Mach Translat (2006) 20:247–266
DOI 10.1007/s10590-007-9029-7

Finding translations for low-frequency words in comparable corpora

Viktor Pekar · Ruslan Mitkov · Dimitar Blagoev · Andrea Mulloni

Received: 3 July 2007 / Accepted: 5 December 2007 / Published online: 23 February 2008

Abstract Statistical methods to extract translational equivalents from non-parallel corpora hold the promise of ensuring the required coverage and domain customisation of lexicons as well as accelerating their compilation and maintenance. A challenge for these methods are rare, less common words and expressions, which often have low corpus frequencies. However, it is rare words such as newly introduced terminology and named entities that present the main interest for practical lexical acquisition. In this article, we study possibilities of improving the extraction of low-frequency equivalents from bilingual comparable corpora. Our work is carried out in the general framework which discovers equivalences between words of different languages using similarities between their occurrence patterns found in respective monolingual corpora. We develop a method that aims to compensate for insufficient amounts of corpus evidence on rare words: prior to measuring cross-language similarities, the method uses same-language corpus data to model co-occurrence vectors of rare words by predicting their unseen co-occurrences and smoothing rare, unreliable ones. Our experimental evaluation demonstrates that the proposed method delivers a consistent and significant improvement on the conventional approach to this task.

V. Pekar · R. Mitkov (✉)
ILP, University of Wolverhampton, Stafford St., Wolverhampton WV1 1SB, UK
e-mail: r.mitkov@wlv.ac.uk

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

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

A. Mulloni
Expert System, Via F. Zeni 8, Rovereto 38068, Italy
e-mail: amulloni@expertsystem.it

Keywords Lexical acquisition · Translational equivalents · Comparable corpora · Distributional similarity · Data sparseness

1 Introduction

Broad-coverage, up-to-date dictionaries are key to multilingual information technology. However, as with any sort of hand-encoded linguistic resources, they are extremely expensive to build and maintain, requiring the effort of specially trained lexicographers. Parallel corpora (texts in two languages where the original is aligned with its translation at the sentence level) can be used to extract new dictionary entries with high accuracy (Dagan and Church 1997; Tiedemann 1998; Melamed 2000). Unfortunately, such methods themselves suffer from the acquisition bottleneck: a large amount of parallel text needs to be produced by translators before it can be used for the discovery of new dictionary entries.

Bilingual comparable corpora are an alternative, which potentially have very attractive properties such as much greater possibilities for domain- and language-portability for a lexical acquisition system. Recognising this potential, in recent years researchers have begun to explore possibilities of using comparable corpora for this purpose (Fung 1995; Tanaka and Iwasaki 1996; Fung and McKeown 1997; Rapp 1999; Déjean et al. 2002; Gaussier et al. 2004; Robitaille et al. 2006; Morin et al. 2007 inter alia). Bilingual comparable corpora are collections of documents that are not translations of each other, but are characterised by the same topical composition and style of presentation. In contrast to their parallel counterparts, it is quite easy to obtain comparable corpora of the required specification (in terms of domain specialisation, discourse type, size, origin, period, etc.). Especially with the advent of methodologies and tools to automatically build customised corpora from the web (e.g. Baroni and Berdardini 2004; Fletcher 2004), one can quickly construct a corpus that would serve as an ample source of usage examples of new terms that are of interest for the lexicographer.

The key assumption behind comparable corpora approaches is that translationally equivalent expressions exhibit similar occurrence patterns in the respective monolingual corpora. The general procedure begins by collecting co-occurrence data on words of potential interest and representing them as context vectors. After that context vectors of different languages are mapped onto a single vector space using a bilingual dictionary. Translation equivalents are then retrieved as pairs of words that have the greatest similarity in their vectors.

Unfortunately, the accuracy of existing methods is decidedly suboptimal, if they were to operate in a fully automatic mode. Fung and McKeown (1997) report a precision of 0.58 when the correct English translation of a source Japanese word is found anywhere among the top-100 candidates. The best method in Gaussier et al. (2004) achieves an F1 score of only 0.32, whereby the retrieval of the equivalent was taken to be successful when it appeared among the 100 highest-ranking candidates. Morin et al. (2007) report 30 and 42% correct equivalents found in the top-10 and top-20 system-proposed candidates. The accuracy of the approach appears to improve considerably when one introduces a high frequency cut-off on the words being aligned. Requiring that words among which a translation is sought have a corpus frequency of

100 or higher, Rapp (1999) obtains very promising results: 72% of correct translations were the first-ranked candidates; 89% of translations were found among the top-10 candidates. Chiao and Zweigenbaum (2002) used the cut-off of 100 on English words and 60 on French words and were able to find the correct translation at the very top of the candidate list in 20% of cases, and among the top-20 candidates in 60% of cases.

The results of such previous work suggest that the extraction of equivalents from comparable corpora is quite unreliable on all but the most frequent words. As is known from work on monolingual lexical acquisition, the amount of corpus data is an important factor for the co-occurrence model of word meaning (Curran 2004). In the bilingual context, the data sparseness problem appears to be even more severe. A possible explanation is that the translation between the matrices of different languages introduces a lot of noise, coming from the many-to-many mapping between the vocabularies of the languages, and only more frequent words remain relatively robust against this noise. The fact that the approach depends on the words being very frequent renders it rather impractical for its intended purpose: it is infrequent words such as neologisms or narrow-domain terminology that one is primarily interested in when compiling or updating a bilingual dictionary.

In this paper we investigate ways to improve the accuracy of retrieval of translation equivalents for low-frequency words from comparable corpora. We develop an extension of the similarity-based method for estimating word co-occurrence probabilities (Pereira et al. 1993; Dagan et al. 1999; Lee 1999) to the problem of modelling additional context features of rare words. Prior to translating the vectors of source words into the vector space of a different language, our method uses same-language corpus data to predict unseen and smooth unreliable co-occurrences of rare words.

The organisation of the paper is as follows. In Sect. 2, we describe in more detail the standard procedure for finding equivalents in comparable corpora. In Sect. 3 we look at possible solutions to deal with the data sparseness problem in the context of the present task. In Sect. 4 we consider the distance-based averaging method, and in Sect. 5 describe its two proposed modifications. Sections 6 and 7 are devoted to the experimental evaluation of our method of dealing with low-frequency words. Section 8 presents a brief overview of related work. Section 9 summarises conclusions from this study.

