Development of Experience Base Ontology to Increase Competency of Semi-automated ICD-10-TM Coding System

Wansa Paoin*, Supot Nitsuwat**

Abstract

The objectives of this research were to create the International Classification of Diseases, 10th edition, Thai Modification - ICD-10-TM experience base ontology, to test usability of the ICD-10-TM experience base with knowledge base in a semi-automated ICD coding system, and to increase competency of the system. ICD-10-TM experience base ontology was created by collecting 4,880 anonymous patient records coded into ICD codes from 32 volunteer expert codes working in different hospitals. Data were checked for misspelling and mismatch elements and converted into experience base ontology using n-triple (N3) format of resource description framework. The semi-automated coding software could search experience base when initial searching from ICD knowledge base yielded no result. Competency of the semi-automated coding system was tested using another data set contain 14,982 diagnosis from 5,000 medical records of anonymous patients. All ICD codes produced by the semi-automated coding system were checked against the correct ICD codes validated by ICD expert coders. When the system use only ICD knowledge base for automated coding, it could find 7,142 ICD codes (47.67%), recall = 0.477, precision =0.909, but when it used ICD knowledge base with experience base search, it could find 9,283 ICD codes (61.96%), recall = 0.677, precision = 0.928. This increase ability of the system was statistical significant (paired T-test p-value = 0.008 (< 0.05). This research demonstrated a novel mechanism to use experience base ontology to enhance competency of semi-automate ICD coding system. The model of interaction between knowledge base and experience base developed in this work could be used as a basic knowledge for development of other computer systems to compute intelligence answer for complex questions as well.

Keyword: experience base, knowledge base, ontology, semi-automated ICD coding system

1. Introduction

Ontology is a data structure, a data representation tool to share and reuse knowledge between artificial intelligence systems which share a common vocabulary. Ontology could be used as a knowledge base for computer system to compute intelligence answer for complex questions like ICD-10-TM (The International Classification of Diseases and Related Health Problems, 10th Revision, Thai Modification) [1] coding.

ICD-10 is a classification that was created and maintained by the World Health Organization –WHO since 1992 [2]. The electronic versions of ICD-10 was released in 2004 as a browsing software in CD-ROM package [3] and as ICD-10 online on WHO website [4]. Both electronic versions provided only a simple word search service that facilitate only minor part of the complex ICD coding processes. Since 2000 AD, some countries add more codes from medical expert opinions into ICD-10 so ICD-10 was modified in some countries e.g. Australia, Canada, Germany etc. In Thailand ICD-10 was modified to be ICD-10-TM (Thai Modification) since 2000 [5] and is maintained by Ministry of Public Health, Thailand.

ICD coding is an important task for every hospital. After a medical doctor complete treatment for a patient, the doctor must summarized all diagnosis of the patient into a form of diagnosis and procedures summary. Then a clinical coder will start ICD coding for that case using manual ICD coding process which use two ICD books as reference sources. All ICD codes for each patient will be used for morbidity and mortality statistical analysis and reimbursement of medical care cost in hospital. Manually ICD coding processes are complex. The ICD coding could not be finished merely by word matching between diagnosis words and list of ICD codes/labels, a clinical coder may assign two different ICD codes for two patients with same diagnosis word based on each patient context. Unfortunately, this complexity of ICD-coding were not recognized by most researchers who tried to develop semi-automated and automated ICD coding systems in the past.

Several research works mentioned automated ICD coding process in their researches. Diogene 2 program [6] built medical terminology table and used it to map diagnosis word into morphosemantem (word-form) layer, then converted the term into concept layer before matching to labels of ICD codes in expression layer. Heja et al [7] did matching diagnosis words with list of ICD code labels and suggested that hybrid model yield better matching results. Pakhomov et al [8] designed an automated coding system to assign codes for out-patient diagnosis using example-based and machine learning techniques. Periera et al [9] built a semi-automated coding

* *Faculty of Information Technology, King Mongkut's University of Technology North Bangkok.*

** *Faculty of Applied Science, King Mongkut's University of Technology North Bangkok.*

help system using an automated MeSH-based indexing system and a mapping between MeSH and ICD-10 extracted from the UMLS metathesaurus. These previous works, only used word matching approach processes and never covered full standard ICD coding processes, which had been summarized in ICD-10 volume-2 [10].

In our previous work [11] we had created ICD-10-TM ontology as a knowledge base for development of semi-automated ICD coding. ICD-10-TM ontology contains 2 main knowledge bases i.e. tabular list knowledge base and index knowledge base with 309,985 concepts and 162,092 relations. Tabular list knowledge base could be divided into upper level ontology, which defined hierarchical relationship between 22 ICD chapters, and lower level ontology which defined relations between chapters, blocks, categories, rubrics and basic elements (include, exclude, synonym etc.) of the ICD tabular list. Index knowledge base described relation between keywords, modifiers in general format and table format of the ICD index.

