

Semantic-Based Information Retrieval of Biomedical Data

Peng Yan¹, Yu Jiao², Ali R. Hurson¹, and Thomas E. Potok²

¹Computer Science and Engineering Dept.
The Pennsylvania State University
University Park, PA 16802
pzy1@psu.edu, hurson@cse.psu.edu

²Oak Ridge National Laboratory
1 Bethel Valley Road
Building 5700, MS 6085
jiaoy, potokte@ornl.gov

ABSTRACT

In this paper, we propose to improve the effectiveness of biomedical information retrieval via a medical thesaurus. We analyzed the deficiencies of the existing medical thesauri and reconstructed a new thesaurus, called MEDTHES, which follows the ANSI/NISO Z39.19-2003 standard. MEDTHES also endows the users with fine-grained control of information retrieval by providing functions to calculate the semantic similarity between words. We demonstrate the usage of MEDTHES through an existing data search engine.

Categories and Subject Descriptors

H.4 [Information Systems Applications]

Keywords

Semantic-based Information Retrieval, MeSH, Thesaurus, Semantic Similarity

1. INTRODUCTION

The effectiveness of information retrieval is assessed by the ability of the system to retrieve relevant documents while, at the same time, suppressing the retrieval of irrelevant documents. A critical problem in information retrieval is that the vocabulary that a user chooses may not be the same as the one used to index the documents. For example, if a user uses a synonym of a word with which a document has been indexed in the query, the document may not be retrieved.

Query expansion is one of the methods that can be used to alleviate this problem [8,15,6]. The expansion terms can be taken from a thesaurus that contains a list of pre-determined terms (called a controlled vocabulary) and their semantic relations. The grammatical forms and spellings of the controlled vocabulary follow certain specifications. For instance, if a term has multiple meanings depending on the context, each meaning is explicitly identified; if multiple terms share the same meaning, one of them is selected as the preferred term and the others, called non-preferred terms, are listed as synonyms of that term. The controlled vocabulary is then organized into a hierarchical structure with semantic relationships defined among them, i.e.

© 2005 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by a contractor or affiliate of the U.S. government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

SAC'06, April, 23-27, 2006, Dijon, France.
Copyright 2006 ACM 1-59593-108-2/06/0004...\$5.00.

broader term, narrower term, synonym, etc.

Thesauri have been created in many application domains. For example, *Roget's Thesaurus* [4] and *WordNet* [14] are two well known thesauri for the general English language; the *NASA Thesaurus* is a tool that can be used to index and retrieve materials in the fields of aerospace, earth science, natural space sciences, and biological sciences [16]; the *ERIC* (The Educational Resources Information Center) *Thesaurus* provides public access to education-related documents, journal articles, and books [5]; and the *Astronomy Thesaurus* standardizes astronomical terms in order to assist astronomers and librarians in cataloging, indexing, searching, and sharing information [21].

The proliferation of biomedical research and the public demand of e-healthcare systems have stimulated the development of biomedical thesauri. Several examples include Medical Subject Headings [13], Unified Medical Language System [22], and Systematized Nomenclature of Medicine [20]. While the existing medical thesauri have helped immensely in information categorization, indexing, and retrieval, two major problems remain:

- Their designs do not follow any international or national thesaurus standard and therefore, result in poor interoperability and reusability.
- They do not provide information regarding the semantic similarities among terms and thus the users are required to possess precise knowledge of the controlled vocabulary in order to make effective use of the thesaurus.

The goal of this work is to alleviate the aforementioned problems and improve the effectiveness of biomedical information retrieval. More specifically, our research objectives and major contributions include the following:

- Establish a medical thesaurus (MEDTHES) that follows the ANSI (American National Standard Institute) Standard for thesaurus design and thus, improve the interoperability with other thesauri that also follow the national standards.
- Provide semantic similarity measures to assist users in performing imprecise queries in which the query term is different than the indexing term of a document.
- Include synonyms of medical terms from a general English thesaurus, WordNet, in order to ease the use of MEDTHES for non-medical professionals.
- Incorporate MEDTHES into an existing mobile agent-based information search engine MAMDAS (Mobile Agent-based Mobile Data Access System) [10] which utilizes an online

thesaurus to improve the precision and recall of information retrieval.

