

Creating Bilingual Lexica Using Reference Wordlists for Alignment of Monolingual Semantic Vector Spaces

Jon Holmlund and Magnus Sahlgren and Jussi Karlgren

Swedish Institute of Computer Science, SICS

Box 1263, SE-164 29 Kista, Sweden

{jon,mange,jussi}@sics.se

Abstract

This paper proposes a novel method for automatically acquiring multilingual lexica from non-parallel data and reports some initial experiments to prove the viability of the approach. Using established techniques for building mono-lingual vector spaces two independent semantic vector spaces are built from textual data. These vector spaces are related to each other using a small *reference word list* of manually chosen reference points taken from available bi-lingual dictionaries. Other words can then be related to these reference points first in the one language and then in the other. In the present experiments, we apply the proposed method to comparable but non-parallel English-German data. The resulting bi-lingual lexicon is evaluated using an online English-German lexicon as gold standard. The results clearly demonstrate the viability of the proposed methodology.

1 Introduction

Using data to build multilingual lexical resources automatically or with minimum manual intervention is an important research goal. Lack of both general-case and domain-specific multi-lingual translation dictionaries hamper the usefulness and effectiveness of multi-lingual information access systems —

most multilingual information retrieval systems use multilingual lexicon lookup coupled with standard search algorithms (for state of the art in multilingual information retrieval, see the proceedings of CLEF¹, TREC², and NTCIR³). Unfortunately, existing multilingual resources have limited coverage, limited accessibility, and are static in nature. Much of the research put into building lexical resources is vectored towards the needs of automatic translation rather than application to information access. The needs for a translation system are very different than those of an information access system: translation puts an emphasis on exact translation and works with fine-grained semantic distinctions, where information access systems typically can make good use of related terms even the match is less than exact. Our aim with the set of experiments presented here is to provide starting steps for the automatic distillation of multi-lingual correspondences from existing data with a minimal amount of processing.

Semantic vector space models are easy to use for experimentation in this vein — they are designed to be efficient, portable, and scalable. Vector space techniques that have been used for multi-lingual lexicon acquisition include Latent Semantic Indexing (Landauer and Dumais, 1997) and Random Indexing or Random Key Indexing (Sahlgren, 2004; Karlgren and Sahlgren, 2001). Both these techniques are originally designed to build semantic models for one language, and are later modified to handle bi-lingual data

¹<http://clef.iei.pi.cnr.it/>

²<http://trec.nist.gov/>

³<http://research.nii.ac.jp/ntcir/index-en.html>

sets. Although the experiments published show surprisingly good results, given their rather limited scope, they typically require large amounts of aligned parallel data, which seriously limits the portability and coverage of the techniques.

In this paper, we propose a novel method for automatically acquiring multi-lingual lexica from non-parallel data. The idea is to make use of established techniques for building mono-lingual vector spaces, and to independently create two (or more) vector spaces, one for each language for which data are available. The basic assumption is that two vector spaces built on similar comparable data from two different languages should be isomorphic to some degree. We exploit this potential isomorphy by aligning the independent vector spaces using a set of reference points from a *reference wordlist* containing a smallish set of established lexical correspondences.

The reference wordlist consists of a set of r bi-lingual pairs of one-word lexemes, which we call *reference words*, and that are thought to have similar semantic roles in the respective languages — e.g. “boat” in English and “boot” in German. The idea is that every word in the different languages could be profiled by measuring their correspondence in the monolingual vector spaces to each of the reference words. Thus, a word’s *reference profile* would consist of one scalar measure for every reference word, resulting in a vector representation in an r -dimensional space, where each dimension is linked to the word’s correspondence to a reference word. The resulting r -dimensional space effectively constitutes a multi-lingual lexicon, where words in different languages are related by their mutual correspondence to the reference words.

In the present experiments, we apply the proposed method to comparable but non-parallel English-German data. We have here used the Random Indexing technology to build the base vector spaces — conceivably other vector space models would work in a similar manner. The resulting bi-lingual lexicon is evaluated using an online English-German lexicon as gold standard.

2 Methodology

Figure 1 demonstrates a simplified schematic of the processing steps for the experiment methodology. The two text collections (1) are non-parallel, and consist of n (some 50 000) unique words each. Cooccurrence statistics for the texts are then collected using the Random Indexing vector space technique, resulting in two $n \times k$ matrices (2). By measuring correspondences between words using some distance measure in the vector space, this space can be represented as a word-by-word matrix (3) of correspondences.

After r words are manually chosen as reference words, a smaller matrix of word-by-word correspondences is produced. If the reference words are well chosen, the strength of the connections in step 4 are assumed to be enough for the matrices to be viewed as representing the same vector space, in a simple approximation of dimensional rotation. This estimate gives us (with n_{en} and n_{de} both $\approx 50\,000$) a $100\,000 \times 100\,000$ matrix (5) of English by German words. In this matrix, the best cross-language correspondences can be found for each row or column.

