Identifying False Friends between Closely Related Languages

Nikola Ljubešić
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr

Darja Fišer
Faculty of Arts
University of Ljubljana
Aškerčeva 2
1000 Ljubljana, Slovenija
darja.fiser@ff.uni-lj.si

Abstract

In this paper we present a corpus-based approach to automatic identification of false friends for Slovene and Croatian, a pair of closely related languages. By taking advantage of the lexical overlap between the two languages, we focus on measuring the difference in meaning between identically spelled words by using frequency and distributional information. We analyze the impact of corpora of different origin and size together with different association and similarity measures and compare them to a simple frequency-based baseline. With the best performing setting we obtain very good average precision of 0.973 and 0.883 on different gold standards. The presented approach works on non-parallel datasets, is knowledge-lean and language-independent, which makes it attractive for natural language processing tasks that often lack the lexical resources and cannot afford to build them by hand.

1 Introduction

False friends are words in two or more languages that are orthographically or semantically similar but do not have the same meaning, such as the noun *burro*, which means *butter* in Italian but *donkey* in Spanish (Allan, 2009). For that reason, they represent a dangerous pitfall for translators, language students as well as bilingual computer tools, such as machine translation systems, which would all benefit greatly from a comprehensive collection of false friends for a given language pair.

False friends between related languages, such as English and French, have been discussed by lexicographers, translators and language teachers for decades (Chacón Beltrán, 2006; Granger and Swallow, 1988; Holmes and Ramos, 1993). However, they have so far played a minor role in NLP and have been almost exclusively limited to parallel data (Inkpen et al., 2005; Nakov and Nakov, 2009). In this paper we tackle the problem of automatically identifying false friends in weakly comparable corpora by taking into account the distributional and frequency information collected from non-parallel texts.

Identifying false friends automatically has the same prerequisite as the problem of detecting cognates – identifying similarly (and identically) spelled words between two languages, which is far from trivial if one takes into account the specificity of inter-language variation of a specific language pair. In this contribution we focus on the problem of false friends on two quite similar languages with a high lexical overlap – Croatian and Slovene – which enables us to circumvent the problem of identifying similarly spelled words and use identical words only as the word pair candidate list for false friends.

Our approach to identifying false friends relies on two types of information extracted from corpora. The first one is the frequency of a false friend candidate pair in the corresponding corpora where the greater the difference in frequency, the more certain one can be that the words are used in different meanings. The second information source is the context from corresponding corpora where the context dissimilarity of the two words in question is calculated through a vector space model.

The paper is structured as follows: in Section 2 we give an overview of the related work. In Section 3 we describe the resources we use and in Section 4 we present the gold standards used for evaluation. Section 5 describes the experimental setup and Section 6 reports on the results. We conclude the paper with final remarks and ideas for future work.

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 69–77,
Sofia, Bulgaria, 8-9 August 2013. ©2010 Association for Computational Linguistics

2 Related Work

Automatic detection of false friends was initially limited to parallel corpora but has been extended to comparable corpora and web snippets (Nakov et al., 2007). The approaches to automatically identify false friends fall into two categories: those that only look at orthographic features of the source and the target word, and those that combine orthographic features with the semantic ones.

Orthographic approaches typically rely on combinations of a number of orthographic similarity measures and machine learning techniques to classify source and target word pairs to cognates, false friends or unrelated words and evaluate the different combinations against a manually compiled list of legitimate and illegitimate cognates. This has been attempted for English and French (Inkpen et al., 2005; Frunza and Inkpen, 2007) as well as for Spanish and Portuguese (Torres and Aluísio, 2011).

Most of the approaches that combine orthographic features with the semantic ones have been performed on parallel corpora where word frequency information and alignments at paragraph, sentence as well as word level play a crucial role at singling out false friends, which has been tested on Bulgarian and Russian (Nakov and Nakov, 2009). Work on non-parallel data, on the other hand, often treats false friend candidates as search queries, and considers the retrieved web snippets for these queries as contexts that are used to establish the degree of semantic similarity of the given word pair (Nakov and Nakov, 2007).