The rest of this paper is organized as follows: Section 2 discusses the background and related work. Section 3 provides the design and implementation details of MEDTHES. Section 4 demonstrates the integration of MEDTHES and MAMDAS. Finally, section 5 concludes our work and points out some future research directions.

2. BACKGROUND AND RELATED WORK

In this section, we *i)* briefly outline the ANSI/NISO standard for thesauri construction; *ii)* describe the two thesauri that have served as the foundation of our MEDTHES, MeSH and WordNet; and *iii)* introduce the concept of semantic similarity.

2.1 The ANSI/NISO Z39.19-2003 Standard

The ANSI/NISO Z39.19-2003 standard [17], entitled *Guidelines for the Construction, Format, and Management of Monolingual Thesauri Abstract*, was developed by the National Information Standards Organization (NISO) and approved by the American National Standard Institute (ANSI). It provides guidelines for the design and use of thesauri, including rules for term selection, thesaurus structure, relation definitions, and thesaurus maintenance [17]. Three types of semantic relationships between terms are distinguished in this standard: equivalence, hierarchical, and related. The equivalence relation establishes the link between a non-preferred term and its corresponding preferred term; the hierarchical relationship provides links between terms that reflect general concepts (broader terms) and those that represent more specific information (narrower terms); and the related relationship exists among terms that have similar meanings or are often used in the same context but do not have hierarchical relationships. The design of MEDTHES follows this standard.

2.2 MeSH

The Medical Subject Headings (MeSH) [13] thesaurus is the standardized vocabulary developed by the National Library of Medicine for indexing, cataloging, and searching the medical literature. Currently, it contains approximately 22,000 terms (called descriptors) that describe the biomedical concepts used in health-related databases such as MEDLINE [12], which is an online bibliographic database for medicine, nursing, health services, etc. All descriptors in MeSH are organized into 15 categories. Each category is then further divided into more specific subcategories. Within each category, descriptors are organized in a hierarchical fashion of up to eleven levels. In addition to the hierarchical structure, MeSH uses “Entry Term” or “See” references to indicate semantic relations such as synonyms, near-synonyms, and related concepts of some terms.

Although MeSH is comprehensive and well maintained, it has several drawbacks. First, the synonymous relationship is not clearly listed and not differentiated from the related term relation in MeSH. Second, many descriptors do not have corresponding “Entry” vocabularies listed, which means that synonyms cannot be found for many terms in MeSH. Third, the design of MeSH does not follow the ANSI thesaurus standard, which results in problems of interoperability and reusability.

2.3 WordNet

WordNet is an online thesaurus that models the lexical knowledge of the English language [1]. It organizes English nouns, verbs, adjectives, and adverbs into synonym sets, called synsets. In other words, a synset is a list of synonymous terms. Each term in WordNet may have one or more meanings, and each meaning has a synset. Different synsets are connected through hierarchical relationships.

In sum, WordNet is comprehensive and designed with the goal to include every English word, and it makes a number of fine-grained distinctions among word meanings. Thus, we decided to take advantage of the well-defined synonyms of WordNet and use them to complement the MeSH thesaurus.

2.3 Semantic Similarity

Synonyms and related terms obtained from a thesaurus are often used in query expansion for the purpose of improving the effectiveness of information retrieval [19]. However, in order to improve the quality of document ranking, a more fine-grained measure is needed to describe the *degree of semantic similarity*, or more generally, the *relatedness*, between two lexically expressed concepts [3]. Naturally, *semantic distance* is the inverse of semantic similarity. For example, the semantic distance between synonyms can be defined as 0, and that between antonyms can be defined as infinity.

If a thesaurus provides functions that calculate the semantic similarity between terms, the users can perform fine-tuned queries by specifying the scope of the search using the semantic distance between the keyword and the search results. The user can indicate how closely the returned terms should be related to the keyword (searched term) by selecting preferred semantic distance values.

Two main categories of algorithms for computing the semantic distance between terms organized in a hierarchical structure (e.g. WordNet) have been proposed in the literature: distance-based approaches and information content-based approaches. The general idea behind the distance-based algorithms [9,18,23] is to find the shortest path between two terms in terms of number of edges and then, translate this distance into semantic distance. Information content-based approaches [11,18] are inspired by the perception that pairs of words which share many common contexts are semantically related. Thus, the basic idea of these methods is to quantify the frequency of the co-occurrences of words within various contexts.

