Visualisation Techniques for Analysing Meaning

Dominic Widdows, Scott Cederberg, Beate Dorow
*Center for the Study of Language and Information
Stanford University, California
{dwiddows,cederber,beate}@csli.stanford.edu
`http://infomap.stanford.edu`

April 8, 2002

Abstract

Many ways of dealing with large collections of linguistic information involve the general principle of mapping words, larger terms and documents into some sort of abstract space. Considerable effort has been devoted to applying such techniques for practical tasks such as information retrieval and word-sense disambiguation. However, the inherent structure of these spaces is often less well-understood.

Visualisation tools can help to uncover the relationships between meanings in this space, giving a clearer picture of the natural structure of linguistic information. We present a variety of tools for visualising word-meanings in vector spaces and graph models, derived from co-occurrence information and local syntactic analysis. Our techniques suggest new solutions to standard problems such as automatic management of lexical resources, which perform well under evaluation.

The tools presented in this paper are all available for public use on our website.

1 Introduction

Large text corpora are used for many purposes in computational linguistics and NLP. Dictionaries can be built and enriched automatically or semi-automatically using corpora [6]. Bilingual texts can be used to enrich multilingual dictionaries [4]. Word-sense disambiguation systems can benefit from analysing distributional clusters in large corpora [9]. Information retrieval systems are built from large document collections in order to organise and access the information therein [1].

These systems all share the following property. Information derived from the text is built into some mathematical or conceptual model, and it is the model rather than the text itself which is used to solve the problem in question. Partly due to the focus of traditional NLP tasks such as parsing, this point has often been overlooked. Tremendous effort has been devoted to understanding the way the text itself should be processed, based upon sound linguistic principles. However, the properties of the resulting models are often less well-understood: knowledge of formal logic remains much more part of a traditional linguist's training than knowledge of different kinds of mathematical spaces and models.

This paper presents some simple techniques which help researchers and users to understand the spaces they are working with more clearly, using techniques for visualising information that have been developed by the CSLI Infomap project. Our methods are specifically designed to uncover the meanings of words and word groups. We focus on two main types of mathematical spaces: vector spaces and graphs.

*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.

Vector spaces are the underlying spaces in the theory of linear algebra [10]. Points in the space can be specified by giving co-ordinates which measure the amount to which certain features or axes contribute to the point. One typical use of vector spaces is for information retrieval, where the points are words and the 'features' are documents. Using documents which are translated into more than one language, vector spaces can be built which encode multilingual information.

A *graph* in this paper means a set of nodes and a collection of links between those nodes [2]. Undirected graphs have been used to describe semantic networks and directed acyclic graphs have been used to describe ontological hierarchies. Connected with both of these models is the idea that proximity in the model reflects semantic similarity between word meanings.

In this paper we will describe how to build examples of both of these types of model automatically from text-corpora, and describe the tools we have built to enable users to interact with and visualise the results.

2 Vector Spaces

In this section we describe ways in which words can be mapped into vector spaces in such a way that the similarity between two words can be measured. We describe this process for both monolingual and bilingual corpora. We then present a technique for visualising the structure of the resulting space by projecting onto the 2 most significant dimensions.

2.1 Building vector models from corpora

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 [7, 1]. 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* [3].

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 [1, 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* [1, Ch 5].

A variant of the traditional term-document matrix was developed by [8] 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.

To build a bilingual vector model, we proceed 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. [1] We have also built a bilingual vector model from the parallel French/English

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

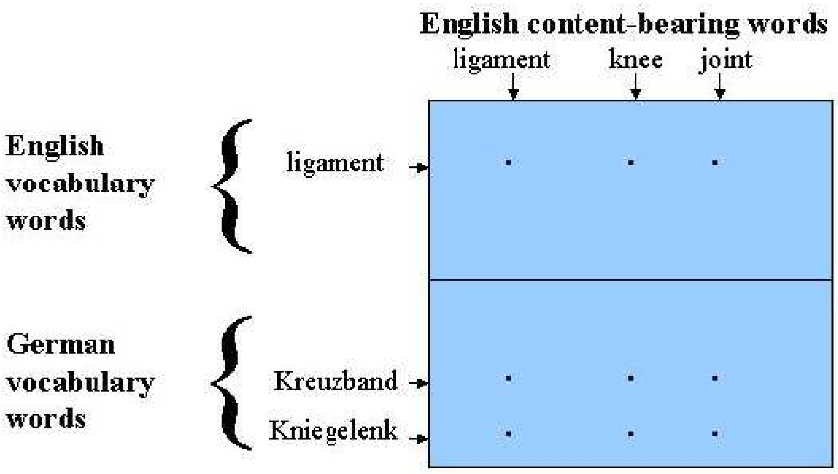

Figure 1: Each vocabulary word in each language is given co-ordinates based on 1,000 English 'Content-Bearing Words'

