

Passage relevancy through semantic similarity

Luis Tari¹, Phan Huy Tu¹, Barry Lumpkin¹, Robert Leaman¹,
Graciela Gonzalez² and Chitta Baral¹

¹Department of Computer Science and Engineering, Arizona State University

²Department of Biomedical Informatics, Arizona State University

Abstract

Questions that require answers in the form of a list of entities and the identification of diverse biological entity classes present an interesting challenge that required new approaches not needed for the TREC 2006 Genomics track. We added some components to our automatic question answering system, including (i) a simple algorithm to select which keywords extracted from natural language questions should be treated as essential in the formation of queries, (ii) the use of different entity recognizers for various biological entity classes in the extraction of passages (iii) determining relevancy of candidate passages with the use of semantic similarity based on MeSH and UMLS semantic network. We present here an overview of our approach, with preliminary analysis and insights as to the performance of our system.

1. Introduction

The task of TREC 2007 Genomics track is to implement a system that can answer the given biological questions through the retrieval of passages from 162K full-text HTML articles. The introduction of a new set of questions (36 official topics and 14 sample topics) in TREC 2007 Genomics track poses an interesting and a significantly different challenge from the one in 2006. Unlike the template-based questions in 2006, the questions this year required answers in the form of a list of entities that belong to the requested entity classes. For instance, antibodies, which is marked by brackets, is the requested entity type of the question "What [ANTIBODIES] have been used to detect protein TLR4?". In addition, the questions cover a range of 14 different biological entity classes such as genes, proteins, mutations, diseases etc. The participants were provided with 14 sample questions together with a sample passage for each of the questions, and the goal is to build a system that can find passages for the 36 official questions based on the techniques developed in answering the sample questions.

Similar to our system for TREC 2006 Genomics track, our system this year can also be divided into three major components: preprocessing, document retrieval and passage retrieval. While the overall system is composed of the same main components, the subcomponents have been substantially revised to deal with finding answers that are in the form of list entity classes. We first provide an overview of these three major components and describe some of the innovative features of our system in detail.

2. Preprocessing

Rather than using the HTML full-text articles, we used the structured XML format of the articles that were processed by the BioSemantics group¹, which were made available for 2006 participants. Part of the preprocessing step is to resolve acronyms in the full-text articles using a popular acronym resolution algorithm [1]. The corresponding MeSH terms for each article are obtained from PubMed as well. These XML files are then indexed using both Lucene [2] and Indri [3] indexing systems. The reason behind using two indexing systems is to explore if the different rankings of the documents retrieved by the two systems would make any impact in the precision of the extracted passages.

¹ The full document collection was converted into a XML format by Martijn Schuemie from Erasmus University Medical Center at Rotterdam. It was obtained from <http://www.biosemantics.org/TREC2006>.

3. Document Retrieval

The document retrieval component involves the processing of natural language questions to form queries and expansion of biological entities identified by the question processor. A variety of techniques were used in the expansion of keywords, including one to determine the specificity of a keyword. In this section, we first describe our approach for question processing, then for keyword expansion.

3.1. Question Processing

Before queries are generated, keywords have to be extracted from the natural language questions. A fairly naïve approach was used in extracting keywords by finding the noun phrases from the questions and check against resources such as Entrez Gene and UMLS Metathesaurus to recognize their types. It is typical that different parts of the word sequences in the noun phrases can belong to different entity classes. In that case, such noun phrases are broken down into separate query terms. For instance, the noun phrase “lysosomal abnormalities” are broken down into “lysosomal” and “abnormalities”, since “lysosomal” is recognized as part of a body region and “abnormalities” as a finding according to UMLS Metathesaurus.

Other than noun phrases, verbs and dependencies of word relations are also extracted. For questions that have certain verbs such as “detect” and “measure”, the corresponding extracted passages are preferred to have mentions of experimental methods. Dependency between a noun phrase and the question requested entity type is extracted in an attempt to narrow the scope of the requested entity type. For instance, in the question “What serum [PROTEINS] change expression in association with high disease activity in lupus?”, articles assigned with the MeSH term “serum protein” are preferred.

