Ontology Learning for Medical Digital Libraries

Chew-Hung Lee, Jin-Cheon Na, and Christopher Khoo

Division of Information Studies, School of Communication and Information, Nanyang Technological University, Singapore
lchewhun@dso.org.sg, {tjcna, assgkhoo}@ntu.edu.sg

Abstract. Ontologies play an important role in the Semantic Web as well as in digital library and knowledge portal applications. This project seeks to develop an automatic method to enrich existing ontologies, especially in the identification of semantic relations between concepts in the ontology. The initial study investigates an approach of identifying pairs of related concepts in a medical domain using association rule induction and inferring the type of semantic relation using 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 manually tagging. However, natural language processing of medical abstracts is likely to improve the identification of semantic relations.

1 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 computers 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 labour intensive approach is unsuitable for developing a large ontology as it is likely to give rise to inconsistencies. An alternative is to use automatic or semi-automatic methods to extract the concepts and relations [4, 5]. 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.

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) [6] semantic network. 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. Section 2 highlights related works and our framework for ontology learning. Section 3 discusses the results of an initial experiment involving the colon cancer domain, and Section 4 concludes the paper.

2 Ontology Learning

Blake and Pratt [2] mined semantic relationships among medical concepts from medical texts. They focused on "Breast Cancer Treatment" using association rule mining to find associated concept pairs like magnesium-migraines. They were mainly interested in mining the existence of relationships between medical concepts (i.e., finding treatment methods for breast cancer) and not in identifying specific semantic relations for the associated concept pairs. For example, the relationship between magnesium and migraines pair could be one of the following semantic relations: treat, prevents, disrupts, 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 as a seed ontology. 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, and (iii) the Specialist lexicon providing lexical information on concepts.

Our ontology learning process is shown in Figure 1. Abstracts of medical research papers are first collected from MedLine through the PubMed interface [6] using a specific medical query such as "Colon Cancer Treatment". 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 (we also plan to extract important terms by processing the domain corpus using text mining techniques). 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. The UMLS semantic network provides information about the set of basic semantic types that may be assigned to concepts in the Metathesaurus. It also defines the set of relationships 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.

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

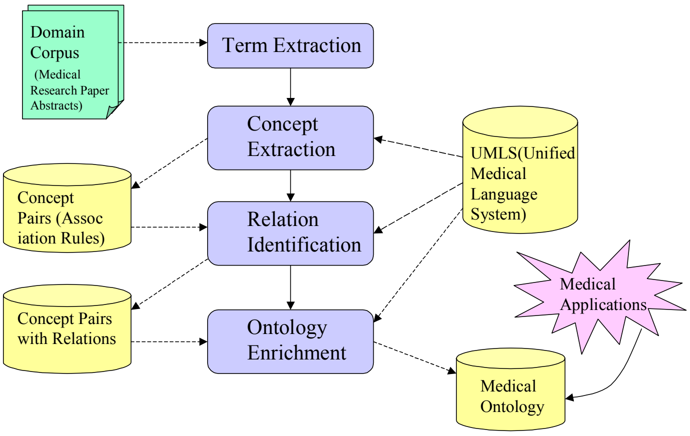
Domain Corpus (Medical Research Paper Abstracts)
Term Extraction
Concept Extraction
UMLS(Unified Medical Language System)
Concept Pairs (Association Rules)
Relation Identification
Medical Applications
Concept Pairs with Relations
Ontology Enrichment
Medical Ontology

Fig. 1. Ontology Learning Processes

3 Results of Initial Experiment

In our experiment, we extracted the association rules 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 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 second and third authors also manually tagged each association rule with a semantic relation after examining 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*.

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. However, as our result shows, the seed ontology is not a panacea and analysis of medical texts such as medical abstracts is needed to identify both missing relations as well as to select an appropriate relation from a set of identified relationships.

4 Conclusion

The major benefit of this project will be the provision of a new tool for ontology engineers to create ontology automatically or semi-automatically. 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. Our next step is to investigate the use of natural language processing (NLP) of medical abstracts to identify the appropriate relation. As associated concept generally occurs within the same compound noun, or in two noun phrases linked by a verb, this suggests that NLP could be used to identify the relations between the concepts.

The generated ontology will be helpful for building the digital library applications like updating a medical treatment website with new treatments identified in the ontology and navigating medical digital encyclopedias using the generated 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 June 2003.
4. A. Maedche. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002.
5. 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.

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