Using Parallel Corpora to enrich Multilingual Lexical Resources

Appeared in Third International Conference on Language Resources and Evaluation (LREC 3),
Las Palmas, May 2002, pages 240-245

Dominic Widdows, Beate Dorow, Chiu-Ki Chan
[*]Center for the Study of Language and Information
Stanford University, California
{dwiddows,beate,ckchan}@csli.stanford.edu
http://infomap.stanford.edu

March 21, 2003

Abstract

This paper describes the use of a bilingual vector model for the automatic discovery of German translations of English terms. The model is built by analysing co-occurence patterns in a parallel corpus of English and German medical abstracts, and can also be used for Cross-Lingual Information Retrieval. The model generates candidate German translations of English words using the cosine similarity measure between terms in the bilingual vector space. The correct translations could be added to UMLS, the multilingual dictionary in question. The accuracy of the translations is evaluated by measuring how many of the existing UMLS translations are correctly predicted by the vector translations. The model also detects synonymy, particularly acronyms. An online public demonstration of the model is available.

1 Introduction

Hand-built lexical resources are expensive to construct and maintain, vary in coverage, and often lack new and domain-specific terms. Hence automatic methods of lexical acquisition are important, especially for multilingual dictionaries and ontologies, where a particular opportunity exists to fill gaps within resources for one language with information from resources for another language.

Various approaches to this problem exist, many of which are related to other aspects of work on parallel corpora (Véronis 2000). A parallel corpus is a collection of documents translated into more than one language, and parallel corpora are very rich sources of information about the translation process. Moore (2001) describes a statistical approach to learning translational relationships, and summarises the general method of choosing the 'highest-scoring partner' as the potential translation, using some suitable similarity score. Fung (2000) adapts correlation scores to extract terms from non-parallel corpora.

The problem has often been approached using bilingual text-alignment methods, such as the bitext mappings of Melamed (1996), because words which are directly aligned to one another are very likely to have the same meaning. Gaussier (1998) describes the collection of possible alignments between parallel sentences in terms of network-flows, and demonstrates that this general approach can be used to extract lexical information even from very small corpora. Many alignment method rely on (or can be improved using) a core of known translations, sometimes referred to as a *seed-lexicon*. It follows that using alignment methods for lexical extraction can be a circular approach, as Melamed (1996) points out.

The approach taken in this paper draws upon methods used for Cross-Lingual Information Retrieval (CLIR) rather than text alignment. Because of this, our method does not rely on the corpus being any more closely aligned than at the document level. (For this reason, our method can also be used to generate the seed-lexicons necessary to support more detailed processing.) The alignment between pairs of translated documents is used to map words from each language into a single bilingual vector space, a technique first developed for CLIR (Dumais, Landauer, and Littman 1996). Using the standard cosine sim-

[*]This research was supported in part by the Research Collaboration between the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University, and by EC/NSF grant IST-1999-11438 for the MUCHMORE project.

ilarity measure in this bilingual vector space, it is possible to pick out cross-lingual pairs which are similar in usage, and if the words are close enough to one another under this metric we can be confident that they refer to the same concept. This technique can therefore be used to fill in some of the gaps in multilingual lexical resources. A similar process was used by Brown, Carbonall, and Yang (2000) to develop term-substitution for CLIR.

We evaluate this method at the task of adding missing German translations of English words in the Unified Medical Language System (UMLS), a publicly available medical language resource. [1] Though the coverage of UMLS for German is better than for any language other than English, there are many gaps in the lexicon where there is no corresponding German equivalent of an English term. We are seeking to improve this situation to facilitate multilingual information management, as part of the MUCHMORE project. [2] Our results show that accuracy of translation for high-scoring pairs can exceed 90%, and that the method also finds synonyms of terms which are already contained in UMLS.

We also discuss the possibility of using many-to-one mappings of English words to German compounds, and the uses of the bilingual vector space to model this kind of semantic composition.

