Construction of a Condensed Thesaurus for Building Radiology Ontology

Seung-Bin Han, M.S., Jinwook Choi, M.D., Ph.D
Department of Biomedical Engineering, College of Medicine, Seoul National University, Korea
{sbinhan, jinchoi}@snu.ac.kr

Abstract

The building of thesauri for large domains, especially for medicine, is a costly affair. However, in many domains thesauri can be constructed on an ontological basis [Wielinga , Schreiber, 2001]. We are developing an ontological information retrieval system for the retrieving of medical records from an electronic medical record system (EMR). We decided to use the UMLS as a basis for building this system, but its many synonyms and complex relationships unnecessarily complicate the retrieval process. In this paper, we present a method of constructing a condensed thesaurus, which contains terms practically used in medical records. It is hoped that this condensed thesaurus will effectively support the document indexing and retrieval processes.

1 Introduction

The Unified Medical Language System (UMLS), which was developed by the National Library of Medicine, is a useful basis for the building medical ontologies. The main components of the UMLS are a Metathesaurus (MT) and a Semantic Network (SN). The purpose of the SN is to provide a consistent categorization of the concepts represented in the UMLS MT and a set of relationships between these concepts. The goal of the UMLS MT is to provide linkages between different vocabularies, and to expand terminologies for applications in such areas as information retrieval (IR) to facilitate efficient searching when terms are used by searchers that differ from those used by original authors. We are currently developing an information retrieval system based on applied ontology for retrieving medical records from an electronic medical record system (EMR).

Medical vocabularies in medical records involve the use of synonymous terms and expressions, and search mismatches may be caused by synonymy, where different terms have identical meanings, like hypertension and high blood pressure. Thus, query expansion using a thesaurus enhances the recall of medical information retrieval, whether searching patient records or the medical literature [Hersh, 2000]. For this reason we chose the UMLS as a useful thesaurus for our purposes. However, it contains too many synonyms to allow for such expansions, even for a single terms or concepts. Thus, we undertook to abridge the UMLS and to construct an IR efficient ontologically based system.

In this paper, we present a way of constructing a condensed thesaurus for use in IR systems.

2 Methods

2.1 Experimental Environment

For our experimental evaluations we used 40,000 Brain CT/MRI radiology reports at Seoul National University Hospital (SNUH) written between 1999.7~2003.1. Patient's names were removed for privacy. The texts were obtained in Korean and English and most of the keywords were medical terms written in English. We divided the whole collection into 8 different document sets of 5,000 documents. The reason for this division was to construct a thesaurus or ontology module capable of expansion from other corpora. On average, there were 84 index terms per document and about 7,100 unique index terms per 5000 document set. 74 queries were selected from frequent diagnoses in the document collection. Queries consist of one to three words like *meningioma*, *angina pectoris*, *or systemic lupus erythematosus*. In our experience, queries are composed of a set of terms presented in a Boolean structure.

2.2 Condensed Thesaurus Builder

We exploited the UMLS 2003AA version of the Metathesaurus which contains 875,255 concepts described by 2.14 million terms [NLM, 2003]. Queries were expanded according to their concept area of in MRCON table of UMLS MT since synonymous terms or expressions are linked to the same concept identifier (CUI).

As mentioned earlier, the UMLS MT contains too many synonyms and relationships to allow for query expansion. Thus, in this study, we developed a condensed thesaurus builder (CTB), which utilizes the UMLS, and which incorporates a query expansion methodology facility. As illustrated in figure 1, we prepared index files consisting of vocabulary lists and posting lists for the 8 document sets. And then we submitted 74 test queries to the CTB which expands each queries terms from UMLS. CTB recorded

term frequencies (tf) when expanded terms from the UMLS appeared in the document set. For example, for the query 'angina pectoris', the following MT synonyms appeared in a document set (n=5,000); angina-49 times, angina pectoris-30 times, and ischemic chest pain-twice, whereas anginal pain, angor pectoris, stenocardia, and anginal syndrome did not occur in the document set, though they are synonymous in the UMLS, thus, indicating substantial differences in the frequencies of the terms used.

