Semantics-based Text Mining of Biomedical Concepts in Scientific Publications

CS224 Final Project Report, June 4, 2010

Saeed Hassanpour
Stanford University, Stanford, CA 94305 USA
saeedhp@stanford.edu

Siddharth Taduri
Stanford University, Stanford, CA 94305 USA
staduri@stanford.edu

ABSTRACT

Searching publications for prior work on scientific concepts is central to the research process. The relevant parts of retrieved publications are typically found and evaluated manually. In the field of biomedicine, due to rapidly growing numbers of publications and the of lack standard scientific terminologies, this task is particularly challenging, complex and time consuming. Prior information retrieval methods, using term match and term frequency, have been developed to find sections of document text that contain the definitions and associations of scientific concepts being studied. In this work, we present a novel method to extract parts of publications that are most relevant to biomedical concepts that have been defined by domain ontologies and rules. We use hierarchal clustering to identify the parts of the texts that contain the most relevant terms to the concepts and used vector space modeling to extract these texts. We then use cosine similarity to compute the texts correlation to the concepts, rank them and return the most relevant ones. We have applied our method to a knowledge base of autism phenotype definitions, which are modeled using the web ontology language, OWL, and its rule language, SWRL. We compared the accuracy and relevance of three levels of semantics using the ontology hierarchy, the rule definition, or both. Our results showed that the ontology hierarchy provided the highest accuracy (73%) in finding text that defined the domain concept, whereas a combined ontology hierarchy and rule definition approach provided the highest accuracy (76%) in retrieving any text that referred to the domain concept. The results indicate that a semantics-based text mining approach can be useful in speeding the process of reviewing definitions of and associations with biomedical concepts.

Keywords

Natural Language Processing, Information Retrieval, Text Mining, Semantic Similarity, Ontology, Rules, OWL, SWRL

1. INTRODUCTION

The number of scientific publications is increasing by an incredibly fast rate. Scientists, as a part of their research, need to keep up with this rapid growth of knowledge and be updated in their fields of research. Search engines for scientific publication repositories and general-purpose web search engines can help users to find the most relevant publications to the search concepts and search terms. However, users still need to review the publications and confirm that they are relevant to their search. As part of this process, they need to find the parts within the publications relevant to the concept being searched. Considering the number of publications and the lack of standard scientific terminologies, a large amount of time and effort is spent on publication review. One solution to this problem is an automatic method that finds the most relevant parts of a document to the required concepts. Prior methods have been developed to perform this task, mainly based on term matching and term frequencies, but perform poorly in finding semantically implicit parts of a text relevant to a concept. In our work, we present a novel text mining method that uses domain ontologies and formal definitions of concepts rules to discover semantically related document text and retrieve the most accurate and informative results.

This work is motivated by the needs of developers of an ontology of autism phenotypes who want to automatically find publications that define the domain concepts and find text that relates the concepts to other findings [23]. The ontology contains both an information model that represents research or clinical data collected through standardized instruments and a domain ontology that defines terms and relationships among nine major categories of autism phenotypes, such as language, social interaction, and behavioral abnormalities. The clinical domain knowledge for encoding phenotypes was initially encoded by domain experts who manually reviewed articles in PubMed on clinical phenotypes [24, 25]. The domain experts then defined classes, properties and rules in the ontology that encode the 156 unique phenotype concepts they found.

The autism ontology is specified using the Web Ontology Language (OWL). OWL is the World Wide Web Consortium (W3C) standard for developing ontologies for the Semantic Web [21], and it is the most common language among large scientific communities for developing ontologies. OWL ontologies contain hierarchies of classes and properties, which model the domain concept abstractions and their relationships. OWL Individual presents members of OWL classes and they are the real world examples of the ontology abstractions. Rules within the autism ontology are specified using the Semantic Web Rule Language (SWRL), a horn-like language extension to OWL for specifying rules [22]. A SWRL rule is the logical conjunction of unary and binary predicates where the unary predicates are OWL classes and the binary predicates are OWL properties. SWRL also provides built-in predicates to provide mathematical and logical functionalities. The arguments in a predicate refer to individuals in the ontology or data values. SWRL thus provides a powerful way to formally define concepts, such as autism phenotypes, in terms of logical relationships to classes and properties in an OWL ontology.

