Finding Translations for Low-Frequency Words in Comparable Corpora

Viktor Pekar[1], Ruslan Mitkov[1] Dimitar Blagoev[2], and Andrea Mulloni[1]

[1] ILP, University of Wolverhampton, WV1 1SB, United Kingdom
{V.Pekar,R.Mitkov,Andrea2}@wlv.ac.uk
[2] University of Plovdiv, Department of Informatics, 4003 Plovdiv, Bulgaria
gefix@pu.acad.bg

Abstract. The paper proposes a method to improve the extraction of low-frequency translation equivalents from comparable corpora. Prior to performing the mapping between vector spaces of different languages, the method models context vectors of rare words using their distributional similarity to words of the same language to predict unseen co-occurrences as well as to smooth rare, unreliable ones. Our evaluation shows that the proposed method delivers a consistent and significant improvement on the conventional approach to this task.

1 Introduction

The distributional hypothesis, the idea that words with similar meaning tend to occur in similar contexts, can be extended to the bilingual scenario, so that distributional similarity between words of different languages is used to discover translationally equivalent words. The assumption underlies a growing body of research into automated compilation of bilingual lexicons from bilingual comparable corpora [3, 4, 7, 9, among others]. The general procedure implementing this idea begins by collecting co-occurrence data on words of potential interest from monolingual corpora 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 from different languages that have the greatest similarity of their vectors.

A well-known limitation of this approach is that it performs quite unreliably on all but the most frequent words. Even if one ensures that the bilingual dictionary has sufficient coverage so that every occurrence pattern is matched with its equivalent in the other language, a lot of evidence on the words is still lost because of the many-to-many mapping between the two sets of co-occurrence features, resulting from polysemy and synonymy in both languages. As a consequence, only frequent words remain relatively robust against the noise introduced during translation.

In this paper we aim to improve the accuracy of retrieval of translation equivalents for rare words from comparable corpora. We describe an extension of the similarity-based method for estimating word co-occurrence probabilities [2] to the problem of modeling context features of rare words prior to translating their vectors into the vector space of a different language.

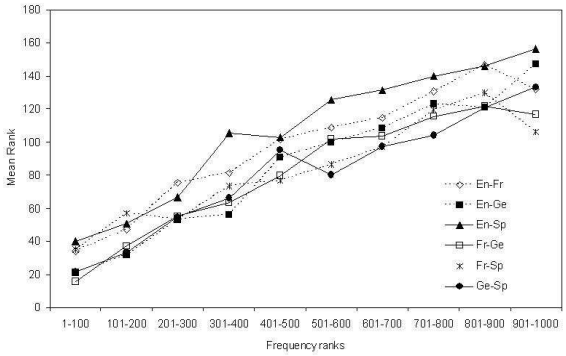

Fig. 1. The performance of the standard algorithm with respect to words with different corpus frequencies. The x-axis shows frequency ranks of source words, the y-axis – the mean rank of their correct translations as assigned by the algorithm.

2 Dealing with Data Sparseness

To verify the effect of word frequency on the standard algorithm, we run a pilot experiment on six pairs of comparable corpora in different languages. We extracted a sample of 1000 pairs of translation equivalents from each pair of corpora and divided them into 10 equal-size bands according to their frequency (Section 5 contains a detailed description of this experimental setup). Figure 1 depicts the mean rank achieved by the algorithm for each language pair, within each frequency band.

For all the 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 (frequency ranks 1 to 100), the correct equivalent typically appears at ranks between 20 and 40, while for the least frequent ones, one can expect to find it only between ranks 100 and 160. The shapes of the performance functions are also very much the same across the language pairs.

This observation calls for a certain way to estimate 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's back-off, being the standard approaches. Comparative studies of methods for smoothing bigrams [1, 2, 5] suggest that class-based smoothing, and distance-based averaging, and methods for reconstructing word frequencies from the web are among the best choices. Class-based smoothing [8] relies on a broad-coverage taxonomy of semantic classes. Such resources may not be readily available for a given language, and dependence on them would greatly limit the portability of the overall approach. Web-based estimation of bigram counts [5] appears impractical for a large-scale smoothing exercise. Therefore in this study, we opt for distance-based averaging techniques.

