Automatic Extraction of Ontology Relations From Medical Abstracts

Jin-Cheon Na, Christopher Khoo, Chew-Hung Lee
Division of Information Studies
School of Communication and Information
Nanyang Technological University, Singapore
65-6790-4621
{tjcna, assgkhoo}@ntu.edu.sg, lchewhun@dso.org.sg

ABSTRACT

Ontologies play an important role in the Semantic Web as well as in knowledge management. This project seeks to develop an automatic method to build ontologies, especially in a medical domain. The initial study investigates an approach of identifying pairs of related concepts using association rule induction and identifying semantic relations between concepts using an existing medical knowledge base, the UMLS (Unified Medical Language System) semantic net. This is evaluated by comparing the result with manually assigned semantic relations based on an analysis of medical abstracts containing each pair of concepts. Our initial finding shows that the automatic process is promising, achieving a 68% coverage compared to manual tagging. We also discuss about probable approaches for the improvement of the identification of semantic relations by employing natural language processing techniques.

INTRODUCTION

The Semantic Web [1] is a vision to extend the current Web into an environment where computers can cooperate with people to perform sophisticated tasks. This environment relies on information provided with well-defined meanings that computer agents can process and use. Ontologies as formal knowledge bases provide such machine-processable semantics. An issue facing the Semantic Web community is the lack of rich ontologies as the creation of ontologies is non-trivial requiring analysis of domain sources, background knowledge, and consensus among the users of the ontologies. The conventional approach in constructing an ontology is to manually enumerate the concepts and relations found in a domain from domain sources. This approach is not suitable for developing a large ontology as it is labour intensive and is likely to give rise to inconsistencies. An alternative approach is to use automatic or semi-automatic methods to extract the concepts and relations [6, 7, 8].

We have embarked on a project to develop an automatic method to enrich existing ontologies, especially the identification of semantic relations between concepts in the ontology, by analyzing domain texts. The major benefit of this project will be the provision of a new tool for ontology engineers to enrich an ontology automatically. With the generated domain ontology, we can then build interesting Semantic Web applications, such as a knowledge management system based on domain ontologies.

As an initial study, we carried out a small experiment using a sample of abstracts of medical articles to identify pairs of related concepts related to "Colon Cancer Treatment" and inferred the semantic relations between the terms in each pair using the UMLS (Unified Medical Language System) semantic network [10]. The purpose was to find out how effective this simple method is in identifying ontological relationships, and to what extent natural language processing techniques need to be applied to the text to infer relationships between the concepts.

The rest of the paper is organized as follows. Sections 2 and 3 briefly describe related works and our framework for ontology learning respectively. Section 4 discusses the results of an initial experiment of ontology learning in the colon cancer domain, and Section 5 concludes the paper with discussions of probable approaches for the improvement.

RELATED WORKS

Maedche [6] and Navigli, Velardi, and Gangemi [7] worked on (semi-)automatic methods to extract the concepts and relations. They investigated building ontologies from general domain documents, such as travel related documents. Since target domains are very broad, the generated ontologies seem not have deep structured ontologies compared to the one manually generated by domain experts. Some projects use Word Net [4] as an existing domain knowledge base to overcome the problem. However, it could be too general for specific domain documents, such as medical documents. It is necessary to use existing domain knowledge bases when building domain specific significant ontologies.

Blake and Pratt [2] worked on mining semantic relationships among medical concepts (or terms) from medical texts. They focused on "Breast Cancer Treatment" using association rules (a data mining technique) to find associated concept pairs like magnesium-migraines. They were mainly interested in mining the *existence* of a relationship between medical concepts and not in identifying the specific type of semantic relation for the associated concept pairs. For example, the relationship between magnesium and migraines pair could be one of the following semantic relations: treat, prevent, disrupt, and cause. Because identifying specific semantic relations is very important for ontology learning, our work focuses more on finding specific semantic relations.

For the ontology learning, we use UMLS, an existing medical knowledge base maintained by the NLM (National Library of Medicine), as a seed ontology. The UMLS consists of three components: (i) the Metathesaurus containing information about biomedical concepts and terms from many controlled vocabularies and classification systems used in medical information systems, (ii) a semantic network providing a consistent categorization of all concepts represented in the UMLS Metathesaurus (the links between the semantic types provide the structure for the Network and represent important relationships in the biomedical domain), and (iii) the Specialist lexicon providing lexical information on concepts.

ONTOLOGY LEARNING

Our ontology learning process is shown in Figure 1. Abstracts of medical research papers are first collected from MedLine through the PubMed interface using a specific medical query such as "Colon Cancer Therapy". Important terms are then extracted from the medical abstracts. Currently we use the MeSH (Medical Subject Headings) terms used in indexing the abstracts as important terms. Next we map each extracted term to a medical concept in the UMLS, and an association rule tool [3] is applied to the concepts to find associated concept pairs.