Canadian Hansard corpus. [2]

Each German/English document pair was treated as a single 'compound document' for the purpose of recording term-term co-occurrence. After stopwords were removed [1, 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 2.1. This bilingual model can be used to represent translational relationships between words, and can therefore be used for the automatic creation and enrichment of multilingual dictionaries, achieving an accuracy of over 90% in cases where the similarity score between translation pairs is high [11].

In this way, words in one or more languages are mapped into a single 1,000-dimensional vector space. Singular value decomposition (LSI) is 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.

2.2 Visualisation by Planar Projection

This 100 dimensional vector space still contains far too many words and too many dimensions to be visualised at once. To produce a meaningful diagram of results related to a particular word or query, we perform two extra steps. Firstly, we restrict attention to a given number of closely related words (determined by cosine similarity of word vectors), selecting a "local context" of up to 100 words and their word vectors for deeper

[2] http://www.ldc.upenn.edu/Catalog/LDC95T20.html.
This model is currently under development and will be publicly available by the time of the TSD conference.

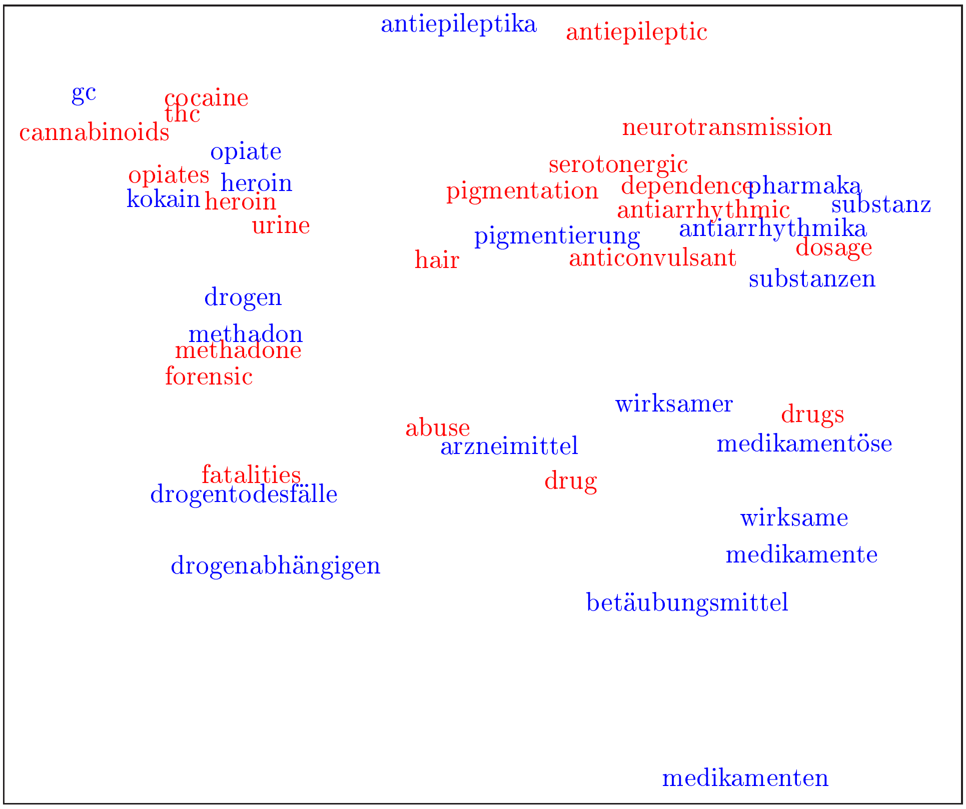

Figure 2: Planar projection of the words similar to the English word *drug* in the bilingual Springer vector model.

analysis. This is done by selecting those words which are most similar to a particular target word, using cosine similarity.

A second round of Latent Semantic Analysis is then performed on this restricted set, giving the most significant directions to describe this local data. The 2 most significant axes determine the plane which best represents the data. (This process can be regarded as a higher-dimensional analogue of finding the line of best-fit for a normal 2-dimensional graph.)

