Widening the Discussion on "False Friends" in Multilingual Wordnets

Hugo Gonçalo Oliveira, Ana R. Luís
University of Coimbra, DEI-CISUC, University of Coimbra, CELGA-ILTEC
hroliv@dei.uc.pt, aluis@fl.uc.pt

Abstract

Cognates are words that have similar meaning and spelling in two or more languages (Carroll, 1992), such as *impossible* (English) and *impossível* (Portuguese) or *education* (English) and *educação* (Portuguese). Although Portuguese is a Romance language and English is a Germanic language, they share an extraordinary high number of cognates which are essentially of Latin and Greek origin (Domíngues, 2008).

Portuguese and English also have *false friends*, namely pairs of words that appear similar but have a different meaning. Examples include *push* (English) and *puxar* (Portuguese), meaning 'to pull'; *library* (English) and *livraria* (Portuguese), meaning 'bookshop'; or *beef* (English) and *bife* (Portuguese), meaning 'steak'. Among them, some pairs of words are 'partial' *false friends* as they may have different meanings only in some contexts. Examples include *medicine* (English), which is cognate with *medicina* (Portuguese), but can also mean 'substance used to treat an illness'; or *figure* (English), which is cognate with *figura* (Portuguese), but can also mea 'number'.

Cognates have been successfully identified with Natural Language Processing techniques using methods such as orthographic similarity and semantic similarity, combined with machine learning (Bradley & Kondrak, 2011). While the identification of cognates has made much progress, the identification of *false friends* is still an under-researched area. But available studies show that it is an area from which other areas that support Natural Language Processing, including the development of computational lexical resources, could benefit (Hefler, 2017; Castro, Bonanata & Rosá, 2018).

Focusing on wordnets for different languages, and towards multilingual processing, two main strategies have been adopted for alignment with the Princeton WordNet (Fellbaum, 1998): the expand or the merge approach (Vossen, 2002). In both cases, *false friends* can be a source of errors, either in the application of automatic steps (e.g., for automatic translation) or simply due to a lack of knowledge of the people involved, influenced by orthographic similarity.

For instance, with a quick search in the Portuguese wordnet OpenWordNet-PT (Paiva, Rademaker & Melo, 2012) a few errors of this kind are identified. In Table 1, we present some of them, with the id and the words of the English synset, followed by the words of the aligned Portuguese synset, followed by a brief explanation of the problem. Although the focus of this exploratory exercise was on Portuguese and English, such problems are common among other pairs of languages.

Motivated by these problems, we aim to open a discussion on the potential benefits of further research on *false friends* in the development of wordnets and other multilingual linked resources. Possible tasks to tackle could exploit lists of *false friends* from the literature for cleaning multilingual wordnets. A simple thing to do would be to remove *false friends* from linked synsets, or even to remove the connection between those synsets. Moreover, an RDF property could perhaps be used for explicitly linking pairs of lexical items, in different languages, that are *false friends*. Besides enabling other lines of research, this information could also be considered in further expansions of the wordnet.

Synset ID	Portuguese	English	Explanation
02275799-v	*pretender*	*pretend*	*pretend* (EN) means 'deceive', while *pretender* (PT) means 'want, intend
02374914-a	*simpático*	*sympathetic*	*sympathetic* (EN) means 'showing compassion', while *simpático* (PT) means 'nice, friendly'
00074558-v	*constipar*	*constipate*	*constipate* (EN) is related to 'difficulty in emptying the bowels', while *constipar* (PT) means 'getting a cold'
10661216-n	*estrangeiro*	*stranger*	*stranger* (EN) is a person who is unkown, while *estrangeiro* (PT) is a 'foreigner'

Table 1 : False friend-related issues in OpenWordNet-PT

Keywords: false friends, cognates, multilingual wordnets

Bibliographical References

Bradley, H. and Kondrak, G. (2011). Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 865–873.

Carroll, S. (1992). On cognates. *Second Language Research* 8 (2), 93–119

Castro, S., Bonanata. J. and Rosá, A. (2018). A High Coverage Method for Automatic False Friends Detection for Spanish and Portuguese. Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, 29–36.

Domínguez, P. (2008). Semantics and Pragmatics of False Friends. New York, London: Routledge.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.
Hefler, M. (2017). False Friends between English and Italian. Submitted in partial fulfilment of the requirements for the B.A. in English Language and Language, University of Rijeka.

Paiva, V. D., Rademaker, A., & Melo, G. D. (2012). OpenWordNet-PT: An open Brazilian wordnet for reasoning. COLING 2012.

Vossen, P. (2002). EuroWordNet: general document.