Apart from the web snippets, comparable corpora have also been used to extract and classify pairs of cognates and false friends between English and German, English and Spanish, and French and Spanish (Mitkov et al., 2007). In their work, the traditional distributional approach is compared with the approach of calculating n-nearest neighbors for each false friend candidate in the source language, translating the nearest neighbors via a seed lexicon and calculating the set intersection to the N nearest neighbors of the false friend candidate from the target language.

A slightly different setting has been investigated by Schultz et al. (2004) who built a medical domain lexicon from a closely related language pair (Spanish-Portuguese) and used the standard distributional approach to filter out false friends from cognate candidates by catching orthographically most similar but contextually most dissimilar word pairs.

The feature weighting used throughout the related work is mostly plain frequency with one case of using TF-IDF (Nakov and Nakov, 2007) whereas cosine is the most widely used similarity measure (Nakov and Nakov, 2007; Nakov and Nakov, 2009; Schulz et al., 2004) while Mitkov et al. (2007) use skew divergence which is very similar to Jensen-Shannon divergence.

The main differences between the work we report on in this paper and the related work are:

1. we identify false friends on a language pair with a large lexical overlap – hence we can look for false friends only among identically spelled words, such as *boja*, which means *buoy* in Slovene but *colour* in Croatian, and not among similarly spelled words, such as the Slovene adjective *bučen* (*made of pumpkins* and *noisy*) and its Croatian counterpart *bučan* (only *noisy*);

2. we inspect multiple association and similarity measure combinations on two different corpora pairs, which enables us to assess the stability of those parameters in the task at hand;

3. we work on two different corpora pairs which we have full control over (that is not the case with web snippets), and are therefore able to examine the impact of corpus type and corpus size on the task;

4. we use three categories for the identically spelled words:
   (a) we use the term *true equivalents* (TE) to refer to the pairs that have the same meaning and usage in both languages (e.g. adjective *bivši*, which means *former* in both languages),
   (b) the term *partial false friends* (PFF) describes pairs that are polysemous and are equivalent in some of the senses but false friends in others (e.g. verb *dražiti*, which can mean either *irritate* or *make more expensive* in Slovene but only *irritate* in Croatian), and
   (c) we use the term *false friends* (FF) for word pairs which represent different concepts in the two languages (e.g. noun *slovo*, which means *farewell* in Slovene and *letter of the alphabet* in Croatian)

By avoiding the problem of identifying relevant similarly spelled words prior to the identification of false friends, in this paper we focus only on the latter and avoid adding noise from the preceding task.

3 Resources Used

In this paper we use two types of corpora: Wikipedia corpora (hereafter WIKI) which have gained in popularity lately because of their simple construction and decent size and web corpora (hereafter WAC) which are becoming the standard for building big corpora.

We prepared the WIKI corpora from the dumps of the Croatian and Slovene Wikipedias by extracting their content, tokenizing and annotating them with morphosyntactic descriptions and lemma information. The web corpora of Croatian and Slovene were built in previous work of Ljubešić and Erjavec (2011). They were created by crawling the whole top-level Slovene and Croatian domains and applying generic text extraction, language identification, near-duplicate removal, linguistic filtering and morphosyntactic annotation and lemmatization.

In terms of content, it is to expect that web corpora are much richer genre-wise while articles in Wikipedia corpora all belong to the same genre. As far as topics are concerned, web corpora are believed to be more diverse but contain a less uniform topic distribution than Wikipedia corpora. Finally, it is to expect that Wikipedia corpora contain mostly standard language while web corpora contain a good portion of user-generated content and thereby non-standard language as well.

Some basic statistical information on the corpora is given in Table 1.

