Natural Language Query in the Biochemistry and Molecular Biology Domains Based on Cognition Search™

Elizabeth J. Goldsmith[†||], Saurabh Mendiratta[†], Radha Akella[†,] and Kathleen Dahlgren[||§]
[†]Department of Biochemistry, The University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816. [§] Cognition Technologies, Inc, 6133 Bristol Parkway, Culver City, CA 90230.

Abstract

Motivation: With the increasing volume of scientific papers and heterogeneous nomenclature in the biomedical literature, it is apparent that an improvement over standard pattern matching available in existing search engines is required. Cognition Search Information Retrieval (CSIR) is a natural language processing (NLP) technology that possesses a large dictionary (lexicon) and large semantic databases, such that search can be based on meaning. Encoded synonymy, ontological relationships, phrases, and seeds for word sense disambiguation offer significant improvement over pattern matching. Thus, the CSIR has the right architecture to form the basis for a scientific search engine.

Result: Here we have augmented CSIR to improve access to the MEDLINE database of scientific abstracts. New biochemical, molecular biological and medical language and acronyms were introduced from curated web-based sources. The resulting system was used to interpret MEDLINE abstracts. Meaning-based search of MEDLINE abstracts yields high precision (estimated at >90%), and high recall (estimated at >90%), where synonym, ontology, phrases and sense seeds have been encoded. The present implementation can be found at http://MEDLINE.cognition.com.

Contact:
Elizabeth.goldsmith@UTsouthwestern.edu
Kathleen.dahlgren@cognition.com

Introduction

The goal in search is to create software that finds all of the desired information (full recall) without introducing undesired information (high precision), despite the inherent heterogeneity of language usage. Some of the major problems that can be handled computationally are synonym relationships, ontological relationships, morphology, sense selection, and phrase recognition. Each of these requires databases specifically describing relationships. Several laboratories and companies offer technologies for recognizing nouns and phrases (named entity recognition) in biomedical terminology (1-5). A few free sources are visible on the web e.g. Google Scholar (http://scholar.google.com/), Highwire press (http://highwire.stanford.edu/lists/freeart.dtl) whereas other relatively commercial sources of this information is present at Scopus (http://www.scopus.com/scopus/home.url), Ovid (http://www.ovid.com/site/index.jsp), and Infotrieve (http://www4.infotrieve.com/newMEDLINE/search.asp). Many of these technologies use aspects of linguistic processing (4, 6-15) such as synonymy and ontology. These features improve recall. The problem of disambiguating abbreviations has also been addressed (16, 17), which improves precision. CSIR is unique in that the word sense selection occurs in an offline indexing process. This offline process improves both sense selection and mapping of synonyms at query time. CSIR is also unique in that words of ordinary English have been introduced with substantial curation.

Architecture of CSIR™

CSIR™ is an NLP technology that possesses a broad semantic map of English based on word senses, synonyms (8) and hypernyms (higher nodes in an ontology) (9). It also possesses a database of sense seeds which are used to identify a particular word meaning in text. The CSIR Indexer uses its NLP component to build a cognitive model of the text in which all of the concepts (word meanings) of a document are indexed in an offline job. The indexer relies on the dictionary, semantic map, morphological and syntactic tags, word seeds and database of synonyms and ontological relationships (Fig.1). During indexing, linguistic modules determine word and phrase boundaries, recover stem forms from variants (such as "catch" from "caught", or "phosphorylate" from "dephosphorylation", parse phrases such as "LIM-domain-interacting-RING-finger-protein", disambiguate words and abbreviations such as "base" and "CAP", and place word senses (concepts) in the semantic map. Synonyms and hypernyms are selected which can be used for semantic reasoning during search. At search time, CSIR interprets the query for meaning, and searches for the meaning of the query in the concept index. The patented meaning-based architecture and methods have been described previously (18-20).

