On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship

Joke Daems, Lieve Macken, Sonia Vandepitte
Department of Translation, Interpreting and Communication
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
E-mail: joke.daems@ugent.be, lieve.macken@ugent.be, sonia.vandepitte@ugent.be

Abstract

In order to improve the symbiosis between machine translation (MT) system and post-editor, it is not enough to know that the output of one system is better than the output of another system. A fine-grained error analysis is needed to provide information on the type and location of errors occurring in MT and the corresponding errors occurring after post-editing (PE). This article reports on a fine-grained translation quality assessment approach which was applied to machine translated-texts and the post-edited versions of these texts, made by student post-editors. By linking each error to the corresponding source text-passage, it is possible to identify passages that were problematic in MT, but not after PE, or passages that were problematic even after PE. This method provides rich data on the origin and impact of errors, which can be used to improve post-editor training as well as machine translation systems. We present the results of a pilot experiment on the post-editing of newspaper articles and highlight the advantages of our approach.

Keywords: Post-editing, Machine Translation, Translation Quality Assessment

1. Introduction

Evaluation of machine translation (MT) output is fundamental to the efficient improvement of MT systems. While automatic evaluation metrics such as the widely used BLEU (Papineni et al., 2002) can be used to compare the overall quality of different systems, a more detailed error analysis is necessary to identify specific strengths and weaknesses (Berka et al., 2012; Stymne & Ahrenberg, 2012). Knowledge of the types of errors that an MT system makes has even been said to reduce post-editing time (Martínez, 2003).

As post-editing MT is an important step towards high quality translations, we analyzed both MT errors and errors after post-editing (PE errors) and their relationship in order to determine the weaknesses and strengths of MT followed by PE. These insights can help define necessary improvements of MT systems to make the PE task easier as well as suggestions for the improvement of post-editor training. We are mainly interested in answering the following questions: What (and how many) MT errors are solved by post-editors, what (and how many) problems occur in post-edited MT and which (and how many) of these originate from MT?

The study presented in this paper is a pilot study of the ROBOT-project[1], whose design includes an analysis of the differences between human translation and post-editing for general text types. In the following sections, we will first discuss how quality is assessed within the ROBOT-project, followed by how we applied the method to MT output, and we will finish with a comparison of MT errors and student PE errors and their relationship.

2. Assessing Quality

Within the ROBOT-project, we designed a two-step translation quality assessment (TQA) approach. While borrowing some error categories from existing metrics (LISA, 2006; SAE-J2540, 2001), our categorization is more fine-grained and divided into acceptability and adequacy errors. Acceptability errors take the target language and the target text as a whole into account. Adequacy errors concern the relationship between source text and target text. Our approach contains categories that other metrics lack, such as terminological issues, coherence issues and text type-specific issues. An overview of the categorization can be found in Tables 1 and 2. Each category receives an error weight from 0 to 4, based on the text type and the impact the error would have on readability and accuracy of information. For example, contradictions receive a weight of 4, but capitalization errors receive a weight of 1 since capitalization problems hardly affect readability. A more detailed overview of the guidelines and categorizations can be found in (Daems & Macken, 2013). Depending on the text type and the goal of the assessment, the error weights can be changed to suit evaluation needs. For example, when working with technical texts, where terms are crucial, terminology issues and cases of hyperonymy or hyponymy would receive higher error weights.

Though the approach was specifically designed for the analysis of English and Dutch translations, the structure of the categorization is generic enough to allow for addition of language-specific evaluation needs. Languages requiring cases, such as Russian, can be analyzed by adding a subcategory 'incorrect case' to the category of grammar and syntax. Depending on the goal of the assessment, more specific (sub)categories can be defined, such as incorrect pronoun suffixes for imperative verbs in Italian.

The approach was tested on human translations and

[1] http://www.lt3.ugent.be/en/projects/robot/

post-edited MT from English into Dutch made by 16 master's students of translation (Daems et al., 2013). The corpus consisted of four different newspaper articles of approximately 250 words each. Students were told to provide translations of publishable quality, for an audience comparable to that of the source text. There were no time constraints for the translation or post-editing tasks.