CORPUS	MWORDS	MTOKENS	DOC #
HR.WIKI	31.21	37.35	146,737
SL.WIKI	23.47	27.85	131,984
HRWAC	787.23	906.81	2,550,271
SLWAC	450.06	525.55	1,975,324

Table 1: Basic statistics about the corpora used

Both types of corpora are regularly used in today's NLP research and one of the tasks of this paper is to compare those two not only in relation to the specific task of false friends identification, but on a broader scale of exploiting their contextual and frequency information as well.

4 Gold Standards

The gold standards for this research were built from identically spelled nouns, adjectives and verbs that appeared with a frequency equal or higher than 50 in the web corpora for both languages.

The false friend candidates were categorized in the three categories defined in Section 2: false friends, partial false friends and true equivalents.

Manual classification was performed by three annotators, all of them linguists. Since identifying false friends is hard even for a well-trained linguist, all of them consulted monolingual dictionaries and corpora for both languages before making the final decision.

The first annotation session was performed by a single annotator only. Out of 8491 candidates, he managed to identify 117 FFs, 110 PFFs and 8264 (97.3%) TEs. All the identified FFs and PFFs as well as 380 TEs were then given to two more annotators, shrinking the dataset to be annotated by the other two annotators down to 607 entries, i.e. to only 7% of the initial dataset. The agreement between all three annotators on the smaller dataset is given in Table 2.

ANNOTATORS	INTERSECTION	KAPPA
A1 A2	0.766	0.549
A1 A3	0.786	0.598
A2 A3	0.743	0.501
average	0.765	0.546

Table 2: Inter-annotator agreement on building the gold standards

The obtained average kappa inter-annotator agreement is considered moderate and proves the problem to be quite complex, even for humans well trained in both languages with all the available resources at hand. Since we did not have sufficient resources for all the annotators to revise their divergent annotations, we proceeded by building the following two gold standards:

1. the first gold standard (GOLD1) contains only FFs and TEs on which all the three annotators agreed (60 FFs and 324 TEs) and

2. the second gold standard (GOLD2) contains all entries where at least the first and one of the other two annotators agreed (81 FFs, 33 PFFs and 351 TEs).

We consider GOLD1 to be simpler and cleaner while GOLD2 contains the full complexity of the task at hand.

5 Experimental Setup

We experimented with the following parameters: corpus type, corpus size, association measure for feature weighting, similarity measure for comparing context vectors and gold standard type.

We ran our experiments on two pairs of corpora:

1. one pair originating from local Wikipedia dumps (WIKI) and
2. one pair originating from the top-level-domain web corpora of the two languages (WAC)

We took under consideration the following association measures:

1. TF-IDF (TF-IDF) is well known from information retrieval but frequently applied on other problems as well; we consider context vectors to be information entities and calculate the IDF statistic for a term t and vector set V as follows:

$$IDF(t, V) = \log \frac{|V|}{|\{v \in V : t \in v\}|}$$

2. log-likelihood (LL) (Dunning, 1993) which has proven to perform very well in a number of experiments on lexicon extraction i.e. finding words with the most similar context, performing similarity well as TF-IDF and
3. discounted log-odds (LO) first used in lexicon extraction by Laroche and Langlais (2010), showing consistently better performance than LL; it is calculated from contingency table information as follows:

$$LO = \log \frac{(O_{11} + 0.5)(O_{22} + 0.5)}{(O_{12} + 0.5)(O_{21} + 0.5)}$$

The following similarity measures were taken into account:

1. the well-known cosine measure (COSINE),
2. the Dice measure (DICE), defined in (Otero, 2008) as DiceMin, which has proven to be very good in various tasks of distributional semantics (v_{1f} is the feature weight of feature f in vector v_1):

$$DICE(v_1, v_2) = \frac{2 * \sum_f min(v_{1f}, v_{2f})}{\sum_f v_{1f} + v_{2f}}$$

3. and the Jensen-Shannon divergence (JENSHAN) which shows consistent performance on various tasks:

$$JS(v_1, v_2) = \frac{KL(v_1|v_2)}{2} + \frac{KL(v_2|v_1)}{2}$$

$$KL(v_1|v_2) = \sum_f v_{1f} \log \frac{v_{1f}}{v_{1f} + v_{2f}}$$

We used the standard approach for extracting context and building context vectors and calculated the frequency distribution of three content words to the left and to the right of the headword without encoding their position. We did not perform any cross-lingual feature projection via a seed lexicon or similar, but relied completely on the lexical overlap between the two similar languages.

