Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3001–3007
Marseille, 11–16 May 2020

Automatically Building a Multilingual Lexicon of False Friends With No Supervision

Ana-Sabina Uban, Liviu P. Dinu
Human Language Technologies Research Center,
Faculty of Mathematics and Computer Science, University of Bucharest
ana.uban@gmail.com, ldinu@fmi.unibuc.ro

Abstract

Cognate words, defined as words in different languages which derive from a common etymon, can be useful for language learners, who can leverage the orthographical similarity of cognates to more easily understand a text in a foreign language. Deceptive cognates, or false friends, do not share the same meaning anymore; these can be instead deceiving and detrimental for language acquisition or text understanding in a foreign language. We use an automatic method of detecting false friends from a set of cognates, in a fully unsupervised fashion, based on cross-lingual word embeddings. We implement our method for English and five Romance languages, including a low-resource language (Romanian), and evaluate it against two different gold standards. The method can be extended easily to any language pair, requiring only large monolingual corpora for the involved languages and a small bilingual dictionary for the pair. We additionally propose a measure of "falseness" of a false friends pair. We publish freely the database of false friends in the six languages, along with the falseness scores for each cognate pair. The resource is the largest of the kind that we are aware of, both in terms of languages covered and number of word pairs.

Keywords: cognates, false friends, database, language acquisition, word embeddings, low-resource languages, unsupervised

1. Introduction

Cognates are words in genetically related languages (languages descending from a common ancestor) with a common proto-word. For example, the Romanian word *victorie* and the Italian word *vittoria* are cognates, as they both descend from the Latin word *victoria* (meaning *victory*) – see Figure 1. Cognates can be useful for language learners, who can have a prior understanding of what a word in a foreign language means only through knowing its cognate: a Romanian speaker can easily learn the Italian word *vittoria*, and, for related languages where much of the vocabulary consists of cognates, non-native speakers of the related language can even achieve basic understanding of texts in the foreign language without ever having studied it.

In most cases, cognates have preserved similar meanings across languages, but there are also exceptions. In some cases, the meanings of cognates has diverged from the common etymon through their use in each of the two languages, and their meanings became different from each other. These are called *deceptive cognates* or, more commonly, *false friends*. Here we use the definition of cognates that refers to words with similar appearance and some common etymology, while *true cognates* is used to refer to cognates which also have a common meaning, and *deceptive cognates* or *false friends* refers to cognate pairs which do not have the same meaning (anymore).

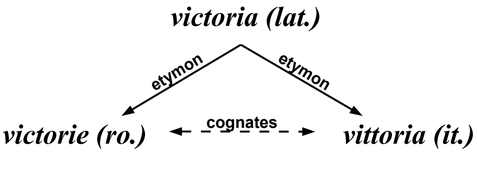

Figure 1: Example of cognates and their common ancestor.

Dominguez and Nerlich (2002) distinguish between *chance false friends*, which have similar form but different etymologies as well as different meanings in different languages, and *semantic false friends*, which share the etymological origin, but their meanings differ (to some extent) in different languages. In this study we focus on semantic false friends (or deceptive cognates), by leveraging a previously published resource of cognate words extracted from etymological dictionaries. Semantic false friends are more common and often more difficult to correctly detect, nevertheless the same method can in principle be used to identify chance false friends.

Many false friends have diverged into entirely different meanings. There are many examples, however, for which the changes in meaning are more subtle, at the level of connotations, and more difficult to detect even for humans. The notion of semantic equivalence used to define false friends is in itself ambiguous and difficult to treat as a binary property, and we propose in this paper that the quality of a cognate pair of being in a *false friends* relationship should also be treated as a spectrum.

