INTERNATIONAL WORKSHOP

ACQUISITION AND MANAGEMENT OF MULTILINGUAL LEXICONS

held in conjunction with the International Conference

RANLP - 2007, September 27-29, 2007, Borovets, Bulgaria

PROCEEDINGS

Edited by

Viktor Pekar, Diana Inkpen and Andrea Mulloni

Borovets, Bulgaria

30th September 2007

TABLE OF CONTENTS

[Nakov & колектив, 2007b] Nakov S., Nakov P., Paskaleva E. "Cognate or False Friend? Ask the Web!", *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, held in conjunction with RANLP 2007*, pages 55–62, Borovets, Bulgaria, 2007

Cognate or False Friend? Ask the Web!

Svetlin Nakov
Sofia University
5 James Boucher Blvd.
Sofia, Bulgaria
nakov@fmi.uni-sofia.bg

Preslav Nakov
Univ. of Cal. Berkeley
EECS, CS division
Berkeley, CA 94720
nakov@cs.berkeley.edu

Elena Paskaleva
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str.
Sofia, Bulgaria
hellen@lml.bas.bg

Abstract

We propose a novel unsupervised semantic method for distinguishing cognates from false friends. The basic intuition is that if two words are cognates, then most of the words in their respective local contexts should be translations of each other. The idea is formalised using the Web as a corpus, a glossary of known word translations used as cross-linguistic "bridges", and the vector space model. Unlike traditional orthographic similarity measures, our method can easily handle words with identical spelling. The evaluation on 200 Bulgarian-Russian word pairs shows this is a very promising approach.

Keywords

Cognates, false friends, semantic similarity, Web as a corpus.

1 Introduction

Linguists define *cognates* as words derived from a common root. For example, the *Electronic Glossary of Linguistic Terms* gives the following definition [5]:

> 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 [4, 22, 25], we adopt a simplified definition, which ignores origin, defining *cognates* (or *true friends*) as words in different languages that are translations and have a similar orthography. Similarly, we define *false friends* as words in different languages with similar orthography that are not translations. Here are some identically-spelled examples of false friends:

- **pozor** (*позор*) means *a disgrace* in Bulgarian, but *attention* in Czech;
- **mart** (*март*) means *March* in Bulgarian, but *a market* in English;
- **Gift** means *a poison* in German, but *a present* in English;
- **Prost** means *cheers* in German, but *stupid* in Bulgarian.

And some examples with a different orthography:

- **embaraçada** means *embarrassed* in Portuguese, while **embarazada** means *pregnant* in Spanish;
- **spenden** means *to donate* in German, but **to spend** means *to use up* or *to pay out* in English;
- **bachelier** means *a person who passed his bac exam* in French, but in English **bachelor** means *an unmarried man*;
- **babichka** (*бабичка*) means *an old woman* in Bulgarian, but **babochka** (*бабочка*) is *a butterfly* in Russian;
- **godina** (*година*) means *a year* in Russian, but **godzina** is *an hour* in Polish.

In the present paper, we describe a novel semantic approach to distinguishing cognates from false friends. The paper is organised as follows: Sections 2 explains the method, section 3 describes the resources, section 4 presents the data set, section 5 describes the experiments, section 6 discusses the results of the evaluation, and section 7 points to important related work. We conclude with directions for future work in section 8.

2 Method

2.1 Contextual Web Similarity

We propose an unsupervised algorithm, 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 formalised 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}.

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 three words on either side of it. We filter out the words that do not appear on the Bulgarian side of G. 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 of 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 between their corresponding Bulgarian vectors, V_{bg} and $V_{ru \to bg}$.

2.2 Reverse Context Lookup

The *reverse context lookup* is a modification of the above algorithm. The original algorithm implicitly assumes that, given a word w, the words in the local context of w are semantically associated with it, which is often wrong due to Web-specific words like *home*, *site*, *page*, *click*, *link*, *download*, *up*, *down*, *back*, etc. Since their Bulgarian and Russian equivalents are in the glossary G, we can get very high similarity for unrelated words. For the same reason, we cannot judge such navigational words as true/false friends.

The reverse context lookup copes with the problem as follows: in order to consider w associated with a word w_c from the local context of w, it requires that w appear in the local context of w_c as well[1]. More formally, let $\#(x, y)$ be the number of occurrences of x in the local context of y. The strength of association is calculated as $p(w, w_c) = \min\{\#(w, w_c), \#(w_c, w)\}$ and is used in the vector coordinates instead of $\#(w, w_c)$, which is used in the original algorithm.

2.3 Web Similarity Using Seed Words

For comparison purposes, we also experiment with the *seed words* algorithm of Fung&Yee'98 [12], which we adapt to use the Web. We prepare a small glossary of 300 Russian-Bulgarian word translation pairs, which is a subset of the glossary used for our contextual Web similarity algorithm[2]. Given a Bulgarian word w_{bg} and a Russian word w_{ru} to compare, we build two vectors, one Bulgarian (V_{bg}) and one Russian (V_{ru}), both of size 300, where each coordinate corresponds to a particular glossary entry (g_{ru}, g_{bg}). Therefore, we have a direct correspondence between the coordinates of V_{bg} and V_{ru}. The coordinate value for g_{bg} in V_{bg} is calculated as the total number of co-occurrences of w_{bg} and g_{bg} on the Web, where g_{bg} immediately precedes or immediately follows w_{bg}. This number is calculated using Google page hits as a proxy for bigram frequencies: we issue two exact phrase queries "w_{bg} g_{bg}" and "g_{bg} w_{bg}", and we sum the corresponding numbers of page hits. We repeat the same procedure with w_{ru} and g_{ru} in order to obtain the values for the corresponding coordinates of the Russian vector V_{ru}. Finally, we calculate the semantic similarity between w_{bg} and w_{ru} as the cosine between V_{bg} and V_{ru}.