As an example the following SWRL rule defines the autism phenotype concept of “no delayed development in word acquisition” based on items on Autism Diagnostic Interview-Revised (ADI-R) instrument [25]. This rule indicates that if a word is acquired at the age of 24 months or earlier, then there is

no delayed development in the word acquisition and asserts this finding in the record of the subject who completed the ADI-R questionnaire.

```
ADI-2003(?a) ^
adi-r2003:ADI_2003_acqorlossoflang_aword(?a, ?wordage) ^
swrlb:lessThanOrEqual(?wordage, 24) ^
adi-r2003:SubjectKey(?a, ?subjectID) ^
adi-r2003:ADI_2003_interview_date(?a, ?date) ^
swrlx:createOWLThing(?phenorecord, ?subjectID) →
'Phenotype record'(?phenorecord) ^
temporal:hasValidTime(?phenorecord, ?date) ^
autism-core:is_derived_from(?phenorecord,
"ConcludeNotDelayedWord") ^
autism-core:subjectId(?phenorecord, ?subjectID) ^
autism-
core:subject_has_quality_or_disposition(?phenorecord,
'Not delayed word')
```

The goal of our text-mining method is to find document text within a retrieved publication that relates to concept definitions based on the ontology hierarchy and rule definition. For example, the definition of "no delayed development in word acquisition" based on the example SWRL rule would correspond to the highlighted section of a publication [23], as shown in Figure 1. In this paper, we present our method and evaluate its accuracy in text retrieval using three different levels of semantics from the ontology.

(MZ) twins for RRSBs on the ADI but found familial clustering of Nonverbal IQ and verbal/nonverbal status. Nevertheless, markedly different levels of impairment were common. In contrast, Kolevzon et al (2004) found that ADI-R Communication and Social domains showed significantly decreased variance within MZ twins compared with other sibships. Szatmari et al (1996) found concordance in IQ and level of adaptive functioning in affected siblings and then, using a larger sample (MacLean et al 1999), reported a moderate degree of family resemblance for Nonverbal IQ and social and communicative adaptation, with Nonverbal Communication and verbal–nonverbal status being the only ADI/ADI-R measures to show familial aggregation.

Stratifying Samples by Language Acquisition

One construct commonly used to stratify samples is age of language acquisition, based on age of first words or phrases. Delayed language is defined on the ADI-R by age of first words ≥ 24 and age of first phrases ≥ 33–36 months. As shown in Table 1, several research groups have found increased logarithm of the odds (LOD) scores for various chromosomal areas using subsamples of families with phrase speech delay (PSD; Bradford et al 2001; Buxbaum et al 2001; Shao et al 2002). Wassink et al (2004) found evidence for linkage in the nondelayed, but not the PSD, subsample. Although each research group used PSD as a stratification variable to form subgroups, their initial samples varied by study. For example, Shao and colleagues' (2002) sample was restricted to families with children with autism, whereas Buxbaum et al's (2001) sample included families who met less stringent criteria for an ASD. Furthermore, Bradford et al's (2001) findings only emerged when they also incorporated a history of language-related difficulties in the parents.

Identifying Quantitative Trait Loci

Rather than stratifying families based on probands' characteristics, some researchers have attempted to identify loci that affect

Figure 1. The highlighted text is the definition of the concept "no delayed development in word acquisition" from a scientific publication [23] and the most relevant part of the publication to the concept

2. RELATED WORK

In general, a document or a piece of text can be summarized in different ways based on different readers' points of view. In scientific publications, abstracts are meant to summarize the documents. However, abstracts may not provide relevant information for all of the user's needs. Text summarization has been part of natural language processing and information retrieval fields for many years. The early work in this field goes back to 1958 by IBM [10].