After finding associated concept pairs, we proceed to extract specific relations. In the UMLS, the semantic network provides information about the set of basic semantic types (the nodes in the network) that may be assigned to concepts in the Metathesaurus. It also defines the set of relationships (the links in the network) that may hold between the semantic types. The 2003AA release of the semantic network contains 125 semantic types and 54 relationships. The relations are stated between high level semantic types in the semantic network whenever possible, and are generally inherited via the "is-a" link by all the children of those types. In some cases there will be a conflict between the placement of types in the semantic network and the link to be inherited. Thus, the inheritance of the link also can be blocked.

In the initial experiment reported in this paper, the semantic relations between associated concepts are inferred from this semantic network. First each concept in a concept pair is mapped to one of the 135 semantic types, and the direct or indirect semantic relations that are predefined between the two semantic types in the semantic network are taken as the semantic relation for the target concept pair.

Our approach for the identification of the semantic relations is similar to the one by Rindflesch and Aronson [9]. However their concept extraction method is different from ours. They mainly process a document sentence by sentence, find concepts co-occurring in a sentence, and infer their semantic relationship using the UMLS semantic network and natural language processing techniques. Nonetheless mining of important (i.e., frequently occurring) concepts and their associations from multiple documents looks more appropriate for ontology building. Otherwise, the result ontology might become very big with insignificant information.

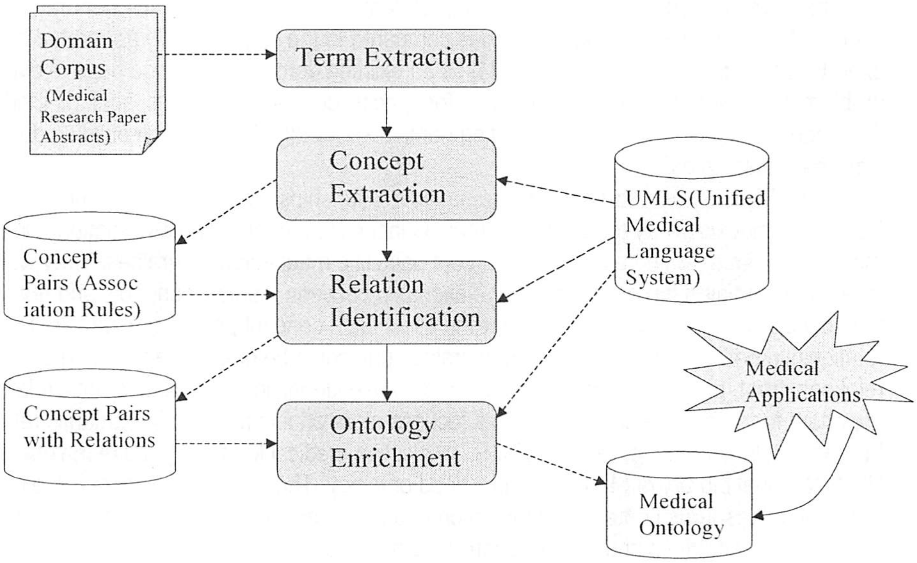
Domain Corpus (Medical Research Paper Abstracts)
Term Extraction
Concept Extraction
UMLS(Unified Medical Language System)
Concept Pairs (Association Rules)
Relation Identification
Medical Applications
Concept Pairs with Relations
Ontology Enrichment
Medical Ontology

Figure 1. Ontology Learning Processes

Finally, at the ontology enrichment stage, we merge the extracted concepts and their semantic relations with the seed ontology, the UMLS semantic network. The generated ontology can then be used as a domain knowledge base for various medical applications.

RESULTS OF INITIAL EXPERIMENTS

In our experiment, we extracted the association rules (i.e. pairs of associated concepts) from a sample of 387 medical abstracts following the framework outlined above. These rules had at least 2% support and 80% confidence -- i.e. both concepts occurred in at least 2% of the abstracts, and of the abstracts containing the first concept, 80% also contained the second (associated) concept. We also filtered out rules involving "human", "mice" and "rats" as these concepts yielded trivial rules, such as "Mice, Inbred-> Mice", and we are interested in rules relating to colon cancer and treatment. The remaining 34 rules were tagged automatically with UMLS semantic relations using the inferencing method outlined earlier. The first and second authors also manually tagged each association rule with a semantic relation after examining a sample of 10 abstracts containing the pair of concepts.

Of the 34 rules, 11 rules had no matching semantic relation using the automatic method. Four rules were automatically tagged with a relation, and 19 rules were automatically tagged with multiple relations. In the manual tagging of semantic relations, all 34 rules had semantic relations assigned to them, indicating that a semantic relation between the concepts was expressed in at least one of the abstracts examined. 19 of the rules were manually assigned 1 relation, and 15 rules had multiple relations assigned.

The automatically tagged relations were compared with the manually assigned relations. As mentioned earlier, 11 rules (or 32%) were not tagged with a semantic relation by the automatic method. Of the remainder, 4 rules (12%) were assigned the same semantic relation by both the automatic and manual tagging. 19 rules (56%) had partial matches – the automatic and manual tagging had at least 1 relation in common.