2.1 Choice of Reference Words

The reference word pairs need to be frequent in the texts to get reliable statistics. We therefore only consider words with a frequency above 75 occurrences. Finding one-to-one translations, however, has not been a priority. If the reference words for the most part have similar meanings in both languages their correspondence in occurrence will match even allowing for some usage difference and some occurrences of synonyms to blur the pictures somewhat. Above all, we have taken care to avoid obviously polysemous reference words. We have sought words as prototypically monosemous as possible. In practice, it has often been a struggle between the demands of rich coverage and occurrence on the one hand and monosemy on the other.

2.2 Random Indexing

Random Indexing, or Random Key Indexing, first developed and published by Kan-

		English	German
1	Text data	n_{en}	n_{de}
RANDOM INDEXING			
2	Mono-lingual vector space	$n_{en} \times k$	$n_{de} \times k$
VECTOR SPACE DISTANCE DEFINITION			
3	Mono-lingual correspondence thesaurus	$n_{en} \times n_{en}$	$n_{de} \times n_{de}$
ESTABLISHING REFERENCE DIMENSIONS			
4	Mono-lingual reference list correspondence	$n_{en} \times r$	$n_{de} \times r$
ALIGNING VECTOR SPACES			
5	Bi-lingual cross-language correspondence	$(n_{en} + n_{de}) \times (n_{en} + n_{de})$	

Figure 1: Steps of the proposed method

erva (Kanerva et al., 2000), and later applied by Sahlgren and Karlgren (Karlgren and Sahlgren, 2001) to information access, is a technique for producing *context vectors* for words based on cooccurrence statistics. Random Indexing differs from other related reduced-dimension vector space methods, such as Latent Semantic Indexing/Analysis (LSI/LSA; (Deerwester et al., 1990; Landauer and Dumais, 1997)), by not requiring an explicit dimension reduction phase in order to construct the vector space. Instead of collecting the data in a word-by-word or word-by-document cooccurrence matrix that need to be reduced using computationally expensive matrix operations, Random Indexing incrementally collects the data in a *context* matrix with fixed dimensionality k , such that $k \ll D < V$, where D is the size of the document collection, and V is the size of the vocabulary. The fact that no dimension reduction of the resulting matrix is needed makes Random Indexing very efficient and scalable. Using document-based cooccurrences with Random Indexing, which we do in the present experiments, is a two-step operation:

1. A unique k -dimensional *index vector* consisting of a small number of randomly selected non-zero elements is assigned to each document in the data.
2. Context vectors for the words are produced by scanning through the text, and each time a word occurs in a document, the document’s k -dimensional index vec-

tor is added to the row for the word in the context matrix. When the entire text has been scanned, words are represented in the context matrix by k -dimensional context vectors that are effectively the sum of the words’ contexts.

In the present set of experiments, we set $k = 2\,000$, with 20 non-zero elements (10 negative and 10 positive unit values) randomly distributed in the 2 000-dimensional index vectors. Sahlgren has reported (Sahlgren, 2004) that in his original experiments on Random Indexing dimensionalities around 2 000 seem to be optimal, and our initial experiments have confirmed this to be a useful starting value.

2.3 Similarity Measure

The vector space model offers a convenient way to statistically measure similarity between words, but exactly how this similarity is best represented is far from self-evident, especially in the case where semantic correspondences from one larger semantic space are mapped onto values in a smaller semantic space. The question what an ideal distribution of the semantic correspondence function would be is open. If there were a clear distinction where words tended to be either close or far away, a binary representation might very well be preferable; this property could be approximated by using some function with a non-linear increase at close proximities to better enhance the effect of thematic closeness.

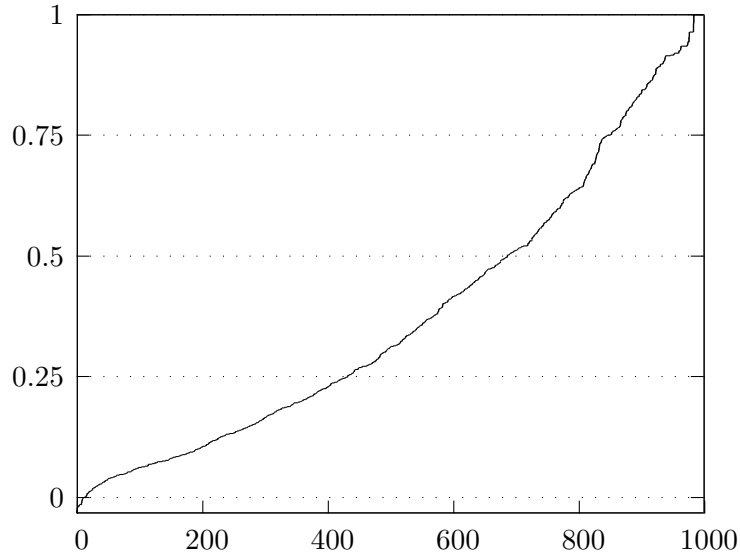


Figure 2: Distribution of correspondence (measured as $\cos\theta$) between 1 000 random word pairs

