Automatic Bilingual Lexicon Acquisition Using Random Indexing of Aligned Bilingual Data

Magnus Sahlgren

Swedish Institute of Computer Science, SICS
Box 1263, SE-164 29 Kista, Sweden
mange@sics.se

Abstract

This paper presents a very simple and effective approach to automatic bilingual lexicon acquisition. The approach is cooccurrence-based, and uses the Random Indexing vector space methodology applied to aligned bilingual data. The approach is simple, efficient and scalable, and generate promising results when compared to a manually compiled lexicon. The paper also discusses some of the methodological problems with the prefered evaluation procedure.

1. Introduction

There is a growing need for lexical resources in natural language processing. An increasing number of systems and applications rely on lexica to function; examples range from automatic speech recognizers to word sense disambiguation systems. The need is especially pressing in multilingual tasks and applications, such as machine translation, where multilingual lexica are the arguably most critical components.

Unfortunately, lexica in general, and multilingual lexica in particular, are hard to come by. Manual approaches to lexicon construction vouch for high quality results, but are time-consuming, costly, and static (i.e. they are not easily updated with new information, or tuned to new domains). Automatic lexicon acquisition techniques, on the other hand, provide fast, cheap and dynamic alternatives to manual approaches, but have yet to prove their viability.

This paper investigates a simple and effective approach to automatic multilingual lexicon acquisition. The methodology is based on cooccurrence statistics, and uses aligned bilingual data to find related words across languages. The approach is efficient, fast and scalable, and is easily adapted to new domains and new data.

The proposed methodology is evaluated by extracting a small bilingual lexicon from aligned bilingual English-German data, and by comparing it to a manually compiled gold standard lexicon. The results demonstrate the viability of the approach, and they also point to some of the methodological problems related to evaluating lexical resources.

2. Cooccurrence-based multilingual lexion acquisition

Cooccurrence-based techniques for automatic semantic knowledge acquisition have gained much recognition in recent years. For example, techniques such as Latent Semantic Analysis/Latent Semantic Indexing, LSA/LSI (Deerwester et al., 1990; Landauer and Dumais, 1997), and Hyperspace Analogue to Language, HAL (Lund et al., 1995) use simple cooccurrence statistics to acquire semantic information. This is done by representing the data in a cooccurrence matrix such that the rows represent the words and the columns represent the contexts (or *cooccurrence regions*) used in the model. LSA/LSI uses documents as contexts, and HAL uses words. The cells of the matrix are the cooccurrence counts of a given word in, or with, a given context. The point of this representation is that the rows can be interpreted as *context vectors* for the words, making it possible to express distributional similarity between words in terms of vector similarity.

In this paper, the cooccurrence-based semantic knowledge acquisition methodology is applied to the problem of multilingual lexicon acquisition. This is done by using aligned data, and by defining a context as an alignment region — typically the documents or the sentences in the data. Translations are then defined as words in different languages that have occurred in the same aligned documents or sentences. This means that if an English word and a German word occurs with exactly the same frequency in exactly the same aligned documents, they will get identical context vectors, and we will assume (probably correctly) that they are translations of each other.

The cooccurrence-based methodology, despite its apparent simplicity, has proven to be a surprisingly powerful tool for semantic knowledge acquisition (Dumais et al., 1988; Lund and Burgess, 1996; Landauer and Dumais, 1997; Karlgren and Sahlgren, 2001). However, there are some problematic issues with the methodology. The arguably most serious problem is that the methodology is not very scalable, and that it will become computationally intractable for large data. This problem will be especially severe when using multilingual data, since this requires us to accomodate two different vocabularies.

To alleviate the problem of scalability, cooccurrence-based models normally use some form of dimension reduction. Commonly used techniques include factor analytic methods such as singular value decomposition (used in LSA/LSI) and principal component analysis. Unfortunately, statistical dimension reduction techniques tend to be computationally very costly, and typically can not accomodate dynamic data[1].

[1] The reason is simply that once the dimensionality of the data has been reduced using a factor analytic method, it is not trivial to include new data in the model.

2.1. Random Indexing

One alternative to computationally expensive dimension reduction techniques is the Random Indexing approach (Kanerva et al., 2000; Karlgren and Sahlgren, 2001), which uses *distributed* representations to accumulate context vectors. This is done by representing the contexts in the data (documents or words) by random *index* vectors that constitute a unique representation for each context. These random index vectors are high-dimensional and sparse, which means that they consist of a very small number of randomly distributed non-zero elements (an equal amount of $+1$s and -1s).