2 Translation equivalents in comparable corpora

While quite a broad range of approaches are described in the literature, most of them follow the same general algorithm, which can be described as follows. Given a source word n and a set of words from the target language M, the goal is to find a word $m \in M$ that would be translationally equivalent to n.

In the first step, a context vector for n is created by going over all its occurrences in a corpus and counting words that appear in its context (e.g. words that are syntactically related to n, such as verbs of which the noun n is a modifier). The value of each feature v in the vector is the conditional probability $p(v|n)$ estimated from the corpus counts. The vector for n is represented as follows:

$$C(n) = \{p(v_1|n), p(v_2|n), \ldots, p(v_i|n)\} \tag{1}$$

where $v_1 \ldots v_i$ are all unique occurrence patterns of n found in the corpus. Context vectors for words in M are prepared in a similar manner.

Second, a bilingual dictionary is used to create a translation matrix D, where rows correspond to unique features extracted from the source language corpus, columns to features of the target language corpus, and cells to the translation relation between the two sets of words provided in the dictionary, with the value in each cell being either binary or weighted. The matrix is used to map the vector for n into the vector space of the target language.

Finally, a (dis)similarity measure such as cosine or Euclidean distance is used to find the word in M whose vector has the greatest similarity to the vector for n. This word is taken to be the translation of n.

3 Dealing with data sparseness

To verify the effect of word frequency on this algorithm, we ran a pilot experiment on six pairs of comparable corpora. We extracted a sample of 1,000 pairs of translation equivalents from each corpus pair and divided it into 10 equal-sized bands according to their frequency (Sect. 6 contains a detailed description of this experimental setup). Figure 1 shows the mean rank of the correct equivalent achieved for each language pair, in each of the frequency bands.

For all the six language pairs, we indeed find large differences in the algorithm's performance in relation to words belonging to different frequency ranges. For example, for the most frequent words in the sample, the correct equivalent typically ranks between 20 and 40, while for the least frequent ones, it can be expected to be found

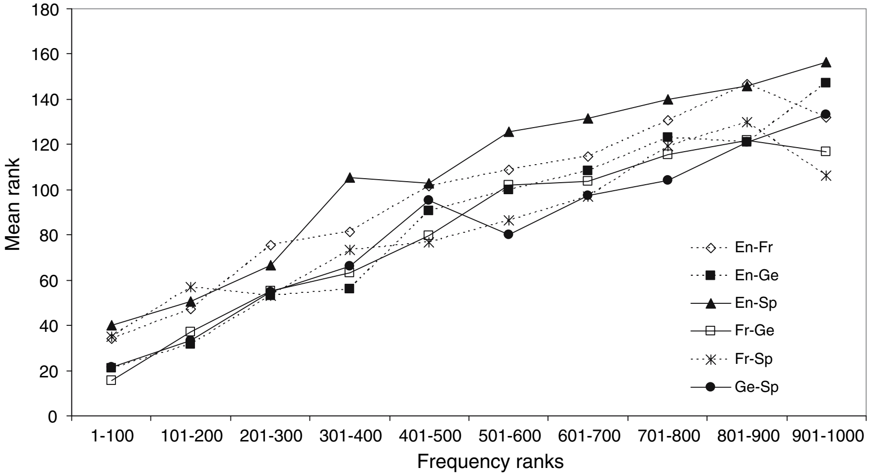

Fig. 1 The performance of the standard algorithm with respect to words with different corpus frequencies. The x-axis shows frequency ranks of source words, with the y-axis showing the mean rank of their correct translations as assigned by the algorithm

only between ranks 100 and 180. The general shape of the performance function also appears to be consistent across language pairs.

This observation calls for a certain way of estimating the probability of occurrence of rare words in contexts where they failed to occur or occurred too few times. Overcoming data sparseness by smoothing corpus frequencies is a familiar problem in NLP, with some techniques, such as Good-Turing and Katz back-off, being the standard approaches. Comparative studies of methods for estimating bigram probabilities (Dagan et al. 1999; Brockmann and Lapata 2003; Keller and Lapata 2003) suggest that class-based smoothing, distance-based averaging, and methods for reconstructing word frequencies from the web are among the best choices. Class-based smoothing (Resnik 1993) relies on a broad-coverage taxonomy of semantic classes. Such resources may not be readily available for any given language, and dependence on them would greatly limit the portability of the overall approach. Web-based estimation of bigram counts (Keller and Lapata 2003) appears impractical for a large-scale smoothing exercise. Therefore in this study, we opt for distance-based averaging techniques.

4 Distance-based averaging

In the distance-based averaging framework (Pereira et al. 1993; Dagan et al. 1999; Lee 1999; Lee and Pereira 1999), the probability of co-occurrence of two words is modelled by analogy with other words that are distributionally similar to the given ones. In this study we employ the nearest neighbour variety of the approach, where the set of distributionally similar words is created ad hoc for each bigram, rather than using fixed sets obtained by clustering. Lee and Pereira (1999) compared the two methods on a pseudo-word disambiguation task, but could not conclusively demonstrate that one method serves as a better model of word co-occurrence than the other. We chose nearest neighbour averaging because of efficiency considerations.

If the probability of a word n appearing with a context word v cannot be estimated because of a zero co-occurrence count, the nearest neighbour method computes the estimate $p^*(v|n)$ as a weighted average of known probabilities $p(v|n')$, where each n' is a close neighbour of n. The weight with which each neighbour influences the average is determined by its similarity to n:

$$w(n, n') = 10^{-\beta \cdot \mathrm{sim}(n,n')} \tag{2}$$

where $\mathrm{sim}(n, n')$ is the distance between the distributional vectors of n and n', and β is a parameter that diminishes the effect of distant neighbours (in all our experiments experimentally set to 0.13). The probability estimate is calculated based on K nearest neighbours as follows:

$$p^*(v|n) = \sum_{n' \in K} \frac{p(v|n') \cdot w(n, n')}{\mathrm{norm}(n)} \tag{3}$$

where $\text{norm}(n) = \sum_{n' \in K} w(n, n')$ is a normalisation factor to ensure probabilistic context.

5 Constructing smoothed context vectors

We wish not only to predict probabilities for unseen co-occurrences, but also to smooth known, but unreliable probabilities for low frequency words. In the latter case, the corpus-estimated probability $p(v|n)$ participates in the calculation of the average p^*, with the weight γ:

$$p^*(v|n) = \gamma \cdot p(v|n) + (1 - \gamma) \cdot \sum_{n' \in K} \frac{p(v|n') \cdot w(n, n')}{\text{norm}(n)} \tag{4}$$

Here, γ controls the amount by which the corpus-estimated probability is smoothed. We propose and evaluate two ways to estimate this variable.