According to Uban et al. (2019), a *hard false friend* is a pair of cognates for which the meanings of the two words have diverged enough such that they no longer have the same meaning, and should not be used interchangibly (as translations of one another). In this category fall most known examples of false friends, such as the French-English cognate pair *attendre* / *attend*: in French, *attendre* has a completely different meaning than in English, which is *to wait*. A different and more subtle type of false friends can result from more minor semantic shifts between the cognates. In such pairs, the meaning of the cognate words may remain roughly the same, but with a difference in nuance or connotation. Such an example is the Romanian-Italian cognate pair *amic* / *amico*. Here, both cognates mean *friend*, but in Italian the conotation is that of a closer friend, whereas the Romanian *amic* denotes a more distant friend, or even aquaintance. A more suitable Romanian translation for *am-*

ico would be *prieten*, while a better translation in Italian for *amic* could be *conoscente*.

In the case of soft false friends, their meaning is roughly the same (and since they can be considered ”partial” translations, they can even be found in bilingual dictionaries as correct translations), nevertheless, translating one word for the other would be an inaccurate use of the language. In these cases, instead of helping non-natives to more easily understand a text in a foreign language, cognates can instead cause more confusion and deceive the language learner into misunderstanding the text. This is why having a method to correctly detect even the more subtle kind of false friends is very important for assisting with language learning and text comprehension in a foreign language.

Moreover, identifying false friends can be useful not only for language acquisition, but also in downstream applications relying on cognates, such as machine translation.

In this paper we propose using a fully automatic and unsupervised algorithm in order to detect false friends, and we generate a lexicon of false friends, along with falseness scores for each pair, for every language pair among six languages considered (Romance languages and English). Our method is based on the false friend detection algorithm relying on cross-lingual word embeddings introduced in Uban et al. (2019), to which we add a more extensive evaluation of the resulted false friends pairs, including the extended list of over 3,000 cognate sets (instead of the smaller 305 words list evaluated in the previous study) and additionally include an evaluation and analysis of the falseness measure. We publish freely the resulted database comprised of false friend pairs for each pair of languages considered, and the falseness score for each pair.

2. Related Work

A comprehensive list of cognates and false friends for every language pair is difficult to find and expensive to manually build. Moreover, dictionaries grow outdated and it is difficult to continuously update them to incorporate new words in the vocabulary. This is why applications have to rely on automatically identifying false friends.

There have been a number of previous studies attempting to automatically extract pairs of true cognates and false friends from corpora or from dictionaries. Most methods are based either on orthographic and phonetic similarity (Inkpen et al., 2005), or require large parallel corpora (Nakov et al., 2009) or dictionaries (St Arnaud et al., 2017). Inkpen et al. (2005) use orthographic features to extract cognate pairs for French-English, but do not take semantic similarity into account. Torres and Aluísio (2011) also rely on orthographic and phonetic features, to which they add a semantic feature extracted from a bilingual dictionary. They additionally release a lexicon of Spanish-Portuguese false friends and true cognates, obtained through manual annotation, that they use to evaluate their algorithms. Nakov et al. (2009) identify false friends pairs in Bulgarian and Russian by making use of sentence-aligned parallel corpora. In (Aminian et al., 2015) the authors propose using a model of identifying false friends from parellel corpora in order to improve English-Egyptian statistical machine translation.

Cross-lingual semantic word similarity consists in identifying words that refer to similar semantic concepts and convey similar meanings across languages (Vulic and Moens, 2013). Some of the most popular approaches rely on probabilistic models (Vulic and Moens, 2014) and cross-lingual word embeddings (Søgaard et al., 2017). There have been few previous studies using word embeddings for the detection of false friends or cognate words, usually using simple methods on only one or two pairs of languages (Castro et al., 2018; Torres and Aluísio, 2011). Castro et al. (2018) detect false friends in Spanish-Portuguese, employing a classifier that learns from features extracted from multilingual embedding spaces. In (Mitkov et al., 2007) the authors use a method based on distributed representations of words in a continuous space built using comparable corpora, as well as a taxonomy-based approach, to identify false friends in four language pairs involving English, French, German and Spanish.

Uban et al. (2019) propose a method for identifying and correcting false friends, as well as define a measure of their “falseness”, using cross-lingual word embeddings. We base our study on the method proposed here.

3. Detecting False Friends

In the following section we describe the algorithm used for detecting false friends automatically, in an unsupervised manner, based on a seed set of cognate sets.