In order to eliminate potential bias introduced by context selection, we chose to implement three distance-based algorithms in the MEDTHES prototype: the Edge Counting [18], Leacock & Chodorow [9], and Wu & Palmer [23] algorithms.

(1) The Edge Counting Algorithm

In the Edge Counting algorithm, the semantic distance is defined as the number of edges (nodes) along the shortest path between any two terms.

(2) The Leacock & Chodorow Algorithm

The relatedness measure proposed by Leacock & Chodorow also relies on the shortest path between two terms, t_1 and t_2 . The relatedness between two terms t_1 and t_2 is calculated as follows:

$$relatedness(t_1, t_2) = -\log \frac{len(t_1, t_2)}{2D} \quad (1)$$

where $relatedness(t_1, t_2)$ is the similarity of terms t_1 and t_2 ; $len(t_1, t_2)$ is the length of shortest path between two terms (using edge counting); D is the maximum depth of the structure. Semantic distance is the inverse of $relatedness(t_1, t_2)$, i.e.,

$$\frac{1}{relatedness(t_1, t_2)}$$

(3) The Wu & Palmer Algorithm

The Wu & Palmer algorithm uses the terminology *score* to define how two terms are related to each other. It measures the score by considering the depth of the two terms t_1 and t_2 in the tree structure, along with the depth of the LCA (Least Common Ancestor). The formula used to calculate the score is shown below.

$$score(t_1, t_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \quad (2)$$

Where N_1 is the length of shortest path from t_1 to LCA; N_2 is the length of shortest path from t_2 to LCA; N_3 is the length of shortest path from LCA to the root. The range of $relatedness$ is $0 < score(t_1, t_2) \leq 1$. The $score(t_1, t_2)$ is 1 if t_1 and t_2 are the same. Semantic distance is the inverse of $score(t_1, t_2)$, i.e.,

$$\frac{1}{score(t_1, t_2)}$$

3. MEDTHES

The taxonomy defined in MeSH is the foundation of MEDTHES. However, several major changes to MeSH have been made: *i*) the semantic relations of MeSH were reconstructed according to the ANSI standard; *ii*) the synonym set of each entry in MeSH was enriched by synonyms extracted from WordNet; *iii*) three algorithms of semantic distance calculation were implemented in order to provide users fine-grained control over the query results.

MEDTHES adopts the three standard relationships suggested by the ANSI/NISO standard for thesaurus construction: equivalence relationship, hierarchical relationship, and associative relationship. Terms in MeSH are arranged hierarchically in a tree structure, top down from general to more specific. The broader term (BT) and narrower term (NT) relations can be easily extracted from this hierarchical structure. A program, "MeSHFileParser", is developed to automatically parse such information.

In MeSH, synonyms and related terms (RT) are not clearly differentiated. The definitions of synonyms are neither accurate nor complete. As a result, MeSH is not suitable to be used directly to obtain synonyms. Since the well defined synonyms are one of the major strengths of WordNet, it was used as a reference when adding synonyms to MEDTHES. A term is selected from the synonym set as the "preferred term", which means that term is used for (Used For, UF in short) indexing other terms in the same set. The reverse relation is "Use", which means that if a keyword

is a non-preferred term, it is substituted with the preferred term before searching.

A term may exist in one or more categories in MeSH. In order to establish a link between a term and the category it belongs to, an additional relationship, "Subject Categories (SC)", was also defined. Table 1 summarizes the relationships used in MEDTHES.