[1] These contexts are collected using a separate query for w_c.

[2] We chose those 300 words from the glossary that occur most frequently on the Web.

3 Resources

3.1 Grammatical Resources

We use two monolingual dictionaries for lemmatisation. For Bulgarian, we have a large morphological dictionary, containing about 1,000,000 wordforms and 70,000 lemmata [29], created at the Linguistic Modeling Department, Institute for Parallel Processing, Bulgarian Academy of Sciences. Each dictionary entry consists of a wordform, a corresponding lemma, followed by morphological and grammatical information. There can be multiple entries for the same wordform, in case of multiple homographs. We also use a large grammatical dictionary of Russian in the same format, consisting of 1,500,000 wordforms and 100,000 lemmata, based on the Grammatical Dictionary of A. Zaliznjak [35]. Its electronic version was supplied by the Computerised fund of Russian language, Institute of Russian language, Russian Academy of Sciences.

3.2 Bilingual Glossary

We built a bilingual glossary using an online Russian-Bulgarian dictionary[3] with 3,982 entries in the following format: a Russian word, an optional grammatical marker, optional stylistic references, and a list of Bulgarian translation equivalents. 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. As a result, we obtained a glossary G of 4,563 word-word translation pairs (3,794 if we exclude the stop words).

3.3 Huge Bilingual Glossary

Similarly, we adapted a much larger Bulgarian-Russian electronic dictionary, transforming it into a bilingual glossary with 59,583 word-word translation pairs.

4 Data Set

4.1 Overview

Our evaluation data set consists of 200 Bulgarian-Russian pairs – 100 cognates and 100 false friends. It has been extracted from two large lists of cognates and false friends, manually assembled by a linguist from several monolingual and bilingual dictionaries. We limited the scope of our evaluation to nouns only.

[3] `http://www.bgru.net/intr/dictionary/`

As Table 1 shows, most of the words in our pairs constitute a perfect orthographic match: this is the case for 79 of the false friends and for 71 of the cognates. The remaining ones exhibit minor variations, e.g.:

- *ы* → *и* (*r. р<u>ы</u>ба* → *b. р<u>и</u>ба*, '*a fish*');
- *э* → *е* (*r. <u>э</u>таж* → *b. <u>е</u>таж*, '*a floor*');
- *ь* → ∅ (*r. кост<u>ь</u>* → *b. кост*, '*a bone*');
- *double consonant* → *single consonant* (*r. програ<u>мм</u>а* → *b. програ<u>м</u>а*, '*a programme*');
- etc.

4.2 Discussion

There are two general approaches to testing a statistical hypothesis about a linguistic problem: (1) from the data to the rules, and (2) from the rules to the data. In the first approach, we need to collect a large number of instances of potential interest and then to filter out the bad ones using lexical and grammatical competence, linguistic rules as formulated in grammars and dictionaries, etc. This direction is from the data to the rules, and the final evaluation is made by a linguist. The second approach requires to formulate the postulates of the method from a linguistic point of view, and then to check its consistency on a large volume of data. Again, the check is done by a linguist, but the direction is from the rules to the data.

We combined both approaches. We started with two large lists of cognates and false friends, manually assembled by a linguist from several monolingual and bilingual dictionaries: Bulgarian [2, 3, 8, 19, 20, 30], Russian [10], and Bulgarian-Russian [6, 28]. From these lists, we repeatedly extracted Russian-Bulgarian word pairs (nouns), cognates and false friends, which were further checked against our monolingual electronic dictionaries, described in section 3. The process was repeated until we were able to collect 100 cognates and 100 false friends.

Given an example pair exhibiting orthographic differences between Bulgarian and Russian, we tested against our electronic dictionaries the corresponding letter sequences substitutions. While the four correspondences in section 4.1 have proven incontestable, other have been found inconsistent. For example, the correspondence between the Russian *-оро-* and the Bulgarian *-ра-* (e.g. *r. г<u>оро</u>х* → *b. г<u>ра</u>х*, '*peas*'), mentioned in many comparative Russian-Bulgarian studies, does not always hold, as it is formulated for root morphemes only. This correspondence fails in cases where these strings occur outside the root morpheme or at morpheme boundaries, e.g. *r. плодо<u>родие</u>* → *b. плодо<u>родие</u>*, '*fruitfulness, fertility*'. We excluded all examples exhibiting inconsistent orthographic alternations between Bulgarian and Russian.

