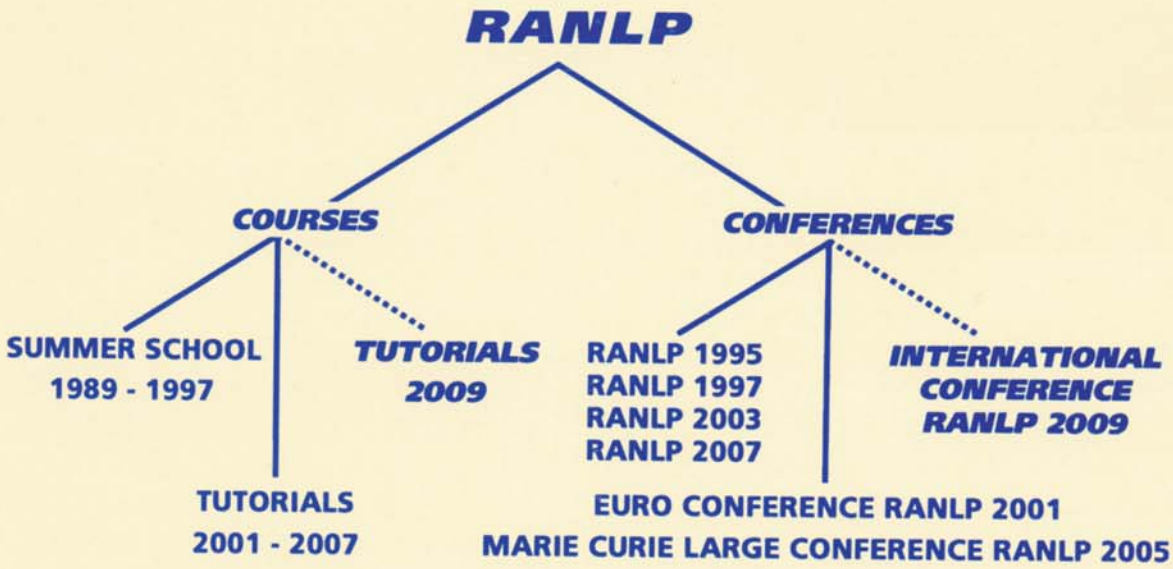
INTERNATIONAL CONFERENCE

RECENT ADVANCES IN

NATURAL LANGUAGE PROCESSING

PROCEEDINGS

Edited by
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov



14-16 September 2009, Borovets, Bulgaria

[Nakov & колектив, 2009a] Nakov S., Nakov P., Paskaleva E. "Unsupervised Extraction of False Friends from Parallel Bi-Texts Using the Web as a Corpus", *Proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP 2009)*, pages 292-298, Borovets, Bulgaria, 2009

Unsupervised Extraction of False Friends from Parallel Bi-Texts Using the Web as a Corpus

Svetlin Nakov and Preslav Nakov*
Department of Mathematics and Informatics
Sofia University "St Kliment Ohridski"
5, James Bourchier Blvd.,1164 Sofia, Bulgaria
{*svetlin.nakov, preslav.nakov*}@*fmi.uni-sofia.bg*

Elena Paskaleva
Linguistic Modeling Department
Institute for Parallel Processing
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev St., 1113 Sofia, Bulgaria
hellen@lml.bas.bg

Abstract

False friends are pairs of words in two languages that are perceived as similar, but have different meanings, e.g., *Gift* in German means *poison* in English. In this paper, we present several unsupervised algorithms for acquiring such pairs from a sentence-aligned bi-text. First, we try different ways of exploiting simple statistics about monolingual word occurrences and cross-lingual word co-occurrences in the bi-text. Second, using methods from statistical machine translation, we induce word alignments in an unsupervised way, from which we estimate lexical translation probabilities, which we use to measure cross-lingual semantic similarity. Third, we experiment with a semantic similarity measure that uses the Web as a corpus to extract local contexts from text snippets returned by a search engine, and a bilingual glossary of known word translation pairs, used as "bridges". Finally, all measures are combined and applied to the task of identifying likely false friends. The evaluation for Russian and Bulgarian shows a significant improvement over previously-proposed algorithms.

Keywords

Cognates, false friends, cross-lingual semantic similarity, Web as a corpus, statistical machine translation.