3.2. Related Terms through Definitions

Keywords are typically expanded with their synonyms, hyponyms and hypernyms. However, definitions of terms defined in ontologies such as MeSH can be useful for the expansion of keywords. Unlike [4] which reported the use of definitions for keyword expansion decreases the performance of their system, our system does not utilize the whole definitions as expansion, but rather using only the recognized terms identified in the definitions. This can better prevent arbitrary and irrelevant terms from being used in keyword expansion.

We define term t_1 is related to term t_2 if t_1 appears in the definition of t_2 . As an example, “neurodegenerative disease” and “nervous system” are terms in the MeSH ontology. The definition of “neurodegenerative disease” is “Hereditary and sporadic conditions which are characterized by progressive nervous system dysfunction”, according to MeSH. From the definition, we infer that “nervous system” is related to “neurodegenerative disease”.

By applying this rule to the MeSH ontology, relations of terms were built and utilized for keyword of expansion. Term t_2 is used as keyword expansion of term t_1 if t_1 is related to t_2 . Using the above example, “neurodegenerative disease” is used as an expansion of the keyword “nervous system”.

3.3. Specificity of keywords

Some keywords such as “nervous system” tend to be too general to be used for queries, as relevant documents can mention entities that are related to “nervous system” rather than the actual mention of “nervous system”. For instance, a relevant document can have mentions of a specific part of the nervous system (such as neurons), or a specific disease related to nervous system (such as Parkinson’s disease). Inclusion of such general keywords in the queries can potentially miss out relevant documents. It is therefore important to be able to determine the specificity of keywords in an automatic manner.

Our approach is to determine the specificity of a keyword based on the MeSH hierarchy. The intuition of the approach is that if a keyword has more hyponyms (i.e. more specific terms)

than hypernyms (i.e. more general terms), then the keyword should be considered as general. Given the MeSH hierarchy, we define term t' is one level apart from term t if t is a parent of t' . Let t_1 , t_2 and t_3 be terms belonging to the MeSH hierarchy. Let k be the number of levels that separate t_1 from t_2 such that t_1 is an ancestor of t_2 and there is no ancestor for t_1 . Let m be the number of levels that separate t_2 from t_3 such that (i) t_2 is an ancestor of t_3 (ii) t_3 does not have any descendants (iii) there is no other descendant of t_2 that is $> m$ levels below t_2 . We define t_2 as *general* if $k < m$, otherwise t_2 is considered as *specific*.

With respect to the formation of queries, we call keywords that are required to appear in documents as *essential keywords*, while keywords that are preferred but not required as non-essential. In other words, documents are considered as relevant even in the absence of the *non-essential keywords*. Keywords that are considered as general are treated as non-essential keywords in queries, while specific keywords are used as essential keywords. Using the Lucene query syntax as an illustration, essential keywords are preceded by the “+” operator as follows:

+lysosomal +abnormalities “nervous system”

The above query indicates that a relevant document must contain the words “lysosomal” and “abnormalities”, and it is preferred but not required to have the phrase “nervous system”.

3.4. Keyword Expansion

Our system utilizes the following types of keyword expansion: synonyms, hypernyms, hyponyms, related terms through definitions (as described in section 2.2) and lexical variants. The MeSH ontology is used in expanding keywords with synonyms, hypernyms, hyponyms and related terms. For gene names, lexical variants are generated by recognizing breakpoints as in [5]. Breakpoints of a gene name are positions in which a space or a hyphen can be inserted in order to generate variants. An example of a breakpoint for the gene “Sec61” is the position in between the letter “c” and the digit “6”. With this breakpoint, variants such as “Sec-61” and “Sec 61” are generated. Another way of generating lexical variants was through the use of ADAM database [6] to get frequently occurring abbreviations.

