Effectiveness of UMLS Semantic Network as a Seed Ontology for Medical Domain Ontology Building

Hock Leng Neoh and Jin-Cheon Na
Division of Information Studies, School of Communication and Information, 31 Nanyang Link,
Nanyang Technological University, Singapore 689918
+65 97428910
{ps7677798e, tjcna}@ntu.edu.sg

Abstract

In the Semantic Web, ontologies play an important role in knowledge portal applications especially in the medical field. Of interest in this project is the manual information extraction process known as the Knowledge Engineering approach to identify relationship pairs of "Colon Cancer Treatment". The concept and semantic types of these relationship terms extracted are determined using the Unified Medical Language System (UMLS) which is used as a seed ontology. The manual information extraction process yielded some interesting patterns that could further improve the relevance of the Information Extraction (IE) systems. Protégé was used to build and extend the ontology or knowledge base. From this study, it was found that the semantic relationship extracted and compared against UMLS semantic network, only for half of the relationships holds in UMLS. Some semantic types mapped as classes were too general and subclasses were proposed to make it more specific.

Introduction

Much effort has been carried out by researchers to scale the infinite number of knowledge sources to extract relevant information from the World Wide Web, digital library and knowledge portal applications. The Semantic Web is one effort by W3C which was an idea initiated by Tim Berners Lee to link up information to be easily processable by machines on a global scale [Berners-Lee, Hendler and Lassila, 2001]. These machine processable semantics could provide formal knowledge bases termed as ontologies. The lack of rich ontologies is an issue facing the Semantic Web community. This is due to the fact that ontology building requires analysis of domain sources, background knowledge, and consensus among the users of ontologies [Lee, Na and Khoo, 2003]. Analysis of domain sources involves the extraction of relevant information either by using the Knowledge Engineering approach or the Automatic Training approach. The Knowledge Engineering approach is utilized in this study whereby a sample of abstracts in the "Colon Cancer Treatment" domain obtained from Medline is analyzed manually to extract relevant relationship terms. These relationship pairs are then entered into the Unified Medical Language System Knowledge Server (UMLSKS) to identify its concept and semantic relationships. The manual extraction of the concepts from the abstracts yielded some interesting patterns. These patterns are intended to be used to generate Information Extraction (IE) patterns. Based on these semantic relationship patterns obtained using UMLS as a seed ontology, an ontology is generated using Protege. The goals of building this ontology are to determine the effectiveness of UMLS in predicting the extracted concept relationships and consequently enrich the seed ontology based on the extracted concept and semantic relationships. The generated ontology is hoped to be used as a domain knowledge base for medical digital library application. In addition, it would provide faster retrieval of relevant and useful information for treatment of colon cancer.

Literature Review

Information Extraction (IE) is a process which takes unseen texts as input and produces fixed-format, unambiguous data as output [Cunningham, 1999]. It is a subfield of natural language processing that is concerned with identifying predefined types of information from text [Riloff, 1999]. To put it in a nutshell, developing systems that can identify information in a document that contains relevant information to a prescribed task, extracting the information and in turn relating these pieces of information by means of filling a structured template or a database record is the ultimate goal of information extraction systems [Khoo, Chan and Yun, 2003; Appelt and Israel, 1999]. One type of IE approaches is the Knowledge

Engineering approach which utilizes the expertise of a knowledge engineer to develop the grammars used by a component of the Information Extraction system. This approach is considered a manual approach.

In the medical field, UMLS is used widely for building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research. UMLS is an existing medical knowledge base that is maintained by the National Library of Medicine (NLM). The UMLS consists of three components namely the Metathesaurus, the Semantic Network and the Specialist Lexicon. The Metathesaurus contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases, and expert systems. The Semantic Network provides a consistent categorization of all concepts represented in the UMLS Metathesaurus through its semantic types. The structure of the Network and representation of important relationships in the biomedical domain are provided by the links between the semantic types.

One particular work relating to the design, extension and the learning of domain ontologies in the medical field is related to a domain specific knowledge base in the area of blood transfusion that was built based on the UMLS knowledge sources. The goal of this team of researchers is to introduce a part of UMLS concepts and relations in the knowledge representation language to design their knowledge-based system (KBS) [Achour, Dojat, Brethon, Blain and Lepage, 1999]. The approach was based on the Protégé II project that allows domain experts to directly enter or modify existing knowledge in a Decision Support System (DSS) using domain ontology. Another effort in the medical field was carried out and an automatic method to enrich ontologies was developed for the purpose of identification of semantic relations between concepts in ontology [Lee, Na and Khoo, 2003]. The specific domain ontology of interest in this project is colon cancer treatment. The initial study investigated an approach of identifying pairs of related concepts using association rule induction and inferring the type of semantic relations using the UMLS semantic net.