1 Introduction

Words in two languages that are orthographically and/or phonetically similar are often perceived as mutual translations, which could be wrong in some contexts. Such words are known as *cognates*[1] when they are mutual translations in all contexts, *partial cognates* when they are mutual translations in some contexts but not in other, and *false friends* when they are not mutual translations in any context.

For example, the Bulgarian *слънце* (*slynce*) and the Russian *солнце* (*solnce*) are cognates, both meaning *sun*. However, the Bulgarian *син* (*sin*) and the Russian *сын* (*syn*) are only partial cognates: while they can both mean *son* in some contexts, the Bulgarian word can also mean *blue* in other. Finally, the Bulgarian *бистрота* (*bistrota*) and the Russian *быстрота* (*bystrota*) are false friends meaning *clearness* and *quickness*, respectively.

False friends are important not only for foreign language learning, but also for various natural language processing (NLP) tasks such as statistical machine translation, word alignment, automated translation quality control, etc.

In this paper, we propose an unsupervised approach to the task of extracting pairs of false friends from a sentence-aligned corpus (a bi-text). While we experiment with Bulgarian and Russian, our general method is language-independent.

The remainder of the paper is organized as follows: Section 2 provides an overview of the related work, Section 3 present our method, Section 4 lists the resources we are using, Sections 5 and 6 describe the experiments and discuss the evaluation results, Section 7 concludes and suggests directions for future work.

*Also: Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417, nakov@comp.nus.edu.sg

[1] We should note that linguists define *cognates* as words derived from a common root regardless of whether they differ in meaning or not. For example, the *Electronic Glossary of Linguistic Terms* gives the following definition [3]:

> Two words (or other structures) in related languages are cognate if they come from the same original word (or other structure). Generally cognates will have similar, though often not identical, phonological and semantic structures (sounds and meanings). For instance, Latin *tu*, Spanish *tú*, Greek *sú*, German *du*, and English *thou* are all cognates; all mean 'second person singular', but they differ in form and in whether they mean specifically 'familiar' (non-honorific).

Following previous researchers in *computational* linguistics [2, 24, 25], we will adopt a simplified definition which ignores origin, defining cognates as words in different languages that are mutual translations and have a similar orthography.

2 Related work

Previous work on extracting false friends from text can be divided into the following categories: (1) methods for measuring orthographic and phonetic similarity, (2) methods for identifying cognates and false friends from parallel bi-texts, and (3) semantic methods for distinguishing between cognates and false friends.

Most of the research in the last decade has focused on orthographic methods for cognate identification that do not try to distinguish between cognates from false friends. While traditional orthographic similarity measures like *longest common subsequence ratio* and *minimum edit distance* have evolved over the years towards machine learning approaches for identifying cross-lingual orthographic transformation patterns [2, 26]), recent research has shown interest in using semantic evidence as well [26, 27].

However, little research was conducted on extracting false friends from parallel bi-texts. Very few authors proposed such algorithms [31], while most research focused on word to word alignment [39] and extracting bilingual lexicons with applications to identifying cognates [9].

2.1 Orthographic/phonetic similarity

In this subsection, we describe some relevant methods based on orthographic and phonetic similarity.

2.1.1 Orthographic similarity

The first methods proposed for identifying cognates were based on measuring orthographic similarity. For languages sharing the same alphabet, classical approaches include *minimum edit distance* [22], *longest common subsequence ratio* [25], and variants of the Dice coefficient measuring overlap at the level of character bigrams [5].

The Levenshtein distance or *minimum edit distance* (MED) is defined as the minimum number of INSERT, REPLACE, and DELETE operations at the character level needed to transform one string into another [22]. For example, the MED between the Bulgarian word *пръвият* ('*the first*') and the Russian word *первый* ('*first*') is 4: there are three REPLACE operations, namely, *ъ* → *е*, *у* → *ы*, and *я* → *й*, and one DELETE operation (to remove *т*). To be used as a similarity measure, it is typically normalized by dividing it by the length of the longer word; this normalized measure is known as the *minimum edit distance ratio* (MEDR):

$$\text{MEDR}(\textit{пръвият}, \textit{първый}) = 1 - 4/7 \approx 0.43$$

The *longest common subsequence ratio* (LCSR) [25] is another classic normalized orthographic similarity measure. It is calculated as the ratio of the length of the longest common subsequence of the two words and the length of the longer word. For example, LCS(*пръвият*, *първый*) = *прв*, and thus we have:

$$\text{LCSR}(\textit{пръвият}, \textit{първый}) = 3/7 \approx 0.43$$

Other approaches to measuring orthographic similarity between two words have been proposed by [1] who calculate the Dice coefficient for character bigrams. Their idea is further extended by [5], who used a weighting version of the Dice coefficient, and by [19], who proposed a generalized n-gram measure.

