Discovering Biomedical Relations Utilizing the World-Wide Web

Sougata Mukherjea and Saurav Sahay

Pacific Symposium on Biocomputing 11:164-175(2006)

DISCOVERING BIOMEDICAL RELATIONS UTILIZING THE WORLD-WIDE WEB

SOUGATA MUKHERJEA
IBM India Research Lab,
Hauz Khas, New Delhi, India
E-mail: smukherj@in.ibm.com

SAURAV SAHAY
College of Computing, Georgia Institute of Technology,
Atlanta, Ga, USA
E-mail: ssahay@cc.gatech.edu

To crate a Semantic Web for Life Sciences discovering relations between biomedical entities is essential. Journals and conference proceedings represent the dominant mechanisms of reporting newly discovered biomedical interactions. The unstructured nature of such publications makes it difficult to utilize data mining or knowledge discovery techniques to automatically incorporate knowledge from these publications into the ontologies. On the other hand, since biomedical information is growing explosively, it is difficult to have human curators manually extract all the information from literature. In this paper we present techniques to automatically discover biomedical relations from the World-wide Web. For this purpose we retrieve relevant information from Web Search engines using various lexico-syntactic patterns as queries. Experiments are presented to show the usefulness of our techniques.

1. Introduction

A Semantic Web for Life Sciences storing information about all the biomedical concepts as well as relations between them will enable researchers and autonomous agents to efficiently retrieve information as well as discover unknown and hidden knowledge. However, the current situation of the Semantic Web is one of a vicious cycle in which there is not much of a Semantic Web due to the lack of semantic markup of data, and there is such a lack because there is no easy way to semantically annotate biological knowledge.

One of the goals of Semantic Web research is to incorporate most of

the knowledge of a domain in an ontology that can be shared by many applications. Various ontologies and knowledge bases have been developed for Life Sciences including *Unified Medical Language System (UMLS)*[1] and Gene Ontology[2]. These ontologies organize information of various biological concepts, each with their attributes, and describe simple relationships like *is-a* and *part-of* between concepts. However, these ontologies are not up-to-date and may not have the information about newly discovered biomedical entities. Moreover, they generally do not incorporate complex relationships between biomedical entities. For example, although UMLS contains details about many diseases, viruses and bacteria, it does not incorporate relations between diseases and the causes of the diseases. Therefore, representing these ontologies and knowledge sources in Semantic Web languages like OWL will not be sufficient to create a Semantic Web for Life Sciences.

Journals and conference proceedings represent the dominant mechanisms of reporting biomedical results. The unstructured nature of such publications makes it difficult to utilize automated techniques to extract knowledge from these sources. Therefore information about new biomedical entities or relations between them need to be added to the ontologies manually. However, because of the very large amounts of data being generated, it is difficult to have human curators extract all these information and keep the ontologies up-to-date.

A large section of the research literature is available online and is therefore searchable by Web Search engines like Google. Although databases like PubMed are not readily accessible to the Google crawler, many of the PubMed abstracts has been crawled by Google (by following the links to these abstracts specified in other Web pages). Moreover, publications available from researchers' homepages or from conference Websites as well as biomedical information sources other than research publications can be also accessed by Google.

This paper presents a technique to automatically discover biomedical relations from the World-wide Web. We first query Web search engines with hand-crafted lexico-syntactic patterns to retrieve relevant information. The knowledge extracted from the search results can be used to augment the ontologies and knowledge bases and create a Semantic Web for Life Sciences. Different types of relations between biomedical entities can be discovered by this technique. For example, given a biomedical term and a class, one can determine whether the entity belongs to the class. Our technique is efficient and does not require any Web page download.

The paper is organized as follows. The next section cites related work.

Section 3 explains how our technique can be used to classify biomedical terms. Section 4 discusses how we can automatically discover any arbitrary relation between biomedical entities. Finally Section 5 is the conclusion.

2. Related Work

2.1. *Biomedical Information Extraction*

Our objective is to automatically discover biomedical information. Automatic extraction of useful information from online biomedical literature is a challenging problem because these documents are expressed in a natural language form.

The first task is to recognize and classify the biological entities in scientific text. Biological term extraction systems can be broadly divided into two types: those with a rule base and those with a learning method. In [3] protein names are identified in biological papers using hand-coded rules. On the other hand, in [4] supervised learning methods based on Hidden Markov Models are used. We have developed the BioAnnotator system[5] which uses rules and dictionary lookup for identifying and classifying biological terms.

