QRev: Machine Translation of User Reviews: What Influences the Translation Quality?

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

Abstract

This project aims to identify the important aspects of translation quality of user reviews which will represent a starting point for developing better automatic MT metrics and challenge test sets, and will be also helpful for developing MT systems for this genre. We work on two types of reviews: Amazon products and IMDb movies, written in English and translated into two closely related target languages, Croatian and Serbian.

1 Description

Data sets used for MT research include mainly "formal written text" (such as news) and "formal speech" (such as TED talks). Recently, there has been an increase of interest in the translation of "informal written text" which focuses on very noisy texts originating from sources like WhatsApp, Twitter and Reddit. On the other hand, other types of "informal written text" such as user reviews have not been investigated thoroughly, although they are important both from commercial and from a user perspective – user reviews of products have become an important feature that many customers expect to find.

This project focusses on user reviews in order to investigate which new challenges this "mid-way" kind of text poses for current MT systems. The main goal is to identify important quality aspects for MT of user reviews which will enable:

- development of appropriate automatic evaluation metrics;
- design of test suites specialised for important factors;
- definition of directions for improving MT systems.

Although the focus of the project are user reviews translated into Serbian and Croatian (as a case involving mid-size less-resourced morphologically rich European languages), the proposed evaluation strategy is completely genre/domain/language independent, so it can be applied to any genre, domain and language pair.

2 Data sets

We are working with two types of publicly available user reviews:

- IMDb movie reviews[1]
- Amazon product reviews[2]

3 MT systems

The main goal of the project is to find the common aspects important for the translation quality, and not to evaluate or compare particular MT systems. We are currently analysing MT outputs[3] of three on-line systems: Google Translate[4], Bing[5] and Amazon translate[6]. We are also developing our own system using publicly available data, which will be analysed in the later stages of the project.

[1] https://ai.stanford.edu/~amaas/data/sentiment/
[2] http://jmcauley.ucsd.edu/data/amazon/
[3] generated at the end of January 2020
[4] https://translate.google.com/
[5] https://www.bing.com/translator
[6] https://aws.amazon.com/translate/

4 Evaluation procedure

Our evaluation procedure is based on comprehensibility and fidelity (adequacy) (Roturier and Bensadoun, 2011), and it is being carried out on the review level (not on the sentence level). It should be noted that comprehensibility is not fluency – a fluent text can be incomprehensible, and vice versa. The novelty of our procedure is asking the annotators to concentrate on problematic parts of the text and to mark them, without assigning any scores or classifying errors. The procedure can also be guided by other evaluation criteria, not only comprehensibility and adequacy. The annotators were computational linguistics students and researchers fluent in the source language and native speakers of the target language. The annotation consisted of two independent subsequent tasks with the following guidelines:

Comprehensibility A monolingual task without access to the original source language text. Which parts of the translated review are not understandable? Distinguish two levels: "completely incomprehensible" and "not fully clear due to grammatic or stylistic errors".

Fidelity (Adequacy) A bilingual task with access to the original source language text. Which parts of the translated review do not correspond to the meaning of the original? Distinguish two levels: "the meaning of the original text is changed" and "not an optimal translation choice". If there are any problems in the source language, mark it, too (spelling or other errors, incomprehensible, unfinished, etc.).

The annotation started on 2 February 2020 and finished in April 2020. The annotated texts will be further analysed in order to identify common mistakes and linguistic phenomena which have the largest influence on comprehensibility and adequacy. The main aim of the analysis is to find the most important patterns and aspects which then can serve as a basis for automatic metrics, test suites, as well as for system improvements. In addition, the analysis will show in which way and to which extent particular phenomena contribute to comprehensibility and adequacy.

In total, 28 IMDb and 122 Amazon reviews (16807 untokenised English source words) are covered in this evaluation. However, not all generated MT hypotheses (6 for each review) are included. Each of the 270 included hypotheses is annotated by two annotators. The annotated data sets will be publicly released under the Creative Commons CC-BY licence.

5 First results: inter-annotator agreement and percentage of issues

Inter-annotator agreement (IAA) is shown in Table 1 in the form of F-score and normalised edit distance (WER).

IAA (%)	C	F (A)
F-score ↑	85.5	86.6
WER ↓	27.2	23.9

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

Percentage of words with issues for the two target languages is shown in Table 2.

% of issues		C	F (A)
hr	major	9.0	8.0
	minor	12.3	12.5
sr	major	13.1	12.1
	minor	19.4	14.4

Table 2: Percentages of words problematic for comprehensibility (C) and fidelity/adequacy (F (A)).

Acknowledgments

This research is being conducted with the financial support of the European Association for Machine Translation 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

Lohar, Pintu, Maja Popović, and Andy Way. 2019. Building English-to-Serbian Machine Translation System for IMDb Movie Reviews. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, Florence, Italy, August.

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