The linguistic check against the grammatical dictionaries further revealed different interactions between orthography, part-of-speech (POS), grammatical function, and sense, suggesting the following degrees of *falseness*:

- **Absolute falseness:** same lemma, same POS, same number of senses, but different meanings for all senses, e.g. *r. жесть* ('*a tin*') and *b. жест* ('*a gesture*');
- **Partial lemma falseness:** same lemma, same POS, but different number of senses, and different meanings for some senses. For example, the *r. бас* ('*bass, voice*') is a cognate of the first sense of the *b. бас*, '*bass, voice*', but a false friend of its second sense '*a bet*' (the Russian for *a bet* is *пари*). On the other hand, the *r. пари* is a false friend of the first sense of the *b. пари* ('*money*'); the Russian for *money* is *деньги*. In addition, the *b. пари* can also be the plural for the *b. пара* ('*a vapour*'), which translates into Russian as *испарения*. This quite complex example shows that the falseness is not a symmetric cross-linguistic relation. It is shown schematically on Figure 1.
- **Partial wordform falseness:** the number of senses is not relevant, the POS can differ, the lemmata are different, but some wordform of the lemma in one language is identical to a wordform of the lemma in the second language. For example, the *b. хотел* ('*a hotel*') is the same as the inflected form *r. хотел* (past tense, singular, masculine of *хотеть*, '*to want*').

Our list of true and false friends contains only absolute false friends and cognates, excluding any partial cognates. Note that we should not expect to have a full identity between all wordforms of the cognates for a given pair of lemmata, since the rules for inflections are different for Bulgarian and Russian.

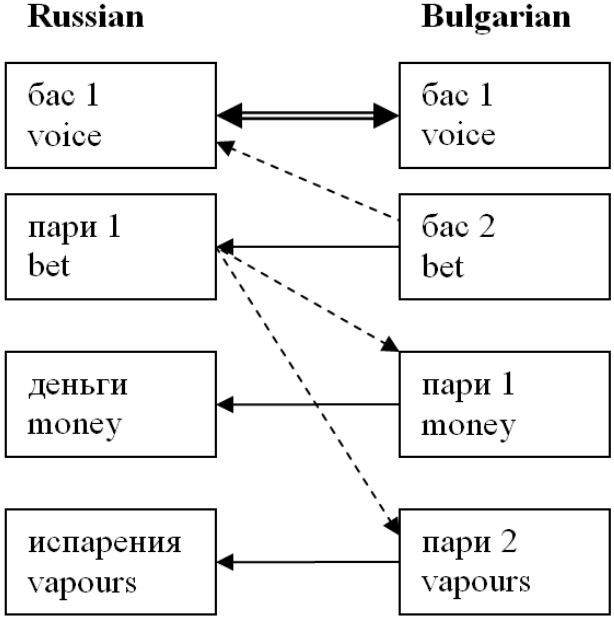

Fig. 1: Falseness example: *double lines link cognates, dotted lines link false friends, and solid lines link translations.*

5 Experiments and Evaluation

In our experiments, we calculate the semantic (or orthographic) similarity between each pair of words from the data set. We then order the pairs in ascending order, so that the ones near the top are likely to be false friends, while those near the bottom are likely to be cognates. Following Bergsma&Kondrak'07 [4] and Kondrak&Sherif'06 [18], we measure the quality of the ranking using *11-point average precision*. We experiment with the following similarity measures:

- **Baseline**
  - BASELINE: random.
- **Orthographic Similarity**
  - MEDR: *minimum edit distance ratio*, defined as $\text{MEDR}(s_1,s_2) = 1 - \frac{|\text{MED}(s_1,s_2)|}{\max(|s_1|,|s_2|)}$, where $|s|$ is the length of the string s, and MED is the minimum edit distance or Levenshtein distance [21], calculated as the minimum number of edit operations – INSERT, REPLACE, DELETE – needed to transform s_1 into s_2. For example, the MED between *b.* **мляко** ('*milk*') and *r.* **молоко** ('*milk*') is two: one REPLACE operation (**я** → **о**) and one INSERT operation (of **о**). Therefore we obtain MEDR(**мляко**, **молоко**) = $1 - 2/6 \approx 0.667$;
  - LCSR: *longest common subsequence ratio* [24], defined as $\text{LCSR}(s_1,s_2) = \frac{|\text{LCS}(s_1,s_2)|}{\max(|s_1|,|s_2|)}$, where $\text{LCS}(s_1, s_2)$ is the longest common subsequence of s_1 and s_2. For example, the LCS(**мляко**, **молоко**) = **мло**, and therefore LCSR(**мляко**, **молоко**) = $4/6 \approx 0.667$.
- **Semantic Similarity**
  - SEED: our implementation and adaptation of the *seed words* algorithm of Fung&Yee'98 [12];
  - WEB3: the Web-based similarity algorithm with the default parameters: local context size of 3, the smaller bilingual glossary, stop words filtering, no lemmatisation, no reverse context lookup, no TF.IDF-weighting;
  - NO-STOP: WEB3 without stop words removal;
  - WEB1: WEB3 with local context size of 1;
  - WEB2: WEB3 with local context size of 2;
  - WEB4: WEB3 with local context size of 4;
  - WEB5: WEB3 with local context size of 5;
  - WEB3+TF.IDF: WEB3 with context size of 1;
  - LEMMA: WEB3 with lemmatisation;
  - LEMMA+TF.IDF: WEB3 with lemmatisation and TF.IDF-weighting;
  - HUGEDICT: WEB3 with the huge glossary;
  - HUGEDICT+TF.IDF: WEB3 with the huge glossary and TF.IDF-weighting;
  - REVERSE: WEB3 with reverse context lookup;
  - REVERSE+TF.IDF: WEB3 with reverse context lookup and TF.IDF-weighting;
  - COMBINED: WEB3 with lemmatisation, huge glossary, and reverse context lookup;
  - COMBINED+TF.IDF: COMBINED with TF.IDF-weighting.

