Ontology-Based Text Mining of Concept Definitions in Biomedical Literature

Saeed Hassanpour, Amar K. Das

Stanford Center for Biomedical Informatics Research,
Stanford, CA 94305, U.S.A.
{saeedhp, amar.das}@stanford.edu

Abstract. Many developers of biomedical knowledge bases typically validate and update formalized knowledge based on reviews of full-text scientific articles, but finding text relevant to domain concepts can be tedious and prone to errors. Prior methods have automated this process by matching term-based patterns within a single sentence. In our work developing a knowledge base of autism phenotypes, specified using Semantic Web standards, we are interested in finding multi-sentence sections of text that contains complex phenotype definitions. In this paper, we present a text-mining method that incorporates both ontology- and rule-based semantics to determine which section is relevant. We evaluated our method in undertaking text extraction for the set of full-text articles used to create the knowledge base. We show that our method has higher precision and recall than a term-based approach in identifying definitions that contain complex patterns and occur across sentence boundaries.

Keywords: Information Extraction, Text Analysis, Semantic Web, Ontology, Rules Base, OWL, SWRL

1 Introduction

Biomedical knowledge resources, such as terminologies and ontologies, are important for community-based annotation and sharing of data. Creating and maintaining these resources is challenging given the rapid growth of scientific knowledge. Generally, scientists, annotators and developers try to keep up by using search engines that find publications relevant to given concepts in the knowledge resource. However, users still need to review the publications and find sections within the documents that relate to the concept being searched. One solution to this challenge is to automatically identify the relevant parts of a full-text document. Prior methods, such as Textpresso [1], have focused on finding individual sentences that match the terms of biomedical concepts and of properties that connect concepts. Such approaches do not find sections of an article—including multiple sentences—that are semantically and implicitly relevant to the definition of a concept. In our work, we present a novel text mining method that retrieves the most semantically informative text in a document using definitions of concepts modeled as rules in a domain ontology, and we compare the precision and recall of our method against a term-based approach.

Our work is motivated by the needs of developers of an ontology of autism phenotypes [2, 3]. As part of these efforts, experts want to easily find text within a publication that relates to the definition of a phenotype concept, both to find new definitions of that concept and to annotate the document section as the relevant text to the concept. For example, in a paper on autism genetics, Hus et al. [4] define Savant-positive and Savant-negative phenotype concepts as:

> *The Savant Skills Factor was based on … current and ever scores of four ADI-R items: visuospatial ability, memory skill, musical ability, and computational ability. Item scores were summed and divided by total number of items to generate a score between 0 and 1. … Participants were then divided into two groups: Savant-positive and Savant-negative … .*

The autism ontology uses the Web Ontology Language (OWL) [5] to model concepts and hierarchical relationships and the Semantic Web Rule Language (SWRL) [6] to define phenotype concepts as value restrictions on data collected through standardized instruments, such as the Autism Diagnostic Instrument-Revised (ADI-R).

2 Related Work

Finding text relevant to a search term is undertaken by some web search engines, which provides a few lines of site description or snippet for a search result to indicate the relevance of a web page to the search query. Google, for example, uses the description provided by meta tags, references to the web pages, Open Directory [7], and the text around the query keywords on web pages to provide informative search result descriptions [8]. We argue that structured domain knowledge can be used to enhance the relevance of snippets to the queries as well and provide the most semantically relevant parts of web page contents in result snippets.

Another related work in this field is question-answering systems, which return a part of a text from a corpus as the answer to a specified question. These techniques rank the snippets from the relevant documents by criteria such as: containing expected types of named entities, the percentage of overlap with question terms, containing lexical patterns, and using information from lexicon dictionaries [9-11]. Other work has tried to retrieve descriptive phrases from free text by using pattern matching, word counting, and sentence location without using domain knowledge [12]. In our work, we address the broader problem of extracting text that is semantically relevant to domain concepts. Our approach leverages the structured and axiomatic forms of knowledge in ontologies and rules, which contain richer semantic relationships than lexical databases.

3 METHODS

In our work, we find the most relevant parts of science publications to domain concepts using existing OWL ontologies and SWRL rules. As noted, both provide formal definitions of domain concepts and their relationships to other concepts.

