Mach Translat (2007) 21:29–53
DOI 10.1007/s10590-008-9034-5

Methods for extracting and classifying pairs of cognates and false friends

Ruslan Mitkov · Viktor Pekar · Dimitar Blagoev · Andrea Mulloni

Received: 18 January 2007 / Accepted: 27 February 2008 / Published online: 17 May 2008

Abstract The identification of cognates has attracted the attention of researchers working in the area of Natural Language Processing, but the identification of false friends is still an under-researched area. This paper proposes novel methods for the automatic identification of both cognates and false friends from comparable bilingual corpora. The methods are not dependent on the existence of parallel texts, and make use of only monolingual corpora and a bilingual dictionary necessary for the mapping of co-occurrence data across languages. In addition, the methods do not require that the newly discovered cognates or false friends are present in the dictionary and hence are capable of operating on out-of-vocabulary expressions. These methods are evaluated on English, French, German and Spanish corpora in order to identify English–French, English–German, English–Spanish and French–Spanish pairs of cognates or false friends. The experiments were performed in two settings: (i) assuming 'ideal' extraction of cognates and false friends from plain-text corpora, i.e. when the evaluation data contains only cognates and false friends, and (ii) a real-world extraction scenario where cognates and false friends have to first be identified among words found in two comparable corpora in different languages. The evaluation results show that the

R. Mitkov (✉) · V. Pekar · A. Mulloni
Research Institute for Information and Language Processing, University of Wolverhampton, Stafford Street, Wolverhampton WV1 1SB, UK
e-mail: R.Mitkov@wlv.ac.uk

V. Pekar
e-mail: V.Pekar@wlv.ac.uk

A. Mulloni
e-mail: Andrea2@wlv.ac.uk

D. Blagoev
Mathematics and Informatics Department, University of Plovdiv, 4003 Plovdiv, Bulgaria
e-mail: gefix@pu.acad.bg

developed methods identify cognates and false friends with very satisfactory results for both recall and precision, with methods that incorporate background semantic knowledge, in addition to co-occurrence data obtained from the corpora, delivering the best results.

Keywords Cognates · Faux amis · Orthographic similarity · Distributonal similarity · Semantic similarity · Translational equivalence

1 Introduction

Cognates and false friends play an important role in the teaching and learning of a foreign language, as well as in translation. The existence of *cognates*, which are words that have similar meaning and spelling in two or more languages (e.g. Eng. *colour*, Spa. *color* and Fre. *couleur* 'colour', or Ger. *Bibliothek*, Fre. *Bibliothèque* and Spa. *Biblioteca* 'library'), helps students' reading comprehension and contributes to the expansion of their vocabularies. *False friends* (also known as *faux amis*), however, create problems and have the opposite effect, as they have similar spellings but do not share the same meaning (e.g. Eng. *library* as opposed to Spa. *librería* or Fre. *librairie* 'bookshop').

The identification of cognates and false friends finds application in many NLP tasks involving bilingual lexical knowledge, such as acquisition of bilingual lexicons from comparable corpora (Koehn and Knight 2002) and statistical machine translation (Simard et al. 1992; Melamed 1999). Other application areas include second language acquisition, where the ability of a student to distinguish cognates from false friends can contribute considerably to the process of learning a foreign language. In translation studies, cognates and false friends contribute to the notorious problem of source language interference for translators, and tools for automated quality control for translators that are able to highlight potential difficulties involving them are being developed, such as the TransCheck project.[1]

Unfortunately, a comprehensive list of cognates and false friends for a given pair of languages is often difficult to find, especially if one is interested in languages outside a few most widely spoken ones. Manual preparation of these lists requires a considerable amount of effort from trained lexicographers. Therefore, an attractive alternative is to automatically retrieve cognates or false friends from corpora for some given languages. The problem of automatic identification cognates has been in the focus of research for quite some time now (Brew and McKelvie 1996; Melamed 1999; Danielsson and Muehlenbock 2000; Kondrak and Dorr 2004; Bergsma and Kondrak 2007b, *inter alia*). The identification of false friends, however, has not received much attention, possibly because in many applications the discovery of translationally equivalent vocabulary items was the main goal and no special distinction has been made between false friends and pairs of simply non-equivalent expressions. This assumption seems also to be the reason why most of the approaches have primarily used orthographic and phonetic evidence for finding cognates, while the semantic

[1] http://rali.iro.umontreal.ca/Traduction/TransCheck.en.html

evidence has often been overlooked. A number of researchers specifically aimed to identify false friends, but used such evidence as co-occurrence of the expressions in parallel corpora (Brew and McKelvie 1996), similarity of dictionary definitions of the expressions (Kondrak 2001), and orthographic cues (Barker and Sutcliffe 2000). Only a few studies exist that attempt to compare the semantics of words in a pair in order to tell cognates from false friends, without the recourse to comprehensive lexical resources (Schulz et al. 2004; Mulloni et al. 2007; Nakov et al. 2007).

Our aim is to investigate ways of identifying pairs of cognates and false friends, that is, pairs of non-equivalent expressions that have etymologically motivated similarity in their orthography. By adopting such a view on false friends, we also aim to achieve an improved accuracy for identifying cognates. With the ultimate goal being to develop a tool for translators, language learners and lexicographers that would be capable of extracting lists of cognates and false friends from available texts, we propose and investigate a general framework for the discovery of such expressions from comparable corpora. The main practical advantages of the framework we propose are that it does not depend on the availability of large lexical resources, except for seed lists of known cognates and false friends and a bilingual dictionary encoding equivalence between only the basic vocabularies of the two languages. It includes a novel method for learning orthographic transformation rules that enhance the quality of extraction of words with etymologically motivated spelling and new methods for establishing translational equivalence between words of different languages based on comparable corpora.

The paper is organised as follows. In the next section we review related work. In Sect. 3 we describe the methodology behind our work, and the methods to measure orthographic and semantic similarity of words belonging to different languages. In Sect. 4 we present a method for learning orthographic transformation rules which is aimed at improving recognition of pairs of words with etymologically motivated similarities in orthography. In Sect. 5 we describe a number of methods to measure the semantic similarity between such expressions. Section 6 is concerned with the experimental evaluation of this method and of the framework as a whole. In Sect. 7 we discuss the results of the study and draw conclusions.

2 Previous work

The previous work on cognate identification can be split into three general areas of research: *orthographic approaches, phonetic approaches* and *semantic approaches*, which use some from of semantic evidence in addition to orthography or phonetics.

2.1 Orthographic approaches

Approximate string matching is the oldest method used to detect cognates. A simple and well-known approach is to measure the Edit Distance (ED) (Levenshtein 1965) between words of different languages sharing the same alphabet. ED returns a value corresponding to the minimum number of deletions, insertions and substitutions needed to transform the source language word into the target language word. Many

current approaches still implement ED in some form in their algorithms. Another popular and effective technique is the Longest Common Subsequence Ratio (LCSR) (Melamed 1999), the ratio of the length of their longest (not necessarily contiguous) common subsequence (LCS) and the length of the longer token.

Simple orthography-based heuristics for recognising cognates appear to work well for some tasks. Advocating the use of cognates to align sentences in bilingual corpora, Simard et al. (1992) recognise cognates just by considering the first letters of the words; they are taken to be cognates if their first four characters are identical. Though quite simple and straightforward, their algorithm has proven to result in better and more efficient sentence alignment. Danielsson and Muehlenbock (2000) detect cognates starting from aligned sentences in two languages. The match of two words is calculated as the number of matching consonants, allowing for one mismatched character.

Some studies have used manually constructed lists of orthographic transformation rules that assist identification of cognates prior to the use of a certain string matching technique (Barker and Sutcliffe 2000; Koehn and Knight 2002).

