Automatic Model Structuring from Text using BioMedical Ontology

Rohit Joshi[*], Xiaoli Li [*§], Sreeram Ramachandaran[*], Tze Yun Leong [*§]

* Medical Computing Laboratory,
School of Computing,
National University of Singapore,
3, Science Drive 2, Singapore 117543
{dcsrj, dcssr, dcslty}@nus.edu.sg

§ Singapore-MIT Alliance, Computer Science
National University of Singapore,
3, Science Drive 2, Singapore 117543
smalxl@nus.edu.sg

Abstract

Bayesian Networks and Influence Diagrams are effective methods for structuring clinical problems. Constructing a relevant structure without the numerical probabilities in itself is a challenging task. In addition, due to the rapid rate of innovations and new findings in the biomedical domain, constructing a relevant graphical model becomes even more challenging. Building a model structure from text with minimum intervention from domain experts and minimum training examples has always been a challenge for the researchers. In the biomedical domain, numerous advances have been made which may make this dream a possibility now. We are currently trying to build a general purpose system to automatically extract the model structure from scientific articles using a combination of ontological knowledge and data mining with natural language processing. This paper discusses the prototype system that we are working on. Previously, systems have used keyword features to extract knowledge from text. We, like Blake et al [4], argue that the choice of features used to represent a domain has a profound effect on the quality of model produced. Our system uses concepts and semantic types rather than keywords. We map complete sentences in the medical text to a conceptual level and a semantic level. We then, use Association Rule Mining (ARM) to extract relationships from text. Rules are then filtered and verified to improve precision of the obtained rules. Preliminary results applied to Colorectal Cancer medical domain are presented, which suggest the feasibility of our approach.

Introduction

Bayesian Networks and Influence Diagrams are effective methods for structuring clinical problems, encoding objective evidence, representing a clinician's subjective judgments and expressing a patient's preferences to derive optimizing solutions in diagnostic, therapeutic, and prognostic management. However, it is usually an arduous task to process and integrate all the knowledge needed for model construction. Constructing a relevant structure without the numerical probabilities in itself is a challenging task. In addition, due to the rapid rate of innovations and new findings in the biomedical domain, domain experts may not always be up- to-date on all the latest advancements in a particular field, so constructing a relevant graphical structure becomes even more challenging. Such information is best captured in the scientific literature. To build a Bayesian network or an influence diagram automatically from the text with minimum intervention of a domain expert has been a dream for many researchers. In many fields, it is not possible yet. However, in recent years, in the biomedical domain, we have seen emergence of large repositories such as MEDLINE [1] that index over 12 million citations, ontology such as the Unified Medical Language System [2] that have over 800,000 medical concepts, and various tools to aid in extraction of knowledge from text, all of which may help to bring this dream closer to reality now. We believe that the complete automation of model construction might not be yet possible, however with the use of biomedical ontology, a model can still be derived that can act as an starting point for the clinicians. We are currently working on a general purpose system that uses a combination of biomedical ontology and data mining techniques with Natural Language Processing (NLP) to model structure from the scientific articles. In the past, various works have tried to extract knowledge from the medical literature; however they have normally been specific to a domain such as radiology or pathology. Moreover, they use either hand-crafted pre-defined patterns or specific training examples, which were produced with huge effort and time spent by domain experts. We are currently developing a general purpose system with minimum training examples to address this problem. Since our aim is to use minimum training examples, many knowledge extraction or relationship extraction methods developed in Information extraction and AI communities become inapplicable. Moreover, NLP based techniques are usually used for a specific task such as sub-cellular localization [17]. However, Bayesian Network structure can change according to the clinical problem.

