Ongoing Developments in Automatically Adapting Lexical Resources to the Biomedical Domain

Fifth International Conference on Language Resources and Evaluation (LREC 2006),
Genoa, Italy, May 24-26, 2006.

Dominic Widdows[1], Adil Toumouh[2], Beate Dorow[3], Ahmed Lehireche[2]

[1] MAYA Design, Inc., Pittsburgh, Pennsylvania
[2] Computer Science Departement, University Djillali Liabes
[3] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Abstract

This paper describes a range of experiments using empirical methods to adapt the WordNet noun ontology for specific use in the biomedical domain. Our basic technique is to extract relationships between terms using the Ohsumed corpus, a large collection of abstracts from PubMed, and to compare the relationships extracted with those that would be expected for medical terms, given the structure of the WordNet ontology. The linguistic methods involve the use of a variety of lexicosyntactic patterns, that enable us to extract pairs of coordinate noun terms, and also related groups of adjectives and nouns, using Markov clustering. This enables us in many cases to analyse ambiguous words and select the correct meaning for the biomedical domain. While results are often encouraging, the paper also highlights evident problems and drawbacks with the method, and outlines suggestions for future work.

1. Introduction

Lexical resources such as ontologies are important for creating knowledge-rich NLP applications. Because of this, large projects such as the Princeton WordNet have been funded for many years to build such resources manually. However, developers and knowledge engineers often find such general resources to be inappropriate for their needs when creating applications in specific domains, since they contain many senses that are not appropriate and actually misleading for their domain. In such circumstances, developers often resort to the manual creation of completely new domain-specific resources. Not only is this a costly process, but it also prevents interoperation between NLP applications in different domains and the gradual emergence of semantically stable standards. Ideally, it would be possible for developers to automatically select those parts of a general resource that are appropriate for their needs, obtaining a 'tailor-made' lexical resource that maintains general interoperability without costly manual intervention.

This paper describes automatic techniques for adapting the WordNet noun taxonomy (Fellbaum, 1998) to the medical domain, by comparing the available senses of nouns given by WordNet with the distribution of words in the Ohsumed corpus (Hersh et al., 1994), a large selection of documents from the PubMed catalogue. Though specific in nature, our experiments are based on a very small set of assumptions about lexicosyntactic patterns in the English language, and thus there is hope that they may be easily adapted to other domains and other languages where lexicosyntactic patterns are a reliable guide to semantic usage.

2. General Strategy of these Experiments

Since Aristotle wrote the *Categories* and the analytical works that introduced the study of logic, conceptual structures describing things in the world have tended to focus on creating inheritance taxonomies. (For a thorough introduction to this topic, see Sowa (2000).) Thus, the most widely studied relationship between concepts is the inheritance relationship between a more general class of objects and a more specific subclass, described as the 'hyponymy' relationship in WordNet (Miller, 1998a). While other relationships have been considered at some length (see (Fellbaum, 1998), (Bean and Green, 2001)), the most comprehensive resources that combine both lexical and conceptual knowledge remain largely taxonomic, examples including WordNet (used in our experiments) and EuroWordNet (Vossen, 1998).

One well-developed family of empirical methods for finding relationships between concepts from texts that discuss those concepts is to use *lexicosyntactic patterns*. Such pattern-matching techniques were pioneered in the work of Hearst (1992), who noted that patterns like

$$\text{“}A \text{ such as } B\text{”} \quad \text{and} \quad \text{“}B \text{ and other } A\text{”}$$

often indicate that B is a kind of A (sometimes written as the symbolic relationship $B \sqsubseteq A$). However, such direct references to inheritance structure in written text appear to be comparatively few and far between, when compared with textual evidence for *similarity*. At the very least, it is certainly the case that pattern matching techniques that search for similar terms as well as taxonomically related terms are likely to obtain higher recall, and it is demonstrably possible to improve hyponymy-extraction engines by considering related terms (Cederberg and Widdows, 2003). For example, if one encounters the pattern

$$\text{“}A, B, \text{ and } C\text{”}$$

and one already has an ontology that contains the relationship $A \sqsubseteq X$ for some X, it is a reasonable hypothesis that the relationships $B \sqsubseteq X$ and $C \sqsubseteq X$ are also valid.

