Using error annotation to evaluate machine translation and human post-editing in a business environment

Lucia Comparin
Universidade de Lisboa
Centro de Linguística
da Universidade de Lisboa
Unbabel, Lisboa, Portugal
lcompa@gmail.com

Sara Mendes
Universidade de Lisboa
Centro de Linguística
da Universidade de Lisboa
Faculdade de Letras da Universidade de Lisboa
s.mendes@campus.ul.pt

Abstract

Quality Assessment currently plays a key role in the field of Machine Translation (MT) and in the organization of the translation market. Besides allowing to rank the players providing MT services, accurately assessing the quality of translation results is also a valuable step to improve the performance of automatic systems. In this study, we present the results of a study involving an error annotation task of a machine translated corpus from English into Italian. The data obtained allowed us to identify frequent and critical errors, and to observe their prevalence at different stages of the translation process, a most valuable analysis to outline strategies to automatically detect and correct the most relevant and prevalent errors in MT results. Accomplishing this is a crucial future step towards being able to guarantee the quality of results and a cost-effective workflow to obtain them.

1 Introduction

Research in machine translation (henceforth MT) has increased in the last decades, and MT systems have been increasingly integrated as part of the workflow adopted by translation providers in the market. Despite the development and improvements in MT systems and the continuous research done in the field, the quality of the results is still variable and dependent on many aspects such as the MT system used and the type of texts translated. This makes post-editing a necessary step when MT is part of the translation process adopted by a company. At the same time, the variability of results highlights the importance of evaluating the performance of MT systems. Error annotation, i.e. the identification and categorization of errors present in a text, is used to assess the results of a MT system in terms of quality. Assessing quality of machine translated texts through error annotation is useful not only to evaluate the quality of the results produced by a MT system, but also to outline strategies to improve them and reduce the number of errors in the output produced. Such strategies can lead to the definition of specifications to implement in the system, or rules to automatically correct errors in the post-editing stage.

In the work presented in this paper, we performed error annotation of machine translated texts in order to provide data for improving translation results. The study was carried out within Unbabel, a startup company that offers almost real-time translation services by combining MT and human post-editing. Taking into account that Unbabel's translation workflow involves human post-edition, being able to identify and characterize the errors human editors are confronted with, and to which extent they persist after a first edition is crucial to outline strategies that aim at improving translation results in a cost-effective way.

2 Related Work

Due to the increasing adoption of MT systems in the translation process and to the development of different MT systems, quality assessment and the evaluation of MT systems have become an important field of research.

Quality assessment can be either performed by humans or automatic systems. Typically, in the former case, a human annotator identifies errors in

translation results, categorizes them and provides an analysis for them as described in Daems et al. (2014) and Stymne and Ahrenberg (2012). The latter, on the contrary, are based on the comparison of MT results with a human translation that is considered a high-quality reference. The most widely used systems are BLEU (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009).

Research done in the analysis and description of MT errors is extensive and mostly related to the annotation and analysis of all errors present in texts that were translated using a particular MT system, in order to improve its performance (e.g. Kirchhoff and Yang (2007) and Vilar et al. (2006)). The classification of errors is usually based on error taxonomies such as those presented in Vilar et al. (2006), and Popovic and Burchardt (2011). As annotation can be used to assess the quality of a translation for different purposes and in different contexts, error taxonomies are adapted to the purpose of the research. When they are used to assess the service provided by a company, they can be customized as described in the framework presented by Lommel (2015) under the scope of the Quality Translation 21 project. In Costa et al. (2015) an error taxonomy is presented to classify translation errors from English into European Portuguese, and a linguistic motivation for the selection of categories is provided. While these studies contemplate the description and categorization of errors, Hermjakob et al. (2008) concentrated on error detection, studying named entity translation errors, and improving an on-the-fly NE transliterator that is integrated into a statistical machine translation system.

3 Methodology

In the study presented here, we considered the language pair English-Italian and performed human annotation of a corpus. The corpus consisted of text provided by Unbabel clients and included web content such as travel descriptions and Customer Service emails, which were translated from English into Italian using the Google Translator API. In the translation process adopted by Unbabel, texts are firstly translated by the MT system, and then edited online by a community of human translators. Depending on the content of the text and on its length, one or more post-editions of the same text is performed. The corpus considered in this work included texts of 100 to 700 words. The motivation for using texts with this length lies on the fact that, in order for the annotation to be accurate, texts have to be long enough for the annotator to understand the content, but short enough so that the task is not too time-consuming and demanding.

