Building a Corpus of Errors and Quality in Machine Translation: Experiments on Error Impact

Ângela Costa, Rui Correia, Luísa Coheur
INESC-ID and CLUNL, INESC-ID and LTI/CMU, INESC-ID and IST
Rua Alves Redol, 9
1000-029 Lisboa
angela@l2f.inesc-id.pt, rui.correia@l2f.inesc-id.pt, luisa.coheur@inesc-id.pt

Abstract

In this paper we describe a corpus of automatic translations annotated with both error type and quality. The 300 sentences that we have selected were generated by Google Translate, Systran and two in-house Machine Translation systems that use Moses technology. The errors present on the translations were annotated with an error taxonomy that divides errors in five main linguistic categories (Orthography, Lexis, Grammar, Semantics and Discourse), reflecting the language level where the error is located. After the error annotation process, we accessed the translation quality of each sentence using a four point comprehension scale from 1 to 5. Both tasks of error and quality annotation were performed by two different annotators, achieving good levels of inter-annotator agreement. The creation of this corpus allowed us to use it as training data for a translation quality classifier. We concluded on error severity by observing the outputs of two machine learning classifiers: a decision tree and a regression model.

Keywords: machine translation, translation errors, translation quality

1. Introduction

Error analysis is a linguistic discipline that has been mostly connected to language learning, but more recently has been used to evaluate automatic language production, like Speech Recognition, Summarization, but specially in Machine Translation (MT).

In this paper, we describe the creation of a corpus annotated with errors and translation quality. First, using a specifically designed taxonomy (Costa et al., 2015), we annotate the errors present on translations generated by four different systems: two mainstream online translation systems – Google Translate (Statistical) and Systran (Hybrid Machine Translation) – and two in-house MT systems that use Moses technology. This is done in three scenarios representing different translation challenges from English (EN) to European Portuguese (EP). We then evaluate translation quality using a 4-level classification system based on comprehension.

The motivation to build a corpus annotated with errors and translation quality was to help determine future research directions in the MT area, allowing, for instance, to highlight which kind of errors hinder comprehension the most.

Finally, for this paper, using the presented corpus and the types of errors from the taxonomy of choice as features, we set up different machine learning classifiers to predict translation quality (decision trees and regression models). Then, by looking into the outputs and parameters of the classifiers, we infer the severity of different error types and show how their presence impacts on quality.

In Section 2. we present previous work on error annotation and translation quality. Section 3. describes the text sources, the tools and the annotation process (errors and quality). Section 4. addresses error gravity, showing the classification results. Finally, in Section 5., we highlight the main conclusions and point to future work directions.

2. Related work

Several studies have been developed with the goal of classifying translation errors and different taxonomies have been suggested.

One of the most used in MT is the hierarchical classification proposed by (Vilar et al., 2006), followed up by (Bojar, 2011). They extend the work of (Llitjs et al., 2005) and split errors into five categories: *Missing Words* (when some of the words are omitted in the translation), *Word Order* (when the words in the target sentence are wrongly positioned), *Incorrect Words* (when the system is unable to find the correct translation for a given word), *Unknown Words* (when an item is simply copied from the input word to the generated sentence, without being translated), and *Punctuation*.

Another approach, this time for human errors, comes from (H. Dulay and Krashen, 1982). They suggest two major descriptive error taxonomies: the Linguistic Category Classification (LCC) and the Surface Structure Taxonomy (SST). LCC is based on linguistic categories (general ones, such as morphology, lexis, and grammar and more specific ones, such as auxiliaries, passives, and prepositions). On the other hand, SST focuses on the way surface structures have been altered by learners (e.g., omission, addition, misformation, and misordering). These two approaches are originally presented as alternative taxonomies, although we found work in the literature pointing to the advantages on their combination (James, 1998). In this work, we extend and unify these approaches in a taxonomy shown on Section 3.3..

Regarding error gravity, i.e. how different errors impact translation quality, literature focuses on two different five point scales that judge fluency (native-like performance) and adequacy (how much of the original meaning is expressed in the translation) (Callison-Burch et al., 2007; Koehn and Monz, 2006). When judging fluency, a 5 is at-

tributed to flawless English and 1 to incomprehensible English. The best score for adequacy is 5 and this means that the whole meaning of the reference is also expressed in the translation, while 1 means that none of the original meaning was kept. However, it seems that people have difficulty in evaluating these two aspects of language separately. Also (Callison-Burch et al., 2007) points out to the difficulty of defining objective scoring guidelines, for example, how many grammatical errors (or what sort) distinguish between the different levels.