More recently a number of studies have attempted to apply statistical or machine learning techniques to learn orthographic correspondences between cognates in a certain language pair. Mann and Yarowsky (2001) investigate one function learned with stochastic transducers and another learned with a hidden Markov model in order to induce translation lexicons between cross-family languages via third languages, for subsequent expansion to intra-family languages using cognate pairs and cognate distance. Inkpen et al. (2005) identify cognates by testing several measures of orthographic similarity individually and then combine them using several different machine learning classifiers. In their exhaustive evaluation they obtain an accuracy on a test set as high as 95.39%. Mulloni and Pekar (2006) propose a methodology for the automatic detection of cognates based on the orthographic similarity of the two words. From a set of known cognates, their method induces rules that capture regularities in the orthographic mutation that a word undergoes when migrating from one language to the other. Bergsma and Kondrak (2007a) use Integer Linear Programming to form sets of cognates across groups of languages, introducing a transitivity constraint into the procedure, with the final goal of inducing the clustering of cognate pairs by supporting correct positive/negative decisions and penalising incorrect ones.

2.2 Phonetic approaches

Another group of approaches aims to recognise cognates based on the similarity in the phonetic form of the candidate cognates rather than their orthography. Guy (1994) represents the first attempt to implement this idea; the algorithm estimates the probability of phoneme correspondences by employing a variant of the chi-square statistic on a contingency table, which indicates how often two phonemes co-occur in words of the same meaning. Nerbonne and Heeringa (1997) developed distance functions based on binary features to compute relative distance between words from 40 pairs of Dutch dialects. They consider 14 methods, all based on simple (atomic character) and complex (feature vector) variants of ED. One aspect of their work worth particular

mention focused on the representation of diphthongs, with the results indicating that feature representations are more sensitive.

There are a number of approaches that use statistical and machine learning techniques operating on the phonetic representations of words. Knight and Graehl (1998) showed that it is possible to find phonetic cognates even between languages whose writing systems are as different as those of English and Japanese. Their weighted finite-state automaton efficiently represents a large number of transliteration probabilities between words written in the katakana and Latin alphabets, showing that standard finite-state techniques can efficiently find the most likely path through the automaton from a Japanese word written in katakana to an English word. Kondrak's (2000) algorithm for the alignment of phonetic sequences called ALINE determines a set of best local alignments that fall within a range of the optimal alignment in order to calculate the similarity of phonetic segments of two words.

Kondrak and Dorr (2004) combine orthographic and phonetic evidence when addressing the problem of the confusion between drug names that sound and look alike. They find that combining similarity of character n-grams with a phonetic module yields very high accuracy.

2.3 Semantic approaches

The next group of approaches represent work on combining evidence as to translational equivalence (i.e. similarity of their meanings) between cognates with their orthographic or phonetic similarity. These approaches also make it possible to detect false friends—in the sense that is adopted in the linguistics and language acquisition: words with etymologically motivated similarities in spelling, but different meanings.

In Brew and McKelvie (1996), the source of evidence about translational equivalence of the words is an aligned parallel corpus. They describe an application of sentence alignment techniques and approximate string matching to the problem of extracting lexicographically interesting word–word pairs from parallel corpora. They analyse six variants of the Dice coefficient, and find out that a variant called XXDice, combined with an association strength score such as likelihood ratio, achieves very good precision. A parallel corpus is used in Frunza and Inkpen (2006) in order to train a classifier to distinguish between cognate or false friend senses of a known partial cognate in a specific monolingual context. In Kondrak (2001), semantic similarity between the words is modelled by the similarity of their dictionary definitions.

The work that is most closely related to our proposed methods to determine the semantic similarity in a pair are the studies by Schulz et al. (2004) and Nakov et al. (2007), and as well as our own previous work described in Mulloni et al. (2007). Schulz et al. (2004) refine the list of cognates discovered with the help of purely orthographic techniques by measuring the distributional similarity between candidate cognates. Nakov et al. (2007) similarily use word co-occurrences extracted from unrelated monolingual text, exploiting the text snippets delivered by a search engine for this purpose. Mulloni et al. (2007) combine orthographic and semantic similarity measures, whereby semantic similarity between candidate cognates is approximated using both taxonomic and distributional similarity.

In this study we develop and evaluate new methods to measure semantic similarity in a pair of candidate cognates. Unlike previous work based on translating co-occurrence data into a different language, our methodology requires the translation of a much smaller set of words to establish equivalence in a pair. Our evaluation results show that this methodology frequently outperforms the methods based on translating co-occurrence vectors.

3 Methodology

The methodology for automatic identification of cognates and false friends that we propose in this paper is based on a two-stage process. The first stage involves the *extraction* of candidate pairs from non-parallel bilingual corpora whereas the second stage is concerned with the *classification* of the extracted pairs as cognates, false friends or unrelated words. This methodology has been tested on four language pairs: English–French, English–Spanish, English–German and Spanish–French.

The extraction of candidate pairs is based on comparing *orthographic similarity* between two words, whereas the classification of the extracted pair of two words is performed on the basis of their *semantic similarity*. Semantic similarity has been computed from taxonomies, or approximated from corpus data employing *distributional similarity* algorithms. In the following we introduce the measures of orthographic and semantic similarity employed in this study.

3.1 Orthographic similarity

Given that the extraction process involved comparison of any word from the first language with any word from the second, speed and efficiency was a major consideration, and hence the choice of a suitable, in our case, orthographic similarity measure/algorithm.[2] In this study, two orthographic similarity measures have been experimented with.

LCSR, as proposed by Melamed (1999), is computed by dividing the length of the longest common sub-sequence (LCS) by the length of the longer word, as in (1):

$$LCSR(w_1, w_2) = \frac{|LCS(w_1, w_2)|}{\max(|w_1|, |w_2|)} \quad (1)$$

For example, LCSR (*example, exemple*) = $\frac{6}{7}$ (their LCS is "e-x-m-p-l-e").[3]

Normalised Edit Distance (*NED*), a version of ED proposed in (Inkpen et al. 2005), is calculated by dividing the minimum number of edit operations needed to transform

[2] Measuring orthographic similarity is a commonly used method for distinguishing pairs of unrelated words from pairs of cognates and false friends. Inkpen et al. (2005) present a study of different measures and their efficiency.

[3] It should be noted that Longest Common Subsequence is different from Longest Common Substring (e.g. Oakes 1998), which for this pair would be "mple".

one word into another by the length of the longer string. Edit operations include substitutions, insertions and deletions.

3.2 Semantic similarity

There exist a number of semantic similarity measures that exploit the taxonomic structure of a thesaurus such as WordNet (see Budanitsky and Hirst 2001 for an overview), which are based on the intuition that semantic similarity between two words can be estimated from their distance in the taxonomic structure of a resource like WordNet. In this study we experiment with measures from Leacock and Chodorow (1998) and Wu and Palmer (1994), which performed better than other techniques in our pilot experiments. As the semantic thesaurus of the four languages under study we use EuroWordNet (Vossen et al. 1998).

Leacock and Chodorow's measure uses the normalised path length between the two concepts c_1 and c_2 and is computed as in (2):

$$sim_{LC}(c_1, c_2) = -\log\left[\frac{len(c_1, c_2)}{(2 \times MAX)}\right] \tag{2}$$

where *len* is the number of edges on the shortest path in the taxonomy between the two concepts and *MAX* is the depth of the taxonomy.

Wu and Palmer's measure is based on edge distance but also takes into account the most specific node dominating the two concepts c_1 and $c_2$as shown in (3):

$$sim_{WP}(c_1, c_2) = \frac{2 \times d(c_3)}{d(c_1) + d(c_2)} \tag{3}$$

where c_3 is the maximally specific superclass of c_1 and c_2, $d(c_3)$ is the depth of c_3 (the distance from the root of the taxonomy), and $d(c_1)$ and $d(c_2)$ are the depths of c_1 and c_2.