There is an on-line public demonstration of our system which can be used for term-translation, query expansion and document retrieval. [3]

2 Building a Bilingual Vector Model from a Parallel Corpus

In this section we describe how English and German terms were encoded as points in a single abstract vector space. [4] This space could be used to represent semantic similarity, because terms with similar or related meanings are usually close to one another in the vector space.

First we review the standard processes whereby such a vector space can be built from monolingual documents. The first examples of such spaces were pioneered for Information Retrieval (Salton and McGill 1983; Baeza-Yates and Ribiero-Neto 1999). Counting the number of times each word occurs in each document gives a *term-document matrix*, where the i, j^{th} matrix entry records the number of times the word w_i occurs in the document d_j. The rows of this matrix can then be thought of as *word-vectors*. *Document vectors* are then generated by computing a (weighted) sum of the word-vectors of the words appearing in a given document. The dimension of this vector space (the number of co-ordinates given to each word) is therefore equal to the number of documents in the collection. Typically, such *term-document matrices* are extremely sparse. The information can be concentrated in a smaller number of dimensions using singular-value decomposition, projecting each word onto the n-dimensional subspace which gives the best least-squares approximation to the original data. This represents each word using the n most significant 'latent variables', and for this reason this process is called *latent semantic analysis* (Deerwester, Dumais, Furnas, Landauer, and Harshman 1990).

Such techniques are used in information retrieval to measure the similarity between words (or more general query statements) and documents, using a similarity measure such as the cosine of the angle between two vectors (Baeza-Yates and Ribiero-Neto 1999, p 27). A less-well known but natural corrolary is that this technique can be used to measure the similarity between pairs of terms. Term-term similarities of this sort can be used for the process of *automatic thesaurus generation* (Baeza-Yates and Ribiero-Neto 1999, Ch 5). The underlying idea of this paper is that with a bilingual vector model, such term-term similarities can be used to detect which pairs of words are translations of one another.

A variant of the traditional term-document matrix was developed by Schütze (1997) specifically for the purpose of measuring semantic similarity between words. Instead of using the documents as column labels for the matrix, semantically significant *content-bearing words* are used, and other words in the vocabulary are given a score each time they occur within a context window of (say) 15 words of one of these content-bearing words. Thus the vector of the word *football* is determined by the fact that it frequently appears near the words *sport* and *play*, etc. This method has been found to be well-suited for semantic tasks such as word-sense clustering and disambiguation.

Drawing upon these techniques, our bilingual vector model was built as follows. A corpus consisting of 9640 German abstracts from medical documents and their English translations (*ca* 1.5 million words) was obtained from the Springer Link information service. [5] Each German/English document pair was treated

[1] http://umls.nlm.nih.gov

[2] http://muchmore.dfki.de

[3] http://infomap.stanford.edu/bilingual

[4] There are several good introductions to vectors and vector spaces, including that of Vallejo (1993).