3 Distance-Based Averaging

In the distance-based averaging framework [2], the probability of co-occurrence of two words is modeled by analogy with other words that are distributionally similar to the given ones. In this study we employ the nearest neighbor variety of the approach, where the set of neighbors, i.e. distributionally similar words, is created ad hoc for each bigram, rather than using fixed sets of similar words obtained by clustering.

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 neighbor method computes the estimate $p^*(v|n)$ as a weighted average of known probabilities $p(v'|n)$, where each n' is a close neighbor of n. The weight with which each neighbor influences the average is determined by its similarity to n:

$$w(n, n') = 10^{-\beta \cdot SimScore(n,n')} \tag{1}$$

where β is a parameter that diminishes the effect of distant neighbors (in our experiments fine-tuned to .13). The probability estimate is calculated based on K nearest neighbors as follows (K is set experimentally):

$$p^*(v|n) = \sum_{n' \in K} \frac{p(v|n') \cdot w(n, n')}{norm(n)} \tag{2}$$

where $norm(n) = \sum_{n' \in K} w(n,n')$ is a normalization factor used to ensure that the conditional probabilities for n sum to 1.

4 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:

$$p^*(v|n) = \gamma \cdot p(v|n) + (1 - \gamma) \cdot \sum_{n' \in K} \frac{p(v|n') \cdot w(n, n')}{norm(n)} \tag{3}$$

Here, γ controls the amount by which the corpus-estimated probability is smoothed. We believe that γ should be a function of the frequency of n: the less frequent is n, the more its corpus-estimated probabilities should be smoothed with data from its neighbors. We propose and evaluate two ways to estimate this function.

The first one is a heuristic that 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 neighbors:

$$\gamma = \frac{\log f(n)}{\log \max_{x \in N} f(x)} \tag{4}$$

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 mean rank for a certain novel word, given its frequency. γ is then determined as the ratio between the predicted rank of n and that of the most frequent word in the data:

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

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

5 Experimental Setup

Dictionary. We evaluate the proposed method on translationally equivalent nouns in six language pairs, all pairwise combinations between English, French, German and Spanish. As the gold standard, we use pairs of nouns extracted from synsets and the multilingual synset index in EuroWordNet (EWN)[3]. 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. If, during the translation, a context word that had multiple equivalents in the target language according to the dictionary, it was mapped into all its equivalents, with its original probability equally distributed among them.

Corpus Data. As comparable corpora, we use newspaper texts from the *Wall Street Journal* (1987-89) for English, *Le Monde* (1994-96) for French, *die tageszeitung* (1987-89 and 1994-98) for German, and *EFE* (1994-95) for Spanish. The English and Spanish corpora were processed with the Connexor FDG parser, French with Xerox Xelda, and German with Versley's parser. From the parsed corpora we extracted verb–direct object dependencies, where the noun is the head of the modifier phrase.

Evaluation Nouns. To ensure that the evaluation data for all the language pairs contain an equal number of nouns from similar frequency ranges, we used the following sampling procedure. For each language pair, we first created a list of all translation equivalents that are present both in EWN and in both monolingual corpora with frequency above 5. The pairs were then sorted according to the count of the noun which was 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, 1000 pairs were selected from equidistant locations in this ordered list, and divided into 10 equal-size frequency bands, such that the first band included the top 100 most frequent pairs, the second one – pairs with frequency ranks between 101 and 200, and so on.

Assignment Algorithm. Once the similarities between the source word and the target words have been computed (we use the Jensen-Shannon Divergence to measure

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

(dis)similarity of context vectors, for a discussion of the function, cf. [2]), the problem is to select the most likely translation for the source word. To determine optimal assignment for the entire set of source words, we employ the Hungarian (also known as Kuhn-Munkres) algorithm [6], which efficiently finds such matching of source and target words that maximizes the sum of similarity scores in the bipartite graph made up of the two sets of words.

Evaluation Measure. Following the evaluation procedure adopted in [10], 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. A mean rank appears an intuitive evaluation measure, since it describes how soon a correct translation for a source word can be found by a lexicographer who revises translations proposed by the system.

Baseline. The baseline in our experiments is the standard algorithm without any smoothing of the data. Its performance achieved on different language pairs with respect to different frequency bands is shown in Figure 1. In the following sections, we report differences to the baseline attained by configurations of the extended algorithm.