Each word, however, can have one or more meaning(s) (sense(s)) mapping to different concepts in the taxonomy. Using $s(w)$ to represent the set of concepts in the taxonomy that are senses of the word w, the word similarity can be defined as in (4) (cf. Resnik 1999):

$$wsim(w_1, w_2) = \max_{c_1 \in s(w_1),\, c_2 \in s(w_2)} [sim(c_1, c_2)] \tag{4}$$

3.3 Distributional similarity

Since taxonomies with wide coverage are not readily available, semantic similarity can also be modelled via word co-occurrences in corpora. Every word w_j is represented by the set of words $w_{i1\ldots n}$ with which it co-occurs. For deriving a representation of w_j, all occurrences of w_j and all words in the context of w_j are identified and counted. To account for the context of w_j, two general approaches can be taken: window-based

and syntactic. In the first, context is marked out by defining a window of a certain size around w_j, e.g. Gale et al. (1992) used a thousand-word window. In the second approach, the context is limited to words appearing in a certain syntactic relation to w_j, such as direct objects of a verb (Pereira et al. 1993; Grefenstette 1996). Once the co-occurrence data is collected, the semantics of w_j are modelled as a vector in an n-dimensional space where n is the number of words co-occurring with w_j, and the features of the vector are the probabilities of the co-occurrences established from their observed frequencies, as in (5):

$$C(w_j) = \langle P(w_j|w_{i1}), P(w_j|w_{i2}), \ldots, P(w_j|w_{in}) \rangle \tag{5}$$

Semantic similarity between words is then operationalised via the distance between their vectors. In the literature, various distance measures have been used including Euclidean distance, the cosine, Kullback-Leibler divergence and Jensen-Shannon divergence (see Manning and Schütze 1999 for an overview).

4 Extraction of candidate pairs

During the extraction stage the orthographic similarity between each noun from the first language (S) and a noun from the second language (T) is computed. Thus, the present method for the extraction of pairs that are either cognates or false friends depends on a certain affinity between the alphabets of the two languages. Following the computation of orthographic similarity, a list of the most similar word pairs is compiled, which we expect to contain pairs of cognates or false friends, but due to pre-processing errors, may contain pairs of unrelated words or errors such as words which are not orthographically similar or are not of the same part of speech.

While a high degree of orthographic similarity may well indicate that two words belonging to different languages are cognates,[4] many unrelated words may also have great similarity in spelling (e.g. Eng. *black* and Ger. *Block*). And vice versa, two words may be cognates, but their spellings may have little in common (e.g. Eng. *cat* and Ger. *Katze*). Our algorithm for the identification of candidate word pairs is based on the intuition that between any given pair of languages there are certain regularities in which the spelling of a word changes once it is borrowed from one language to the other.

In the following sections we describe an algorithm which learns orthographic transformation rules capturing such regularities from a list of known cognates (Sect. 4.1) and an algorithm for applying the induced rules to the discovery of potential cognates in a corpus (Sect. 4.2).

4.1 Learning algorithm

The learning algorithm involves three major steps: (a) the association of edit operations to the actual mutations that occur between two words known to be cognates (or

[4] Or, in fact, borrowings.

false friends); (b) the extraction of candidate rules; (c) the assignment of a statistical score to the extracted rules signifying their reliability. Its input is a list of translation pairs, which is then passed on to a filtering module based on NED; this allows for the identification of pairs of cognates/false friends C in two languages S and T, each consisting of a, word $w^S \in S$ and a word $w^T \in T$. The output of the algorithm is a set of rules R. At the start, two procedures are applied to the data: (i) edit operations between the two strings of the same pair are identified; (ii) NED between each pair is calculated in order to assign a score to each cognate pair. NED is calculated by dividing ED by the length of the longer string. NED—and normalisation in general—allows for more consistent values, since it was noticed that when applying standard ED, word pairs of short length (2 to 4 words each) would be more prone to be included in the cognate list even if they are actually unrelated (e.g. *at/an, tree/treu*). Sample output of this step for three English–German pairs is shown in Fig. 1.

At the next stage of the algorithm, a candidate rule c_r is extracted from each edit operation of each word pair in the training data. Each candidate rule consists of two sequences of letters, the former referring to language S and the latter pointing to its counterpart in language T. To construct a candidate rule, for each edit operation detected we use k symbols on either side of the edited symbol in both S and T. The left-hand side refers to the language S sequence, while the right-hand side gives the corresponding sequence in language T with the detected mutations. Table 1 illustrates rules detected in this manner. Candidate rules are extracted using different values of k for each kind of edit operation, with each value having been set experimentally. Substitution rules are created without considering the context around the letter being substituted, i.e. taking into account only the letter substitution itself, while deletions and insertions are sampled with k symbols on both sides. After extensive testing, k was empirically set to 2, whereas the candidate rule would vary in length both on the left-hand side and the right-hand side depending on the number of insertions and deletions it accounts for. This decision was supported by the fact that longer 'rules' are less frequent than shorter 'rules', but they are nonetheless more precise. In fact, because of the task at stake and the further areas we want to apply the algorithm to, we were somewhat more inclined towards obtaining higher precision rather than higher recall.

```
toilet/toilette
t     |o     |i     |l     |e     |t     |      |
t     |o     |i     |l     |e     |t     |t     |e
MATCH|MATCH|MATCH|MATCH|MATCH|MATCH|INS |INS

tractor/traktor
t     |r     |a     |c     |t     |o     |r
t     |r     |a     |k     |t     |o     |r
MATCH|MATCH|MATCH|SUBST|MATCH|MATCH|MATCH

absolute/absolut
a     |b     |s     |o     |l     |u     |t     |e
a     |b     |s     |o     |l     |u     |t     |
MATCH|MATCH|MATCH|MATCH|MATCH|MATCH|MATCH|DEL
```

Fig. 1 Edit operation association between English and German cognates

Table 1 Top-10 rules detected by the algorithm for the English–German language combination, along with the associated chi-square scores

Rule	Chi-square
c/k	386.878
d/t	345.699
ary/är	187.930
my/mie	187.930
hy/hie	187.930
gy/gie	172.510
ty/tät	167.517
et/ett	162.597
sh/sch	157.750
ive/iv	148.277

At the final stage, statistical scores are assigned to each unique candidate rule extracted, so that the significance of the rule can be assessed according to its actual occurrence in free text. After exploring different scoring functions (Fisher's exact test, chi-square, odds ratio and likelihood ratio), the chi-square for measuring the strength of the association between the left-hand side and the right-hand side of the candidate rule was chosen. The chi-square was calculated by considering the left, right and joint probabilities of each letter sequence compared to the total number of rules found. Once every candidate rule has been associated with a chi-square value, candidates falling below a specific threshold on the chi-square value are filtered, thus giving the final set of output rules.

4.2 Testing algorithm

The learning algorithm provides a set of rules which account for the orthographic behaviour of words between a source language and a target language. The second part of the algorithm (i.e. the testing algorithm) tries to deploy this kind of information (input) in the candidate extraction process.

Once the input data is made available, we proceed to apply the rules to each possible word pair, i.e. we substitute relevant n-grams in the rules with their counterpart in the target language. LCSR is then computed for every pair, and the top most similar pairs are added to the candidate cognate list. A case in point is represented by the English–German entry "*electric*/*elektrisch*"; the original LCSR is 0.700, but if the rules "c/k" and "*ic*/*isch*" detected earlier in the algorithm are applied, the new LCSR is 1.000.

5 Classification

The goal of the next stage of the methodology is to separate cognates from false friends from the lists obtained in the previous stage. To this end, the semantic similarity between the words in each pair in the training data is computed, and a numerical measure (threshold) is estimated. This threshold is later used to label the test data; if the similarity between the words in the test pair is lower than the threshold, the

words are returned as false friends—otherwise, they are returned as cognates. After a threshold is estimated, all pairs in the test data are labelled in this manner.

The presented methodology for separating cognates from false friends is language-independent in that it operates on any texts in any language-pair regardless of their typological relatedness. The texts also do not have to be parallel, although better results are expected if the corpora are comparable in their composition, size, period, and genre.

In order to establish the similarity between words in different languages, we experimented with four methods: Method 1 operating in 3 different variants (Method 1 without taxonomy, Method 1 with Leacock and Chodorow, and Method 1 with Wu and Palmer), and Method 2. These are outlined below.