In this paper, we present a prototype system that we are building to model structure from the scientific articles. We use MEDLINE keyword terms and Unified Medical language System (UMLS) to aid in automatically extracting facts and relations from the scientific literature. Our system needs a semantic template as input. The system then downloads the relevant scientific papers using a search engine. It uses a novel document classification technique to classify documents without

labeled training examples. Our Text Summarizer selects the best sentences in text to extract relevant relationships. We, then, map the sentences into concepts and concepts to semantic level and use Association rule mining techniques to explore relationships between the concepts. Derived rules are then filtered and verified using a full syntactic parse of sentences and NLP techniques. Obtained rules, after filtering and verification, are transformed into the required model. Preliminary results have been presented using Colorectal Cancer medical domain. We believe that our method is general enough to be applied to other bio-medical domains. However, no results have been presented for this claim.

Problem Formulation

Given skeleton templates such as those in Probabilistic Relational Models (PRM), how do we discover model structure from text? Specifically, how can we generate all the instantiations of the attributes of the nodes from the available scientific literature? Such a model can be used as an aid to build Bayesian networks or influence diagrams in a clinical setting. We do not want to elicit conditional probabilities from text. Rather, we limit ourselves to discovering only the nodes or the structure. We illustrate the problem in figure below:

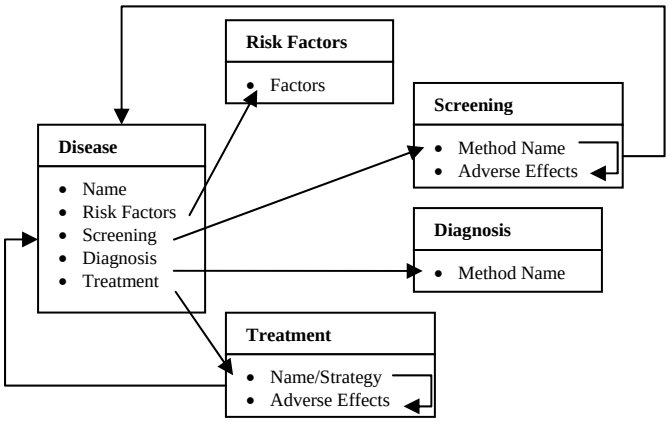

Figure 1: Semantic template as in PRMs

From the diagram, the questions that we would like to get answers automatically for from the text are

- What are the risk factors of a disease?
- What are the screening methods for a given disease?
- Do these screening methods have side effects? If so, what are the harmful effects?
- Which methods can we use to diagnose this disease?
- What are the available treatment strategies?
- What are their adverse effects?

This is a kind of knowledge extraction task, which has normally been best addressed using Natural Language Processing and Information Extraction (IE). However, NLP and IE techniques use either hand-crafted pre-defined patterns or specific training examples, which were produced with huge effort and time spent by domain experts. Moreover, NLP methods are usually trained for a very specific task such as disease and treatment. In our scenario, the skeleton template of PRM can change according to the question that needs to be addressed. Hence, method based on purely NLP techniques may not be quite effective. We believe that in such situations, combination of ontological knowledge and data mining techniques with NLP can produce effective results. We address this challenge by utilizing the Unified Medical Language System medical ontology and Association Rule Mining.

Preliminaries

UMLS: The Unified Medical Language System (UMLS) [2] is a compilation of more than 60 controlled vocabularies in the biomedical domain. The UMLS is structured around three separate components: Metathesaurus, SPECIALIST Lexicon and Semantic Network. The *UMLS Metathesaurus* provides a representation of biomedical knowledge consisting of concepts (more than 800,000 concepts) classified by semantic type and both hierarchical and non-hierarchical relationships among the concept. English terms from the Metathesaurus are included in the *SPECIALIST Lexicon*, which contains more than 140,000 entries of general and medical terms and stipulates morphological and syntactic facts about English verbs, noun, adjectives and adverbs. Each concept in Metathesaurus is assigned a semantic category, which appears in the *Semantic Network.* The UMLS Semantic Network is a high level categorization of the biomedical domain. It is composed of 134 semantic types and 54 relationships binding them together.