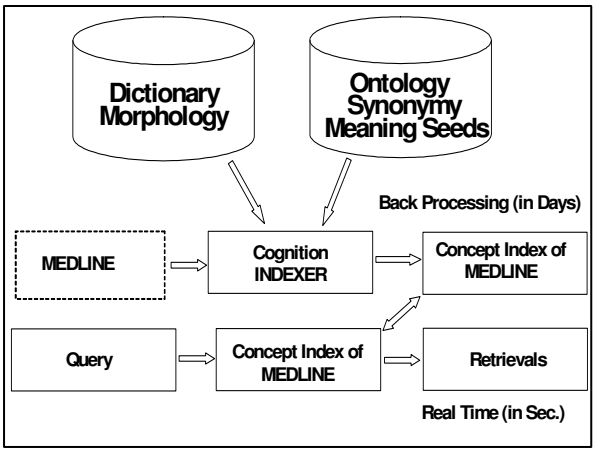

Figure 1: Architecture of CSIR

Improvements since the original description of this work, includes sense disambiguation using sense seeds (10), phrase parsing (21), data compression and speed upgrades (22). The morphology and phrase identification components were built in-house (patent pending). The software also uses simple algorithms for phrasal parsing and concept clustering to improve document relevancy (precision). Demonstrations of CSIR are available at http://medline.cognition.com and http://wikipedia.cognition.com. The search engine should be used asking a straightforward question that might be answered in MEDLINE, such as "Oxidative stress in plants," "spectroscopy of amidohydrolases," or "Depression in aging." Retrieval time on the 17 million MEDLINE abstracts is sub-second on Xeon Dual Core 3.0 GHz computers with 1 GB of RAM.

Methods

Lexicon and Concept Thesaurus

Databases encoding desired biomedical terminology were identified. These databases were crawled using a Python crawler, with the fields being captured tailored to the specific database. For example, fields were identified as base terms or phrases, synonyms or ontological classes, and were extracted into a database which was familiar to an automated lexical acquisition program. The potential data to be entered was curated. Vocabulary unknown to the existing CSIR was checked for frequency in the MEDLINE, so that the most frequent unknown words could be added along with definitions by curators. Improper synonyms could occur with the automated lexical data entry. Additional curation checked suspiciously large synonym classes. Words available online without ontological attachments were also a common problem. Trial ontological attachments were formed computationally using the longest strings in a synonym class, which were then curated. Sense contexts for acronyms were garnered only from text containing the spell-outs to improve the accuracy of seed generation (11).

Ontology

To augment the ontology for biochemistry and molecular biology, a top ontology was constructed by hand, based upon our own domain knowledge, and just using a very simple text discrimination of nodes and leaves. Nodes are given a unique word form, but mapped to synonymous ordinary words and phrases. Websites of curated biomedical terminology were crawled for their ontological attachments. Again, specialized programs were written to crawl each website. The ontological attachments were then mapped to our top ontology by hand.

Precision and Relative Recall Test of CSIR vs PubMed

Queries were formulated in formats consistent with either Cognition or PubMed (as a question for Cognition and as Boolians for PubMed). The total number of CSIR retrievals was recorded, and the relevance evaluated for the top 10 and top 20 retrievals, as assessed by the UT Southwestern team. Only the top 10 and 20 retrievals were evaluated due limitations of time (a standard practice in text retrieval evaluations)(23).The same queries were posed to PubMed for comparison (in a Boolean format: "genetic" AND "interaction" AND "BCL2"). To make the evaluation manageable, we used the "relative recall" technique, wherein full recall is estimated as the greatest number of retrievals achieved by either search engine. For example, one query was "genetic correlates of alcoholism". Of the first twenty CSIR retrievals, 16 were relevant. Thus CSIR's precision was 16/20 or 0.8. The number of retrievals for CSIR was 1,436. To extrapolate the good retrievals, we multiplied the precision ratio 0.8 times 1,436 to yield extrapolated recall of 1,149. The queries used here can be seen on the Goldsmith webpage(http://hhmi.swmed.edu/Labs/bg/Cognition).