In the first annotation step, two evaluators received only the target texts, and they annotated the products for acceptability. In the second step, the evaluators compared the source sentences with the translated sentences and they annotated the products for adequacy.

Following the suggestion by Stymne and Ahrenberg (2012) that inter-annotator agreement could benefit from joint discussion of examples and detailed guidelines, we measured agreement before and after a consolidation phase during which the evaluators discussed each other's annotations. Inter-annotator agreement did indeed benefit from such a consolidation phase (from 39% with κ=0.32 to 67% with κ=0.65 for acceptability, and from 42% with κ=0.31 to 82% with κ=0.79) for adequacy.

Having established the validity of the approach, we then compared the most common errors for human translation with those in post-edited MT.

Grammar & syntax	**Lexicon**	**Spelling & typos**	**Style & register**	**Coherence**
Article	Wrong preposition	Capitalization	Register	Conjunction
Comparative/ superlative	Wrong collocation	spelling mistake	Untranslated	Missing info
Singular/plural	Word non-existent	Compound	Repetition	Logical problem
Verb form		Punctuation	Disfluent structure/sentence	Paragraph
Article-noun agreement		Typo	Short sentences	Inconsistency
Noun-adjective agreement			Long sentence	Other
Subject-verb agreement			Text type	
Reference			Other	
Missing constituent/preposition				
Superfluous word/ constituent				
Word order				
Structure				
Other				

Table 1: Overview of the acceptability error categories and their subcategories.

Adequacy		
Contradiction	Quantity	Addition
Word sense disambiguation	Time	Explicitation
Hyponymy	Meaning shift caused by punctuation	Coherence
Hyperonymy	Meaning shift caused by misplaced word	Inconsistent terminology
Terminology	Deletion	Other meaning shift

Table 2: Overview of the adequacy error subcategories.

3. Adding MT to the Analysis

To better understand the cause of post-editing errors, a detailed analysis of the MT errors is needed. We applied the two-step TQA approach to the MT output that was used in the post-editing experiment described above. Google Translate[2] was used as the MT system, and the corpus consisted of four newspaper articles, selected from the Dutch Parallel Corpus (Macken et al., 2011).

Since it was the second time we applied the approach, inter-annotator agreement was higher for the MT annotations: 53% with κ=0.49 for acceptability and 57% with κ=0.46 for adequacy before consolidation and 84% with κ=0.83 for acceptability and 94% with κ=0.92 for adequacy after consolidation.

The MT annotations allowed us to compare the MT error score per word with the error score after student PE per word for each text, averaged over the number of post-editors, as can be seen in Figure 1.

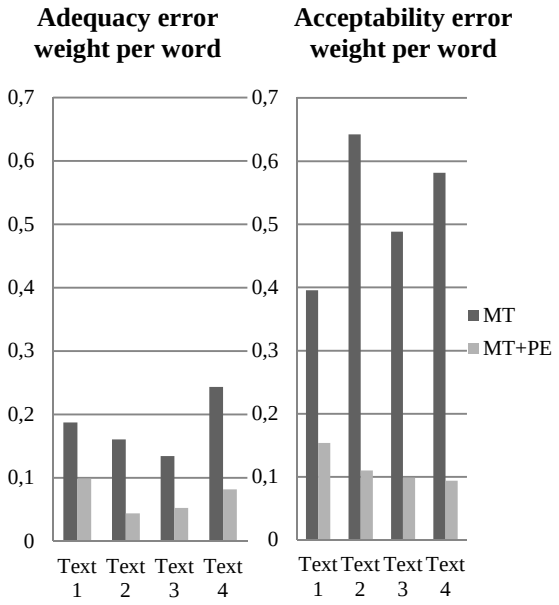

Figure 1: average MT error scores and scores after PE per text per word for adequacy and acceptability.

[2] translate.google.com

What can be derived from these graphs is that student PE indeed leads to a serious increase in quality, compared to the initial MT quality, both for adequacy and acceptability errors. The difference is greatest for acceptability problems, with error reductions of up to 83%.

The fine-grained analysis also allows us to look at the main error categories present in MT and the final post-edited product. The most common MT errors are displayed in percentages in Figure 2. Only the errors that account for at least 5% of all MT errors are shown. The most common errors after student post-editing can be seen in Figure 3.