In this work, we annotated the texts of the corpus both immediately after MT and after the first human post-edition. This allows us to calculate the amount of errors that are corrected in post-editing and to figure out which errors generated by MT systems go on unnoticed at the following stage. The information resulting from the type of work described in this paper can therefore be used to outline strategies to improve post-editing and guarantee high-quality translations (Comparin, 2016; Comparin and Mendes, 2017).

In order to perform the annotation of the corpus, we considered the error taxonomy used at Unbabel. Work already done in the area and the analysis of each category were the starting point to better define the task perfomed in this study and its specifications. Data collected in the annotation of machine translated texts and in that of edited texts were then compared and analyzed, setting the grounds for the design of strategies to address the issues in the post-editing stage, as proposed in Comparin and Mendes (2017).

4 Error Annotation

The documents and guidelines used as a basis in order to define the typology used in the annotation were the MQM framework (Lommel, 2015) and TAUS documents (www.taus.net). The former is a model developed in the Quality Translation 21 project, funded by the European Union Horizon 2020 research and innovation program, whose goal is to overcome language barriers to encourage flow of ideas, commerce and people within the EU. TAUS is a resource center offering support to translation service providers by making available different tools, such as software, metrics, and knowledge. A framework was developed in the scope of the QT21 project in order to define task-specific translation metrics, that help assessing the translation performed by a MT system or by a company.

The tool used in this study was created by Unbabel and used to assess the quality of the texts delivered to clients in different language pairs on a weekly basis. The tool shows the source text, the target text, the annotations, and the glossary terms.

When the annotator selects a word or a sequence of words in the target text, possible error types appear in a box and the relevant one can be selected. Additionally, the annotator can also assess the fluency of the entire text, using a scale of 0 to 5.

In order to design an error taxonomy suitable to an annotation task with the goals of the one discussed in this work, some prerequisites have to be considered. Taking into account the standards and the work already done in the annotation field not only to define the set of useful error types, but also to guarantee that annotation is accurate, first, all errors that can be generated in MT should be covered, but the number of error types should be limited, to avoid noise in data annotation and to make the annotation process affordable both in terms of time dedicated to the task and in terms of its cost. Secondly, error types should be clearly distinguished from one another.

4.1 Error Types and Penalty system

The 41 error types included in the taxonomy used at Unbabel and considered in this study are divided into 7 major categories: accuracy, fluency, style, terminology, wrong language variety, named entities, and formatting and encoding. Below we briefly define the aforementioned error categories considered in the typology.

ACCURACY: errors in this category concern the relationship between the source text and the target text and the extent to which the latter maintains the meaning and the information of the former

FLUENCY: errors in this category regard the quality of a text, assessing whether it is well-written and easy to read, and if it accomplishes its communication purpose in the target language

STYLE: issues concerning register and fluency

TERMINOLOGY: mistranslation of terminology

WRONG LANGUAGE VARIETY: use of a word or expression from a different language variety.

NAMED ENTITIES: wrong translation of proper nouns

FORMATTING AND ENCODING: issues concerning the segmentation of sentences and paragraphs

In addition to the categorization of errors, a penalty is also available to be associated to each error annotated. By doing this, a numerical quality score can be calculated by the tool for each translation, and can be used as an indicator of its quality and of the improvements still to be made. Additionally, such a score is used in the industry to position a company in the market. The penalty system was set up based on the system used at Google LQE (Localization Quality Evaluation) and in the MQM. The errors annotated were divided according to their severity into minor, major and critical errors, following the definitions below.

Minor: Errors that do not change nor compromise the information provided in the source text. They do not prevent the reader of the target text to understand it in a clear way and they do not generate confusion or doubts. They can nonetheless affect fluency. The penalty associated to minor errors is 0.5 points.

Major: Errors that make the target text either confusing or ambiguous. They make it more difficult for the reader to clearly understand the text, although the target text conveys the message. In some cases, the meaning of the target text can slightly change, however general comprehension is guaranteed. The penalty associated to this type of error is 1 point.

Critical: Critical errors change the meaning of the source text. Not only they prevent the reader from understanding the information provided in the text, but also they can cause damage to the reputation of a company and carry health, safety or legal implications. The penalty associated to this type of error is 3 points.

4.2 Some remarks on the annotation performed in this study

Before discussing the data obtained in the annotation task, we would like to discuss a few aspects related to annotation and make a few notes regarding the task in this particular case. Human annotation can be a challenging task, as it is related to the annotators understanding and categorization of an error. In this study each annotation was performed by a single annotator, which made the definition of clear guidelines to help in the task a necessary step. In those cases in which a single error simultaneously involved different error types, the type that provided more information about the phenomenon at stake was preferred. For instance, when a conjunction was omitted, the error category *conjunctions* was selected instead of *omission*.