The focus of text summarization is to extract the most relevant sentences from a text based on user queries. Different methods have been developed for this purpose. Statistical methods are developed to remove the redundancies from texts, so core parts of text stand out [11-12]. Some other statistical methods rely on the frequency of occurrence of the user query terms to outline the most relevant portions of the text. Reeve et al. proposed the use of domain concept frequency to identify important parts of a text [14]. Among statistical text summarizers, Open Text Summarizer (OTS) is an open source text summarizer, which is widely being used and is usually being considered for benchmarking other text summarizers [13]. OTS is based on extracting the most frequent terms in a text and returns the sentences that cover these terms. Some other statistical methods use certain linguistic features based on the argumentative structure of the texts to summarize them [16]. Although frequency based methods are efficient in dealing with a large number of texts, their accuracy is restricted by the representativeness of the presented key terms for each concept. Often it is difficult to capture semantics of a concept in from of several key terms.

Lexical chaining is another method to extract theme and focus of a document. Lexical chains are sequences of related words that do not necessarily follow a grammatical structure [14]. In this method every noun word is considered as a chain element candidate, and the set sets of lexical chains are formed and maintained from these candidates based on their sense. This method is computationally expensive due to the large size of the chaining possibilities, however several efficient methods are presented to compute lexical chains [8, 18-19].

In particular, in the field of biomedicine, due to the vast size of the knowledge domain and the lack of standard terminologies, the general statistical and lexical chaining methods are not very functional. As a solution ontologies have been used as controlled terminologies in statistical summarization and chaining methods for specific domains [4, 7, 9, 15, 20]. Morales et al. used graph-based approach to summarize biomedical literature based on the domain ontologies. In this method, every document is represented as a graph where nodes are representing concepts from the domain ontology, and edges represent the relations between them [2]. In a different approach, Ruch et al. extracted the key sentences from MEDLINE abstracts by training a Bayesian classifier on biomedical literature to classify the text in four argumentative categories: Purpose, Methods, Results and Conclusion [3].

3. METHODS

To find the most relevant parts of science publications to domain concepts, we have chosen to use OWL ontologies and SWRL rules because they provide formal definitions of domain concepts and their relationships to other concepts. Using this knowledge about domain concepts, we seek to compute the correlation between an arbitrary section of text and a particular concept. To

be able to compare concepts and texts, we use vector space modeling as the mathematical presentation of texts and concepts. This vector space modeling captures the semantic similarity of a concept and its relevant concepts in the domain. We use the cosine similarity measure to quantify the correlation between a part of a papers and a concept. We explain our modeling method for concepts and texts and the correlation computation in Sections 3.1 and 3.2.

In searching for parts of a publication relevant to a concept we apply heuristic techniques to prune the search domain and restrict it to potential candidates with desirable properties. These heuristic techniques include restrictions on the length of the relevant parts. We also merge relevant parts of texts that are located close to each other in the document using hierarchal clustering. We present our searching method in detail in Section 3.3, and discuss our evaluation strategy in Section 3.4.

3.1 Vector Space Modeling

Our first step is to quantify the relevance between concepts and pieces of text. As a result, we need a mathematical modeling of concepts and texts to provide a common basis for comparison. For this purpose we chose vector space modeling, a common method in the web search engines for indexing web pages [26].

We use the standard vector space modeling to model the publication texts, which is based on term frequencies. To model a part of a text as a vector, we first removed the stop words, the most common English words that are not informative about the context. We used a common list of 571 stop words in English [27]. Then we applied Porter stemming algorithm, a very common stemming method for English terms [28], to replace different derivations of a word with their root. Then we built 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.

A similar technique is used to model a concept as a vector. We use a specific 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 add one dimension for each class and property mentioned in the rule as relevant concepts.

Besides the concepts mentioned in the rule, we use ontology hierarchies to extract more related concepts. Figure 2 shows a part of the autism class hierarchy. We consider the parents and grandparents of the main concept and its related concepts 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.