After the biological entities are recognized, the next task is to identify the relations between these entities. To determine the relations between biological entities (for example protein-protein interactions), one approach is to use templates that match specific linguistic structures[6]. Natural Language processing techniques that use parsers of increasing sophistication have also been utilized. For example in [7], a bi-directional incremental parsing technique based on combinatory categorical grammar is used. Recently, research has gone beyond treatment of single sentences to look at relations that span multiple sentences through the use of co-reference[8].

Since it is very difficult to extract information from unstructured text, in this paper we introduce a completely different technique of identifying biomedical knowledge which utilizes a Web search engine.

2.2. *Knowledge Extraction from the World-wide Web*

Marti Hearst had suggested that hyponyms could be acquired from Large Text Corpora[9]. For example, consider the sentence *"The bow lute, such as the Bambara ndang, is plucked"*. Even if we have not encountered the terms *bow lute* and *Bambara ndang*, we can infer from the sentence that *Bambara ndang* is a kind of *bow lute*. Thus lexico-syntactic patterns can be utilized to discover information from a large Text corpus.

This technique has been successfully utilized to discover knowledge from the World-wide Web, the largest Text corpus. Oren Etzioni introduced the metaphor of an *Information Food Chain* where Search engines are herbivores "grazing" on the Web and intelligent agents are *"information carnivores"* that consume output from various herbivores[10].

Several systems have been built based on this principle. Instead of gathering information from the Web directly, these systems utilize Web search engines which have already crawled and indexed the information. For example, Know-it-all[11] was able to extract thousands of facts automatically using Web search engines. Similarly, PANKOW[12] could automatically discover names of countries, cities and rivers. We believe that ours is the first system that utilizes this technique to discover biomedical knowledge. Moreover, we have extended the technique to identify relations between entities.

3. Classifying Biomedical Terms

Biological knowledge sources like UMLS can be utilized to create a Semantic Web for Life Sciences by representing them using languages like RDF[13] and RDFS[14]. The biological concepts in UMLS can be represented as RDF resources and the Semantic Network classes can be represented as RDFS classes in the Semantic Web. The *RDF:type* property will link a concept to the classes it belongs to. However UMLS is not comprehensive and does not contain information about all biological terms present in the research literature. In this section we discuss a technique for determining the biomedical class of an unknown biological term so that it can be included in the Semantic Web.

3.1. *Methodology*

Marti Hearst had introduced several patterns that indicate the *"is-a"* relation in English text[9]. More patterns have been identified by others[12]. Examples of such patterns together with instances from the biomedical domain are[a]:

- *NP_term* is a *NP_class*
  ... *malaria* is a *disease*

[a]Here *NP_term* indicates the noun phrase for the term and *NP_class* indicates the noun phrase of the class

- *NP_class* such as *NP_term*
  ... *genes* such as *p53*
- *NP_term* or other *NP_class*
  ... *amylase* or other *proteins*
- *NP_class* including *NP_term*
  ... *vitamins* including *riboflavin*
- the *NP_class NP_term*
  ... the *peptide somatostatin*

If a biological term belongs to a particular class, there would be a large number of the above patterns in the World-wide Web. Thus there will be several occurrences of the phrase *"malaria is a disease"* and the phrase *"diseases including malaria"* in the Web. On the other hand there will be very few occurrences of phrases such as *"the hormone malaria"* or *"hormones such as malaria"*.

```
isa(t,c) {
  let PATTERNS be the set of patterns to determine IS-A relationships
  count = 0
  for each pattern in PATTERNS {
     queryString = pattern with NP_term replaced by t
                   and NP_class replaced by c
     resultCount = GoogleSearchResultCount(``queryString'')
     count += resultCount
  }

  if (count <= THRESHOLD)
     return false
  else return true
}
```

Figure 1. Pseudo code to classify a biological term

Based on these observations, we can determine whether a term t belongs to a class c using the procedure *isA(t,c)* as shown in Figure 1. For each pattern that indicates the *isA* relationship, we determine the number of such phrases in the WWW. We utilize the *Google APIs*[15] for searching the Web. If the total number of patterns is greater than a predefined constant (*THRESHOLD*), we consider the term to belong to the particular class.