ICD-10-TM ontology was implemented in semi-automated ICD-10-TM coding software as a knowledge base. The software was distributed by the Thai Health Coding Center, Ministry of Public Health, Thailand [12]. The coding algorithms will search matching keywords and modifiers from the index ontology and diagnosis knowledge base, then verify code definition, include and exclude conditions from tabular list ontology. The program will display all ICD-10-TM codes found or not found to the clinical coder, then the human coder could accept the codes or change to other codes based on her judgment and standard coding guideline. Users survey revealed good results got from ontology search with high user satisfaction (>95%) on well usability of the ontology.
When we tried to use the system to do automate coding i.e. to code all diagnosis before a clinical coder start coding, to reduce number of diagnosis to be coded by clinical coder, we found that the automated coding work based on the ICD-10-TM ontology could successfully code diagnosis words for 24-50% of all diagnosis words. To increase competency of the system, we created another ontology call “experience base” to help the system to be able to code more diagnosis words than previously done.

In this paper, we present ICD-10-TM experience base and the application of a novel mechanism using experience base ontology to enhance competency of semi-automate ICD coding system. The model of interaction between knowledge base and experience base developed in this work could be used as a basic knowledge for development of other computer systems to compute intelligence answer for complex questions as well.

2. Methodology

To create knowledge base, we asked all expert coders in Thailand to volunteer participate in this project. An expert coder must had at least 10 years experience on ICD coding, or had passed the examination for certified coder (intermediate level) by the Thai Health Coding Centers, Ministry of Public Health to be able to participate. The project committee selected 42 expert coders from 198 volunteers based on their ability to devote time for the project, hospital size, hospital location where the coders work and competency on using computer and software.

All selected expert coders attended one day training on how to use the semi-automated coding system. Each of them was assigned to use the system to do ICD coding. They used medical records of patients admitted into their hospital during January to November 2011 as input to the system. The input data did not include patient identification data. Only sex, age and obstetrics condition of each patient must be input into the system since these data elements, as well as all diagnosis words, are essential for ICD code selection by the system. Each expert coder must input at least 100 different cases into the system within 30 days. After finishing their task, each coder sent the saved data to the project coordinator by email. Data from all expert coders were checked for misspelling and mismatch elements (for example, a male patient could not be obstetrics case). Records of patient type with each diagnosis word and ICD code from every cases were created using n-triple (N3) format of resource description framework – RDF [10] to built the experience base ontology. The ontology was built into the system using inverted index structure by transforming into Lucene 3.4 [13] search engine library which is the core engine of the semi-automated ICD coding system. The new semi-automated coding system now has another ontology - ICD experience base created from expert coders work. The automated coding algorithm had one new step. This step will be executed when searching from ICD knowledge base yielded no result. When ICD code was not found after searching from ICD knowledge base, the system will search from ICD experience base. Sometimes ICD code of a diagnosis with the same patient context varies from one expert opinion to another, the system will select the ICD code with highest frequency of expert opinion.

Competency of the semi-automated coding system was tested using another set of patient data. This dataset contains 14,982 diagnosis from 5,000 medical records of patients admitted during January to June 2011, into another hospital which did not participate in the knowledge base creation. Every ICD codes in this dataset were validated for 100% accuracy by another three expert coders. All ICD codes produced by the semi-automated coding system when using knowledge base only and when using knowledge base with experience base were checked against the correct ICD codes in the dataset for accuracy.

3. Results

By the end of the project 4,880 diagnosis words and patient context were collected from 32 expert coders. Ten expert coders did not send the cases within the dateline, so their data were excluded from analysis in this phase. All 4,880 diagnosis words and patient context were used to created the experience base ontology. A python script written and used to transform each record from comma separate value file format to RDF N3 files.

The experience base ontology contains five concepts and four relations as shown in Table 1. Each diagnosis word in a patient record could be uniquely identified. Each ICD expert opinion on the ICD code that should be used for each diagnosis word based on the patient context was an important concept in the ontology. All these concepts and relations were used to construct all RDF statements in the experience base ontology. For example if an expert 'abc123@mymail.com' gave an opinion that a diagnosis word 'disseminated tuberculosis' in a patient context 'man not newborn' should be coded to ICD code 'A18.3', the RDF statements in the N3 format will be written as the following phrase;

dxword: disseminated_tuberculosis word:hasPtDxId ptdxid:001.

ptdxid:001 pt:isA ptcontext: man_not_newborn .

ptdxid:001 icd:codeBy expert:abc123@mymail.com .

ptdxid:001 icd:hasCode icd10:A183 .