[5] http://link.springer.de/

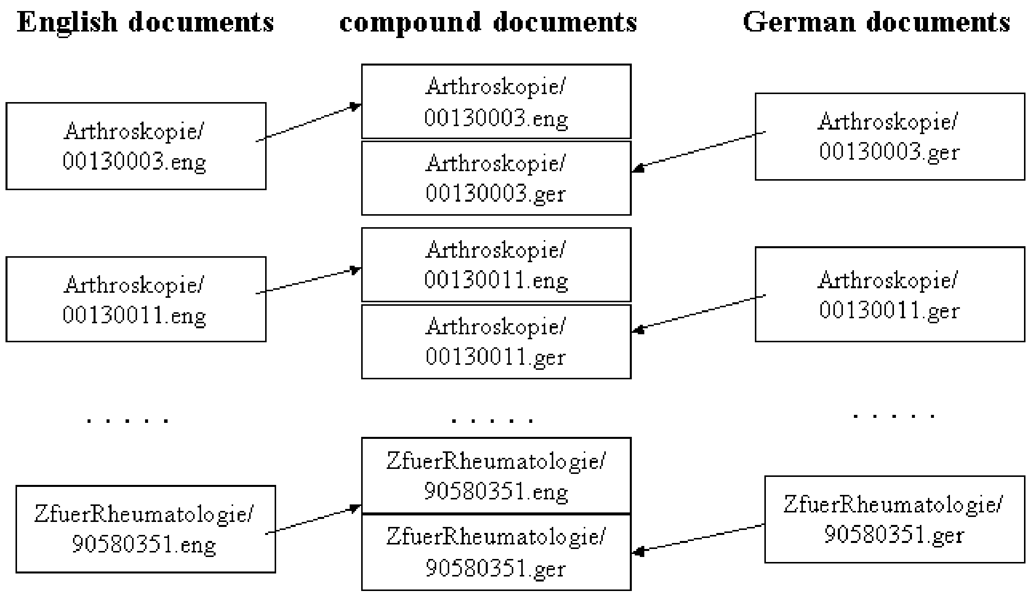
English documents
compound documents
German documents
Arthroskopie/ 00130003.eng
Arthroskopie/ 00130003.eng
Arthroskopie/ 00130003.ger
Arthroskopie/ 00130003.ger
Arthroskopie/ 00130011.eng
Arthroskopie/ 00130011.eng
Arthroskopie/ 00130011.ger
Arthroskopie/ 00130011.ger
. . . . .
. . . . .
. . . . .
ZfuerRheumatologie/ 90580351.eng
ZfuerRheumatologie/ 90580351.eng
ZfuerRheumatologie/ 90580351.ger
ZfuerRheumatologie/ 90580351.ger

Figure 1: Treating pairs of abstracts as a single document for recording co-occurence

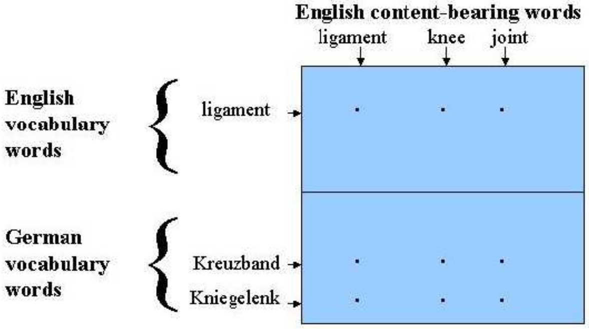

Figure 2: Each vocabulary word in each language is given co-ordinates based on 1,000 English 'Content-Bearing Words'

as a single 'compound document' for the purpose of recording term-term co-occurence (Figure 1). After stopwords were removed (Baeza-Yates and Ribiero-Neto 1999, p 167), the 1000 most frequent English words were selected as content-bearing words. (English words were chosen because semantically significant units are more often single words in English but parts of compounds in German, and because other parallel corpora are more likely to have English as one of the languages.)

English and German words were regarded as co-occurring with a particular content-bearing word if they occurred in the same document as the content-bearing word, or the translation of this document. This avoided the need for in-depth alignment of the corpus, a simplification which was made possible by the brevity of most of the documents (*ca* 150 words on average). (A bilingual corpus of many thousand short documents is naturally much better aligned than a corpus of fewer much longer documents.) This indexing process is illustrated in Figure 2.

In this way, the 10,000 most frequent words in each language were mapped into a single 1,000-dimensional vector space. Singular value decomposition (LSI) was used to reduce the number of dimensions to 100. Semantic similarity between English and German terms could then be computed using cosine similarity in this 100-dimensional bilingual vector space. This method was used to measure term-term similarity throughout the experiments described in this paper.

3 Enriching the German UMLS using the Vector Model

In this section we describe experiments that demonstrate the usefulness of the bilingual model for the task of creating or enriching bilingual lexical resources. Our goal is to add German terms to the UMLS database, [6] to improve bilingual access to medical information in English and German. Large amounts of new knowledge and terminology are always being added to the medical domain, and UMLS is already much richer for English than for any other language, making efficient techniques for automatically extending such a database particularly important.

