International Journal of Advances in Computer Science & Its Applications – IJCSIA
Copyright © Institute of Research Engineers and Doctors
Volume 9 : Issue 2- [ISSN : 2250-3765] - Publication Date: 27 Dec, 2019

Development of Diagnosis Directory and Algorithm to Facilitate Physician Diagnosis Mapping with Concepts in International Classification of Diseases

Wansa Paoin

***Abstract*—The International Statistical Classification of Diseases and Related Health Problem 10th Revision (ICD-10) was used extensively in all public hospitals in Thailand. All outpatients and inpatient diagnosis must be coded to ICD-10. Every year 250 million ICD-10 codes were sent to the information department at Ministry of Public Health Thailand. Manual coding by clinical coders could not be completed in some hospitals. Semi-automated ICD-10 coding software is needed to overcome this obstacle. Development of diagnosis directory and algorithm to facilitate physician diagnosis mapping with concepts in ICD-10 is the first step before semi-automated ICD-10 coding software development. The diagnosis directory in this work was built upon two sources of diagnosis list. The first source is the previous diagnosis list from the author research work in 2009. The second source is the diagnosis list built from ICD-10 alphabetical index. Final check of the diagnosis directory was performed by comparison with the Unified Medical Language System. The mapping algorithm steps are; 1) abbreviation conversion 2) stop words removed or replaced 3) spell checking 4) word counting and selection of appropriate set of diagnosis statement 5) similarity measurement 6) output of results. The diagnosis directory contains 43,331 diagnosis statements and could be divided into 10 subsets. The algorithm was tested using OPD diagnosis data from one community hospital of the Ministry of Public Health. The data contain 400 diagnosis statements. The algorithm could map 252 (63.0%) original diagnosis statements to diagnosis directory with 152 (38.0%) exact match (Cosine similarity = 1.0). this finding mean that in the future we can use the algorithm to reduce 63% of the manual work done by clinical coders.**

***Keywords*—ICD-10 codings, diagnosis direcoty, diagnosis mapping algorithm, semi-automated ICD-10 coding**

I. Introduction

The International Statistical Classification of Diseases and Related Health Problems 10th revision or ICD-10 is the classification of diseases system developed and maintained by the World Health Organization – WHO since 1946. The current edition is the 5th edition released in 2016 [1]. Many countries in the world use ICD-10 as the coding system to capture diagnosis statement in the patient records and produce morbidity and mortality statistics and compare health situation among countries in the same region.

Since 1994, Thailand used ICD-10 in all public hospitals to capture all diagnosis statements for every out-patient department (OPD) visits as well as every in-patient department (IPD) admission episodes. Since 2015, all individual patient data with ICD-10 codes from all public health hospitals were reported to the Ministry of Public Health, Thailand monthly. Each year 230 million OPD ICD-10 codes and 24 million IPD ICD-10 codes were sent to the Ministry for analysis and statistics production.

Current ICD-10 coding in hospitals require a huge amount of work. Most of ICD-10 coding work were done by clinical coders in the hospital. However, the number of competent clinical coders in Thailand is not enough to supply all hospitals. So some hospitals assigned the ICD-10 coding job to some physicians or nurses which may produce many coding errors due to lack of knowledge in clinical coding rules.

The automatic or semi-automatic ICD-10 coding system should be developed to replace the manual methods of ICD-10 coding in hospitals. In order to create the system, the diagnosis directory and algorithm to facilitate physician diagnosis mapping with concepts in ICD-10 must be developed first.

II. Literature Reviews

A. *Diagnosis Statement Lists*

The first international diagnosis statement list was the International Nomenclature of Disease (IND) developed by the Council for International Organization at the World Health Organization in Geneva Switzerland in the 1970s [2]. However, the compilation of all diagnosis statement could not be finished as plan.

The second international diagnosis statement list came from adaptation of the Systematized Nomenclature of Medicine (SNOMED) [3] which was a product created from the United States of America Systemized Nomenclature of Pathology. From 2007, the International Health Terminology Standards Development Organization (IHTSDO) owned and maintained SNOMED to be used by many countries. As of January 2019, it has 349,548 concepts including body structure, event, pathogens, drug etc. besides the diagnosis concepts. Usage of SNOMED-CT in any applications need license from SNOMED International. This requirement limit the opportunity to use the system in developing countries.