2.1.2 Phonetic similarity

The phonetic similarity measures the degree to which two words sound alike. Unlike orthographic similarity, it operates with sounds rather than letter sequences.

Although a number of specialized algorithms for measuring phonetic similarity have been proposed [11, 17], it can be also measured using orthographic similarity methods after the words have been phonetically transcribed. This approach also works for languages that use different alphabets.

2.1.3 Using transformation rules

Recently, some researchers have proposed to apply transformation rules that reflect typical cross-lingual transformation patterns observed for the target pair of languages before measuring orthographic similarity. This is a good idea when the languages do not use exactly the same alphabet or the same spelling systems. Of course, such substitutions do not have to be limited to single letters and could be defined for letter sequence such as syllables, endings, and prefixes. For example, [16] use manually constructed transformation rules between German and English for expanding a list of cognates: e.g., replacing the German letters *k* and *z* by the English *c*, and changing the ending German *-tät* to the English *-ty*.

Other researchers have tried to learn automatically cross-lingual transformation rules that reflect regular phonetic changes between two languages. For example, [27] do that using MED and positive examples provided as a list of known cognates. Unlike them, [2] use both positive and negative examples to learn weights on substring pairings in order to better identify related substring transformations. Starting with MED, they first obtain an alignment at the letter level. They then extract corresponding substrings that are consistent with that alignment, which in turn are used with a support vector machine (SVM) classifier to distinguish between cognates and false friends.

2.2 Using parallel bi-texts

There is little research on extracting false friends from text corpora directly. Most methods first extract candidate cognates and false friends using some measure of orthographic or phonetic similarity, and then try to distinguish between cognates and false friends in a second step [26].

Fung [9] extracted semantically similar cross-lingual word pairs from parallel and comparable corpora using binary co-occurrence vectors – one for each side of the bi-text.

Brew and McKelvie [5] used sentence alignment to extract cognates and false friends from a bi-text using co-occurrence statistics in the aligned sentences.

Nakov and Pacovski [31] extracted false friends from a paragraph-aligned Bulgarian-Macedonian[2] bi-text, assuming that false friends are unlikely to co-occur in paragraphs that are translations of each other, while cognates tend to do so. Several formulas formalizing this assumption have been proposed and evaluated.

2.3 Semantic approaches

2.3.1 Corpus-based approaches

There has been a lot of research during the last decade on measuring semantic similarity with applications to finding cognates and false friends. Most of the proposed approaches are based on the *distributional hypothesis*, which states that words occurring in similar contexts tend to be semantically similar [12]. This hypothesis is typically operationalized using the vector-space model [38], where vectors are built using words from the local context of the target word as coordinates and their frequencies as values for these coordinates. For example, Nakov & al. [33] defined the local context as a window of a certain size around the target word, while other researchers have limited the context to words in a particular syntactic relationship with the target word, e.g., direct object of a verb [7, 23, 28]. These vectors are typically compared using the cosine of the angle between them as in [33], but other similarity measures such as the Dice coefficient have been used as well [28].

Kondrak [18] proposed an algorithm for measuring semantic similarity using WordNet [8]. He used the following eight semantic similarity levels as binary features: gloss identity, keyword identity, gloss synonymy, keyword synonymy, gloss hypernymy, keyword hypernymy, gloss meronymy, and keyword meronymy. These features were combined with a measure of phonetic similarity and used in a naive Bayes classifier to distinguish between cognates and non-cognates.

Mitkov & al. [26] proposed several methods for measuring the semantic similarity between orthographically similar pairs of words, which were used to distinguish between cognates and false friends. Their first method uses comparable corpora in the two target languages and relies on distributional similarity: given a cross-lingual pair of words to compare, a set of the most similar words to each of the two targets is collected from the respective monolingual corpus. The semantic similarity is calculated as the Dice coefficient between these two sets, using a bilingual glossary to check whether two words can be translations of each other. The method is further extended to use taxonomic data from EuroWordNet [40] when available. The second method extracts co-occurrence statistics for each of the two words of interest from a respective monolingual corpus using a dependency parser. In particular, verbs are used as distributional features when comparing nouns. Semantic sets are thus created, and the similarity between them is measured using the Dice coefficient and a bilingual glossary.

