Translation errors from English to Portuguese: an annotated corpus

Angela Costa, Tiago Luís, Luísa Coheur

INESC-ID and CLUNL, INESC-ID, INESC-ID and IST
Rua Alves Redol, 9
1000-029 Lisboa
angela@l2f.inesc-id.pt, tiago.luis@l2f.inesc-id.pt, luisa.coheur@inesc-id.pt

Abstract

Analysing the translation errors is a task that can help us finding and describing translation problems in greater detail, but can also suggest where the automatic engines should be improved. Having these aims in mind we have created a corpus composed of 150 sentences, 50 from the TAP magazine, 50 from a TED talk and the other 50 from the from the TREC collection of factoid questions. We have automatically translated these sentences from English into Portuguese using Google Translate and Moses. After we have analysed the errors and created the error annotation taxonomy, the corpus was annotated by a linguist native speaker of Portuguese. Although Google’s overall performance was better in the translation task (we have also calculated the BLUE and NIST scores), there are some error types that Moses was better at coping with, specially discourse level errors.

Keywords: Error analysis, Machine translation, Evaluation

1. Introduction

Error analysis is a field of research that originally analyses human errors, but nowadays it has become popular to evaluate natural language processing performances, for instance automatic translation tasks. Applying this knowledge to evaluate an automatic translation allows us to understand what type of errors are present in the translation, instead of just obtaining a score like BLUE. There are some works dedicated to the design of taxonomies (Llitjós et al., 2005; Vilar et al., 2006; Bojar, 2011) and others target errors’ identification (Popović and Ney, 2006). In this paper, we will use a new linguistically motivated taxonomy for translation errors that extends previous ones. Contrary to other approaches, our proposal:

- clusters different types of errors in the main areas of linguistics, allowing to precise the information level needed to identify the errors and easing a possible extension process;
- allows to classify errors that occur in Romance languages and not in English (being usually ignored in previous taxonomies);
- allows to take into consideration language’s variations;
- intends to cover both machine and human translation errors.

For this paper we have created a corpus constituted by automatic translations performed by two widely used translation engines (Google Translator and Moses) in three different scenarios representing different challenges in the translation from English to European Portuguese. A linguist native speaker of Portuguese has annotated this corpus using our error taxonomy and carried out an analyses of the type of errors that we have found.

2. Corpus

The error analysis was carried out on a corpus of 150 sentences, composed of:

- 50 sentences taken from a TED talk from Barry Schwartz, called On our loss of wisdom – from now on the TED corpus;
- 50 sentences taken from the “UP Magazine” from TAP (Transportes Aéreos Portugueses) – from now on the TAP corpus.
- 50 questions taken from the corpus made available by Li and Roth (from the TREC collection) (Li and Roth, 2002) – from now on the Questions corpus.

The TED corpus is constituted by TED Talks transcriptions and the EP translations created by volunteers and is available at the TED website. The TAP corpus is constituted by 51 editions of the Portuguese national airline company, divided in 2 100 files for EN and EP. It has almost 32 000 aligned sentences and a total of 724 000 Portuguese words and 730 000 English words. The parallel corpus of questions (EP and EN) consists of two sets of nearly 5 500 plus 500 questions each, to be used as training/testing corpora, respectively. Details on its translation and some experiments regarding statistical machine translation of questions can be found in (Angela Costa et al., 2012). Additional information about this and the previous corpus, can be found on the META-NET page\(^1\), where both corpora are freely available.

Some details on the word distribution of the resulting corpus are shown in Table 1.

\(^1\)http://metanet4u.l2f.inesc-id.pt/
TED and TAP sentences, which contain a lot more words per sentence. The TAP corpus is composed by sentences with a better grammar structure when compared with the TED corpus, which is mainly constituted by the transcription of non-planned speeches.

As shown, the Europarl dataset is much bigger than the 3 corpora, as it has almost 2M sentences. The TAP, TED and Questions corpus have 8462, 158184 and 8914 sentences, respectively. Despite this difference, the interpolation of the Europarl model with the TAP, TED and Questions models, individually, improves the translation, as shown in (Angela Costa et al., 2012). Regarding the translation quality, Table 4 shows the BLEU and NIST scores achieved by both systems. Since Google trained their system with more data, it is able to achieve better results than our Moses system. Moreover, they also incorporate translation errors corrections made by the users in their models, making their models even better.

### 4. Error taxonomy

Inspired by the work of (Vilar et al., 2006), (Bojar, 2011), we now present the taxonomy used.

