Extracting Ontology from Medical Documents with Ontology Maturing Process

Enkhbold Nyamsuren, Kang Dong-Yeop, Su-Kyoung Kim and Ho-Jin Choi
Intelligent Software Engineering and Robotics Lab
Korean Advanced Institute of Science and technology
Daejeon, Republic of Korea
Email: {n_egii,aconqueror,ksk0314,hjchoi}@icu.ac.kr

Abstract

Ontology maintenance is a time consuming and costly process which requires special skill and knowledge. It requires joint effort of both ontology engineer and domain specialist to properly maintain ontology and update knowledge in it. This is specially true for medical domain which is highly specialized domain. This paper proposes a novel approach for maintenance and update of existing ontologies in a medical domain. The proposed approach is based on modified Ontology Maturing Process which was originally developed for web domain. The proposed approach provides way to populate medical ontology with new knowledge obtained from medical documents. This is achieved through use of natural language processing techniques and highly specialized medical knowledge bases such as Unified Medical Language System.

1. Introduction

Within the project done under the Grid Middleware Center we are trying to create a complex medical system for diagnosis of heart disease. The system is called Clinical Decision Support System (CDSS) for Heart Disease and suppose to provide ECG based diagnosis of heart disease and possible treatment recommendation. The core part of this system is a Heart Disease Ontology. Heart Disease Ontology is complex ontology which tries to model the domain of heart disease from the various aspects such as types of heart disease and its possible treatments, structure of the heart and its muscles, ECG based heart disease diagnosis and etc.

Maintenance of such complex ontology is a complicated and challenging task not only for doctors (the main users of the system), but even for ontology engineers. Maintenance of ontology has proven to be not only demanding in terms of skill required, but also time consuming and costly[8]. This is especially true when it comes to a problem of updating existing knowledge within the ontology, where slight change can result in inconsistencies. Therefore solving the problem of maintenance of ontology is essential for the success of whole CDSS.

We are proposing a novel approach that delegates the ontology maintenance process from human to a computer. It provides computer with ability to automatically obtain new knowledge and updated ontology with it.

2. Related works

The Ontology Maturing Process (OMT) [7] is a model that identifies the characteristical maturing transition in collaboratively developing a shared ontology. More easily speaking OMT views the ontology engineering as a sustainable process of continuous evolution, and emphasizes the collaborative approach rather than putting sole responsibility on ontology engineering expert or domain expert.

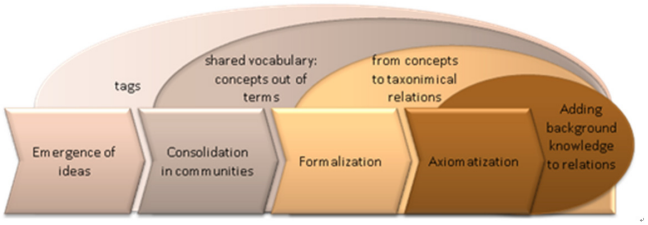

Figure 1. Ontology Maturing Process.

It describes how raw ideas such as simple tag based relations can be transformed into ontological knowledge. The whole process is defined by means of four main phases as you can see from the figure below:

- *Emergence of ideas* – emergence of new concept ideas in ad-hoc manner; ideas are informally communicated and technically typically represented by tags.
- *Consolidation in communities* – through reuse and adaptation of concept symbols (tags) of others, a shared vocabulary emerges within a community; when comparing currently envisioned tags with previously used ones or with tags from other people assigned to same resource, similarities and differences are discovered that allow for creating concepts.

- *Formalization* – within the third phase, concept are organized into relations- both taxonomical and ad hoc relations.
- *Axiomatization* – capturing more domain semantics by adding background knowledge for improving inferencing processes, for query answering, this phase requires high level of competence in logical formalism, requiring the domain expert.

Because OMT views whole ontology development process as a sustainable process with continuous evolution, it supports the update and correction of existing ontologies.

3. Ontology Maturing Process for Medical Domain

The OMT was originally created within the context of Web 2.0 domain which has readily available user community. Then this user community is actively involved in all four phases, especially in the phases one and two where users create ideas (tags) and then try to achieve consolidation through some type of negotiation process.

Problem with directly applying OMT for medical domain in cases such as ours is that we don't have active user community that can support phases of OMT. Therefore in order to apply OMT to medical domain the following questions must be answered:

1. **What are the alternative sources of ideas?**
2. **How to achieve consolidation without human community?**
3. **How to do the formalization and the axiomatization automatically?**

The answer to the first question is in medical documents. Documents are important artifacts in medical domain and contain valuable knowledge. In medicine, the documents are still serving as a major channel of saving and sharing the knowledge[4]. It is possible to say that the ideas emerge within the medical documents and therefore documents can be used as a major source of ideas in OMT.

Consolidation is particularly important phase where users try to negotiate with each other in order to identify proper ideas and create ontological concepts from those ideas. In medical domain such kind of consolidation is occurring in form of medical knowledge base. Lots of effort is put by experts to create common knowledge base for whole domain of medicine. One such knowledge base is Unified Medical Language System (UMLS) [3] which is compendium of many controlled vocabularies in biomedical science. Since UMLS provide some type of standardized knowledge in medical domain, it is possible to use UMLS for validating ideas extracted from the medical documents and creating ontological concepts. Therefore UMLS can be effectively used in phase two of OMT.

