World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:8, 2013

Designing Ontology-Based Knowledge Integration for Preprocessing of Medical Data in Enhancing a Machine Learning System for Coding Assignment of a Multi-Label Medical Text

Phanu Waraporn

***Abstract*—**This paper discusses the designing of knowledge integration of clinical information extracted from distributed medical ontologies in order to ameliorate a machine learning-based multi-label coding assignment system. The proposed approach is implemented using a decision tree technique of the machine learning on the university hospital data for patients with Coronary Heart Disease (CHD). The preliminary results obtained show a satisfactory finding that the use of medical ontologies improves the overall system performance.

***Keywords*—**Medical Ontology, Knowledge Integration, Machine Learning, Medical Coding, Text Assignment.

I. Introduction

We present a knowledge integration method based on the use of distributed medical ontologies and machine learning techniques to enhance a coding assignment of multi-label medical text. Text Categorization or Text Assignment (TA) as part of the Natural Language Processing (NLP) consists the assignment of one or more preexisting categories to a text document [1]. In multi-label assignment, the problem can comprise various classes.

As large unstructured and structured medical databases are being generated momentarily, difficulties accessing, integrating, extracting, and managing knowledge out of them are among many reasons researchers are trying to overcome including utilizing ontologies, a form of knowledge-based systems which are repositories of structured knowledge such as UMLS, MeSH, ICD, etc.

According to Nelson et al. [2], several studies have shown that the use and integration of several knowledge sources improves the quality and efficiency of information systems using the query on the ontology, specifically so in the domain specific such as the health information systems or medicine. An ontology is a specification of a conceptualization that defines and/or specifies the concepts, relationships, and other distinctions that are relevant for modeling a domain. Such specification takes the form of the definitions of representational vocabulary (classes, relations, and so on), which provide meanings to the vocabulary and formal constraints on its coherent use [3]. In compliance with the WHO ICD 10 for Coronary Heart Disease [4], a domain specific ontology based on a distributed architecture is constructed for use in the work.

The rest of the paper is arranged as follows: first, a background section comments on some relevant works in the field; Section III briefly introduces the proposed system architecture; Section IV summarizes the data collection used for pre-processing works. We deliver the conclusions and future work in the epilogue of Section VI.

Phanu Waraporn is with the Department of Computer Science, Faculty of Science & Technology, Suan Sunan Rajabhat Univeristy, Bangkok, Thailand.

II. Background

Our aim of enhancing the automatic assignment of medical coding by automatically integrating ontologies is guided by the use of ontologies in various natural language processing tasks such as automatic summarization, text annotation and word sense disambiguation, among other [5]. Advantage of using ontologies in the area of relevance-feedback, corpus-dependent knowledge models and corpus-independent knowledge models on the domain-specific and domain-independent ontologies all contribute to ameliorate information retrieval systems [6].

Domain-independent ontologies such as Word Net/Medical Word Net though improves a word sense disambiguation, it has so broad coverage that it can be debatable for the ambiguous terms making a domain-specific ontologies, particularly on the part of a terminology which is less ambiguous. Furthermore, it models terms and concepts corresponding to a specific or given domain [7].

III. Proposed System Architecture

Patient records have many components but three areas are fundamental for manually assigning of the ICD 10 codes:

1. A Physician's Discharge Summary Report/Notes including but not limited to Nurses' Notes, etc.
2. A laboratory results
3. A drug usage information

In replication to the human coders gathering of information for use in giving out the ICD code above, the proposed system is depicted in the Fig. 1. This is the current system architecture in use for testing of the automatic assignment of medical coding. The Fig. 2 illustrates the ICD 10 specific to the Coronary Heart Diseases (CHD) or interchangeably the Ischaemic Heart Diseases (ISD) in a hierarchical format.