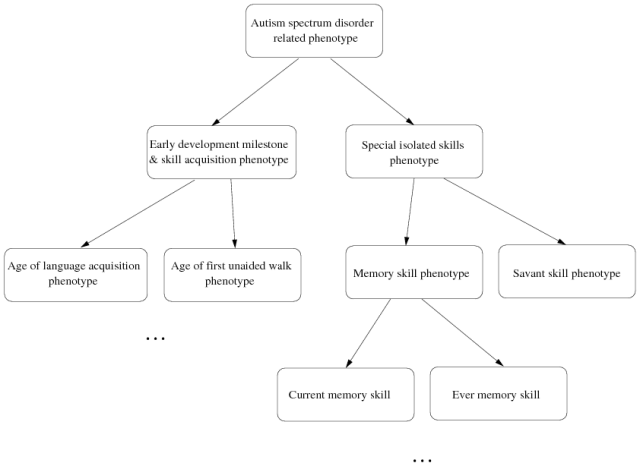

Figure 2. A part of the class hierarchy from the autism ontology

Usually classes and properties have associated metadata in OWL ontologies. These metadata are in form of RDF labels and comments in OWL. These are informative terms or explanations of classes and properties, and are usually added by ontology developers as documentation. We included this metadata in vector space modeling in the same way that we modeled texts and with the same weight as the OWL classes' and properties' that they are attached to.

3.2 Correlation Computation

After we have created a method to present both text words and domain concepts as vectors, we needed 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. The mathematical formula for cosine similarity is:

$$Similarity(a,b) = \cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|}$$

Where a and b are two vectors in the Euclidean space, θ is the angle between them, $a.b$ is their dot product, and $||a||$ and $||b||$ are the magnitudes of the vectors. Since cosine similarity measure normalizes vectors, it does not depend on the size of the vectors and gives an accurate and stable measure of similarity for any vector space dimension.

3.3 Relevant Text Search

Once we compute the correlation between any arbitrary text and a concept, we go through a publication and search for the most relevant parts of the text. Systematic search is not an efficient approach and it does not find informative results. Since extending a piece of text to a longer text does not adversely effect its correlation to a concept, the systematic search eventually will

return the whole publication as the most relevant piece of text to any concept, which is not a desirable result. To tackle the searching problem in publications we define restrictions on the domain to find the most informative candidates efficiently.

Our first step is to search for the most relevant part of the text for a particular concept. We looked 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 went through the publication and marked all the occurrence of the concept terms in the text. We cover occurrences of different forms of a concept terms by applying stemming on both concept terms and publication terms.

After we found the occurrence 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. In general the average English sentence length is between 15 and 20 [26]. In the single linkage clustering we use 30 as a heuristic threshold and in every step we merged the closest clusters that are separated by less than 30 words. Thus, we ensure that a continuous section o text without any concept term is limited to a few sentences and the whole cluster is continuously correlated to the concept.

After finding the clusters, we filter out the clusters with less than 4 words to make sure that the clusters contain complete meaningful phrases or sentences rather than a few words that do not convey any information. We consider the remaining clusters as the candidates for the most relevant parts of the text. We then compute the cosine similarity of each candidate and the concept and ranked them based on their correlations and return the five most relevant candidates as the results. As an example for the "no delayed development in word acquisition" concept that mentioned earlier our method found the actual concept definition highlighted in Figure 1 from the corresponding publication [23] as the most relevant part of the text.

3.4 Evaluation

In this work, we applied our method to the autism phenotype ontology and the papers used to derive those concepts as mentioned in Section 1. We focused our validation on complex domain concepts that had a rule that specified a non-one-to-one mapping between the results of a study instrument. We applied our method on each these domain concept and its related publications. We returned the top five most relevant parts of the publication for the concepts.

Furthermore, to investigate the significance of using both ontological hierarchies and rule bases we applied and compared three variations of our method. In the first experiment we just considered the rules for the concept modeling without incorporating the related concepts and their metadata from the ontology. In the second variations we only considered hierarchal structures and the metadata for each concept without including the concept definitions from the rule base. In the last variation we used the complete information from both the ontology hierarchal structures and metadata, and the rule base as it was described in the method section. The most relevant parts of the texts were reviewed by one of the autism ontology developers. To eliminate bias in the assessment of the performance of the three variations, the corresponding variation used for each set of results was not revealed to the reviewer during the evaluation process.