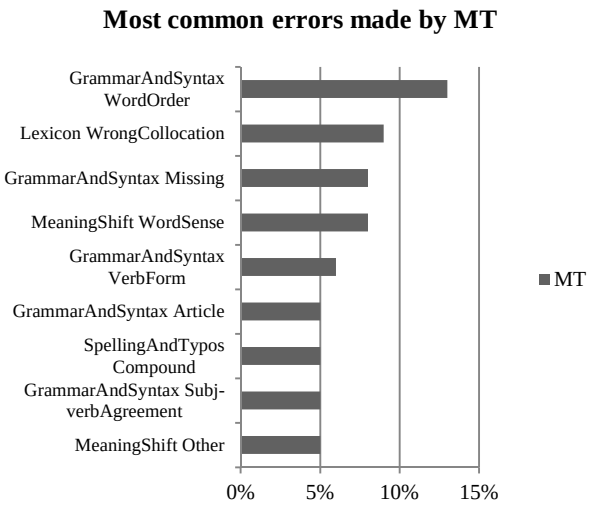

Figure 2: MT errors accounting for at least 5% of all MT errors made.

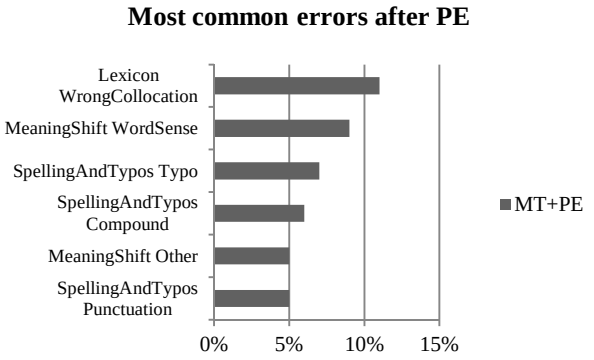

Figure 3: PE errors accounting for at least 5% of all PE errors made.

MT seems to suffer most from grammatical errors, with five of the nine most common errors belonging to the category of grammar and syntax. Problematic as well are cases of wrong collocation or word sense disambiguation errors.
After students' post-editing, grammatical issues seem to be far less problematic. Post-edited texts still suffer from wrong collocations and word sense errors, and contain many spelling errors such as the misspelling of compound nouns or punctuation errors.

4. The Origin of Errors

Though our two-step TQA approach provided us with interesting information on the type of errors found in MT and PE, what it did not tell us, is whether or not there is a relationship between the MT and students' PE errors. How many MT errors are corrected during post-editing? What type of MT errors are corrected? How many of the PE errors were caused by MT? To analyze this relationship, we manually linked the MT and PE errors to the corresponding source text passages and grouped the errors that originate from the same or similar source text passage into source text related error sets.

The following example illustrates the idea of error sets:

ST: ...the report (...) appeared on a celebrity website...
MT: ...het rapport (...) verscheen op een beroemdheid website... (compound) "*The report appeared on a celebritywebsite...*"
PE1: ...dat rapport verschenen (...) op een website voor beroemdheden... (other type of meaning shift) "*...appeared on a website for celebrities...*"
PE2: ...verschenen (...) op een website van een celibrity... (other type of meaning shift + spelling mistake) "*...appeared on the website of a celibrity...*"

The word 'celebrity website' was misspelled by Google Translate (English two-word compounds are usually one-word compounds in Dutch), and two student post-editors introduced an incorrect meaning shift. One of the post-editors also made a spelling mistake. Though the translations and type of errors are different, the errors are related to the same ST passage.

4.1 Corpus

The corpus consists of the four newspaper articles used in the post-editing experiment. Text 1 was post-edited by 3 students, text 2 was post-edited by 8 students, text 3 was post-edited by 7 students and text 4 by 4 students.