Apart from the context and its dissimilarity, there is another, very fundamental source of information that can be used to assess the difference in usage and therefore meaning – the frequency of the word pair in question in specific languages. That is why we also calculated pointwise mutual information (PMI) between candidate pairs.

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1) * p(w_2)}$$

We estimated the joint probability of the two words by calculating the maximum likelihood estimate of the identically spelled word on the merged corpora. We considered this measure to be a strong baseline. For a weak baseline we took a random ordering of pairs of words (RANDOM).

Since the result of the procedure of identifying false friends in this setting is a single ranked list of lemma pairs where the ranking is performed by contextual or frequency dissimilarity, the same evaluation method can be applied as to evaluating a single query response in information retrieval. That is why we evaluated the output of each setting with average precision (AP), which averages over all precisions calculated on lists of false friend candidates built from each positive example upwards.

As three categories were encoded in the GOLD2 gold standard, we weighted FFs with 1, TEs with 0 and PFFs with 0.5. In the GOLD1 gold standard FFs were, naturally, weighted with 1 and TEs with 0.

6 Results

In our initial set of experiments we ran a Cartesian product on the sets of corpora types, gold standards, association measures and similarity measures. The results of those experiments are given in Table 3.

WIKI			
GOLD1	COSINE	DICE	JENSHAN
TF-IDF	0.326	0.349	0.337
LL	0.333	0.401	0.355
LO	0.340	0.539	0.434
PMI			0.634
GOLD2	COSINE	DICE	JENSHAN
TF-IDF	0.376	0.392	0.380
LL	0.390	0.440	0.406
LO	0.442	0.561	0.470
PMI			0.581
WAC			
GOLD1	COSINE	DICE	JENSHAN
TF-IDF	0.777	0.757	0.739
LL	0.773	0.934	0.880
LO	0.973	0.324	0.903
PMI			0.629
GOLD2	COSINE	DICE	JENSHAN
TF-IDF	0.694	0.714	0.659
LL	0.714	0.828	0.782
LO	0.883	0.384	0.837
PMI			0.600
RANDOM GOLD1			0.267
RANDOM GOLD2			0.225

Table 3: Average precision obtained over corpora types, gold standards, association measures and similarity measures

The first observation is that the overall results on the WAC corpus pair are about twice as high as the results obtained on the WIKI corpus pair. Since the first is more than 20 times larger than the second, we assumed the amount of information available to be the main cause for such drastic difference.

We then analyzed the difference in the results obtained on the two gold standards. As expected, the results are better on PMI baselines, the RANDOM baseline and in the distributional approach on the WAC corpus pair. The reverse result was obtained with the distributional approach on the WIKI corpus pair and at this point we assumed that it is the result of chance since the results are quite low and close to each other.

6.1 The Baselines

All the results outperformed the weak RANDOM baseline. On the contrary, the strong PMI baseline, which uses only frequency information, proved to be a better method for identifying false friends in the WIKI corpus pair, while it was outperformed by distributional methods on the WAC corpus pair. An important observation regarding PMI in general is that its results relies solely on frequencies of words and having more information than necessary to make good frequency estimates for all the words analyzed cannot improve the results any further. This is the reason why the PMI scores on both corpora pairs regarding the specific gold standard are so close to each other (0.634 and 0.629 on GOLD1, 0.581 and 0.600 on GOLD2), regardless of the much larger size of the WAC corpora pair. This shows that both corpora pairs are large enough for good frequency estimates of the gold standard entries.