Results

Scale and Scope

At the initiation of this project, a lexical evaluation of MEDLINE showed that CSIR was missing 66,000 words. Estimates of the total number of biomedical terms is over a million, a much larger number, mostly

phrases (12). Here we added about 85,000 protein names, 35,000 chemical names, ontology for biochemistry and molecular biology possessing 2,400 nodes, and over 30,000 biomedical synonym classes. Table 1 represents the detailed description of the entire Cognition semantic map at present with ongoing lexical augmentations.

Cognition's Semantic Map (Based on Computational Linguistic Science)	
Word Stems	**506,000 Word stems**
Words and Phrases	**536,000 Word senses or concepts**
Meanings in context	**4,000,000 Semantic contexts**
Different Word Meanings	**17,000 Ambiguous word definitions**
Complex Word Series Meanings	**191,000 Phrases**
Ontology or Taxonomy	**7,000 Nodes**
Synonyms	**76,000 Thesaural concept groups**

Table 1: Cognition Dictionary by numbers

Ontology for Biochemistry and Molecular Biology

Ontologies need to be established at the desired granularity. We defined a top ontology for the biochemical and molecular biology domain that serves as a basis for capturing finer, more desired ontological nodes. Our top ontology, primarily for molecular entities, resembles SEMEDA (9), or TAMBIS (13). We primarily added an ontology for protein and gene names, but also included some ontology of drug names, biological processes, and laboratory procedures. An intermediate level of protein and gene name ontology was inspired by that in the Alliance for Cell Signalling (AfCS, eg. "binding protein," "g-protein", transcription-factors, etc) , and by an ontology of terms in the Human Genome Nomenclature Committee (HGNC) that categorizes proteins and genes (Table 2). Work predating the present study had already defined ontologies of human anatomy, diseases, medical treatments and a rudimentary tree-of-life.

Table 2A: Ontology of Biochemical and Molecular Biology

A. **Piece of the Top Ontology for Biochemistry**
Macromolecule-node
Protein-stuff
antibody
binding protein
enzyme
Nucleic-acid
Laboratory-procedure
Electrophoresis
Spectroscopy
B. **Ontology for protein kinases**
protein-kinases
protein-histidine-kinases
serine-threonine-kinases
AGC-kinases
STE-kinase
Tyrosine-kinase
ACK-kinase
EGFR-kinase
Tyrosine-Like-Kinase
MLK-kinase
RAF-kinase

Table 2B: Finer grained Protein Kinases ontology.

Introducing new language from existing databases

Web-based sources of biomedical terminology were: acronyms from http://medstract.med.tufts.edu (8), the molecules and genes defined by the AfCS database (24), the Human Genome Nomenclature Consortium (25), the UMLS Metathesaurus and the International Union of Pure and Applied Chemistry (IUPAC) enzyme names. The acronym database and UMLS were selected for their wide coverage. We selected the AfCS and HGNC databases because the curators captured natural word usage, and have encoded a gross molecular ontology as well as some synonymy. The IUPAC database was chosen because the ontology has been constructed carefully. Some of the larger databases were avoided because we noted numerous errors and short and redundant acronyms, requiring too much curation. Since some acronyms were added to the semantic map in earlier projects, a challenge was to add only new senses (26). We curated 16,256 acronyms from the Tufts database, removing rarely used acronyms (usage cutoff of 20), and very redundant acronyms. This resulted in 15,657 acronyms with 16,858 total meanings.

We introduced vocabulary from the UMLS Metathesaurus. We built a map from the Metathesaurus ontology to our existing ontology, and then introduced the UMLS vocabulary into the lexicon automatically. Multi-sense words were inspected by a scientist. Synonyms, with the

appropriate senses, were introduced to the Concept Thesaurus automatically.
This database includes both nouns and verbs covering biological sciences and medicine, amounting to 88,423 word senses, and 76,816 synonyms.
We then obtained additional word senses, all nouns, from the Alliance for Cell Signaling (www.alliance.org) (24). This source is current, curated and offers ontological entries, giving 15,661 new or improved word senses. The adoption of this vocabulary was accomplished through a combination of automated tasks and expert curation. Duplicates were curated. Unknown vocabulary was then added to the semantic map automatically, including ontological attachments and synonyms. Data from the HGNC (www.genenames.org) (25) has also been partially introduced. About 30 ontologies of protein families in HGNC have been imported, including AKAPs, bcl, BRCA, channel proteins, P450s, tubulins, ubiquitin ligases, phosphatases, TNF-receptors, histones, SMADs, and so on. We also introduced the IUPAC enzyme names and EC numbers, over 6,000 names. A difficulty with this augmentation is the lack of natural language usage and lack of synonymy.

