A High Coverage Method for Automatic False Friends Detection for Spanish and Portuguese

Santiago Castro **Jairo Bonanata** **Aiala Rosá**
Grupo de Procesamiento de Lenguaje Natural
Universidad de la República — Uruguay
{sacastro,jbonanata,aialar}@fing.edu.uy

Abstract

False friends are words in two languages that look or sound similar, but have different meanings. They are a common source of confusion among language learners. Methods to detect them automatically do exist, however they make use of large aligned bilingual corpora, which are hard to find and expensive to build, or encounter problems dealing with infrequent words. In this work we propose a high coverage method that uses word vector representations to build a false friends classifier for any pair of languages, which we apply to the particular case of Spanish and Portuguese. The required resources are a large corpus for each language and a small bilingual lexicon for the pair.

1 Introduction

Closely related languages often share a significant number of similar words which may have different meanings in each language. Similar words with different meanings are called *false friends*, while similar words sharing meaning are called *cognates*. For instance, between Spanish and Portuguese, the amount of cognates reaches the 85% of the total vocabulary (Ulsh, 1971). This fact represents a clear advantage for language learners, but it may also lead to an important number of interferences, since similar words will be interpreted as in the native language, which is not correct in the case of false friends.

Generally, the expression false friends refers not only to pairs of identical words, but also to pairs of similar words, differing in a few characters. Thus, the Spanish verb *halagar* ("to flatten") and the similar Portuguese verb *alagar* ("to flood") are usually considered false friends.

Besides traditional false friends, that are similar words with different meanings, Humblé (2006) analyses three more types. First, he mentions words with similar meanings but used in different contexts, as *esclarecer*, which is used in a few contexts in Spanish (*esclarecer un crimen*, "clarify a crime"), but not in other contexts where *aclarar* is used (*aclarar una duda*, "clarify a doubt"), while in Portuguese *esclarecer* is used in all these contexts. Secondly, there are similar words with partial meaning differences, as *abrigo*, which in Spanish means "shelter" and "coat", but in Portuguese has just the first meaning. Finally, Humblé (2006) also considers false friends as similar words with the same meaning but used in different syntactic structures in each language, as the Spanish verb *hablar* ("to speak"), which does not accept a sentential direct object, and its Portuguese equivalent *falar*, which does (**yo hablé que . . . / eu falei que . . .* , *"I spoke that . . . "). These non-traditional false friends are more difficult to detect by language learners than traditional ones, because of their subtle differences.

Having a list of false friends can help native speakers of one language to avoid confusion when speaking and writing in the other language. Such a list could be integrated into a writing assistant to prevent the writer when using these words. For Spanish/Portuguese, in particular, while there are printed dictionaries that compile false friends (Otero Brabo Cruz, 2004), we did not find a complete digital false friends list, therefore, an automatic method for false friends detection would be useful. Furthermore, it

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 29–36
Santa Fe, New Mexico, USA, August 20, 2018.

is interesting to study methods which could generate false friends lists for any pair of similar languages, particularly, languages for which this phenomenon has not been studied.

In this work we present an automatic method for false friends detection. We focus on the traditional false friends definition (similar words with different meanings) because of the dataset we count with and also to present our method in a simple context. We describe a supervised classifier we constructed to distinguish false friends from cognates based on word embeddings. Although for the method development and evaluation we used Spanish and Portuguese, the method could be applied to other language pairs, provided that the resources needed for the method building are available. We do not deal with the problem of determining if two words are similar or not, which is prior to the issue we tackle.

The paper is organized as follows: in Section 2 we describe some related work, in Section 3 we introduce the word embeddings used in this work, in Section 4 we describe our method, in Section 5 we present and analyze the experiments carried out. Finally, in Section 6, we present our conclusions and sketch some future work.

2 Related Work

Previous work use a combination of orthographic, syntactic, semantic and frequency-based features. Frunza (2006) worked with French and English, focusing only on orthographic features via a supervised machine learning algorithm. While this method can work in some cases — e.g. to detect true cognates with a common root, such as *inaccesible* in Spanish and *inacessível* in Portuguese (“inaccessible”), that come from the Latin word *inaccessibilis* — it does not take into account the meanings of the words.