In original OMT the formalization and axiomatization phases are supposed to be done under human supervision, so alternative to human supervision should be found in order to do those steps automatically. Such alternative are UMLS Metathesaurus and UMLS Semantic Network. UMLS Metathesaurus is a collection of biomedical terms and concepts from various controlled vocabularies structured in taxonomical hierarchies. Those hierarchies can be exploited in formalization phase as an alternative to the human supervision. UMLS Semantic Network is a set of categories and entries that are used to classify and relate the entries in the UMLS Metathesaurus. UMLS Semantic Network is suitable for using in axiomatization phase for creating the non-taxonomical relationships[1].

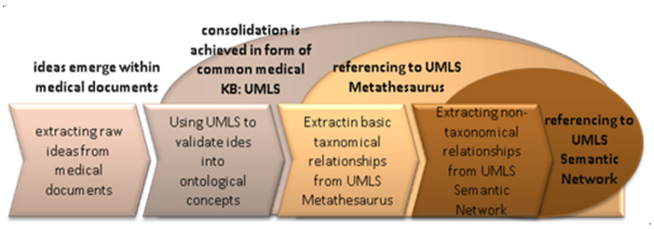

Figure 2. Modified Ontology Maturing Process for medical domain.

In a Figure 2 you can see four phases of modified OMT which was adapted to medical domain.

4. Proposed algorithm based on OMT

In the previous sections the OMT was described in terms of four phases shown in Figure 2. Those phases are rather superficial and general therefore need to be broken into more detailed steps in order to form a valid algorithm. Those steps are shown in Figure 3.

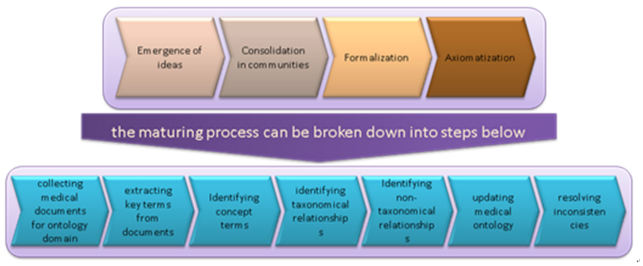

Figure 3. Steps within OMT for medical domain.

Although due to limitation in page number this paper will not describe in details each step shown in Figure 3, let us provide very brief description of them:

1. Collecting medical document for ontology domain – this step is involved with collecting the medical document which might contains useful knowledge that can be added to the existing ontology; it is

important to make sure that the domain of the document matches with the domain of ontology.

2. Extracting key terms from documents – given a medical document, through the use of NLP techniques[6] the keyterms and noun phrases are extracted from the document; those keyterms are not yet valid ontological concepts.
3. Identification of concept terms – given the set of keyterms extracted in previous step, those terms are validated through matching to the terms in UMLS Metathesaurus; this ensures that extracted keyterms are valid medical concepts that can be used in ontology.
4. Identification of taxonomical relationships – the previous step produces the set of ontological concepts, in this step the possible taxonomical relationships among those concepts are identified through the utilization of hierarchical structure within UMLS Metathesaurus;
5. Identification of non-taxonomical relationships – taxonomical relationships alone are not enough for building a good ontology, therefore the ontology should also contain non-taxonomical relationships; we are trying to use UMLS Semantic Network[1,3,5] and SemRep tool for identifying such relations among ontological concepts.
6. Updating medical ontology – the previous steps produces the set of ontological concepts with taxonomical and non-taxonomical relationships in it; this set is inserted into existing medical ontology; important thing to consider here is to find correct place within ontology for each newly inserted concept, this is done through utilization of ontology aligning methods
7. Resolving inconsistencies – it is possible that newly inserted concepts and relationships can result in inconsistencies within the ontology; easiest way to identify and remove such inconsistencies is to use ontological reasoner.

5. Conclusion

In this paper we have described a novel approach for automatically populating existing medical ontology with new knowledge extracted from medical documents. Although the approach was originally developed for domain of heart diseases, we believe that it can be easily applied to other medical domains. The preliminary tests of the proposed algorithm have shown promising results, but final implementation and test are still should be done as a future work.

This work was done as a part of bigger research on developing Clinical Decision Support System for Heart Diseases, which is still ongoing research.

6. Acknowledgement

This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement)(IITA-2009-C1090-0902-0014)

REFERENCES

[1] Wang W, Min Feng, Dawei Hu, Liu Wenyin, “Automatic Clinical Question Answering based on UMLS Relation”, IEEE Computer Society, Third International Conference on Semantics, Knowledge and Grids, 2007

[2] Erik Tjong Kim Sang, Gosse Bouma, Maarten de Rijke, “Developing Offline Strategies for Answering Medical Questions”, 2005, American Association for Artificial intelligence

[3] http://www.nlm.nih.gov/research/umls/

[4] Yun Niu, Graeme Hirst, “Analysis of Semantic Classes in medical text for Question Answering”, Department of Computer Science, University of Toronto, Canada

[5] Spela Vintar, Ljupco Todorovski, Daniel Sonntagm, Paul Buitelaar, “Evaluating Context for Medical Relation Mining”, European Workhop on Data Mining and text Mining for Bioinformatics, ECML/PKDD, 2003

[6] Quig T Z., Goryachev S., Wess S., Sordo M., Murphy S. N, Lazarus R.,“Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of natural language processing system”, BMC Medical Informatics and Decision Making, 2006

[7] Simone Braun, Andreas Schmidt, Andreas Walter, “Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering”, WWW2007, May 2007, Banff, Canada

[8] Nataya F. Noy, “Semantic Integration: Survey of ontology-based approaches”, SIGMOD Record, Vol. 33, No. 4, December 2004