We also utilized the fuzzy match and wildcard match features of Lucene to generate additional lexical variants of gene names other than the two approaches mentioned above. A fuzzy match is to find matches with small edit distances from the original term. This allows matches such as “MMS2” with “hMMS2” from the document collection. On the other hand, wildcard matches allow “Raf” to be matched with “Raf1” from the document collection. For fuzzy and wildcard matches of gene names, we picked the top 5 matches that had the shortest edit distances from the original term and disallowed matches that are common English words. For lexical variants of non-gene names, we only perform lemmatization to obtain their singular form.

3.5. Ranking of documents

To merge the list of documents retrieved by Lucene and Indri, we used a weighted combination of the normalized scores to score and rank documents. Let d_i be a document in a set of documents D . Let $r = \{“lucene”, “indri”\}$ to indicate the retrieval systems Lucene and Indri, and $ds_r(d_i)$ be the score for the document in which d_i is retrieved by r . Let $DS_r = \{ ds_r(d_1), \dots, ds_r(d_n) \}$, where $n = |D|$.

$$score_doc(d_i) = \alpha \times \frac{ds_{lucene}(d_i)}{\max DS_{lucene}} + (1 - \alpha) \times \frac{ds_{indri}(d_i)}{\max DS_{indri}}, \text{ where } 0 \leq \alpha \leq 1.$$

4. Passage Extraction

Candidate passages are extracted from the retrieved documents by the passage extraction component, which utilizes various entity recognizers and verifies the relevancy of candidate passages with the use of semantic similarity based on MeSH and UMLS semantic network. Certain requirements of the candidate passages have to be met before they are considered as valid passages for their corresponding questions and submitted as the final results.

A passage is defined as a contiguous list of sentences from a paragraph. In our case, we limited the maximum number of sentences in a passage to be 3. Our passage extraction component takes top- k ranked articles that are relevant to the question and retrieve sentences that have the essential keywords. We call such sentences as the *seed sentences*. Neighboring sentences of a seed sentence are merged to form a candidate passage. Various entity recognizers are utilized to tag and identify entity classes of interest in candidate passages. Semantic similarity based on the MeSH ontology and the UMLS semantic network is verified between the tagged entities in the candidate passages and the non-essential keywords in the corresponding question to determine relevancy of passages. Candidate passages that do not satisfy certain requirements are not considered as valid passages, and therefore would be filtered out from the final results. Valid passages are then ranked before submitting as final results. A passage p is *valid* with respect to question q if it satisfies the following requirements:

- (i) p contains all essential keywords of q .
- (ii) p contains a term that belongs to the requested entity type of q .
- (iii) p contains a term t such that t and non-essential keyword w of q are semantic similar.

We detail each of the main steps in extracting valid passages in this section.

4.1. Recognition of Entities

Once candidate passages are generated based on seed sentences, recognition of entities is performed for the candidate passages to identify keywords with entity classes of interest. To recognize gene or protein names, a statistical learner named BANNER [7] was used to handle the wide variation of gene names due to the frequent use of authors' preferred way of naming genes rather than using the standardized gene names [8, 9]. Mutation is another complicated domain that lacks standard convention of naming mutations. Our system utilizes MutationFinder [10], which is based on a large set of regular expressions describing mutations, to recognize mentions of mutations. In the case of antibodies, mentions of antibodies are commonly prefixed as "anti-" and keywords such as "antibody" and "serotype". We also noticed it is common to mention antibodies in the form of catalog numbers from manufacturers of antibodies. We collected the catalog numbers from the major manufacturers and represented them in regular expressions. For other entity classes, we rely on the use of MetaMap [11], which is based on comprehensive UMLS Metathesaurus and is one of the few recognizers that can identify a huge variety of entity classes in one package. We list the UMLS semantic types used in recognizing various entity classes of interest in table 1.