Missing words by frequency

The numbers of words or tokens present in MEDLINE by missing in the Cognition dictionary were counted. Unknown works with frequency greater than 100 were curated; there were only 800 of these. The remainder gave the frequency distribution shown in Fig. 2. As can be seen, capturing the words with frequency greater than 20 is desirable. At this writing, we have introduced most words with frequency greater than 50.

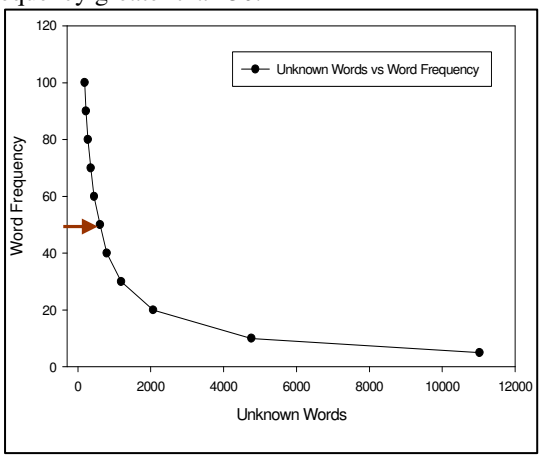

Figure 2 shows the Coverage of MEDLINE

MEDLINE abstracts were also searched to find verbs, which were curated to find words (such as express, silence, translocate, spin, bait, prey) that have domain specific-meanings. This project has led to 225 new word senses. The added verb definitions contribute to improved precision through word sense disambiguation, and will be useful when full sentence parsing is included in CSIR (14).

Precision and Recall Test

Fifty queries for MEDLINE were formulated as simple questions in the areas of biochemistry, molecular biology and medicine. The UT Southwestern team tabulated the relevance of the retrievals in http://MEDLINE.cognition.com and compared them with those of PubMed (http://pubmed.com) retrievals. As described in the methods, queries were formatted to conform to the two different search engines, and the relative recall method was used for evaluation. As can be seen Cognition did better by both precision and recall measures (Table 3). The reader, however, is perhaps the best judge of the relative performance of the search engines.

Table 3 Precision and Recall: Comparison between Cognition and Pubmed.

Cognition vs MEDLINE search	Cognition good/20	Cognition bad/20	Total	Pubmed good/20	Pubmed bad/20	Total
Genetic correlates of alcoholism	18	2	1436	6	14	44
DNA repair and aging	17	3	1220	11	9	1265
Drugs for fibromyalgia	17	3	1484	9	11	220
Genetic interactions of BCL2	18	2	876	8	11	19
Oxidative stress in plants	18	2	3122	9	11	3197
spectroscopy of amidohydrolases	17	3	861	7	13	1142
Benzene induced neuropathy	18	2	220	6	1	7
Birth defects from glycol ether	16	4	20	13	7	61
Depression in aging	19	1	13381	7	13	3658
Symptoms of type II diabetes mellitus	18	2	241	7	13	24704
Menopause and depression	18	2	696	11	9	1146
Treatment for bronchiectasis	18	2	2163	6	14	3207
OCD and anorexia	20	0	176	14	6	247
Proteolysis in SARS virus entry	4	0	4	2	0	2
Total	280	60	18433	125	127	34080
	Cognition			MEDLINE		
Precision	0.90			0.50		
Recall (*Assume total recall is the total of the cognition retrievals)	0.99			0.54		

Bootstrapping ontological attachments

Most of the vocabulary derived from the acronym database and the UMLS had poor (very general) ontological attachments (eg, “amino-acid”). About 80,000 of 136,000 protein names were poorly attached. Attachments of well-classified words were spread to their synonyms resulting in 20,000 better attachments. A bootstrapping method took substrings as triggers; for example, “helix-loop-helix” as a substring of “transcription-factor-15-basic-helix-loop-helix” suggests an attachment to the node “helix-loop-helix.” This attachment was then assigned to the synonym “bHLH-EC2-protein”.

