

Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap

Nipun Bhatia¹, Nigam H. Shah, MBBS, PhD², Daniel Rubin, MD, PhD²,
Annie P. Chiang PhD and Mark A. Musen, MD, PhD

¹Department of Computer Science, Stanford University, ²Stanford Center for Biomedical Informatics Research, School of Medicine, Stanford University

Abstract

The National Center for Biomedical Ontology is developing a system for automated, ontology-based access to online biomedical resources. The system's indexing workflow processes the text metadata of diverse resources such as datasets from GEO and ArrayExpress to annotate and index them with concepts from appropriate ontologies. This indexing requires the use of a concept-recognition system to identify the presence of ontology concepts in the resource metadata. In this paper, we present a comprehensive comparison of two concept recognizers – NIH's MetaMap and the University of Michigan's MGREP. We utilize a number of data sources and dictionaries to evaluate the concept recognizers in terms of precision, recall, speed of execution, scalability and customizability. Our evaluations demonstrate that MGREP has a clear edge over MetaMap for large-scale applications. Based on our analysis we also suggest areas of potential improvements for MGREP.

Introduction

There continues to be a tremendous increase in the amount, diversity, and rate of generation of high-throughput datasets as well as exponential growth in the biomedical literature. Since 1999, Gene Ontology (GO) annotations of data resources, when they are available, have enabled queries to accurately identify gene products associated with a particular cellular component, biological process or a molecular function. Similarly, creation of data annotations based on other shared ontologies would enable researchers to locate datasets, tissue samples, and clinical trials that relate to a given disease. This capability would permit a whole new class of integrative analyses.^{1, 2} However, due to the size of the data and the complexity of the task involved, adding ontology-based annotations to online data repositories manually on a case-by-case basis is unlikely ever to scale.³

At the National Center of Biomedical Ontology, we are developing methods to annotate large numbers of

data resources automatically, and have developed a prototype system for ontology-based annotation and indexing of biomedical data.⁴ The key functionality of this system is to provide a service that enables users to locate biomedical data resources related to particular ontology concepts. The system processes the text metadata of diverse biomedical data resources (such as gene-expression data sets, descriptions of radiology images, clinical-trial reports, and PubMed abstracts), annotating and indexing them with concepts from appropriate ontologies.

A critical step that our system performs is to recognize a given ontology concept in the metadata of some online data resource. This task is generally referred to as *concept recognition*. A core aspect of concept recognition is a taxonomy or ontology to which text is mapped. In the biomedical domain, the UMLS is an extensive resource that incorporates a number of disparate terminologies and ontologies and that provides a cross-referencing of related concepts. However, efforts to map public, open biomedical resources to semantically rich thesauri such as UMLS have been scattered. Barring a few initiatives,^{1, 2, 5} most efforts to date have focused on mapping text from patient records to UMLS, rather than on mapping metadata from online biomedical resources.^{6, 7}

Most previous work in concept recognition in bioinformatics has been restricted to the identification of protein and gene names,⁸⁻¹⁰ with a few groups attempting to identify concepts representing relationships among entities.¹¹ This trend is obvious when looking at popular tools such as EBIMed and TextPresso, all of which identify genes or proteins in documents, but struggle to identify disease names.^{11, 12} The same emphasis was visible in the BioCreative text-processing challenge, which was primarily concerned with recognizing gene and protein names.⁸

In the field of clinical informatics, the efforts to recognize concepts in text have focused on finding disease names in electronic medical records, discharge summaries, clinical guideline descriptions,

and clinical-trial summaries.^{6, 7, 13} However, electronic medical records are seldom made “public” as online biomedical resources. As a result, current methods and tools are usually not portable across a different problem category – such as processing the metadata of public, open biomedical resources.

In recent times, there has been a shift in the focus of research from individual genes and proteins to entire biological systems.¹⁴ As a result, researchers need services that can process the metadata of diverse resources to annotate and index them with concepts from appropriate ontologies, and that can enable the researchers to locate resources related to particular ontology concepts. Concept recognition is a key step for such systems.

NLM’s MetaMap was one of the first tools for recognizing UMLS concepts.¹⁵ It is widely regarded as the gold standard for this task. Recently, there have been a number of tools such as MGREP¹⁸ and MTag¹⁹ that also perform concept recognition. The advent of these new tools has made the task of evaluating concept recognizers particularly important.