The Unified Medical Language System (UMLS) is a medical term lists that was developed and maintain by the U.S. National Library of Medicine (NLM) [4]. The system contains

Wansa Paoin MD. Ph.D. *(Author)*
Faculty of Medicine, Thammasat University
Thailand

SEEK DIGITAL LIBRARY

5 million biomedical concepts name. The license for usage of the metathesaurus could be obtained free from NLM for non-commercial use.

The ICD-10 concepts are different from the other terminology systems. Each concept of ICD-10 represent a class of disease, so only 14,200 concepts was published in volume 1 of ICD-10. Classification of disease in ICD-10 was based on patient context and diagnosis statement [5]. So one diagnosis statement of different patient context could be mapped to different ICD-10 codes. For example; a diagnosis statement of *acute cystitis* could be mapped to N30.0 for a male patient but the same diangosis *acute cystitis* must be mapped to O23.1 for a pregnant woman.

The first diagnosis statement list for Thailand was created by the author in 2009 [6]. Data from discharge summary of the patients from 11 hospitals during July 2004 to August 2005 were used to build a common diagnosis statement list comprised 23,024 diagnosis statements.

B. *Diagnosis Statement Text Processing*

Text mining or text analytics is the process of deriving high-quality information from text. In order to get information from text effectively, we need natural language processing algorithms so a computer program can interpret what was written in natural text using computation linguistic methods.

Diagnosis statement text processing can use the same building blocks for natural text processing [7]. Main steps are 1) Segmentation and Tokenisation 2) Lemmatization 3) Stemming 4) Compound splitting 5) Abbreviation detection and expansion 6) Spell checking and spelling error correction. However, diagnosis statements contain more noise in the form of incomplete word, misspelled word and non-standard abbreviation that can make the text processing difficult.

Text similarity analysis measure how two entities of text are close or far apart from each other. To do statement similarity analysis between the input diagnosis statement and diagnosis statement in the diagnosis directory, we need measurements methods which support analyzing of noun phrases similarity [8]. There are three metrics for this purpose i.e. cosine similarity, Hellinger-Bhattacharya distance and Okapi BM25 ranking.

The cosine similarity method gives us the cosine angle between two diagnosis statement represent in the vectorized forms. Two diagnosis statement vectors having similar orientation will have scores closer to 1 (cos0°).

The Hellinger-Bhattachrya distance (HB distance) could be used to measure frequency distributions distance between two diagnosis statement so a value of 0 indicates perfect similarity and high value indicate some dissimilarity.

The Okapi BM25 ranking is a document ranking and retrieval function bases on a Bag of Words-based mode. The input diagnosis statement BM25 score could be computed and score comparison could be done with the diagnosis statement in the directory. Top N most relevant diagnosis statement could be displayed ranked by BM25 scores. The most relevant diagnosis statement get the highest score.

III. Development Methodology

A. *Diagnosis Directory*

Diagnosis directory is the list of diagnosis statements. A diagnosis statement is the noun phrases that the physician recorded in the section “Diagnosis” or “Impression” of the patient medical record. The diagnosis statements are the physician opinions about the disease names of the patient. One diagnosis statement may contain 1-6 words from medical terms or common terms. For examples, the diagnosis statement may be *pneumonia, acute tonsillitis, chronic kidney disease, chronic obstructive pulmonary disease, open intertrochanteric fracture of left femur* etc.

To facilitate physician diagnosis mapping with concepts in ICD-10, the diagnosis directory should contain diagnosis statements that could be mapped to the original physician diagnosis on one side as well as mapped to the concepts in ICD-10 on another side easily. However, this is the difficult task because the number of all possible diagnosis statements in English is around 300,000 statements, while the number of concepts in ICD-10 is 14,200 concepts. The diagnosis directory must contain “common” diagnosis statements that can be used to map various diagnosis statement with similar concepts into one common diagnosis statement. The diagnosis directory must also contain mapping between common diagnosis statement and common ICD-10 concepts as shown in Table 1.

TABLE 1 : Original diagnosis, Common diagnosis and Common ICD-10 concepts

Original Diagnosis	*Common Diagnosis*	*ICD-10 Concepts*
Adenocarcinoma of colon	Carcinoma of colon	Malignant neoplasm of colon
Carcinoma of colon	Carcinoma of colon	Malignant neoplasm of colon
Duke B Cancer of colon	Cancer of colon	Malignant neoplasm of colon

The diagnosis directory in this work was built upon two sources of diagnosis list. The first source is the previous diagnosis list from the author research work in the past [3]. The second source is the diagnosis list built from ICD-10 alphabetical index [4]. Final check of the diagnosis directory was performed by comparison with the Unified Medical Language System [5].

