ICBO: International Conference on Biomedical Ontologies · Buffalo, NY, USA
Doctoral & Post-Doctoral Consortium, July 27, 2011

Applications for a Translational Biomedical Ontology Model

Robert Yao, Graciela Gonzalez

Arizona State University

1 Introduction

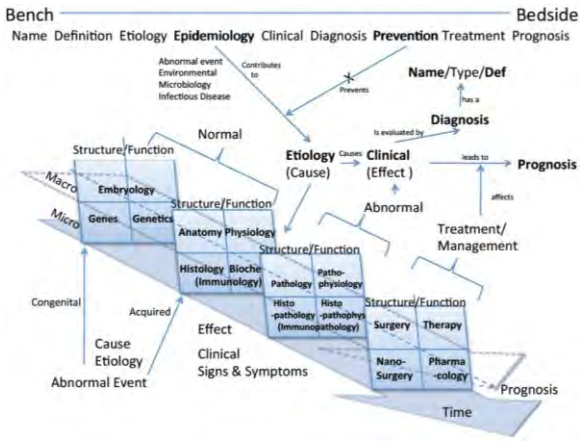
In the past two centuries, the amount of scientific (especially medical) information has exponentially increased. As more became discovered, new paradigms required novel vocabularies and terms. This created new specialties and subspecialties of scientists and health professionals. They became subcultures with different languages. For instance, in general medicine, 'idiopathic' means: unknown cause; while in neurology, it means: presumed genetic.

2 Problem

Sub-specialization coupled with the inundation of data has been problematic. First the current infrastructure does not take into account semantic similarities between these specialties. Second, because a coherent semantic infra-structure is missing, experts in those fields struggle to adequately understand disease and define syndromes. This is especially the case in neurology. For instance, inferences made from new findings could change the fundamental understanding of Alzhiemer's disease. This in turn affects the kinds of questions researchers need to ask. Additionally, the criteria with which to define syndromes in Epilepsy are subject to change. This affects the ability for clinical researchers to locate patients that fit their criteria. In other words, this all creates an inefficient scenario where too much information builds up and keeping up with the literature proves difficult. Thus, understanding of the disease is incomplete, which then adversely affects a researcher's ability to uniquely research an unanswered question and/or to locate the right subjects for clinical research.

3 Background

Techniques in information retrieval like machine learning have been proven to be effective at clustering and classifying data. However, without specific objectives or rules pertinent to the domain, the use of machine learning alone has run into various limitations in medicine. Ontologies which specify how concepts are represented are attempting to semantically standardize how information is handled and provide rules for the categorization of terms and defining their relationships. By connecting concepts and relationships to specific instances and information, a knowledgebase is instantiated using the ontology as its schema. While various biological and medical ontologies exist, finding the correct mapping combinations to answer clinical questions is still a major challenge.



4 Proposal

In an effort to solve this dilemma on several fronts, this work proposes a translational biomedical ontology for human disease. Rather than classifying where a disease name fits in a hierarchy (like most disease ontologies), it provides a robust ontologic infrastructure with which to help define any specific disease and its progression through various states that develop through time. Additionally, it does this from the perspective of a basic scientist to medical doctor to a clinical researcher. Thus it potentially aids

the organization, understanding, and transfer of information from bench to bedside and back.

5 Current Projects

In one use case, part of the ontology is being used on a corpus to train and test a machine learning algorithm that recognizes and semantically tags sentences in scientific and medical text about Alzheimer's Disease. The purpose of this is to build an algorithm for a system that automatically updates its own knowledgebase. It can scan sentences and automatically classify information about a disease and thus update that information in a knowledgebase. The growing knowledgebase is then used as the training and testing corpus and helps to further refine the algorithm for future scans. The knowledgebase also serves as a semantically searchable database that can be filtered for information that answers a specific question or helps someone to better understand the disease or the direction that future research must take.

In another use case, the terms and relationships present in the ontology itself are currently being explored in the definition of Epileptic seizures and syndromes. This is to be used with clinical data in i2b2 for the purpose of identifying subjects for clinical studies and trials. Information gathered at this stage will be mined, semantically tagged, and added to the knowledgebase. This will add to the corpus used to the machine learning model and hopefully result in more accurate tagging which will lead to better understanding.

6 Methods

As a first step towards proving this, efforts to automatically populate instances were pursued. The ontology was used in a machine learning classifier to semantically tag sentences automatically. In this project, over 1000 diseases were manually reviewed and semantically tagged. Categories that define disease in the ontology were chosen to be the semantic tags: Etiology, Epidemiology, Clinical, Diagnosis (Tests), Prevention, Treatment, and Prognosis. For the purpose of this experiment, sentences pertaining to Alzheimer's disease in Harrison's and Cecil's Textbooks of Internal Medicine and Robbin's Pathology were manually curated and tagged to produce a corpus. A program was created to do a 10x10 cross validation of the corpus using Support Vector Machine and Logistic Regression classifiers in WEKA to test and train a model for automatic classification.

7 Results

The experiment is ongoing and results preliminary. Precision, recall, and f-measure of the model indicate high precision (85-100%) and low recall (7-20%) which adversely affects the f-measure. The low recall can be improved by the removal of stop words, stemming, adding more tagged sentences to the corpus. Perhaps active learning, or including human feedback in the process, may improve results. All of these can be attempted to maximize f-measure. Given that clinical text tends to focus on discussing epidemiology, clinical signs and symptoms, diagnostic tests and procedures, and prognosis of disease more than etiology, it is not surprising that etiology had the lowest precision. This may be improved further by adding more basic science text into the corpus.

8 Conclusion

The high precision indicates that while much work still needs to be done to improve recall, the use of learning algorithms to automatically tag ontology terms is a viable process. This provides the means to formalize a difficult problem that needs further testing. The next step is to expand the use of the other semantic terms and to determine if the tagged sentences from these are able to be searched for answers to biomedical questions.

	Etiology		Epidemiology		Clinical		Diagnosis		Prevention		Treatment		Prognosis	
	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR
Precision	86%	87%	95%	98%	98%	98%	92%	91%	100%	99%	99%	91%	92%	99%
Recall	20%	20%	11%	11%	13%	11%	21%	21%	4%	4%	6%	8%	12%	12%
F-measure	32%	32%	19%	20%	22%	20%	34%	35%	8%	9%	12%	14%	22%	22%