3.1. Cross-lingual Word Embeddings

Word embeddings are vectorial representations of words in a continuous space, built by training a model to predict the occurrence of a target word in a text corpus given its context. Based on the distributional hypothesis stating that similar words occur in similar contexts, these vectorial representations can be seen as semantic representations of words and can be used to compute semantic similarity between word pairs (representations of words with similar meanings are expected to be close together in the embeddings space).

In our study we make use of word embeddings computed using the FastText algorithm, pre-trained on Wikipedia for the six languages in question. The vectors have 300 dimensions, and were obtained using the skip-gram model described by Bojanowski et al. (2016) with default parameters. These pre-trained embeddings are suitable for our multilingual study since: they are trained on large amounts of text, which minimizes the amount of noise in the vectors, making them good approximators of word meanings; and they are trained on text that is relatively uniform in style and topic - ensuring any differences in the structure of the embedding spaces of different languages is dependent on the language, rather than an artifact of the topic or genre of the corpus. Nevertheless, even high quality embeddings can be noisy or biased and this should be kept in mind when interpreting the results of our experiments.

To compute the semantic divergence of cognates across sister languages, we need to obtain a multilingual semantic space, which is shared between the cognates. Having the representations of both cognates in the same semantic space, we can then compute the semantic distance between them using their vectorial representations in this space. For a given pair of languages among the six considered, we can then accomplish this following the steps below:

1. **Step 1.** Train a word embeddings model for each of the two languages.

2. **Step 2.** Obtain a shared embedding space, common to the two languages. This is accomplished using an alignment algorithm, which consists of finding a linear transformation between the two spaces that on average optimally transforms each vector in one embedding space into a vector in the second embedding space, minimizing the distance between a few seed word pairs (which are assumed to have the same meaning), based on a small bilingual dictionary. The linear nature of the transformation guarantees distances between words in the original spaces (within each language) are preserved. For our purposes, we use the publicly available FastText multilingual word embeddings pre-aligned in a common vector space (Conneau et al., 2017).[1]

3. **Step 3.** Compute the semantic distance for the pair of cognates in the two languages, using a vectorial distance (we chose cosine distance) on their corresponding vectors in the shared embedding space.

3.2. Cognates Dataset

In order to identify false friends (deceptive cognates), we start from a database of cognates, defined as words with common etymology and similar orthography. As our data source, we use the list of cognate sets in Romance languages published in (Ciobanu and Dinu, 2014). It contains 3,218 complete cognate sets in Romanian, French, Italian, Spanish and Portuguese, along with their Latin common ancestors, extracted from online etymology dictionaries. A subset of 305 of these sets also contains the corresponding cognate (in the broad sense, since these are mostly borrowings) in English.

One complete example of a cognate set for the word “architect” in the Romance languages is illustrated in Table 1.

3.3. Deceptive Cognates and Falseness

The multilingual embedding spaces as defined above can be used to measure the semantic distances between cognates in order to detect pairs of false friends, which are simply defined as pairs of cognates which do not share the same meaning. More specifically, following the false friends detection and correction algorithm in Uban et al. (2019), we consider a pair of cognates to be a false friend pair if in the shared semantic space, there exists a word in the second language which is semantically closer to the original word than its cognate in that language (in other words, the cognate is not the optimal translation). The arithmetic difference between the semantic distance between these words and the semantic distance between the cognates will be used as a measure of the *falseness* of the false friend. The closest neighbor of the target word in the second language's embedding space is considered to be the correct semantic equivalent of the target word in the second language, and can be provided as a ”correction” for the false friend.

The algorithm for detecting (and correcting) false friends, as well as measuring their degree of falseness, can be described as shown in Algorithm 1.

Algorithm 1 Detection and correction of false friends

```
Given the cognate pair $(c_1, c_2)$ where $c_1$ is a word in $lang_1$ and $c_2$ is a word in $lang_2$:
Find the word $w_2$ in $lang_2$ such that for any $w_i$ in $lang_2$, $distance(c_1, w_2) < distance(c_1, w_i)$
if $w_2 \neq c_2$ then
    $(c_1, c_2)$ is a pair of false friends
    Degree of falseness = $distance(c_1, w_2) - distance(c_1, c_2)$
    return $w_2$ as potential correction
end if
```

4. False Friends Dataset

We use the algorithm described in the previous section to build a database of false friends pairs for each language pair among the six languages considered, which we make freely available [2].