[2] There is a heated linguistic and political debate about whether Macedonian represents a separate language or is a regional literary form of Bulgarian. Since no clear criteria exist for distinguishing a dialect from a language, linguists remain divided on that issue. Politically, Macedonian remains unrecognized as a language by Bulgaria and Greece.

2.3.2 Web-based approaches

The idea of using the Web as a corpus is getting increasingly popular and has been applied to various problems. See [15] for an overview. See also [20, 21] for some interesting applications, and [14, 29] for a discussion on some issues.

Many researchers have used the Web to identify cognates and false friends. Some used Web search engines as a proxy for n-gram frequency, i.e., to estimate how many times a word or a phrase is met on the Web [13], whereas others directly retrieved contexts from the returned text snippets [33]. There have been also some combined approaches, e.g., that of Bollegala & al. [4], who further learned lexico-syntactic templates for semantically related and unrelated words using WordNet, which were used for extracting information from the text snippets returned by the search engine.

3 Method

We propose a method for extracting false friends from a bi-text that combines statistical and semantic evidence in a two-step process: (1) we extract cross-lingual pairs of orthographically similar words, and (2) we identify which of them are false friends.

For the sake of definiteness, below we will assume that our objective is to extract pairs of false friend between Bulgarian and Russian.

3.1 Finding candidates

We look for cross-lingual pairs of words that are perceived as similar and thus could be cognates or false friends. First, we extract from the bi-text[3] all cross-lingual Bulgarian-Russian word pairs (w_{bg}, w_{ru}). We then measure the orthographic similarity between w_{bg} and w_{ru}, and we accept the pair as a candidate only if that similarity is above a pre-specified threshold. In the process, we ignore part of speech, gender, number, definiteness, and case, which are expressed as inflections in both Bulgarian and Russian.

We measure the orthographic similarity using the *modified minimum edit distance ratio* (MMEDR) algorithm [30]. First, some Bulgarian-specific letter sequences are replaced by Russian-specific ones. Then, a weighted minimum edit distance that is specific to Bulgarian and Russian is calculated between the resulting strings. Finally, that distance is normalized by dividing it by the length of the longer word and the result is subtracted from 1 so that it can become a similarity measure.

Let us consider for example the Bulgarian word *пъръвият* ('*the first*') and the Russian one *первый* ('*first*'). First, they will be normalized to *пъръви* and *първи*, respectively. Then, a single vowel-to-vowel REPLACE operation will be performed with the cost of 0.5. Finally, the result will be normalized and subtracted from one:[4]

[3] We extract pairs from the whole bi-text ignoring sentence alignment, i.e., the words in a pair do not have to come from corresponding sentences.

[4] Compare this result to MEDR and LCSR, which severely underestimate the similarity: they are both 0.43 for *пъръвият* and *первый*.

$$\text{MMEDR}(\textit{първият}, \textit{первый}) = 1 - 0.5/7 \approx 0.93$$

Although MMEDR is an orthographic approach, it reflects phonetics to some extent since it uses transformation rules and edit distance weights that are motivated by regular phonetic changes between Bulgarian and Russian as reflected in the spelling systems of the two languages. See [30] for more details.

While the MMEDR algorithm could be improved to learn transformation rules automatically, e.g., following [2, 26, 27], this is out of the scope of the present work. Our main focus below will be on distinguishing between cognates and false friends, which is a much more challenging task.

3.2 Identifying false friends

Once we have collected a list of orthographically and/or phonetically similar cross-lingual word pairs, we need to decide which of them are false friends, i.e., distinguish between false friends and cognates.[5] We use several approaches for this purpose.

3.2.1 Sentence-level co-occurrences

First, we use a statistical approach, based on statistics about word occurrences and co-occurrences in a bi-text. The idea is that cognates tend to co-occur in corresponding sentences while false friends do not. Following [31], we make use of the following statistics:

- $\#(w_{bg})$ – the number of Bulgarian sentences in the bi-text containing the word w_{bg};
- $\#(w_{ru})$ – the number of Russian sentences in the bi-text containing the word w_{ru};
- $\#(w_{bg}, w_{ru})$ – the number of corresponding sentence pairs in the bi-text containing the word w_{bg} on the Bulgarian side and the word w_{ru} on the Russian side.