Thus, we included in the CTB an instruction that a term must occur more than once in a document set (tf >=1). From each 8 document sets, each condensed thesauri (CT) were generated. Then Accumulator in CTB combined each and every CT without duplication (Figure 1). In this way, we constructed a condensed thesaurus, which consists of terms practically used in radiology reports in SNUH. This development represents the groundwork required to build a radiologic ontology which can be applied to IR systems.

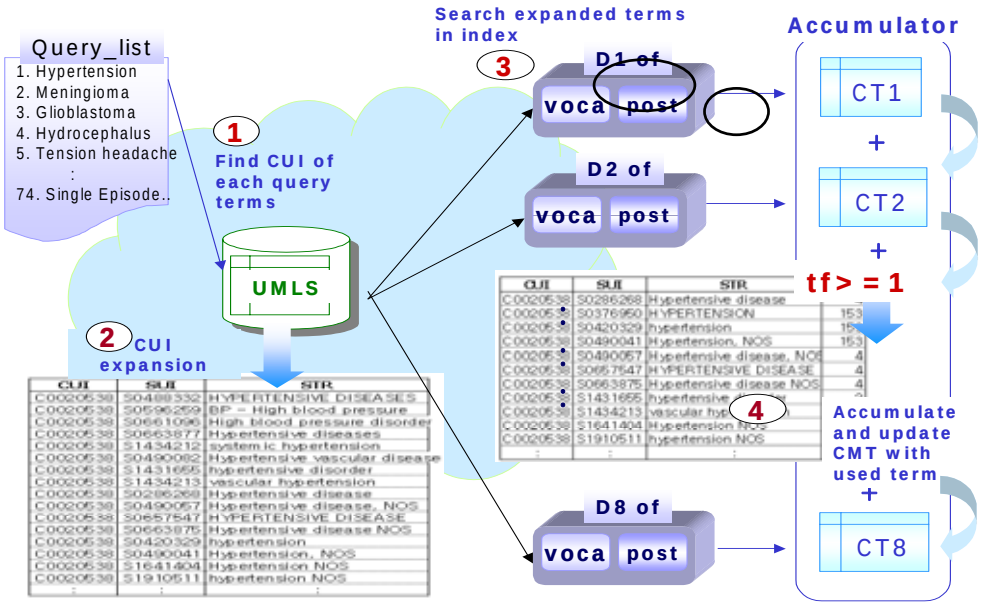

Figure 1. The process of Building the Condensed Thesaurus

Then we compared the retrieval effectiveness of three cases, described as follows:

- MT(-) : IR system has no thesaurus for query expansion, this case provided a baseline.
- MT(+) : IR system applied the whole MT for synonymy expansion.
- CT : IR system used our retrieval model containing a condensed thesaurus.

In our experiment, the retrieved documents of MT(+) outperformed non MT implemented system in terms of document retrieval. Query expansion using the thesaurus substantially improved the search performance in IR system by offering conceptual search. In the third case, however, CT had retrieved as much as MT(+). However, MT(+) took 3.5 times longer to retrieve documents that the MT(-) IR system. When we applied CT to the IR system, the retrieval time reduced by 50% versus MT(+) implemented IR system.

The results show that condensed thesaurus which is constituted of practically used terms in medical records enhanced the retrieval performance and improved the expansion time.

3. Conclusion and Future Work

The UMLS provides a comprehensive basis for the building of medical ontologies, but the relations it contains are too complicated, and these hinder its application. Here, we present a way of condensing the thesaurus to one containing practically used terms. We will later link relationships between extracted terms in a condensed thesaurus using the MT and SN of UMLS. This condensed thesaurus offers an effective method of query expansion in IR systems, and can reduce the effort and cost of building domain ontologies.

Acknowledgments

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (02-PJ1-PG6-HI03-0004).

References

[Wielinga , Schreiber, 2001] B.J. Wielinga, A.Th. Schreiber, J. Wielemaker, From Thesaurus to Ontology J.A.C. Sandberg, 2001.

[Hersh, 2000] W. R. Hersh, S. Price, L. Donohoe, Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus, Proceedings of AMIA Fall Symposium 2000, pp344-348.

[NLM, 2003] Unified Medical Language System, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, National Institutes of Health. http://www.nlm.nih.gov/research/UMLS/