A Corpus of Machine Translation Errors Extracted from Translation Students Exercises

Guillaume Wisniewski*, Natalie Kübler†, François Yvon*

*Université Paris Sud, LIMSI-CNRS
91 403 Orsay, France
guillaume.wisniewski@limsi.fr, francois.yvon@limsi.fr

†EILA/CRILAC, Université Denis Diderot
75 013 Paris, France
nkubler@eila.univ-paris-diderot.fr

Abstract

In this paper, we present a freely available corpus of automatic translations accompanied with post-edited versions, annotated with labels identifying the different kinds of errors made by the MT system. These data have been extracted from translation students exercises that have been corrected by a senior professor. This corpus can be useful for training quality estimation tools and for analyzing the types of errors made MT system.

Keywords: Translation Error Corpus, Post-Edition, Error Analysis

1. Introduction

The lack of automatic diagnostics tools that could help sort out and assess the impact of the various causes of errors is, today, a major bottleneck for the development of high-quality Machine Translation systems: for lack of such diagnoses, it is difficult to figure out which components of the system require the most urgent attention.

Several methods have recently been proposed to automatically detect Machine Translation errors (Zhou et al., 2008; Popović and Ney, 2011; Zeman et al., 2011; Bach et al., 2011) which rely on Machine Learning methologies. This means that the development and evaluation of these error detection techniques crucially depends on the availability of annotated corpora, containing MT outputs in which errors have been identified and labeled such as the one described by Fishel et al. (2012). Unfortunately, such resources are still rare, and collecting them is an expensive and error-prone task. To illustrate this fact, a recent attempt, made in the context of the QT Launchpad project,[1] only managed to collect and annotate a few hundreds examples. This is not enough to use Machine Learning approaches that may require to estimate several hundreds parameters. When analyzing the reasons of this (relative) failure, Burchardt et al. (2013) note that (emphasis ours):

> "Error analysis is considerably more time-consuming than anticipated. Rather than analyzing a few thousands of sentences in our pilot phase, we were able to have a few hundred analyzed. While spees would improve with training and experience, *detailed analysis is a labor-intensive task and large-scale annotation would require either many annotators* (raising problems of inter- annotator consistency) *or much time*."

Building on this experience, we adopt here another approach for collecting an error corpus that avoids these difficulties. Rather than building a corpus specifically for the task at hand, which would require the training of annotators who have no prior knowledge of MT error identification, we propose to take advantage of exercises made by students in Translation Studies, part of which consist precisely in identifying, labeling and discussing the errors contained in translations. All these exercises have been corrected by senior professors, which guarantees the quality of the data. This paper describes the construction of a corpus of post-edited translations extracted from apprentice translators exercises. These translations are annotated with the type of errors made by the MT system. The corpus is freely available from our website.[2] The rest of this article, is organized as follows: the corpus will be described in a first section. In a second section we will detail the different classes of errors that have been identified. We will conclude by presenting several ways in which the resource we are providing could be exploited.

2. Building the resource

2.1. Context

The corpus we have gathered has been extracted from the exercises of translation students, taking part in a master program in specialized translations.[3] These exercises consist in post-editing the translation of a technical document (be it a scientific article, a technical manual, an entry in an encyclopedia, etc.) produced by a rule- based machine translation system. All the documents are translated from English into French. A subset of the considered documents also contain a detailed analysis of the error made by the MT system. All the exercises have been corrected and annotated by a senior professor.

[1] http://www.qt21.eu/launchpad/

[2] http://perso.limsi.fr/Individu/wisniews/ressources

[3] Master *Industrie de la langue et traduction spécialisée* of the Université Paris Diderot.

Both the original student works and the professor commentaries are stored in Microsoft Word documents. These documents are organized in tables: each row in a table describes a sentence of a source document, its translation by a MT system, its post-edition by a student and information about the errors the MT translation contain. Comments by the professor are stored in *Word commentaries*. The rows appear in the same order as in the original document and, generally, correspond to a complete document or, at least, to a large portion of it. Figure 1 displays an example of such a document.
Using the Microsoft Word API, we have extracted all the data contained in the student exercises and stored them in an JSON document more amenable to automatic processing by standard NLP tools. In particular, the following informations have been extracted:

- the source document (special care was taken to keep the original document structure);
- its automatic translation by a rule-based MT system;
- the post-edited translation made by a student in Translation Studies;
- possibly an analysis of the errors of the automatic translations;
- the correction of the post-editions made by a professor.