The first one is a heuristic based on the idea that γ should be a function of the frequency of n: the less frequent n is, the more its corpus-estimated probabilities should be smoothed with data from its neighbours. It computes γ as a ratio between the log-transformed counts of n and the most frequent word in the data. This has the effect that the most frequent word will not be smoothed at all, while the least frequent ones will be mainly estimated from the data on their neighbours:

$$\gamma = \frac{\log f(n)}{\log \max_{x \in N} f(x)} \tag{5}$$

The motivation for log-transforming the counts is the same as in many versions of the tf.idf indexing functions: it is meant to downplay differences between high-frequency nouns, which are likely to be less important than the same differences between low-frequency nouns.

The second method estimates γ based on the performance of the algorithm on a held-out set of translation equivalents. First, the held-out word pairs divided into a number of frequency ranges are used to find out the mean rank of the correct translation for each frequency range. Then, function $g(x)$ is interpolated along the points corresponding to the mean ranks in order to predict the rank for some new word, given its frequency. γ is determined from the ratio between the predicted rank of n and the random rank (RR), which is taken to be the lowest possible bound on the mean rank:

$$\gamma = 1 - \frac{g(n)}{\text{RR}} \tag{6}$$

Figure 2 illustrates the two smoothing methods, showing γ values computed for the English–French and German–Spanish pairs using these methods. As one can see, the shape of each function is very similar for the two different language pairs. The smoothing methods, however, appear to prescribe different amounts of smoothing for rare words: the performance-based method is more conservative than the heuristic

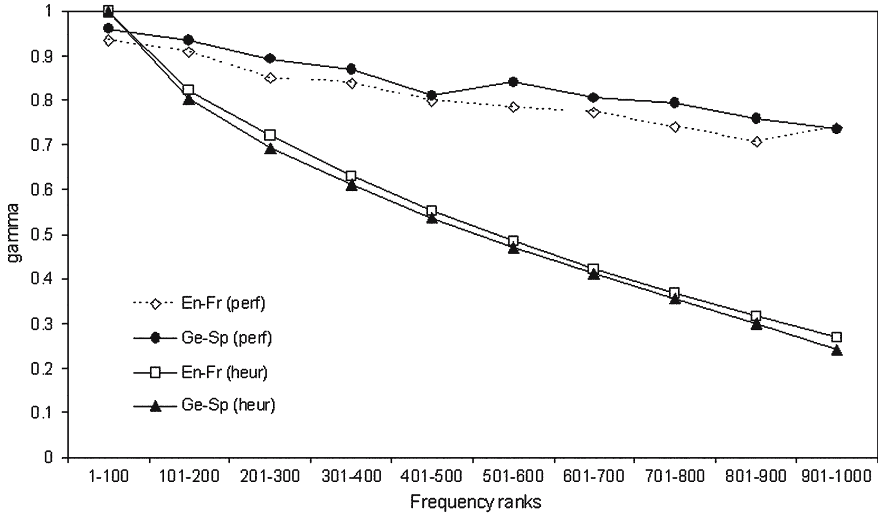

Fig. 2 The values of γ computed using the heuristic and the performance-based method, for the English–French and the German–Spanish datasets

method. For example, for the bottom frequency range the former method will smooth the words by approximately 25%, while the latter will do so by 75% with data on the nearest neighbours of the word.

Another modification of the standard algorithm that we introduce aims to reflect the intuition that infrequent neighbours are likely to decrease the quality of the smoothed vector, because of their unreliable corpus-estimated probabilities. We study the effect of discarding those neighbours that have a lower frequency than the word being smoothed.

6 Experimental setup

6.1 Dictionary

We evaluated the proposed method on translation equivalents for nouns in six language pairs, all pairwise combinations between English, French, German and Spanish. As the gold standard, we used pairs of nouns extracted from synsets and the multilingual synset index in EuroWordNet (EWN).[1]

In a similar manner we extracted pairs of equivalent verbs from EWN for the six language pairs. These were used to construct the translation matrix necessary for mapping context vectors into different languages. During the translation, if a context word had multiple equivalents in the target language according to the dictionary, we followed the previous practice (e.g. Fung and McKeown 1997) and mapped the source context word into all its equivalents, with its original probability equally distributed

[1] http://www.illc.uva.nl/EuroWordNet/

among them. The average number of translations for each source verb was approximately the same across the language pairs, and varied between 3.7 (English–French) and 4.9 (French–Spanish).

6.2 Corpus data

As comparable corpora, we use newsfeed texts from the *Wall Street Journal* (1987–1989) for English, *Le Monde* (1994–1996) for French, *die tageszeitung* (1987–1989 and 1994–1998) for German, and *EFE* (1994–1995) for Spanish. The English and Spanish corpora were processed with the Connexor FDG parser (Tapanainen and Järvinen 1997), French with Xelda[2] from Xerox, and German with Versley's parser (Versley 2005). From the parsed corpora we extracted verb–noun dependencies, where the noun was the head of the direct object 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').

6.3 Evaluation nouns

To be able to compare experimental results across language pairs, we needed to make sure that the evaluation samples contained an equal number of nouns from various frequency ranges and, as far as possible, that the frequency distributions of the nouns were similar in all the samples.

We sampled the evaluation nouns in the following manner. For each language pair, we first created a list of all translation equivalents that were present both in EWN and in both monolingual corpora with a frequency of at least 5.[3] The pairs were then sorted according to the count of the noun which was the less frequent of the two, on the assumption that the less frequent word is the better indicator of the difficulty of finding its equivalent. After that, 1,000 pairs were selected from equidistant locations in the sorted list, and divided into 10 equal-sized frequency bands, such that the first band included the top-100 most frequent pairs, the second one contained 100 pairs with frequency ranks between 101 and 200, and so on. Table 1 presents some descriptive statistics on the evaluation sets produced by this sampling procedure. As one can see, the average frequency within each range is very similar across language pairs, with the exception of the two most frequent ranges.

It should be noted that we did not pre-filter polysemous nouns prior to sampling, and, on average, each noun had between 1.06 (English, in English–Spanish) and 1.15 (French, in French–German) equivalents in the opposite language within the sample.