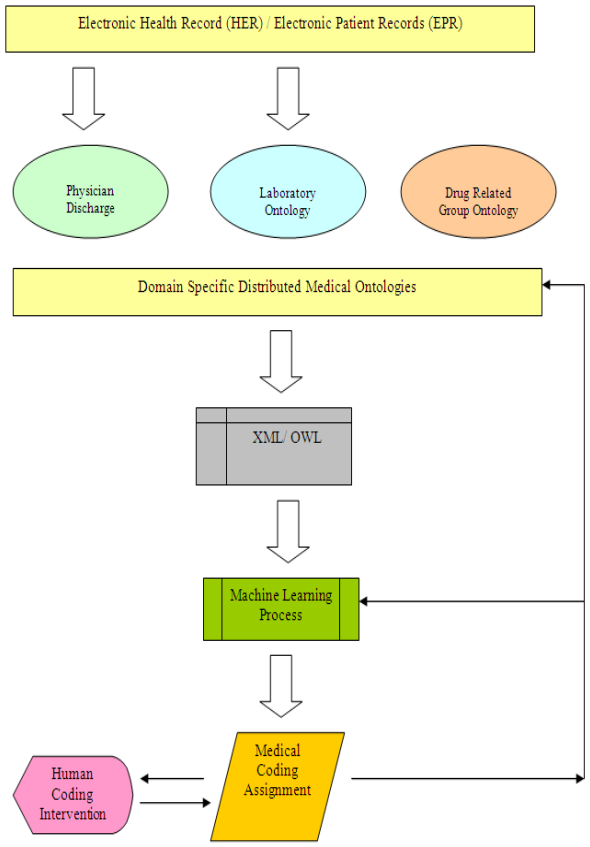

Fig. 1 Proposed System Architecture for automatic assignment of medical coding

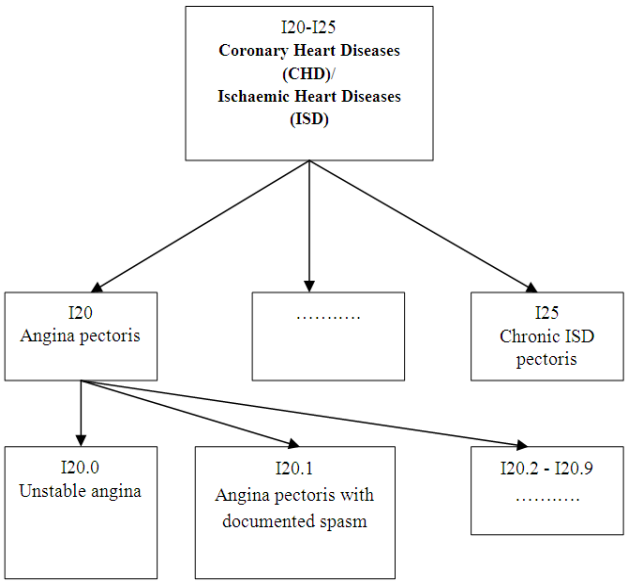

Fig. 2 ICD 10 Hierarchical structure of Coronary Heart Diseases (CHD)

IV. Understanding Medical Data Prior to Pre-Processing of Machine Learning Algorithm

In this section, we present, from a sample of data, its attributes, extracted from distributed medical ontologies. Some attributes are appeared but do not strengthen the process of automatic assignment of ICD 10 code. However, they are significant from an ontological point of view as they are part of the concepts and relations necessary for use in the ontological engineering process [8]. Fig. 3 shows the XML/OWL equivalents for the sample data.

	Onset	Previous CAD	New or symptom worsen	History of CHF	Chest pain	Enzyme	EKG	Site of EG change	Chest X-Ray	Echocardiogram Main	Echocardiogram Sub-Wall Motion Abnorm	Drug 1	Drug 2	Drug 3	Drug 4	Cath result	Outcome
I200	1	1	1	1	1	1	3	1	3	2	2	1	1	1		4	1
I201	1	0	1	0	1	1	4	2	1	1	0	1	0	0		0	1
I201	1	0	1	0	1	1	4	2	1	1	0	1	0	0		0	1
I209	1	1	0	0	1	1	1	0	1	1	0	0	0	0		0	1
I209	2	0	0	0	1	1	1	0	1	1	0	0	0	0		0	1
I210	1	0	1	0	1	2	4	2	2	2	2	1	0	1		4	2
I210	1	0	1	0	2	2	5	2	1	2	2	1	0	1		2	1
I211	1	1	1	1	3	2	4	1	3	2	2	1	1	1		3	1
I211	1	0	1	0	2	2	4	1	1	2	2	1	0	1		1	1
I212	1	1	1	0	1	2	4	3	0	2	2	1	0	1		1	1

Fig. 3 Sample Patient data extracted from XML/OWL based medical ontology in a modified tabular format

V. Conclusion and Future Works

A prototype system has given satisfactorily a preliminary result that is used in further advanced experimental work on machine learning techniques. After extensive laboratory works are being carried out and human expert/medical professional coders' confirmations of the final result, a system will be rolled out to other university hospital in the Bangkok area for testing before rolling out on a wide scale. However, since the proposed system is based on a particular disease, this generic model will need to be substantiated for other vital diseases as well.

ACKNOWLEDGMENTS

We would like to extend our sincere thanks to the Mahidol University's Faculty of Medicine Siriraj Hospital Division of Molecular Genetics, National Center for Genetic Engineering and Biotechnology (Biotec) and Human Language Technology Laboratory of the National Electronics and Computer Technology Center (NECTEC) for all resources and advices deployed in this preliminary study and experimentations. In addition, many thanks to Mizoguchi Laboratory of the Institute of Scientific and Industrial Research, Osaka University for free accesses to Hozo, an ontology editor.

REFERENCES

[1] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv., 34*(1), 1-47.

[2] Nelson, S.J., et al. (2001). Relationships in medical subject headings. In C.A. Bean & R. Green (Eds.), *Relationship in the Organization of Knowledge*. New York: Kluwer Academic Publishers, (pp. 171-184).

[3] Gruber, T. (1995). Toward Principles for the Design of Ontologies used for Knowledge Sharing. *International Journal of Human-Computer Studies, 43*, 907-928.

[4] WHO. World Health Organization. www.who.int, www.who.int/ classifications/icd/en/

[5] Martin-Valdivia, M.T. (2009) Expanding terms with medical ontologies to improve a multi-label text categorization system. In P.Violaine and R. Mathieu (Eds.), *Information Rerieval in Biomedicine: Natural Language Processing for Knowledge Integration,* (pp 38-57).

[6] Bhogal, J., et al. (2007) A review of ontology based query expansion. *Information Processing & Management, 43*(4), July 2007, 866-886.

[7] Waraporn, P. (2008). Proposed framework for interpreting medical diagnosis records using adopted WordNet/Medical WordNet. *Proceedings of Technology and Innovation for Sustainable Development Conference (TISD2008), 05_004_2008I,* 433-436.

[8] Waraporn, P. (2008). Distributed Ontological Engineering and Integrated Development of Medical Diagnosis Coding Ontology for State Hospitals in Thailand, *Proceedings of National Conference on Computer and Information Technology (NCIT 2008)*.