Monolingual and Bilingual Concept Visualization from Corpora

Demonstration presented at HLT/NAACL 2003, Edmonton, Canada, HLT/NAACL conference companion, pages 31–32

Dominic Widdows **Scott Cederberg**
Center for the Study of Language and Information, Stanford University
{dwiddows,cederber}@csli.stanford.edu

As well as identifying relevant information, a successful information management system must be able to present its findings in terms which are familiar to the user, which is especially challenging when the incoming information is in a foreign language (Levow et al., 2001). We demonstrate techniques which attempt to address this challenge by placing terms in an abstract 'information space' based on their occurrences in text corpora, and then allowing a user to visualize local regions of this information space. Words are plotted in a 2-dimensional picture so that related words are close together and whole classes of similar words occur in recognizable clusters which sometimes clearly signify a particular meaning. As well as giving a clear view of which concepts are related in a particular document collection, this technique also helps a user to interpret unknown words.

The main technique we will demonstrate is planar projection of word-vectors from a vector space built using Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997; Schütze, 1998), a method which can be applied multilingually if translated corpora are available for training. Following the method of Schütze (1998), we assign each word w 1000 coordinates based on the number of times w occurs in a 15-word window with one of 1000 'content-bearing words', chosen by frequency, and the number of coordinates is reduced to 100 'latent dimensions' using LSA.

This is still far too many words and too many dimensions to be visualized 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 group of up to 100 words and their word vectors for deeper analysis. A second round of Latent Semantic Analysis is then performed on this restricted set, giving the most significant directions to describe this local information. 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 a summary of the areas of meaning in which a word is actually used in a particular document collection.

This is particularly effective for visualizing words in more than one language. This can be achieved by building a single latent semantic vector space incorporating words from two languages using a parallel corpus (Dumais et al., 1996; Widdows et al., 2002b). We will demonstrate a system which does this for English and German terms in the medical domain. The system is trained on a corpus of 10,000 abstracts from German medical documents available with their English translations[1]. In the demonstration, 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.

An example output for the English query word *drug* is shown in Figure below.[2] 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 Springer abstracts corpus, the system has been trained to work with the parallel English/French Canadian Hansard corpus and several large monolingual corpora. Other functionalities of this system include automatic thesaurus generation, clustering of terms to determine different context areas, query refinement, and document retrieval.

As well as LSA, which only uses broad 'bag of words'

[1] Available from the Springer Link website, http://link.springer.de/

[2] In the actual demonstration, English results appear in red and German results in blue; for the description here we have used different fonts instead.

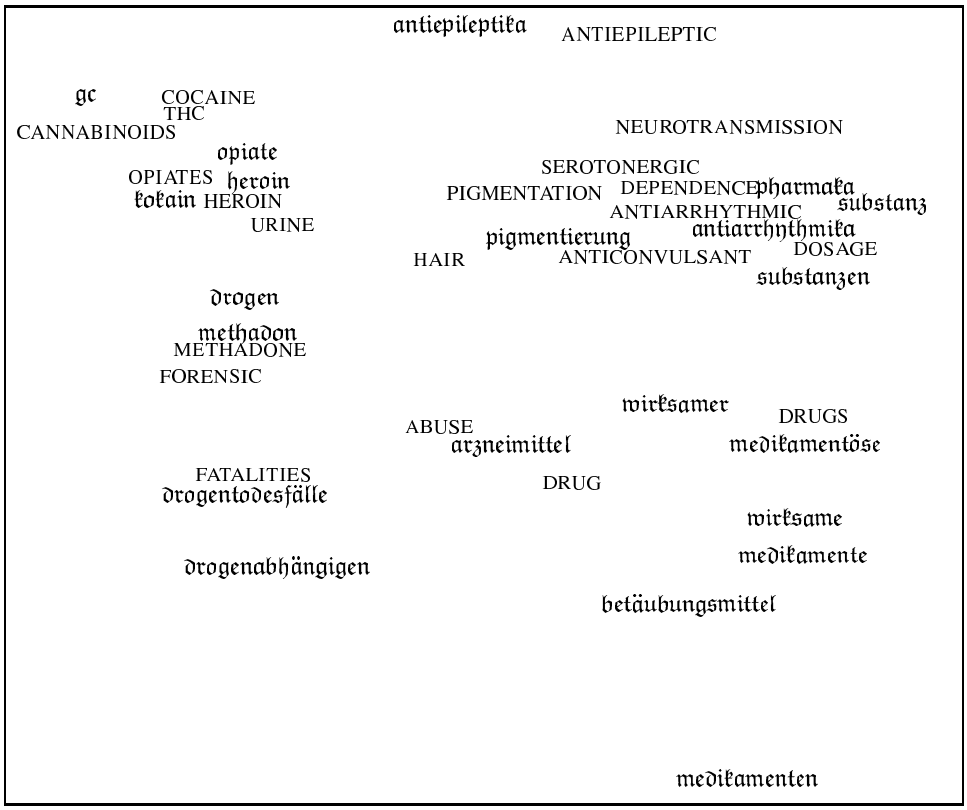

Figure 1: ENGLISH and German terms related to the English word *drug* in the Springer medical abstracts.

coocurrence to define similarities, mathematical models can be built using local coordination of terms based on syntactic properties. For example, lists of nouns such as "apples, pears, and oranges" can be used as information that these words are all linked, and these links can be recorded in a database which can also be analyzed using visualization techniques (Widdows et al., 2002a) and will be included in the demonstration.

Demonstration website

Versions of these demonstrations are publicly available through the CSLI Infomap project website, (`http://infomap.stanford.edu/`).

Acknowledgments

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.

References

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

Thomas Landauer and Susan Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition. *Psychological Review*, 104(2):211–240.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2001. Rapidly retargetable interactive translingual retrieval. In *Human Language Technology Conference (HLT 2001), San Diego, CA*.

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

Dominic Widdows, Scott Cederberg, and Beate Dorow. 2002a. Visualisation techniques for analysing meaning. In *Fifth International Conference on Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 2448, pages 107–115, Brno, Czech Republic, September. Springer.

Dominic Widdows, Beate Dorow, and Chiu-Ki Chan. 2002b. Using parallel corpora to enrich multilingual lexical resources. In *Third International Conference on Language Resources and Evaluation*, pages 240–245, Las Palmas, Spain, May.