[2] http://www.xrce.xerox.com/competencies/past-projects/platforms/xelda.html

[3] In the following, by "corpus frequency of a noun" we mean the length of the feature vector, i.e. the number of non-zero distributional features of a noun that we were able to extract from the corpus.

Table 1 Average number of distributional features within each frequency range, for each of the six evaluation sets (the first column describes the ranges in terms of frequency ranks)

Ranges	En–Fr	En–Ge	En–Sp	Fr–Ge	Fr–Sp	Ge–Sp
1–100	561.2	1114.13	1395.54	644.68	616.31	1591.19
101–200	182.22	257.08	275.54	224.69	211.77	370.64
201–300	95.5	130.92	121.57	125.2	110.83	166.8
301–400	54.21	73.93	61.77	76.34	64.7	90.7
401–500	33.23	45.05	36.28	47.96	39.09	52.35
501–600	21.37	27.43	22.77	29.79	24.97	32.26
601–700	14.52	18.03	15.15	19.33	16.52	20.88
701–800	10.18	12.31	10.32	12.98	11.25	13.82
801–900	7.44	8.53	7.3	8.89	7.9	9.15
901–1000	5.46	5.76	5.4	5.86	5.56	5.9

6.4 Assignment algorithm

To measure the similarity between a source word and a target word we use Jensen–Shannon Divergence (Rao 1982), which has often been shown to achieve superior results in comparative studies (e.g. Dagan et al. 1999). Jensen–Shannon Divergence between words n and m is computed as:

$$J(n,m) = \frac{1}{2}[D(n||\mathrm{avg}_{n,m}) + D(m||\mathrm{avg}_{n,m})] \tag{7}$$

where $D(x||y)$ is the Kullback–Leibler divergence between two probability distributions x and y over a set of features V:

$$D(x||y) = \sum_{v \in V} p(v|x) \log \frac{p(v|x)}{p(v|y)} \tag{8}$$

and $\mathrm{avg}_{n,m}$ is the average of the distributions x and y.

Once the similarities between the source word and the target words have been computed, the problem is to select the most likely translation for the source word. A simple and computationally inexpensive solution is the ‘greedy’ algorithm which simply assigns the target word with the greatest similarity as the translation for the source word. However, because in our experiments each source word from a test pair is assigned to a single target word (either can be present in multiple pairs), the greedy algorithm does not guarantee optimal assignment for the entire set of source words and its performance may vary greatly depending on the order in which the words are processed. Instead, we employ the Hungarian (also known as Kuhn–Munkres) algorithm (Kuhn 1955), which efficiently finds such matching of source and target words that maximises the sum of similarity scores in the bipartite graph made up of the two sets of words.

6.5 Evaluation measure

Following the evaluation procedure adopted in (Utsuro et al. 2003), we note the system-assigned rank of the correct translation for each source word and compute a mean rank over all the pairs in sample.

As was mentioned in Sect. 6.3, the evaluation samples we created contain source words that have multiple translations among the target words. There is no widely accepted solution in the literature as to how such cases should be evaluated. For example, Rapp (1999) considered the correct translation for a polysemous source word to be successfully identified if at least one of its translations was discovered. A number of researchers required that all translations for a source word appear within the top-N candidates output by the system (Fung and McKeown 1997; Chiao and Zweigenbaum 2002; Gaussier et al. 2004). In our study, since the usage context of a polysemous source word is not available and it is impossible to decide among its various translations, each particular translation was taken to have the rank of the highest-ranking translation of that word.

Additionally, in Sect. 7.6 we evaluate the methods in terms of the proportion of source words in the evaluation set, for which the correct translation was among the top-N candidates generated by the system.

7 Results

The baseline in our experiments is the standard algorithm (Sect. 2) without any prior smoothing of rare words. Its performance achieved on different language pairs with respect to different frequency bands is shown in Fig. 1. In the following sections, we report differences to the baseline obtained by different configurations of the extended algorithm.

7.1 Nearest neighbour smoothing

We first examined how nearest neighbour smoothing affects the performance of the standard algorithm. The smoothing of the probability in the vector for each noun was carried out according to Eq. 4, with γ set to 0, and the noun being smoothed was included into the nearest neighbour set.

The nearest neighbours are determined from the entire set of nouns extracted from the monolingual corpus, not only from nouns included into the evaluation sample. In the experiment, we varied k, the number of nearest neighbours, between 1 and 1,000. Table 2 shows the differences in the mean rank achieved by the most optimal values of k.

Most of the time, smoothing noun vectors with nearest neighbours actually harmed the performance. While there are a few ranges for some language pairs where a lower mean rank was reached in comparison to the baseline, the average over frequency ranges was worse than that of the baseline (the mean rank increased by 2.9–14%), with the exception of the German–Spanish pair where the decrease in the mean rank was hardly noticeable (0.2%). These results indicate that corpus data on the nearest

Table 2 Changes of the mean rank of the correct translation with respect to the baseline after nearest neighbour smoothing

	En–Fr	En–Ge	En–Sp	Fr–Ge	Fr–Sp	Ge–Sp
1–100	+14.6	+8.7	+13.4	+6.1	+4.9	+6.6
101–200	+10.5	+11.3	+7.3	+1.9	−3.0	+6.2
201–300	+9.2	+2.3	+18.0	−5.7	−5.7	−7.7
301–400	+14.5	+3.8	+8.7	−2.4	+5.5	−12.2
401–500	+16.3	+14.3	+13.4	+2.7	+10.9	−13.7
501–600	+24.9	+7.5	+9.3	−0.6	+4.4	+1.4
601–700	+9.4	+2.4	+6.6	+14.2	+9.5	+12.2
701–800	+25.9	+12.6	+13.2	+17.2	−4.4	+2.4
801–900	+14.8	+10.8	+14.8	+5.1	−3.8	+4.7
901–1000	+19.2	+2.6	+16.4	+6.8	+6.9	−2.0
Average	+15.9	+7.6	+12.1	+4.5	+2.5	−0.2

neighbours cannot completely replace the data on a particular word of interest, which may possibly be explained by insufficient quality of the data available on the neighbours as well as the fact that for many words, the actual corpus-attested probabilities constitute the most valuable evidence for building the model of their meanings.

7.2 Removing less frequent neighbours

Our next experiment consisted of smoothing vectors as in the previous experiment, but excluding those nouns from the set of nearest neighbours that had a corpus frequency below that of the noun being smoothed. Less frequent neighbours of a word are going to 'dilute' its vector, rather than supply missing corpus data. We therefore chose to use all neighbours for smoothing that were more frequent than the word being smoothed. After removing infrequent nearest neighbours, we expanded the set of neighbours accordingly. Table 3 describes the effect of this modification.