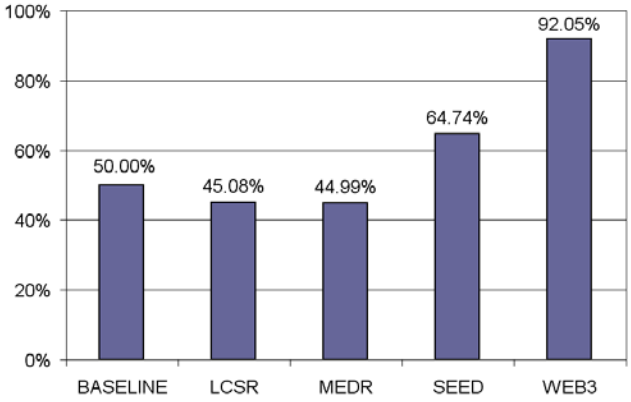

Fig. 2: Evaluation, 11-point average precision. *Comparing* WEB3 *with* BASELINE *and three old algorithms –* LCSR*,* MEDR *and* SEED.

First, we compare two semantic similarity measures, WEB3 and SEED, with the orthographic similarity measures, LCSR and MEDR, and with BASELINE; the results are shown on Figure 2. The BASELINE algorithm achieves 50% on 11-point average precision, and outperforms the orthographic similarity measures LCSR and MEDR, which achieve 45.08% and 44.99%. This is not surprising since most of our pairs consist of identical words. The semantic similarity measures, SEED and WEB3 perform much better, achieving 64.74% and 92.05% on 11-point average precision. The huge absolute difference in performance (almost 30%) between SEED and WEB3 suggests that building a dynamic set of textual contexts from which to extract words co-occurring with the target is a much better idea than using a fixed set of seed words and page hits as a proxy for Web frequencies.

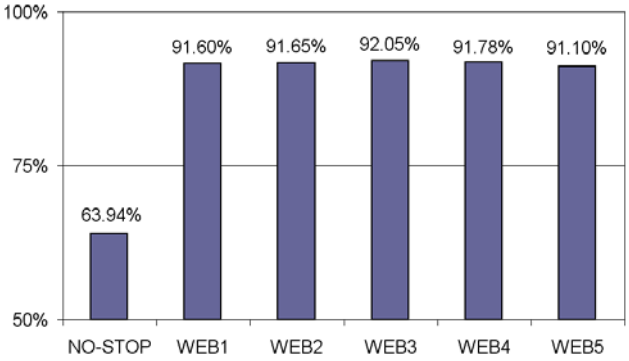

Fig. 3: Evaluation, 11-point average precision. *Different context sizes; keeping the stop words.*

The remaining experiments try different variations of the contextual Web similarity algorithm WEB3. First, we tested the impact of stop words removal. Recall that WEB3 removes the stop words from the text snippets returned by Google; therefore, we tried a version of it, NO-STOP, which keeps them. As Figure 3 shows, this was a bad idea yielding about 28% absolute loss in accuracy – from 92.05% to 63.94%. We also tried different context sizes: 1, 2, 3, 4 and 5. Context size 3 performed best, but the differences are small.

We also experimented with different modifications of WEB3: using lemmatisation, TF.IDF-weighting, reverse lookup, a bigger glossary, and a combination of them. The results are shown on Figure 5.

First, we tried lemmatising the words from the snippets using the monolingual grammatical dictionaries described in section 3.1. We tried both with and without TF.IDF-weighting, achieving in either case an improvement of about 2% over WEB3: as Figure 5 shows, LEMMA and LEMMA+TF.IDF yield an 11-point average precision of 94.00% and 94.17%, respectively.

In our next experiment, we used the thirteen times larger glossary described in section 3.3, which yielded an 11-point average precision of 94.37% – an absolute improvement of 2.3% compared to WEB3 (Figure 5, HUGEDICT). Interestingly, when we tried using TF.IDF-weighting together with this glossary, we achieved only 93.31% (Figure 5, HUGEDICT-TF.IDF).

We also tried the reverse context lookup described in section 2.2, which improved the results by 3.64% to 95.69%. Again, combining it with TF.IDF-weighting, performed worse: 94.58%.

Finally, we tried a combination of WEB3 with lemmatisation, reverse lookup and the huge glossary, achieving an 11-point average precision of 95.84%, which is our best result. Adding TF.IDF-weighting to the combination yielded slightly worse results: 94.23%.

Figure 4 shows the precision-recall curves for LCSR, SEED, WEB3, and COMBINED. We can see that WEB3 and COMBINED clearly outperform LCSR and SEED.