Nakov & Pacovski [31] tried various combinations of these statistics; in their experiments, the best-performing formula was the following one:

$$F_6(w_{bg}, w_{ru}) = \frac{\#(w_{bg}, w_{ru}) + 1}{\max\left(\frac{\#(w_{bg})+1}{\#(w_{ru})+1}, \frac{\#(w_{ru})+1}{\#(w_{bg})+1}\right)}$$

Now, note that we have the following inequalities:

$$\#(w_{bg}) \geq \#(w_{bg}, w_{ru})$$
$$\#(w_{ru}) \geq \#(w_{bg}, w_{ru})$$

Thus, having a high number of co-occurrences $\#(w_{bg}, w_{ru})$ should increase the probability that the words w_{bg} and w_{ru} are cognates. At the same time, a big difference between $\#(w_{bg})$ and $\#(w_{ru})$ should increase the likelihood of them being false friends. Based on these observations, we propose the following extra formulas (E_1 and E_2):

$$E_1(w_{bg}, w_{ru}) = \frac{(\#(w_{bg}, w_{ru}) + 1)^2}{(\#(w_{bg}) + 1)(\#(w_{ru}) + 1)}$$

$$E_2(w_{bg}, w_{ru}) = \frac{(\#(w_{bg}, w_{ru}) + 1)^2}{P \times Q}$$

where

$$P = \#(w_{bg}) - \#(w_{bg}, w_{ru}) + 1$$
$$Q = \#(w_{ru}) - \#(w_{bg}, w_{ru}) + 1$$

Finally, unlike [31], we perform lemmatization before calculating the above statistics – Bulgarian and Russian are highly inflectional languages, and words are expected to have several inflected forms in the bi-text.

3.2.2 Word alignments

Our next information source are lexical probabilities for cross-lingual word pairs: they should be high for cognates and low for false friends. Such probabilities can be easily estimated from word alignments, which in turn can be obtained using techniques from statistical machine translation.

We start by tokenizing and lowercasing both sides of the training bi-text. We then build separate directed word alignments for Bulgarian→Russian and Russian→Bulgarian using IBM model 4 [6], and we combine them using the *intersect+grow heuristic* described in [34]. Using the resulting undirected alignment, we estimate lexical translation probabilities $\Pr(w_{bg}|w_{ru})$ and $\Pr(w_{ru}|w_{bg})$ for all Bulgarian-Russian word pairs that co-occur in aligned sentences in the bi-text. Finally, we define a cross-lingual semantic similarity measure as follows:

$$lex(w_{bg}, w_{ru}) = \frac{\Pr(w_{bg}|w_{ru}) + \Pr(w_{ru}|w_{bg})}{2}$$

Note that the above definition has an important drawback: it is zero for all words that do not co-occur in corresponding sentences in the bi-text. Thus, we will never use $lex(w_{bg}, w_{ru})$ alone but only in combination with other measures.

3.2.3 Web similarity

Next, we use an algorithm described in [33], which, given a Russian word w_{ru} and a Bulgarian word w_{bg} to be compared, measures the semantic similarity between them using the Web as a corpus and a glossary G of known Russian-Bulgarian translation pairs, used as "bridges". The basic idea is that if two words are translations, then the words in their respective local contexts should be translations as well. The idea is formalized using the Web as a corpus, a glossary of known word translations serving as cross-linguistic "bridges", and the vector space model. We measure the semantic similarity between a Bulgarian and a Russian word, w_{bg} and w_{ru}, by construct corresponding contextual semantic vectors V_{bg} and V_{ru}, translating V_{ru} into Bulgarian, and comparing it to V_{bg}.

[5] Since our ultimate objective is to extract pairs of false friends, we do not need to distinguish between true and partial cognates; we will thus use the term *cognates* to refer to both.

The process of building V_{bg}, starts with a query to Google limited to Bulgarian pages for the target word w_{bg}. We collect the resulting text snippets (up to 1,000), and we remove all stop words – prepositions, pronouns, conjunctions, interjections and some adverbs. We then identify the occurrences of w_{bg}, and we extract the words from its local context – three words on either side. We filter out the words that do not appear on the Bulgarian side of G. We do this for all text snippets. Finally, for each retained word, we calculate the number of times it has been extracted, thus producing a frequency vector V_{bg}. We repeat the procedure for w_{ru} to obtain a Russian frequency vector V_{ru}, which is then "translated" into Bulgarian by replacing each Russian word with its translation(s) in G, retaining the co-occurrence frequencies. In case there are multiple Bulgarian translations for some Russian word, we distribute the corresponding frequency equally among them, and in case of multiple Russian words with the same Bulgarian translation, we sum up the corresponding frequencies. As a result, we end up with a Bulgarian vector $V_{ru \to bg}$ for the Russian word w_{ru}. Finally, we calculate the semantic similarity between w_{bg} and w_{ru} as the cosine of the angle between their corresponding Bulgarian vectors, V_{bg} and $V_{ru \to bg}$, as follows:

$$sim(w_{bg}, w_{ru}) = \frac{V_{bg}.V_{ru \to bg}}{|V_{bg}| \times |V_{ru \to bg}|}$$

3.2.4 Combined approach

While all three approaches described above – sentence-level co-occurrences, word alignments, and Web similarity – are useful for distinguishing between cognates and false friends, each of them has some weaknesses.

Using sentence-level co-occurrences and word alignments is a good idea when the statistics for the target words are reliable, but both work poorly for infrequent words. Unfortunately, most words (and thus most word pairs) will be rare for any (bi-)text, according to the Zipf law [41],

Data sparsity is less of an issue for the Web similarity approach, which uses statistics derived from the largest existing corpus – the Web. Still, while being quite reliable for unrelated words, it sometimes assigns very low scores to highly-related word pairs. The problem stems from it relying on a commercial search engine like Google, which only returns up to 1,000 results per query and rates too high sites about news, e-commerce, and blogs, which introduces a bias on the local contexts of the target words. Moreover, some geographical and cultural notions have different contexts on the Web for Bulgarian and Russian, despite being very related otherwise, e.g., person names and goods used in e-commerce (due to different popular brands in different countries).

A natural way to overcome these issues is to combine all three approaches, e.g., by taking the average of the similarity each of them predicts. As we will see below, this is indeed quite a valuable strategy.

4 Resources

For the purpose of our experiments, we used the following resources: a bi-text, lemmatization dictionaries, a bilingual glossary, and the Web as a corpus.

4.1 Bi-text

We used the first seven chapters of the Russian novel *Lord of the World* by Alexander Belyaev and its Bulgarian translation consisting of 759 parallel sentences. The text has been sentence aligned automatically using the alignment tool *MARK ALISTeR* [36], which is based on the Gale-Church algorithm [10].

4.2 Morphological dictionaries

We used two large monolingual morphological dictionaries for Bulgarian and Russian created at the Linguistic Modeling Department of the Institute for Parallel Processing in the Bulgarian Academy of Sciences [35]. The Bulgarian dictionary contains about 1M wordforms and 70K lemmata, and the Russian one contains about 1.5M wordforms and 100K lemmata.

4.3 Bilingual glossary

We built a bilingual glossary by adapting an online Russian-Bulgarian dictionary. First, we removed all multi-word expressions. Then we combined each Russian word with each of its Bulgarian translations – due to polysemy/homonymy some words had multiple translations. We thus obtained a glossary of 59,583 pairs of words that are translations of each other.

4.4 Web as a corpus

In our experiments, we performed searches in Google for 557 Bulgarian and for 550 Russian wordforms, and we collected as many as possible (up to 1,000) page titles and text snippets from the search results.

5 Experiments and evaluation

We performed several experiments in order to evaluate the above-described approaches – both individually and in various combinations.[6]

First, we extracted cross-lingual pairs of orthographically similar words using the MMEDR algorithm with a threshold of 0.90. This yielded 612 pairs, each of which was judged by a linguist[7] as being a pair of false friends, partial cognates, or true cognates. There were 35 false friends (5.72%), 67 partial cognates (10.95%), and 510 true cognates (83.33%).

Then, we applied different algorithms to distinguish which of the 612 pairs are false friends and which are cognates (partial or true). Each algorithm assigned a similarity score to each pair and then the pairs were sorted by that score in descending order. Ideally, the false friends, having a very low similarity, should appear at the top of the list, followed by the cognates.

[6] Some of the experiments have been published in [32].

[7] The linguist judged the examples as cognates or false friends by consulting Bulgarian-Russian bilingual dictionaries.

Following [2] and [33], the resulting lists were evaluated using *11-pt average precision*, which is well-known in information retrieval [37]; it averages the precision at 11 points corresponding to recall of 0%, 10%, 20%, ..., 100%, respectively.