4. RESULTS

The autism ontology contains 1726 classes and properties, and it includes 156 SWRL rules. Our selection criteria for complex domain concepts and corresponding rules returned 49 rules. These rules were referenced in the ontology to be present in 7 of the 26 publications used to create the autism ontology. We thus applied our method on each of the 49 concepts and the seven related publications, and we returned the top 5 most relevant parts of the publication for review by the domain expert. Altogether 735 sections of text were reviewed and evaluated as to whether they were relevant to the corresponding concepts. Table 1 shows the accuracy, for each variation of our method, whether the returned section referred to the concept. In this case every concept was defined in the corresponding publication.

Table 1. The precision of three different levels of semantics in finding texts relevant to concepts

Level of Semantics	Precision
Only rules	62%
Only ontology hierarchies	73%
Both rules and ontology hierarchies	76%

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 contains a clear definition. Table 2 shows the accuracy of the different variations of our method in finding the definition of the concepts in the corresponding publication text.

Table 2. The precision of three different levels of semantics in identifying concept definition in the publication text

Level of Semantics	Precision
Only rules	57%
Only ontology hierarchies	73%
Both rules and ontology hierarchies	69%

In the ideal case we expect that the concept definitions will appear on the top of the list of the most relevant texts to a concept. For the concept definitions found among the top five most relevant parts of texts, we extracted their rank, ranging from 1 to 5 and compute the average concept definition rank. Table 3 shows the average concept definition average rank for our three different methods.

Table 3. The average rank of the concept definitions among the five most relevant texts returned by three different levels of semantics

Level of Semantics	Average Rank
Only rules	2.75
Only ontology hierarchies	2.42
Both rules and ontology hierarchies	2.09

We also looked at the average length of the text that each semantics level returned as the relevant texts for these concepts. As it is shown in figure 3, the average length of the relevant texts for only rules and only ontology methods were approximately 70 words. However, for the method that used both rules and ontology, the average result length was about 140 words.

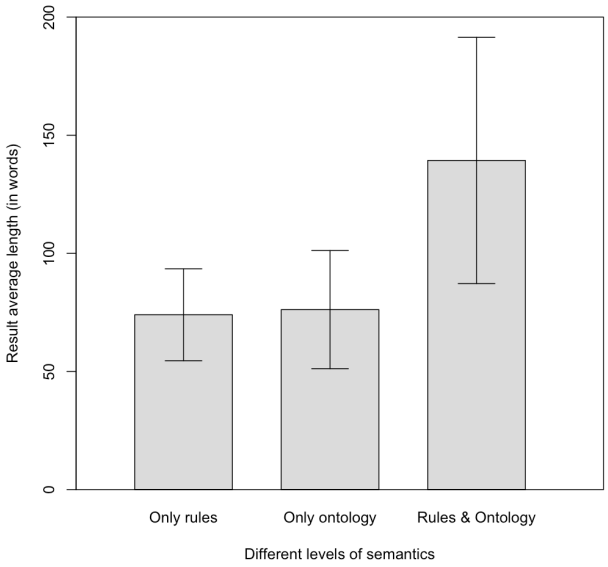

Figure 3. The average length of the returned relevant texts for 49 concepts in words for three levels of semantics

5. DISCUSSION

In this paper, we have presented a novel method to find the relevant parts of publications for biomedical concepts. Biomedical terms usually have overlapping definitions, borders and relationships with different concepts. Very often different terms and terminologies are used to refer to similar concepts in the field. These characteristics of the domain make it difficult for general statistical text extraction and summarization to be precise and accurate in biomedicine.

To tackle this problem, we have used domain ontologies and rule bases to clarify the semantics among terms. Rule bases present clear definitions of the concepts in a formal form. We used ontologies to capture semantically related concepts in the hierarchy and incorporated semantic similarity through weighted terms in the search for the relevant parts of texts. These ontology-based relationships between domain concepts and definitions provide advantages in capturing semantics-based relationships, which are missing in purely statistical methods.