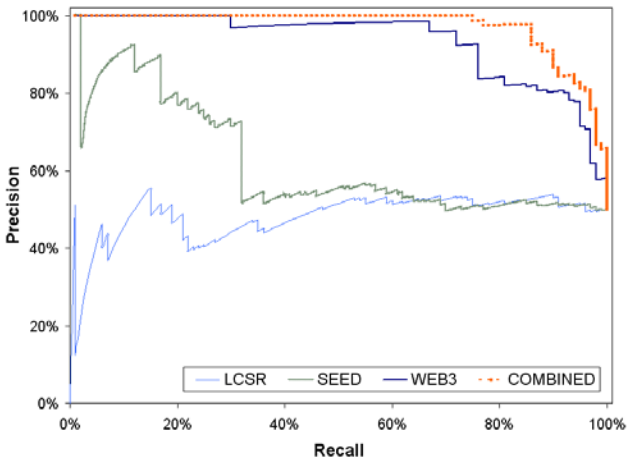

Fig. 4: Precision-Recall Curve. *Comparing* WEB3 *with* LCSR, SEED *and* COMBINED.

6 Discussion

Ideally, our algorithm would rank first all false friends and only then the cognates. Indeed, in the ranking produced by COMBINED, the top 75 pairs are false friends, while the last 48 are cognates; things get mixed in the middle.

The two lowest-ranked (misplaced) by COMBINED false friends are *врата* ('*a door*' in Bulgarian, but '*an entrance*' in Russian) at rank 152, and *абитуриент* ('*a person who just graduated from a high school*' in Bulgarian, but '*a person who just enrolled in an university*' in Russian) at rank 148. These pairs are problematic for our semantic similarity algorithms, since while the senses differ in Bulgarian and Russian, they are related – a door is a kind of entrance, and the newly admitted freshmen in an university are very likely to have just graduated from a high school.

The highest-ranked (misplaced) by COMBINED cognate is the pair *b. гордост / r. гордость* ('*pride*') at position 76. On the Web, this word often appears in historical and/or cultural contexts, which are nation-specific. As a result the word's contexts appear misleadingly different in Bulgarian and Russian.

In addition, when querying Google, we only have access to at most 1,000 top-ranked results. Since Google's ranking often prefers commercial sites, travel agencies, news portals, etc., over books, scientific articles, forum posts, etc., this introduces a bias on the kinds of contexts we extract.

7 Related Work

Many researchers have exploited the intuition that words in two different languages with similar or identical spelling are likely to be translations of each other.

Al-Onaizan&al.'99 [1] create improved Czech-English word alignments using probable cognates extracted with one of the variations of LCSR [24] described in [34]. Using a variation of that technique, Kondrak&al.'03 [17] demonstrate improved translation quality for nine European languages.

Koehn&Knight'02 [15] describe several techniques for inducing translation lexicons. Starting with unrelated German and English corpora, they look for (1) identical words, (2) cognates, (3) words with similar frequencies, (4) words with similar meanings, and (5) words with similar contexts. This is a bootstrapping process, where new translation pairs are added to the lexicon at each iteration.

Rapp'95 [31] describes a correlation between the co-occurrences of words that are translations of each other. In particular, he shows that if in a text in one language two words A and B co-occur more often than expected by chance, then in a text in another language the translations of A and B are also likely to co-occur frequently. Based on this observation, he proposes a model for finding the most accurate cross-linguistic mapping between German and English words using non-parallel corpora. His approach differs from ours in the similarity measure, the text source, and the addressed problem. In later work on the same problem, Rapp'99 [32] represents the context of the target word with four vectors: one for the words immediately pre-