The random index vectors are then used to accumulate context vectors by incrementally summing them whenever a word occurs in a particular context. This means that the context vectors have the same dimensionality as the index vectors, and that they contain traces of every context that a word has occurred in — in effect, the context vectors are the sum of the index vectors of the contexts that a word has occurred in. The important thing to note is that the dimensionality of the context vectors does not increase in the accumulation process. It is merely the values of the elements in the context vectors that increases.

Using the Random Indexing approach thus avoids initial sampling of the entire data —- i.e. there is no need to construct a huge cooccurrence matrix as in LSA/LSI or HAL, since the dimensionality of the context vectors is independent of, and much smaller than, the number of contexts in the data. The advantage of using this approach is a significant gain in processing time and memory consumption. Furtermore, the technique is extremely scalable, since new data does not increase the dimensionality of the context vectors. Mathematically, the Random Indexing approach is equivalent to Random Mapping (Kaski, 1999), and Random Projections (Papadimitriou et al., 1998).

3. Experimental setup

3.1. Training data

As training data, the document-aligned English-German Europarl corpus was used (Koehn, 2002)[2]. This data contains some 20 million words in 63,973 aligned documents in each language. The data was lemmatized using the freely available TreeTagger[3].

3.2. Applying Random Indexing

To extract a bilingual lexicon using the Random Indexing approach, one random index vector was assigned to each aligned document pair. Context vectors were then accumulated by adding a document's index vector to the context vector for a given word every time the word occurred in the document. Lexicon entries were created by simply computing the correlation between the context vector for a randomly selected English word, and the context vectors for all the German words, and the German word whose context vector had the highest correlation to the context vector for the English word was entered into the lexicon as a translation of the English word[4].

3.3. Evaluation metric

The quality of the automatically extracted lexicon is characterized in terms of the overlap between a manually compiled gold standard lexicon and the automatically extracted lexicon. The overlap was computed as the precision (i.e. the number of correct entries divided by the total number of entries) of the automatically extracted lexicon. TU Chemnitz' German-English dictionary was used as gold standard[5]. The coverage of this resource is by no means complete, which means that a fair amount of correct, and partially correct, translations will not be featured in the gold standard, and will therefore not be counted as correct entries in the evaluation. This and other problems with the evaluation metric are further discussed below.

4. Experiments and results

Three different sets of English-German experiments were conducted. In the first set of experiments, the relationship between a word's frequency and the quality of its translation was investigated. The second set of experiments investigated the effects of using different dimensionalities of the vectors, and the third and final set of experiments looked at how good translation candidates the second, third, fourth and fifth most similar words in the other language are.

In each set of experiments, the results are reported using average precision over 5 different runs. This is done in order to counter the effects of randomness — since the index vectors are chosen at random, each new run with a different set of index vectors will produce slightly different results.

4.1. Frequency effects

(Grefenstette, 1993) notes that cooccurrence-based techniques for automatic lexicon acquisition are liable to frequency effects — that is, the methods tend to work better for words with high and medium frequency. This might not be very surprising, since high-frequency words provide better statistics, and will therefore get more reliable cooccurrence estimates, than low-frequency words.

In order to investigate whether such a correlation can be identified in the present data, the precision for translations of 100 randomly selected English words were computed using 9 different frequency ranges. These experiments use 1,300-dimensional vectors, and the index vectors contain 6 non-zero elements (three $+1$s and three -1s)[6]. The results are displayed in figure 1.

[2] The Europarl corpora consists of parallel texts from the proceedings of the European Parliament, and is available in 11 European languages. The data is available at http://www.isi.edu/ koehn/europarl/

[3] The TreeTagger is available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] Correlation between vectors was computed as the cosine of the angles between the vectors:

$$d_{cos}(x,y) = \frac{\vec{x} \cdot \vec{y}}{\mid \vec{x} \mid\mid \vec{y} \mid} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

[5] http://dict.tu-chemnitz.de/

[6] These parameters were empirically determined — see the next set of experiments.