Of course, this rule-of-thumb is a long way from being a generally valid syllogism, and while demonstrating improvements in recall, the work in Cederberg and Widdows (2003) also highlighted many of the pitfalls that these methods encounter. One particularly lexical concern is ambiguity. For example, from the phrase

mass and other religious services

the method deduced the relationship *mass* $\sqsubseteq$ *religious service*, and from the phrase

mass, charge and spin

deduced that *charge* and *spin* are similar terms to *mass*. However, the conclusion that *charge* and *spin* are also *religious services* is mistaken, because the relationships in question are between different senses of the ambiguous term *mass*.

Though such examples are a problem for any attempt to infer relationships naïvely between words, they are not altogether negative. A system encountering the phrase "mass and charge" may well use the cooccurrence of these terms to infer than, in this context, *mass* refers to a physical property not a religious service. As an alternative example, a document containing the phrase "doctor and nurse" is probably using *doctor* to mean *physician* rather than *learned person*. This observation was made by Resnik (1999) and used for the purpose of word sense disambiguation.

This line of research led us to the main purpose of this paper. In order to successfully adapt general ontologies to particular domains, learning new domain-specific terms (ontology enrichment) is not sufficient. It is also necessary to work out which senses of terms given in the general ontology are valid in the specific domain. This task may be described as *ontology pruning*. As a practical task, ontology pruning is increasing in importance as lexical resources increase their coverage. Ten years ago, the refrain "ontologies don't contain the stuff I need" may have been common. Nowadays, the complaint "ontologies introduce a whole load of stuff I don't need" is just as valid.

Progress in both automatic relation extraction and word sense disambiguation over the past 15 years led us to believe that the problem of ontology pruning could be addressed by comparing the results of lexicosyntactic pattern matching experiments with the structure of the WordNet noun taxonomy. It should also be noted that experiments in domain adaptation of lexical resources have been performed using broader statistical analyses (see for example (Buitelaar and Sacaleanu, 2001)), and this general field of research seems set to expand in importance.

The methods discussed in the following sections include extraction through noun-noun coordination patterns and by Markov clustering of adjective-noun pairs. The methods were all applied using the combination of the WordNet noun taxonomy and the Ohsumed corpus (Hersh et al., 1994), which contains several years of abstracts from the PubMed catalogue of the US National Library of Medicine.

3. Noun Coordination Pattern Ontology Pruning System

Two or more nouns are described to be in *coordination* (Huddleston and Pullum, 2002, p. 1287) if they occur in a list, e.g., if they are separated by a conjunction or are comma-separated. For example, in the phrase "apples, pears and ackees," each of the three nouns occurs in coordination with the other two. This information can be used to great effect, for example, to build a graph in which each noun is a node and two nodes are linked if they occur in a coordination pattern in some corpus (Widdows and Dorow, 2002), (Widdows, 2004, Ch 2, 4), and this technique has led to successful results in lexical acquisition tasks. (For example, you may not have known that an *ackee* is a fruit, but you might be tempted to hazard a guess that it is, and in this case, you would be correct.)

Because of the relative maturity and success of systems based on noun coordination pattern extraction, an end-to-end prototype for selecting medical senses of WordNet terms was constructed and evaluated. Our system architecture is described in Figure 1.

