Natural Language Query In The Biomedical Domain Based On The Cognition Search^{TM}

Saurabh Mendiratta[1], Kathleen Dahlgren[2] and Elizabeth J. Goldsmith[1]

[1]UTSouthwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX, 75390 [2]Cognition Technologies, Inc, 6133 Bristol Parkway, Culver City, CA 90203

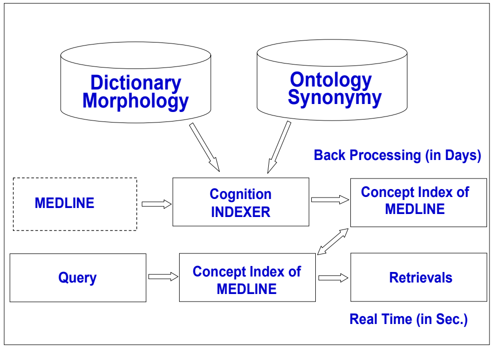
Abstract

Natural language processing technology is required to properly access the biomedical literature. Cognition semantic NLP technology has unraveled the full complexity of ordinary English. The architecture of the software and databases are such that multiple meanings of ordinary words and synonymy are resolved. To improve access to MEDLINE, several sources of biomedical language and acronyms were incorporated semi-automatically into the Cognition lexicon. Websites used in these projects include The Alliance for Cell Signaling (AfCS) and databases from the website http://medstract.med.tufts.edu, The Human Genome Nomenclature Consortium (HGNC), The United Medical Language System (UMLS) Meta-thesaurus, and The International Union of Pure and Applied Chemistry (IUPAC). These websites were chosen for their vocabulary (terms, phrases and acronyms), synonyms along with their the ontological relationships. The Cognition Search engine uses downward reasoning synonymy and word morphology to improve recall. The software also uses word sense selection and concept clustering which improve precision. Cognition was employed as a search engine and the resulting system was used to read and interpret MEDLINE abstracts. Meaning-based search of MEDLINE abstracts yields high precision (estimated at ≥80%), and high recall (estimated at >90%), where synonym information has been encoded. The present implementation can be found at http://medline.cognition.com.

Cognition's Semantic Map
(Based on Computational Linguistic Science)

Word Stems	506,000 Word Stems
Words and Phrases	536,000 word senses or concepts
Different Word Meanings	17,000 Ambiguous Word Definitions
Ontology or Taxonomy	7,000 Nodes
Synonyms	76,000 Thesaural Concept Groups

Architecture



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 as well as word strings. The NLP component relies on its dictionary, semantic map, and morphological and syntactic tags. At search time, CSIR interprets the query meaning, and searches for this meaning in its concept index rather using statistical word pattern matching. Therefore, the results are more complete and relevant.

Retrieval Features of Cognition Search

1. **Synonymy improves recall.**
   e.g. CD116 and granulocyte /macrophage-colony-stimulating factor GM-CSF-R-alpha.

2. **Sense disambiguation improves precision.**
   e.g. MBP stands for both mylein basic protein and maltose binding protein.

3. **Downward reasoning improves recall.**
   e.g. MAP kinase type ERK and P38 alpha.

4. **Morphology improves recall.**
   e.g. phosphorylate and phosphorylation.

5. **Phrase Recognition improves precision.**
   eg. Pyruvate dehydrogenase kinase.

Sample Queries

Cognition vs MEDLINE search	Cognition good/20*	Cognition bad/20*	Total	Pubmed good/20*	Pubmed bad/20*	Total
Genetic correlates of alcoholism	16	4	1436	6	14	44
DNA repair and aging	13	7	1220	11	9	1265
Drugs for fibromyalgia	15	5	1484	9	11	220
Genetic correlates of prostate cancer	15	5	2301	13	7	60
Genetic interactions of BCL2	14	6	876	8	11	19
Oxidative stress in plants	15	5	3122	9	11	3197
Spectroscopy of amidohydrolases	15	5	861	7	13	1142
Benzene induced neuropathy	14	6	220	6	1	7
Birth defects from glycol ether	14	6	20	13	7	61
Depression in aging	17	3	13381	7	13	3658
Symptoms of type II diabetes mellitus	16	4	241	7	13	24704
Dioxin and birth defects	12	8	111	7	13	294
Menopause and depression	17	3	696	11	9	1146
Genetic correlates of OCD	18	2	224	6	3	9
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	0.80			0.50		
Recall**	0.98			0.54		

*Percentage within the top 20 retrievals

**Assume total recall is the total of the Cognition retrievals

Precision - Specificity of the result.
Recall - Number of relevant retrievals.

PubMed

Missed the Target

Overload
- **many false positives – ambiguity**
- **lack understanding that words have multiple meanings**

Underload
- **miss relevant information – synonymy lack understanding of multiple words with similar meanings**

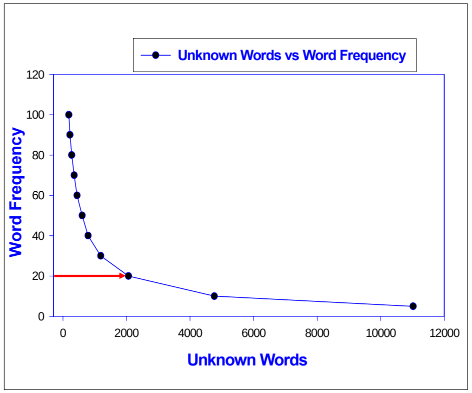
CognitionSearch
Data • Information • Knowledge • Understanding

Better Precision, Better Recall

Increased precision
- **retrieves fewer irrelevant documents manages ambiguity**
- **understands meaning of words and phrases**

Increased recall
- **more relevant documents retrieved**
- **understands synonymy**

Coverage of Medline Words



Currently, we are adding words that are missing in the lexicon (sorted by the frequency of occurrence). 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. This work contributes to precise interpretation of biomedical texts for research and data mining.

Acknowledgments

THE UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER AT DALLAS

HEALTH
COGNITION
TRY IT FOR YOURSELF