Concept: A Concept is a grouping of synonymous words and phrases defined in the UMLS Metathesaurus. E.g. Concept Name {*Adverse Effect*} is a grouping of the following synonymous terms: [*adverse effects, injurious effects, side effects, therapy adverse effects, treatment adverse effect, treatment harmful effects, treatment side effects, undesirable effects*]. Such concept groupings have been predefined in UMLS Metathesaurus.

Semantic Type: A UMLS Semantic type is a category that comes from the UMLS Semantic Network. Each UMLS Concept has been labeled with one or more UMLS semantic types. E.g. Colorectal Neoplasm has a semantic type of Neoplastic Process.

Semantic Relation: A Semantic relation is a relation pattern at the semantic level defined in the UMLS Semantic Network. In all, 54 relationships bind 134

semantic types. E.g. *{Pharmacological Substance} treats {Neoplastic Process}*

MeSH Terms: MEDLINE contains over 12 million citations to biomedical journal articles. These MeSH terms are keywords that are manually assigned to each MEDLINE citation by trained individuals.

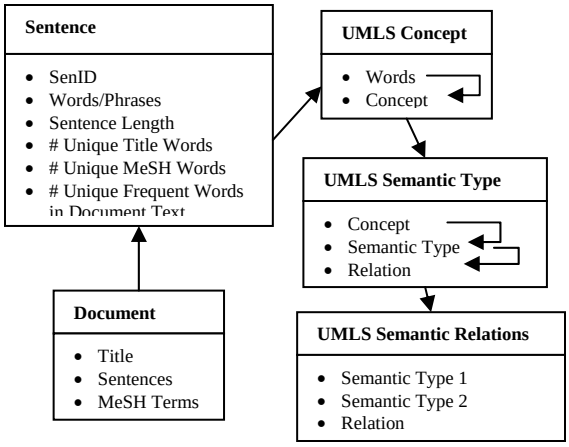

Figure 2: Document and Sentence Model
(Arrows show relationships and chain of relationships)
{Document.Sentences} is related to {Sentence.Words}
{Sentence.Words} => {UMLS Concept.Words.Concept}
=> {UMLS Concept.Words.Concept.Semantic Type}

Document and Sentence Model: We divide each MEDLINE citation into title, abstract text and MeSH terms. Each sentence in abstract text is tokenized into words and phrases. Each phrase has a concept and each concept has a semantic type. Two semantic types are related to each other by semantic relations. Figure above represents the document and sentence model that we follow.

Underlying System Framework

The underlying system framework is divided into six different modules: Query Generation Module, Document Classification Module, Automatic Text Summarization Module, Mapping Module, Relationship Extraction Module and the Rule Verification & Visualization Module. Figure on the next page shows the complete system framework.

Query Generation Module

Users commonly type only a few keywords to retrieve the results from a search engine. In our system, we tried to depict a common user usage of a search engine. We used Google search engine on MEDLINE website to retrieve the scientific articles e.g. to retrieve the scientific literature text related to the disease-treatment model, we used "*Colorectal cancer*" and "*Treatment*" as the query terms.

Document Classification Module

Information Retrieval based Document retrieval using just a few query terms is usually noisy i.e. it may contain documents that actually belong to other categories e.g. a document text may contain a *treatment* word; however it describes about the *diagnosis of a disease*; such a document may also be present in the documents retrieved using "*Colorectal cancer*" and "*Treatment*" query terms. Hence, the document classification technique is beneficial. Normally the task of document classification involves the explicit representation of positive and negative data examples. However, in our case such an explicit representation is not needed. We employ a novel mutual reinforcing algorithm to classify documents without labeled training examples. We first utilize the search results of a general search engine as original training data. We then apply a mutually reinforcing learning algorithm (*MRL*) to mine the classification knowledge and to "clean" the training data. With the help of a set of established domain-specific ontological terms or keywords, the MRL mining step derives the relevant classification knowledge. The MRL cleaning step then builds a Naive Bayes classifier based on the mined classification knowledge and tries to clean the training set. The MRL algorithm is iteratively applied until a clean training set is obtained. This algorithm is detailed in [22]. Results show that it is quite effective.