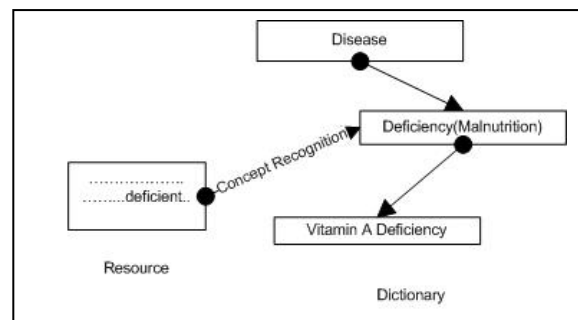
We conducted a survey of existing concept recognizers based on their published reports, and selected MetaMap and MGREP as the two tools to evaluate for our purposes. This paper provides comparison of NLM’s MetaMap and the University of Michigan’s MGREP.¹⁶ We choose MGREP because it is claimed to be a fast and scalable tool for concept recognition with a high degree of customizability vis-à-vis dictionaries and resources. Considering the vast number of biomedical resources and ontologies available, factors of speed, scalability, and customizability are of prime concern in developing a concept-recognition system⁴.

In the remaining part of the paper, we first give a brief outline of the concept-recognition task and discuss our data sources and dictionaries. We explain the evaluation methodology adopted and the results obtained. We then discuss the performance of concept recognizers based on a number of performance metrics such as precision and recall. We also analyze the suitability of a concept recognizer based on a number of subjective parameters such as ease of use, ability to customize, and scalability. We conclude with a summary of our findings.

Concept Recognition

In the domain of Biomedical informatics, the task of concept recognition can be understood as mapping biomedical text to a representation of biomedical knowledge consisting of inter-related concepts,

usually codified as an ontology or a thesaurus. Figure 1 illustrates the task of a concept recognizer. Most concept recognizers take as input a resource and a dictionary – which can be a flat list or taxonomy of hierarchically related terms – and produce annotated files. The concept recognizer in Figure 1 recognizes the string ‘deficient’ in the resource and maps it to the concept ‘Deficiency’ in the dictionary. Most concept recognizers leverage natural-language



processing and computational linguistic techniques to some extent.

Figure 1. The figure shows the working of a generic concept recognizer, which maps the text ‘deficient’ to the concept of ‘Deficiency’ in a hierarchical dictionary of concepts.

Methods

Data Sources and Dictionaries

There are many online resources in biomedicine, ranging from data repositories such as Array Express and the Gene Expression Omnibus (GEO), radiology image repositories such as GoldMiner, which stores published images and their figure captions, to clinical trial repositories and Medline. Each of these resources is assumed a particular type of biomedical knowledge. Comprehensive evaluations of concept recognizers would require several of these resources and their annotations to be evaluated. Also, it would be important to find out if a particular concept recognizer is more efficient in processing the textual annotations of certain resources. The variation in the sizes of the resources helps us to compare the scalability of a concept recognizer. For example, the size of the entire MedLine download is ~10.4 Gigabyte, while the size of ClinicalTrials.gov is only of the order of 99 Megabytes. This variation allows a performance benchmark on the scalability of the concept recognizers as well as an evaluation of the effect of data size on the execution time. Due to the generally large size of most biomedical resources, it is very important to see how scalable a concept recognizer is with respect to size. Table 1 gives a brief overview of the data resources we used in evaluating the concept recognizers. In each case, we

used the title and description of an element from the resource as our input for concept recognition.

Resource	Elements	Size
Array Express www.ebi.ac.uk/arrayexpress	3212	3.5MB
ClinicalTrials.gov www.clinicaltrials.gov	50303	99MB
Gene Expression Omnibus www.ncbi.nlm.nih.gov/geo/	2085	0.7MB
PubMed (Subset) www.ncbi.nlm.nih.gov/pubmed	2827	3.7MB
Gold Miner (Subset) http://goldminer.org	2085	0.5MB

Table 1: Size and Number of Elements of Data Sources

Dictionaries

A dictionary with respect to a concept recognizer can be described to be the set of terms or concepts that we aim to recognize in the data. Analogous to data sources, dictionaries can be specialized along different axes, such as diseases and anatomical parts. As most of the work in Biomedical informatics has primarily focused in recognizing genes or proteins⁸ the dictionaries for genomics and proteomics are comprehensive and extensively evaluated. The same is not true for dictionaries pertaining to diseases, body parts, biological processes, drug names, and so on.

We performed evaluations of concept recognizers using a number of different dictionaries. Thus, we could identify if a particular concept recognizer–dictionary combination is best suited for a particular semantic class of entities, such as diseases or body parts. Further, as in the case of data-sources, varying size of the dictionary helps to evaluate the scalability of the concept recognizers. As the data and the dictionary are both critical inputs to a concept recognizer, we note the effect of sizes of both in the performance of concept recognizers. In tune with the above notions – we performed evaluations using four different dictionaries (Table 2) of varying sizes. The ‘diseases’ dictionary comprises all the concepts in the UMLS that are of semantic type ‘disease or syndrome’. The ‘biological processes’ dictionary comprises all the GO biological processes contained in the UMLS.