The simplest way to define similarity in the vector space is to use the standard cosine measure between length-normalized vectors. Measured distributions of cosine-values appear to be close to linear, as seen in Figure 2 (which shows 1 000 sorted results from randomly selected word pairs in a document-coded model with 2 000 dimensions, trained on the texts used in our experiments). There is no non-arbitrary point in the value set where a threshold could be set, and in the present experiments we use the unmodified cosine function.

Only some values below 0 (angles above 90°) and a few hits on 1 (where the angle is 0, and the words identical) can be found in the data. The distribution only slightly favors low and high values, but is comparable to a linear function. For our experiments, we settle with subtracting the median value, to get a good distribution of positive and negative values. Testing randomly selected words suggests a value of 0.33, and thus we will use the function $\cos\theta - 0.33$ throughout, where θ is the angle between the vectors in the 2 000-dimensional space.

2.4 Evaluation Criteria

A possible gold standard in this type of experiment is comparison against a manually compiled English-German lexicon, something that we will be using in these experiments. The cross-references found in such a lexicon are however not typical of any results one could realistically expect. Experiments that sort out the closest matches in a vector space tend to find words that are similar in a very general sense. In going from a vector space of higher dimensionality to a smaller one, it is unavoidable to lose some information. This means that an absolute upper limit for the results would be the results our larger model delivers on one language, or possibly what could be reached by Random Indexing when document-trained on aligned bi-lingual texts. In experiments made on parallel texts Karlgren and Sahlgren (Karlgrén and Sahlgren, 2001) report an approximate precision of 73%.

3 Experiment

3.1 Data

We have used a parallel and lemmatized English and German section of the translated

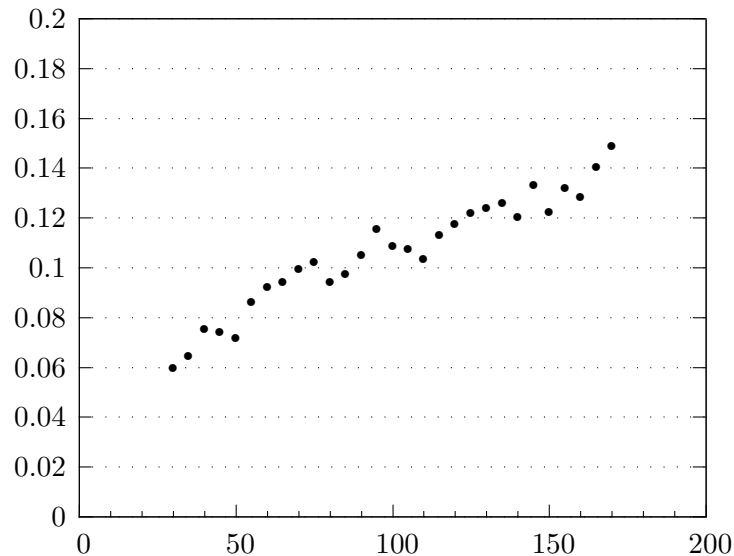


Figure 3: Precision by number of reference words, strict evaluation scheme.

proceedings from the European Parliament of Europarl⁴. 31 250 documents for each language have been chosen from the corpus — if a document has been chosen for the one language, the translation of it was not included. The English documents average 325 words per document and the German documents a bit over 300.

3.2 Reference Wordlist

The reference wordlist is selected through judicious manual selection, by examining the topic matter and terminology of the training data and perusing independent published dictionaries. We have kept the domain of the reference words as close as possible to those discussed in the European Parliament. No independent assessment of reference term translation quality was made; it would be desirable to have a more principled approach to reference term selection, and it would seem to be possible to use term occurrence and distributional characteristics for this purpose.

⁴Europarl consists of parallel texts from the plenary debates of the European Parliament. The data is available at <http://people.csail.mit.edu/people/koehn/publications/europarl/>

3.3 Gold Standard

We use the English-German dictionary of Technische Universität Chemnitz⁵ to evaluate the closest word in the reduced vector space. Every German word given in the dictionary as a translation candidate to the English one is regarded as an equally valid hit in the conducted experiments.