We performed the following experiments:

- BASELINE – word pairs in alphabetical order: it behaves nearly like a random function and is used as a baseline;
- COOC – the sentence-level co-occurrence algorithm of [31] with their formula F_6;
- COOC+L – the algorithm COOC with lemmatization;
- COOC+E1 – the algorithm COOC with the formula E_1;
- COOC+E1+L – the algorithm COOC with the formula E_1 and lemmatization;
- COOC+E2 – the algorithm COOC with the formula E_2;
- COOC+E2+L – the algorithm COOC with the formula E_2 and lemmatization;
- WEB+L – the Web-based semantic similarity with lemmatization;
- WEB+COOC+L – averaging the values of WEB+L and COOC+L;
- WEB+E1+L – averaging the values of WEB+L and E1+L;
- WEB+E2+L – averaging the values of WEB+L and E2+L;
- WEB+SMT+L – the algorithm WEB+L combined with the statistical machine translation similarity measure by averaging the values of WEB+L and the estimated lexical translation probability;
- COOC+SMT+L – the algorithm COOC+L combined with the machine translation similarity by averaging the similarity scores;
- E1+SMT+L – the algorithm E1+L combined with the machine translation similarity by averaging the similarity scores;
- E2+SMT+L – the algorithm E2+L combined with the machine translation similarity by averaging the similarity scores;
- WEB+COOC+SMT+L – averaging WEB+L, COOC+L, and the machine translation similarity;
- WEB+E1+SMT+L – averaging WEB+L, E1+L, and the machine translation similarity;
- WEB+E2+SMT+L – averaging WEB+L, E2+L and the machine translation similarity.

The results are shown in Table 1. We also tried several ways of weighting the statistical and semantic components in some of the above algorithms, but this had little impact on the results.

Algorithm	**11-pt average precision**
BASELINE	4.17%
E2	38.60%
E1	39.50%
COOC	43.81%
COOC+L	53.20%
COOC+SMT+L	56.22%
WEB+COOC+L	61.28%
WEB+COOC+SMT+L	61.67%
WEB+L	63.68%
E1+L	63.98%
E1+SMT+L	65.36%
E2+L	66.82%
WEB+SMT+L	69.88%
E2+SMT+L	70.62%
WEB+E2+L	76.15%
WEB+E1+SMT+L	76.35%
WEB+E1+L	77.50%
WEB+E2+SMT+L	78.24%

Table 1: *Performance of the different methods sorted by 11-pt average precision (in %).*

6 Discussion

Table 1 shows that most algorithms perform well above the baseline, with the exception of E_1 and E_2 when used in isolation. However, when combined with lemmatization, both E_1 and E_2 perform much better than the original formula F_6 (the COOC algorithm) of [31]: Bulgarian and Russian are highly inflectional languages and thus applying lemmatization is a must. Not surprisingly, the best results are obtained for the combined approaches.

Overall, we observe significant improvements over the original algorithm of [31], but the results are still not perfect.

7 Conclusions and future work

We have presented and compared several algorithms for acquiring pairs of false friends from a sentence-aligned bi-text based on sentence-level co-occurrence statistics, word alignments, the Web as a corpus, lemmatization, and various combinations thereof. The experimental results show a significant improvement over [31].

There are many ways in which the results could be improved even further. First, we would like to try other formulas for measuring the semantic similarity using word-level occurrences and co-occurrences in a bi-text. It would be also interesting to try the contextual mapping vectors approach of [9]. We could try using additional bi-texts to estimate more reliable word alignments, and thus, obtain better lexical probabilities. Using non-parallel corpora as in [26] is another promising direction for future work. The Web-based semantic similarity calculations could be improved using syntactic relations between the words as in [4]. Finally we would like to try the algorithm for other language pairs and to compare our results directly with those of other researchers – on the same datasets.

Acknowledgments

The presented research is supported by the project FP7-REGPOT-2007-1 SISTER.

References

[1] G. W. Adamson and J. Boreham. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10(7-8):253–260, 1974.

[2] S. Bergsma and G. Kondrak. Alignment-based discriminative string similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 656–663, Prague, Czech Republic, 2007.

[3] A. Bickford and D. Tuggy. Electronic glossary of linguistic terms. http://www.sil.org/mexico/ling/glosario/E005aiGlossary.htm, 2002.

[4] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, pages 757–766, New York, NY, USA, 2007. ACM.