Additionally, due to technical constraints regarding the platform used at Unbabel - the annotation tool used does not allow the association of more than one error type to the same expression - , when one word or sequence of words contained more than one error, only the most relevant one was marked. Since data collected from annotation, in this specific case, were used to improve translation results through the definition of a set of rules for automatic post-edition and/or automatic checking of machine translated results, errors types involving grammar phenomena were preferred, such as *agreement*, *tense/mood/aspect*, *word order*, *sentence structure*, *prepositions*, *conjunctions*, or *determiners*. If the purpose of the annotation were to study spelling mistakes in MT, then *orthography* errors would be selected as more relevant.

Since in this work we concentrated on errors after MT and after the first post-edition, a high number of errors, and particularly of critical errors, was observed in the target text. Given this, and even if a penalty was assigned to each error during annotation, we do not discuss this aspect here as the high number of errors and the great impact they have on translation quality does not allow for clear and insightful distinctions in terms of severity for a great part of the annotated errors.

The guidelines defined in the MQM framework (Burchardt and Lommel, 2014), which highlight the fact that the annotator should be as precise as possible both in the selection of the text containing the error, and in the selection of the error type, were taken into account in this study, as long as the specifications of the annotation tool used at Unbabel allowed the annotator to do so, which was not always the case, as mentioned above.

5 Annotation Data

The errors annotated both after MT and after the first post-edition are presented in the tables below. In Table 1, absolute and relative frequency of annotated errors per error category in the typology is presented.

The data in Table 1 show that the number of errors in machine translated texts is high and not evenly distributed among the different error categories. This is certainly related to the fact that it was not possible to mark two errors in the same word or sequence of words, and, in such cases, the error with the greatest impact on the quality of the translation and particularly on the access to the content of the text was marked, and thus the categories mentioned in section 4.2 were preferred.

With regard to the number of errors in the two stages considered in our study, MT and the first post-edition, there is an 85% error reduction between the two stages. However, the impact of human post-edition on error reduction is variable between different error categories: e.g while *fluency errors* lower their relative frequency from 77% to 49%, *accuracy errors* actually increase their relative frequency (the absolute number of errors decreases significantly in both cases, naturally: 90.2% for *fluency* and 76.7% for *accuracy*). The significant increase of the relative frequency of errors in error types more related to style and client specifications (e.g. *inconsistent register*, *repetitive style*, or *noncompliance with client's glossary and vocabulary*) is due to the fact that, in many cases in which an error belonging to these types occurred after MT, more severe errors were present in the translated texts, and were thus the ones marked. As the first post-edition tends to correct the most severe errors, those related to the creative use of language and style become in turn visible. Let us now consider the most frequently marked error categories in more detail, i.e. *accuracy errors* and *fluency errors*.

The error type with the highest number of errors annotated in machine translated texts is *determiners*, followed by *lexical selection*, *agreement*, *tense/mood/aspect*, and *word order*. Errors belonging to these error types, in the majority of the cases, do not allow the reader to understand the text clearly, and therefore have a major or critical impact on the quality of the translation. Two error types that have a lower number of errors but are still crucial for the quality of translation results are *sentence structure* and *prepositions*. Errors in sentence structure, in particular, have a great impact on translation, because they often result in a sentence that cannot be understood without knowledge of the source language and the sentence structures commonly used in it. Additionally, such errors require a major intervention of the editor, since the text has to be rewritten in the majority of the cases, which takes significantly more time than just changing a morpheme or a word. The time spent in the correction of errors involving prepositions is also considerable, because, when the wrong preposition is selected, the meaning of the text often cannot be fully and accurately

	MT		First edition	
Main error types	abs. freq.	rel. freq.	abs. freq.	rel. freq.
Accuracy errors	236	0.21	55	0.32
Fluency errors	848	0.77	83	0.49
Style errors	1	0	3	0.02
Terminology errors	0	0	14	0.08
Wrong language variety errors	0	0	0	0
Named entities errors	19	0.02	15	0.09
Formatting and encoding errors	0	0	0	0
Total	1104	1	170	1

Table 1: Absolute and relative frequency of annotated errors per error category after MT and first human edition

	MT		First edition	
Accuracy errors	abs. freq.	rel. freq.	abs. freq.	rel. freq.
Mistranslation				
Overly literal	9	0.01	4	0.02
False friend	0	0	0	0
Should not have been translated	18	0.02	3	0.02
Lexical selection	165	0.15	37	0.22
Omission	6	0.01	0	0
Untranslated	27	0.02	9	0.05
Addition	11	0.01	2	0.01
Total	236	0.21	55	0.32

Table 2: Absolute and relative frequency of accuracy errors after MT and first human edition