Discussion

We think that the natural language approach of CSIR has an important role in future access to textual information in the biomedical domain. This effort is our first pass at introducing biochemical and molecular biology terms into the CSIR lexicon. Other sources of new words will come from tracking user queries, evaluation of MEDLINE, and other curated databases. CSIR works equally well on full-text as on abstracts. It can be used to read full-length papers and other databases containing text. This work contributes to precise interpretation of biomedical texts for purposes of search (1, 3, 27), research (4) and data mining (2, 28). Cognition Search has features in common with other NLP software (6, 8, 29), but unique to Cognition Search are its very large hand built lexical resources with synonymy, ontology, built with all linguistic features encoded, and the linguistically-based morphology, sense disambiguation and parsing that draw upon these lexical resources. The present work relied upon the existing hand-built lexical resources, bootstrapping them for semi-automated lexical acquisition in the biomedical domain.

Uses and Applications of CSIR

It is useful to review which linguistic processes produce these improved results. Morphology improves recall, so that the user can state a query term in one of its morphological variants, and CSIR automatically finds all other forms, as in phosphorylate and phosphorylation. Synonymy improves recall because one member of a synonym class retrieves documents with any of its members, as in “CD116,” "GMHCFS receptor alpha subunit," etc. Ontological reasoning improves recall as the software reasons down from higher-level concepts to lower-level concepts. For example, you can query "what MAP kinase phosphorylates ATF2" and get documents with “ERK” and “p38” which are kinds of MAP kinases. Sense disambiguation improves precision because only the documents that contain the query terms in the meanings intended by the user are retrieved. Phrase parsing improves both precision and recall. It improves precision by avoiding retrievals that happen to contain parts of a phrase in various positions, but not as the phrase. So "RNA", "binding" and "protein" might all appear in an abstract that has nothing to do with RNA binding proteins. It improves recall because it enables the mapping of synonym relations between phrases, and between phrases and acronyms, as in "TUBB" and "beta-tubulin". Biomedical language also possesses ontological relationships for proteins, genes, the Tree-of-Life animals, diseases, etc. CSIR includes the function of downward reasoning in ontologies.

Areas for improvement

It will be relatively easy to address missing terms by frequency. We will use the methods of Tsuruoka (30) for future term recognition, synonymy expansion and evaluation of coverage. Automatic discovery of additional normalization rules (mapping different spelling variants to each other), as in Wellner and Yoshimasa (31, 32) would be a further step. Efforts directed toward database integration may provide useful definitions, synonymy and ontology in molecular biology (15). We also plan to introduce additional parsing functions (29), (14) which should improve the precision of Cognition Search.

Acknowledgements

We thank Ron Taussig for pointing out the Alliance for Cell Signaling website and other discussions. UMLS resources licensed (number 21817A334). The work in E. J. Goldsmith’s group was carried out under contract with Cognition, Technologies, Inc.

References

1. Vanhecke TE, Barnes, M.A., Zimmerman, J., Shoichet, S. . PubMed vs. HighWire Press: A head-to-head comparison of tow medical literature search engines. Computers in Biology and Medicine. 2007; 37:1252-8.
2. Divoli A, Attwood, T.K. . . “BiolE sentences - Extracting informative sentences from the biomedical literature.” Bioinformatics 2005; 21(9):2138-9.
3. Doms A, Schroeder, M. “GoPubMed: exploring PubMed with the Gene Ontology”. . Nucleic Acids Research 2005;33.