3.4 Procedure

For each run, 50 different words were chosen from the English documents by randomly selecting them from the texts. Words already used in the run and reference words were discarded and new words drawn. We ran evaluation using two schemes: *strict* and *lenient*.

Strict evaluation was done by selecting the single closest German word in the combined cross-language correspondence vector space. If the German word was given as a candidate translation to the English word in the gold standard dictionary, it was counted as a hit. The precision of a run was calculated as the proportion of successful translations of the 50

⁵Technische Universität Chemnitz' German-English Dictionary contains a bit over 170 000 entries and can be found at: <http://dict.tu-chemnitz.de>

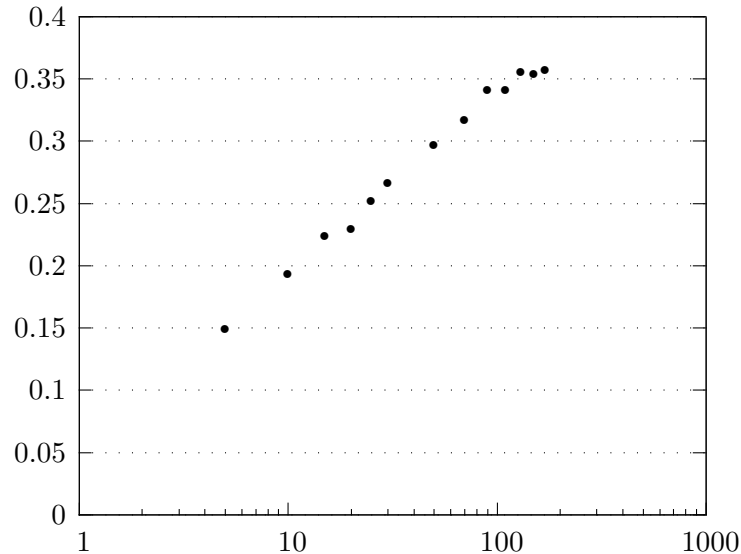


Figure 4: Precision by number of reference words, lenient evaluation scheme.

words.

Lenient evaluation was done by selecting the *ten* closest German words in the combined cross-language correspondence vector space. Also, if an English word did not appear in the gold standard dictionary at all it was discarded and a new word drawn. If any of the ten German words was given as a candidate translation to the English word in the gold standard dictionary, it was counted as a hit. The precision of a run was again calculated as the proportion of successful translations of the 50 words.

The size of the reference wordlist is varied: for each run, r reference words were chosen randomly from the 170 word full reference word list. The test was performed with 50 runs of 50 random words with several different values for r tested, and for each r the average percentage of hits is recorded.

4 Results

Figure 3 shows the precision for the strict evaluation as the reference wordlist size varies from 30 to 170.

Figure 4 shows the precision for the lenient evaluation as the reference wordlist size varies

from 5 to 170.

5 Discussion

The number of dictionary hits ranges from three to twelve out of fifty, depending on size of reference word list and on evaluation scheme. There are a number of observations that can be made from this seemingly low score.

Firstly, the results are amazingly high for reference word lists sizes of *five* and *ten*! With only a very small number of fix points several of the suggestions hit candidates from the lexicon. It is an impressive tribute to the underlying topical similarity of the data that the alignment can be done to that level with only a handful of reference points.

Secondly, the reference word lists were selected manually, and purposely kept small and of high quality. Some preliminary experiments with sloppier and larger word lists (not reported here) gave unsatisfactory results. The quality of the reference word list is — unsurprisingly — a major determining factor for the quality of the results. A principled method for selection and evaluation of the reference pairs would be desirable and the

design of such a method is yet an open question. The coverage and scope of the set of words, the characteristics of the translations correspondences over the languages, the topical characteristics of the words themselves, and the domain orientedness of the set are all factors that have not been systematically studied by us.

Thirdly, the evaluation scheme is very straight-laced. As discussed above, translation dictionaries are designed for purposes different from those we envision for the resource we are developing. Related words of similar meaning but opposite polarity; variants along a semantic dimension; archaic turns of phrase; subsumption hierarchies are none counted as hits by the current scheme.

Fourthly, while this study makes first steps towards evaluating the effects of the reference list on the result quality, no examination of the effects of the quality of the original vector space on the result have been investigated.

In spite of these limitations and reservations, the results are surprisingly promising even given the narrow base of the data. Our belief is that these are the first steps towards a computationally tractable, cognitively plausible, and task- and application-wise reasonable solution for the problem of multi-lingual lexical resources.

References

- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.
- Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036. Erlbaum.
- Jussi Karlgren and Magnus Sahlgren. 2001. From words to understanding. In Yoshinori Uesaka, Pentti Kanerva, and Hideki Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications.
- Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.
- Magnus Sahlgren. 2004. Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data. In *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*.