4.2 Analysis

Based on the error sets, we identified issues that were only found in MT, issues that were only found in PE, and issues that were found in both. Of the 103 ST-passages found to be problematic after MT, 61 were still problematic after students' post-editing. Of the 107 ST-passages problematic for student post-editors, 46 were not problematic for MT, and the errors were thus introduced during the post-editing process. To be able to count the types of problems that occurred both in MT and PE and the types of problems that only occurred in either MT or PE, we added a normalized weight to each error category: each translation method (MT or PE) received an equal weight (1) which was proportionally divided over all translators that could have made the error, and the error categories of the actual errors made. An illustration of this principle, based on the abovementioned example, can be seen in Figure 4. The total for the MT errors for the ST-passage under scrutiny would be 1 for misspelling of

compounds, whereas the PE normalized weight (divided over the four post-editors that could have made errors on this ST-passage) would result in a weight of 0.375 for other meaning shifts and 0.125 for spelling mistakes.

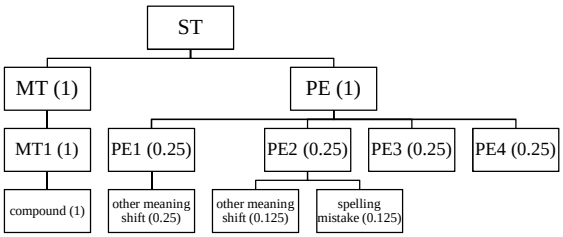

Figure 4: Example of the quantification within a source text-related error set.

4.3 Results

The normalized weights allow for a more accurate representation of the actual impact of an error. Figure 5 focuses on the ten most common MT errors. The average total normalized weight for all MT errors is represented by the bars' total length. The lower part of the bar (the bar minus the top section) represents the MT errors that occurred in ST-passages that were problematic for at least one student post-editor after post-editing. The actual impact of the errors on PE is then represented by the lowest part of the bar: the average total normalized weight for all PE errors found in the subset, reflecting the number of student post-editors that failed to solve the MT error.

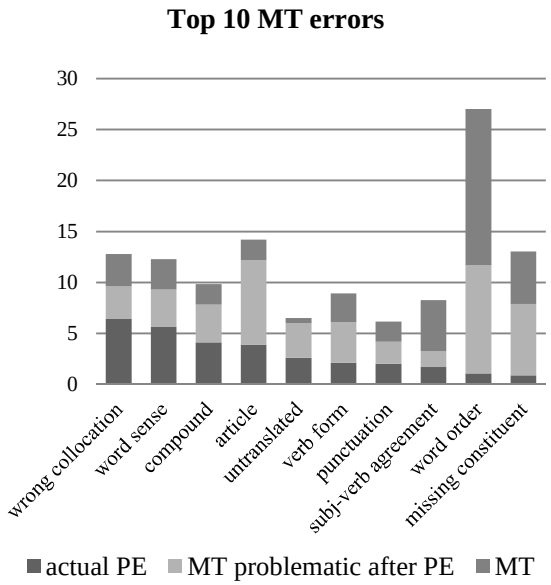

Figure 5: Overview of 10 most common MT error categories, proportion of the errors problematic for at least one student post-editor and the actual impact on PE. Values expressed in total normalized weight. Categories sorted from highest to lowest actual impact on PE.

It is striking that five out of ten most common MT errors are grammatical errors (superfluous or missing articles, incorrect verb forms, agreement issues, word order problems, missing constituents), with word order issues being the most common MT error. Yet these errors do not seem to be the errors that are the most problematic for post-editors. Though most MT errors are problematic for at least one post-editor (with the exception of subject-verb agreement issues and word order problems), the most problematic categories still present after PE are wrong collocations, word sense errors, and the spelling of compounds.

The large number of syntactic errors demonstrates that SMT systems could greatly benefit from some kind of rule-based post-processing step, an idea that has been proven successful for English-Czech translations (Mareček et al., 2011). This would allow the student post-editors to focus more on lexical issues (collocations) and adequacy errors, such as word sense disambiguation.

Figure 6 reveals the origin of the most common PE errors. Most student PE errors seem to be, in fact, caused by MT errors. Word sense errors, wrong collocations, misspelled compounds, incorrect verb forms, incorrect or missing articles and misplaced words barely ever occur without a corresponding error in MT.