Research Method and Results

The manual extraction process went through 4 distinct stages to obtain the best possible relationship pattern. The first stage involved manual reading of each abstract and the relevant terms related to “Colon Cancer Treatment” were extracted. In this stage, the relationship term was also determined. This was followed by the second stage whereby, the concept and semantic types were determined based on the terms extracted in the first stage using the Unified Medical Language Sources Knowledge Server (UMLSKS). In this way, each relationship term is mapped to a concept and semantic type in UMLS and thus making it possible to determine the effectiveness of UMLS. The third stage involved refining the semantic relationships by analyzing the structure of the relationships. Stage four involves determining similar patterns and summarizing the patterns. The third and fourth stage is for the purpose of determining similar patterns for “treat” relationships and also to recognize the relationship terms for Information Extraction purposes. Part of the summarized semantic relationship pattern and the number of relationships from the 109 abstracts analyzed is shown in Table 1. The “*” symbol signifies terms existing before or after the relationship terms in the sentence that contains “Colon Cancer Treatment” relationships.

Semantic Type 1	Relationship Term	Semantic Type 2	Number of Relationships
<Therapeutic or Preventive Procedure>	* Of *	<Neoplastic Process>	7
<Therapeutic or Preventive Procedure>	* For *	<Neoplastic Process>	6
<Pharmacologic Substance>	* Inhibit *	<Neoplastic Process>	4
<Pharmacologic Substance>	* Against *	<Neoplastic Process>	3
<Organic Chemical>	* Against *	<Neoplastic Process>	3

Table 1: Semantic Relationships Extracted

A total of 113 relationships were extracted from the 109 abstracts analyzed. The highest number semantic relationship extracted was the <Therapeutic or Preventive Procedure> *of* <Neoplastic Process> with 7 relationships extracted. This is followed by the <Therapeutic or Preventive Procedure> *for* <Neoplastic Process> relationship with 6. In terms of relationship patterns, the most relationship terms for

"Colon Cancer Treatment" that were extracted were verbs and prepositions which accounted for 41%-42% of the 113 relationship extracted.

The semantic relationships extracted are then used to build or extend an ontology using Protégé. The domain ontology build is based on "Colon Cancer Treatment" using UMLS as a seed ontology. All the 154 semantic types are mapped into the Protégé as classes or subclasses depending on its hierarchy in the UMLS semantic network. Concepts of these semantic types are then populated as instances of the classes. Figure 1 shows part of the structure of the semantic types and concepts that are mapped using Protégé. A total of 108 semantic relationship pairs were mapped and 119 instances populated. The main semantic relationship pairs mapped are the "Therapeutic or Preventive Procedure" treats "Neoplastic Process" and "Pharmacologic Substance" treats "Neoplastic Process" having 27 and 20 pairs respectively. The semantic relationship pairs extracted involved 4 relationship terms that are "treats", "manages", "prevents" and "brings_about" and 96 of them involved the relationship term "treats". Only 55 or 50.9% of the semantic relationships mapped holds true based on the UMLS semantic network. The rest of the semantic relationships that were mapped does not exist in the UMLS semantic network. An example of semantic relationship that does not hold in UMLS would be <Clinical Drug> "treats" <Neoplastic Process>.

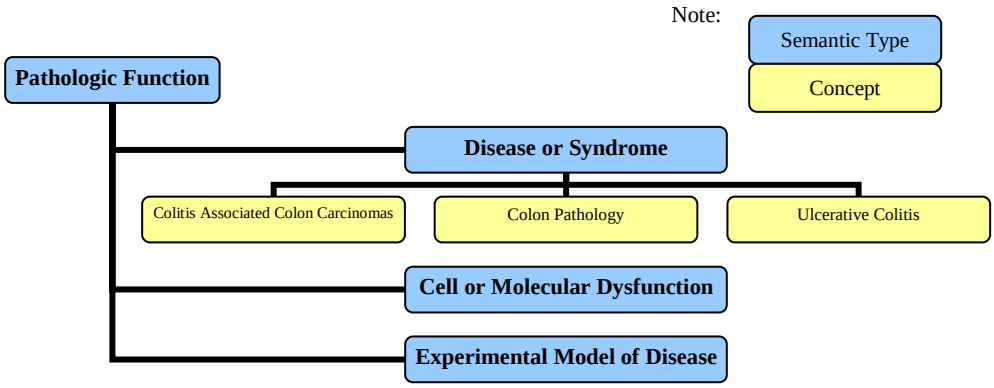

Figure 1: Part of the Structure of the Semantic Types and Concepts Mapped in Protégé