The experience ontology concepts and relations can be presented as a graph data as in Figure 1.

Table 1. *All experience base concepts and relations in RDF N3 format.*

Experience Base Concepts	*Ontology type*	*RDF Format*	*Example*
Diagnosis Word	Concept	dxword:	dxword: disseminated_tuberculosis
PatientDiag ID	Concept	ptdxid:	ptdxid:001
Patient Context	Concept	ptcontext:	ptcontext:man_not_newborn
Expert	Concept	expert:	expert:abc123@mymail.com
ICD10 Code	Concept	icd10:	icd10:A183
hasPtDxId	Relation	word:hasPtDxId	dxword:dyslipidemia word:hasPtDxId ptdxid:101
isA	Relation	pt:isA	ptid:101 pt:isA ptcontext: man_not_newborn
codeBy	Relation	icd:codeBy	ptid:101 icd:codeBy expert:abc
hasCode	Relation	icd:hasCode	Ptdxid:101 icd:hadCode icd10:E78.5

The system was used to automate coded 14,982 diagnosis in the test dataset. When the system use only ICD knowledge base, it could find 7,142 ICD codes (47.67%), but when it used ICD knowledge base with experience base search, the system could find 9,283 ICD codes (61.96%). This increase ability was tested for statistical significant using paired T-test with alpha value = 0.05. T-stat = -79.30 with p-value = 0.008 (< 0.05).

Recall and precision of the system were calculated. The recall and precision value when the system used ICD knowledge base only were 0.477 and 0.909 , while the recall and precision value when the system used ICD knowledge base with experience base were 0.677 and 0.928.

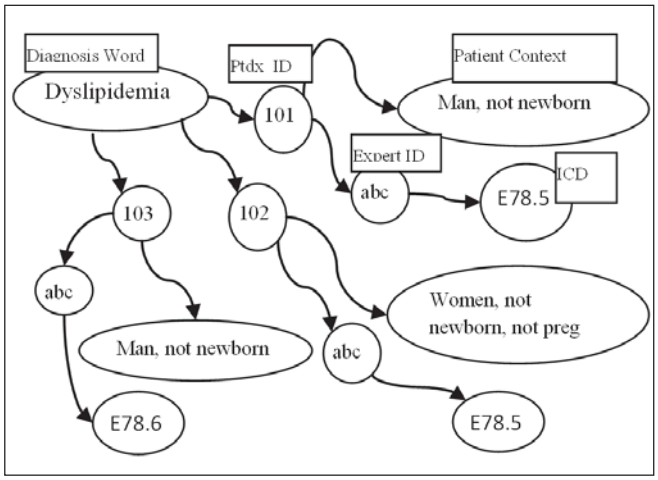

Figure 1. A part of the ICD experience base. A diagnois word "Dyslipidemia" in each patient record could be code to various ICD codes, based on each expert opinion and each patient context.

4. Discussion

ICD-10 coding is not a simple word matching process. Qualified human ICD coders will never do simple diagnosis word search or browse the diagnosis term from a list of ICD codes and labels. Unfortunately, research on semi-automated and automated ICD coding system in the past [6-9] never recognize this important concept. This finding explained why there is no real workable automated ICD coding system until now.

ICD index and tabular list of disease were created since 1992, diagnosis words in ICD did not include every synonym, alternative name or some specific diagnosis in highly specialized medical service. On the other hand, ICD classification added some patient context into classification scheme, this made coding for one disease name may produced different ICD codes if the patient context change. For example an ICD code for diagnosis "internal hemorrhoids" would be O22.4 when the patient was a pregnant woman, but the code will be I84.2 for an adult man patient. These facts made ICD coding a complex job and need human coders. A clinical coder must know how to change some diagnosis word when first round searching could not find the code. She must had patient records in hand all the time she was coding to check necessary patient context that may affected correct ICD code choosing.

Our semi-automated ICD coding system was not developed to replace all the clinical coders work on ICD coding. But if the system could find initial ICD codes for some diagnosis word summarized by the medical doctor, the coder works will be reduced in some extents. Our system used ICD ontology created from ICD-10-TM alphabetical index and tabular list of disease as knowledge bases to search for correct ICD code for each diagnosis word + patient context. Automated coding base on

this knowledge could code 47.67% of all diagnosis with good accuracy (90.9%).

Recall ability of the old system was low because in real world medical records there are many varieties of words that the doctors may used for diagnosis. Some words are new words which occured after ICD-10 creation, for examples "dyslipidemia, chronic kidney disease, diabetes mellitus type 2" are more common used by doctors today than the old words "hyperlipidemia, chronic renal failure, non-insulin dependent diabetes mellitus" found in ICD-10.