The removal of infrequent neighbours resulted in a noticeably better performance in lower frequency ranges: for ranges 301–400 and above the reduction was generally more than 10 points for all language pairs. In the top two ranges, smoothing still often led to higher mean ranks.

Considering the performance on the entire sample (the last row in the table), discarding infrequent neighbours entailed a modest reduction of the mean rank with respect to the baseline for all the language pairs (between 0.7 and 15.1 points, 0.9 and 18% in relative mean rank reduction). According to a two-tailed paired t-test,[4] the reduction was significant in three pairs at $p < 0.001$: French–German ($t = 6.78$), French–Spanish ($t = 4.73$), and German–Spanish ($t = 8.08$), but in the other three pairs the test failed to indicate any significance in the improvement.

[4] df (degrees of freedom) = 1,000 in all the tests reported below.

Table 3 Changes of the mean rank with respect to the baseline, after the removal of infrequent neighbours

	En–Fr	En–Ge	En–Sp	Fr–Ge	Fr–Sp	Ge–Sp
1–100	+2.3	+9.1	+10.5	+4.7	+3.7	+5.5
101–200	+1.5	+8.2	+4.2	−7.3	−2.8	−2.4
201–300	−1.4	−4.7	+4.7	−9.5	−10.6	−11.8
301–400	−11.1	−11.3	−10.0	−22.4	−7.6	−20.2
401–500	−18.7	−13.5	−10.2	−20.2	−7.0	−37.1
501–600	−9.1	−14.2	−9.1	−35.3	−16.5	−15.0
601–700	−0.2	−7.5	−25.9	−22.6	−21.1	−23.6
701–800	−5.1	−12.2	−6.4	−17.9	−34.4	−30.0
801–900	−10.4	−9.8	−4.7	−24.8	−25.7	−32.7
901–1000	−13.6	−26.7	−12.1	−15.6	−4.9	−27.4
Average	−1.0	−0.7	−1.6	−13.5	−8.9	−15.1

Table 4 Changes of the mean rank for the heuristic estimation of γ with respect to the baseline

	En–Fr	En–Ge	En–Sp	Fr–Ge	Fr–Sp	Ge–Sp
1–100	+1.1	+1.8	+11.2	−0.3	−0.1	+3.9
101–200	−4.2	−2.0	−3.1	−10.6	−6.6	−5.5
201–300	−13.4	−17.9	−6.9	−20.1	−15.0	−15.8
301–400	−24.0	−22.6	−23.4	−29.0	−15.9	−30.2
401–500	−36.9	−31.7	−25.0	−35.9	−17.0	−45.0
501–600	−38.7	−41.4	−30.2	−49.1	−29.6	−30.9
601–700	−36.0	−39.5	−39.5	−40.3	−33.3	−33.5
701–800	−39.2	−47.2	−30.1	−37.8	−41.3	−38.2
801–900	−39.4	−34.8	−20.4	−41.8	−31.3	−45.9
901–1000	−32.3	−47.8	−33.1	−32	−15.8	−34.6
Average	−23.3	−26.0	−16.9	−27.7	−18.4	−25.3

In all the following experiments, less frequent neighbours were excluded from the set of nearest neighbours.

7.3 Heuristic estimation of γ

We next examined the performance of the algorithm when the gamma in Eq. 4 was set to be a function of the frequency of the noun being smoothed. Table 4 describes the mean ranks achieved when γ was calculated heuristically according to Eq. 5.

We see that conditioning γ on the frequency of the word being smoothed leads to even better results. With the exception of the most frequent band, all frequency ranges for all language pairs demonstrate lower mean ranks compared with the baseline. In

Table 5 Changes of the mean rank for the performance-based estimation of γ with respect to the baseline

	En–Fr	En–Ge	En–Sp	Fr–Ge	Fr–Sp	Ge–Sp
1–100	+5.1	+2.0	+12.7	−2.8	+0.2	+3.4
101–200	−2.6	+1.4	−2.2	−9.9	−6.5	−5.7
201–300	−11.6	−16.6	−5.0	−20.3	−15.3	−14.5
301–400	−24.2	−22.2	−23.1	−29.8	−14.9	−29.0
401–500	−38.2	−32.7	−24.6	−37.7	−17.3	−45.4
501–600	−37.7	−45.2	−29.7	−55.7	−29.2	−31.0
601–700	−33.0	−39.9	−40.4	−43.2	−34.3	−33.2
701–800	−32.6	−49.3	−25.8	−33.5	−40.1	−37.7
801–900	−31.4	−33.6	−13.6	−36.3	−29.8	−43.2
901–1000	−18.8	−46.7	−30.6	−27.7	−10.4	−35.9
Average	−17.2	−23.5	−14.2	−26.3	−16.7	−24.5

general, it seems that better improvements are achieved on words with lower frequencies: while for the 101–200 range the improvement is under 10 points, for the 201–300 range, it is between 10 and 20 points, and for ranges above 301 it is often over 30 points.

Comparing the mean rank on the entire sample against the one achieved with the baseline, we see improvements for all language pairs; the mean rank is reduced by between 13.7 and 25.5%. The improvement is statistically significant at $p < 0.001$ across the board: English–French ($t = 14.92$), English–German ($t = 17.34$), English–Spanish ($t = 10.09$), French–German ($t = 18.95$), French–Spanish ($t = 13.41$), German–Spanish ($t = 17.32$).

7.4 Performance-based estimation of γ

We then examined an alternative method of computing γ, based on a function estimated from the performance of the method on a held-out set of words (Eq. 6). Table 5 describes the results of this experiment. We observe results similar to those obtained with the heuristic computation of γ: infrequent nouns tend to benefit more from this smoothing technique, and only in the topmost range does the performance exhibit a slight degradation.

Considering the mean rank for the entire sample, it is also substantially lower than the baseline, the mean rank reduction being between 11.8 and 24.5%. The differences are significant at $p < 0.001$ for all language pairs: English–French ($t = 8.89$), English–German ($t = 12.89$), English–Spanish ($t = 7.73$), French–German ($t = 15.34$), French–Spanish ($t = 10.5$), German–Spanish ($t = 16.09$).

