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 (Ângela 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.
This set of sentences was translated with Google Translate and Moses. The next section shows more details about this step. Some examples of sentences from these corpora can be found in Table 2. The questions corpus has small sentences with a fixed and simple grammar structure, unlike

[1] http://metanet4u.l2f.inesc-id.pt/

Dataset	Language	Tokens
TAP	EN	984
	PT	1136
TED	EN	844
	PT	854
Questions	EN	377
	PT	375

Table 1: Data used in the error analysis.

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.

TED	The publisher bears no responsibility for return of unsolicited material and reserves the right to accept or reject any editorial and advertising material. No parts of the magazine may be reproduced without the written permission of up. The opinions expressed in this magazine are those of the authors and not necessarily those of the auditor.
TAP	They're the things you would expect: mop the floors, sweep them, empty the trash, restock the cabinets. It may be a little surprising how many things there are, but it's not surprising what they are.
Questions	Who developed the vaccination against polio? What is epilepsy? What year did the Titanic sink? Who was the first American to walk in space?

Table 2: Examples of sentences.

3. Systems and tools

3.1. Machine Translation Systems

Both Google Translate and Moses were used in our experiments. Concerning Google Translate, we have no control in its models. However, in what respects Moses, we created its models in order to have them adapted, as much as possible, to the three different scenarios. As usual, the directional word alignments were produced by GIZA++ (Och and Ney, 2003), using the IBM M4 model and combined using the grow-diagonal-final heuristic. The resulting word alignments were then used to create phrase-based translation models. We started by training and tuning a baseline phrase-based system for the EN-PT direction, using only data from the Europarl parallel corpus. Next, we trained and tuned SMT models with the training and development set from the different parallel corpus, and combined these (both translation and language models) with the Europarl models, during decoding using a set of weights tuned with MERT. In this way, the Moses decoder tries to gather the translation hypotheses from the questions models, TAP models and TED talk corpus, respectively, and collects additional options from the Europarl models. If the same translation hypotheses (in terms of identical input phrase and output phrase) is found in both models, separate translation hypotheses are created for each occurrence, but with different scores. The weights of the models were tuned with Minimum Error Rate Training (MERT). Table 3 shows the statistics of the corpus used to train the models.

Dataset	Language	Tokens
Europarl	EN	54,720,731
	PT	53,799,459
TAP	EN	171,338
	PT	169,974
TED	EN	1,306,938
	PT	1,233,616
Questions	EN	89,264
	PT	95,462

Table 3: Data used to train the Moses system.

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 (Ângela 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.

Dataset	System	BLEU	NIST
Questions	Moses	41.52	5.77
	Google	63.33	6.90
TAP	Moses	18.84	4.96
	Google	26.33	6.09
TED	Moses	19.75	5.49
	Google	27.54	5.77

Table 4: BLEU and NIST scores achieved by the Moses and Google systems when evaluated on each test dataset.

3.1.1. UAM CorpusTool

Our corpus was annotated using UAM CorpusTool[2], a state-of-the-art environment for annotation of text corpora (see Figure 1).

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

[2]http://www.wagsoft.com/CorpusTool

Type of Study: Describe each file | Aspect of Interest: Feature Coding | Counting: Local

Unit: errors | Show

Feature	ted-trad-mose N	ted-trad-mose Percent	que-trad-mose N	que-trad-mose Percent	tap-trad-mose N	tap-trad-mose Percent
Total Units	242		110		399	
ERRORS-TYPE	N=242		N=110		N=399	
- substance	4	1.65%	0	0.00%	4	1.00%
- lexis	109	45.04%	73	66.36%	256	64.16%
- grammar	87	35.95%	26	23.64%	96	24.06%
- semantics	40	16.53%	11	10.00%	35	8.77%
- discourse	2	0.83%	0	0.00%	8	2.01%
SUBSTANCE-TYPE	N=4		N=0		N=4	
- punctuation	4	100.00%	0	0.00%	0	0.00%
- capitalization	0	0.00%	0	0.00%	2	50.00%
- spelling	0	0.00%	0	0.00%	2	50.00%
LEXIS-TYPE	N=109		N=73		N=255	
- omission	23	21.10%	17	23.29%	36	14.12%
- addition	28	25.69%	12	16.44%	26	10.20%
- untranslated	42	38.53%	35	47.95%	101	39.61%
- wrong-lexical-choic	16	14.68%	9	12.33%	92	36.08%
GRAMMAR-TYPE	N=87		N=26		N=96	

Figure 1: Statistical Study

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.

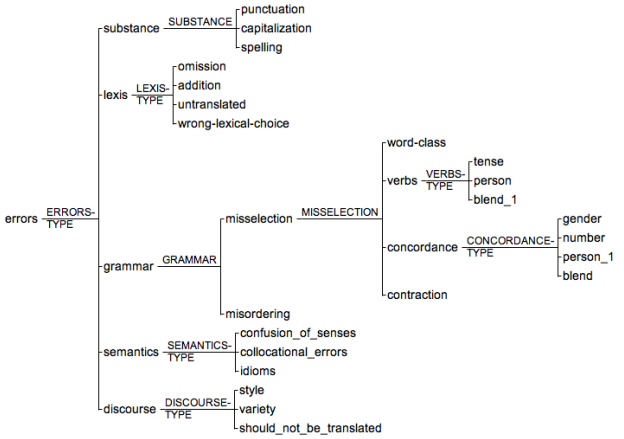

Figure 2: Taxonomy

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.

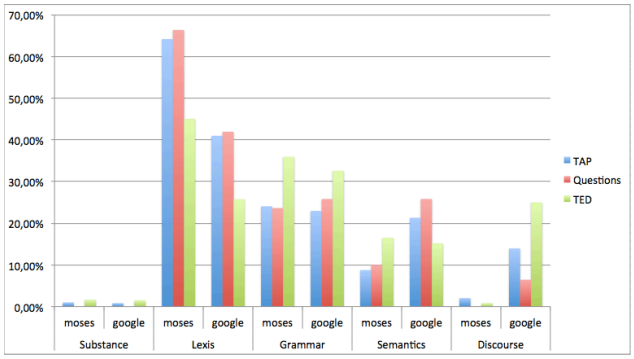

Figure 3: Errors in Moses and Google, by errors' type

Table 5 contrasts Google results with Moses. In all the entries, the first element is the number of Moses errors and the second, the number of Google errors.

From this table, we can see that:

- Moses has much more `Lexical` (and `grammar`) errors that Google in all the corpora translations.
- On the `Semantic Level` of analysis, Moses behaves better than Google in the translation of the TAP

Moses/Google	TAP	Questions	TED	Errors (total)
Substance	4/1	0/0	4/2	8/3
Lexis	256/50	73/26	109/34	438/110
Grammar	96/28	26/16	87/43	209/87
Semantics	35/26	11/16	40/20	86/62
Discourse	8/17	0/4	2/33	10/54

Table 5: Moses vs. Google errors

and Questions corpora. In what concerns the Questions corpus, as previously explained, the training adaptation with the corpus created by (Ângela 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 quł*, *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/LA0021/2013. Angela Costa was supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/85737/2012).

8. References

Ângela 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.