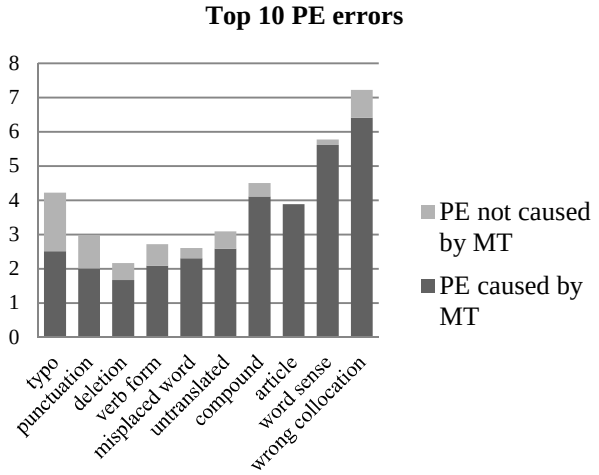

Figure 6: Overview of 10 most common PE error categories and their origin in MT. Values expressed in total normalized weight. Categories sorted from smallest to largest difference between normalized weight of both origin types.

Combining the information from Figures 5 and 6, it seems that student post-editors would especially benefit from some kind of training to spot wrong collocations, word sense errors and missing or superfluous articles in English-Dutch translations, since these are three of the five most common MT and PE errors that are mainly caused by errors in the MT output. A good post-editing environment should also contain a spell-checker, to reduce the large number of typos and spelling errors such as compound nouns. Our data for the post-editing experiment was gathered with PET (Aziz et al., 2012), which does not contain a spell-checker, and this probably accounts for the many typos found.

On the MT end, it could be interesting to integrate the error set information into MT confidence information. A

good post-editing tool would perhaps benefit from warnings whenever certain awkward collocations or polysemous words could occur.

5. Conclusions and Future Work

We extended the fine-grained two-step TQA approach described in Daems et al. (2013) to include annotations for MT errors. These annotations were used to identify the main problems for MT and subsequent PE by student translators. In a second step, the annotations were grouped into ST-related error sets to identify the relationship between MT and student PE errors.

We believe that the method here presented can lead to a better understanding of the relationship between MT quality and post-editing. Knowledge of the type of errors that are most easily corrected by post-editors and the type of errors that are mainly caused by MT output may contribute to the improvement of the PE process. This can be done both by improving the MT system and by focussing on typical MT errors during post-editor training. We are convinced that the annotated MT can be used for other research purposes as well, for example, a more fine-grained error annotation might be useful for the development of automatic quality assessment systems, which currently mainly focus on the sentence level (Specia & Soricut, 2013).

The only drawback of the approach in its current state is the fact that it is highly time-consuming (45 minutes annotation time on average per 150 words for new texts), which is why we restricted ourselves to two annotators only. Speed is higher for adequacy than for acceptability, and increases if the text is already familiar. We are currently trying to optimize the annotation method to reduce effort, and we will try to automate (part of) the identification process of source text-related error sets in order to further reduce the amount of manual effort and increase total process speed.

Currently, we can only draw conclusions for post-editing by student translators, but future experiments will include the same analysis of texts post-edited by professional translators.

6. References

Aziz, W., Sousa, S., & Specia, L. (2012). PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*, 3982-3987.

Berka, J., Bojar, O., Fishel, M., Popovic, M., & Zeman, D. (2012). Automatic MT error analysis: Hjerson helping Addicter. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*, 2158-2163.

Daems, J., & Macken, L. (2013). Annotation guidelines for English-Dutch translation quality assessment, version 1.1 *LT3 Technical Report*.

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice*, 63-71.

LISA. (2006). LISA QA Model 3.1.: Localization Industry Standards Association.

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: a balanced copyright-cleared parallel corpus. *Meta, 56*(2), 374-390.

Mareček, D., Rosa, R., Galuščáková, P., & Bojar, O. (2011). Two-step translation with grammatical post-processing. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, 426-432.

Martínez, L. G. (2003). *Human translation versus machine translation and full post-editing of raw machine translation output.* Dublin City University, Dublin, Ireland.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-j. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318.

SAE-J2540. (2001) *Quality Metric for Language Translation.*: Society of Automotive Engineers.

Specia, L., & Soricut, R. (2013). Quality estimation for machine translation: preface. *Machine Translation, 27*(3-4), 167-170.

Stymne, S., & Ahrenberg, L. (2012). On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*, 1785-1790.