Some of the classes in the UMLS may be too general and hence subclasses are created in order to have a more specific categorization of the semantic type. Two of the semantic types that were found to be too general were "Neoplastic Process" and "Therapeutic and Preventive Procedure". In UMLS, "Neoplastic Process" encompasses all types of cell growth that are cancerous and non-cancerous which is too general. This also holds true for "Therapeutic and Preventive Procedure" where it includes a broad range of treatment methods. In this study, under "Neoplastic Process", 5 subclasses were added while under "Therapeutic and Preventive Procedure", 2 subclasses were added as shown in Figure 2.

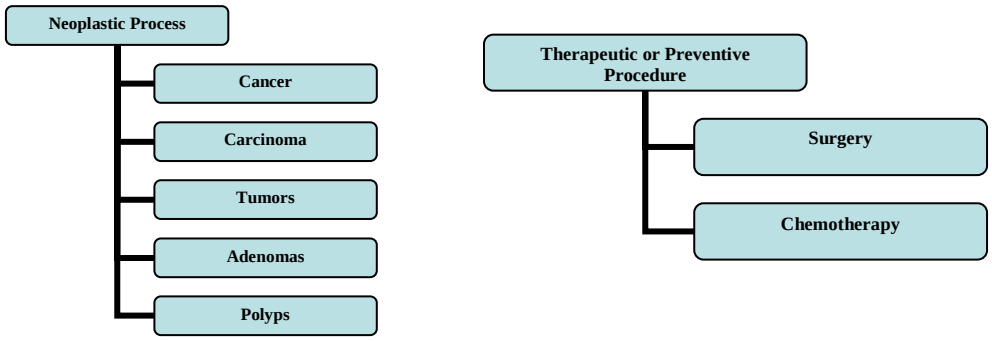

Figure 2: Subclasses Added Under "Neoplastic Process" and "Therapeutic and Preventive Procedure"

Conclusion

Manual information extraction yielded interesting patterns that could be used to generate common patterns for more relevant and easier information extraction. The UMLS semantic network is an extensive knowledge source that is useful in the medical domain. However, it was found that information extraction using UMLS concepts and semantic types were not sufficient to predict the relationships. This is due to the fact that only half of the semantic relationships extracted hold true based on UMLS. In addition, some semantic types in UMLS was found to be too general and its definitions are too broad. Therefore, subclasses are created to have a more specific categorization of two of the semantic types. In general, the effectiveness of the UMLS knowledge source could be further enhanced with the above findings.

References

1. Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. In *Scientific American.com*.
2. Cunningham, H. (1999). Information Extraction - A User Guide (2nd ed.). In Research Memo CS-99-07, *General Architecture of Text Engineering*, Hamish Cunningham.
3. Riloff, E. (1999). Information Extraction as a Stepping Stone toward Story Understanding. In *Computational Models of Reading and Understanding*, Ashwin Ram and Kenneth Moorman, (Eds.), The MIT Press.
4. Khoo S. G. C., Chan S., & Yun N. (2003). Extracting Causal Knowledge from Medical Database Using Graphical Patterns. *Singapore Journal of Library & Information Management, 28*, pp. 48-63.
5. Appelt, D. E., & Israel D. J. (1999). Artificial Intelligence Centre, SRI International. Introduction to Information Extraction Technology. *A tutorial prepared for IJCAI-99*.
6. Archour, S., Dojat, M., Brethon, J. M., Blain, G., & Lepage, E., (1999). The Use of the UMLS Knowledge Sources for the Design of a Domain Specific Ontology: a Practical Experience in Blood Transfusion. In Proceeding of. *American Medical Association Annual Symposium*, 6-10 Nov, Washington DC, pp 249-253.
7. Lee, C. H., Na, J. C., & Khoo, C., (2003). Ontology Learning for Medical Digital Libraries. In Proceeding of *ICADL (International Conference on Asian Digital Libraries) '2003*, Malaysia, Kuala Lumpur, December, pp. 302-305.