False friends for Romance languages are extracted from the original 3,218 cognate sets, resulting in 500 to 1,200 of detected false friends for each language pair. For English, the original cognate resource contains a smaller set of only 305 cognate sets, which results in smaller false friends lists for language pairs involving English.

Table 2 shows the number of false friends pairs generated for each language pair, and included in the published resource.

5. Evaluation

In order to evaluate the quality of the false friends dataset generated with our algorithm, we first test its accuracy against a multilingual dictionary, for this study we choose to use Open Multilingual WordNet (Miller, 1998; **?**). A pair of words with common etymology are considered true cognates if they belong to the same WordNet synset (are synonyms), and false friends if they are not synonyms. Using this standard, the obtained measured accuracy is between 73% and 81%, depending on the language pair considered. Table 3 presents a breakdown of the obtained performance per language pair considered. Romanian is the only language missing from the evaluation since it is not represented in multilingual WordNet. Since English is only available for a subset of the cognates, evaluation for Romance languages may be more robust.

We select a few results of the algorithm to show in Table 4, containing examples of extracted false friends, along with the suggested correction and the computed degree of falseness. The tables shows some examples of the algorithm correctly identifying and correcting false friends pairs - such as the Romanian-Italian pairs *tânăr* (meaning *young*)

[1] https://github.com/facebookresearch/MUSE

[2] https://github.com/ananana/false_friends_resource

Romanian	French	Italian	Spanish	Portuguese	Latin
arhitect	architecte	architetto	arquitecto	arquiteto	architectus

Table 1: An example of a cognate set: “architect” in Romance languages.

Languages	FF Pairs	Languages	FF Pairs
ES-IT	739	IT-ES	727
ES-PT	490	PT-ES	502
FR-IT	921	IT-FR	925
FR-ES	886	ES-FR	905
FR-PT	1,023	PT-FR	1,060
IT-PT	795	PT-IT	848
RO-FR	1,258	FR-RO	1,596
RO-IT	1,286	IT-RO	1,654
RO-ES	1,229	ES-RO	1,647
RO-PT	1,227	PT-RO	1,640
EN-PT	148	PT-EN	137
EN-ES	158	ES-EN	136
EN-IT	153	IT-EN	139
EN-FR	150	FR-EN	133
EN-RO	205	RO-EN	161

Table 2: Number of datapoints in false friends database

	Accuracy	Precision	Recall
ES-IT	73.69	43.27	38.06
IT-ES	73.58	43,12	37.73
ES-PT	79.09	36.05	26.49
PT-ES	78.65	32.32	24.35
FR-IT	74.43	33.39	57.40
IT-FR	74.77	34.32	58.68
FR-ES	76.25	42.02	51.94
ES-FR	75.13	40.27	51.78
IT-PT	74.58	33.20	44.73
PT-IT	73.61	31.69	49.31
EN-PT	77.25	59.81	86.48
PT-EN	79.82	64.70	85.71
EN-ES	76.58	63.88	88.46
ES-EN	80.48	71.57	83.95
EN-IT	77.40	61.73	87.65
IT-EN	74.89	61.90	76.47
EN-FR	77.09	57.89	94.28
FR-EN	81.05	66.32	86.66

Table 3: Performance for all language pairs using WordNet as gold standard.