Entity Type	UMLS semantic types
Biological substances	Biologically Active Substance, Neuroreactive Substance or Biogenic Amine, Hormone, Enzyme, Vitamin, Immunologic Factor, Receptor, Steroid
Body parts	Body Part, Organ, or Organ Component, Body System
Cell or tissue types	Cell, Tissue
Diseases	Mental or Behavioral Dysfunction, Disease or Syndrome
Drugs	Pharmacologic Substance, Organic Chemical, Clinical Drug, Antibiotic
Experimental methods	Laboratory Procedure, Molecular Biology Research Technique
Molecular functions	Molecular Function, Genetic Function
Signs and Symptoms	Signs or Symptoms, Organism attribute, Clinical Attribute
Strains	Fungus, Virus, Rickettsia or Chlamydia, Bacterium, Archaeon
Tumor types	Neoplastic process

Table 1 - List of entity classes and their corresponding UMLS semantic types used in MetaMap

4.2. Semantic Similarity

Typical expansion techniques such as synonyms, hypernyms and hyponyms rely on the use of ontologies. However, relevant documents can be left out when general keywords as well as their expanded forms by typical expansion techniques are used as part of the queries. Terms related to such general keywords should be used for expansion instead, but expansion of general keywords at the query level can be difficult when related terms are considered. There can be thousands of terms that are related to a general keyword, but not every related term is relevant to the question. Inclusion of irrelevant terms in the process of query formation can affect the precision of document retrieval. However, verifying the relevancy of passages by checking the existence of semantic relations between the terms appearing in the passages and the general keywords in the question can avoid the inclusion of irrelevant passages. Assessing passage relatedness can improve the precision of the passage retrieval component of question answering systems.

We developed a novel and scalable method based on logic programs to find semantic similarity [12] between a pair of terms by utilizing the MeSH ontology and UMLS semantic network. Semantic similarity based on the MeSH ontology relies on the MeSH hierarchy and related terms as described in section 2.2. Let t and t' be MeSH terms. The relations are represented as logical facts, so that a is-a relation in the MeSH hierarchy is represented as $is_a_desc(t, t')$, where t is a descendant of t' . Likewise, related terms are represented as $is_mesh_related(t, t')$, in which t appears in the definition of t' . The predicate symbol $is_related(t, t')$ indicates that t is semantic similar to t' . Facts, denoted as F_{mesh} , with predicates $is_a(t, t')$ and $is_mesh_related(t, t')$ form a knowledge base and rules, denoted as R_{mesh} , are defined to infer relations. Below are some of the rules defined for R_{mesh} :

```
is_related(X, Y) :- is_a_desc(X, Y).  
is_related(X, Y) :- is_mesh_related(X, Y).  
is_related(X, Y) :- is_a_desc(X, Z), is_mesh_related(Z, Y).
```

The rules for R_{mesh} are written in Prolog style, and the symbol “:-” indicates if. The left side of the “:-” is the head of the rule and the right side of the rule is the body or the condition. The first rule means that X is semantic similar to Y if X is a descendant of Y , where X and Y are MeSH terms. The third rule indicates that X is semantic similar to Y if X is a descendant of Z and Z is related to Y through definition of Y , where X , Y and Z are MeSH terms. We say that two terms t_1 and t_2 are *semantic similar* if $F_{mesh} \cup R_{mesh} \models is_related(t_1, t_2)$.

The UMLS semantic network describes semantic relations such as causes between two semantic types. As in the approach for semantic similarity based on MeSH, logic programs are also used to find semantic similarity between two terms by utilizing the UMLS semantic network. Term t belonging to semantic type st is represented as $is_a(t, st)$. $causes(st_1, st_2)$ is used to represent the fact that semantic type st_1 causes semantic type st_2 . For instance, the semantic type “cell dysfunction” has the semantic relation “causes” with the semantic type “neoplastic process”. These facts in their logical forms (denoted as F_{UMLS}) and rules (denoted as R_{UMLS}) are used to infer relations between two terms. Some of the rules defined for R_{UMLS} are as follows:

```
is_related(X, Y) :- is_a(X, ST1), is_a(Y, ST2), causes(ST1, ST2).  
is_related(X, Y) :- is_a(X, ST1), is_a(Y, ST2), is_a(Z, ST3), produces(ST1, ST3),  
affects(ST3, ST2).
```

The first rule says that term X is semantic similar to term Y if X belongs to semantic type $ST1$ and Y belongs to semantic type $ST2$, and $ST1$ has the semantic relation “causes” with $ST2$. We say that two terms t_1 and t_2 are *semantic similar* if $F_{UMLS} \cup R_{UMLS} \models is_related(t_1, t_2)$.