3.1 Semantic Concept Modeling

As the first step, we need a formal representation of domain concepts. In this work, we use vector space modeling, a common method in the web search engines for indexing web pages [13], and a structured knowledgebase as a basis of the concept modeling. The concepts in the knowledgebase may be formally defined in logical form of SWRL rules and saved as a part of an OWL ontology, as in the case of the autism ontology. We thus consider rules' components as relevant concepts and incorporate them in our modeling for better presentation of the main concept. Therefore, we have one dimension for each ontology class and property mentioned in the rule as relevant concepts.

Besides the classes and properties that are mentioned in the rule, we use ontology hierarchies to extract more related concepts and incorporate them in the concept presentation. We consider the parents and grandparents of the main concept and its related concepts extracted from the corresponding rule as potential related concepts that can strengthen our concept vector modeling. However, the relevance of these concepts from the ontology hierarchy decreases by their distance from the main concept in the hierarchy graph. Therefore, we weight these related terms in the vector presentation less than the main class and the related concepts explicitly mentioned in the rule that defines the concepts. As a heuristic choice to capture these differences, we count the frequencies of the parent classes or properties as half of the actual frequencies, and the frequencies of grandparent classes or properties as one-quarter of the actual frequencies.

3.2 Relevant Text Finding

After we model the concept, we go through a publication to find the most relevant parts of the text for a particular concept. As the first step, we look at the vector representation of the concept and found all the terms associated with that concept as the concept terms. Concept terms are the terms that have weights greater than zero in the concept vector presentation. We then go through the publication and mark all the occurrence of the concept terms in the text. We cover occurrences of different forms of a concept terms by applying, Porter stemming algorithm, a common stemming method for English terms [14], on both concept terms and publication terms.

Given the occurrences of concept terms in a publication, we treat them as indicators of relevant parts of the text and use single linkage hierarchal clustering to find the candidates for the most relevant parts of the publication. The average sentence length in our corpus is 20 words. In the single linkage clustering we use 30 words as a heuristic threshold and in every step we merge the closest clusters that are separated by less than 30 words. Thus, we ensure that a continuous section of text without any concept term is limited to a few sentences and the whole cluster is continuously correlated to the concept. We consider these clusters as the candidates for the most relevant parts of the text.

3.3 Text Modeling and Correlation Computation

In this work, our goal is to quantify the relevance between concepts and pieces of text. Therefore, we need a mathematical modeling of texts. We use vector space modeling again to provide a common basis for comparison. Vector space modeling for documents' text is based on term frequencies. To model a part of a text as a vector, we first remove the stop words, the most common English words that are not informative about the context. We use a common list of stop words in English [15]. Then we apply Porter stemming algorithm to replace different derivations of a word with their root. Then we build a vector with one dimension for each term in the text and assign the frequency of that term in the text as the value of that dimension in the vector.

After we present both text words and domain concepts as vectors, we need to compute the correlations between them in order to find the most relevant parts of a publication for a concept. To do that, we use cosine similarity as the measure of correlation between texts and concepts. The cosine similarity for two vectors is the cosine of the angle between them. Similarity values range from 0 for orthogonal vectors to 1 for parallel vectors.

3.4 Evaluation Strategy

In this work, we applied our method on the autism phenotype ontology and the papers used to derive those concepts as mentioned in Section 1. We examined only the top five most relevant parts of the publication for each concept and had an autism ontology expert review these text sections to determine the efficacy and accuracy of whether each section was related or not to the definition of the concept. To investigate the significance of using ontological hierarchies and rule bases, we compared our method to a baseline, which is a term-only method. The baseline method is a variation of our method that only uses the terms in the semantic concept-modeling step. That is, our baseline approach does not include concepts from the ontology or rules that are related to the term. To eliminate bias in the assessment of the performance of the two approaches, the expert was blind to which method produced the extracted text.

4 Results

The autism ontology contains 1726 classes and properties, and it includes 156 SWRL rules that correspond to 145 phenotype definitions. The ontology and rules were based on a review of 26 publications that had been undertaken by one of the authors (AKD) and other domain experts in autism [3]. For this study, we selected 49 domain concepts that had rules using multiple criteria to define a phenotype (such as the example concept of Savant positive given in Section 1). We excluded phenotype definitions where the concept directly corresponded to the value of a single item on a clinical assessment. We applied both our ontology-based text extraction method and the term only method on each of the 49 concepts, and we returned the top 5 most

relevant parts of the publication for review by the domain expert. Altogether 338 sections of text were reviewed and evaluated by the autism ontology expert as to whether they were relevant to the corresponding phenotype concept. Table 1 shows the precision of our ontology-based method and the concept-based method—that is, the percentage of returned sections that refer to the concepts.