5.1 Exploiting distributional similarities between words from the same language

Method 1 is based on the premise that if two words have the same meanings (and are cognates), they should be semantically close to roughly the same set of words in both (or more) languages, whereas two words which do not have the same meaning (and are false friends) will not be semantically close to the same set of words in both or more languages. If the two words being compared come from different languages, then the similarity between the two sets of nearest neighbours can be established with the help of a bilingual dictionary. Only the nearest neighbours, rather than the two words being compared, need to be present in the dictionary, which ensures that the method is capable of operating on out-of-vocabulary words.[5]

Method 1 can be formally described as follows. Start with two words in the two languages ($w^S \in S, w^T \in T$). Then calculate the N most similar words for each of the two words according to a chosen distributional similarity function. In this study, *skew divergence* (Lee 1999) was selected for Method 1 because it performed best during our pilot tests. Then build two sets of N words $W^S(w_1^S, w_2^S, \ldots, w_N^S)$ and $W^T(w_1^T, w_2^T, \ldots, w_N^T)$, such that w_i^S is the i-th most similar word to w^S and w_i^T is the i-th most similar word to w^T. Then a Dice Coefficient (Manning and Schütze 1999) function is applied over the two sets to determine their similarity. Those words from S and T are taken to be part of the intersection between the two sets of nearest neighbours if at least one translation of these words is found in the opposite language. Currently multiple translations of the same word are treated as having the same weight in the similarity function. Figure 2 illustrates the computation of the similarity between two cognates Eng. *article* and Fre. *article* 'article' (a) and two false friends Eng. *tenant* and Fre. *tenant* 'supporter, upholder' (b). For each word being compared the figure shows the sets of their 10 most similar words in the respective languages (N = 10); the arrows indicate equivalent pairs that have been looked up in a bilingual dictionary. In the first case, the similarity score is $(8 + 5)/20 = 0.65$.; in the second, it is $(1 + 1)/20 = 0.1$.

Note that N is a crucial parameter. If the value of N is too small, similar words may appear as very distant, because their common synonyms may not be present in

[5] Of course, if the two words themselves are in the dictionary, these can be just looked up quite straightforwardly.

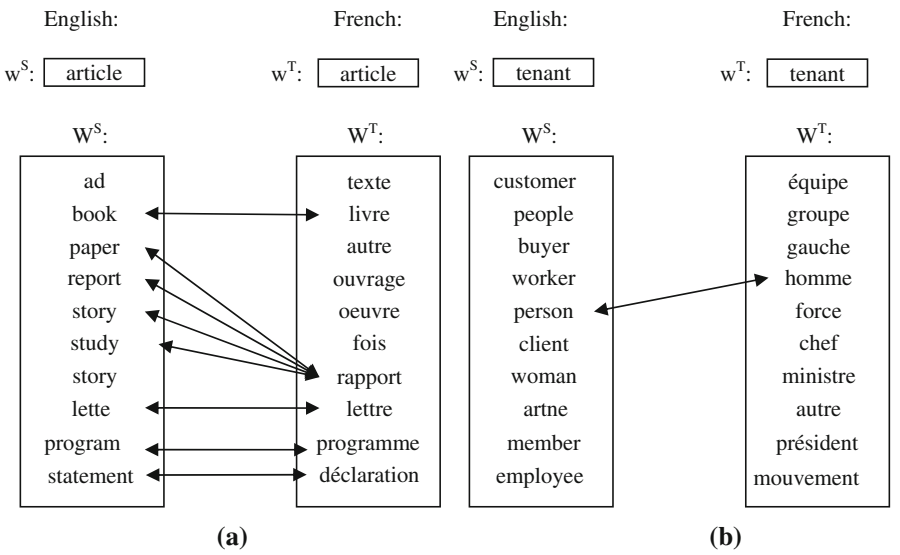

Fig. 2 Computing similarity between cognates and false friends using Method 1

both sets (or at all). If the value of N is too big, the sets of distant words may be filled with synonym word pairs that are distant from the initial one, thus making the words in the initial pair appear more similar than they actually are. The dictionary used can further affect the results.

In the evaluation section, the variant of Method 1 which employs skew divergence as the distributional similarity function as outlined above, is referred to simply as *Method 1*.

5.2 Exploiting distributional similarities between words across different languages

Method 2, which is inspired by work on acquisition of translation lexicons from comparable corpora (Tanaka and Iwasaki 1996; Fung 1998; Rapp 1999; Gaussier et al. 2004), models semantic similarity between words of two languages by mapping the space of distributional features of one language onto that of the other using a bilingual dictionary. The algorithm of Method 2 can be described as follows. First, co-occurrence data on each word of interest in both languages are extracted from the respective monolingual corpus parsed with a dependency parser. In the present study the semantics of nouns is represented by their co-occurrences with verbs as the heads of noun phrases serving as direct objects to the verbs. Thus, verbs are used as distributional features of the nouns. Once co-occurring verbs for each noun of interest have been extracted and counted, vector spaces for both sets of words are created, as described in Sect. 3.3. Then, given two words (w^S, w^T) in the languages S and T, $(w^S \in S, w^T \in T)$, the feature vector of w^S is translated into T with the help of a dictionary and then added to the co-occurrence data for T. The result is co-occurrence data that contains vectors for all words in the target language (T) plus the translated vector of the source

word (w^S). All words in T are then ranked according to their distributional similarity to the translated vector of w^S (using *skew divergence* again) and the rank of w^T is noted (R_1). The same measure is used to rank all words according to their similarity from w_2, taking only the rank of w^S(R_2). This is done because skew divergence is not symmetrical. The final result is the average of the two ranks $(R_1 + R_2)/2$.

Here the quality of the dictionary used for translation is essential. Besides the source data itself, one parameter that can make a difference is the direction of translation, i.e. which of the languages is the source and which is the target.

Method 1 (without taxonomy) and Method 2 can be regarded as 'related' in that the distributional similarity techniques employed both rely on context, but we felt it was worth experimenting with both methods with a view to comparing performance. In addition to making use of co-occurrence data, both methods also rely on dictionaries. However, one major difference is in the type of dictionaries needed. For the current task of comparing cognates/false friends which are nouns, Method 1 requires pairs of equivalent nouns from the bilingual dictionary, whereas Method 2 requires pairs of equivalent verbs.

5.3 Exploiting taxonomic similarities between words of the same language

If a pair of cognates/false friends under consideration are present in a monolingual taxonomical thesaurus, computing the semantic similarity directly from the taxonomy using a method such as Leacock and Chodorow (1998) or Wu and Palmer (1994) promises to be the best way forward. However, the absence of words in the thesaurus could result in a high precision but low recall. To overcome this limitation, a hybrid method was developed which works as follows; for any two words, their presence in a specific taxonomy is checked and if they are present, a taxonomical semantic similarity measure is employed to compute a similarity value. Otherwise distributional measures are used (as in Method 1 without taxonomy) to obtain a list of the N nearest neighbours for each word, i.e. words with the greatest distributional similarity to the given one. The taxonomy is used in this case instead of a dictionary.

The variant of Method 1 which employs Leacock and Chodorow's similarity measure is referred to in the evaluation section as *Method 1 with Leacock & Chodorow*, whereas the variant which computes similarity according to Wu and Palmer's measure is referred to as *Method 1 with Wu & Palmer*.

5.4 Threshold estimation

In order to determine the threshold on the similarity between the words in a pair that will be used to classify the pair as either cognates or false friends, the following threshold estimation techniques were used with the methods outlined above:

- *Mean average:* The distances between the words in both training sets (cognates and false friends) are measured using the chosen method. The mean average of the distances of the cognates and the mean average of the distances of the false friends are computed. The threshold is the average of the two means.

- *Median average:* As above, but median average is used instead of mean average.
- *Max accuracy:* All distances from the training data sets are analysed to find a distance which, when used as a threshold, would supposedly give a maximal accuracy for the test data. The accuracy is first separately computed for cognates and false friends, as in (6):

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

This is then averaged for the two types of word pairs. In calculating the accuracy of cognates, *TP* (true positives) is the number of cognates labelled as such by the system, *TN* (true negatives) is the number of non-cognates that the system did not label as cognates, *FP* (false positives) are non-cognates incorrectly labelled as cognates and *FN* (false negatives) are cognates that the system failed to identify. For false friends, the accuracy is computed in a similar manner. It should be noted that these threshold estimation methods are based on the assumption that cognates and false friends have equal importance for the classification task. This assumption is also reflected in the way the classification methods are evaluated; to compute the overall precision and recall, we average individual rates for cognates and false friends. If one wished to maximise the effectiveness with respect to only one class of expressions, the thresholds as well as the evaluation measures should be computed differently.

An evaluation framework was created, which simplifies and speeds up the process. The evaluation process splits the list of known cognates and false friends, and creates the training and test pairs for each test. The threshold estimation process uses only the training pairs to find a threshold that is sent to the classification process. The distance measurement process receives a word pair, and returns the distance according to the method and parameters used.