/ *tenero* (meaning *tender*), with the Italian correction *giovane* (*young*), or *inimă* (*heart*) / *anima* (*soul*), corrected to *cuore* (*heart*). The falseness scores also reflect the degree of semantic drift between the false friends, with the *tânăr*/*tenero* pair being more dissimilar than *inimă*/*anima*. The *amic*/*amico*/*amichetto* set, which refers to different degrees of friendship, is awarded the lowest falseness score. It is valuable to note the algorithm also selects word pairs which can technically be considered true cognates (*long*/*luengo* – meaning *long*), but are not used as such in current speech: *largo* is more frequently used than *luengo*. This is to be expected since the algorithm is based on word *usage* in language (since this is the basis of the embedding training algorithm). We also illustrate an example where the algorithm makes a mistake: in the case of *stânga* (*left*)/*stanco* (*tired*), the algorithm rightly identifies this as a false friends pair, but provides an erroneous correction: *destra* is the Italian word for *left*, not *right*. This error can also be traced back to the nature of semantic similarity as captured by word embeddings: related but not equivalent words (and sometimes even antonyms) can have similar embedding vectors due to their similar occurence patterns in corpora.

Cognate	False Friend	Correction	Falseness
long (FR)	luengo (ES)	largo	0.50
face (FR)	faz (ES)	cara	0.39
change(FR)	caer (ES)	cambia	0.46
stânga (RO)	stanco (IT)	destra	0.52
tânăr (RO)	tenero (IT)	giovane	0.41
inimă (RO)	anima (IT)	cuore	0.13
amic (RO)	amico (IT)	amichetto	0.04

Table 4: Extracted false friends and falseness.

In a second experiment, we measure accuracy of false friend detection on a manually curated list of false friends and true cognates in Spanish and Portuguese, used in a previous study (Castro et al., 2018), and introduced in (Torres and Aluísio, 2011). This resource is composed by 710 Spanish-Portuguese word pairs: 338 true cognates and 372 false friends. We also compare our results to the ones reported in this study, which uses a method similar to ours (using a simple classifier that takes embedding similarities as features to identify false friends) and shows improvements over results in previous research. The results are shown in Table 5. We also compute the same metrics using a falseness threshold as a lower bound to decide whether two words are false friends; results show a trade-off between recall and precision when using a threshold. The following section discusses in more detail the use of falseness thresholds.

In this second experiment, WordNet is used as a baseline algorithm for false friend identification instead of a gold standard. Its relatively poor results as compared to the automatic methods may stem from the lower coverage it has as compared to corpus-based methods. Castro et al. (2018) show that only 55% of the word pairs in the evaluation set used here are found in WordNet synsets. This shows that using WordNet as an evaluation standard has its limits, and that corpus-based methods such as the one we propose have an advantage over dictionary-based methods.

5.1. Falseness as a Spectrum

The measure of falseness that we provide for every detected pair of false friends can be useful not only for a better un-

	Accuracy	Precision	Recall
Our method (ft=0)	81.81	78.69	80.80
Our method (ft=0.1)	82.62	92.37	66.06
(Castro et al)	77.28	-	-
(Sepulveda et al)	76.37	-	-
WN Baseline	69.57	85.82	54.50

Table 5: Performance for Spanish-Portuguese using curated false friends test set, compared to previous attempts.

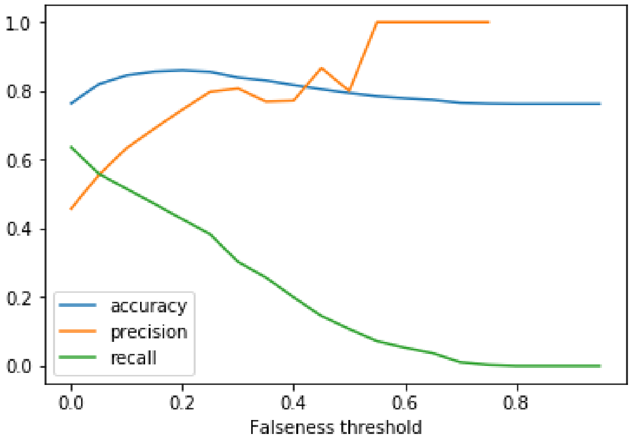

Figure 2: Performance with falseness threshold

derstanding of the linguistic phenomenon behind the semantic divergence of the cognates, but also for a more flexible integration with downstream applications. When our resource of false friends, a custom threshold of falseness could be set and used to filter out false friends in a more coarse or fine-grained way, depending on the needs of the application. For example, for applications where capturing subtle changes in meaning is important, maintaining a low threshold of falseness is useful. On the other hand, when simply identifying false friends which have entirely different meanings, choosing a high threshold may be sufficient and might also ensure a lower rate of false positives, by filtering out the delicate cases of false friend pairs with similar meanings, which lie at the boundary between true and deceptive cognates.