The vector model was used to suggest German translations for English words. This was done by computing the nearest German neighbour of each English term in the vector model. Recalling that each word was represented as a vector, we could find the nearest neighbour of any word by comparing the cosine distances between it and all other vectors, and retrieve those with the highest similarity score. This process is illustrated in Table 1, which shows the first few English and German neighbours of the English word *bone* in the bilingual vector space. Note that the highest scoring German neighbour is the word *knochen*, which is a correct translation of the English *bone*. The relatively high similarity score of over 0.82 indicates a high confidence that these two words really do share the same meaning.

To evaluate the accuracy of our translation method, we compared the results with those translation which are already in UMLS. Among the English terms in the bilingual vector model, 9213 are recognised UMLS concepts. UMLS represents each concept with a unique concept identifier (CUI). To extract the UMLS translation of the English terms, we did a lookup in UMLS for the German terms that shares a CUI with the English term. Only 6823 of these English terms had German translations in UMLS. We compare these known UMLS German translations with the translation predicted by our bilingual vector model to calculate the accuracy of our translation. When the two translations agreed, we marked the vector translation as correct. [7] When the vector translation disagreed with the UMLS translation, we assumed that the vector trans-

[6] http://www.nlm.nih.gov/research/umls/

[7] There is a certain amount of synonymy in UMLS, so some words are given several possible translations. In these cases we considered a translation to be correct if it obtained one of the possible synonyms.

English Neighbours	Similarity	German Neighbours	Similarity
bone	1.000000	knochen	0.823083
cancellous	0.700623	knochens	0.708817
osteoinductive	0.671816	knochenneubildung	0.699606
demineralized	0.648947	spongiosa	0.635176
trabeculae	0.639279	knochenresorption	0.595616
formation	0.595301	allogenen	0.594648
periosteum	0.562293	knöcherne	0.590172
osteoporotic	0.561281	knochenheilung	0.578918
autoclaved	0.559798	bone	0.569451
augmentation	0.543297	knochentransplantate	0.565430
substitute	0.532057	knochentransplantaten	0.564502
hydroxyapatite	0.528326	trabekulären	0.555980
ridge	0.526757	knochentransplantation	0.548806
osteoclast	0.523437	aufgefüllt	0.545810
marrow	0.523071	hydroxylapatitkeramik	0.542906
resorption	0.516087	knochenregeneration	0.531353

Table 1: English and German Neighbours of the English word *bone*

lation was wrong. This gave us a conservative estimate of our accuracy.

The results of this evaluation experiment are displayed in Figure 3. There was a strong correlation between the similarity score between an English vector and its nearest German neighbours, given by the vector model, and the likelihood that this translation was correct according to UMLS. However, there were still more highly-scoring translations that were marked incorrect than we had hoped for. (In pictorial terms, we would like the 'Wrong Translations' curve in Figure 3 to continue its downward gradient so that when we reach a similarity score of 1 (an exact match) the probability of error is zero.)

Having used those UMLS concepts with both English and German versions as a benchmark, we were in a position to estimate the accuracy with which the bilingual vector space translated the 2350 English terms with no German counterpart in UMLS. Over 160 were translated with a confidence of 75% or above. These results were independently checked by a human annotator, who confirmed that in fact, over 88% of our translations were correct and could be added to our version of UMLS. Precision was thus much higher than expected from the estimated accuracy derived from known UMLS translations. This led us to perform a more detailed error analysis on those high-scoring translation candidates that had been marked as 'wrong' by our evaluation method.