5.5 Evaluation settings

The evaluation is performed using 10-fold cross-validation. Given a test set of word pairs, the task is to classify each pair contained within as cognates or false friends, based on the measure of similarity and the threshold, obtained from the training set. The results were evaluated in terms of macroaverage recall and precision, since we wish to assign equal importance to the performance of the methods in relation to both cognates and false friends. To calculate a macroaverage recall, first recall rates for cognates and false friends are computed and then the two figures are averaged. Similarly, precision is computed as the average of the precision in identifying cognates and the precision in identifying false friends. The reported figures are averages of the macroaveraged precision and recall, calculated over the ten folds.

6 Experiments, evaluation and discussion

The experiments, extracting candidate pairs and classifying them, covered bilingual texts in 4 pairs of languages: English–French, English–German, English–Spanish and

French–Spanish. They were performed in two settings: (i) an 'ideal' environment where the pairs to be classified were either cognates or false friends, and (ii) in a real-world environment where the pairs to be classified as cognates and false friends were extracted from bilingual corpora. The former task would be the same as the latter in the case of perfect pre-processing, or as the task of classifying a pair as cognates or false friends from lists of orthographically (and semantically) close words.

In order to conduct the experiments described below, as well as bilingual dictionaries, co-occurrence statistics were needed to compare sets of similar words for two candidate cognates (Method 1), and perform the mapping between the vector spaces of two languages (Method 2). To extract co-occurrence data, we used the following corpora: the Wall Street Journal (1987–1989) part of the AQUAINT corpus for English, the Le Monde (1994–1996) corpus for French, the Tageszeitung (1987–1989) corpus for German and the EFE (1994–1995) corpus for Spanish. The English and Spanish corpora were processed with the Connexor FDG parser (Tapanainen and Järvinen 1997), French with Xerox Xelda, and German with Versley's parser (Versley 2005). From the parsed corpora we extracted verb–direct object dependencies, where the noun was used as the head of the modifier phrase. Because German compound nouns typically correspond to multiword noun phrases in the other three languages, they were split using a heuristic based on dictionary look-up and only the main element of the compound was retained (e.g. *Exportwirtschaft* 'export economy' was transformed into *Wirtschaft* 'economy').

EuroWordNet was used as the source of the four bilingual dictionaries. For Method 1, we extracted pairs of equivalent nouns, and for Method 2 pairs of equivalent verbs. In the latter case, the pairs were used to construct the translation matrix necessary for mapping distributional vectors into different languages. If, during the translation, a context word had multiple equivalents in the target language according to the dictionary, we followed previous practice (e.g. Fung and McKeown 1997) and mapped the source context word into all its equivalents, with its original probability equally distributed among them.

6.1 Extracting candidate pairs

For the task of extracting candidates from corpora, all combinations of word pairs for each of the four language pairs were compared in terms of LCSR orthographic similarity. The 500 most similar pairs were chosen as sample data for each language pair.

The lists were then manually annotated by trained linguists to provide both training data for the classifier and evaluation results for the extraction stage. The annotators were instructed to mark the pairs in terms of four categories:

- *Cognates* (word pairs with etymologically motivated similarities in both orthography and meaning, which would include both genetic cognates, e.g. Eng. *hound* and Ger. *Hund*, and borrowings from one language into another, e.g. Fre. *cognac* and Eng. *cognac*),

Table 2 Pairs returned in the automatic extraction process

Language pair	Cognates	False friends	Unrelated	Errors	Precision (%)
English–French	381	67	11	41	89.60
English–German	389	19	85	7	81.60
English–Spanish	370	47	28	55	83.40
French–Spanish	345	69	43	43	82.80
Average	371.25	50.5	41.75	36.5	84.35

- *False friends* (words that have etymologically motivated similarities of their orthography, but whose meanings have diverged so much that they are not translationally equivalent, e.g. Eng. *advertisement* and Spa. *advertencia* = 'warning'),
- *Unrelated* (words with no etymologically justified similarities in meaning or orthography), and
- *Errors* (words tagged with different or incorrect parts of speech, i.e. which resulted from incorrect PoStagging or parsing).

The accuracy of extraction of candidates was computed as the number of cognates and false friends divided by the total number of pairs (in our case 500), as this task is concerned with the identification of pairs that are either cognates or false friends. The results are satisfactory with an average precision of 84.35% (see Table 2). If the task is limited to finding cognates then the employed methodology can be quite effective, as the ratio of extracted cognates to false friends is more than 5:1.[6]

6.2 Classifying pairs as cognates or false friends

The extracted pairs were classified as either cognates or false friends. The evaluation was carried out in parameter-driven fashion, with the impact on the performance results of the following parameters being investigated.

- *Threshold estimation method*: How does the choice of the threshold estimation method affect the results?
- *The influence of N for Method 1*: What is the optimal number of nearest neighbours when measuring the semantic similarity between candidate cognates?
- *Direction of translation for Method 2*: When mapping feature vectors from one language to another, does the direction of translation have an effect on the performance?
- *The effect of errors in the extraction stage*: How does this affect the classification methods?

For each evaluation parameter (method, threshold estimation, etc.), the evaluation procedure splits the two samples (cognates and false friends) into 10 parts each, and runs a 10-fold cross-validation using the specified parameters.

[6] Note the much lower ratio of cognates to false friends for English–German, about 20:1. This cannot be a matter solely of language relatedness, especially as the ratio for French–Spanish is the highest. Trying to come up with a good explanation for this diversity is an avenue for future work.

As mentioned above, the evaluation was conducted in two settings: (i) *ideal setting* where the classification of cognates or false friends operated on a 'perfect input' of pairs that were either cognates or false friends, and (ii) *real-world setting* where the pairs to be classified as cognates and false friends (or unrelated) were automatically extracted from the corpora, and given some pre-processing errors, not all pairs were necessarily cognates or false friends.

6.2.1 Classification of cognates and false friends in ideal extraction settings

The evaluation under ideal extraction conditions was based on the assumption that the cognates/false friends extraction process was 100% accurate. For this purpose, the automatically extracted lists were post-edited to contain only cognates and false friends.

Given that the cognates are by far the largest class of pairs in this experiment, the baseline consists of assigning all pairs to the Cognates class. The recall rate for this baseline will be 50% for all language pairs (an average of 100% recall for cognates and 0% recall for false friends). The precision for cognates will be the proportion of true cognates among cognates and false friends found by the extraction step (i.e. the sum of the two columns in Table 2). These rates are 85% for English–French, 95% for English–German, 89% for English–Spanish, and 83% for French–Spanish. For false friends the precision is 0% for all the language pairs. Averaging over the two types of words yields 42.5%, 47.5%, 44.5%, and 41.5% overall precision rates for the respective language pairs.

Table 3 contains the precision and recall results achieved by Method 1 (M1), Method 2 (M2), Method 1 with Leacock & Chodorow (M1 + LC) and Method 1 with Wu & Palmer (M1 + WP) respectively, when using lists containing only either cognates or false friends.[7] The first column describes the classification methods, and each cell describes the best precision and recall rates achieved by the methods and their corresponding configuration settings (threshold estimation methods, the number of neighbours for Method 1, and the direction of translation for Method 2), for each language pair. The figures shown in bold represent the best results achieved on each language pair.

We see that on all language pairs, all methods invariably beat the baselines; the best performing methods gain 10 to 45% in terms of precision and 15 to 37% in terms of recall. The combinations of M1 with background semantic knowledge (M1+LC and M1+WP) produce better results than the methods that use only corpus data, with the exception of the English–French pair. Therefore, M1+LC leads to the best precision rates overall, while M1+WP gives the best recall rates. The differences between the four methods are not considerable; the difference between the best and the worst performing methods in precision is maximum 11% (English–Spanish), and in recall is maximum 16% (English–Spanish, as well).

Considering the performance of the methods on different language pairs, there does not seem to be any correlation between the typological closeness of the vocabularies

[7] Note that here and in all following tables, the statistical significance of the different results was not calculated.