Our evaluation showed that ontology hierarchies and metadata have a bigger impact in identifying the relevant parts of the text than the rules. We believe this might be because of the informative taxonomies and strong relationships between the concepts in this field captured in the domain ontology. However, our evaluation showed that the relevance of the results to the concepts is enhanced by adding the information from the rule base to the ontology hierarchy and metadata information in the concept modeling process. Our evaluation also found that using ontologies has the highest accuracy in finding the concept definitions among the top five most relevant parts of texts. Rules may include many concepts that are directly related to the definition such as asserting a finding for a subject, as shown in the example rule in Section 1. Such assertions might reduce the degree of the relevance of the top five results in the case that we consider the rules in the concept modeling.

We consider the concept definition as the most relevant part of a publication to a concept. We observed in all variations of our method the concept definition text is ranked in the top half of the 5 returned results. We also observed that the average rank for the concept definition text is slightly better when we used both rules and anthologies, which indicates the positive effect of rules in boosting the relevance of the concepts and the returning concept definitions.

To search for the relevant pieces of text in a publication we used hierarchal clustering to group the related terms to a concept as continues pieces of texts. Clustering limits the search domain and makes the search process very efficient. Using the combination of ontology and rules increases the concept terms, which increases the size of the continuous texts, which cover concept terms. This leads to longer relevant text results. As it is shown in figure 3 the average size of the returned texts in the method that used both rules and ontologies was twice of the size of the average results in the variations that used either of them.

6. FUTURE WORK

In addition to incorporating semantics from domain ontologies and rule bases, we are planning to use the text's syntactic structures through constituent and dependency parsing methods on the publications. 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 have richer presentation of texts and find relationship between the publication text and the queried concept in the field of biomedicine.

We are planning to provide our text-mining methods as a web service in the Phenologue project, an online community resource for cataloging, inserting, and editing ontology-based phenotypes. In our web service, we will provide the facility to upload ontologies, rule bases and scientific publications. Then the user can query the system by different concepts and our method use the uploaded ontologies and rule bases to find and extract the most relevant parts of the publications to the queried concepts. We propose to combine this service with other scientific literature search engines, thus giving scientists an array of tools to discover and review the most relevant knowledge published on scientific concepts in their field.

7. ACKNOWLEDGMENTS

The authors would like to thank Amar Das for his help on the method evaluation, Martin O'Conner, Samson W. Tu, Lakshika Tennakoon, Richard Waldinger and Joachim Hallmayer for their help on the ontology development and Christopher Manning and Noah Zimmerman for their comments on this work.

8. REFERENCES

[1] Afantenos, S., Karkaletsis, V., and Stamatopoulos, P. 2003. Summarization from medical documents: a survey. Artificial Intelligence in Medicine, Volume 33, Issue 2. Pages 157-177.

[2] Morales, L. P., Esteban, A. D., and Gervás, P. 2008. Concept-graph based biomedical automatic summarization using ontologies. In Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms For Natural Language Processing. August, 2008. 53-56.

[3] Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbuhler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., and Veuthey, A. 2007. Using argumentation to extract key sentences from biomedical abstracts, International Journal of Medical Informatics, Volume 76, Issues 2-3. Connecting Medical Informatics and Bio-Informatics - MIE 2005. February-March 2007, Pages 195-200.

[4] Ling, X., Jiang J., He, X., Mei Q., Zhai, C., and Schatz, B. 2007. Generating gene summaries from biomedical literature: A study of semi-structured summarization. Information Processing & Management, Volume 43, Issue 6. November 2007. Pages 1777-1791.

[5] Rodriguez-Esteban, R. 2009. Biomedical Text Mining and Its Applications. PLoS Computational Biology 5(12).

[6] Ananiadou, S., Kell, D., B., and Tsujii, J. 2006. Text mining and its potential applications in systems biology. Trends in Biotechnology. Volume 24, Issue 12, December 2006, Pages 571-579.