Note that only the search result count is sufficient for our purpose; we do not need to download any Web pages. Therefore the technique is quite efficient.

3.2. *Experiments*

Although the objective of our technique is to classify terms not in the ontologies, it is not possible to evaluate the effectiveness of the classification of unknown biological terms without the help of domain experts. Therefore, we have evaluated our technique with terms that have already been classified by UMLS. We randomly selected 100 UMLS terms belonging to 10 classes that have many instances in the biomedical literature including *gene, protein, lipid, vitamin*, etc. We utilized our technique to determine whether the term belongs to some of these 10 classes.

We calculated the following statistics from our experiments:

- **True Positive (TP)**: If a term t belongs to a class c and *isa(t,c)* returns *true*.
- **True Negative (TN)**: If a term t does not belong to a class c and *isa(t,c)* returns *false*.
- **False Positive (FP)**: If a term t does not belong to a class c but *isa(t,c)* returns *true*.
- **False Negative (FN)**: If a term t belongs to a class c but *isa(t,c)* returns *false*.
- **Precision** $P = \frac{TP}{TP+FP}$
- **Recall** $R = \frac{TP}{TP+FN}$
- **F-measure** $F = \frac{2*P*R}{P+R}$

Note that we designed our experiment so that there were an equal number of positive and negative examples. Thus $TP + FN = TN + FP$. Table 1 shows the results of our experiments.

Table 1. Precision and Recall of the Classifier

Threshold	Precision	Recall	F-score
0	0.615	0.798	0.695
25	0.875	0.702	0.779
50	0.877	0.596	0.71

The best results were obtained at a *THRESHOLD* of 25 when our tech-

nique could classify biological terms with a precision of 87.5% and recall of 70.2%. At a lower threshold there were many false positives which reduced the precision. At a higher threshold there were many false negatives reducing the recall.

False Negatives occur when we find very few matching patterns for a term that belongs to a particular class. (Since at threshold 0 recall is not 100%, it indicates that in some cases not even a single matching pattern could be found). This mostly occurs for uncommon terms like *dipalmitoyllecithin* (a lipid). Moreover, some terms have many synonyms. If a synonym is not common, we may not be able to classify it. Thus we could not classify *riboflavine* as a vitamin but *riboflavin* could be correctly classified.

False Positives occur mostly because sometimes an IS-A pattern may occur in a different context. For example, the sentence "*diseases caused by viruses such as aids*" matches our IS-A pattern "*NP_class* such as *NP_term*" which indicates that aids is a virus. Patterns like "*the aids virus*" are also common.

4. Discovering Relations between Biomedical Terms

In UMLS Semantic Network the 135 biomedical classes are linked by a set of 54 semantic relationships (like *prevents*, *causes*). However there are no relationships between the biomedical concepts themselves. To develop a comprehensive Semantic Web, discovering relations between the biomedical concepts is essential. In this section we will discuss how the World-wide Web can be utilized to discover such relationships.

It should be noted that classification of biomedical terms is the determination of IS-A relation between the term and a Biomedical class. However, identifying any arbitrary relation between two biomedical entities is much more challenging. It would be very difficult to determine patterns that are true for any relation between biological terms. On the other hand, if we try to determine some particular types of relations, specifying the patterns is much easier.

Let us assume that our objective is to discover causal relationship between a disease and a biological entity. Given a disease d and a biomedical entity e, we can query Google with phrases like *"e causes d"* or *"d is caused by e"* and count the number of results that are retrieved. However, there are thousands of entities (viruses, bacteria, parasites, etc.) that can cause a disease. Querying Google for each of them is not efficient. It would be more useful if given a disease we can discover the likely causes of the disease.

```
relationIdentifier(t,patterns) {
  initialize a Hash Map resultEntities
  for each pattern in patterns {
    queryString = pattern with NP_term replaced by t
    results = GoogleSearchResultSnippets(``queryString'')
    for each result in results {
      bioAnnotatedResult = BioAnnotate(result)
      relAnnotatedResult = RelationAnnotate(bioAnnotatedResult)
      entity = relationEntity(relAnnotatedResult,t)
      resultEntities{entity}++
    }
  }

  return resultEntities
}
```

Figure 2. Pseudo code to determine the entity that has the relations specified in *patterns* with term t

We have implemented a generic framework for discovering relationships between biomedical entities. Patterns that indicate each of these relations have been identified. Figure 2 shows the pseudocode to determine the entity that takes part in relations specified by *patterns* with term t. For example, if we want to discover causal relationship between a disease and a biological entity, *patterns* may consist of phrases like *"NP_term causes"* or *"is caused by NP_term"*. In this case just the number of results retrieved by Google for the queries is not sufficient. However, downloading the result pages will make the process very slow. Therefore, we utilize the *result snippets* (the small section of the result pages that contain the query string that is returned with a Google search).