Automatic Text Summarization Module

One key observation about the scientific literature text in MEDLINE is that it has "one sense per discourse" property. Each abstract normally reveals only a few key relationships that are normally captured in 1 or 2 sentences e.g. normally the "Conclusion" section captures the gist of the document. This module uses this observation as the base and removes redundant sentences from the text. The selection of sentences central to the theme of the document improves both the performance as well as the speed of the system. Our sentence extraction method works by scoring each sentence as a candidate to be a part of the summary, and then selects the highest scoring 'd' sentences. Parameter 'd' is predefined (d can be= 1, 2 or 3).

Features that we used to score a sentence are:

- Number of unique Title keywords in a sentence
- Number of unique MeSH terms in a sentence,
- Number of unique most frequent words in Abstract text in a given sentence,
- Sentence length, and
- Sentence location

A higher number of the Title words and the MeSH words in a sentence may mean that the sentence captures the central idea of the document. Similarly, sentence location e.g. first or the last sentence may capture important relations in the text. Different weights were assigned to the different features. If a sentence is too long, we

decrease its weight. Sentences were then ranked according to their scored and the top d sentences were selected.

Mapping Module

Mapping Sentences to Concept Terms:

The goal of this module is to map the phrases of sentences to the UMLS Concepts. To achieve this, the sentence text is tokenized into the word tokens. These tokens are matched against terms from the SPECIALIST Lexicon to combine the word tokens into multi-word terms. These terms are passed through noun phrase chunker to tokenize them into phrases. Variants including synonyms, spelling variants, acronym and abbreviation, derivational inflections, meaningful combination of these, are retrieved for the words. Candidates are then evaluated against several criteria. This gives a best mapping for the sentence text to UMLS concept terms. If no mapping is found for a phrase, the phrase is discarded. We use a freely available API's from UMLS website: MetaMap Transfer technology (MMTx) [3] to achieve this. MMTx is known to achieve a high level of accuracy in mapping.

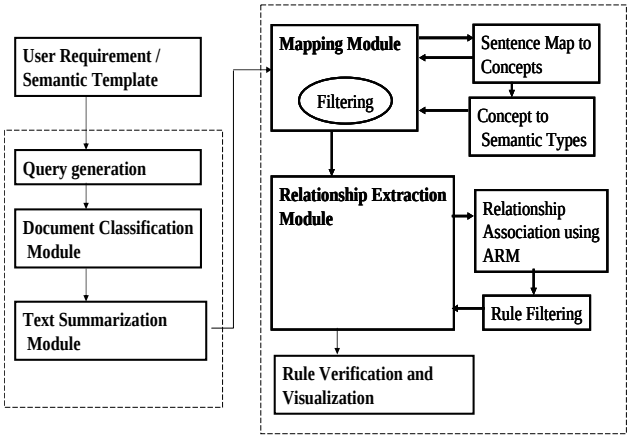

Figure 3: System Framework

Mapping Concept Terms to Semantic Types and Filtering:

This module maps conceptual terms (UMLS Concepts) for a sentence to their corresponding UMLS semantic types. After mapping the concept terms to the semantic types, we put a semantic constraint filter to remove the uninteresting concepts from the text. Interesting concepts for a particular relationship were obtained from the Semantic Network.The example below will further help to understand this concept. This sentence has been taken from "treatment" category. In this example, sentence has been broken to [Concept/Semantic type] mapping e.g. Fluorouracil is the concept and Pharmacologic Substance is the Semantic Type. Here [initial] can be filtered as it has an uninteresting semantic type [temporal concept].

Sentence:
5-Fluorouracil (5-FU) has been the mainstay of systemic therapy for colorectal cancer since its initial development 40 years ago.
Sentence mapping to Concepts / Semantic Types:
[Fluorouracil / Pharmacologic Substance] [systemic therapy / Therapeutic or Preventive Procedure] [therapeutic procedure / Therapeutic or Preventive Procedure] [colorectal cancer / Neoplastic process] [initial / temporal concept]

Relation Extraction Module

This module consists of two sub-modules. In the first sub-module, we use the Association rule mining to extract the knowledge rules from the concepts. In the second, we use some filtering methods to prune off redundant rules.