Mitkov et al. (2007) used both a distributional and taxonomy-based approach to multiple language pairs: English–French, English–German, English–Spanish and French–Spanish. For the former approach, they build vectors based on the words that appear in a window in the corpus, computing the co-occurrence probability. Then they defined two methods for classification: one that considers the N nearest neighbors for each word in the pair and computes the Dice coefficient to determine the similarity between both[1], and another one that is similar but using syntactically related words instead of the adjacent words. Additionally, they evaluated a method which uses a taxonomy to classify false friends, and fails back to the distributional similarity for words not included in the taxonomy. They achieved better results under this experiment than only using the distributional similarity. Based on the former technique, Ljubešic et al. (2013) focused on detecting false friends in closely related languages: Slovene and Croatian. Likewise, they exploited a distributional technique but also propose the use of Pointwise Mutual Information (PMI) as an effective way to classify false friends via the frequencies in the corpora.

Sepúlveda and Aluísio (2011) tackled this task for Portuguese and Spanish, taking the same orthographic approach as Frunza (2006). Nonetheless, they carried out an additional experiment in which they added a new feature whose value is the likelihood of one of the words of the pair to be a translation of the other one. This number was obtained from a probabilistic Spanish-Portuguese dictionary, previously generated taking a large sentence-aligned bilingual corpus.

3 Word Vector Representations

As seen in the previous section, some authors (Mitkov et al., 2007; Ljubešic et al., 2013) represented words as vectors by counting occurrences or by building tf–idf vectors, among other techniques. Similarly, Mikolov et al. (2013a) proposed an unsupervised technique, known as *word2vec*, to efficiently represent words as vectors from a large unlabeled corpus, which has proven to outperform several other representations in tasks involving text as input (LeCun et al., 2015). As it is a vector-based distributional representation technique, it is based on computing a vector space in which vectors are close if their corresponding words appear frequently in the same contexts in the corpus used to train it. Interesting relationships and patterns are learned in particular with this method, e.g. the result of the vector calculation $vector(``Madrid'') - vector(``Spain'') + vector(``France'')$ is closer to $vector(``Paris'')$ than to any other word vector (Mikolov et al., 2013a). Additionally, Mikolov et al. (2013c) has shown a technique

[1] Note that for this approach a bilingual dictionary is needed.

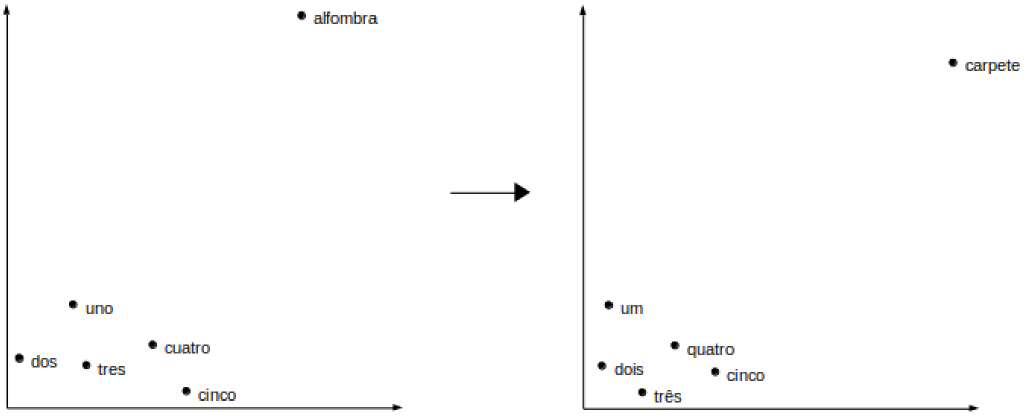

Figure 1: Example showing word2vec properties. The 2D graphs represent Spanish and Portuguese word spaces after applying PCA, scaling and rotating to exaggerate the similarities and emphasize the differences. The left graph is the source language vector space (in this case Spanish) and the right one is the target language vector space (Portuguese).

to detect common phrases such as "New York" to be part of the vector space, being able to detect more entities and at the same time enhancing the context of others.