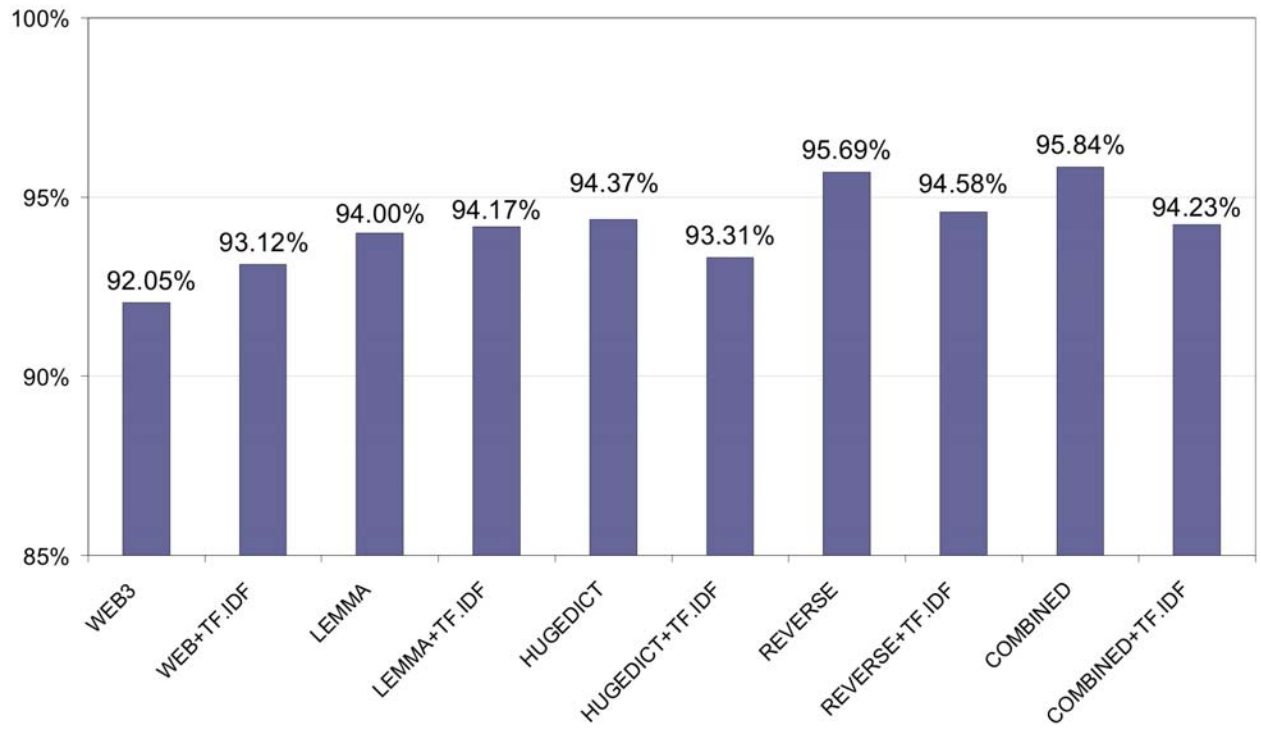

Fig. 5: Evaluation, 11-point average precision. *Different improvements of* WEB3.

ceding the target, another one for the ones immediately following the target, and two more for the words one more word before/after the target.

Fung&Yee'98 [12] extract word-level translations from non-parallel corpora. They count the number of sentence-level co-occurrences of the target word with a fixed set of "seed" words in order to rank the candidates in a vector-space model using different similarity measures, after normalisation and TF.IDF-weighting. The process starts with a small initial set of seed words, which are dynamically augmented as new translation pairs are identified. As we have seen above, an adaptation of this algorithm, SEED, yielded significantly worse results compared to WEB3. Another problem of that algorithm is that for a glossary of size $|G|$, it requires $4 \times |G|$ queries, which makes it too expensive for practical use. We tried another adaptation: given the words A and B, instead of using the exact phrase queries `"A B"` and `"B A"` and then adding the page hits, to use the page hits for `A and B`. This performed even worse.

Diab&Finch'00 [9] present a model for statistical word-level translation between comparable corpora. They count the co-occurrences for each pair of words in each corpus and assign a cross-linguistic mapping between the words in the corpora such that it preserves the co-occurrences between the words in the source language as closely as possible to the co-occurrences of their mappings in the target language.

Zhang&al.'05 [36] present an algorithm that uses a search engine to improve query translation in order to carry out cross-lingual information retrieval. They issue a query for a word in language A and tell the search engine to return the results only in language B and expect the possible translations of the query terms from language A to language B to be found in the title and summary of returned search results. They then look for the most frequently occurring word in the search results and apply TF.IDF-weighting.

Finally, there is a lot of research on string similarity that has been applied to cognate identification: Ristad&Yianilos'98 [33] and Mann&Yarowsky'01 [22] learn the MED weights using a stochastic transducer. Tiedemann'99 [34] and Mulloni&Pekar'06 [26] learn spelling changes between two languages for LCSR and for MEDR respectively. Kondrak'05 [16] proposes longest common prefix ratio, and longest common subsequence formula, which counters LCSR's preference for short words. Klementiev&Roth'06 [14] and Bergsma&Kondrak'07 [4] propose discriminative frameworks for string similarity. Rappoport&Levent-Levi'06 [23] learn substring correspondences for cognates, using string-level substitutions method of Brill&Moore'00 [7]. Inkpen&al.'05 [13] compare several orthographic similarity measures. Frunza&Inkpen'06 [11] disambiguate partial cognates.

While these algorithms can successfully distinguish cognates from false friends based on orthographic similarity, they do not use semantics and therefore cannot distinguish between equally spelled cognates and false friends (with the notable exception of [4], which can do so in some cases).

Unlike the above-mentioned methods, our approach:

- uses semantic similarity measure – not orthographical or phonetic;
- uses the Web, rather than pre-existing corpora to extract the local context of the target word when collecting semantic information about it;
- is applied to a different problem: classification of (nearly) identically-spelled false/true friends.

8 Conclusions and Future Work

We have proposed a novel unsupervised semantic method for distinguishing cognates from false friends, based on the intuition that if two words are cognates, then the words in their local contexts should be translations of each other, and we have demonstrated that this is a very promising approach.

There are many ways in which we could improve the proposed algorithm. First, we would like to automatically expand the bilingual glossary with more word translation pairs using bootstrapping as well as to combine the method with (language-specific) orthographic similarity measures, as done in [27]. We also plan to apply this approach to other language pairs and to other tasks, e.g. to improving word alignments.

Acknowledgments. We would like to thank the anonymous reviewers for their useful comments.

References

[1] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. Statistical machine translation. Technical report, 1999.

[2] L. Andreychin, L. Georgiev, S. Ilchev, N. Kostov, I. Lekov, S. Stoykov, and C. Todorov, editors. *Explanatory Dictionary of the Bulgarian Language*. 3 edition, 1973 (Л. Андрейчин, Л. Георгиев, Ст. Илчев, Н. Костов, Ив. Леков, Ст. Стойков и Цв. Тодоров, "*Български тълковен речник*". Издателство "Наука и изкуство", София, 1973).