Association Rule Mining

Association rule mining (ARM), since its introduction, has become one of the core data mining tasks. The prototypical application of ARM is market-basket analysis. Zaki [5] provides a good survey on the different techniques used for ARM.

Here, we seek to find the relations among concepts that manifest semantic relation patterns such as treat(pharmacologic substance, neoplastic process) using the ARM technique. ARM is usually used to discover relationships of certain type such as "co-occur" using the keywords in the text. Blake et al [4] have earlier argued that UMLS concept representation can capture much better rules than those based on words or keyword features from MeSH terms. Our approach maps the sentence to a conceptual model. We reason that co-occurrence of concepts at the conceptual level may reveal interesting relationships. ARM can then be successfully applied to reveal associations. We restrict the rules discovered by defining a pattern based on the "must-have" concepts such as the list of consequents must contain come specific concept terms (e.g. colorectal cancer) and the list of antecedents must contain therapeutic procedures (treatment concept) or viceversa. This list is derived from the input semantic template. We then get the rules that manifest semantic relation patterns such as treat(pharmacologic substance, neoplastic process).

We use a low support and high confidence criteria. A low support criterion is useful to capture a good coverage of rules. Moreover, it can also capture special scenarios such as a recently introduced drug might be mentioned in very few scientific papers. A high confidence criterion helps to extract only the relevant relationships.. We also link back the rules to the sentence from which they were derived. This helps the user to verify the relationships obtained. These rules were further filtered in our rule filtering phase.

Rule filtering

We follow low support criteria. Therefore, we use a rules

filtering module to remove the redundant relationships and concepts. Currently, we have developed two filters – relationship filter and concept filter. Relationship-filter groups all the nodes in a rule into the semantic categories and tries to verify if these categories determine a valid relation. We use concept filter to rectify errors induced due to unwanted frequent concept sets discovered.

Rule Verification and Visualization Module

Rule Verification module tries to score each rule discovered from a sentence. In this module, we do a full syntactic parse & build a concept graph [23] of the sentence. We are experimenting with relational learning functions and NLP techniques such as in [23] to verify the relationship between the nodes of the rules at the semantic level. We start with small set of known relationships at the semantic level and iteratively build the learning knowledge. Visualization Module consists of converting the elements in the rules discovered to nodes. These are currently under development and are not described here.

Results

We present here only the preliminary results. Common query terms were used for downloading the scientific articles - "Colorectal Cancer and Treatment", "Colorectal Cancer and Screening", "Colorectal Cancer and Diagnosis". Here, we present results from the 100 documents downloaded for the category "Colorectal Cancer and Treatment". 70 documents were selected after categorization. Parameter d was set to 2. Total sentences summarized were 70*2=140. The sentences were mapped. There were total 3 mapping errors e.g. the word "*correct*" was wrongly mapped to the "*Correct*", a pharmacological substance. ARM was then run on these mapped concepts and only the maximal frequent pattern rules were mined. 42 such rules were discovered. Rules are defined as interesting if they capture a relationship concept e.g. {*Pemetrexed, therapeutic procedure=> Colorectal Cancer*} is an interesting rule as *Pemetrexed* is a drug used in treatment of Colorectal Cancer (CRC). 32 of the rules were found interesting rules and 10 of them were uninteresting. Concept filter removed 7 of them. Relationship filtered removed 1 of these. Out of these 10, 1 discovered rule was due to the error in mapping. 2 of these uninteresting rules were not filtered out because they had related concepts that were very general e.g. "Adjuvant Immunologic" concept term is a concept related to the treatment of colorectal cancer and has a high support, however it is a general term. We categorized such rules as "difficult to filter out". From the interesting rules discovered, a total of 47 unique relationships or nodes in a model were found. 24 of them were correct and directly related to the Colorectal Cancer e.g. {*Irinotecan treats Colorectal Cancer*}. 10 of them were of similar semantic type (*neoplastic process*) as Colorectal Cancer (CRC), 5 of which were other type of related cancers e.g. *liver, stomach and breast cancer*. 2 were concepts similar to CRC like CRC metastatic. 3 were general terms e.g. advanced cancer. Out of the remaining 13 nodes, 6 were filtered out by our concept filter. 2 of them were due to wrong mapping of concepts e.g. liver was mapped to Liver Extract, a pharmacological substance. 5 of them were difficult to filter out. 3 of which were concepts related to treatment strategies e.g. "systemic therapy". 2 errors were induced due to related general terms such as Adjuvant Immunologic. We conclude that by utilizing ontological knowledge to map text at a higher level of abstraction, ARM can successfully reveal relationships.