4.3. Passage Ranking

Among the list of valid passages, the passages are ranked based on the following criteria: (i) keyword density (ii) section rank (iii) request entity type density. The scoring criterion for section rank (denoted as *score_origin*) is described as follows:

$$score_origin(p_i) = \begin{cases} 1 & \text{if } p_i \text{ is originated from abstract, conclusion of an article} \\ 0.5 & \text{if } p_i \text{ is originated from the title, introduction of an article} \\ 0 & \text{if } p_i \text{ is originated from the method and results section of an article} \end{cases}$$

The final score of a valid passage is simply the summation of all the four scoring criteria, and the list of passages is ranked according to the final score.

5. Results and Preliminary Analysis

We submitted 3 runs, in which 2 as automatic (denoted as *run1* and *run2*) and the other as interactive (denoted as *run3*). In an attempt of extracting passages of high passage MAP, only the top 50 documents retrieved by Lucene (i.e. $\alpha=1$ for document ranking) were used in the extraction of passages in *run1*. For *run2*, top 75 documents retrieved by Lucene and Indri with α set as 0.75 for document ranking. We give higher preference to documents retrieved by Lucene since our Indri component was in an early stage of development. *run3* is based our automatic run *run1* with the modification of the queries of 8 topics (topics 205, 206, 215, 216, 224, 229, 230, 231) after reviewing the extracted passages. Our results are listed in table 2.

	Passage2 MAP	Passage MAP	Aspect MAP	Doc MAP
All	0.0377	0.0565	0.1311	0.1897
Run 1	0.0157	0.0287	0.1302	0.0737
Run 2	0.0140	0.0351	0.1102	0.0932
Run 3	0.0268	0.0416	0.1782	0.0892

Table 2 – The median of our runs for four different measures compared with the median of all runs (denoted as “All”).

Here we list the 8 queries we modified for our interactive run, and group them in terms of methods used for their modification:

- Extra knowledge from web resources
 - Topic 205: What [SIGNS OR SYMPTOMS] of anxiety disorder are related to coronary artery disease?
 - Inclusion of the underlined terms in the query as general symptoms of anxiety disorder according to [13]: “Anxiety disorder can have massive and lasting effects on the sufferers ability to enjoy life, sufferers commonly experience symptoms such as panic attacks, phobias and many physiological anxiety disorder symptoms like shortness of breath, lethargy, insomnia and many more.”
 - Topic 206: What [TOXICITIES] are associated with zoledronic acid?
 - Inclusion of the underlined terms in the query as synonyms of zoledronic acid according to Wikipedia: “Zoledronate (zoledronic acid, marketed by Novartis under the trade names Zometa and Reclast) is a bisphosphonate.”
 - Topic 216: What [GENES] regulate puberty in humans?
 - Inclusion of seed genes that have the word “puberty” in the gene description according to Entrez Gene, such as GPR54.
 - Topic 224: What [GENES] are involved in the melanogenesis of human lung cancers?
 - Dropping the keyword “human” and inclusion of the underlined terms in the query as synonyms of melanogenesis according to Wikipedia: “Melanocytes are cells located in the bottom layer, the basal lamina, of the skin's epidermis and in the middle layer of the eye, the uvea. Through a process called melanogenesis, these cells produce melanin.”
 - Topic 229: What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?

- Inclusion of symptoms listed in [14]: rash, sore throat, slight fever, upset stomach, headache, fatigue, itching
- Topic 231: What [TUMOR TYPES] are found in zebrafish?
 - Inclusion of the underlined terms in the query according to [15]: “Although a few reports have described chemically induced zebrafish tumors (1, 2), naturally occurring tumors in zebrafish have not been identified. Chemically induced tumors include papillomas of the skin, hemangiomas, hemangiosarcomas, leiomyosarcomas, neural sheath tumors, and seminomas.”
- Lexical variants
 - Topic 230: What [PATHWAYS] are involved in Ewing's sarcoma?
 - Inclusion of lexical variants of Ewing's sarcoma in the query: ewing sarcoma, ews
- Alternative of keywords based on biological knowledge
 - Topic 215: What [PROTEINS] are involved in actin polymerization in smooth muscle?
 - Inclusion of alternatives for the keyword “polymerization”: oligomerization, repolymerization, copolymerization.