An important concept in this process is the Lowest Common Ancestor (LCA) of a pair of terms in the WordNet taxonomy. The LCA corresponds to the Least Upper Bound or *join* in lattice theory and the disjunction in logic. The LCA of two example noun pairs is shown in Figure 2. As described by Resnik (1999), the LCA concept can be used to select between several senses of an ambiguous term in a coordination pattern. The LCA of the terms A and B can be written $A \vee B$, using lattice theoretic notation. For an ambiguous term A with senses A_1, A_2, etc., occurring in the coordination pattern "A and B", we compute the LCA $A_j \vee B$ for each of the senses A_j. The sense which is most closely related to B is usually the sense A_j whose LCA $A_j \vee B$ is lowest in the hierarchy. If B is a term from the medical domain, it then makes sense to assume that the sense A_j is the most relevant to the medical domain.

Step by step, the algorithm for using noun coordination patterns to find such medical senses is as follows:

1. In Stage 1, we start by processing the Ohsumed corpus in order to extract pairs of nouns separated by a conjunctions, from which we constitute a set of word pairs (the two nouns of the conjunction). This method extracted word pairs, including general and medical terms. (Words that occur more in Ohsumed than in the British National Corpus (`www.natcorp.ox.ac.uk`) are considered to be more domain-specific.) These two tasks are performed by the *Noun_CJC_Noun Extraction Module* and *Terms Frequencies Extraction Module*. Stage 1 therefore outputs a collection of pairs of words which are important to the medical domain, though it does nothing to indicate which senses of these words are important in the medical domain.

2. The most important part is Stage 2. The *Similarity Module* attempts to find the Lowest Common Ancestor (LCA) for each word pair, dividing the results into a "success set" of pairs for which an LCA was found, and a "failure set" of pairs for which no LCA was found. (This set largely consists of pairs containing words which are absent from WordNet.) The success set is further partitioned into 3 sets as follows:

   - Those terms for which a single LCA was found which gave a unique sense of the input terms
   - Those terms for which multiple LCAs were found