#### 4.1. Substance level

Substance level errors include all the errors concerning misuse of punctuation and misspelling of words, so are not simply dealing with orthographic errors. We divide phrase and output phrase (in terms of identical input
substance level errors into three types: punctuation, capitalization and spelling.

4.2. Lexis level

Under this category we considered all errors affecting lexical items. It should be clear that, contrary to spelling errors that respect the characters used within a word, lexical errors concern the way each word, as a whole, is translated. Thus, the following types of errors at the lexis level are taken into account: omission, addition, untranslated and wrong lexical choice. Moreover, all these errors are then analysed considering the type of words they affect: content words and function words.

4.3. Grammar level

Grammar level errors are deviations in the morphological and syntactical aspects of language. On this level of analysis we identified two types of errors: misselection errors and misordering errors.

4.4. Semantic level

By semantic errors we understand problems that regard the meaning of the words and subsequent wrong word selection. We have individuated three different types of errors: confusion of senses, collocational and idiomatic. We should not confuse “wrong lexical choice” with “confusion of senses”, an example of the first case is, for instance, the translation of “care” as “conta” (check), there is no semantic relation between these two words. As for the translation of “glasses” as “óculos” (glasses) is a predictable “confusion of senses”, as the English word “glasses” can be translated into two different words in Portuguese: glasses to drink (“copos”) and glasses to see (“óculos”).

4.5. Discourse level

By discourse level errors, we consider the phenomenon that could be considered as a discursive option more than an error. We consider three different situations at the discourse level: style, variety and should not be translated. In all this cases, the meaning is preserved (thus, they are not semantic errors), but the chosen word is not the best choice.

In Figure 2 we resume the taxonomy previously presented. To simplify the readings, the subdivision between content and function words, although annotated in our corpus, is not present in the scheme.

5. Error Analysis

5.1. Google vs. Moses general overview by errors type

Figure 3 shows the errors found in Moses and Google, considering the different errors’ types proposed in our taxonomy. From this chart we can conclude that most errors occur on the Lexical and Grammatical level for both engines, independently of the type of text that it is translated.
and Questions corpora. In what concerns the Questions corpus, as previously explained, the training adaptation with the corpus created by (Angela Costa et al., 2012) overcome the problem of translating the Wh-words. For instance, What can be translated into Portuguese as O que but also as Qual, O qual, Quais, A que. This training corpus allowed Moses to make less errors of Confusion of Senses type than Google in this particular sub-corpus.

- At the Discourse Level of analysis Moses always behaved better than Google. We should underline that the Brazilian Portuguese (BP) was considered an error, as our goal was to reach a correct translation in European Portuguese (EP). As Google uses much data in BP, its translations use some vocabulary and grammatical forms that are only correct in BP, which contributed to these errors.

6. Conclusions

This work aimed at building three corpora from a TED-talk, the TAP magazine and a corpus of questions, each one representing a specific translation challenge. These corpora were then automatically translated by Moses and Google Translator and errors were manually annotated, according to an error taxonomy, allowing us to make a straightforward comparison between the two systems in the different corpora. We have seen that Google behaves better than Moses in almost every scenario. Moses main weakness are lexical errors: it does not really know how to translate many words. However, it can be adapted to specific lexicon/syntax, which is its most important feature. Google translator is particularly bad with (Portuguese) variety errors. Probably, as much of its sources of training are BP and it is not distinguishing between the two varieties.

7. Acknowledgements

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA00021/2013. Angela Costa was supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/85737/2012).

8. References

Angela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes, and Luísa Coheur. 2012. An english-portuguese parallel corpus of questions: translation guidelines and application in smt. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), pages 2172–2176, Istanbul, Turkey, may. European Language Resources Association (ELRA).

O. Bojar. 2011. Analysing error types in english-czech machine translation. The Prague Bulletin of Mathematical Linguistics, pages 63–76.

X. Li and D. Roth. 2002. Learning question classifiers. In Proc. 19th Int. Conf. Computational linguistics, pages 1–7. ACL.

Ariadna Font Llitjós, Jaime G Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In In Proceedings of EAMT 10th Annual Conference, pages 30–31.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Comput. Linguist., 29:19–51, March.

Maja Popović and Hermann Ney. 2006. Error analysis of verb inflections in spanish translation output. In TC-STAR Workshop on Speech-to-Speech Translation, pages 99–103, Barcelona, Spain.

David Vilar, Jia Xu, Luis Fernando D’Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In International Conference on Language Resources and Evaluation, pages 697–702, Genoa, Italy, May.