We determine the entity that is related to term t from these result snippets. For this purpose we first use BioAnnotator[5] to determine the biomedical entities in the strings. After that a Relation Annotator discovers the relations between the biomedical entities. It uses templates for patterns which specify relationships in sentences. For example some common templates are:

- *Subject Verb_Group Object* (For example, *"HIV causes AIDS"*)
- *Object Passive_Verb_Group Subject* (For example, *"AIDS is caused by HIV"*)

- *Noun (Nominal form of verb) Object Subject* (For example, *"causing of AIDS by HIV"*)

If a template is matched it is assumed that a relation of the matching verb group (or nominal form) has been identified. Note that if there are noun phrases or adjectives between the biological entities and the verb groups in the sentences they are considered as qualifiers for the biological entities.

The combination of BioAnnotator and Relation Annotator creates an annotated string from which the entity taking part in the relation with the term t can be easily identified. For example given the result snippet *"AIDS is caused by HIV"*, BioAnnotator will recognize *AIDS* and *HIV*, a Part-of-speech tagger is used to recognize *"is caused by"* as the Verb Group and the Relation Annotator recognizes *HIV* as the entity that is in causal relationship with *AIDS*. On the other hand for the more complex result snippet *"Metabolic bone disease is caused by the lack of Vitamin D3"*, the Relation Annotator recognizes *"Vitamin D3"* as the entity that is in causal relationship with *"Metabolic bone disease"* with the qualifier *"the lack of"*.

Different authors will express the same semantics in different ways. Therefore there will be variations in the snippets that are retrieved by Google. For example, one snippet may state that *AIDS* is caused by *HIV* while another may state that the disease is caused by *Human Immunodeficiency Virus*. However, BioAnnotator will map them to the same biological entity using ontologies (UMLS). Therefore, Relation Annotator will identify the same biological entity from the two snippets. However, this may not be true for all snippets. For example, if one snippet states that *Metabolic bone disease* is caused by *"the lack of Vitamin D3"* and another states that it is caused by *"Calcium deficiency"*, our annotators will not be able to match the two entities. Therefore, as shown in Figure 2, a hash map that has the entities that have the specified relation with the given concept along with the number of occurrences for each of them are returned from the *relationIdentifier* procedure.

4.1. *Experiments*

We have utilized our technique to identify various types of relationships between biomedical entities. However, a formal evaluation of our technique is difficult because there are no test data sets that can be used for the evaluation. For determining the efficiency of our technique, we determined five types of relations. Besides Semantic Network properties *causes*, *diagnoses*, *consists of* and *affects* we also extracted *binds* relations for several entities

of UMLS class *Amino Acids, Peptides or Proteins*. Table 2 shows several biomedical relations determined by our technique. Thus we could identify the cause of *Thyphoid* (*Bacterium Salmonella Typhi*) as well as entities that affect *Statin* (*Lipitor, Gemfibrozil, Niaspan*).

Table 2. Some Biomedical Relations determined by our technique

PROPERTY	UMLS ENTITY	RELATION ENTITY
causes	Typhoid	Bacterium Salmonella Typhi
diagnoses	Cyst	Ultrasonography
consists of	Butane	Liquefied Petroleum Gas
affects	Statin	Lipitor, Gemfibrozil, Niaspan
binds	Rhodopsin	Lys296, Transducin

For each property, we determined relations for several entities of some particular UMLS class which has that property. To test the system impartially we have included common as well as rare concepts in our experiments. In the absence of domain experts, we did a literature survey to determine whether the relations identified by our system are correct. We calculated the following statistics for each property from our experiments:

- **N**: Total number of biomedical entities for which we tried to identify relations.
- **F**: The number of entities for which at least one relation was identified by our system.
- **C**: The number of entities for which at least one relation that was identified by our system is correct.
- **Precision (P)** $P = \frac{C}{F}$
- **Recall (R)** $R = \frac{F}{N}$

Table 3 displays the results for each property and the corresponding UMLS class. The results show the promise of our technique, with the precision and recall values exceeding 70% for all properties. The quality of the Relation Identifier is affected by various factors:

- The recall is affected by Google's inability to identify complex class associations such as chemicals, genes, proteins and their relationships. For example, Google is unable to retrieve any results on our queries such as *"binds Auxin Response Factor 1"* or *"Nephroptosis is diagnosed by"*.