The resulting diagrams give an accurate summary of the contexts in which a word is used in a particular document collection. This is particularly effective for visualizing words in more than one language. Users submit a query statement consisting of any combination of words in English or German, and are then able to visualize the words most closely related to this query in a 2-dimensional plot of the latent semantic space. English words appear in red and German words appear in blue. An example output for the English query word "drug" is shown below. Such words are of special interest because the English word "drug" has two meanings which are represented by different words in German (*medikament* = prescription drug and *drogen* = narcotic). The 2-dimensional plot clearly distinguishes these two areas of meaning, with the English word "drug" being in between. Such techniques can enable users to recognize and understand translational ambiguities.

As well as the bilingual corpus, the system has been trained to work on several (larger) monolingual corpora. Figure 3 shows the corresponding visualisation for the word *drug* built from the New York Times

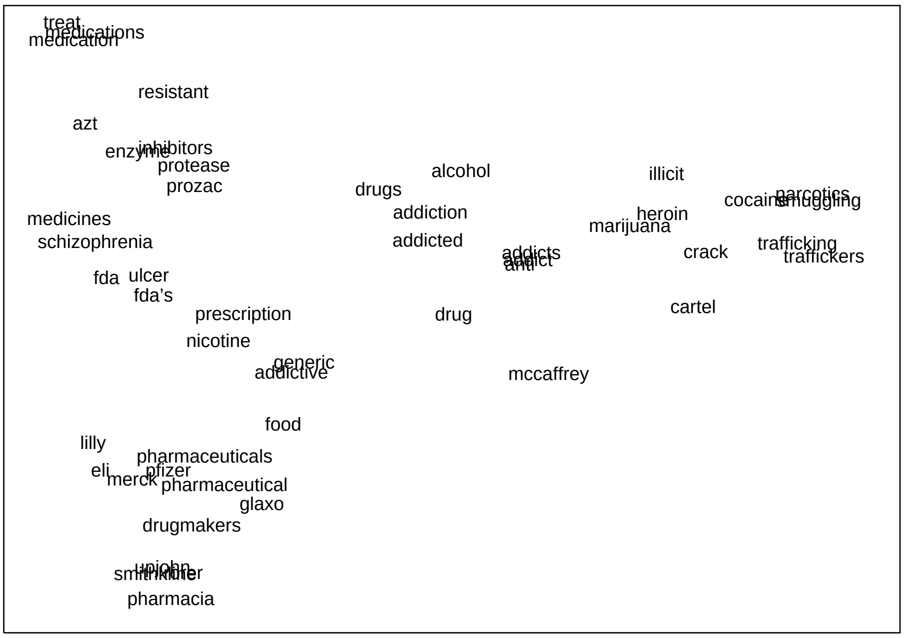

Figure 3: Planar projection of the words similar to the word *drug* in the New York Times vector model.

text corpus. [3] Note again that the different regions correspond to different sense of *drug*.

These models are clearly adept at gathering words into contexts-of-use.

3 Graph models built using local syntactic information

The vector methods above are good at collecting together words which appear in similar contexts. However, they fail to distinguish between words in different semantic classes. So *drug* does appear with the words *pharmaceutical* and *alcohol*, but also words like *illicit*, *trafficking* and the names of drug companies such as *glaxo*.

We demonstrate that these types can be successfully distinguished using part-of-speech information, by building a *semantic graph*. The model was built using the British National Corpus [4] which is automatically tagged for parts of speech.

Each noun in the corpus is taken to be a node in the graph. A link is placed between two nodes if they co-occur in lists, separated by the words *and*, *or* or a comma. For example, consider the following sentences from the BNC:

> But she began to gather their limbs together and put them in order, **head**, **body**, **arms** and **legs**.
>
> A possible reason is that it was difficult to get **arms** and **ammunition** to the right place, despite the virtual absence of border controls between Germany and its western neighbours.

[3] http://www.ldc.upenn.edu/Catalog/LDC95T21.html
[4] http://www.hcu.ox.ac.uk/BNC/