Finally, (Daems et al., 2013)'s work merges both error classification and error gravity, allowing to assign weights to each type of error. Herein, weights should be hand-tuned by the user so that problems that have a larger impact on comprehension receive a higher weight. For example, for the task of translating terminology texts, the impact of a lexical error can be set up as more severe than other types of errors (such as grammatical). Contrastingly, we believe that creating a corpus annotated with errors and translation quality, will allow to find out these weights empirically, observing from data which errors most affect comprehension.

3. Corpus

In this section, we start by describing the selection of the corpus (Section 3.1.) and MT systems used to translate it (Section 3.2.). In Section 3.3., we describe the task of annotating the errors obtained after translation. Lastly, in Section 3.4., we outline the quality assessment process.

3.1. Text sources

The corpus is composed of 300 sentence pairs (source/translation) evenly distributed (25 pairs each) between **(a)** TED-talks transcriptions (and respective EP subtitles)[1], **(b)** texts from the bilingual Portuguese national airline company magazine UP[2], and **(c)** the translated TREC evaluation questions (Li and Roth, 2002; Costa et al., 2012).

The TED corpus is composed of TED talk subtitles and corresponding EP translations. These were created by volunteers and are available at the TED website. As we are dealing with subtitles (and not transcriptions), content is aligned to fit the screen, and, thus, some pre-processing was needed. Therefore, we manually connected the segments in order to obtain parallel sentences.

The TAP corpus is constituted of 51 editions of the bilingual Portuguese national airline company magazine, divided into 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 corpus, respectively. Details on its translation and some experiments regarding statistical machine translation of questions can be found in (Costa et al., 2012).

[1] http://www.ted.com/

[2] http://upmagazine-tap.com/

3.2. Translation tools

For our experiments we used 4 different MT systems. Two are mainstream online translation systems[3]: Google Translate[4] and Systran[5] (further referred to as **G** and **S**, respectively). The other two are in-house MT systems, both trained using Moses. One of them uses a phrase-based model (Koehn et al., 2007) and the other an hierarchical phrase-based model (Chiang, 2007) (further **PSMT** and **HSMT**). Both in-house systems share the same training, approximately 2 million sentence pairs from Europarl (Koehn, 2005).

3.3. Error Annotation

Having decided on the corpora (see Section 3.1.) and MT systems to use (see Section 3.2.), we generated the corresponding translations. Given that we had 3 different sources of data (each with 25 sentences) and 4 MT systems, the final set of translations is composed of 300 sentences. This material was annotated by a linguist with the errors represented in figure 1.

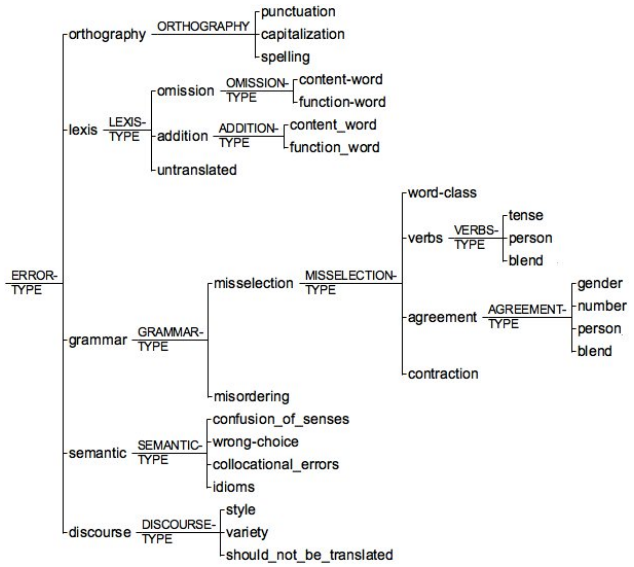

Figure 1: Taxonomy of Translation Errors

This taxonomy is divided in 5 main categories, reflecting the language level where the error is located. `Orthography` level errors include all the errors concerning misuse of punctuation, capitalization, and misspelling of words. At the `exical` level we identify `omissions`, `additions` and `untranslated` items. `Omissions` and `additions` are analysed considering the type of words they affect (content vs. function words). The category `untranslated` is similar to *Unknown Words* from (Vilar et al., 2006)'s work.