4. Fontelo P, Liu, F., Ackerman, M.. "askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. BMC Medical Informatics and Decision-Making 5:5. 2005.
5. Matthew E. Falagas*, 1, Eleni I. Pitsouni*, George A. Malietzis* and Georgios Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. The FASEB Journal. 2008; 22:338-42.
6. Ohta T, Tsuruoka, Y, Takeuchi, J, Jin-Dong Kim, JD, Miyao, Y, Yakushiji, A, Yoshida, K, Tateisi, Y, Ninomiya, T, Masuda, K, Hara, T, Tsujii, J., editor. An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. . Proceedings of the COLING/ACL; 2006.
7. Tsai R, Wu, S-H., Hsu, W-L., editor. Exploitation of Linguistic Features Using a CRF-Based Biomedical Named Entity Recognizer. ACL Workshop on Linking Biological Literature; 2005.
8. Wren JD, Chang JT, Pustejovsky J, Adar E, Garner HR, Altman RB. Biomedical term mapping databases. Nucleic Acids Res. 2005 Jan 1; 33(Database issue):D289-93.
9. Kohler J, Schulze-Kremer S. The semantic metadatabase (SEMEDA): ontology based integration of federated molecular biological data sources. In Silico Biol. 2002;2(3):219-31.
10. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics. 2001;17 Suppl 1:S97-106.
11. Yu H, Kim, W., Hatzivassiloglou, V., and Wilbur, W. Disambiguating biomedical abbrevations. ACM Transactions on Information Systems (TOIS). 2006;24(3):380-404.
12. Bodenreider O. Lexical, terminological and ontological resources for biological text mining. Ananiadou S, McNaught, J., editor.: Artech House; 2006.
13. Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A. An ontology for bioinformatics applications. Bioinformatics. 1999 Jun; 15(6):510-20.
14. Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput. 2002b:362-73.
15. Philippi S, Kohler J. Using XML technology for the ontology-based semantic integration of life science databases. IEEE Trans Inf Technol Biomed. 2004 Jun; 8(2):154-60.
16. Xu H MM, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics. 2006;7:334.
17. Yu H, Kim W, Hatzivassiloglou V, Wilbur WJ. Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. J Biomed Inform. 2007 Apr; 40(2):150-9.
18. Dahlgren K, McDowell, J., and Stabler, E.P. Knowledge Representation for Commonsense Reasoning with Text. Computational Linguistics. 1989; 15:149-70.
19. Dahlgren K. Interpretation of Textual Queries Using a Cognitive Model: Ehrlbaum; 1992.
20. Dahlgren K, editor. Improving Precision and Recall with Linguistic Semantics. Proc Semantic Technology Conference; 2007; San Jose, CA.
21. Kornai A. Mathematical Linguistics.: Springer.; 2008.
22. Witten IH, Moffat, A.M., and Bell, T.C. Managing Gigabytes of Data. New York, NY.: Morgan Kaufmann.; 1999.
23. Gaithersburg M, editor. Proceedings Sixteenth Text REtrieval Conference (TREC 2007); November, 2007.
24. Gilman AG. Cross talk: interview with Al Gilman. Mol Interv. 2001 Apr;1(1):14-21.
25. Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. Nucleic Acids Res. 2002 Jan 1;30(1):169-71.
26. Wren JD, Garner HR. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods Inf Med. 2002; 41(5):426-34.
27. Eaton AD. HubMed: a web-based biomedical literature search interface. Nucleic Acids Research 2006; 34.
28. Lee S, Yang, L., Jianrong, L., Friedman, C., Lussier, Y.A. . "Discovery of protein interaction networks shared by diseases". . Pacific Symposium on Biocomputing; 2007. 2007. p. 76-87.
29. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics. 2001;17 Suppl 1:S74-82.
30. Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics. 2007 Oct 15; 23(20):2768-74.
31. Wellner BC, J and Pustejovsky, J. . "Adaptive string similarity metrics for biomedical reference resolution". . Proc ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics; 2005. 2005. p. 9-16

32. Yoshimasa T, McNaught, J and Ananiadou, S. . "Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinformatics 9(Suppl 3). 2008(S2).