Table 1. Relationship Definitions in MEDTHES

ANSI/NISO Relationship	MEDTHES Representation	Abbreviation
Equivalence	Use	USE
	Used For	UF
Hierarchical	Broader Term	BT
	Narrower Term	NT
Associative	Related Term	RT
	Subject Category	SC

4. APPLICATION OF MEDTHES- A CASE STUDY

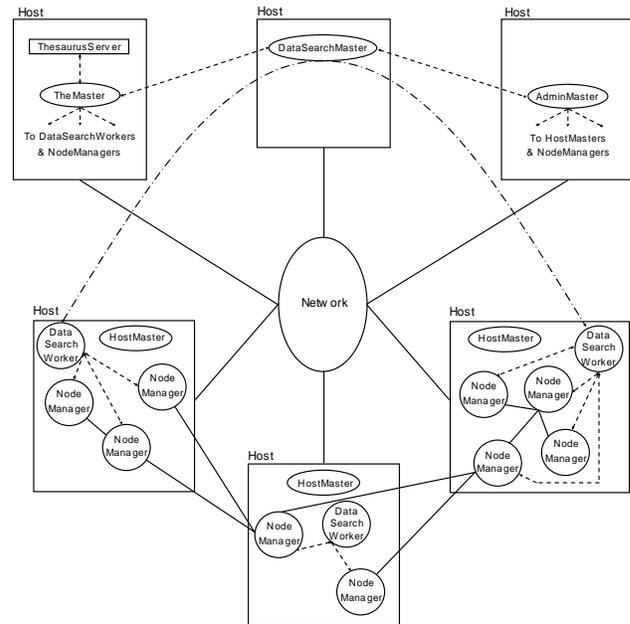


Figure 1. The Overall Architecture of MAMDAS.

In this section, we integrate MEDTHES with an existing mobile agent-based information search engine, MAMDAS (Mobile Agent-based Mobile Data Access System) [10]. MAMDAS was established based on the mobile agent technology and a hierarchical multi-database organization model called the summary schemas model [2]. It utilizes a thesaurus to resolve semantic ambiguities of queries. Any thesaurus that follows the ANSI standard can be used as a plug-in of the MAMDAS search engine. Figure 1 illustrates the overall architecture of MAMDAS.

MAMDAS consists of five major components: the administrator, the thesaurus, the node, the host, and the user. Each component is represented by one or more agents, stationary or mobile. Computers connected through the networks are called hosts. Each host can maintain several data sources each of which is called a node. The Administrator (AdminMaster Agent) is in charge of the

initial setup of the system, and the thesaurus server (ThesMaster Agent) responds to all inquiries regarding the semantic distance between terms.

When MAMDAS is used in different application domains, the only modification required is to change the thesaurus in a plug-and-play fashion. In other words, we should choose a thesaurus plug-in that is the most appropriate for that domain. In our study, we integrated MEDTHES with MAMDAS in order to resolve queries in the biomedical domain. Figure 2 shows an example of the MAMDAS data search GUI.

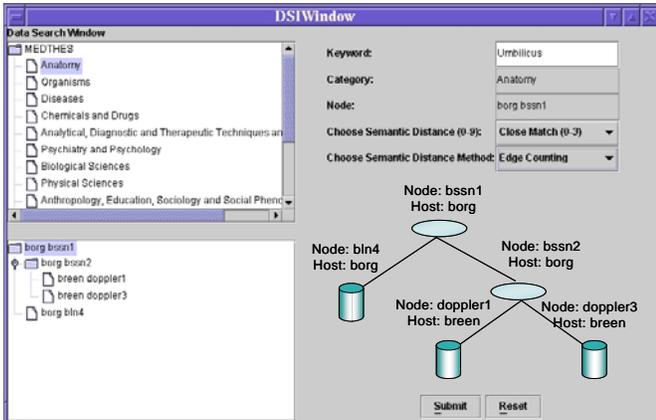


Figure 2. Data Search GUI.

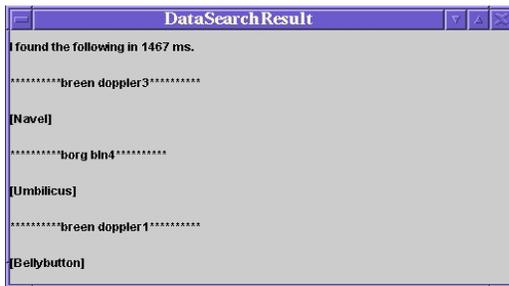


Figure 3. Search Results

The top left window shows the categories defined by MEDTHES, and the lower left window contains the current multi-database configuration. The tree structure shown on the right hand side is an equivalent representation of the multi-database hierarchy shown on the left. In this example, the search keyword is “umbilicus” in the “anatomy” category, and the search starts at the root of the multi-database hierarchy. In other words, every node (data source) in the hierarchy should be searched. The user wishes the search engine to return all terms that have a semantic distance of less than or equal to 3 with respect to the keyword, according to the edge counting algorithm. The search result is shown in Figure 3. Three terms, “navel”, “umbilicus”, and “belly button”, were found from three different data sources. Although terms “navel” and “bellybutton” are not the exact lexical match of “umbilicus”, they are returned because they are semantically closely related to “umbilicus”.