Table 1. The precision of the term- and ontology-based methods in finding texts relevant to phenotype definitions

Method	Precision (%)
Term based	68 %
Ontology based	76 %

In our evaluation strategy, we knew that every concept had been defined in the corresponding publication. For further investigation of the relevance strength in our results, we asked the reviewer to identify which of the five most relevant parts of the publications for a concept contained a clear definition. We used this to calculate the recall for each method, which is the percentage of concepts that their definitions were found. Table 2 shows the recall of the concept- and ontology-based methods in finding the definitions of the concepts in the corresponding publication text.

Table 2. The recall of the term- and ontology-based methods in identifying phenotype definitions in the publication text

Method	Recall (%)
Term based	39 %
Ontology based	69 %

5 Discussion

In this paper, we present a novel method to find parts of text in scientific publications that relate to definitions of biomedical concepts. In comparison to methods that do term matching to find individual sentences that contain a single concept or pairwise sets of concepts, our ontology-based approach addresses the challenge of finding a concept definition that occurs across multiple sentences or that is semantically similar to predefined concepts. Our approach was particularly driven by the need to identify text related to complex domain concepts like autism phenotypes, in which use different terms and terminologies refer to similar concepts. Our evaluation shows that ontology hierarchies and rules have a large impact on identifying the relevant parts of the text. This is because of the informative nature of ontological hierarchies and the inter-relationship of concepts maintained in rule bases.

As future work, we are planning to improve upon our method by using the text's syntactic structures through constituent and dependency parsing methods. The syntactic and dependency information can be used in the text modeling to improve the concept relevance detection. Also, we will consider further addition of name entity

recognition methods, which can extract the information about the biomedical concepts outside of the ontologies in texts. We are planning to use this information to develop a richer presentation of text and find relationships between the publication text and the queried biomedical concept.

Acknowledgments. The authors would like to acknowledge Martin O'Connor and Siddharth Taduri for their comments on the approach. This research was supported in part by grant R01 MH87756 from the National Institutes of Health.

References

1. Muller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol. 2(11):e309. doi:10.1371/journal.pbio.0020309 (2004)
2. Young, L., Tu, S.W., Tennakoon, L., Vismer, D., Astakhov, V., Gupta, A., Grethe, J.S., Martone, M.E., Das, A.K., McAuliffe, M.J.: Ontology Driven Data Integration for Autism Research. 22nd IEEE International Symposium on Computer Based Medical Systems, pp. 1–7, Albuquerque, NM (2009)
3. Tu, S.W., Tennakoon, L., Das, A.K.: Using an Integrated Ontology and Information Model for Querying and Reasoning about Phenotypes: The Case of Autism. AMIA Annual Symposium, pp. 727–731, Washington, DC (2008)
4. Hus, V., Pickles, A., Cook, E.H., Risi, S., Lord, C.: Using the Autism Diagnostic Interview-Revised to Increase Phenotypic Homogeneity in Genetic Studies of Autism. Biol Psychiatry. 61(4), 438–448 (2007)
5. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview. W3C Recommendation, <http://www.w3.org/TR/2004/REC-owl-features-20040210/> (2004)
6. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML. <http://www.w3.org/Submission/SWRL/> (2004)
7. Open Directory Project, http://www.dmoz.org/
8. Google support on snippets, http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35624
9. Cooper, W.S.: Fact Retrieval and Deductive Question-Answering Information Retrieval Systems. J. ACM. 11(2), 117–137 (1964)
10. Miliaraki, S., Androutsopoulos, I.: Learning to Identify Single-snippet Answers to Definition Questions. 20th International Conference on Computational Linguistics, Geneva, Switzerland (2004)
11. Radev, D.R., Prager, J., Samn, V.: Ranking Suspected Answers to Natural Language Questions Using Predictive Annotation. 6th Conference on Applied Natural Language Processing, pp.150–157, Seattle, WA (2000)
12. Joho, H., Sanderson, M., Retrieving Descriptive Phrases from Large Amounts of Free Text. 9th ACM Conference on Information and Knowledge Management, pp. 180--186, McLean, VA (2000)
13. Salton, G., Wong, A., Yang C.S.: A Vector Space Model for Automatic Indexing. Commun ACM. 18(11), 613–620 (1975)
14. Porter stemmer, http://tartarus.org/~martin/PorterStemmer
15. List of English stop words, http://members.unine.ch/jacques.savoy/clef