Challenges on the Automatic Translation of Collocations

Ângela Costa
L2F/INESC-ID
CLUNL
angela@l2f.inesc-id.pt

Teresa Lino
CLUNL
tlino@mail.telepac.pt

Luísa Coheur
L2F/INESC-ID
IST
luisa.coheur@l2f.inesc-id.pt

Abstract

As a linguistic phenomenon, collocations have been the subject of numerous researches both in the fields of theoretical and descriptive linguistics, and, more recently, in automatic Natural Language Processing. In the area of Machine there is still improvements to be done, as major translation engines do not handle collocations in the appropriate way and end up producing literal unsatisfactory translations. Having as a starting point our previous work on machine translation error analysis (Costa et al., 2015), in this article we present a corpus annotated with collocation errors and their classification. To our believe, to have a clear understanding of the difficulties that the collocations represent to the Machine Translations engines, it is necessary a detailed linguistic analysis of their errors.

1 Introduction

According to (MELćUK, 1998), collocations are particularly relevant in the context of lexical combinatory as they are "the absolute majority of phrasemes and represent the main challenge for any theory of phraseology". (Tutin Agnés, 2002) defines them as "a privileged lexical co-occurrence of two (or more) linguistic elements that establish a syntactic relationship between them". (Hausmann, 1989), (Hausmann, 1985) and (Hausmann, 1984) observed that the status of the constituents are not similar, registering between them an hypotactic relationship. Hausmann calls "base" to the word that determines the choice of the co-occurring element and "collocate" the determined constituent.

The relationship between base and collocate is, in most cases, unpredictable, and does not demonstrate a particularly clear semantic motivation that can explain it. This idiosyncratic character and the fact that they cannot yet be considered lexicalized expressions, standing between lexicon and grammar, makes them very complex structures, from the production point of view. In fact, (Cruse, 2004) considers them "idioms of encoding", as they do not particularly cause problems from the decoding perspective, being relatively transparent constructions and syntactically regular. The problem lies on producing them, since the relationship between the base and collocate is, in most cases, arbitrary. Considering the translation task, we can imagine the number of problems that can occur, as a word-by-word translation may not always be the best choice. For instance, *break a record* cannot literally be translated into French *casser un record*, but as *battre un record* (lit. to beat a record).

In this article we briefly describe the role of collocations on machine translation, then we describe the corpus and error typology used in our study, finally we present the error analyses and the conclusions.

2 Collocations in Machine Translation

Collocations have been the subject of numerous researches both in the fields of theoretical and descriptive linguistics, and, more recently, in Natural Language Processing (NLP), as they can be useful for many language processing tasks, like parsing, word sense disambiguation, text generation and machine translation.

Although there are several methods for the extraction of collocations from corpora and evaluation of extraction results, the area of post-processing of this structures and their application to various branches of NLP is still at the beginning, especially in the area of machine translation (Seretan and Wehrli, 2007). Because of their semantic irregularities, collocations cannot always be translated word-by-word, creating a problem for automatic translation. In this example of a

Google translation the collocation *high wind* was literally translated to *vento alto* (lit. tall wind) instead of *vento forte*. On the other side, sometimes a literal translation may be correct, *make the bed* was translated to *fazer a cama* which is correct. Just as for a student learning a foreign language, also for an MT system is not always easy to know when the correct option is a word-by-word translation.

Error analysis of collocations in machine translation is still lacking. For instance two of the most used error taxonomies by (Bojar, 2011) and (Vilar et al., 2006) do not consider collocational errors on their classification. As previously mentioned, collocations have at least two elements, so the errors may concern any of the elements of the collocation (base, collocate) or the collocation as a whole. Finding the error within is compositional parts can help improve the translation of these structures.

3 Error Analysis

3.1 Corpus

Having as a starting point our previous work on machine translation errors (Costa et al., 2015), the error analysis of collocations was carried out on a corpus generated by four different systems: Google Translate[1] (Statistical), Systran[2] (Hybrid Machine Translation) and two in-house Machine Translation systems trained using Moses[3], and the two popular models: the phrase-based model (Koehn et al., 2007) (PSMT) and the hierarchical phrase-based model (Chiang, 2007) (HSMT), in three scenarios representing different challenges in the translation from English to European Portuguese:

- 250 sentences taken from TED talks[4];
- 250 sentences taken from the bilingual Portuguese national airline company: TAP magazine "UP"[5];
- 250 questions taken from a corpus made available by (Li and Roth, 2002), from the TREC collection (Li and Roth, 2002; Costa et al., 2012).

[1] http://translate.google.com
[2] http://www.systranet.com/translate
[3] http://www.statmt.org/moses
[4] http://www.ted.com/
[5] http://upmagazine-tap.com/

The Ted talks, in the original text in English had 3.346 tokens, the TAP and the corpus of Questions had 3.346 and 1.856, respectively. We were able to find a total of 172 collocations: 41 were found on the TED corpus, 84 on the TAP magazines and 47 on the Questions corpus. As previously mentioned, the three datasets were translated by four translation engines, so in total we have evaluated 164 collocations on the TED corpus, 336 on the TAP corpus and 188 on the Questions corpus.

3.2 Error types

To assess the errors that we have found, we used the location dimension of (Wanner et al., 2011) taxonomy to evaluate students errors when producing collocations. The first two categories show errors that were found on one of the two elements of the collocation (cf. (1) wrong collocate use and (2) wrong base use) and the third type problems that affected the collocation as a whole (cf. (3)).