To exploit multi-language capabilities, Mikolov et al. (2013b) developed a method to automatically generate dictionaries and phrase tables from small bilingual data (translation word pairs), based on the calculation of a linear transformation between the vector spaces built with word2vec. This is presented as an optimization problem that tries to minimize the sum of the Euclidean distances between the translated source word vectors and the target vectors of each pair, and the translation matrix is obtained by means of stochastic gradient descent. We chose this distributional representation technique because of this translation property, which is what our method is mainly based on.

These concepts around word2vec are shown in Fig. 1. In the example, the five word vectors corresponding to the numbers from "one" to "five" are shown, and also the word vector "carpet" for each language. More related words have closer vectors, while unrelated word vectors are at a greater distance. At the same time, groups of words are arranged in a similar way, allowing to build translation candidates.

4 Method Description

As false friends are word pairs in which one seems to be a translation of the other one, our idea is to compare their vectors using Mikolov et al. (2013b) technique. Our hypothesis is that a word vector in one language should be close to the cognate word vector in another language when it is transformed using this technique, but far when they are false friends, as described hereafter.

First, we exploited the Spanish and Portuguese Wikipedia's (containing several hundreds of thousands of words) to build the vector spaces we needed, using Gensim's skip-gram based word2vec implementation (Řehůřek and Sojka, 2010). The preprocessing of the Wikipedia's involved the following steps. The text was tokenized based on the alphabet of each language, removing words that contain other characters. Numbers were converted to their equivalent words. Wikipedia non-article pages were removed (e.g. disambiguation pages) and punctuation marks were discarded as well. Portuguese was harder to tokenize provided that the hyphen is widely used as part of the words in the language. For example, *bem-vindo* ("welcome") is a single word whereas *Uruguai-Japão* ("Uruguay-Japan") in *jogo Uruguai-Japão* ("Uruguay-Japan match") are two different words, used with an hyphen only in some contexts. The right option is to treat them as separate tokens in order to avoid spurious words in the model and to provide more information to existing words (*Uruguai* and *Japão*). As the word embedding method exploits the text at the level of sentences (and to avoid splitting ambiguous sentences), paragraphs were used as sentences, which still keep semantic relationships. A word had to appear at least five times in the corresponding Wikipedia to be considered for construction of the vector space.

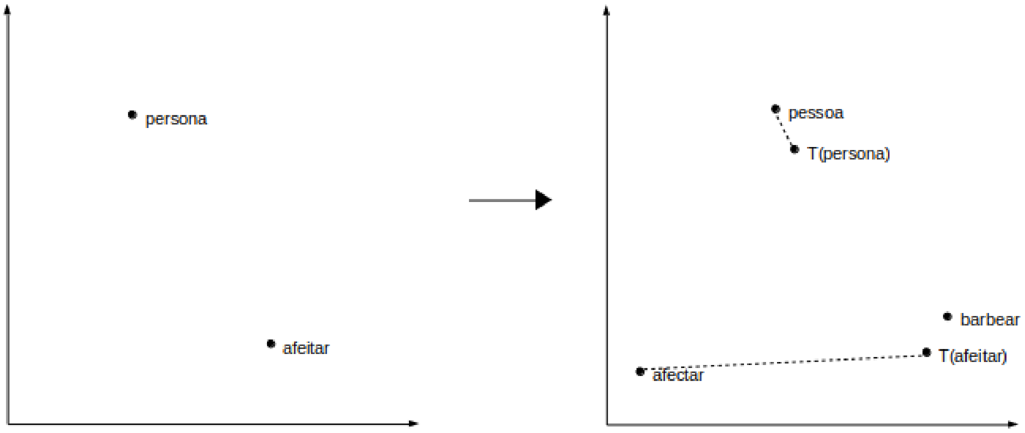

Figure 2: Example showing our method's main idea. The 2D graphs represent the word spaces after applying PCA, scaling and rotating to emphasize the differences. The left graph is the source language vector space (in this case Spanish) and the right one is the target language vector space (Portuguese).