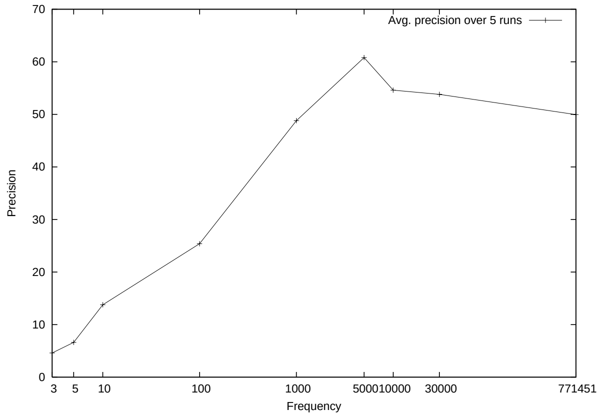

Figure 1: Average precision over 5 runs for 9 different frequency ranges.

As can be seen in figure 1, there is a strong tendency that words with high and medium frequency generate better results than words with low frequency. Top scores are produced for words with a frequency around 5,000 occurrences. Words with a frequency below 100 do not produce reliable statistics and should therefore be excluded from the experiments.

4.2. The effects of dimensionality

In theory, the Random Indexing approach should perform better the closer the dimensionality of the vectors are to the number of aligned documents in the data (Kaski, 1999). In practice, however, it is sometimes the case that several optimal dimensionalities can be found. In this type of application, where the data is two-fold, and very high-dimensional, we could gain considerably in efficiency and scalability by using as low-dimensional vectors as possible. It is therefore important to determine empirically what the optimal dimensionality for this particular data is.

In order to evaluate the effects of dimensionality, 100 English words with frequency between 100 and 100,000 were randomly selected, and the most similar word in German was extracted to each of the English words. 18 different dimensionalities of the vectors were investigated, ranging from 100 to 5,000 dimensions. The index vectors consisted of 2 to 50 (depending on the dimensionality) randomly distributed $+1$s and -1s. The results are displayed in 2.

Figure 2 shows that the results peak when the dimensionality of the vectors is either 1,300 (with 6 non-zero elements), or 3,000 (with 30 non-zeros). Both the best individual score — 68% — and the best average score over 5 runs — 63.28% — were produced with both dimensionalities. The fact that we discover two optimal parameter settings for the present data demonstrates that the method is very sensitive to parameter settings, and that it is important to determine empirically the optimal parameters for a particular data.

Since using the 3,000-dimensional vectors require slightly more memory than using the 1,300-dimensional ones, we use the lower-dimensional ones in the other experiments.

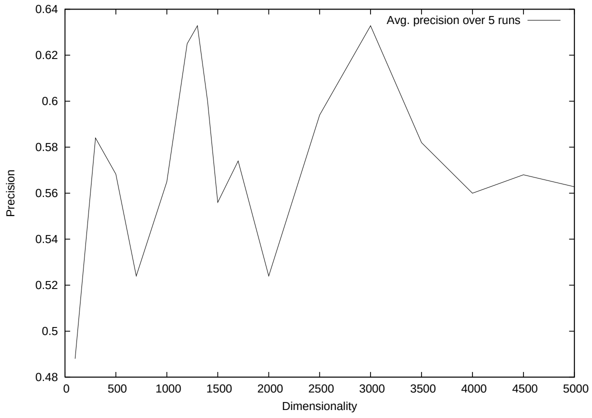

Figure 2: Average precision over 5 runs for dimensionalities ranging from 100 to 5,000.

4.3. Increasing the number of translation candidates

It is a well known fact that some words may have not only one, but several possible, translations in another language (Edmonds, 1998). The vector space representation is particularly suited to handle this situation, since it makes it straightforward to extract several translation candidates to a given word. To do this, we simply compute the correlation between the context vector for a randomly selected English word, and the context vectors for all the German words. We then define the translation candidates as the k most correlated German words. In these experiments, we use $k = 5$.

In order to investigate the quality of the translation candidates, 100 English words with frequency between 100 and 100,000 were randomly sampled from the English-German data, and the five most similar words in German were extracted to each of the English words. Precision was then calculated for each of the translation candidates. The results are shown in 3.

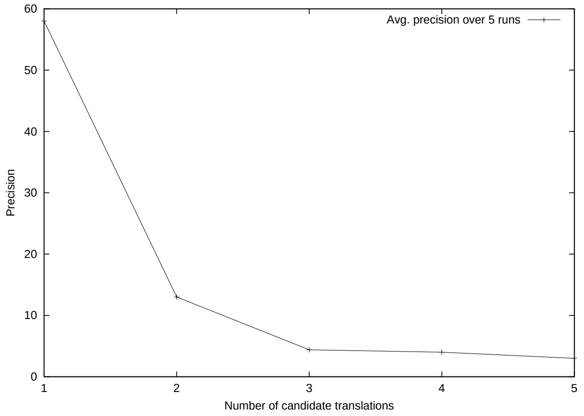

Figure 3: Average precision over 5 runs for 1 to 5 translation candidates.

