were included in the manual annotation, too.

The annotation had been carried out in three rounds. In the first round, in February 2020, a small portion of 8 reviews translated by Amazon and Google were given to the annotators in order to get familiar with the method and clarify any potential questions and doubts. In the second (and largest) round, in spring 2020, the outputs of the three on-line systems were annotated, and the outputs of the two in-house systems were annotated in the third round, in October 2020.

Percentage of marked issues

The percentage of marked issues for each MT output as well as for the entire evaluated text are shown in Table 2. Although the purpose of the project was not to compare the systems, an overview of percentages of issues in different MT outputs gives an idea of how many errors have been spotted in the texts.

Inter-annotator agreement

Inter-annotator agreement (IAA) is shown in Table 3 in terms of Krippendorf's α, F-score and normalised edit distance (also known as Word Error Rate).

[7] https://evalnlg-workshop.github.io/papers/EvalNLGEval_2020_paper_3.pdf

% on the word level		comprehension			adequacy		
	system	major	minor	none	major	minor	none
en→hr	amazon	8.0	11.9	80.1	6.6	11.7	81.7
	bing	15.0	16.0	69.0	13.2	17.0	69.8
	google	7.3	10.8	81.9	6.8	10.6	82.6
	dcu-gen	7.1	7.5	85.4	6.1	8.1	85.8
	dcu-rev	6.7	6.6	86.7	4.7	8.0	87.3
en→sr	amazon	13.2	19.9	66.9	10.0	15.6	74.4
	bing	17.9	19.7	62.4	17.3	14.4	68.3
	google	9.7	18.4	71.9	10.4	13.7	75.9
	dcu-gen	7.6	10.8	81.6	4.9	9.4	85.7
	dcu-rev	8.1	9.3	82.6	4.6	8.9	86.5
entire evaluated text		10.2	13.9	75.9	8.7	12.2	79.1

Table 2: Percentages of words highlighted as errors regarding comprehensibility (left) and adequacy (right) for each target language and MT system, as well as for the entire evaluated text.

IAA (%)	$\alpha \uparrow$	F $\uparrow$	ed. dist. $\downarrow$
comprehension	0.51	79.1	25.8
adequacy	0.61	83.2	21.4

Table 3: Inter-annotator agreement (IAA) for comprehensibility and adequacy: Krippendorf's α, F-score and normalised edit distance.

It can be noted that IAA is higher for adequacy than for comprehensibility. The probable reason is that adequacy is guided by the original source text while comprehensibility is more subjective.

4 Analysis of marked issues

After the annotation of issues was done, we analysed the nature of the annotated issues. We did not perform a classical error classification but tried to identify linguistically motivated phenomena related to the annotated errors. We did not have any pre-defined error/phenomena scheme, but we started by looking into data and identifying the issue types on fly. The general two guiding criteria for assigning an issue type to a (group of) word(s) were :

- at least one evaluator marked the words as errors
- it is possible to define a type/cause for these errors

The very first step towards this analysis was carried out on a portion of the annotated data including both target languages and all three on-line systems. The main goal was to estimate how many major adequacy errors were not perceived as errors during annotation of comprehensibility issues. We found out that there are about 30% of such errors, however we did not found any issue types specific to this type of "masked" adequacy errors. The details can be found in the paper published at CoNLL conference (Popović, 2020b).

Results

Since the goal of the project was not to compare different MT systems nor to identify their strong and weak points, the results of our analysis are presented for all analysed MT outputs together in Table 4.

The numbers in the first column should be interpreted as follows: from all major comprehension errors, 17.32% are due to ambiguous source word, 1.12% are due to incorrect verb aspect, 1.76% are due to incorrect case, etc. The other columns are to be interpreted in the same way (second column: "from all minor comprehension errors", etc.)