Secondly, WordNet (Fellbaum, 1998) was used as the bilingual lexicon to build the linear transformation between the vector spaces by applying the same technique described in (Mikolov et al., 2013b), taking advantage of the multi-language synset alignment available in NLTK (Bird et al., 2009) between Spanish (Gonzalez-Agirre et al., 2012) and Portuguese (de Paiva and Rademaker, 2012), based on Open Multilingual WordNet (Bond and Paik, 2012). We generated this lexicon by iterating through each of the 40,000 WordNet synsets and forming pairs taking their most common Spanish word and Portuguese word. Note that this is a small figure compared with the corpus sizes, and we show in the next section that it could be considerably lower. We also show that the transformation source needs not to be WordNet (we used it just for convenience), which is an expensive and carefully handcrafted resource; it could be just a bilingual dictionary.

Finally, we defined a method to distinguish false friends from cognates. We defined a binary classifier for determining the class, *false friends* or *cognates*, for each pair of similar words.

Given a candidate pair $(source\_word, target\_word)$, and the corresponding vectors $(source\_vector, target\_vector)$, the first step consists of transforming $source\_vector$ to the space computed for the target language, using the transformation described above. Let $T(source\_vector)$ be the result of this transformation.

Then, to determine if $source\_word$ and $target\_word$ are cognates (if one of them is a possible translation of the other one), we analyzed the relationship between $T(source\_vector)$ and $target\_vector$. According to Mikolov et al. (2013b), the transformation we compute between the vector spaces keeps semantic relations between words from the source space to the target space. So, if $(source\_word, target\_word)$ is a pair of cognates, then $T(source\_vector)$ should be close to $target\_vector$. Otherwise, $source\_word$ and $target\_word$ are false friends.

The method is illustrated in Fig. 2. In the example, the pair $(persona, pessoa)$ are cognates (meaning "person" in English) while the pair $(afeitar, afectar)$ are false friends (meaning "to shave" and "that affects", respectively). If we transform the source word vectors ($persona$ and $afeitar$) and thus obtain vectors in the target vector space, $T(persona)$ and $pessoa$ are close while $T(afeitar)$ and $afectar$ are far from each other (while a valid translation of $afeitar$, $barbear$, is close to $T(afeitar)$).

Following this idea, a threshold needs to be established by which two words are considered cognates. In addition to this, we wanted to see if similar properties help to constitute an acceptable division. Hence, we trained and tested by means of cross-validation a supervised binary Support Vector Machines classifier, based on three features:

- Feature 1: the cosine distance between $T(source\_vector)$ and $target\_vector$.
- Feature 2: the number of word vectors in the target vector space closer to $target\_vector$ than

$T(source\_vector)$, using the cosine distance. We believe that in some cases the distance for cognates may be larger but what it counts is if the transformed vector lays within the closest ones to the target vector.

- Feature 3: the sum of the distances between $target\_vector$ and $T(source\_vector_i)$ for the five word vectors $source\_vector_i$ nearest to $source\_vector$, using the cosine distance. The idea here is that the first feature may be error prone since it only considers one vector, so considering more vectors (by taking both the context from the source vector and the one from its transformed vector) should reduce the variance, as neighbor word vectors from the source word should be neighbors of the target word.

We carried out different experiments alternating the language we used as the source and the language we used as the target, and also other parameters, which we show in the next section.

The source code is public and available to use.[2]

5 Experimental Analysis

Unfortunately, we are not able to compare our method to several others presented by other authors as they are not only based on non-public code, but also on non-public datasets which are not directly comparable with the one used here. Nevertheless, we compare our technique against several methods, for the particular case of Spanish and Portuguese and show it is solid. First, we set a simple baseline that does the following: it checks if there exist a WordNet synset which contains both pair words within the Spanish and Portuguese words of it, and if it is does, then they are considered cognates. Then, we compare to the Machine Translation software Apertium[3]: we take one of the pair words, translate it and check if the translation matches the other word. We chose this software since it can be accessed offline and it is freely available. Apart from this, we compare with Sepúlveda and Aluísio (2011, experiment 2 and 3.2) method and also with a variant of our method that adds a word frequency feature (the relative number of times each word appeared in the corpus). Word frequencies are used by other authors and we believe they are a different data source from what the word2vec vectors can provide.

For these experiments we use the same data set as in (Sepúlveda and Aluísio, 2011).[4] This resource is composed by 710 Spanish-Portuguese word pairs: 338 cognates and 372 false friends. The word pairs were selected from the following resources: an online Spanish-Brazilian Portuguese dictionary, an online Spanish-Portuguese dictionary, a list of the most frequent words in Portuguese and Spanish and an online list of different words in Portuguese and Spanish. There are not multi-word expressions and roughly half of the pairs are composed of identically spelled words. It was annotated by two people.