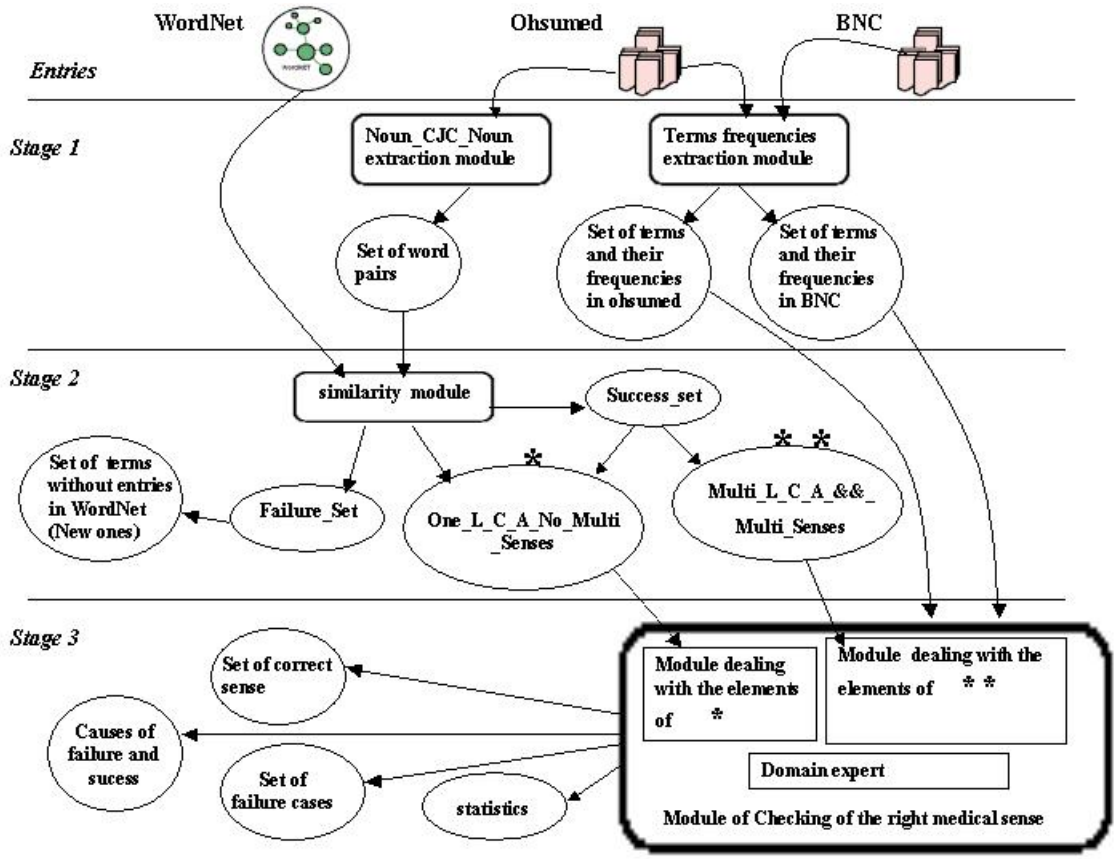

Figure 1: System architecture for coordinate noun based ontology adaptation

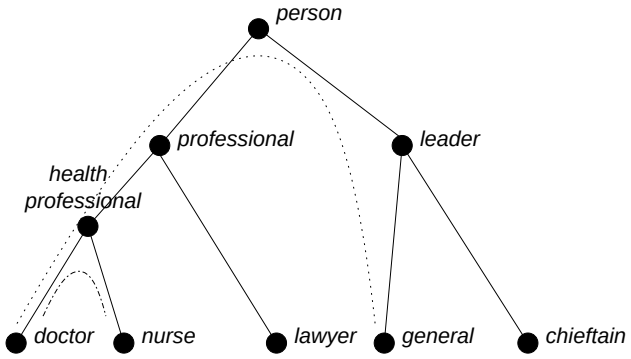

Figure 2: A fragment of the WordNet hierarchy, showing that the LCA of *doctor* and *nurse* is *health professional*, whereas the LCA of *doctor* and *general* is *person*.

- Those terms for which a single LCA was found which subsumed multiple senses of the input terms

3. In Stage 3, we evaluate the results in the success set output by stage 2. The *Right Medical Sense Checking Module* is made up of two modules and assisted by an expert of the domain.

3.1. Results of Noun Coordination Experiments

We extracted from Ohsumed 57887 word pairs (the two nouns of conjunctions)(115774 words). In Stage 2, the cardinality of the Success_Set is 34220 pairs, and that one of the Failure_Set (for which no LCA was found) is 23667 pairs. Of the Success_Set, 26399 words had multiple senses in WordNet but a unique sense was chosen using our system. This was the largest collection of results, and also the one for which the best system performance was observed. Due to the fact that Success_Set is large, we used the first 126 pairs as a sample for expert evaluation. This sample contained 84 pairs for which a unique sense was proposed for ambiguous words. It includes 136 domain-specific words and 32 general ones. For the domain-specific terms (those occurring more frequently in the Ohsumed corpus than the BNC), success in keeping the right sense for the medical domain was 72%. Performance for the more general terms was less effective. Similarly, performance on those terms for which multiple LCAs were found was less good (slightly over 50%). These results are described in much more detail by Toumouh et al. (2006).

4. Adjective-Noun Clustering Experiments

In this section, we describe a preliminary experiment which investigates the potential of syntagmatic relationships for improving the recognition of a noun's medical senses. The experiments follow the pattern of building a combinatoric model of words and their relationships based upon some

shared occurrence pattern in a corpus, and using clustering techniques on the resulting graph (Dorow, 2006).

Many of the coordinated noun pairs in the Failure_Set are not related by (co-)hyponymy as assumed, but instead by, for example, entailment, meronymy (part-to-whole) or a noun-to-attribute relation (e.g., *dose and efficacy* and *memory and its decline*, *a vascular renin-angiotensin system and its role in modulating sympathetic activity*). Such terms are hence distant in the WordNet noun hierarchy, though they are still related. (This inadequacy with purely taxonomic representations of meaning is sometimes described as the *Tennis Problem* (Fellbaum, 1998).)