Since frequency was not directly encoded in the distributional approach, it seemed reasonable to combine the PMI results with those obtained by the distributional approach. We therefore performed linear combinations of the PMI baseline and various distributional results. They yielded no improvements except in the case of TF-IDF, which still performed worse than most other distributional approaches.

The conclusion regarding PMI is that if one does not have access to a large amount of textual data, pointwise mutual information or some other frequency-based method could be the better way to approach the problem of false friend identification. However, having a lot of data does give advantage to distributional methods. We will look into the exact amount of the data needed to outperform PMI in subsection 6.5.

6.2 Document Alignments on the WIKI Pair

Since PMI performed so well, especially on the WIKI corpus pair on which we have access to document alignments as well, we decided to perform another experiment in which we use that additional information. We calculated the joint probability $p(w_1, w_2)$ not by calculating the maximum likelihood estimate of the identically spelled words in a merged corpus but by taking into account the number of co-occurrences of the identically spelled words in aligned documents only. Naturally, this produced much lower joint probabilities than our initial PMI calculation.

The results of this experiment showed to be as low as the random baseline (0.189 on GOLD1 and 0.255 on GOLD2). The reason was that low-frequency lemmas, many of which are TEs, never occurred together in aligned documents giving those pairs the lowest possible score. When removing the entries that never co-occur, the results did rise slightly over the initial PMI score (0.669 on GOLD1 and 0.549 on GOLD2), but roughly half of the lemma pairs were excluded from the calculation.

To conclude, identifying false friends with a simple measure like pointwise mutual information in case of a limited amount of available data cannot benefit from the additional structure like the Wikipedia document alignments. Having much more data, which would be the case in larger Wikipedias, or applying a more sophisticated measure that would be resistant to scarce data, could prove to be beneficial and is considered a direction for future work.

6.3 Association and Similarity Measures

We continued our analysis by observing the interplay of association and similarity measures. First, we performed our analysis on the much better results obtained on the WAC corpus pair. DICE and LL turned out to be a once-again winning combination. TF-IDF underperformed when compared to LL, showing that LL is the superior association measure in this problem as well. JENSHAN showed a very high consistency, regardless of the association measure used, which is an interesting property, but it never obtained the highest score.

The big surprise was the LO association measure. On the WAC corpus pair it resulted in the overall best score when used with COSINE, but failed drastically when combined with DICE. The situation got even more puzzling once we compared these results with those obtained on the WIKI corpus pair where DICE and LO gave the best overall result. Laroche and Langlais (2010) report to get slightly better or identical results when using LO with COSINE in comparison to DICE.

Trying to find an explanation for such variable results of the LO association measure, we decided to analyze the strongest features in the context vectors of both LO and LL on both corpora pairs. We present our findings in Table 4 on the example of the word *gripa* which means *flu* in both languages. We analyzed the 50 strongest features and classified them in one of the following categories: typo, foreign name, rare term and expected term.

The presented data does shed light on the underlying situation, primarily on the LO association measure, and secondly on the difference between the corpora pairs. LL is a very stable association measure that, regardless of the noise present in the corpora, gave the highest weight to the features one would associate with the concept in question. On the contrary, LO is quite good at emphasizing the noise from the corpora. Since more noise is present in web corpora than in Wikipedia corpora, LO got very good results on the WIKI corpus pair but failed on the WAC corpus pair.

	WIKI			
	SL-LO	SL-LL	HR-LO	HR-LL
typo	0.24	0.00	0.56	0.16
foreign	0.06	0.00	0.22	0.08
rare	0.10	0.00	0.04	0.00
ok	0.60	1.00	0.18	0.76
	WAC			
	SL-LO	SL-LL	HR-LO	HR-LL
typo	0.62	0.00	0.72	0.12
foreign	0.20	0.00	0.26	0.00
rare	0.04	0.00	0.00	0.00
ok	0.14	1.00	0.02	0.88

Table 4: Results of the analysis of the 50 strongest features in the eight different LL and LO vectors