Using the gold standard passages released by TREC, we investigated the reason behind the low document MAP achieved by our runs. Rather than the document MAP after passage extraction as described in table 2, we measured document MAP based on the top 100 documents retrieved by document retrieval for each of the topics. This allows us to determine if our passage extractor plays a role in achieving low document MAP. We realized that the document MAP is 0.1509 for our first run (*run1*) and 0.1204 for our second run (*run2*). This suggests that we should perhaps increase the number of documents used in passage extraction and take document rank as part of the rank of passages.

From our preliminary analysis, we also noticed that the semantic types we used have been too restrictive. For instance in topic 206, the requested entity type is “toxicity” and gold standard passages of this topic include osteronecrosis and fatigue. In the future, we should expand the way we search for entity classes to include related entity classes. We also noticed that our system performed above the median for 5 of the 36 topics (topics 201, 202, 204, 205, 219) that include general keywords. This suggests a positive effect of performance for our semantic similarity component. A thorough error analysis is needed to see how the performance of the entity recognizers affects our passage extraction component. We will detail the analysis of the impact of each of the components used in our system in the TREC final paper.

Acknowledgement

We thank Santa Cruz Biotechnology Inc. for kindly providing their list of catalog numbers for the antibodies they manufacture.

References

1. Schwartz A, Hearst M: **A simple algorithm for identifying abbreviation definitions in biomedical texts**. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)* 2003, **8**:451-462.
2. Lucene [<http://lucene.apache.org/java/docs/>]
3. Metzler D, Croft WB: **Combining the Language Model and Inference Network Approaches to Retrieval**. *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval* 2004, **40**(5):735-750.
4. Hersh W, Price S, L LD: **Assessing thesaurus-based query expansion using the UMLS metathesaurus**. *Journal of the American Medical Informatics Association* 2000:334-348.
5. Huang X, Zhong M, Si L: **York University at TREC 2005 Genomics Track**.
6. Zhou W, Torvik VI, Smalheiser NR: **ADAM: Another Database of Abbreviations in MEDLINE**. *Bioinformatics* 2006, **22**(2):2813-2818.
7. Leaman R, Gonzalez G: **BANNER: An executable survey of advances in biomedical named entity recognition**. *Pacific Symposium of Biocomputing (PSB) 2008*:(to be published).

8. Chen L, Liu H, Friedman C: **Gene name ambiguity of eukaryotic nomenclatures.** *Bioinformatics* 2005, **21**:248-255.
9. Wilbur J, Smith L, Tanabe T: **BioCreative 2 Gene Mention Task.** . *Proceedings of the Second BioCreative Challenge Workshop 2007*, **7-16**.
10. Caporaso JG, Baumgartner WA, Jr., Randolph DA, Cohen KB, Hunter L: **MutationFinder: a high-performance system for extracting point mutation mentions from text.** *Bioinformatics* 2007, **23**(14):1862-1865.
11. **MetaMap: Mapping Text to the UMLS Metathesaurus**
[<http://skr.nlm.nih.gov/papers/references/metamap06.pdf>]
12. Budanitsky A, Hirst G: **Evaluating WordNet-based measures of semantic distance.** *Computational Linguistics* 2006, **32**(1):13-47.
13. **Anxiety disorder symptoms** [<http://www.panic-anxiety.com/anxiety-disorder-symptoms.htm>]
14. **Parvovirus infection: Signs and Symptoms - mayoclinic.com**
[<http://www.mayoclinic.com/health/parvovirus-infection/DS00437/DSECTION=2>]
15. Smolowitz R: **A Three-Year Retrospective Study of Abdominal Tumors in Zebrafish Maintained in an Aquatic Laboratory Animal Facility.** *Biol Bull* 2002, **203**:265–266.