We perform an analysis of the effect of varying the threshold of falseness used to discriminate between true cognates and false friends, by re-evaluating the generated false friends against WordNet using a varying threshold of falseness to discriminate between false friends and true cognates (as opposed to the simple evaluation in the previous chapter, where no threshold was set, which is equivalent to using a falseness threshold of 0). In this way, we are able to discover the optimal falseness threshold to use in order to maximize performance relative to the WordNet standard. We choose the threshold which leads to maximum average accuracy across all language pairs on a separate training set of 80% of the word pairs. The rest of 20% of word pairs are used to evaluate the method now employing a falseness threshold, set to the optimal value according to the training phase. The optimal falseness threshold $ft\star$ is found to be 0.2, and the average overall accuracy with this threshold is 85.85%. Table 6 shows the difference in accuracy when using the optimal threshold, and Figure 2 illustrates the variation of accuracy, precision and recall (on average for all language pairs) when varying the threshold between 0 and 1.

Falseness threshold	0 (None)	0.2 (optimal)
Accuracy	80.57	85.85

Table 6: Best overall accuracy of our method

$$ft\star = argmax_{ft\in(0,1)} \frac{1}{|LP|} \sum_{l_1,l_2\in LP} Acc(ft, l_1, l_2) \quad (1)$$

where ft is a falseness threshold and LP is the set of all language pairs:

$$LP = \{(l_1, l_2) || l_1, l_2 \in \{RO, ES, PT, FR, IT, EN\}\} \quad (2)$$

The fact that in WordNet the optimal falseness is positive means that many of the extracted false friends with very low falseness make up for most of false positives (actual true cognates), which are not identified as such not necessarily because they are actually different in meaning, but rather because of artifacts of the embedding space.

We then perform the same experiment, but this time evaluate using the curated cognates sets in Spanish-Portuguese. In this case, the optimal threshold is found to be 0.1, and the threshold of 0.2 found in the previous experiment leads to worse results than not using a threshold at all. In order to confirm the difference stems from the the different definition of cross-lingual synonymy in the two datasets and is not specific to just the language pair, we compute the optimal falseness threshold relative to WordNet specifically for Spanish-Portuguese and find an optimal falseness of 0.3. This difference between the optimal threshold values for the two different gold standards suggests the two resources were built with different assumptions about meaning equivalence, and confirms that the availability of the falseness measure can be useful for tuning the false friend detection algorithm to the specific task and standards of the particular application.

5.2. Error Analysis and Discussion

As suggested in the previous section, a significant source of error relative to the WordNet standard are low-falseness pairs of detected false friends. Figures 3 and 4 show the distribution of falseness scores across all word pairs in all languages. We separately show the distribution of false friends extracted with our method that were evaluated as actual false friends using WordNet, and the pairs of extracted false friends that are actually true cognates according to WordNet. The much lower falseness values for word pairs in the second category (false positives in the evaluation using WordNet) suggest that many of the false positives produced by the algorithm fall in the range of word pairs with very subtle differences in meaning. These might stem from imperfections in the embedding space or from the too strong

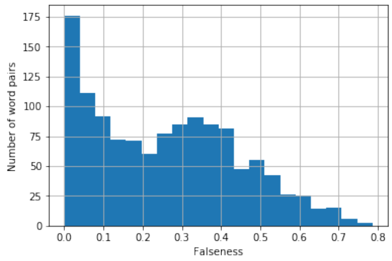

Figure 3: Falseness in correctly detected false friends

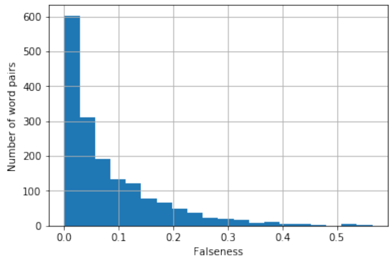

Figure 4: Falseness in incorrectly detected false friends

assumption that the closest word in the multilingual embedding space is the correct translation. Some of the examples in Table 4 illustrate these types of errors, such is the case of the previously discussed pair *stânga* / *stanco*, with the mistaken correction *destra*. More subtle inaccuracies can consist for example of mismatched parts of speech: such as the case of *change* (noun) / *caer* (infinitive verb) / *cambia* (indicative verb).