6 Results

6.1 Nearest Neighbors Smoothing

We first examined how nearest neighbors smoothing affects the performance of the standard algorithm. The smoothing of the probability in the vector for each noun was carried out according to Equation 3, with γ set to 0 and the noun being smoothed was included into the nearest neighbor set. The nearest neighbors 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 neighbors, between 1 and 1000. Table 2 shows the differences in the mean rank achieved by the most optimal values of k in comparison to the baseline algorithm.

Most of the time smoothing noun vectors with nearest neighbors resulted in a higher mean rank in comparison to the baseline, i.e., the performance degraded. While there are a few ranges for some language pairs where a lower mean rank was reached, the average over frequency ranges was higher than that of the baseline, with the exception of the German-Spanish pair where it was only slightly lower.

6.2 Ignoring Less Frequent Neighbors

Our next experiment consisted in smoothing vectors as in the previous experiment, but excluding those nouns from the set of nearest neighbors that had corpus frequency below that of the noun being smoothed. After infrequent nearest neighbors have been removed, the set of neighbors has been expanded accordingly. Table 3 shows the results.

The removal of infrequent neighbors 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 neighbors entailed a modest reduction of the mean rank wrt

Table 1. Changes of the mean rank of the correct translation with respect to the baseline after nearest-neighbor 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

the baseline for all the language pairs (between 0.7 and 15.1 points, .9% and 18%). According to a two-tailed paired t-test[4], the reduction was significant in three pairs at $p < .001$ (French-German, French-Spanish, German-Spanish), but in the other three pairs the test failed to show any significance of the improvement. In all the following experiments, less frequent neighbors were excluded from the set of nearest neighbors.

6.3 Heuristical Estimation of γ

We next examined the performance of the algorithm when γ in Equation 3 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 Equation 4.

We see that making γ dependent 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 general, it seems that better improvement 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 improvement for all the language pairs. The improvement is statistically significant at $p < .001$ across for all pairs.

6.4 Performance-Based Estimation of γ

We then examined the alternative method to compute γ, based on a function estimated from the performance of the method on a held-out set of words (Equation 5). These results are similar to those obtained with the heuristical computation of γ: infrequent

[4] df = 1000 in all the tests reported below.

Table 2. Changes of the mean rank wrt the baseline, after the removal of infrequent neighbors.

	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 3. Changes of the mean rank for the heuristical estimation of γ wrt 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

nouns tend to benefit more from this smoothing technique, and only in the topmost range the performance shows a slight degradation. Considering the mean rank for the entire sample, it is also significantly lower than the baseline, at $p < .001$ for all the language pairs. Comparing the two ways to compute γ, we find that the heuristical approach delivers consistently lower mean ranks, the difference being significant for all the language pairs at $p < .025$.

7 Conclusion

Our study was carried out in the framework which models translational equivalence between words via similarity of their occurrence patterns found in monolingual corpora. In order to improve the retrieval of equivalents for low-frequency words, which are particularly vulnerable to noise introduced during the cross-linguistic mapping

Table 4. Changes of the mean rank for the performance-based estimation of γ wrt 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

of context vectors, we proposed a method which models occurrence patterns of words on analogy with words that are distributionally similar to them. The method extends the distance-based averaging technique to predict not only unseen word co-occurrences, but also to obtain more reliable probability estimates for rare bigrams. Our experimental evaluation has showed that the method yields a significant improvement on the conventional approach in relation to low frequency words and has a considerable positive effect on the overall retrieval accuracy.

References

1. Carsten Brockmann and Mirella Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proc. EACL*, pages 27–34, 2003.
2. Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.
3. Hervé Déjean, Éric Gaussier, and Fatiha Sadat. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proc. COLING*, 2002.
4. Pascale Fung and Kathleen McKeown. Finding terminology translations from non-parallel corpora. In *The 5th Annual Workshop on Very Large Corpora*, pages 192–202, 1997.
5. Frank Keller and Mirella Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484, 2003.
6. Harold W. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.
7. Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora. In *Proc. ACL*, 1999.
8. Phillip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993.
9. Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro. Compiling French-Japanese terminologies from the web. In *Proc. EACL*, 2006.
10. Takehito Utsuro, Takashi Horiuchi, Takeshi Hamamoto, Kohei Hino, and Takeaki Nakayama. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *Proc. EACL*, pages 355–362, 2003.