1. **wrong collocate**: *cores preliminares*, lit. "preliminary colors" (instead of *cores primárias*, "primary colors"), *cabelo cinzento*, lit. "gray hair" (instead of *cabelo grisalho*, "gray hair"), *terra nativa*, lit. "native land" (instead of *terra natal*, "native land")

2. **wrong base**: *perspectiva obtusa*, lit. "obtuse perspective" (instead of *ângulo obtuso*, "obtuse angle"), *começar uma faixa*, lit. "start a strip" (instead of *começar uma banda*, "start a band"), *meta cardíaca*, lit. "heart goal" (instead of *ritmo cardíaco*, "heart rate"), *flopped miseravelmente*, lit. "flopped miserably" (instead of *falhar miseravelmente*, "failed miserably")

3. **wrong collocation**: *pagamento de separação*, lit. "payment of separation" (instead of *indemnização*, "compensation"), *ter ceia*, lit. "have supper" (instead of *jantar*, "have diner")

   The errors found on a collocation can be rooted in the lexicon or in the grammar. A lexicon error concerning the base or the collocate consists in the incorrect translation of one of the two elements or both. This error can be caused by a literal translation from English that does not work in the context of the collocation, a near-synonym or even the non-translation of an element (see examples (1)

and (2)). When the error concerns the whole collocation, we found that new expressions with the structure of a collocation were created, meanwhile a single word should have been used (see examples (3)).

Grammatical errors can also affect the collocation as a whole or all of its parts (base and collocate). We were able to find four types: erroneous absence or presence of determiner, wrong number use, wrong order of the words and wrong government; cf:

4. **determiner**: *pedir a ajuda*, lit. "ask the help" (instead of *pedir ajuda*, "ask for help").

5. **number**: *mudar os canais*, lit. "change the channels" (instead of *mudar o canal*, "change the channel").

6. **reordering**: *chá de conjunto*, lit. "tea of set" (instead of *conjunto de chá*, "set of tea").

7. **government**: *sede para conhecimento*, lit. "thirst for knowledge" (instead of *sede de conhecimento*, "thirst for knowledge"), *carreira solo*, lit. "career solo" (instead of *carreira a solo*, "solo career").

4 Results

Figure 1 shows the number of errors present on each translation engine per error type. The correct translations are not represented on the graphic, but they were the majority of the cases, as Google, HSMT, PSMT and Systran produced 144, 114, 111 and 92 correct translation, respectively. From Figure 1, we can observe that:

- choosing the correct base of the collocation is not as problematic as deciding on the collocate, as this is the most common error for all engines;
- between 14% and 19% of the errors affect the collocation as a whole;
- determinant, number, reordering and government errors are not so common.

5 Conclusions

From this study we could observe that only between 14% and 19% of the errors affect the collocation as a whole. Determinant, number, reordering and government errors are not so common, as

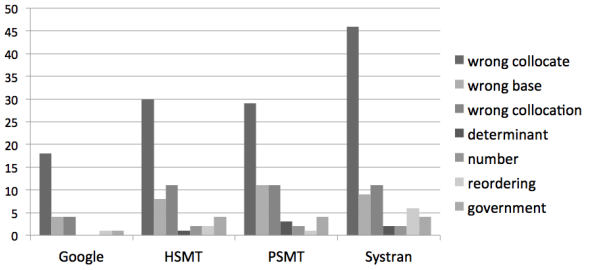

Figure 1: Number of collocation errors per system.

there is a relativity high congruence between English and Portuguese, although this may not be valid for other languages.

On all four MT systems the majority of the errors occur when choosing the collocate. This was also observed on foreign language learners on the already mentioned study by (Wanner et al., 2011). The source of the errors are literal translations of the collocate ("grey" - cinzento), use of a wrong synonym ("angle" - *perspectiva*) or untranslations (e.g. "flopped").

Although our analysed corpus is still very small, we think that it is a good contribution to have a clear understanding of the difficulties that the collocations represent to Machine Translations engines. Only after a detailed linguistic analysis of the errors, we can implement solutions, like finding and automatically correcting collocations.

Acknowledgments

This work is supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project UID/CEC/50021/2013. Ângela Costa is supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/85737/2012).

References

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

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

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).

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

D. A. Cruse. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press.

Franz Josef Hausmann. 1984. Wortschatzlernen ist kollokationslernen. zum lehren und lernen franzsischer wortverbindungen. *Praxis des neusprachlichen Unterrichts 31*, pages 395–406.

Franz J. Hausmann. 1985. Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In Henning Bergenholtz and Joachim Mugdan, editors, *Lexikographie und Grammatik: Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984*, Lexikographica, pages 118–129. Max Niemeyer, Tübingen.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations (artikel 95). In Wiegand H.E. Zgusta L. Hausmann F.J., Reichmann O., editor, *Wörterböcher - Dictionaries - Dictionnaries. Ein internationales Handbuch zur Lexicographie. Erster teilband.* Walter de Gruyter, Berlin/New York.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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.

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

Igor MELćUK. 1998. Collocations and lexical functions. *2001 [1998]*, pages 23–54.

Violeta Seretan and Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 401–410.

Grossmann Francis Tutin Agnés. 2002. Collocations réguliéres et irréguliéres : esquisse de typologie du phénoméne collocatif. *Revue française de linguistique appliquée*, VII:7–25.

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.

Leo Wanner, M Alonso Ramos, Orsolya Vincze, Rogelio Nazar, Gabriela Ferraro, Estela Mosqueira, and Sabela Prieto. 2011. Annotation of collocations in a learner corpus for building a learning environment. *Sylviane Granger/Gaëtanelle Gilquin/Fanny Meunier (edd), Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead. Corpora and Language in Use. Proceedings*, 1.