[5] C. Brew and D. McKelvie. Word-pair extraction for lexicography. In *Proceedings of the 2nd Int. Conference on New Methods in Language Processing*, pages 44–55, 1996.

[6] P. Brown, V. D. Pietra, S. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[7] J. R. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 59–66, 2002.

[8] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[9] P. Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, 1998.

[10] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993.

[11] J. B. M. Guy. An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. *Journal of Quantitative Linguistics*, 1(1):35–42, 1994.

[12] Z. S. Harris. Distributional Structure. *Word*, 10:146–162, 1954.

[13] D. Z. Inkpen. Near-synonym choice in an intelligent thesaurus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'07)*, pages 356–363, Rochester, New York, USA, 2007.

[14] F. Keller and M. Lapata. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29:459–484, 2003.

[15] A. Kilgariff and G. Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347, 2003.

[16] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL workshop on Unsupervised Lexical Acquisition*, pages 9–16, 2002.

[17] G. Kondrak. Identifying complex sound correspondences in bilingual wordlists. In *In Proceedings the 4th International Conference on Computational Linguistics and Intelligent Text Processing (COLING'03)*, pages 432–443, 2003.

[18] G. Kondrak. Combining evidence in cognate identification. In *Advances in Artificial Intelligence, 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2004*, volume 3060 of *Lecture Notes in Computer Science*, pages 44–59, London, Ontario, Canada, 2004. Springer.

[19] G. Kondrak and B. J. Dorr. Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, 36(1):29–42, 2006.

[20] M. Lapata and F. Keller. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'04)*, 2004.

[21] M. Lapata and F. Keller. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.*, 2(1):3, 2005.

[22] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965.

[23] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, 1998.

[24] G. Mann and D. Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of the Annual Meeting of the North American Association for Computational Linguistics (NAACL'01)*, pages 1–8, Pittsburgh, PA, USA, 2001.

[25] D. Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, 1999.

[26] R. Mitkov, V. Pekar, D. Blagoev, and A. Mulloni. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53, 2007.

[27] A. Mulloni and V. Pekar. Automatic detection of orthographic cues for cognate recognition. In *Proceedings of the conference on Language Resources and Evaluation (LREC-06)*, pages 2387–2390, 2006.

[28] A. Mulloni, V. Pekar, and R. M. D. Blagoev. Semantic evidence for automatic identification of cognates. In *Proceedings of the RANLP'2007 workshop: Acquisition and management of multilingual lexicons.*, pages 49–54, Borovets, Bulgaria, 2007.

[29] P. Nakov and M. Hearst. A study of using search engine page hits as a proxy for n-gram frequencies. In *Proceedings of Recent Advances in Natural Language Processing (RANLP'2005)*, pages 347–353, Borovets, Bulgaria, September 2005.

[30] P. Nakov, S. Nakov, and E. Paskaleva. Improved word alignments using the Web as a corpus. In *Proceedigs of Recent Advances in Natural Language Processing (RANLP'07)*, pages 400–405, Borovets, Bulgaria, 2007.

[31] P. Nakov and V. Pacovski. *Readings in Multilinguality, Selected Papers from Young Researchers in BIS-21++*, chapter Acquiring False Friends from Parallel Corpora: Application to South Slavonic Languages, pages 87–94. Incoma Ltd, Shumen, Bulgaria, 2006.

[32] S. Nakov. Automatic identification of false friends in parallel corpora: Statistical and semantic approach. *Serdica Journal of Computing*, 3(2):133–158, 2009.

[33] S. Nakov, P. Nakov, and E. Paskaleva. Cognate or false friend? Ask the Web! In *Proceedings of the RANLP'2007 workshop: Acquisition and management of multilingual lexicons.*, pages 55–62, Borovets, Bulgaria, 2007.

[34] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[35] E. Paskaleva. Compilation and validation of morphological resources. In *Workshop on Balkan Language Resources and Tools (Balkan Conference on Informatics)*, pages 68–74, 2003.

[36] E. Paskaleva and S. Mihov. Second language acquisition from aligned corpora. In *Proceedings of International Conference "Language Technology and Language Teaching"*, pages 43–52, 1998.

[37] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[38] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[39] J. Tiedemann. Word to word alignment strategies. In *Proceedings of the 20th international conference on Computational Linguistics (COLING'04)*, page 212, 2004.

[40] P. Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[41] G. Zipf. The psycho-biology of language. Houghton Mifflin, Boston, MA, 1935.