[3] Y. Baltova and M. Charoleeva, editors. *Dictionary of the Bulgarian Language*, volume 11. 2002 ("*Речник на българския език*". под ред. на Ю. Балтова анд М. Чаролеева. Т. 11, Академично издателство "Проф. Марин Дринов", София, 2002).

[4] S. Bergsma and G. Kondrak. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[5] J. A. Bickford and D. Tuggy. Electronic glossary of linguistic terms (with equivalent terms in Spanish). http://www.sil.org/mexico/ling/glosario/E005aiGlossary.htm, April 2002. version 0.6. Instituto Lingüístico de Verano (Mexico).

[6] D. Bozhkov, V. Velchev, S. Vlahov, H. E. Rot, et al. *Russian-Bulgarian Dictionary*. 1985-1986 (Д. Божков, В. Велчев, С. Влахов, Х. Е. Рот и др., "*Руско-български речник*". Издателство "Наука и изкуство", София, 1985-1986).

[7] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of ACL*, pages 286–293, 2000.

[8] K. Cholakova, editor. *Dictionary of the Bulgarian Language*, volume 1-8. 1977-1995 ("*Речник на българския език*". под ред. на Кр. Чолакова. Т. 1-8, Издателство на БАН, София, 1977-2005).

[9] M. Diab and S. Finch. A statistical word-level translation model for comparable corpora. In *Proceedings of RIAO*, 2000.

[10] A. Evgenyeva, editor. *Dictionary of Russian in Four Volumes*, volume 1-4. 1981-1984 ("*Словарь русского языка в четырех томах*". под ред. А.П. Евгеньевой. Т. 1-4, Издательство "Русский язык", Москва, 1981-1984).

[11] O. Frunza and D. Inkpen. Semi-supervised learning of partial cognates using bilingual bootstrapping. In *Proceedings ACL'06*, pages 441–448, 2006.

[12] P. Fung and L. Y. Yee. An IR approach for translating from nonparallel, comparable texts. In *Proceedings of ACL*, volume 1, pages 414–420, 1998.

[13] D. Inkpen, O. Frunza, and G. Kondrak. Automatic identification of cognates and false friends in french and english. In *Proceedings of RANLP'05*, pages 251–257, 2005.

[14] A. Klementiev and D. Roth. Named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 82–88, New York City, USA, June 2006. Association for Computational Linguistics.

[15] 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.

[16] G. Kondrak. Cognates and word alignment in bitexts. In *Proceedings of the 10th Machine Translation Summit*, pages 305–312, Phuket, Thailand, September 2005.

[17] G. Kondrak, D. Marcu, and K. Knight. Cognates can improve statistical translation models. In *Proceedings of HLT-NAACL 2003 (companion volume)*, pages 44–48, 2003.

[18] G. Kondrak and T. Sherif. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of the Workshop on Linguistic Distances*, pages 43–50, Sydney, Australia, July 2006. Association for Computational Linguistics.

[19] V. Kyuvlieva-Mishaykova and M. Charoleeva, editors. *Dictionary of the Bulgarian Language*, volume 9. 1998 ("*Речник на българския език*". под ред. на В. Кювлиева-Мишайкова и М. Чаролеева. Т. 9, Академично издателство "Проф. Марин Дринов", София, 1998).

[20] V. Kyuvlieva-Mishaykova and E. Pernishka, editors. *Dictionary of the Bulgarian Language*, volume 12. 2004 ("*Речник на българския език*". под ред. на В. Кювлиева-Мишайкова анд Е. Пернишка. Т. 12, Академично издателство "Проф. Марин Дринов", София, 2004).

[21] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, (10):707–710, 1966.

[22] G. Mann and D. Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL'01*, pages 1–8, 2001.

[23] G. Mann and D. Yarowsky. Induction of cross-language affix and letter sequence correspondence. In *Proceedings of EACL Workshop on Cross-Language Knowledge Induction*, 2006.

[24] D. Melamed. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198, 1995.

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

[26] A. Mulloni and V. Pekar. Automatic detection of orthographic cues for cognate recognition. In *Proceedings of LREC-06*, pages 2387–2390, 2006.

[27] P. Nakov, S. Nakov, and E. Paskaleva. Improved word alignments using the web as a corpus. In *Proceedings of RANLP'07*, 2007.

[28] K. Panchev. *Differencial Russian-Bulgarian Dictionary*. 1963 (К. Панчев, "*Диференциален руско-български речник*". под ред. на С. Влахов и Г.А. Тагамлицка. Издателство "Наука и изкуство", София, 1963).

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

[30] E. Pernishka and L. Krumova-Cvetkova, editors. *Dictionary of the Bulgarian Language*, volume 10. 2000 ("*Речник на българския език*". под ред. на Е. Пернишка и Крумова-Цветкова. Т. 10, Академично издателство "Проф. Марин Дринов", София, 2000).

[31] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of ACL*, pages 320–322, 1995.

[32] R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of ACL*, pages 519–526, 1999.

[33] E. Ristad and P. Yianilos. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):522–532, 1998.