Issue types which contribute to at least 2% of errors are presented in bold. These issue types will be discussed in the next section.

	comprehension		adequacy	
issue type	major	minor	major	minor
ambiguity	**15.48**	**7.73**	**15.84**	**7.75**
aspect	0.60	1.27	0.58	1.27
case	1.06	**3.51**	1.08	**3.48**
conjunction	0.74	0.76	0.75	0.75
determiner	0.05	0.07	0.05	0.07
extra word	0.51	0.46	0.50	0.46
gender	0.86	**4.20**	0.88	**4.23**
hallucination	0.08	0	0.08	0
”-ing” word	1.30	1.32	1.31	1.32
mistranslation	**7.87**	1.70	**8.07**	1.64
mood	0.37	0.65	0.38	0.65
named entity	**4.22**	**4.94**	**4.18**	**4.90**
negation	1.49	1.18	1.50	1.17
non-existing word	**2.68**	0.96	**2.66**	0.97
noun phrase	**9.17**	**6.75**	**9.17**	**6.61**
number	0.17	0.84	0.17	0.82
omission	**3.49**	**2.38**	**3.71**	**2.39**
order	0.34	1.13	0.34	1.10
passive	0.44	0.71	0.43	0.76
person	1.36	**2.32**	1.48	**2.30**
POS ambiguity	1.26	0.54	1.26	0.55
preposition	1.14	0.90	1.20	0.89
pronoun	1.10	1.87	1.12	1.86
repetition	0.25	0.23	0.25	0.23
rephrasing	**19.34**	**16.18**	**18.98**	**16.71**
source error	**2.38**	0.52	**2.48**	0.50
tense	0.09	0.44	0.09	0.43
untranslated	**3.89**	0.63	**4.08**	0.55

Table 4: Percentages of issue types perceived as major and minor errors for comprehensibility and adequacy.

Discussion

The most prominent issue types (in alphabetical order) are:

- **ambiguity**

  The obtained translation for the given word is in principle correct, but not in the given context (word sense error).

  Ranked as the second frequent reason both for major and for minor issues, although much more frequently perceived as a major issue.

- **case**

  Morphological (inflectional) form of a word denoting incorrect case.

  Mostly perceived as a minor issue.

- **gender**

  Morphological (inflectional) form of a word denoting incorrect gender.

  Mostly perceived as a minor issue.

- **mistranslation**

  The generated translation for the given word/phrase is incorrect.

  Mostly perceived as a major issue.

- **named entity**

  A named entity in the target language is incorrect for some of the following reasons (or a combination): 1) incorrectly translated 2) incorrectly transcribed 3) unnecessarily translated 5) incorrect case, gender, number 4) there are inconsistencies regarding original version and transcription.

  Triggers both major and minor issues.

- **non-existing word**

  A word in translation which exists neither in the source nor in the target language.

  Mostly perceived as a major issue.

- **noun phrase**

  An English sequence consisting of a head noun and additional nouns and adjectives is incorrectly translated.

  This type of issue is relevant for any target language which does not have the same formation rules for noun phrases as English.

  Ranked as the third frequent reason both for major and for minor issues.

- **omission**

  Something is missing in the translated text.

  Perceived both as minor and major issue.

- **person (subject-verb agreement)**

  A verb inflection in the translation denoting person does not correspond to the subject.

- **rephrasing**

  A sequence of source words is not translated properly for some of the following reasons (or their combination): 1) the translation choice of each word looks random, both lexically and morphologically, without taking any context into account 2) rephrasing is needed in the target language but the translation follows the structure of the source language 2) incorrect rephrasing in the target language.

  This type of issue usually comprises several consecutive different but interwined types of errors, such as morphological (case, gender, person/tense/mood/aspect, etc.), lexical (ambiguity, mistranslation, multi-word expression), word order, etc.

  Ranked as the most frequent reason for both major and minor issues.

- **error in source**

  A word in the original text in the source language has spelling or grammar errors which result in incorrect translation. This type of issue is especially relevant for user-generated content.

  Mostly perceived as a major issue.

- **untranslated**

  A word in the source language is simply copied to the translated text.

  Mostly perceived as a major issue.

References

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

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

Popović, M., Poncelas, A., Brkić, M., and Way, A. (2020). Neural machine translation for translating into Croatian and Serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020)*, pages 102–113, Barcelona, Spain (Online).