Comparing the two ways of computing γ, we find that the heuristic approach delivers consistently lower mean ranks. The difference is significant for all language pairs: at $p < 0.001$ for English–French ($t = 6.83$), English–German ($t = 3.6$), and

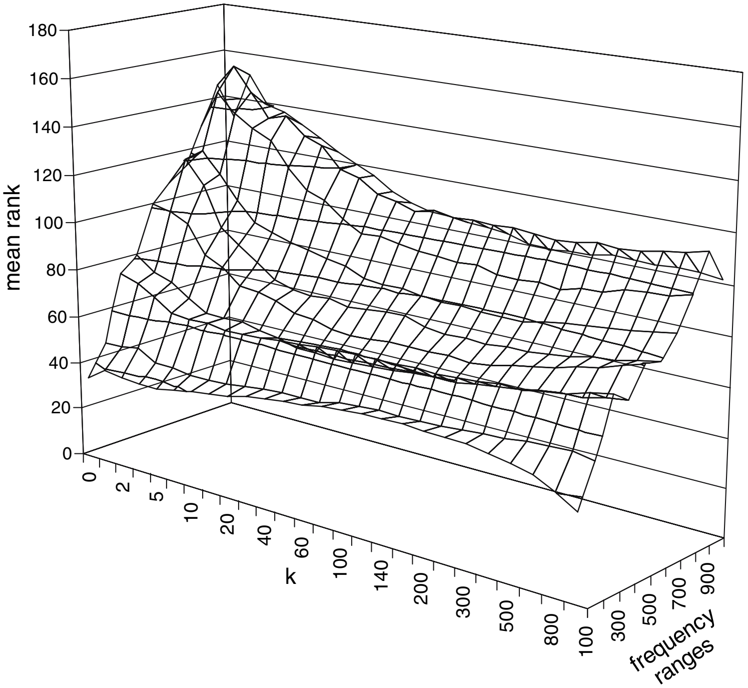

Fig. 3 The relationship between the number of nearest neighbours used for smoothing (k), the mean rank of the correct equivalent and the frequency rank, illustrated on the English–French data when using the heuristic method for estimating γ

French–Spanish ($t = 3.56$); at $p < 0.01$ for English–Spanish ($t = 2.57$) and at $p < 0.025$ for French–German ($t = 2.31$).

7.5 Number of nearest neighbours

Figures 3 and 4 depict the relationships between the mean rank, the frequency range of the noun, and the number of nearest neighbours used for smoothing.[5] For both the heuristic and the performance-based smoothing techniques, we see that the highest mean ranks are achieved for the least frequent words when no smoothing is performed or where very few neighbours are used for smoothing. The mean ranks for low-frequency words decrease steeply with an increase in k and they generally plateau at k between 40 and 100. The more frequent the words, the less they benefit from smoothing: the mean rank for the most frequent words does not appear to change much with alterations of k. Although infrequent words demonstrate the greatest reduction of the mean rank, they still perform worse than the frequent words.

[5] We show the figures only for two language pairs, English–French and German–Spanish, which illustrate the relationships between the three parameters that were found to be similar across all language pairs.

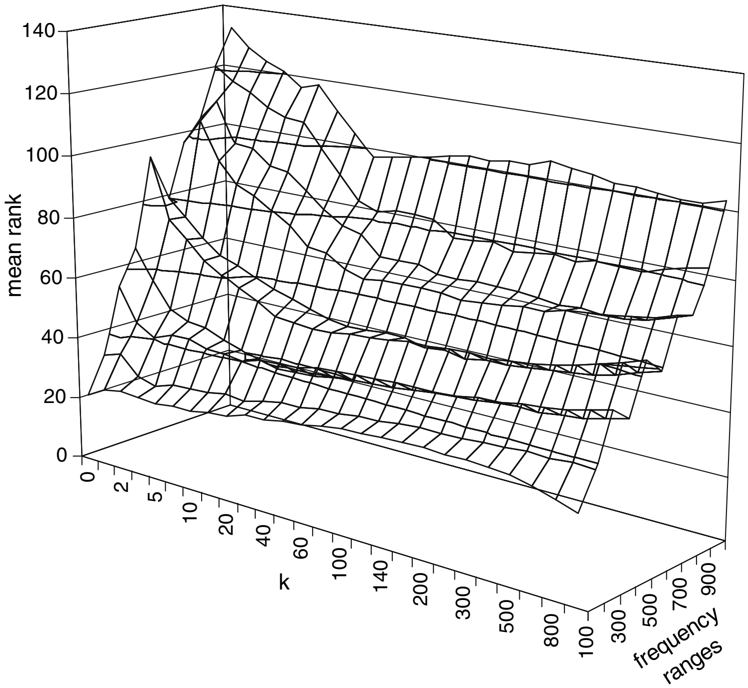

Fig. 4 The relationship between the number of nearest neighbours used for smoothing (k), the mean rank of the correct equivalent and the frequency rank, illustrated on the German–Spanish data when using the performance-based method for estimating γ

7.6 Top-N candidates

Following a number of previous studies (e.g. Fung and McKeown 1997; Chiao and Zweigenbaum 2002; Gaussier et al. 2004), we additionally evaluated the smoothing methods in terms of accuracy, measuring the proportion of source words in the sample, for which the correct translation was among top-N candidates generated by the system. Table 6 shows the accuracy scores achieved by the baseline method and the two smoothing methods (for both, 40 nearest neighbours) for different values of N, for different language pairs.

We see that when considering only the top-most candidate proposed by the system, there is almost no difference between the accuracy scores of the baseline and those of the smoothing methods. On different language pairs, all three methods retrieve the correct equivalent for about 10% of all source words. However, as one increases the candidate list, the positive effect of smoothing begins to be seen: at greater values of N, both smoothing methods gain by up to 15 points over the baseline. These results seem to suggest that cases when the highest-ranked candidate is the correct translation involve the most frequent words, for which smoothing does not help much. Translations for less frequent words, however, are very seldom found at the very top of

Table 6 Accuracy scores for the heuristic and the performance-based smoothing methods, together with the baseline method, considering the top-N candidate translations