Table 3 Best configurations of the methods in 'ideal extraction' settings

	En–Fr		En–Ge		En–Sp		Fr–Sp	
	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
M1	**65.80**	**62.87**	74.14	55.12	85.02	68.78	63.82	72.33
	Mean, 15	**M. Acc., 70**	Mean, 40	Median, 40	Median, 15	Median, 15	Mean, 150	M. Acc., 50
M2	59.72	57.03	63.46	53.07	70.26	71.20	58.82	72.29
	Mean, R	M. Acc., L	Mean, R	Mean, R	Mean, L	M. Acc., L	Mean, L	M. Acc., R
M1+LC	63.75	57.78	74.68	**57.75**	83.86	**89.28**	67.49	**74.75**
	Mean, 5	M. Acc., 150	Mean, 30	**Mean, 150**	Mean, 2	**M. Acc., 2**	Mean, 10	**M. Acc., 100**
M1+WP	61.62	56.30	**75.17**	55.03	**86.99**	80.36	**68.30**	65.56
	Mean, 15	Median, 7	**Median, 40**	Median, 40	**M. Acc., 50**	M. Acc., 40	**Mean, 3**	M. Acc., 15
Baseline	50.00	42.50	50.00	47.50	50.00	44.50	50.00	41.50

Table 4 Best configurations of the methods in 'real-world extraction' settings

	En–Fr		En–Ge		En–Sp		Fr–Sp	
	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
M1	**65.80**	**55.36**	74.14	50.52	85.02	58.79	63.82	52.38
	Mean, 15	**Mean, 15**	Mean, 40	Median, 40	Median, 15	Median, 15	Mean, 150	Mean., 150
M2	59.72	51.31	63.46	45.96	70.26	61.36	58.82	**60.08**
	Mean, R	Mean, R	Mean, R	Mean, R	Mean, L	M. Acc., L	Mean, L	**M. Acc., R**
M1+LC	63.75	53.53	74.68	51.94	83.86	**64.88**	67.49	54.79
	Mean, 5	Mean, 5	Mean, 30	Median, 2	Mean, 2	**M. Acc., 2**	Mean, 10	Median, 70
M1+WP	61.62	52.07	**75.17**	**52.63**	**86.99**	63.08	**68.30**	54.99
	Mean, 15	Mean, 15	**Median, 40**	**Mean, 15**	**M. Acc., 50**	M. Acc., 100	**Mean, 3**	Mean, 100
Baseline	50.00	42.50	50.00	47.50	50.00	44.50	50.00	41.50

of the languages and the results achieved; the best results overall are attained on the English–Spanish pair, languages which are generally considered to be less related than English–French or French–Spanish, for example, where results are noticeably lower. The most likely reason behind this observation seems to be the fact that there are other factors at play, namely those that contribute to the 'comparability' of the corpora in each language pair as well as the amount of corpus evidence and the quality of the semantic resource used.

6.2.2 Classification of cognates and false friends in the fully automatic mode

Table 4 reports the classification performance in a fully automatic mode, that is, involving extraction of pairs using orthographic similarity before classification and, as a result, including pairs resulting from linguistic processing errors. The best results and configuration settings for each language pair are shown in bold.

As expected, these results are lower than those achieved in the ideal extraction settings, but are still higher than the baseline by a considerable margin. The improvements on the baseline in terms of precision are 5 to 20% and in terms of recall 15 to 37%. As in the ideal extraction settings, the methods incorporating background semantic knowledge fare generally better than those using corpus data alone, with the exceptions being the English–French pair, where M1 leads to the best results, and the French–Spanish pair, where M2 achieves the best precision rates. The drop in precision in comparison with Table 3 is between 5% (English–French) and 24% (English–Spanish). The recall rates for the best configurations, however, remain largely the same.

6.2.3 Parameters of the methods

In this section, we present the evaluation results in relation to each of the parameters of the methods described in Sect. 6.2: the effect of the threshold estimation method, the number of nearest neighbours for Method 1 and the direction of translation in Method 2. All these experiments were carried out in the real-world extraction setting.

Threshold estimation: Table 5 compares different methods of estimating the threshold on the similarity measure to separate cognates from false friends. For each of the four classification methods, it describes the results achieved with three different threshold estimation methods, with the best results among the three shown in bold.

We see that *Mean* and *Median* methods very often lead to the best results, and that the differences between them are rather insignificant. Contrary to our expectations, the thresholds that maximise the accuracy of the methods on the training data do not deliver the best results; most of the time the results are clearly inferior to those of the other two methods and in only a few cases they are only slightly better than the alternatives.

The influence of N for M1: The number of the most similar words (N) which M1 uses to measure similarity between pairs is an essential parameter for this method. Figures 3 and 4 report recall and precision for each N when evaluating English–

Table 5 A comparison of threshold estimation methods

			En–Fr		En–Ge		En–Sp		Fr–Sp	
			Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
M1	10	Mean	62.16	53.33	57.51	**48.81**	**78.74**	55.69	61.29	**51.65**
		Median	**64.83**	**54.41**	**63.19**	47.27	78.14	**55.98**	**61.34**	51.57
		M. Acc.	50.00	38.11	50.00	38.91	50.00	37.01	50.69	36.87
M2	L	Mean	**56.59**	**49.43**	**55.49**	44.45	**70.26**	53.15	**58.82**	48.78
		Median	55.67	49.14	53.66	**45.30**	66.80	50.72	57.53	**49.71**
		M. Acc.	52.58	48.66	49.87	38.96	55.57	**61.36**	50.55	37.33
M1+LC	10	Mean	**59.02**	**51.03**	**68.42**	**50.70**	**79.79**	58.05	**67.49**	54.14
		Median	58.57	50.68	66.77	50.66	77.81	57.00	67.14	**54.52**
		M. Acc.	49.61	38.28	50.04	39.34	76.67	**63.98**	57.39	48.64
M1+WP	10	Mean	57.44	**50.36**	**67.24**	**51.88**	85.12	59.41	67.29	54.49
		Median	**57.54**	50.16	60.87	49.91	**86.55**	59.81	**67.33**	**54.50**
		M. Acc.	50.00	38.11	50.00	38.91	82.33	**60.64**	63.62	51.20

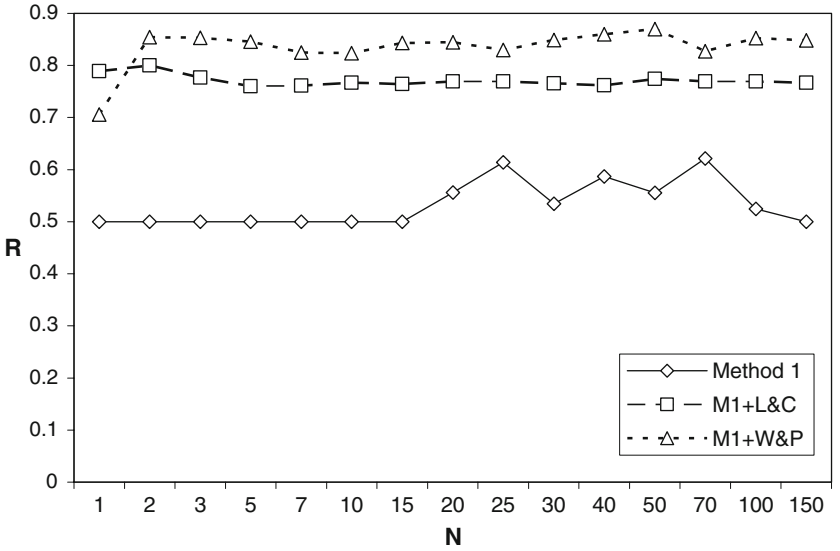

Fig. 3 The recall of M1 for English–Spanish without pre-processing errors, using max. accuracy estimation

Spanish[8] cognates with Max accuracy as threshold estimation. When M1 incorporates background semantic knowledge (M1+LC and M1+WP), precision rises for smaller values of N ($N > 5$), and further increases in N do not bring any improvement. A similar picture is observed for recall, except that for M1+LC, varying N does not lead to substantially different results. It is noteworthy that at different values of N, the two methods behave very similarly.

[8] In this section we chose to report the evaluation results for one language pair, since the relationships between the number of N, on the one hand, and the precision and recall rates, on the other, are similar for different language pairs.