Related Works

Natural Language Processing (NLP) has been applied to the biomedical text for decades. Spyns [6] provides a broad overview of NLP in medicine. Traditional knowledge extraction systems from the text in medicine have concentrated on the different kind of language material such as the patient records, radiology reports [7] [8] [9]. With advance in molecular biology, many researchers have focused on extracting knowledge from MEDLINE scientific articles in the molecular biology field. Bruijn [10] provides an excellent survey on the different aspects in mining knowledge from the biomedical literature. They modularize text mining and divide the whole process into four different tasks – Document Categorization, Named-Entity Tagging, Fact extraction and Collection-wide analysis. Going by their division, we implement the first three modules namely document categorization, named-entity tagging and fact extraction. By collection-wide analysis they imply-combining facts to form a novel insight. This is not our aim. Moreover, our Named-entity Tagging task is much more sophisticated than that applied in other systems. In specific, we tag each phrase in the sentence at multiple levels (on the concept level as well as on the semantic level). Such multiple-level tagging is more useful in extracting facts from the text.

There have mainly been four kinds of approaches used in relationship extraction from MEDLINE documents. Frequent co-occurrence approach [11] [12] seems to easier and popular. Frequent co-occurrence approach focuses on the co-occurrences of two specific entity names such as disease and treatment, or protein names with a verb that indicates an association between them. Weeber et al [13] used statistical word frequency analysis to find association between words, but they restrict themselves to only finding the side-effects of a drug. Ding et al [14] tested co-occurrence of entities on abstracts, sentences and phrases in molecular biology articles to see which one provides the best place to identify the relations. Working with phrases gave the best precision and working with sentences gave the better recall. The second approach uses fixed regular expression linguistic

templates (normally hand-crafted) [15] [16] to search for a specific interaction verb and the surrounding entity names. Third approach uses Machine Learning techniques such as HMM [17] to learn some linguistic templates. Others [18] [19] try to discover relationship using a full syntactic parse and relations between syntactic components are inferred. Our approach is a frequent co-occurrence approach but we work on co-occurrence of concepts rather than words.

Blake et al [4] showed that features of different semantic richness have an effect on the plausibility or usefulness of association rules. Cimino's group in Columbia University has done some extensive works [20] using co-occurrence of MeSH terms and UMLS semantic types. They have successfully applied this knowledge in document retrieval as well as for knowledge extraction. This is most similar to ours. However, they have used set of manually pre-defined rules to extract knowledge from specific citations. Our work builds on the initial work of Ai-Ling Zhu et al [21] in our group. In their work, they had presented the feasibility of using co-occurrence of MeSH terms to find some useful relationships in Medical literature using ARM.

Conclusion

In this paper, we presented a prototype system that we are building to model structure from the scientific articles. We use MEDLINE keyword terms and Unified Medical language System (UMLS) to aid in automatically extracting facts and relations from the scientific literature, rather than the hand-crafted rules or the annotated corpora providing training examples. With a semantic template as input, our system downloads the relevant scientific papers using a search engine. It classifies and summarizes the text to the relevant sentences in text. We, then, map the sentences into the concepts and the concepts to the semantic level and use Association rule mining techniques to explore relationships between the concepts. Preliminary results were presented using the Colorectal Cancer medical domain.