It is important to consider that the word coverage is a concern in this task since every method can only works when the pair words are present in their resources (in other words, they are not out of a method's vocabulary). The accuracy thus only takes into account the covered pairs. The coverage for the simple baseline can be measured by counting the pairs were both words are present in WordNet. Sepúlveda and Aluísio (2011, experiment 2) only considers orthographic and phonetic differences, so always covers all pairs. Sepúlveda and Aluísio (2011, experiment 3.2) uses a dictionary, then the pairs that are in it count towards the coverage. The words that could not be translated by Apertium are counted against the coverage of its related method. Finally, the pairs that cannot be translated into vectors are counted as not covered by our methods.

Results are shown in Table 1. It can be appreciated that our method provides both high accuracy and coverage, and that word embedding information can be further improved if additional information, such as the word frequencies, is included. We also tested a version of our method that only uses Feature 1 via logistic regression, which reduced the accuracy by 3% roughly, showing that the other two features add some missing information to improve the accuracy. As an additional experiment, we tried exploiting

[2] https://github.com/pln-fing-udelar/false-friends

[3] https://www.apertium.org

[4] This data set is available at http://ec.europa.eu/translation/portuguese/magazine/documents/folha47_lista_pt.pdf

Method	Accuracy	Coverage
WN Baseline	68.18	55.38
Sepúlveda 2	63.52	100.00
Sepúlveda 3.2	76.37	59.44
Apertium	77.75	66.01
Our method	77.28	97.91
With frequencies	79.42	97.91

Table 1: Results (%) obtained by the different methods. *WN Baseline* and *Apertium* methods were measured using the whole dataset, whereas our method's evaluation was carried out with a five-fold cross-validation.

Method	Accuracy
es-400-100-1	**77.28**
es-800-100-1	76.99
es-100-100-1	76.98
es-200-100-1	76.84
es-200-200-1	76.55
pt-200-200-1	76.13
es-200-800-1	75.99
pt-400-100-1	75.99
pt-100-100-1	75.84
es-100-200-1	75.83
es-100-100-2	74.98

Table 2: Results obtained under different configurations. The method name complies with the format: `[source language]-[Spanish vectors dimension]-[Portuguese vectors dimension]-[phrases max size]`. All configurations present the same coverage as before.

WordNet to compute taxonomy-based distances as features in the same manner as Mitkov et al. (2007) did, but we did not obtain a significant difference, thus we conclude that it does not add information to what already lays in the features built upon the embeddings.

As Mikolov et al. (2013b) did, we wondered how our method works under different vector configurations, hence we carried out several experiments, varying vector space dimensions. We also experimented with vectors for phrases up to two words. Finally, we evaluated how the election of the source language, Spanish or Portuguese, affects the results. Accuracy obtained for the ten best configurations, and for the experiment with two word vectors are presented in Table 2. For the experiment we used the vector dimensions 100, 200, 400 and 800; source vector space Spanish and Portuguese; and we also tried with a single run with two-word phrases (with Spanish as source and 100 as the vector dimension), summing up 33 configurations in total. As it can be noted, there are no significant differences in the accuracy of our method when varying the vector sizes. Higher dimensions do not provide better results and they even worsen when the target language dimension is greater than or equal to the source language dimension, as Mikolov et al. (2013b) claimed. Taking Spanish as the source language seems to be better, maybe this is due to the corpus sizes: the corpus used to generate the Spanish vector space is 1.4 times larger than the one used for Portuguese. Finally, we can observe that including vectors for two-word phrases does not improve results.