Syntagmatic relationships, such as adjective-noun, verb-subject and verb-object relationships, have been successfully used to assess similarity between words (Lin, 1998). A logical next step in our experiments with the Ohsumed corpus was to try to extract relations based upon lexicosyntactic patterns that tend to reflect syntagmatic relationships. To assess and compare the discriminative power of syntagmatic relationships with that of simple coordination information, we proceeded as follows.

- We used pattern matching on the PoS tags to collect pairs of adjectives and the nouns they modify. We extracted a total of $206,595$ adjective-noun pairs.
- Rarely occurring pairs (frequency ≤ 3) and weak associations (mutual information ≤ 6) were discarded.
- This left $35,244$ significant adjective-noun pairs (involving 4,615 nouns and 4,722 adjectives).

The set of adjective-noun relationships naturally gives rise to a bipartite graph, a graph which consists of two sets of nodes, adjectives and nouns, and in which links connect an adjective with the nouns it modifies and vice versa. From this bipartite adjective-noun graph, we can construct a projected noun-noun graph, by linking any two nouns if they share a critical number of adjectives (5 in our experiment). Ignoring isolated nodes, the projected noun-noun graph contains 662 nodes and $2,168$ links. This is much smaller than the earlier noun-noun graph from coordination patterns alone, mainly because the filtering conditions were quite strict. Recall could easily be improved by relaxing these conditions, though our original goal was to quickly compare the precision of this method with that of the noun-noun coordination method.

Using Markov Clustering (van Dongen, 2000), we divided the noun-noun graph into 120 (non-overlapping) dense regions representing groups of semantically similar words, some of which are listed in Table 1.

Our assumption is that within a cluster, each noun assumes only one sense, namely its medical sense. In order to find this medical sense, our strategy was to repeat the LCA-based algorithm described in earlier, for each other cluster member of the noun whose sense we were trying to identify. For a first comparison between this adjective-noun based method and the noun coordination based method described earlier, we compared the senses given by this process with those given by the earlier method on words that were already in the 126 trial set. The (rather sparse) results of this evaluation are presented in Table 2. Of the pairs tested, 35 gave the correct medical sense, 22 gave an incorrect sense, and 95 gave no LCA at all. Even though there are more correct labels than incorrect ones, this is certainly discouraging as far as recall goes, especially considering that the noun clusters already contained comparatively few nouns (662).

It was therefore decided not to pursue this method before reconsidering the approach in general. This is partly because at least two observations were made during the work that question the relevance of the results.

Firstly, since different senses were obtained for the same term using different cluster members, it questions our assumption that the correct medical sense can be obtained on a type level rather than a token level. Ambiguity of terms

Table 1: Examples of noun clusters

hospital institution care practice practitioner center unit clinic program setting editorial agency medicine department physician service application facility
number weight magnitude quantity size extent severity degree potency amount prevalence percentage affinity numbers count proportion frequency distance
decline accumulation loss reduction increase difference shift decrease drop rise improvement progress elevation fall increment
lung heart pancreas islet liver brain hepatocyte kidney myocardium erythrocyte
removal gastrectomy excision ablation revascularization extraction resection correction reversal
episode depression complication event morbidity stroke death occurrence mortality toxicity life perio
pain headache incontinence diarrhoea diarrhea attack fever
serum plasma blood hematocrit globulin
malformation anomaly anatomy patholog

Table 2: Nouns occurring in the adjective-noun clusters and in the 126 sample terms from manual evaluation

Word	**Correct**	**Incorrect**	**No LCA**
female	15	0	18
treatment	6	6	19
cell	4	2	7
plasma	1	0	0
antagonist	8	6	9
pressure	0	3	3
stage	0	1	2
head	1	2	24
risk	0	0	2
stroke	0	2	9
contraction	0	0	2

in the medical domain appears to be considerably less than ambiguity of English words in general, as can be seen by comparing some of the experiments in the MUCHMORE project (Widdows et al., 2003) with some of the more general SENSEVAL experiments (Kilgarriff and Rosenzweig, 2000). However, it is not negligible, and so far our experiments have not reflected this.