`Grammar` errors are deviations at the morphological and syntactical level. Herein we distinguish between `misordering` (similar to *Word Order* from (Vilar et al., 2006)'s work), and `misselection` (morphological misformations or agreement issues). The latter occur regard-

[3] Translated on 22/10/2014

[4] http://translate.google.com

[5] http://www.systranet.com/translate

ing word class (an adjective instead of a noun), verb form, agreement, and contractions (for instance, in EP "in the" (*em o*) should be contracted to *no*).
`Semantic` errors address meaning problems: `Confusion of senses` (choosing the wrong word sense[6]), `Wrong choice` (using words with no apparent relation to the source word), `Collocational`, and `Idiomatic` errors.
Finally, `Discourse` encompasses unexpected discursive options, dealing with problems of `style` (bad stylistic choice of words, such as the repetition of a word in a near context) and `variety` (in our case, typically the confusion between EP and Brazilian Portuguese (BP)[7]). Errors in the `should not be translated` class represent cases where words in the source language should not have been translated (such as movie titles, company names, etc).
Table 1 shows the number of errors generated by each MT system across the 5 main categories. We can see the impact of the quantity of training data on performance (Google consistently outperforms the other systems), with the only exception being `Discourse`, where Google suffers from Portuguese variety problems, as it does not distinguish between European and Brazilian Portuguese. This is a problem as there are spelling, vocabulary and grammatical divergences between the two varieties.

	G	S	PSMT	HSMT	Total
Orthography	34	69	218	233	554
Lexis	380	883	606	700	2,569
Grammar	404	629	649	713	2,395
Semantics	334	783	486	489	2,092
Discourse	186	134	37	36	393

Table 1: Errors per category across MT systems.

For agreement purposes, a second person also annotated the 300 sentences. The agreement was measured with Cohen's kappa in two dimensions. First, regarding error localization, achieving an agreement of 0.96. Secondly, regarding the type of error at the different four layers of the taxonomy. Annotators achieved high agreement for all layers, with $0.91 \leq \kappa \leq 0.97$.

3.4. Translation Quality Annotation

The last setup step was to assign a quality score to the translations. Therefore, in line with other approaches to quality (Callison-Burch et al., 2007), we defined a scale of quality that distinguishes between 4 comprehension levels: **Q1)** no errors or only minor ones that do not affect comprehension; **Q2)** comprehension problems, but most of the original meaning is still preserved; **Q3)** severe errors, but some meaning can still be inferred; and finally, **Q4)** meaning is completely compromised[8]. It is also important to mention that the decision to use four classes instead of a more fine-grained classification is due to the low agreement reported in the similar task of judging fluency and adequacy on a 5-point scale (Callison-Burch et al., 2007).
Table 2 shows the distribution of quality per system. Google clearly contrasts with the other systems, producing 27 perfect or near-perfect translations and only 3 completely unintelligible ones.

	G	S	PSMT	HSMT	Total
Q1	27	6	13	12	58
Q2	13	21	8	9	51
Q3	32	33	37	30	132
Q4	3	15	17	24	59

Table 2: Translations per quality level across MT systems.

For agreement, the same annotator that annotated the errors on the 300 sentences (Section 3.3.), also carried out this evaluation achieving a kappa of 0.41, which according to (Landis and Koch, 1977) is considered a moderate agreement. Note that this lower agreement was expected given the subjectivity involved in this qualitative evaluation, also observed by (Callison-Burch et al., 2007). For this reason, in the remaining of this work, we will use the judgements of the first annotator only (linguist). By doing this, we preserve the order attributed by the expert, which is expected to be coherent within his interpretation of the quality scale.

4. Error Type vs. Translation Quality

With the corpus that we have created, we were able to build classifiers capable of predicting translation quality based on error presence. The goal of this task is first to understand how an error taxonomy can contribute to automatically assess translation quality, and second to understand what types of errors seem to hinder comprehension the most.
To do this, we use WEKA[9] and two different strategies: decision trees (Section 4.1.) and regression models (Section 4.2.). In both cases, we used the following features: **a)** number of words in the translation, **b)** total number of errors (with no type distinction), **c)** one feature for each error type (at different layers of the taxonomy) representing how many errors of that type are in the translation, and **d)** one feature for each error type representing its proportion, i.e., the number of errors of that type divided by the total number of errors in the sentence. All experiments are trained on 300 translations, with a 10-fold cross-validation strategy.