	MT		First edition	
Fluency errors	abs. freq.	rel. freq.	abs. freq.	rel. freq.
Inconsistency				
Word selection	1	0	1	0.01
Tense selection	0	0	0	0
Coherence	2	0	1	0.01
Duplication	0	0	0	0
Spelling				
Orthography	1	0	1	0.01
Capitalization	52	0.05	19	0.11
Diacritics	0	0	0	0
Typography				
Punctuation	9	0.01	4	0.02
Unpaired quote marks and brackets	1	0	0	0
Whitespace	17	0.02	5	0.03
Inconsistency in character use	0	0	0	0
Grammar Function words				
Prepositions	70	0.06	10	0.06
Conjunctions	12	0.01	1	0.01
Determiners	237	0.21	19	0.11
Word form				
Part-of-speech	30	0.03	1	0.01
Agreement	159	0.14	13	0.08
Tense/mood/aspect	101	0.09	3	0.02
Word order	106	0.10	4	0.02
Sentence structure	50	0.05	1	0.01
Total	848	0.77	83	0.49

Table 3: Absolute and relative frequency of fluency errors after MT and first human edition

understood just by considering the text produced by the MT system. Comparing these more frequent types of errors in the two stages of the translation process, we can identify two types of behavior: some of the most critical errors, such as *tense/mood/aspect*, *word order* and *sentence structure* are almost non-existent after the first human post-edition; on the other hand, errors that are in principle more straightforward to correct, such as *determiners* or *agreement* are visibly reduced, but their relative weight considering all the errors annotated after the first human post-edition is still

considerable. This observation is probably not independent from the fact that these are errors which are easier to be overseen by a human editor, as they often amount to a small variation in the form of the lexical items involved. Finally, some brief remarks regarding errors involving *prepositions* and *lexical selection*, which, respectively, show no reduction and an increase in their relative weight after the post-edition stage, when compared with what was the case after MT. These data make apparent that this type of error persists even after human edition, its weight in the overall amount of errors annotated remaining important by the crucial reduction of other types of error.

6 Results and final remarks

The error annotation presented in this work allowed us to analyze the most significant types of error occuring in machine translated texts from English into Italian using Google Translator API, and their prevalence after the first human post-edition. As expected, the comparison between the errors annotated at these two stages of the translation process is marked by a significant reduction in the absolute number of errors. This comparison also made apparent that there are certain types of error that continue to be present even after human edition. The amount of errors after MT and the prevalence of certain types of error make apparent the need for using the results and analysis of this annotation task to outline strategies to automatically tackle the shortcomings of MT systems and aid human post-edition, as we have proposed and evaluated in Comparin (2016) and Comparin and Mendes (2017).

References

Burchardt, Aljoscha and Arle Lommel. 2014. *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. URL http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf.

Comparin, Lucia. 2016. *Quality in Machine Translation and human post-editing: error annotation and specifications*. MA dissertation, Faculdade de Letras da Universidade de Lisboa, Portugal.

Comparin, Lucia and Sara Mendes. 2017. Error detection and error correction for improving quality in machine translation and human post-editing. *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2017*, Budapest, Hungary, 2017.

Costa, Angela, Wang Ling, Tiago Luis, Rui Correia, and Luisa Coheur. 2015. A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, 29(2): 127-161.

Daems, Joke, Lieve Macken, and Sonia Vandepitte. 2014. On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavik, Iceland, May 26-31, 2014, pages 62-66.

Hermjakob, Ulf, Kevin Knight, and Hal Daume III. 2008. Name Translation in Statistical Machine Translation - Learning When to Transliterate. *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, June 15-20, 2008, Columbus, Ohio, USA, pages 389-397.

Kirchhoff, Katrin and Mei Yang. 2007. The University of Washington machine translation system for the IWSLT 2007 competition. *2007 International Workshop on Spoken Language Translation, IWSLT 2007*, Trento, Italy, October 15-16, 2007, pages 89-94.

Lavie, Alon and Michael J. Denkowski. 2009. The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105-115.

Lommel, Arle. 2015. *Multidimensional Quality Metrics (MQM) Definition*. URL http://www.qt21.eu/mqm-definition/definition-2015-12-30.html.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6-12, 2002, Philadelphia, PA, USA, pages 311-318.

Popovic, Maja and Aljoscha Burchardt. 2011. Error Analysis of Machine Translation Output. *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 265-272, Genoa, Italy. European Association for Machine Translation.

Stymne, Sara and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey, May 23-25, 2012, pages 1785-1790.

Vilar, D., J. Xu, L. D'haro, and H. Ney. 2006. Error Analysis of Statistical Machine Translation Output. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).