5.1 Linear Transformation Analysis

We were intrigued in knowing how different qualities and quantities of bilingual lexicon entries would affect our method performance. We show how the accuracy varies according to the bilingual lexicon size and its source in the Fig. 3. *WN* seems to be slightly better than using *Apertium* as source, albeit they both perform well. Also, both rapidly achieve acceptable results, with less than a thousand entries, and

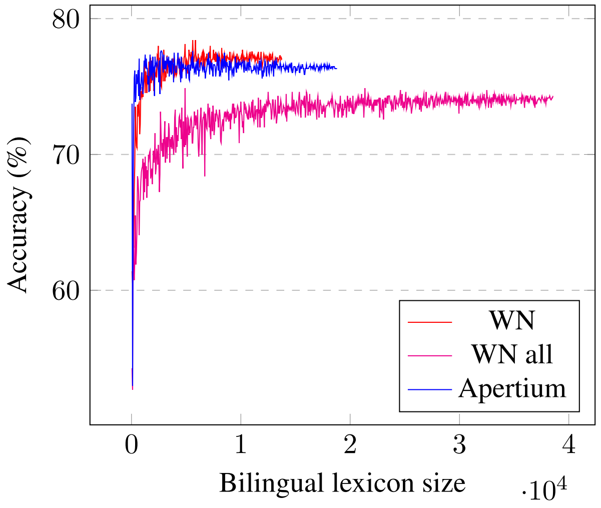

Figure 3: Accuracy of our method with respect to different bilingual lexicon sizes and sources. *WN* is the original approach we take to build the bilingual lexicon, *WN all* is a method that takes every pair of lemmas from both languages in every WordNet synset and *Apertium* uses the translations of the top 50,000 Spanish words in frequencies from the Wikipedia (and that could be translated to Portuguese). Note that the usage of Apertium here has nothing to do with *Apertium* baseline.

yield stable results when the number of entries is larger. This is not the case for the method *WN all*, which needs more word pairs to achieve reasonable results (around 5,000) and it is less stable with larger number of entries.

Even though we use WordNet to build the lexicon, which is a rich and expensive resource, it could also be built with less quality entries, such as those that come from the output of a Machine Translation software or just by having a list of known word translations. Furthermore, our method proved to work with a small number of word pairs, it can be applied to language pairs with scarce bilingual resources.

Additionally, it is interesting to observe that despite the fact that some test set pairs may appear in the bilingual lexicon in which our method is based on, when having changed it (by reducing its size or using Apertium), it still shows great performance. This suggest the results are not biased towards the test set used in this work.

6 Conclusions and Future Work

We have provided an approach to classify false friends and cognates which showed to have both high accuracy and coverage, studying it for the particular case of Spanish and Portuguese and providing state-of-the-art results for this pair of languages. Here we use up-to-date word embedding techniques, which have shown to excel in other tasks, and which can be enriched with other information such as the words frequencies to enhance the classifier. In the future we want to experiment with other word vector representations and state-of-the-art vector space linear transformation such as (Artetxe et al., 2017; Artetxe et al., 2018). Also, we would like to work on fine-grained classifications, as we mentioned before there are some word pairs that behave like cognates in some cases but like false friends in others.

Our method can be applied to any pair of languages, without requiring a large bilingual corpus or taxonomy, which can be hard to find or expensive to build. In contrast, large untagged monolingual corpora are easily obtained on the Internet. Similar languages, that commonly have a high number of false friends, can benefit from the technique we present in this document, for example by generating a list of false friends pairs automatically based on words that are written in both languages in the same way.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Oana Magdalena Frunza. 2006. *Automatic identification of cognates, false friends, and partial cognates*. Ph.D. thesis, University of Ottawa (Canada).

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Philippe Humblé. 2006. Falsos cognados. falsos problemas. un aspecto de la enseñanza del español en brasil. *Revista de Lexicografía*, 12:197–207.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Nikola Ljubešic, Ivana Lucica, and Darja Fišer. 2013. Identifying false friends between closely related languages. *ACL 2013*, page 69.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Workshop at International Conference on Learning Representations*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine translation*, 21(1):29.

María de Lourdes Otero Brabo Cruz. 2004. Diccionario de falsos amigos (español-portugués / portugués-español): Propuesta de utilización en la enseñanza del español a luso hablantes. In *Actas del XV Congreso Internacional de Asele, Sevilla*, pages 632–637. Universidad de Sevilla.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Lianet Sepúlveda and Sandra María Aluísio. 2011. Using machine learning methods to avoid the pitfall of cognates and false friends in spanish-portuguese word pairs. In *8th Brazilian Symposium in Information and Human Language Technology*, pages 67–76.

Jack L Ulsh. 1971. From spanish to portuguese. Washington DC: Foreign Service Institute.