Acknowledgments

This research is partly supported by Research Grant No. R-252-000-111-112/303 from the Agency of Science and Technology (A*Star) and the Ministry of Education in Singapore and partly by the Singapore-MIT Alliance.

References

[1] http://www.ncbi.nlm.nih.gov/PubMed/

[2] http://umlsks.nlm.nih.gov/

[3] http://mmtx.nlm.nih.gov/

[4] C. Blake, W. Pratt, Better Rules, Fewer Features: A semantic approach to selecting features from text, IEEE Intl. Conf. Data Mining, 2001

[5] M. Zaki, Parallel and Distributed Association Mining: A Survey, IEEE Concurrency, 1999

[6] P. Spyns, Natural language processing in medicine: an overview, Methods Inf. Med. 35(4-5) 285-301, 1996.

[7] P. Ruch, R.H. Baud, A.M. Rassinoux, P. Bouillon, G. Robert, Medical document anonymization with semantic lexicon, Proc. AMIA Symp. 729-733 , 2000

[8] R.K. Taira, S.G. Soderland, R.M. Jakobvits, Automatic structuring of radiology free-text reports, Radiographics, 21(1) 237-245, 2001

[9] U. Hahn, M. Romacker, S. Schulz, Creating Knowledge repositories from biomedical reports: the MedSynDiKaTe text mining system, Pac. Symp. Biocomput. 338-349, 2002

[10] B. de Bruijn, J. Martin, International Journal of Medical Informatics, 67 1-18, 2002

[11] C. Blaschke, M.A. Andrade, C. Ouzounis, A. Valencia, Automatic extraction of biological information from scientific text: protein-protein interactions, Proc. Int. Conf. Intell. Syst. Mol. Biol., 30A(2) 60-67, 1999

[12] M. Craven, J. Kumlien, Constructing biological knowledge bases by extracting information from text sources, Proc. Int. Conf. Intell. Syst. Mol. Biol, 77-86, 1999

[13] M. Weeber, R. Vos, Extracting expert medical knowledge from texts, In Working Notes of the Intelligent Data Analysis in Medicine and Pharmocology Workshop, 183-203, 1998

[14] J. Ding, D. Berleant, D. Nettleton, E. Wurtele, Mining MEDLINE: abstracts, sentences, or phrases, Pac. Symp. Biocomput., 326-337, 2002

[15] T.Ono, H. Hishigaki, A. Tanigami, T. Takagi, Automated extraction of protein-protein interactions from the biological literature, Bioinformatics 17(2) 155-161, 2001

[16] T. Sekimizu, H.S. Park, J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, Genome Inf, Ser. Workshop Genome Inf., 9 62-71, 1998

[17] S. Ray, M. Craven, Representing sentence structure in Hidden Markov Models for information extraction, IJCAI, 1273-1279, 2001

[18] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, L. Hunter, EDGAR: extraction of drugs, genes and relations from the biomedical literature, IEEE Intell. Syst., 16(6): 62-67, 2000

[19] J. Pustejovsky, J. Casatano, J. Zhang, M. Kotecki, B. Cochran, Robust relational parsing over biomedical literature: extracting inhibit relations, Pac. Symp. Biocomput., 362-373, 2002

[20] J.J. Cimino, G.O. Barnett, Automatic knowledge acquisition from medline. Methods of Information in Medical, 32(2):120-133, 1998

[21] Ai-Ling Zhu, Jian Li, Tze-Yun Leong, Automated knowledge extraction for decision model construction: a data mining approach, AMIA Annual Symposium, 2003

[22] Xiaoli, Rohit Joshi, Tze-Yun Leong, Classifying Biomedical Citations without labeled training examples, Work In progress (unpublished).

[23] C. Cumby, D. Roth, On Kernel Methods for Relational Learning, ICML, 2003