All these information are aligned. In addition to this raw information, directly extracted from the Microsoft Word documents, we also provide a version of the source and target documents that have been tokenized and segmented in sentences using a simple rule-based method.
This corpus differs from most existing corpora in several ways. First, it contains complete documents, that have been post-edited 'in context', while most existing corpora are made of single sentences, the context of which is not known. As a direct consequence, some post-editions question sentence boundaries: sometimes, two source sentences are translated by a single sentence and sometimes the translation of a single source sentences is split over two target sentences. Second, the post-editions and the error annotations have all been validated by a senior professor in Translation Studies, which guarantees the quality of the data. Third, it is made of technical documents that are using a specialized vocabulary and contain many instances of terminology errors. Lastly, it is, to the best of our knowledge, larger than similar corpora like the one collected by Burchardt et al. (2013).

2.2. Statistics

The corpus presented in this work has been extracted from the work of 46 students. It is made of 4,854 source sentences containing 95,266 words and translated by 4,709 sentences containing 101,951 words (statistics have been computed on the post-edited version of the reference). Errors have been annotated for almost half of the sentences produced by the MT systems.
Sentence boundaries have been changed in less of 5% of the post-editions. The hTER score (Snover et al., 2006) of the system considered is pretty high (close to 40%), which can be expected, given the difficulty of the task: documents come from a technical domain and use a very specific terminology.

3. Typology of Errors

As explained in previous sections, approximately half the post-edited sentences of the corpus contain an additional annotation describing the errors that have been made by the MT system. Two kind of annotations are found.
The first kind of annotations are pretty coarse, as they rely on a simple typology of errors made of 6 different types:

1. lexical errors;
2. morphological errors;
3. syntax errors;
4. semantic errors;
5. format errors (e.g.: error caused by a problem in the tokenization of the source sentence);
6. errors without a clear explanation.

While this typology is not as detailed as the ones already proposed, for instance, by Vilar et al. (2006) or Bojar (2011) or the one used in the WMT'14 shared task on Quality Estimation[4], it still distinguishes the most useful kind of errors.
Besides these annotations, most errors are also analyzed at a fine-grain level. These analyses are more qualitative and given in a semi-structured format: the error is described in a free text field, but its description generally contains the name of the error identified, for instance, by its color or its font and also a possible explanation of the cause of the error. Figure 2 shows several examples of such annotations. Extracting these fine-grain error types is more difficult than for the coarse level description, and has been performed using the following semi-automatic process. In a first step, the error descriptions were normalized using standard preprocessing tools typically applied in texts classification: all stop-words were removed and the remaining words were stemmed. We then extracted the different combination of up to 4 contiguous words that appear in more than one description. These elements correspond, with high probability, to the name of the different errors that have been identified. In a second step, these 'candidates' are manually checked to filtered out non valid names and mapped to one of the 6 error classes of the typology presented above. The distribution of the different error classes is summarized in Table 1.

4. Conclusion

In this paper, we have presented a freely available corpus of translation errors, which contains post-edited translations annotated with labels identifying the different types of errors of the MT system. These data have been extracted from