	1	5	10	25	50	75	100	150	200
English–French									
Baseline	0.078	0.145	0.209	0.303	0.394	0.480	0.538	0.685	0.841
Heuristic	0.084	0.180	0.233	0.342	0.468	0.578	0.650	0.805	1.00
Performance	0.084	0.164	0.239	0.324	0.453	0.544	0.621	0.776	0.949
English–German									
Baseline	0.090	0.194	0.258	0.359	0.467	0.546	0.611	0.741	0.864
Heuristic	0.110	0.247	0.324	0.437	0.562	0.632	0.715	0.840	1.00
Performance	0.093	0.226	0.301	0.421	0.550	0.628	0.695	0.836	1.00
English–Spanish									
Baseline	0.072	0.163	0.201	0.300	0.380	0.450	0.519	0.658	0.772
Heuristic	0.071	0.183	0.239	0.334	0.431	0.510	0.573	0.712	0.862
Performance	0.070	0.177	0.237	0.324	0.417	0.496	0.579	0.707	0.835
French–German									
Baseline	0.096	0.186	0.257	0.363	0.483	0.551	0.615	0.752	0.906
Heuristic	0.103	0.245	0.329	0.462	0.592	0.696	0.758	0.918	1.00
Performance	0.100	0.248	0.326	0.464	0.599	0.685	0.754	0.913	1.00
French–Spanish									
Baseline	0.088	0.194	0.250	0.359	0.472	0.546	0.613	0.730	0.889
Heuristic	0.099	0.212	0.293	0.406	0.529	0.626	0.699	0.835	1.00
Performance	0.099	0.208	0.296	0.404	0.520	0.616	0.690	0.834	1.00
German–Spanish									
Baseline	0.084	0.188	0.250	0.359	0.468	0.550	0.624	0.765	0.911
Heuristic	0.105	0.239	0.311	0.436	0.591	0.688	0.757	0.904	1.00
Performance	0.108	0.223	0.306	0.435	0.578	0.685	0.753	0.896	1.00

the candidate list, with or without smoothing, but smoothing does help to bring their correct translations higher up the ranking list.

8 Related work

The main focus of many previous studies on the topic has been the problem of determining the similarity between co-occurrence vectors of words belonging to different languages. Rapp (1995) represents one of the first attempts to use co-occurrence analysis for translation pair discovery. The approach first constructs two words-by-contexts matrices for the two languages and then, in order to create a mapping between them, permutes the order of words in the source language matrix until the patterns in the two languages correspond. The method of Fung (1995) establishes similarities between vectors belonging to different languages exploiting the principle of context heterogeneity, the idea that translation pairs can be captured by their similarity in terms of the heterogeneity of their context vectors. Fung and Yee (1998) describe an IR-like approach where a lexicon of seed words is employed to detect the environment of an unknown word, for subsequent coupling to the most similar one in the target language.

Most approaches, however, employ a translation matrix, which makes it possible to associate each component of a vector in one language with components of the vector space of the other language. To create a translation matrix, these approaches use an existing, possibly general bilingual dictionary (e.g. Tanaka and Iwasaki 1996; Fung and McKeown 1997; Chiao and Zweigenbaum 2002; Gaussier et al. 2004). To customise the translation matrix to the domain at hand, Rapp (1999) starts with only a small number of seed translation pairs and augments this translation matrix with more dimensions as the algorithm finds more equivalent terms in the corpus. Déjean et al. (2002) enriched the translation matrix prepared from an available dictionary with a hierarchical multilingual thesaurus. A number of studies (Daille and Morin 2005; Robitaille et al. 2006) augmented this approach with techniques for multi-word recognition and alignment to extract equivalents for multi-word expressions from comparable corpora.

There have been only a few attempts to explicitly counter the problem of polysemy and synonymy of context words that are used for constructing the translation matrix. Tanaka and Iwasaki (1996) disambiguated the senses of the context expression using its local context and a bilingual dictionary. Fung and McKeown (1997) remove from the translation matrix those context words that have multiple translations, which themselves translate into more than one word. In order to diminish the effects of the polysemy and synonymy of context words, Gaussier et al. (2004) incorporate probabilistic latent semantic analysis (PLSA) into the standard approach to retrieve translational equivalents. Morin et al. (2007) demonstrated that the quality of co-occurrence vectors can be substantially improved by ensuring domain and discourse comparability of the corpora from which co-occurrences are obtained.

Since these studies all used different experimental tasks and data, they cannot be directly compared with the results of our study. The experimental settings in Gaussier et al. (2004), however, do appear to be close to ours: they used French–English comparable corpora and 1,250 test pairs of words, measuring F1 score on N top-ranking candidates generated by the system. For 100 top-ranking candidates, they found that PLSA helps to raise the F1 score by 4 points (from 0.24 to 0.28) and their newly proposed method for transforming the co-occurrence matrix, also inspired by LSA, increases the F1 score by 8 points (to 0.32) in comparison with the standard approach. Applying the smoothing techniques we studied in this paper, we obtain an 11 point increase in accuracy (from 0.538 to 0.65) for the top-100 candidates, on the same language pair (French–English) and similar improvements are obtained on other pairs. This comparison indicates that the smoothing techniques lead to similar or even greater improvements than the two LSA-inspired matrix transformation methods.

9 Conclusions

In this paper we addressed the problem of automatic acquisition of pairs of translationally equivalent words from comparable corpora. Our study was carried out in a framework which models equivalence between words of different languages via similarity of their occurrence patterns found in the respective corpora. Our specific goal was to improve the accuracy of retrieval of equivalents for low-frequency words, which are particularly vulnerable to noise introduced during translation of co-occurrence vectors from one language to another.

To address this goal, we develop a method which predicts occurrence patterns for rare words on analogy with words that are distributionally similar to them. The method is an extension of the distance-based averaging technique and aims to predict not only probabilities for unseen word co-occurrences, but also to obtain more reliable probability estimates for rare corpus-attested bigrams.

Our main results are that smoothing co-occurrence vectors with data supplied by nearest neighbours harms performance, unless the degree to which a word is smoothed is conditioned on the frequency of the word being smoothed. We studied two ways to assess the required amount of smoothing for a given word, a heuristic one and one that learns the smoothing function from a held-out set of words. Both techniques yield a significant improvement over the performance of the conventional approach: on average, the system-assigned rank for the correct equivalent of a low-frequency word was reduced by up to 47.8 positions (from 146.9 to 99.1, a 32.5% relative improvement) in comparison with the baseline. This also had a significant positive effect on the overall performance of the method across different frequency ranges: for all language pairs, the average rank of the correct equivalent fell by up to 26.3 positions (from 81.0 to 54.7, a 32.4% relative improvement).