Depending on the way that semantic similarity is defined by an algorithm, the same numerical value of semantic distance often means different degrees of semantic closeness in different algorithms. In order to assist users in choosing meaningful values

of semantic distance, we conducted a set of experiments and determined the correlation of the three semantic distance calculation algorithms realized by MEDTHES. Figure 4 shows the data flow chart of the experiments.

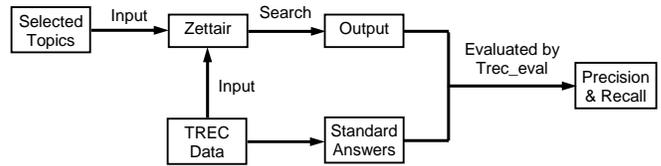


Figure 4. The Data Flow Chart of Experiments.

Queries from the OHSUMED test collection [HBLD94] were expanded with terms returned by MEDTHES for a particular algorithm and a chosen semantic distance value, starting from 0. These expanded queries were then fed into a data search engine called Zettair [25]. Zettair retrieved documents from the TREC9 data set, and the output was evaluated by the trec_eval tool against the standard answers. Then, the semantic distance was increased by a step (the step value differs from one algorithm to another), and the experiment was repeated.

With the *Edge Counting* (EC) algorithm, the semantic distance value is set to 0.0 at the beginning and increased by 1.0 each time. When using the *Leacock & Chodorow* (LCH) and the *Wu & Palmer* (WP) methods, the semantic distance is set to 0.0 initially, and then increased by 0.1 each time. This procedure continued until the average precision drops below a threshold (3%). We grouped similar results into three plots (Figure 5). The value ranges of each algorithm that would generate comparable results are summarized in Table 2. For example, query expansion using the terms returned by MEDTHES with the EC algorithm and semantic distance within the range of [0, 1.0] and with the LCH algorithm and semantic distance within the range of [0, 0.2] would result in similar documents being retrieved.

5. CONCLUSION

This study addressed the issue of semantic-based information retrieval of biomedical data by using a medical thesaurus, MEDTHES. MEDTHES was constructed based on MeSH and WordNet. It follows the specifications defined by the ANSI/NISO standard. This feature significantly improves the interoperability between MEDTHES and other standard thesauri. Moreover, we incorporated three well-known semantic distance calculation algorithms into MEDTHES in order to support imprecise queries. As a demonstration, we integrated it with an existing thesaurus-based data search engine and illustrated the semantic-based imprecise query function by an example. We further quantitatively studied the correlation among the three semantic distance calculation algorithms and recommended the suitable ranges of semantic distance values for each of them. It is envisioned that MEDTHES can be widely used to improve the effectiveness of biomedical information retrieval and to be applied to e-health management systems.

As part of our future work, we plan to implement several information content-based semantic distance calculation algorithms in MEDTHES. This addition will be beneficial for users who work with a particular data source, because the calculation can be tailored to reflect the content of data.

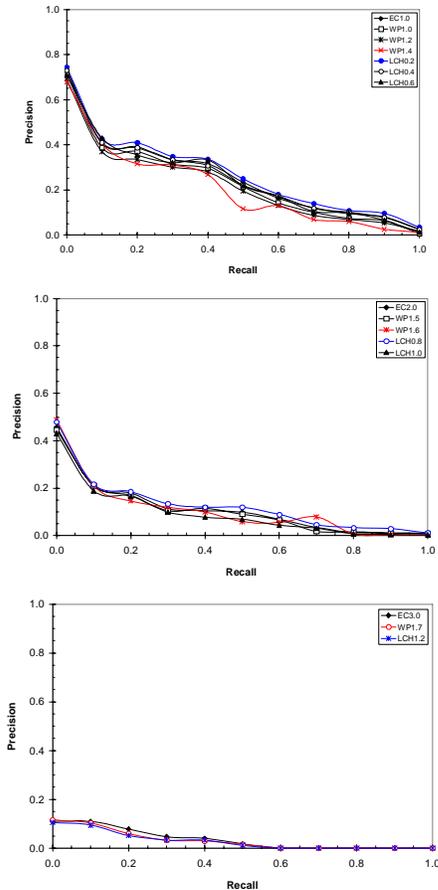


Figure 5. Results Comparison.