[4] http://statmt.org/wmt14/quality-estimation-task.html

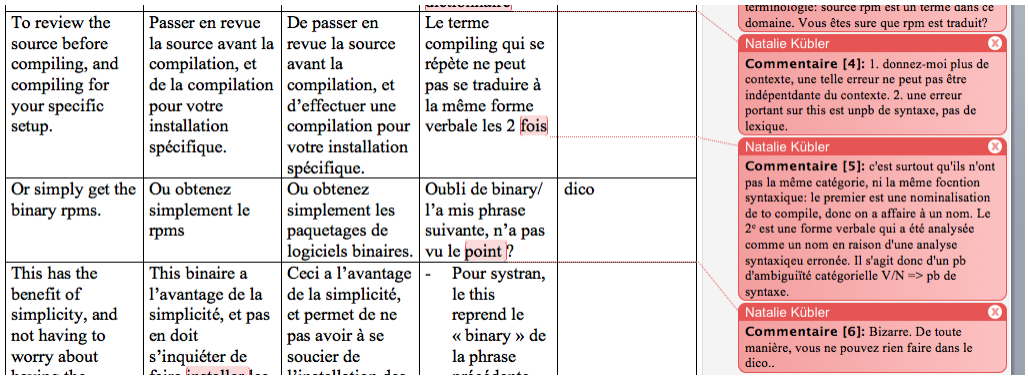
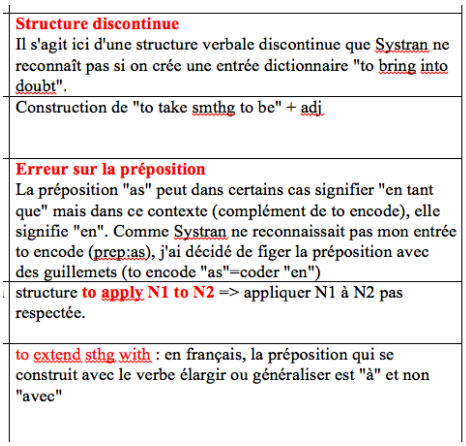


Figure 1: Example of an original Word document we have collected: the first column contains the source text, the second the automatic translation, the third the post-edition and the forth a description of the error. The work is annotated by the professor using the commentary feature provided by Microsoft Word.



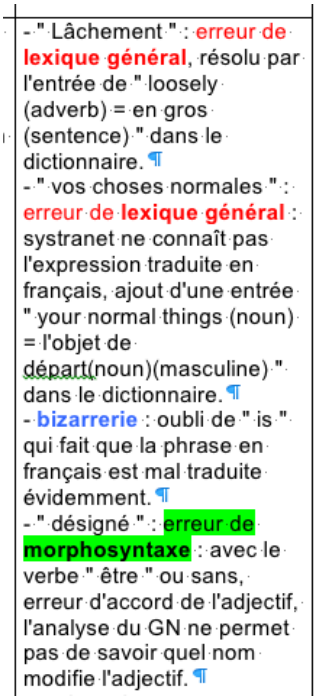


Figure 2: Examples of fine-grain analyses of MT errors

error type	proportion
lexical errors	22%
morphological errors	10%
syntax errors	41%
semantic errors	12%
format errors	5%
other	10%

Table 1: Distribution of error types in the corpus

translation students exercises and corrected by a senior professor.

This corpus can proved useful in several ways. It can be used, for instance, to train systems able to predict if a MT output contain an error, which is of great interest to develop Quality Estimation systems (Specia et al., 2010; Wisniewski et al., 2013). Another interesting question is whether it is possible to automatically identify different classes of errors and, if so, which features are the most effective to sort out the different class of errors. Our future work will tackle all these questions.

Acknowledgments

This work was partly supported by ANR project Transread (ANR-12-CORD-0015). Warm thanks to Andrien Cabaco for his help with the extraction of data from Word documents.

5. References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: a method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 211–219, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ondřej Bojar. 2011. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, March.

Aljoscha Burchardt, Arle Lommel, and Maja Popovic. 2013. Tq error corpus. Technical Report Deliverable D 1.2.1, QT Launchpad Project.

Mark Fishel, Ondřej Bojar, and Maja Popović. 2012. Terra: a collection of translation error-annotated corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Comput. Linguist.*, 37(4):657–688, December.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

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. European Language Resources Association (ELRA).

Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and analysis of a large corpus of post-edited translations: quality estimation, failure analysis and the variability of post-edition. In *Machine Translation Summit (MT Summit 2013)*, pages 117–124, Nice, France, 02/09 au 06/09.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondrej Bojar. 2011. Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96(1):79–88.

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.