Secondly, some of the clusters seem to be semantically coherent, but not really taxonomic enough to correspond well with the WordNet hierarchy. For example, consider the cluster

> *episode depression complication event morbidity stroke death occurrence mortality toxicity life period*

Most of the terms are (arguably) about some medical event, some kind medical event, or a condition that a patient may or may not experience (some of the terms clearly correspond to what genetic epidemiologists would call a *phenotype* (Haines and Pericak-Vance, 1993, p. 53)). However, this intuitive cluster is certainly too fuzzy or contextual to correspond to any particular branch of the WordNet noun hierarchy. The process of combining two syntagmatic adjective-noun relationships in the hope of obtaining a paradigmatic relationship, because our evaluation is based upon paradigmatic relationships, may be overlooking the more subtle opportunities of the situation, in the hope of finding something that we have more experience in evaluating.

4.1. Adjective clusters

Some of the adjective clusters obtained using our method were also of interest, and a selection is presented in Table 3. Some of the larger clusters appear to be diverse groups of adjectives that can be used to modify people or diseases. However, some of the smaller groups are quite specific, and refer to fairly tight clusters of similar medical concepts, for example, the cluster

medial lateral posterior anterior

of terms that that describe anatomical positions. Many of the clusters comprising just 2 members are in fact antonyms, and would perhaps stand up well to a comparison with the antonym organization of modifiers in WordNet (Miller, 1998b).

Table 3: Examples of adjective clusters

young affected female elderly healthy white black obese nondiabetic hypertensive diabetic asymptomatic risk
acute fatal severe mild moderate recurrent idiopathic chronic threatening persistent important critical key distinct main major potential specific minor
intact murine recombinant human bovine normal mutant fetal abnormal
substantial considerable relative similar great limited equal variable comparable
percent total free high average low elevated median
medial lateral posterior anterior
bacterial microbial viral
external internal
routine serial
malignant benign
neonatal newborn
bilateral unilateral
dominant recessive

5. Further Work

There are clearly many directions in which this work could be taken, some of which we intend to pursue. These areas include the following.

Enhanced preprocessing

Many medical terms are not single words, but complex nounphrases. In these cases, it is sometimes (though not always) possible to infer something about the semantics of a term from its constituents. This hypothesis was tested in some detail by Baldwin et al. (2003), using a distributional technique (latent semantic analysis) to try to determine when a complex term could be regarded as a hyponym of its syntactic head. In the medical domain, this hypothesis may be more regularly valid, since non-compositional idioms are presumably rarer in scientific writing. (Alternatively, this hypothesis could be a red herring.)

At the very least, it would be worthwhile to add some nounphrase chunking operation to our preprocessing step, and to repeat our experiments using nounphrases and similar extracted complex terms as well as just words.

Alternative extraction patterns

As well as noun coordinations and adjective-noun patterns, verbal patterns should be considered. The result will be a complex graph structure with (at least) nouns, verbs and adjectives as nodes, and labelled links. Such a semantic model will complement more distributional and probabilistic techniques such as latent semantic analysis and n-gram modelling. In the long term, we hope to provide an empirical semantic complement to purely statistical language models.

Adaptation to other domains and languages

While our work has focussed on the English language and the medical domain, we have deliberately operated with the goal of keeping our linguistic assumptions comparatively simple. For any language with suitable part-of-speech tagged corpus material, it is not difficult to define and use basic lexicosyntactic patterns. Once relationships have been extracted and built into a mathematical model such as a graph, the subsequent combinatoric analysis is language-independent.