Dictionary	Size	Concepts
SNOMED-CT	48MB	1,139,586
Diseases	38MB	764,420

FMA (Body Parts)	4.8MB	93,335
Biological Processes from GO	1.18MB	31,294

Table 2: shows the size and number of concepts in each of the dictionaries SNOMED-CT, Diseases, FMA and Biological Processes.

Evaluation Workflow

We constructed a pipeline for performing the evaluations and to provide a platform to plug in the data sources, on which to run the concept-recognition tools and to map the tool-specific output to a common format. Ideally, this task should have been done using a framework such as IBM’s UIMA, but both MetaMap and MGREP are not available as UIMA components. First, we randomly selected 200 lines from each data source and converted these sources from their native format to a format suitable for input to the concept-recognition tool. For example, the Array Express data are commonly available in XML format; however, University of Michigan’s MGREP requires the data to be in a three-column tab-delimited format. The next step involved running the concept-recognition tool and obtaining the processed file in a format specific to each tool. In the final step, we converted the output files of the different concept recognizers to a common format to ensure uniformity and to aid in performing comparative analysis. A total of four experts examined the resultant files for scoring true positives and false positives. We attempted to estimate recall by assuming a false negative result if no concept was identified. In addition to precision and recall, our evaluation considered customizability and scalability.

Customizability – We define the qualitative measure of customizability of a concept recognizer as the ease with which a dictionary and a data source can be configured for it.

Scalability – We define scalability by how easily a concept recognizer handles different sizes of dictionary and resource.

Results

Tables 3 and 4 provide the numbers of concepts recognized by the two tools with different dictionaries and different data sources as input. Both tools recognize concepts from all resources tested and using all four dictionaries tested. In general, MGREP recognizes a lower number of unique concepts than MetaMap.

Resource	Biological	Diseases
----------	------------	----------

	Process		MG	MM
	MG	MM		
Clinical Trials	10	106	409	710
Gold Miner	12	80	753	1283
GEO	136	188	337	704
MedLine	26	48	22	209

Table 3. Total number of concepts recognized by MGREP and MetaMap across all resources using the biological process and diseases dictionaries. MG = MGREP; MM =MetaMap.

	FMA (Body Parts)		SNOMED	
	MG	MM	MG	MM
Clinical Trials	243	380	1548	1730
Gold Miner	671	1097	3747	3400
GEO	272	818	2228	2372
MedLine	57	132	1320	1088

Table 4. Total number of concepts recognized by MGREP and MetaMap across all resources using the Foundational Model of Anatomy and SNOMED-CT as dictionaries. MG = MGREP; MM =MetaMap.

Table 5 compares the precision for the two tools using the Biological-Process dictionary from the Gene Ontology. To compute recall accurately requires the domain expert to go through each record to identify true and false negatives. We examined the option of estimating recall under the simplifying assumption that a concept should be recognized for every record processed and, if no true positive concept is recognized, then the record constitutes a false negative. This assumption could provide us with an estimate of the lower bound on recall. However, for dictionaries such as Biological Processes and resources such as Figure Captions from radiology images, such an assumption is flawed. Therefore we were unable to estimate recall.

Data Source	MGREP	MetaMap
GEO	0.93	0.73
Gold Miner	0.58	0.33
MedLine	0.77	0.76
Clinical Trials	0.6	0.63

Table 5. Precision of MGREP and MetaMap using Biological Processes as the dictionary.

Table 6 compares the precision for the two tools using the ‘diseases’ dictionary, which contains UMLS concepts that are of semantic type ‘disease or syndrome’. We are unable to calculate recall in this case because some concept was recognized in almost all records and we cannot compute recall using the assumption discussed above.

Data Source	MGREP	MetaMap
GEO	0.88	0.755
Gold Miner	0.73	0.548
MedLine	0.23	0.091
Clinical Trials	0.87	0.71

Table 6. Precision of MGREP and MetaMap using the ‘diseases’ dictionary.

In general, MGREP has a higher precision in recognizing Biological Processes. When considering precision, MGREP outperforms MetaMap in almost all cases, with the exception of an insignificant edge for MetaMap in recognizing Biological Processes in ClinicalTrials.gov.

Discussion

Based on our evaluation and analysis, we identify the following considerations in selecting a concept recognizer for the automated ontology-based annotation: (1) ability to work with non UMLS terminologies; (2) ability to work offline vs. online (annotation of user-submitted data as a service); (3) high speed as well as accuracy in terms of precision and recall.