We had assumed that in the cases where the vector translation and UMLS translation disagreed, the differences were the result of statistical error in the corpus-derived vector model. In many cases, how-

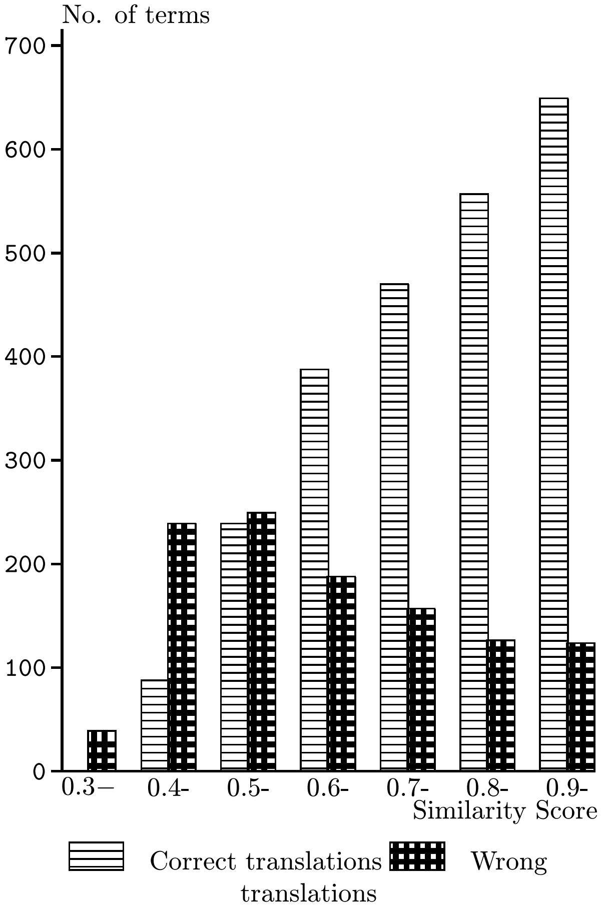

Figure 3: Number of correct / incorrect translations according to UMLS for different ranges of similarity score

ever, the vector translation was a different but valid term for the same concept — in the high-confidence region, around 91% of the vector translations were accurate, giving another 178 correct German translations. This synonymy frequently resulted from the use of acronyms: over 130 acronyms were recognised as correct translations by the bilingual vector model but not by UMLS. It should be possible to use similar techniques to add synonyms to lexical resources using only monolingual corpora.

4 Further Work

This paper clearly demonstrates what can be achieved using bag-of-words statistical methods, and the results are encouraging. Clearly, though, there are many more linguistically motivated techniques that should be integrated into any real-world application for bilingual lexicon extraction.

Preliminary experiments using more detailed sentence-alignment have not shown a definitive improvement in on model. This supports the conjecture that the difference in size between a sentence and a short document is not especially significant for the statistical methods used.

The most obvious first step is morphology for stemming and decompounding, especially for the German abstracts. (A brief glance at Table 1 should convince the reader of this.) Much of this work has been done for the bilingual corpus used in this paper (Špela Vintar, Buitelaar, Ripplinger, Sacaleanu, Raileanu, and Prescher 2002), and further integration with these efforts is underway as part of the MuchMore project. We certainly expect to improve results by combining methods.

However, that good results can be achieved using a comparatively simple baseline method gives the bilingual vector model independent interest. As well as using decompounding to build the model, we could also use the model as it stands to try and model the process of *semantic* composition and compare this with more well-understood results obtained through morphological composition. This follows the the principle that translational relationships between words often involve mappings that are many-to-one rather than just one-to-one (Moore 2001).

Consider the example results in Table 2, which shows the nearest neighbours to the vector given by summing the English *lung* and *transplant*. Using standard addition of vectors as a model for composition clearly works well in this case. However, many times adding vectors is a poor model for semantic composition. Firstly, this operation is commutative (hence the traditional complaint that many IR systems do not distinguish between a *blind Venetian* and a *Venetian blind*). Secondly, vector addition combines co-ordinates which may arise in many contexts. (For example, the query vector *potato* + *chip* can still return documents about silicon chips using just vector addition.)

We therefore propose to investigate different mathematical operations for semantic composition, using known German compounds to evaluate the accuracy of composition of English words.