Table 2. Correlation among the semantic distance algorithms

Edge Counting	Leacock & Chodorow	Wu & Palmer
$D = [0, 1.0]$	$D = [0, 0.2]$	$D = [0, 1.0]$
$D = (1.0, 2.0]$	$D = (0.2, 0.6]$	$D = (1.0, 1.6]$
$D = (2.0, 3.0]$	$D = (0.6, 1.0]$	$D = (1.6, 1.7]$

ACKNOWLEDGEMENTS

The Office of Naval Research under contract N00014-02-1-0282 and the National Science Foundation under contract IIS-0324835 in part have supported this work.

6. REFERENCES

[1] Bates, M.J., Subject access in online catalogs - A design-model. *Journal of the American Society for Information Science*, 37(6): 357-376.

[2] Bright, M.W., Hurson, A.R., and Pakzad, S. Automated resolution of semantic heterogeneity in multidatabases. *ACM Transactions on Database Systems*, 19(2), 212--253, 1994.

[3] Budanitsky, A. and Hirst G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of Workshop on WordNet and Other Lexical Resources*, Second meeting of the North

American Chapter of the Association for Computational Linguistics, Pittsburgh, June 2001.

[4] Chapman, R. (revised) *Roget's International Thesaurus (Fourth Edition)*. New York: Thomas Y. Crowell Company, 1977.

[5] ERIC Thesaurus. <http://www.csa.com/csa/HelpV5/suppl/ericthes.shtml>

[6] Hersh, W., Burksy, C., Leone, T.J., and Hickam, D. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of ACM SIGIR '94*, 1994, 192-201.

[8] Hersh, W., Price, S., and Donohoe, L. Assessing thesaurus-based query expansion using the UMLS metathesaurus. In *Proceedings of the AMIA Symposium*, 2000, 344-8.

[9] Leacock, C. and Chodorow M. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C. (Ed.), *WordNet: A Lexical Reference System and its Application*, MIT Press, Cambridge, MA.

[10] Jiao, Y. and Hurson A.R. Application of mobile agents in mobile data access systems - A prototype. *Journal of Database Management*, 15(4):1-24.

[11] Jiang, J. and Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, 1997.

[12] MEDLINE. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

[13] Medical Subject Headings. <http://www.nlm.nih.gov/mesh>

[14] Miller, G.A. (Ed.) Five papers on WordNet. *International Journal of Lexicology*, 3(4).

[15] Mandala, R., Tokunaga, T., Tanaka, H., Okumura, A., and Satoh, K. Ad hoc retrieval experiments using WordNet and automatically constructed thesauri. In *Proceedings of the 7th Text Retrieval Conference*, 1999, 475-480.

[16] NASA Thesaurus. <http://www.sti.nasa.gov/thesfrm1.htm>

[17] National Information Standards Institute American National Standard Guidelines for the Construction, Format, and Management of Monolingual Thesauri. Bethesda, MD: NISO Press, 1994.

[18] Rada, R., Mili, H., Bicknell, E., and Blettner, M. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17-30.

[19] Shiri A.A., Revie C., Chowdhury G. Thesaurus-assisted search term selection and query expansion: a review of user-centered studies. *Knowl Organ*, 29(1): 1-19.

[20] Systematized Nomenclature of Medicine. <http://www.snomed.org/>

[21] The Astronomy Thesaurus. <http://msowww.anu.edu.au/library/thesaurus>

[22] Unified Medical Language System. <http://www.nlm.nih.gov/research/umls/>

[23] Wu, Z. and Palmer, M. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

[24] Zettair Search Engine, <http://www.seg.rmit.edu.au/zettair>