4.1. Decision Trees

As a first strategy, we built a decision tree (J48 – WEKA's implementation of C4.5 algorithm). Treating quality as a nominal attribute, i.e., assign a quality class to each translation (in Q1 to Q4), we set up a baseline using only the aforementioned features **a)** and **b)**. This setup achieved an *f–measure* $= 0.566$ ($prec = 0.645$; $rec = 0.623$).
When taking into account the error taxonomy, i.e., adding features **c)** and **d)**, we registered a significant improvement in performance. The best setup ended up ignoring feature **a)**, and achieved *f-measure* $= 0.662$ ($prec = 0.665$; $rec =$

[6]Ex: glasses to drink (*copos*) and glasses to see (*óculos*)
[7]Common in Google translations given the supremacy of Brazilian Portuguese content online.
[8]More details on the evaluation guidelines will be provided in the full version of the article.
[9]`http://www.cs.waikato.ac.nz/ml/weka/`

0.670). Table 3 shows the confusion matrix for this setup. We can see that levels Q1 and Q3 are rarely confused, while Q2 and Q4's translations are often attributed to contiguous levels only (the tree never classified a Q4 translation as Q1).

	classified as			
	Q1	Q2	Q3	Q4
Q1	45	6	6	1
Q2	14	20	15	2
Q3	4	14	105	9
Q4	0	2	26	31

Table 3: Confusion matrix for the best tree configuration.

When looking at the tree itself we see that after a first decision based solely on error quantity, it tends to focus on the presence of `semantic` errors, specially `wrong choice`[10] (ex: the rule *If (semantic:wrong_choice* > 0*) then* $Q = 4$, correctly classifies 26 items and fails 5). Further along the tree, we also see addressed `grammar` and `lexical` errors (mostly with no further specification). As expected, errors of `orthography` do not help when classifying quality.

4.2. Regression

We used a support vector machine for regression (SMOreg in WEKA). Herein, quality is seen as a numeric attribute, as it corresponds to a scale of gravity, and, for that reason, we are reporting correlation coefficient (ρ). As previously, we set a baseline with the features **a)** and **b)** alone. This setup produced $\rho = 0.7040$.

When adding the remaining features, we again noticed an improvement in performance. In this case, the best setup was the one with features **b)**, and **d)** only. When adding the proportion of each type of errors with respect to the total number of errors in the sentence, we achieved $\rho = 0.7528$. In Table 4 we show the highest (on the left) and lowest (right) weight assigned by this model. `Wrong choice` appears again having a high impact on the classification. The remaining top features were all from the first layer of the taxonomy, with the exception of `should not be translated`. Unexpectedly, this error proved to be relevant. Not preserving the original form of certain named entities, is not simply a matter of style, often introducing a chain of translation errors[11]. Regarding the lowest ranked features, we see that problems of `agreement`, `punctuation` or `variety` are not indicative of quality.

w	feature	w	feature
0.603	lexis	0.035	verbs: blend
0.480	wrong choice	0.023	punctuation
0.467	grammar	0.020	variety
0.390	semantics	0.013	addition: func word
0.277	should not be translated	0.010	verbs: person

Table 4: Feature weight sample for regression model.

5. Conclusions

In this paper we presented a corpus of automatic translations annotated with both error type and quality. The creation of this corpus can help distinguishing how different error types impact on translation quality. To demonstrate the utility of this corpus, we did an experiment with different machine learning classifiers to predict translation quality. Then, by looking into the outputs and parameters of the classifiers, we conclude that `wrong choice` problems and translating instances that should be kept in the original form, being suitable features to assess translation quality, should be targeted first in future MT efforts.

Future work includes continuing the error and quality annotation effort so as to have more data available.

[10]When a wrong word, without any apparent relation, is used to translate a given source word.

[11]For instance, "Sears building" translated as "to build triggers" (*gatilhos*).

6. Acknowledgements

This work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project UID/CEC/50021/2013. Ângela Costa and Rui Correia are supported by PhD fellowships from FCT (SFRH/BD/85737/2012 and SFRH/BD/51156/2010).

7. Bibliographical References

Bojar, O. (2011). Analysing Error Types in English-Czech Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, pages 63–76.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chiang, D. (2007). Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June.

Costa, A., Luís, T., Ribeiro, J., Mendes, A. C., and Coheur, L. (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).

Costa, A., Ling, W., Luís, T., Correia, R., and Coheur, L. (2015). A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Daems, J., Macken, L., and 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 Sharon O'Brien, et al., editors, *MT Summit XIV Workshop on Post-editing Technology and Practice, Proceedings*, pages 63–71. European Association for Machine Translation.

H. Dulay, M. B. and Krashen, S. D. (1982). *Language Two*. Newbury House, Rowley.

James, C. (1998). *Errors in Language Learning and Use. Exploring Error Analysis*. Applied Linguistics and Language Study. Routledge.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 102–121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Landis, R. J. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–74.

Li, X. and Roth, D. (2002). Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, pages 1–7. ACL.

Llitjs, A. F., Carbonell, J. G., and Lavie, A. (2005). A Framework for Interactive and Automatic Refinement of Transfer-based. In *Machine Translation. European Association of Machine Translation (EAMT) 10th Annual Conference*, pages 30–31.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May.