5 Conclusion

Our experiments show how a bilingual corpus can be used for automatically extending lexical resources. Our results give some indication of the high accuracy that can be attained by simple bag-of-words methods. Our method is sensitive to language use which is not always represented in lexical resources, such as the introduction of new terms for familiar concepts. As document collections grow and new terminology, especially domain-specific terminology, is added faster and faster, we anticipate that automatic methods such as these will assume increasing importance.

Demonstration

An online demonstration of vector term-translation can be accessed publicly on http://infomap.stanford.edu/bilingual.

It is fully interactive - the user enters one or more query terms in English and/or German and receives related terms in both languages ordered by similarity score. The "most similar term" in the other language is very often the correct translation of a query word.

Acknowledgments

The first author would like to thank Chris Manning and Bob Moore for useful guidance.

This research was supported in part by the Research Collaboration between the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University.

This research has in part been supported by EC/NSF grant IST-1999-11438 for the MUCHMORE project.

English Neighbours	Similarity	German Neighbours	Similarity
transplant	0.794835	lungentransplantation	0.730531
lung	0.794835	lunge	0.649870
transplantation	0.608974	transplantiert	0.520466
bronchiolitis	0.586815	lungenemphysem	0.516735
recipients	0.586446	transplantation	0.510346
obliterans	0.545399	lungen	0.510102
actuarial	0.538949	transplantatvaskulopathie	0.496594
rejection	0.521160	lungenfunktion	0.483037
lungs	0.513533	plötzlich	0.466997
pneumonectomy	0.497044	htx	0.458539
orthotopic	0.492516	alveolen	0.442872
allograft	0.477084	pneumonektomie	0.438988
vasculopathy	0.475079	organtransplantation	0.438770
donor	0.472244	ards	0.435389
transplantations	0.460281	lungenerkrankungen	0.434309
bronchial	0.445858	pulmonalen	0.433340

Table 2: English and German Neighbours of the English query *lung* + *transplant*

References

Baeza-Yates, R. and B. Ribiero-Neto (1999). *Modern Information Retrieval*. Addison Wesley / ACM press.

Brown, R. D., J. G. Carbonall, and Y. Yang (2000). Automatic dictionary extraction for cross-language information retrieval. In J. Véronis (Ed.), *Parallel Text Processing*, pp. 275–298. Kluwer.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41(6)*, 391–407.

Dumais, S., T. Landauer, and M. Littman (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR '96 – Workshop on Cross-Linguistic Information Retrieval*, pp. 16–23.

Fung, P. (2000). A statistical view on bilingual lexicon extraction. In J. Véronis (Ed.), *Parallel Text Processing*, pp. 219–236. Kluwer.

Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the joint 17th International Conference on Computational Linguistics and 26th Annual Meeting of the Association for Computational Linguistics*, pp. 444–450.

Melamed, I. D. (1996). Automatic construction of clean broad-coverage translation lexicons. In *2nd Conference of the Association for Machine Translation in the Americas*, Montreal, Canada.

Moore, R. C. (2001, July). Towards a simple and accurate statistical approach to learning translational relationships among words. In *Proceedings of the workshop on data-driven machine translation*, Toulouse. 39th annual meeting of the Associate for Computational Linguistics.

Salton, G. and M. McGill (1983). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill.

Schütze, H. (1997). *Ambiguity resolution in language learning*. Stanford CA: CSLI Publications.

Vallejo, R. J. (1993). *Linear algebra: an introduction to abstract mathematics*. Undergraduate texts in mathematics. Springer-Verlag.

Véronis, J. (2000). From the rosetta stone to the information society: A survey of parallel text processing. In J. Véronis (Ed.), *Parallel Text Processing*, pp. 1–25. Kluwer.

Špela Vintar, P. Buitelaar, B. Ripplinger, B. Sacaleanu, D. Raileanu, and D. Prescher (2002, May). An efficient and flexible format for linguistic and semantic annotation. In *Third International Language Resources and Evaluation Conference*, Las Palmas, Spain. European Language Resources Association.