B. *Mapping Algorithm*

The physicians in Thailand usually write patient diagnosis in English, so mapping algorithm was built by (1) composition and arrangement of English natural language processing algorithm and (2) adding new steps to match ICD-10 context

The algorithm steps are; 1) abbreviation conversion 2) stop words removed or replaced 3) spell checking 4) word counting and selection of appropriate set of diagnosis statement 5) similarity measurement 6) output of results.

1) Abbreviation conversion. In this step the system check if any abbreviation was found in the original input by

SEEK DIGITAL LIBRARY

comparing each word with the abbreviation dictionary. If any abbreviation was found the system change all abbreviation to full word (s).

2) Stop words removed or replaced. In this step the system remove any stop words from the original statement. The stop word list was built using WordNet stop words [7] with combination of ICD-10 stop words (for examples; left of right should be removed while malignant cell types words should be change to cancer or carcinoma).

3) Spell checking. In this step the system correct misspelling word by Peter Norvig algorithm [8] by comparing with WordNet corpus plus ICD-10 word list.

4) Word counting and selection of appropriate set of diagnosis statement. In this step the system count the number of words in the original statement and select the appropriate set of diagnosis statement which was grouped by number of words. For examples, if the original statement was *chronic kidney disease,* the set of diagnosis statement with three words were selected.

5) Similarity measurements. In this step the system used Cosine similarity to measure similarity between the original statement and the diagnosis statement and return the Cosine similarity value.

6) Output of results. In this step the system return the output based on similarity value. If the similarity value is higher than 0.7 the system return the diagnosis statement in the diagnosis directory. Else, the system return "No match was found".

IV. Results and Discussions

The diagnosis directory contains 43,331 diagnosis statements. The directory could be divided into 10 subsets with the number of diagnosis statements shown in Table 2.

TABLE 2 : Number of diagnosis statement in each directory set

Number of words in set	*Number of diagnosis statements*
1	3,682
2	13,654
3	10,697
4	6,006
5	3,357
6	2,291
7	1,389
8	844
9	565
more than 10	846

The algorithm was tested using OPD diagnosis data from one community hospital of the Ministry of Public Health. The data contain 400 diagnosis statements. The algorithm could map 252 (63.0%) original diagnosis statements to diagnosis directory with 152 (38.0%) exact match (Cosine similarity = 1.0).

Development of diagnosis directory and algorithm to facilitate physician diagnosis mapping to diagnosis directory is the first step of building the semi-automate ICD-10 coding software. If the algorithm could map most of the physician original diagnosis statements to the diagnosis statement in diagnosis directory, then the next parts i.e. mapping to ICD-10 could be added easily. In the first testing the algorithm could map 63.0% of the original diagnosis statement. However, this finding mean that in the future we can use the algorithm to reduce 63% of the manual work done by clinical coders.

In order to increase performance of the algorithm, we can use the list of un-match original diagnosis statement to create additional diagnosis statements and append to the previous directory manually or we may add another step in the algorithm to learn from the users who check and add more diagnosis statement into the directory. So these are some issues for improvement in the future research.

References

[1] World Health Organiation, International Statistical Classification of Diseases and Related Health Problems, 10th Revision, 5th ed., vol 1. Geneva: World Health Organization. 2016.

[2] Z. Bankowssi and A. H. T. Robb-Smith, "An international nomenclature of disease," J. R. Soc. Med., vol. 71(5) p. 382, May 1978.

[3] SNOMED International, "Systematized Nomenclature of Medicine – Clinial Term : SNOMED CT " 2019. [Online]. Available: https://www.snomed.org. [Accessed: May. 18, 2019].

[4] National Libray of Medicine, "Unified Medical Language System : UMLS" 2019. [Online]. Available: https://www. uts.nlm.nih.gov/home.html. [Accessed: May. 18, 2019].

[5] W. Paoin, M. Yuenyongsuwan, Y Yokobori et al, "Development of the ICD-10 simplified version and field tests," Health. Inf. Manag. J., vol 27(2) pp. 77-84, May 2018.

[6] W. Paoin, "New diagnosis list for patient diagnosis in hospital computer program," (in Thai) Journal of the Association of Researchers, vol 14(2) pp. 71-81, May 2009.

[7] H. Dalianis, "Basic building blocks for clinical text processing," in Clinical Text Mining,. Cham: Springer Nature, 2018, pp. 55–82.

[8] D. Sarkar, "Text similarity and clustering," Text Analytics with Python,. New York: Apress Media, 2016, pp. 265–317.

SEEK DIGITAL LIBRARY