On the other hand, we believe including the falseness score in the dataset can be useful precisely to remedy this issue when needed, and that in some cases, these low-falseness word pairs could even be considered actual false friends (rather than errors of classification) by a standard of meaning equivalence that is more strict than the one used in WordNet's synsets.

A second source of errors is found in the original cognates data source that we use to then discriminate into true and false cognates. Since it is also an automatically built resource, some of the word pairs are falsely labelled as cognates, and may further perpetuate into false positives in our algorithm.

6. Conclusions

We have built and made freely available a database of false friends in six languages, and evaluated it against WordNet and against a manually curated dataset of false friends, obtaining state of the art results. To the best of our knowledge, the published database is the largest public resource of the kind, both in terms of number of word pairs covered and languages considered. Additionally, the proposed method can be used to generate or detect pairs of false friends for any pair of languages, without requiring expensive manual work or dictionaries, but only large monolingual corpora to train word embeddings on, and small bilingual dictionaries to perform embedding space alignments.

The unsupervised nature of the algorithm proposed also has the advantage of a high coverage of the vocabulary, unlike dictionary-based methods, which are prone to becoming outdated as language evolves. One disadvantage of our embedding-based algorithm is the lack of distinction between different senses of the same word. In the future it would be interesting to continue the study in the direction of considering also context-specific senses of words, in order to be able to better handle partial false friends, which are pairs of cognates which share meaning in some contexts and not others.

Along with false friends pairs, we publish a falseness score for each pair, which can be used to customize the sensitivity to difference in meaning that defines a pair of false friends according to the application. We believe this resource can be very valuable for language learners, for example by incorporating false friends pairs in a tool to aid with language acquisition or text comprehension for non-natives, as well as for machine translation or other applications using natural language processing in a multilingual setting.

7. Bibliographical References

Aminian, M., Ghoneim, M., and Diab, M. (2015). Unsupervised false friend disambiguation using contextual word clusters and parallel word alignments. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 39–48, Denver, Colorado, USA, June. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Castro, S., Bonanata, J., and Rosá, A. (2018). A high coverage method for automatic false friends detection for spanish and portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 29–36.

Ciobanu, A. M. and Dinu, L. P. (2014). Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1038–1043.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dominguez, P. J. C. and Nerlich, B. (2002). False friends: their origin and semantics in some selected languages. *Journal of pragmatics*, 34(12):1833–1849.

Inkpen, D., Frunza, O., and Kondrak, G. (2005). Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

Mitkov, R., Pekar, V., Blagoev, D., and Mulloni, A. (2007). Methods for extracting and classifying pairs of cognates and false friends. *Machine translation*, 21(1):29.

Nakov, S., Nakov, P., and Paskaleva, E. (2009). Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of the International Conference RANLP-2009*, pages 292–298.

Søgaard, A., Goldberg, Y., and Levy, O. (2017). A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017*, pages 765–774.

St Arnaud, A., Beck, D., and Kondrak, G. (2017). Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528.

Torres, L. S. and Aluísio, S. M. (2011). Using machine learning methods to avoid the pitfall of cognates and false friends in spanish-portuguese word pairs. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Uban, A. S., Ciobanu, A., and Dinu, L. (2019). A computational approach to measuring the semantic divergence of cognates. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing*. to be published.

Vulic, I. and Moens, M. (2013). Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 106–116.

Vulic, I. and Moens, M. (2014). Probabilistic Models of Cross-Lingual Semantic Similarity in Context Based on Latent Cross-Lingual Concepts Induced from Comparable Data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 349–362.