This is not to say that our methods will work automatically in other domains and languages, but it does imply that it will comparatively easy to test them in other situations.

Distributed publication

Our results are clearly not yet mature enough to be considered as a trustworthy automatic adaptation of WordNet. However, the trend is going gradually in this direction, not only for economic, but for engineering reasons. As semantic technology matures, part of its value will depend on the potential for deployment to small devices, in which case, pruning an ontology may be seen as not only semantically useful, but practically necessary.

A distributed architecture for publishing subsets of corpora and lexical resources is described by (Balasubramanya et al., 2006), with these specific goals in mind. In this model, entries from resources such as WordNet will be definitively referred to, but researchers will be able to create standoff annotation to the effect that (for example) a particular sense is relevant to the medical domain. Then, instead of making a static choice of which version of WordNet is preferable, researchers and deployed systems can make a dynamic choice of which publishers they trust to create good resources in particular domains, and subscribe to the data created by these publishers as it becomes available.

6. Conclusions

By comparing the senses given for a word in a general lexical ontology with the usages of that word in a domain-specific corpus, it is possible to adapt general lexical resources to a specific domain with high accuracy in some cases, especially using patterns in text that directly reflect paradigmatic similarities.

This can be accomplished using comparatively simple techniques. While it is necessary to improve and extend these techniques to guarantee superior performance, the simplicity of our assumptions gives us reason to hope that our system can be adapted to other languages and domains at little extra cost.

This research is still in its early stages, and pursuing it properly will involve close interactions with other syntactic and semantic work in understanding the nature of relationships beyond taxonomies.

7. References

M. Balasubramanya, M. Higgins, P. Lucas, J. Senn, and D. Widdows. 2006. Collaborative annotation that lasts forever: Using peer-to-peer technology for disseminating corpora and language resources. Genoa, Italy, May 24-26, May.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Carol A. Bean and Rebecca Green, editors. 2001. *Relationships in the Organization of Knowledge*. Kluwer.

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of WordNet and Other Lexical Resources*, NAACL 2001 Workshop, Pittsburgh, PA, June.

Scott Cederberg and Dominic Widdows. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Conference on Natural Language Learning (CoNNL)*, Edmonton, Canada.

Beate Dorow. 2006. *A graph model for words and their meanings*. Ph.D. thesis, IMS, University of Stuttgart.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge MA.

Jonathan L. Haines and Margaret A. Pericak-Vance, editors. 1993. *Approaches to Gene Mapping in Complex Human Diseases*. Wiley-Liss.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, Nantes, France.

William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual ACM SIGIR Conference*, pages 192–201.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1-2):15–48, April.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, Montreal, August.

George A. Miller. 1998a. Nouns in wordnet. In Fellbaum (Fellbaum, 1998), chapter 1, pages 23–46.

Katherine J. Miller. 1998b. Modifiers in wordnet. In Fellbaum (Fellbaum, 1998), chapter 2, pages 47–68.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:93–130.

John F. Sowa. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing, Pacific Grove, CA.

A. Toumouh, A. Lehireche, D. Widdows, and M. Malki. 2006. Adapting wordnet to the medical domain using lexicosyntactic patterns in the ohsumed corpus. In *Proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06)*, Sharjah, UAE, March.

Stijn van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht, May.

Piek Vossen. 1998. Introduction to eurowordnet. *Computers and the Humanities*, 32(2-3):73–89.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099, Taipei, Taiwan, August.

Dominic Widdows, Diana Steffen, Scott Cederberg, Chiu-Ki Chan, Paul Buitelaar, and Bogdan Sacaleanu. 2003. Methods for word-sense disambiguation. Technical report, MUCHMORE project report.

Dominic Widdows. 2004. *Geometry and Meaning*. CSLI publications, Stanford, California.