[7] Reeve, L. H., Han, H., and Brooks, A. D. 2007. The use of domain-specific concepts in biomedical text summarization. Information Processing & Management, Volume 43, Issue 6, Text Summarization, November 2007, Pages 1765-1776.

[8] Silber, H., G., and McCoy, K. F. 2002. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. Computational Linguistics 2002 28:4, 487-496 .

[9] Reeve, L. H., Han, H., Nagori, S. V., Yang, J. C., Schwimmer, T. A., and Brooks, A. D. 2006. Concept frequency distribution in biomedical text summarization. In Proceedings of the 15th ACM international Conference on information and Knowledge Management. November 2006). 604-611.

[10] Luhn, H. P. 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 2 (Apr. 1958), 159-165.

[11] Nenkova, A., and Vanderwende, L. 2005. The impact of frequency on summarization. MSR-TR-2005-101. Redmond, Washington: Microsoft Research.

[12] Carbonell, J., and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval. August 1998. ACM, New York, NY, 335-336.

[13] Rotem, N. 2003. Open text summarizer (OTS). http://libots.sourceforge.net

[14] Wikipedia: http://en.wikipedia.org/wiki/Lexical_chain

[15] Hu, X. 2004. Integration of Cluster Ensemble and Text Summarization for Gene Expression Analysis. Bioinformatics and Bioengineering. IEEE International Symposium on, p. 251.

[16] Neto, J. L, Freitas A. A., and Kaestner., A. A. 2002. Automatic text summarization using a machine learning approach. Advances in Artificial Intelligence. p. 205-215.

[17] Gonzàlez, E., and Fuentes, M. 2009. A New Lexical Chain Algorithm Used for Automatic Summarization. In Proceeding of the 2009 Conference on Artificial intelligence Research and Development: Proceedings of the 12th international Conference of the Catalan Association For Artificial intelligence S. Sandri, M. Sànchez-Marrè, and U. Cortés, Eds. Frontiers in Artificial Intelligence and Applications, vol. 202. IOS Press, Amsterdam, The Netherlands, 329-338.

[18] Silber, H. G., and McCoy, K. F. 2000. Efficient text summarization using lexical chains. In Proceedings of the 5th international Conference on intelligent User interfaces (New Orleans, Louisiana, United States, January 09 - 12, 2000). IUI '00. ACM, New York, NY, 252-255.

[19] Barzilay, R., and Elhadad, M. 1997. Using Lexical Chains for Text Summarization.

[20] Reeve, L., Han, H., and Brooks, A. D. 2006. BioChain: lexical chaining methods for biomedical text summarization. In Proceedings of the 2006 ACM Symposium on Applied Computing. April 2006. 180-184.

[21] McGuinness, D. L., van Harmelen, F., eds. 2004. OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/

[22] SWRL Submission: http://www.w3.org/Submission/SWRL/

[23] Hus V., Pickles A., Cook E. H., Risi S., Lord C. 2007. Using the Autism Diagnostic Interview-Revised to Increase Phenotypic Homogeneity in Genetic Studies of Autism, Biological Psychiatry, Volume 61, Issue 4, Advances in Understanding and Treating Autism Spectrum Disorders, 15 February 2007, Pages 438-448.

[24] Young L., Tu S. W., Tennakoon L., et al. 2009. Ontology-driven data integration for autism research. 22nd IEEE International Symposium on Computer Based Medical Systems, Albuquerque, NM, 2009;1-7.

[25] Tu S., Tennakoon, L., Das A. 2008. Using an integrated ontology and information model for querying and reasoning about phenotypes: the case of autism, American Medical Informatics Association(AMIA) Annual Symposium, Washington, DC 2008;727-731.

[26] Manning CD, Raghavan P, Schüze H. 2008. Introduction to Information Retrieval, Cambridge University Press 2008.

[27] List of English stopwords: http://members.unine.ch/jacques.savoy/clef/

[28] Porter stemmer: http://tartarus.org/~martin/PorterStemmer