- Sometimes the snippet returned by Google may not be able to identify the cause. For example, one snippet retrieved was *"Primary Hypertension is caused by abnormalities of "* with the relevant cause of the disease stripped off.
- Both the precision and recall is affected by the limitations of the Relation Annotator. For example, the Relation Annotator can neither handle complex sentences nor relations expressed in multiple sentences.

Table 3. Precision and Recall of the Relation Identifier

Property	Class	Precision	Recall
causes	Disease	0.82	0.85
diagnoses	Anatomical Abnormality	1.0	0.9
consists of	Organic Chemical	0.75	0.72
affects	Gene	0.8	0.76
binds	Amino Acid, Peptide, or Protein	0.83	0.75

5. Conclusion

This paper introduced a new technique to automatically and efficiently discover biomedical relations. It utilizes the World-wide Web, undoubtedly one of the most comprehensive sources of biomedical knowledge. Since Web Search engines have crawled and indexed most of the information, we query these engines with several lexico-syntactic patterns to retrieve relevant information. This information can be used to classify biomedical terms or discover relations between biomedical entities. Our experiments show the promise of our techniques.

At present we are improving our Relation Annotator system for identifying relations between biomedical entities. We are also extending our system to MedLine to improve the recall. Our ultimate objective is to utilize the discovered relations between biological concepts to develop a Semantic Web for Life Sciences which would store the "meaning" of biological concepts as well as relations between these concepts. This will enable researchers to perform a single semantic search to retrieve all the relevant information about a biological concept.

References

1. UMLS. http://umlsks.nlm.nih.gov.
2. Gene Ontology. http://www.geneontology.org/.
3. K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward Information Extraction: Identifying Protein Names from Biological Papers. In *the Proceedings of the Pacific Symposium on Biocomputing*, pages 707–718, Hawaii, 1998.
4. N. Collier, C. Nobata, and J. Tsujii. Extracting the names of Genes and Gene products with a Hidden Markov Model. In *the Proceedings of the 18th International Conference on Computational Linguistics*, pages 201–207, Saarbrucken, Germany, 2000.
5. L.V. Subramaniam, S. Mukherjea, P. Kankar, B. Srivastava, V. Batra, P. Kamesam, and R. Kothari. Information Extraction from Biomedical Literature: Methodology, Evaluation and an Application. In *the Proceedings of the ACM Conference on Information and Knowledge Management*, New Orleans, Lousiana, 2003.
6. L. Wong. PIES: A Protein Interaction Extraction System. In *the Proceedings of the Pacific Symposium on Biocomputing*, pages 520–531, Hawaii, 2001.
7. J.C. Park, H.S. Kim, and J.J. Kim. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. In *the Proceedings of the Pacific Symposium on Biocomputing*, pages 396–407, Hawaii, 2001.
8. J. Pustejovski, J. Castano, J. Zhang, M. Kotecki, and B. Cochran. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit relations. In *the Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, 2002.
9. M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, July 1992.
10. O. Etzioni. Moving up the Information Food Chain: Softbots as Information Carnivores. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, Oregon, August 1996.
11. O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. Web-Scale Information Extraction in KnowItAll. In *Proceedings of the Thirteenth International World-Wide Web Conference*, New York, NY, May 2004.
12. P. Cimiano, S. Handschuh, and S. Staab. Towards the Self-Annotating Web. In *Proceedings of the Thirteenth International World-Wide Web Conference*, New York, NY, May 2004.
13. Resource Description Format. http://www.w3.org/1999/02/22-rdf-syntax-ns.
14. Resource Description Format Schema. http://www.w3.org/2000/01/rdf-schema.
15. Google APIs. http://www.google.com/apis/.