As an example of interesting relations found through this process, the relation "Leucovorin/administration&dosage *interact_with* Fluorouracil/administration&dosage" with a support of 3% and a confidence of 100% was automatically tagged and concurs with the manual tagging of "*interact_with*". Another interesting rule is the relation between Liver Neoplasms/secondary and Colonic Neoplasms/pathology with a support of 7% and a confidence of 82% although the automatic method was not able to differentiate between the three semantic relations *affects, manifestation_of* and *result_of.*

DISCUSSION

In ontology learning, finding semantic relations between concepts is not an easy problem but the usage of a domain-related seed ontology (e.g. the UMLS semantic network) eases the difficulty of semantic relation identification somewhat. We are able to infer semantic relations between concepts automatically from a seed ontology 68% of the time (23/34), although the method cannot distinguish between a few possible relation types. This suggests that it is feasible to employ natural language processing (NLP) techniques to identify relations between concepts in the medical ontology. Thus, our next step is to investigate the use of NLP of medical abstracts to identify the appropriate relation.

In our manual analysis of medical abstracts, we noticed that generally the pair of associated concepts occurs in the same sentence or in adjacent sentences. Often they occur within the same compound noun, or in two noun phrases linked by a verb. For the case of concepts occurring in the same sentence, we plan to use semantic relation patterns (e.g., treatment of) to identify relations between them. Khoo et al. [5] and Rindflesch and Aronson [9] investigated relation extraction through manually constructed patterns. However, it will be very labor intensive, thus we plan to explore an automatic method for constructing extraction patterns from corpus. Note that since concept pairs can come from multiple documents, when there are multiple different relations, we need to keep all relations (probably with certainty factors). From this, we can see multiple relations, possibly including conflicting ones, between two medical terms. In addition, we will extract important terms by processing the domain corpus using text mining techniques, rather than using the MeSH terms. It will increase the number of cases where the pair of associated concepts occurs in the same sentence.

For concepts occurring across sentences, we plan to incorporate discourse processing, including anaphor and co-reference resolution. Pronouns and references have to be resolved and replaced with the information that they refer to. The interesting aspect of discourse processing is what we refer to as hypothesis confirmation. Sometimes, at the beginning of the abstract, the author hypothesizes a relation. This hypothesis may be confirmed or disconfirmed by another sentence later in the abstract. The system thus has to be able to link the initial hypothetical relation with the confirmation or disconfirmation relation later in the abstract. Finally, there is a case of concepts occurring across documents. Since we extract mainly frequently occurring concept pairs, we think that the case will not be significant.

The eventual contribution of this project will be the provision of a new tool for ontology engineers to create ontology automatically or semi-automatically. The generated ontology will be helpful for building the following Semantic Web applications:

- *Web site (or knowledge portal) creation using domain ontology*: A domain Web structure can be created automatically based on the generated ontology. The Web site might be updated with new domain documents.
- *Mining new treatments from medical documents*: An information system can generate an ontology graph based on the treatment semantic relation. The user can navigate through the ontology to find new treatments for diseases.
- *Query expansion with domain ontology*: For getting improved query results, the system can expand the user input queries using the generated domain ontology.

REFERENCES

[1] T. Bemers-Lee, J. Hendler and O. Lassila. The Semantic Web. *Scientific American,* May 2001, pp. 35-43.

[2] C. Blake and W. Pratt. Better Rules, Fewer Features: A Semantic Approach to Selecting Features from Text. *In Proceedings of the IEEE Data Mining Conference*, San Jose, California, IEEE Press, pp. 59-66.

[3] C. Borgelt, "Apriori Implementation", Available at http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/apriori/, visit on August 2003.

[4] C. Fellbaum. WordNet: An Electronic Lexical Database. The MIT Press, 1998.

[5] C. S.G. Khoo, S. Chan and Y. Niu. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. *In ACL-200: 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, October 2000, pp. 336-343.

[6] A. Maedche. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002.

[7] R. Navigli, P. Velardi, and A. Gangemi. Ontology Learning and its Application to Automated Terminology Translation. *IEEE Intelligent Systems,* the IEEE Computer Society, January/February 2003, pp. 22-31.

[8] B. Omelayenko. Learning of Ontologies for the Web: the Analysis of Existent Approaches. *In Proceedings of the International Workshop on Web Dynamics,* London, UK, January 2001.

[9] T. C. Rindflesch and A. R. Aronson. Semantic Processing for Enhanced Access to Biomedical Knowledge. *Real World Semantic Web Applications,* IOS Press, pp. 157-72.

[10] Unified Medical Language System, National Library of Medicine, Available at http:www.nlm.nih.gov/research/umls, visit on August 2003.

Discussion Points

41. What does the author mean by classification?

42. Does the research involve creation or implementation of a classification scheme?

43. How does the researcher use classification to improve the automated approach?

44. How do these methods compare to current human-generated approaches to classification?

45. How does the reported research expand our understanding of classification?

46. Does the research suggest an improvement over human-generated classification?

47. What do you think are the most important lessons learned in this research?

48. What do you think are the best practices reported in this research?

49. What would you recommend to the researcher as the next step in this approach?

50. Is there other related research that you would recommend the researcher become acquainted with?