As can be seen in figure 3, the quality of the translation candidates decreases rapidly. The average precision for the highest correlated candidate is 58%, while it is only 13% for the second best candidate, and a meager 4.4% for the third best candidate. Thus, it seems wise to only include the highest correlated word as a translation candidate in the bilingual lexicon.

5. Discussion

One of the problems with the preferred method of evaluation is that the results depend on the coverage of the gold standard; if the coverage is insufficient, the results will not be reliable. Even though the coverage of TU Chemnitz' German-English dictionary is fairly extensive[7], it is by no means complete. Some of the generated German translation candidates that were not included in the gold standard are viable translations of the English word, such as: "store"/"einlagern", "establish"/"festlegen", "taxation"/"steuern" and "constantly"/"ständig".

A related problem is German compounds that are not included in the gold standard, but that should count as correct, or at least *partially* correct, translations in the given data. Examples from the present investigation include: "steel"/"stahlindustrie", "taxation"/"steuerpolitik", "working"/"arbeitsgruppe" and "working"/"arbeitszeit". These examples demonstrate that the difference between a compounding langauge (such as German) and a non-compounding language (such as English) needs to be specifically addressed in this type of application — e.g. by using decompounding of the German data and of the lexicon (Hedlund, 2003).

Furthermore, there might be domain specific translations and terms that are not covered in any gold standard. One example of a domain-flavoured translation from the present data is "item"/"tagesordnung", which is not featured in the gold standard (although "tagesordnungspunkt" is). In order to cope with these problems, one would need to perform an ocular sanity check of the results in order to arrive at a more consistent measure of the quality of the translations. Doing so for the best individual run (using 1,300-dimensional vectors with 6 non-zero elements, and a frequency threshold between 100 and 100,000 occurrences) increases the results from 68% to 78% (including compounds as correct translations). This demonstrates the need for more refined evaluation procedures for research in multilingual lexicon acquisition.

We conclude that even though the evaluation procedure used in this paper — counting the overlap between the automatically extracted lexicon and a manually compiled gold standard — has a number of inherent problems, the results still demonstrate the viability of the proposed approach for automatic bilingual lexicon acquisition. The results that the Random Indexing methodology is capable of reaching — 78% precision (when the results are manually corrected *a posteriori*) — are promising, and motivates further research into using cooccurrence-based methodology for automatic multilingual lexicon acquisition.

Acknowledgement The work reported here is funded by the European Commission under contract IST-2000-29452 (DUMAS — Dynamic Universal Mobility for Adaptive Speech Interfaces).

6. References

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman, 1990. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

Dumais, S., G. Furnas, T. Landauer, and S. Deerwester, 1988. Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing*.

Edmonds, P., 1998. Translating near-synonyms: Possibilities and preferences in the interlingua. In *Proceedings of the AMTA/SIG-IL Second Workshop on Interlingua*.

Grefenstette, G., 1993. Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. In *Workshop on Acquisition of Lexical Knowledge from Text*.

Hedlund, T., 2003. *Dictionary-Based Cross-Language Information Retrieval: Principles, System Design and Evaluation*. Ph.D. thesis, Tampere University.

Kanerva, P., J. Kristofersson, and A. Holst, 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Erlbaum.

Karlgren, J. and M. Sahlgren, 2001. From words to understanding. In P. Kanerva Y. Uesaka and H. Asoh (eds.), *Foundations of Real-World Intelligence*. CSLI Publications, pages 294–308.

Kaski, S., 1999. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of the IJCNN'98, International Joint Conference on Neural Networks*. IEEE Service Center.

Koehn, P., 2002. Europarl: A multilingual corpus for evaluation of machine translation.

Landauer, T. and S. 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.

Lund, K. and C. Burgess, 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments and Computers*, 28(2):203–208.

Lund, K., C. Burgess, and R. A. Atchley, 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Erlbaum.

Papadimitriou, C. H., P. Raghavan, H. Tamaki, and S. Vempala, 1998. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM Symposium on the Principles of Database Systems*. ACM Press.

[7] TU Chemnitz' German-English dictionary contains 116,532 entries.