[34] J. Tiedemann. Automatic construction of weighted string similarity measures. In *Proceedings of EMNLP-VLC*, pages 213–219, 1999.

[35] A. Zaliznyak. *Grammatical Dictionary of Russian*. Russky yazyk, Moscow, 1977 (А. Зализняк, *Грамматический словарь русского языка*. "Русский язык", Москва, 1977).

[36] J. Zhang, L. Sun, and J. Min. Using the web corpus to translate the queries in cross-lingual information retrieval. In *IEEE NLP-KE*, pages 414–420, 2005.

r	**Candidate (BG/RU)**	**BG sense**	**RU sense**	**Sim.**	**Cogn.?**	**P@r**	**R@r**
1	муфта	gratis	muff	0.0085	no	100.00	1.00
2	багрене / багренье	mottle	gaff	0.0130	no	100.00	2.00
3	добитък / добыток	livestock	income	0.0143	no	100.00	3.00
4	мраз / мразь	chill	crud	0.0175	no	100.00	4.00
5	плет / плеть	hedge	whip	0.0182	no	100.00	5.00
6	плитка	plait	tile	0.0272	no	100.00	6.00
7	куча	doggish	heap	0.0287	no	100.00	7.00
8	лепка	bur	modeling	0.0301	no	100.00	8.00
9	кайма	minced meat	selvage	0.0305	no	100.00	9.00
10	низ	string	bottom	0.0324	no	100.00	10.00
11	геран / герань	draw-well	geranium	0.0374	no	100.00	11.00
12	печурка	mushroom	small stove	0.0379	no	100.00	12.00
13	ватман	tram-driver	whatman	0.0391	no	100.00	13.00
14	корейка	korean	bacon	0.0396	no	100.00	14.00
15	дума	word	thought	0.0398	no	100.00	15.00
16	товар	load	commodity	0.0402	no	100.00	16.00
17	катран	tar	sea-kale	0.0420	no	100.00	17.00
...	...	...	...	...	...	...	...
76	генератор	generator	generator	0.1621	yes	94.74	72.00
77	лодка	boat	boat	0.1672	yes	93.51	72.00
78	букет	bouquet	bouquet	0.1714	yes	92.31	72.00
79	прах / порох	dust	gunpowder	0.1725	no	92.41	73.00
80	врата	door	entrance	0.1743	no	92.50	74.00
81	клюка	gossip	cammock	0.1754	no	92.59	75.00
...	...	...	...	...	...	...	...
97	движение	motion	motion	0.2023	yes	83.51	81.00
98	компютър / компьютер	computer	computer	0.2059	yes	82.65	81.00
99	вулкан	volcano	volcano	0.2099	yes	81.82	81.00
100	година	year	time	0.2101	no	82.00	82.00
101	бут	leg	rubble	0.2130	no	82.12	83.00
102	заповедник	despot	reserve	0.2152	no	82.35	84.00
103	баба	grandmother	peasant woman	0.2154	no	82.52	85.00
...	...	...	...	...	...	...	...
154	мост	bridge	bridge	0.3990	yes	62.99	97.00
155	звезда	star	star	0.4034	yes	62.58	97.00
156	брат	brother	brother	0.4073	yes	62.18	97.00
157	мечта	dream	dream	0.4090	yes	61.78	97.00
158	дружество	association	friendship	0.4133	no	62.03	98.00
159	мляко / молоко	milk	milk	0.4133	yes	61.64	98.00
160	клиника	clinic	clinic	0.4331	yes	61.25	98.00
161	глина	clay	clay	0.4361	yes	60.87	98.00
162	учебник	textbook	textbook	0.4458	yes	60.49	98.00
...	...	...	...	...	...	...	...
185	корен / корень	root	root	0.5498	yes	54.35	100.00
186	психология	psychology	psychology	0.5501	yes	54.05	100.00
187	чайка	gull	gull	0.5531	yes	53.76	100.00
188	спалня / спальня	bedroom	bedroom	0.5557	yes	53.48	100.00
189	масаж / массаж	massage	massage	0.5623	yes	53.19	100.00
190	бензин	gasoline	gasoline	0.6097	yes	52.91	100.00
191	педагог	pedagogue	pedagogue	0.6459	yes	52.63	100.00
192	теория	theory	theory	0.6783	yes	52.36	100.00
193	бряг / берег	shore	shore	0.6862	yes	52.08	100.00
194	контраст	contrast	contrast	0.7471	yes	51.81	100.00
195	сестра	sister	sister	0.7637	yes	51.55	100.00
196	финанси / финансы	finances	finances	0.8017	yes	51.28	100.00
197	сребро / серебро	silver	silver	0.8916	yes	50.76	100.00
198	наука	science	science	0.9028	yes	50.51	100.00
199	флора	flora	flora	0.9171	yes	50.25	100.00
200	красота	beauty	beauty	0.9684	yes	50.00	100.00
				11-point average precision:		**92.05**	

Table 1: Ranked examples from our data set for WEB3*:* Candidate *is the candidate to be judged as being cognate or not,* Sim. *is the Web similarity score, r is the rank,* P@r *and* R@r *are the precision and the recall for the top r candidates.* Cogn.? *shows whether the words are cognates or not.*