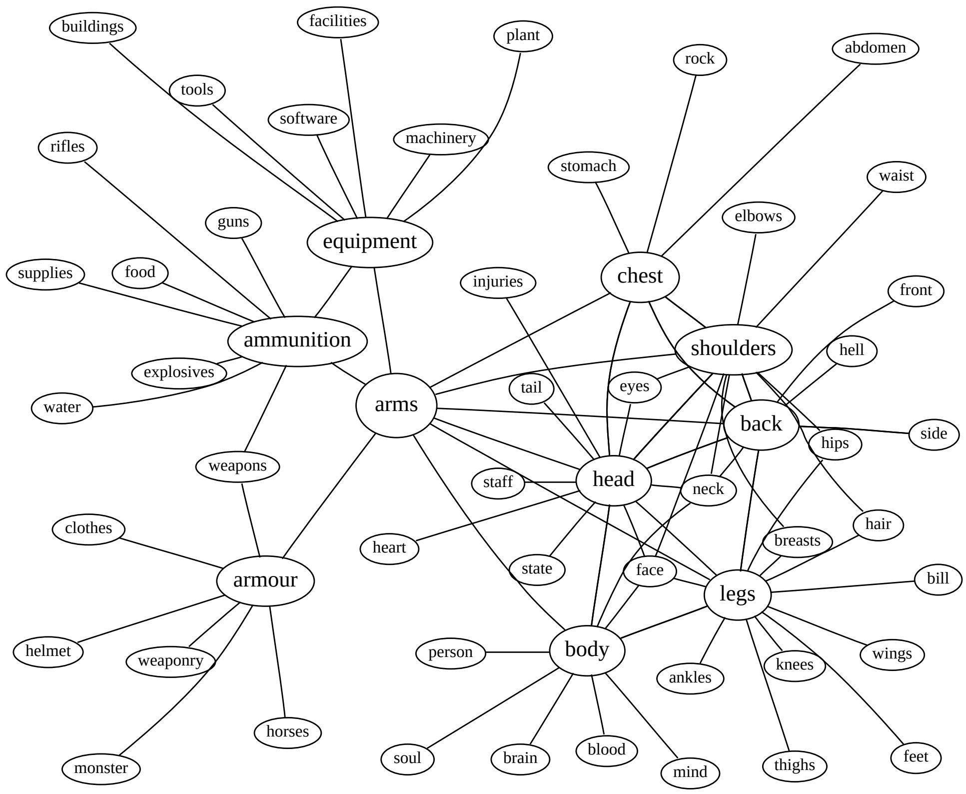

Figure 4: Graph model centred on the word *arms*

Based upon these sentences, we place links between the *arms* node and the *head, body, legs* and *ammunition* nodes. Since lists are usually comprised of objects which are similar in some way, these relationships have been used to extract groups of nouns with similar properties [5] [6].

The links were weighted depending on the number of times the pair of nouns co-occurred. Various cutoff functions were used to determine how many times a relationship must be observed to be counted as a link in the graph. Using a simple rule-of-thumb such as "count two nodes as being linked if they co-occur more than ten times" proved unsatisfactory because of the bias it gives to more frequent words. A better-behaved option was to take the top n neighbours of each word, where n could be determined by the user. More detailed research should reveal optimal techniques for selecting the importance to assign to each link.

As an example, consider the portion of the graph showing the first and second order neighbours of the word *arms* (Figure 4). As well as being an interesting picture, diagrams such as this can be used for practical NLP tasks. Our extremely simple technique has proved to be extremely robust and successful. For example, using an incremental algorithm to add new nodes to clusters, the graph model achieved an accuracy of 82% at a lexical acquisition task similar to that described by Roark and Charniak [6], whose accuracy is only 36%. The overwhelming size of the British National Corpus will account for at least some of this gain. But the gain is also due to our increased understanding of the model we are using, enabled by our visualisation techniques.

The graph model can also be used for ambiguity recognition and resolution, a task which traditionally requires hand-labelled data. This process is both costly and inflexible. Defining what a 'word-sense' is a task that has traditionally been left to lexicographers, with the result that dictionaries omit senses that are relevant in a particular domain, and include senses that are not.

Suppose we want to know which senses of the word *arms* are frequently used in the BNC. Figure 4 can be used to give an empirical answer to this question. Removing the initial node *arms* from Figure 4 leaves two disconnected components, one about *arms* as in parts of the body and one about *arms* as in military equipment. Not only does the model recognise these senses as distinct, it also provides a technique for resolving the ambiguity. Since each sense is empirically derived, we can go back to our empirical observations and annotate them as belonging to one sense or the other. This can then be used as training data for a Bayesian classification. The potential of this system to both recognise *and* resolve ambiguity is currently under investigation. This insight would never have arisen in the first place without the visualisation techniques we have developed.

The semantic graph can also be explored by a user interactively. The user begins with a seed-node and can then select which of its neighbours should be 'expanded' (*i.e.* which should have their own neighbours added to the graph. Figure 5 shows the output of this process, created by a user investigating the links between different countries. This process is potentially valuable as a learning tool — unknown words can be explored until they are linked to known words.