This still did not offer an explanation why LO performed as well as it did on the WAC corpus pair when it was paired with COSINE, or to a smaller extent with JENSHAN. The reason for such behavior lies in the primary difference between DICE and the remaining similarity measures: the latter take into account only the features defined in both vectors while DICE works on a union of the features. Transforming DICE in such a way that it takes into account only the intersection of the defined features did improve the results when using it with LO (from 0.324 and 0.384 to 0.575 and 0.591), but the results deteriorated when used with

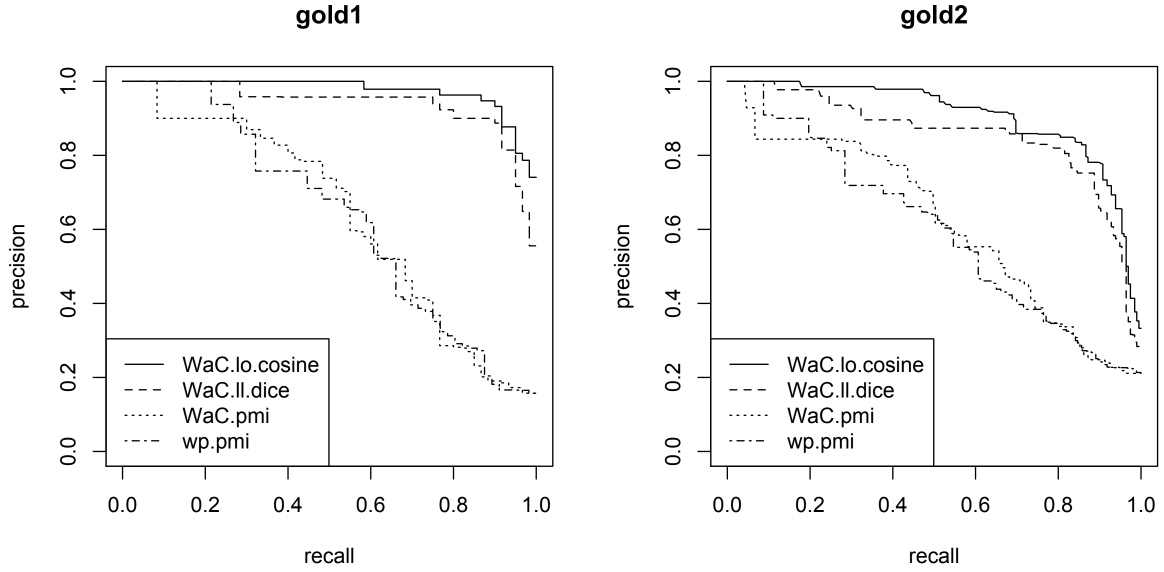

Figure 1: Precision-recall curve of chosen settings on both gold standards

LL (0.934 and 0.828 to 0.768 and 0.719).

We can conclude that LL is a much more stable association measure than LO, but LO performs extremely well as long as the corpora are not noisy or it is not combined with a similarity score that calculates the similarity on a union of the defined features.

6.4 Precision-Recall Curves

We visualized the results obtained with best performing and most interesting settings in Figure 1 with two precision-recall curves, one for each gold standard.

The PR curves stressed the similarity of the results of the PMI method on same gold standards between corpora pairs along the whole precision-recall trade-off spectrum. They also emphasized the significance of the higher quality of the results obtained by the distributional approach on the large WAC corpus pair.

Although somewhat unpredictable, the LO association measure, when coupled with the correct similarity measure, consistently outperformed LL on the whole spectrum on both gold standards.

6.5 Corpus Size

We performed a final set of experiments, which focused on experimenting with the parameter of corpus size. In general, we were interested in the learning curves on different corpora pairs with best performing settings. We also looked for the point where the distributional approach overtakes the frequency approach and a direct comparison between the two corpora pairs.