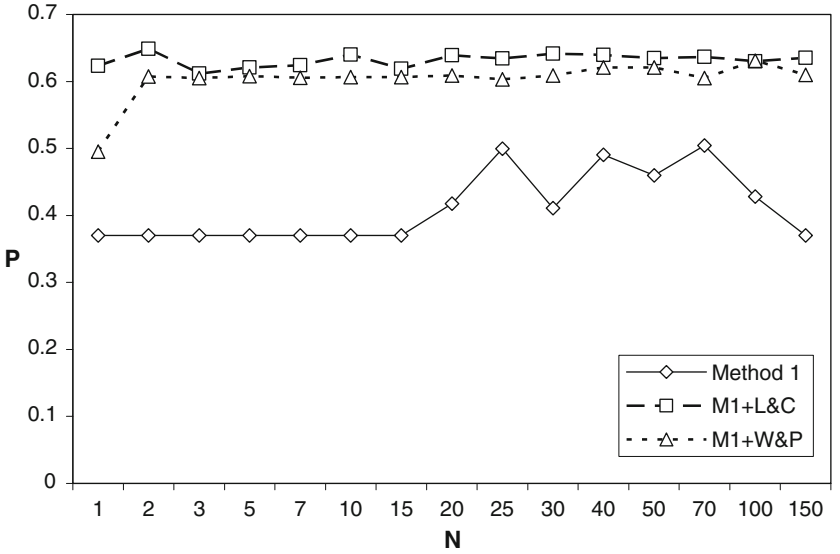

Fig. 4 The precision of M1 for English–Spanish without pre-processing errors, using max. accuracy estimation

When Method 1 is applied in isolation, i.e. without the background knowledge, the picture is quite different; both precision and recall stay mainly the same for $N < 15$, but for larger values of N they improve with the peaks reached at N between 25 and 100. For no values of N, however, does M1 outperform its variants that make use of background knowledge.

Direction of translation for M2: Since the vocabularies of the two corpora in a language pair may be characterised by different degrees of polysemy, it was interesting to examine whether the choice of the direction in which the vector spaces are translated in Method 1 has an effect on the quality of the classification results. Figures 5 and 6 illustrate this effect ("L" stands for left-to-right translation in a language pair and "R" for right-to-left).

We find that in terms of recall there is hardly any consistency in the difference between the two directions of translation; within the same language pair each direction may result in better results than the other depending on the threshold estimation method. A possible exception is the English–German pair when the "left" direction produces better results than the opposite one. The differences depending on the direction are quite small; most of them are not larger than 2–3%, the greatest difference being 7% (English–German pair).

In terms of precision, the differences between the two directions are again not consistent across threshold estimation methods, within each language pair. The choice of translation direction does seem to matter at least for some language pairs; for French–Spanish, the difference is up to 25%, and for English–German it is up to 15%. In most cases, however, the differences are quite small and do not exceed 2–3%.

M1 vs. M2: Method 1 and Method 2 are different ways of measuring the semantic similarity of words from different languages based only on co-occurrence data

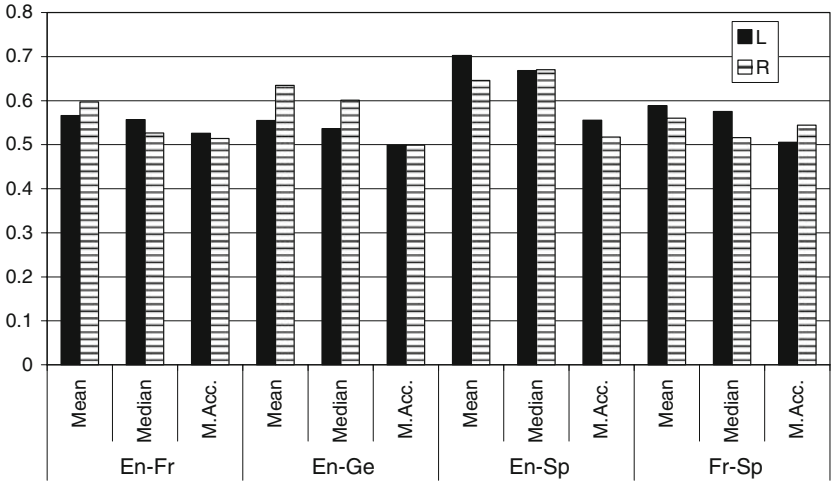

Fig. 5 Recall of M2 for different translation directions

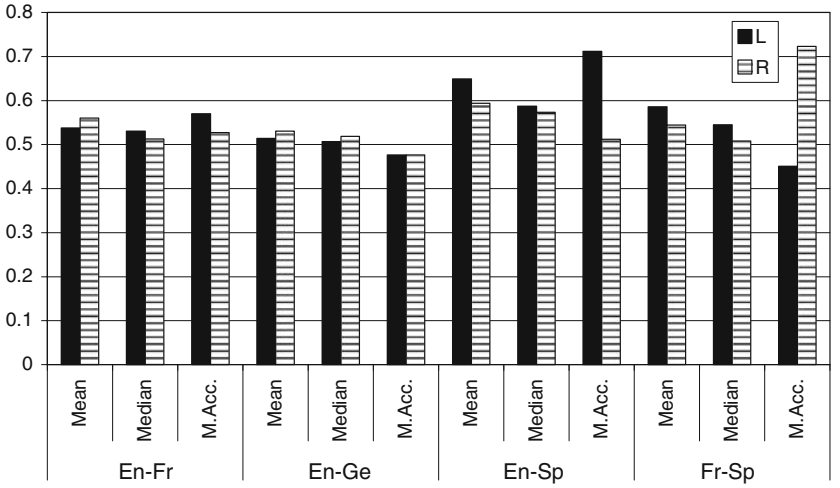

Fig. 6 Precision of M2 for different translation directions

extracted from the corpus and a bilingual dictionary, possibly encoding only a small portion of the vocabularies of the two languages; while Method 1 uses the dictionary to establish equivalence between sets of nearest neighbours determined from their co-occurrence data, Method 2 uses it to map the co-occurrence data of one language onto those of the other. Figure 7 compares these two methods, showing the precision and recall rates achieved by them on different language pairs as well as their most optimal configurations.

The differences in precision are quite small (between 2–3%), with M1 faring better on English–French and English–German pairs and M2 on the other two language pairs. In terms of recall, M1 is clearly superior, gaining up to 25% on M2. This suggests that

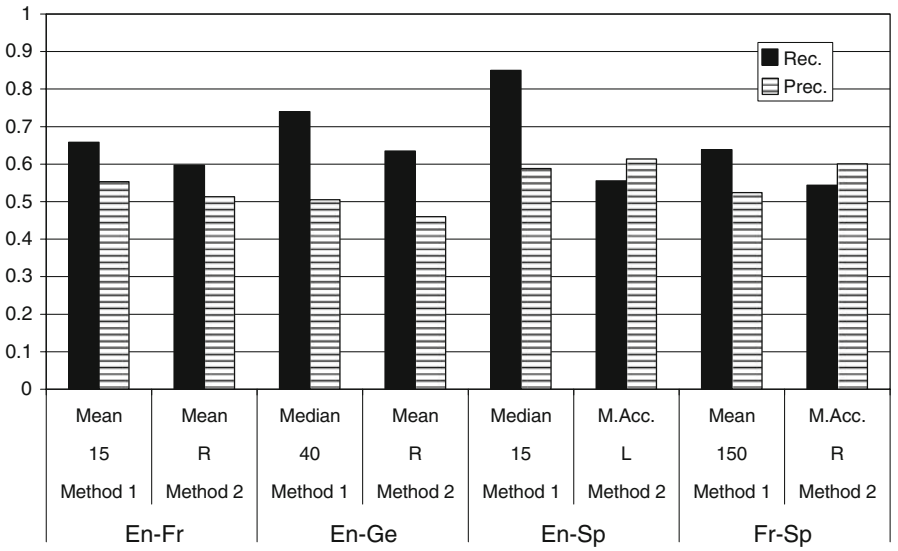

Fig. 7 A comparison of Method 1 and Method 2

while the two methods do not differ much in terms of precision, it is highly advisable to use M1 in cases when high recall is important.