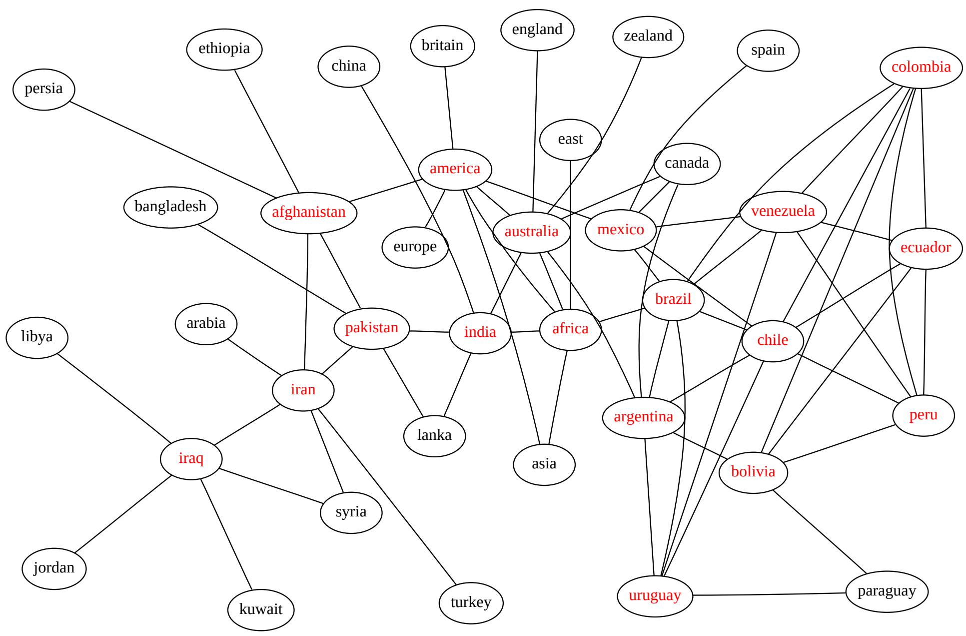

Figure 5: Graph built interactively showing links between countries and regions. The red nodes have been clicked on and expanded by a user exploring the graph.

4 Conclusion

Creating 2-dimensional representations of semantic spaces can provide an excellent means for people to gain a quick, intuitive understanding of how a model built from linguistic information really works. We have also used our visualisation techniques to suggest empirical answers to fundamental challenges in NLP, including results that have already stood up to stringent evaluation criteria.

Methods such as these provide an exciting new extension to the traditional role of corpus linguistics. Rather than deciding our linguistic questions in advance and then approaching corpus material as a statistical resource to provide evidence for our hypotheses, our methods encourage us to observe word meanings with no prior agenda: to hear the corpus speak with its own voice. The tools we have developed enable humans to interpret this information: a tremendous asset when designing new ways to use empirical data in Natural Language Processing.

Demonstration

All of the tools described in this paper are publicly available at `http://infomap.stanford.edu`.

In particular, the bilingual vector model is accessible on `http://infomap.stanford.edu/bilingual` and the graph model on `http://infomap.stanford.edu/graphs`.

The only software needed is a Java-enabled web browser.

References

[1] Ricardo Beaza-Yates and Berthier Ribiero-Neto. *Modern Information Retrieval.* Addison Wesley / ACM press, 1999.

[2] Béla Bollobás. *Modern Graph Theory.* Number 184 in Graduate texts in Mathematics. Springer-Verlag, 1998.

[3] Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

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

[5] Ellen Riloff and Jessica Shepherd. A corpus-based approach for building semantic lexicons. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124. Association for Computational Linguistics, Somerset, New Jersey, 1997.

[6] Brian Roark and Eugene Charniak. Noun-phrase co-occurence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL*, pages 1110–1116, 1998.

[7] Gerard Salton and Michael McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, NY, 1983.

[8] Hinrich Schütze. *Ambiguity resolution in language learning.* CSLI publications, Stanford CA, 1997.

[9] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124, 1998.

[10] Robert J. Vallejo. *Linear algebra: an introduction to abstract mathematics.* Undergraduate texts in mathematics. Springer-Verlag, 1993.

[11] Dominic Widdows, Beate Dorow, and Chiu-Ki Chan. Using parallel corpora to enrich multilingual lexical resources. In *To appear in the Third International Conference on Language Resources and Evaluation.* ELRA, 2002.