The learning curves, calculated on random portions of both corpora pairs on GOLD1, are presented in Figure 2. Both PMI learning curves proved our claim that with a sufficient amount of information required to make good frequency estimates, no further improvement can be achieved. On these datasets good estimates were obtained on 5 million words (both languages combined). The PMI learning curve on the WAC corpus pair was steady on the whole scale and we identified the point up to which PMI is more suitable for identifying false friends than distributional methods somewhere around 130 million words (both corpora combined) from where distributional methods surpass the ~ 0.63 plain frequency result.

The WIKI.LL.DICE and the WAC.LL.DICE curves on the left plot enabled us to compare the suitability of the two corpora pairs for the task of identifying false friends and distributional tasks in general. At lower corpus sizes the results were very close, but from 10 million words onwards, the WAC corpus pair outperformed the WIKI corpus pair, consistently pointing toward the conclusion that web corpora are more suitable for distributional approaches than Wikipedia corpora.

The performance of the two distributional approaches depicted on the second graph evened out

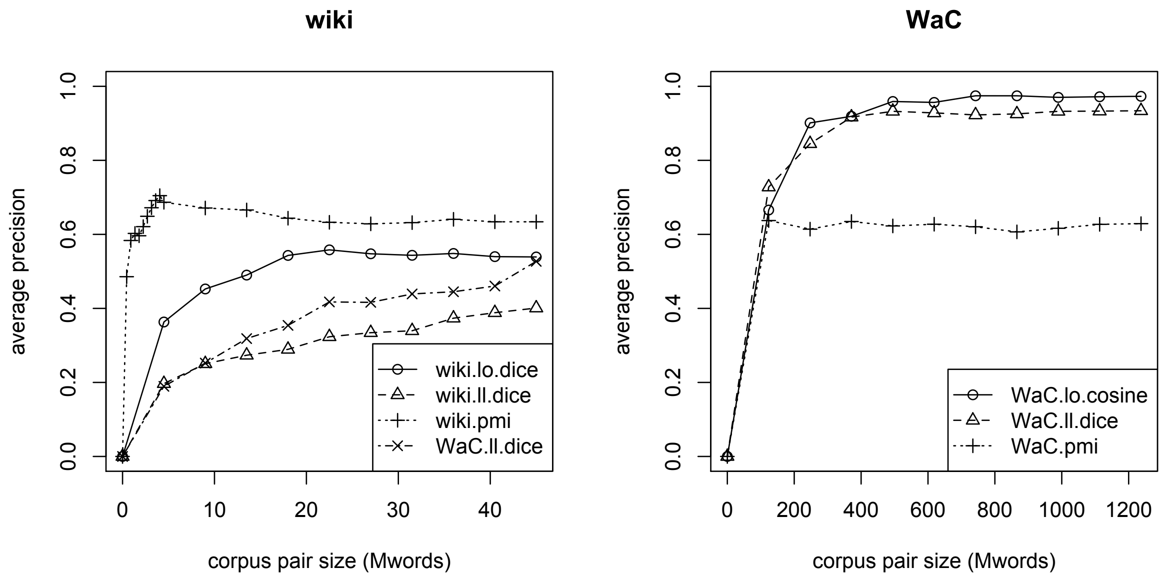

Figure 2: Learning curve on both corpora pairs on GOLD1

around the 500 million word mark, showing that around 250 million words per language should suffice for this task. Having lower-frequency entries in the gold standard would, naturally, call for more data. However, the criterion of 50 occurrences in 500+ million tokens web corpora we used for constructing our gold standards should cover most cases.

Finally, let us point out that the WIKI.LO.DICE curve on the left graph climbed much faster than the WIKI.LL.DICE curve, showing faster learning with the LO association measure in comparison to LL. An interesting observation is that the LO curve obtained its maximum slightly after the 20 million words mark, after which it started a slow decline. Although it could be surprising to see a learning curve declining, this is in line with our previous insights regarding the LO association measure not responding well to many new low-frequency features included in the vector space making the LO+DICE combination struggle. This is one additional reminder that the LO association measure should be used with caution.