Adding experience base created from real world cases into the system could increase recall ability of the system. ICD experience base ontology contains diagnosis words from real medical records with assigned ICD codes for these new words. So the system will search the experience base if first round searching from knowledge base yield no ICD code. Recall ability of the system increased from 0.477 to 0.677 with good precision ability (0.928).

Different expert opinions for same diagnosis were anticipated to be found in the experience base. In facts a consensus of expert opinion was rarely found in ICD coding experience base. Varieties of expert opinions on coding of some diagnosis words were shown in Table 2. The system will choose code with the highest frequency to be used as a "correct" code. This strategy should be good unless there were too few opinions for some rare diagnosis words.

Table 2. *Expert opinion of some diagnosis word in ICD experience base ontology.*

Diagnosis words	***ICD codes from expert opinion***	***Highest frequency code***
Dyslipidemia	E78.5, E78.6, E78.9	E78.5 (64.5%)
Chronic kidney disease	N18.0, N18.9, N19	N18.9 (35.5%)
Triple vessels disease	I21.4, I25.1, I25.9, N18.9	I25.1 (80%)
Diabetes mellitus type 2	E11.9, E11	E11.9 (95.8%)

Although the ICD experience base ontology at this stage contains only 4,880 cases. This experiment encouraged usage of experience ontology to increase recall ability of the semi-automated ICD coding system. In future research work, we plan to add more cases into the experience base and will try to test the ability of the system with more test data.

5. Conclusion

ICD experience base ontology could be created using ICD codes from medical records which was coded by expert coders. This experience base ontology was implemented into the semi-automated ICD coding system. Searching from experience base was very useful when first round searching from knowledge base yielded no result. The recall ability of the system could be increased by adding experience base searching into its algorithm with good precision ability still was preserved.

6. Acknowledgment

This research was supported by the Thai National Health Security Office, Thai Health Standard Coding Center (THCC), Ministry of Public Health, Thailand and Thai Collaborating Center for WHO-Family of International Classification.

7. References

[1] Bureau of Policy and Strategy, Ministry of Public Health, International Statistical Classification of Disease and Related Health Problems, 10th Revision, Thai Modification (ICD-10-TM). Nonthaburi, Thailand: The Ministry of Public Health, 2009.

[2] The World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 10th Revision. Geneva, Switzerland: The World Health Organization, 1992.

[3] The World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 10th Revision, 2nd Edition. Geneva, Switzerland: The World Health Organization, 2004.

[4] The World Health Organization. ICD-10 online [internet]. Geneva, Switzerland: The World Health Organization; 2011 [cited 2011 Jun 30]. Available from http://www.who.int/classifications/icd/en.

[5] Bureau of Policy and Strategy, Ministry of Public Health, Thailand. International Statistical Classification of Disease and Related Health Problems, 10th Revision, Thai Modification (ICD-10-TM). Nonthaburi, Thailand: The Ministry of Public Health, Thailand: 2000.

[6] C. Lovis, R. Buad, A.M. Rassinoux, P.A. Michel and J.R. Scherrer, "Building medical dictionaries for patient encoding systems: A methodology,"*Artificial Intelligence in Medicine*. Heidelberg: Springer, pp. 373-380, 1997.

[7] G. Heja and G. Surjan, "Semi-automatic classification of clinical diagnoses with hybrid approach," *Proceedings of the 15th symposium on computer based medical system - CBMS 2002, IEEE Computer Society Press*, pp. 347-352, 2002.

[8] S.V.S. Pakhomov, J.D. Buntrock and C.G. Chute. "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," J Am Med Inform Assoc, Vol.13, pp. 516 -525, 2006.

[9] S. Periera, A. Neveol , P. Masari and M. Joubert, "Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding" in A. Hasman et al, editors. Ubiquity: Technologies for Better Health in Aging Societies, VA: IOS Press pp. 845-850, 2006.

[10] The World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th Revision, 2nd Edition, Volume 2. Geneva, Switzerland: The World Health Organiztion, pp. 32, 2004.

[11] S. Nitsuwat and W. Paoin, "Development of ICD-10-TM ontology for semi-automated morbidity coding system in Thailand" Methods of Information in Medicine, in press.

[12] Semi-automated ICD-10-TM coding system [internet]. Nonthaburi, Thailand: The Thai Health Coding Center, Ministry of Public Health, Thailand; [cited 2011 Aug 12]. Available from : http://www.thcc.or.th/formbasic/regis.php.

[13] RDF Notation 3 [internet]: The World Wide Web Consortium; [cited 2011 Jun 12]. Available from: http://www.w3.org/DesignIssues/Notation3.

[14] Apache Lucene [internet]: The Apache Software Foundation; [cited 2012 Jan 24]. Available from http://lucene.apache.org/java/docs/index.html.