Because infrequent words are typically the most interesting ones from a lexicographic perspective, we believe that these results will open up possibilities for a broader exploitation of comparable corpora by lexicographers. Although the accuracy of the method is still far short of what would be needed if it were to be directly integrated into fully automatic applications, it offers considerable practical advantages if used within a tool to assist a lexicographer to create bilingual lexicon entries. Presented with an improved ranking of candidate translations for rare source words, lexicographers would need to spend significantly less time scanning through the list of target words in order to define the correct translation. While the method was evaluated on single words, the results we obtained are of direct relevance to the practical task of acquisition of translations for multi-word domain terminology: multi-word terms tend to have low corpus frequencies and algorithms for acquisition of translation from comparable corpora have already been extended to multi-word terms (e.g. Déjean et al. 2002).

Furthermore, the results of this study have a bearing on various multilingual tasks which previously have been shown to profit from translation equivalents mined from comparable corpora, including construction of probabilistic translation lexicons for statistical machine translation (Koehn and Knight 2000), acquisition of equivalent text fragments from comparable corpora (Munteanu and Marcu 2006), cross-lingual document retrieval (Utsuro et al. 2003), and cross-lingual text categorisation (Gliozzo and Strapparava 2006).

While the amount of co-occurrence data on a word is an important factor for methods seeking to find its equivalents in a comparable corpus, other factors, such as the degree of comparability of corpora, the size and domain customisation of the lexicon used to create the translation matrix, and the choice of methods for feature weighting and selection, are likely to have a significant effect on the performance of the method in relation to low-frequency words. Future work may focus on these factors and interaction between them.

References

Baroni M, Bernardini S (2004) BootCaT: bootstrapping corpora and terms from the web. In: Proceedings of the 4th international conference on language resources and evaluation. Lisbon, Portugal, pp 1313–1316

Brockmann C, Lapata M (2003) Evaluating and combining approaches to selectional preference acquisition. In: Proceedings of EACL-03: 10th conference of the European chapter of the Association for Computational Linguistics. Budapest, Hungary, pp 27–34

Chiao Y-C, Zweigenbaum P (2002) Looking for candidate translational equivalents in specialized, comparable corpora. In: Coling 2002, proceedings of the 19th international conference on computational linguistics. Taipei, Taiwan, pp 1–5

Curran J (2004) From distributional to semantic similarity. PhD Thesis, University of Edinburgh, Edinburgh, UK

Dagan I, Church K (1997) Termight: coordinating humans and machines in bilingual terminology acquisition. Mach Translat 12(1–2): 89–107

Dagan I, Lee L, Pereira F (1999) Similarity-based models of word cooccurrence probabilities. Mach Learn 34(1–3): 43–69

Daille B, Morin E (2005) French–English terminology extraction from comparable corpora. In: Proceedings of IJCNLP 2005, second international joint conference on natural language processing. Jeju Island, Korea, pp 707–718

Déjean H, Gaussier E, Sadat F (2002) An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: Coling 2002, proceedings of the 19th international conference on computational linguistics. Taipei, Taiwan, pp 1–7

Fletcher W (2004) Making the web more useful as a source for linguistic corpora. In: Conor U, Upton T (eds) Corpus linguistics in North America 2002. Rodopi, pp 191–205

Fung P (1995) Compiling bilingual lexicon entries from a non-parallel English–Chinese corpus. In: Proceedings of the third workshop on very large corpora. Cambridge, MA, pp 173–183

Fung P, McKeown K (1997) Finding terminology translations from non-parallel corpora. In: Proceedings of the fifth workshop on very large corpora. Hong Kong, pp 192–202

Fung P, Yee LY (1998) An IR approach for translating new words from nonparallel, comparable texts. In: COLING-ACL '98: 36th annual meeting of the Association for Computational Linguistics and 17th international conference on computational linguistics. Montreal, Quebec, Canada, pp 414–420

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

Gliozzo A, Strapparava C (2006) Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. Sydney, Australia, pp 553–560

Keller F, Lapata M (2003) Using the web to obtain frequencies for unseen bigrams. Comput Ling 29(3): 459–484

Koehn P, Knight K (2000) Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In: Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence. Austin, TX, pp 711–715

Kuhn HW (1955) The Hungarian Method for the assignment problem. Naval Res Logistic Quart 2: 83–97

Lee L (1999) Measures of distributional similarity. In: 37th annual meeting of the Association for Computational Linguistics. College Park, MD, pp 25–32

Lee L, Pereira F (1999) Distributional similarity models: clustering vs. nearest neighbors. In: 37th annual meeting of the Association for Computational Linguistics. College Park, MD, pp 33–40

Melamed ID (2000) Models of translational equivalence among words. Comput Ling 26(2): 221–249

Morin E, Daille B, Takeuchi K, Kageura K (2007) Bilingual terminology mining – using brain, not brawn comparable corpora. In: ACL 07, proceedings of the 45th annual meeting of the Association of Computational Linguistics. Prague, Czech Republic, pp 664–671

Munteanu DS, Marcu D (2006) Extracting parallel sub-sentential fragments from non-parallel corpora. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics. Sydney, Australia, pp 81–88

Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words. In: 31st annual meeting of the Association for Computational Linguistics. Columbus, OH, pp 183–190

Radhakrishna Rao C (1982) Diversity: its measurement, decomposition, apportionment and analysis. Sankhya: Indian J Stat 44(A): 1–22

Rapp R (1995) Identifying word translation in non-parallel texts. In: 33rd annual meeting of the Association for Computational Linguistics. Cambridge, MA, pp 320–322

Rapp R (1999) Automatic identification of word translations from unrelated English and German corpora. In: 37th annual meeting of the Association for Computational Linguistics. College Park, MD, pp 519–526

Resnik P (1993) Selection and information: a class-based approach to lexical relationships. PhD Thesis, University of Pennsylvania, Philadelphia, PA

Robitaille X, Sasaki Y, Tonoike M, Sato S, Utsuro T (2006) Compiling French–Japanese terminologies from the web. In: EACL-2006, 11th conference European chapter of the Association for Computational Linguistics, proceedings. Trento, Italy, pp 225–232

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, DC, pp 64–71

Tiedemann J (1998) Extraction of translation equivalents from parallel corpora. In: Proceedings of the 11th Nordic conference on computational linguistics (NODALIDA ‘98). Copenhagen, Denmark, pp 120–128

Utsuro T, Horiuchi T, Hamamoto T, Hino K, Nakayama T (2003) Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In: Proceedings of EACL-03: 10th conference of the European chapter of the Association for Computational Linguistics. Budapest, Hungary, pp 355–362

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