7 Conclusion

In this paper we compared frequency-based and distributional approaches to identifying false friends from two frequently used types of corpora pairs – Wikipedia and web corpora. We have used the PMI method for frequency-based ranking and three association and three similarity measures for distributional-based ranking.

The PMI method has proven to be a very good method if one does not have more than 75 million words available per language, in which case it outperformed the more complex distributional approach. Good frequency estimates for PMI were obtained on 2.5 million words per language, after which introducing more data did not yield any further improvement.

Using document alignments from Wikipedia as an additional source for the frequency-based approach did not perform well because of the small size of the Wikipedias in question (slightly above 100,000 articles), often producing zero joint probabilities for non-false friends. A more thought-through approach that could resist data sparsity or using larger Wikipedias is one of our future research directions.

The DICE+LL similarity and association measures proved to be a very stable combination as is the case on the opposite task of translation equivalence extraction (Ljubešić et al., 2011).

The LO association measure gave excellent results, but only if it was paired with a similarity measure that takes into account only the intersection of the features or if the context vectors were calculated on very clean corpora since LO tends to overemphasize low frequency features. We would recommend using this association measure in distributional approaches, but only if one of the above criteria is satisfied.

The amount of data on which the distributional approach stopped benefitting from more data on this task was around 250 million words per language.

Overall, web corpora showed to be better candidates for distributional methods than Wikipedia corpora for two reasons: 1. the WAC learning curve is steeper, and 2. there are few languages which contain 75 million words per language that are necessary to outperform the frequency-based approach and even fewer for which there are 250 million words per language needed for the learning curve to even out.

Our two primary directions for future research are 1. preceding this procedure with identifying language-pair-specific similarly spelled words and 2. including additional language pairs such as Croatian and Czech or Slovene and Czech.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement no. PIAP-GA-2012-324414 (project Abu-MaTran) and from the Slovene national research programme no. P6-0215.

We would like to thank Daša Berović and Jelena Tušek for their help in annotating the datasets.

References

Keith Allan, editor. 2009. *Concise Encyclopedia of Semantics*. Elsevier Science.

Rubén Chacón Beltrán. 2006. Towards a typological classification of false friends (Spanish-English). *Revista Española de Linguística Aplicada*, 19:29–39.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Oana Frunza and Diana Inkpen. 2007. A tool for detecting French-English cognates and false friends. In *Proceedings of the 14th conference Traitement Automatique des Langues Naturelles, TALN'07,*, Toulouse.

Sylviane Granger and Helen Swallow. 1988. False friends: a kaleidoscope of translation difficulties. *Langage et l'Homme*, 23(2):108–120.

John Holmes and Rosinda Guerra Ramos. 1993. False friends and reckless guessers: Observing cognate recognition strategies. In Thomas Huckin, Margot Haynes, and James Coady, editors, *Second Language Reading and Vocabulary Learning*. Norwood, New Jersey: Ablex.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, pages 251–257.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 617–625, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 395–402. Springer.

Nikola Ljubešić, Darja Fišer, Špela Vintar, and Senja Pollak. 2011. Bilingual lexicon extraction from comparable corpora: A comparative study. In *First International Workshop on Lexical Resources, An ESSLLI 2011 Workshop, Ljubljana, Slovenia - August 1-5, 2011.*

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):29–53.

Svetlin Nakov and Preslav Nakov. 2007. Cognate or false friend? Ask the Web. In *Proceedings of the RANLP'2007 workshop: Acquisition and management of multilingual lexicons*.

Svetlin Nakov and Preslav Nakov. 2009. Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing (RANLP'09)*, pages 292—298.

Stefan Schulz, Kornél Markó, Eduardo Sbrissia, Percy Nohama, and Udo Hahn. 2004. Cognate mapping - A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Lianet Sepúlveda Torres and Sandra Maria Aluísio. 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, STIL'11*.