7 Conclusion

This paper proposes several novel methods for automatic identification of cognates and false friends from corpora, neither of which are dependent on the existence of parallel texts or any other broad-coverage resources. It is sufficient that the bilingual dictionary required by the methods contain only the basic vocabulary of the two languages, which is only needed to establish equivalence between a set of nearest neighbours for test words (Method 1), or between distributional vectors of test words (Method 2). Unlike previous work based on translating co-occurrence data into a different language, our methodology requires the translation of a much smaller set of words in order to establish equivalence in a pair. The extensive evaluation results cover a variety of parameters and show that automatic identification of cognates and false friends from corpora is a feasible task, and all proposed methods perform in a satisfactory manner. The best results which appear to be consistent across the language pairs are obtained by Method 1, which is based on the premise that cognates are semantically closer than false friends, when the variant of the method using a taxonomy together with Wu and Palmer's measure is employed.

Another contribution of the paper is a new method of discovering orthographically similar words across two languages. This method makes it possible to select pairs of potential cognates and false friends from two unrelated monolingual corpora, so that at later stages equivalent expressions are discovered with greater accuracy and coverage. As a result of applying this method prior to the separation of pairs into cognates or false friends, the false friend pairs output by the method are not random pairs of

words, but words that have very similar orthographies and hence closely correspond to the notion of *faux amis* adopted in the field of second language acquisition.

References

Barker G, Sutcliffe R (2000) An experiment in the semi-automatic identification of false-cognates between English and Polish. In: Proceedings of the 11th Irish conference on artificial intelligence and cognitive science. Galway, Ireland, pp 597–606

Bergsma S, Kondrak G (2007a) Multilingual cognate identification using integer linear programming. In: Proceedings of the international workshop on acquisition and management of multilingual lexicons. Borovets, Bulgaria, pp 11–18

Bergsma S, Kondrak G (2007b) Alignment-based discriminative string similarity. In: Proceedings of the 45th annual meeting of the association for computational linguistics. Prague, Czech Republic, pp 656–663

Brew C, McKelvie D (1996) Word-pair extraction for lexicography. In: Proceedings of the second international conference on new methods in language processing. Ankara, Turkey, pp 45–55

Budanitsky A, Hirst G (2001) Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: Proceedings of the workshop on WordNet and other lexical resources second meeting of the North American chapter of the association for computational linguistics (NAACL-2001). Pittsburgh, PA, pp 29–34

Danielsson P, Muehlenbock K (2000) Small but efficient: the misconception of high-frequency words in Scandinavian translation. In: Envisioning machine translation in the information future, 4th conference of the association for machine translation in the Americas (AMTA 2000), LNCS vol 1934. Springer Verlag, Berlin, pp 158–168

Frunza O, Inkpen D (2006) Semi-supervised learning of partial cognates using bilingual bootstrapping. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. Sydney, Australia, pp 441–448

Fung P (1998) Statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In: Machine translation and the information soup, third conference of the association for machine translation in the Americas, LNCS vol 1529. Springer Verlag, Berlin, pp 1–17

Fung P, McKeown K (1997) Finding terminology translation from non-parallel corpora. In: Proceedings of the 5th annual workshop on very large corpora. Hong Kong, August 1997, pp 192-202

Gale WA, Church KW, Yarowsky D (1992) A method for disambiguating word senses in a large corpus. Comput Human 26:415–439

Gaussier E, Renders J-M, Matveeva I, Goutte C, Déjean H (2004) A geometric view on bilingual lexicon extraction from comparable corpora. In: ACL-04: 42nd annual meeting of the association for computational linguistics, proceedings. Barcelona, Spain, pp 526–533

Grefenstette G (1996) Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In: Boguarev B, Pustejovsky J (eds) Corpus processing for lexical acquisition. MIT Press, Cambridge, MA, pp 205–216

Guy J (1994) An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. J Quant Linguist 1(1):35–42

Inkpen D, Frunza O, Kondrak G (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' 05). Borovets, Bulgaria, pp 251–257

Knight K, Graehl J (1998) Machine transliteration. Comput Linguist 24(4):599–612

Koehn P, Knight K (2002) Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In: Proceedings of the 17th national conference on artificial intelligence (AAAI). Austin, TX, pp 711–715

Kondrak G (2000) A new algorithm for the alignment of phonetic sequences. In: Proceedings of NAACL/ANLP 2000: 1st conference of the North American chapter of the association for computational linguistics and 6th conference on applied natural language processing. Seattle, WA, pp 288–295

Kondrak G (2001) Identifying cognates by phonetic and semantic similarity. In: Proceedings of the second meeting of the North American chapter of the association for computational linguistics (NAACL 2001). Pittsburgh, PA, pp 103–110

Kondrak G, Dorr B (2004) Identification of confusable drug names: a new approach and evaluation methodology. In: Coling: 20th international conference on computational linguistics, proceedings. Geneva, Switzerland, pp 952–958

Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. In: Fellbaum C (ed) WordNet: an electronic lexical database. MIT Press, Cambridge, MA, pp 265–283

Lee L (1999) Measures of distributional similarity. 37th Annual meeting of the association for computational linguistics, College Park, MD, 25–32

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

Mann G, Yarowsky D (2001) Multipath translation lexicon induction via bridge languages. In: Proceedings of the second meeting of the North American chapter of the association for computational linguistics (NAACL 2001). Pittsburgh, PA, pp 151–158

Manning C, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, MA

Melamed D (1999) Bitext maps and alignment via pattern recognition. Comput Linguist 25(1):107–130

Mulloni A, Pekar V (2006) Automatic detection of orthographic cues for cognate recognition. In: Proceedings of the 5th international conference on language resources and evaluation (LREC-06). Genoa, Italy, pp 2387–2390

Mulloni A, Pekar V, Mitkov R, Blagoev D (2007) Semantic evidence for automatic identification of cognates. In: Proceedings of the 1st international workshop on acquisition and management of multilingual lexicons. Borovets, Bulgaria, pp 49–54

Nakov S, Nakov P, Paskaleva E (2007) Cognate or false friend? Ask the web! In: Proceedings of the 1st international workshop on acquisition and management of multilingual lexicons. Borovets, Bulgaria, pp 55–62

Nerbonne J, Heeringa W (1997) Measuring dialect distance phonetically. In: Proceedings of the third workshop on computational phonology, special interest group of the association for computational linguistics (SIGPHON-97). Madrid, Spain, pp 11–18

Oakes MP (1998) Statistics for corpus linguistics. Edinburgh University Press, Edinburgh, UK

Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words. 31st Annual meeting of the association for computational linguistics, Columbus, OH, pp 183–190

Rapp R (1999) Automatic identification of word translations from unrelated English and German corpora. 37th Annual meeting of the association for computational linguistics. College Park, MD, pp 519–526

Resnik P (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artif Intel Res (11):95–130

Schulz S, Marko K, Sbrissia E, Nohama P, Hahn U (2004) Cognate mapping—a heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. Coling: 20th international conference on computational linguistics, proceedings Geneva, Switzerland, pp 813–819

Simard M, Foster G, Isabelle P (1992) Using cognates to align sentences in bilingual corpora. In: Fourth international conference on theoretical and methodological issues in machine translation, empiricist vs. rationalist methods in MT, TMI-92, Proceedings, Montreal, Canada, pp 67–81

Tanaka K, Iwasaki H (1996) Extraction of lexical translations from non-aligned corpora. In: Proceedings of COLING 96: the 16th international conference on computational linguistics. Copenhagen, Denmark, pp 580–585

Tapanainen P, Järvinen T (1997) A non-projective dependency parser. In: Proceedings of the 5th conference on applied natural language processing. Washington D.C., Association of Computational Linguistics, pp 64–71

Versley Y (2005) Parser evaluation across text types. In: Proceedings of the 4th workshop on treebanks and linguistic theories (TLT 2005). Barcelona, Spain, pp 209–220

Vossen P, Bloksma L, Boersma P, Verdejo F, Gonzalo J, Rodriquez H, Rigau G, Calzolari N, Peters C, Picchi E, Montemagni S, Peters W (1998) EuroWordNet Tools and resources report. Technical report LE-4003, University of Amsterdam, The Netherlands (http://dare.uva.nl/record/157403)

Wu Z, Palmer M (1994) Verb semantics and lexical selection. 32nd Annual meeting of the association for computational linguistics. Las Cruces, NM, pp 133–138