By design, NIH’s MetaMap is very tightly coupled with the UMLS. This makes mapping text to UMLS concepts very easy. However, generating a custom dictionary for annotation that uses concepts from outside UMLS is non-trivial. MetaMap requires the dictionary to be in a specific format with certain databases always present. Some applications, such as the Open Biomedical Resources system under development by the NCBO^{1, 2}, use a number of different dictionaries from not only UMLS but also other sources for which terms are not present in the UMLS. Formatting such dictionaries into the format required by MetaMap is not always possible. With respect to the input data, MetaMap is very adaptable and easy to customize. It does not require the input sources to be structured in any particular way.

In terms of speed of execution, MetaMap requires much more processing time than does MGREP. For example, MGREP can process 1/5th of the data from ClinicalTrials.gov in 7 seconds, whereas MetaMap runs for over 8 minutes. This makes MetaMap unsuitable for developing an online annotation service. The powerful lexical capability of MetaMap results in MetaMap finding about four times more concepts than MGREP, however.

One of the standout features of MGREP is its fast execution and scalability across all the dictionaries and data resources tested. However, MGREP identifies a large number of concepts that are redundant – concepts recognized at the same position in the input string – and overall the number of unique concepts recognized is less than with MetaMap (Tables 3 and 4).

MGREP is easily customizable to accept variation in the formats of both the input data and the dictionary, making it very easy to use for custom applications. It requires the dictionary to be in an easy to create two-column, tab-delimited file and similarly requires the resources to be in tab-delimited files. MGREP places no rigid requirements on the structure and presence of concepts.

Finally, MGREP shows substantially higher precision than does MetaMap across most resources and dictionary types.

Conclusion

MetaMap places a rigid constraint on the dictionary structure and cannot be used for applications that require dictionaries outside of the UMLS (such as those from the Open Biomedical Ontology library). Because of its slow speed, it cannot be used for many real-time applications or for applications in which either the data sources or the dictionary changes frequently, requiring recurrent reprocessing

MGREP has extremely fast execution speed, but fewer concepts are recognized. If future versions of MGREP provide the ability to generate lexical variants, recall would be enhanced and MGREP could become the concept recognizer of choice for applications that need to process large datasets, that require large dictionaries, or that involve frequent reprocessing.

Acknowledgements

We thank Charles Kahn for providing a sample of Gold Miner data. This work was funded by the National Center for Biomedical Ontology under NIH grant U54 HG004028.

References

- 1 Shah NH, Rubin DL, Espinosa I, et al. Annotation and query of tissue microarray data using the NCI Thesaurus. *BMC Bioinformatics*. 2007;8:296.
- 2 Shah NH, Chiang AP, Butte AJ, et al. Ontology-driven Indexing of Public datasets for Translational Bioinformatics. *AMIA Summit on Translational Bioinformatics*; 2008; San Francisco; 2008.
- 3 Baumgartner WA, Jr., Cohen KB, Fox LM, et al. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*. 2007 Jul 1;23(13):i41-8.
- 4 Jonquet C, Shah NH. *Ontrez Project Report. SMI Technical report*. Stanford. CA 2008.
- 5 Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol*. 2006 Jan;24(1):55-62.
- 6 Reeve LH, Han H. CONANN: An Online Biomedical Concept Annotator. *LECTURE NOTES IN COMPUTER SCIENCE*. 2007;4544:264.
- 7 Hersh W, Leone TJ. The SAPHIRE server: a new algorithm and implementation. *Proc Annu Symp Comput Appl Med Care*. 1995:858-62.
- 8 Hirschman L, Yeh A, Blaschke C, et al. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*. 2005;6 Suppl 1:S1.
- 9 Zhou G, Shen D, Zhang J, et al. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*. 2005;6 Suppl 1:S7.
- 10 Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*. 2005 Jul 15;21(14):3191-2.
- 11 Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. 2004 Nov;2(11):e309.
- 12 Rebholz-Schuhmann D, Kirsch H, Arregui M, et al. Protein annotation by EBIMed. *Nat Biotechnol*. 2006 Aug;24(8):902-3.
- 13 Moskovitch R, Martins SB, Behiri E, et al. A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. *J Am Med Inform Assoc*. 2007 March-April;14(2):164-74.
- 14 Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*. 2006 Feb;7(2):119-29.

- 15 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
- 16 Dai M, Shah NH, Xuan W, et al. An Efficient Solution for Mapping Free Text to Ontology Terms. AMIA Summit on Translational Bioinformatics; 2008; San Francisco, CA; 2008.