

Information Extraction from Handwritten Medical Records and Assigning ICD-10 Codes

Pouya Foudeh, Naomie Salim

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, Johor Bahru
fpouya2@live.utm.my, naomie@utm.my

Abstract. Medical diagnoses for clinical patients are recorded by a physician at the time of admission and discharge from hospital as well as death. Nowadays this information is documented in patient file not only in form of ordinary reports but also in ICD-10 codes to reach more accurate interpretation and use them in statistical surveys. ICD code assignment is a tedious, laborious and highly expensive procedure. Many health care organizations pursue the goal of developing automated procedures. This paper is going to introduce a new method that have machine learned from files with ordinary diagnosis and ICD-10 codes and then will be able to find ICD-10 codes for patients who are not assigned and double checking codes of the others as well. Results of this system will be in form of probabilistic data.

Keywords: diagnosis coding, medical records, ICD-10 codes, information extraction, probabilistic data

1 Introduction

1.1 Classification and Coding of Diseases

Experts and specialists are qualified to deliberate situations and affirm their own discretion and these reports sometimes have a significant influence on individuals' life. Reports, which are written in natural language, can be interpreted arbitrarily so it may cause inaccuracy of reporting. Coding systems are devised to make reports accrue and avoid arbitrary interpretations. They are used in various tasks such as billing systems, accounting transactions and medical records.

First specific list of disease was published in 1893 to determine death causes. This list was revised several times. Ninth version of "International Classification of Diseases" (ICD-9) was published in a conference that had been held by World Health Organization (WHO) in 1975 and finally, ICD-10 was developed between 1983 and 1992, including 14,400 types of disease.

ICD-10 codes are containing two or three parts. An alphabetic character, a two digit number and maybe another one digit number which is separated by a dash. The character is standing for block, or category, of disease, first number and the character for type and optional second number for subtype. For example in "J11-0", J is

standing for “diseases of the respiratory system”, J11 for “influenza, virus not identified” and J11-0 for “influenza with pneumonia, virus not identified”.

In late years, using of ICD-10 codes has become compulsory for clinical patients' records and diagnosis must be noted with ADS-10 at the time of admission, discharge and death. These codes usually are assigned by a medical documentation specialist according to physician diagnosis. Medical records are used in several tasks such as investigation of insurance claims, legal medicine and forensic evidences as well as Statistical surveys on the incidence of diseases of countries or the world. [1][2]

1.2 Information Extraction from Text

Human can store information and knowledge, which are acquired during self experiences or creation, on their mind. They usually write about what is in their mind to avoid oblivion and share it with others, mankind, if they believe it is significant. On the other hand there are some kinds of information which are arranged to be readable by computers. Machines can easily catch this information, which are named databases, and are able to answer our queries. The task of information extraction from text enables computers to acquire information from plain text as well as they acquire data from databases. Indeed human understand information from their environment and inscribe it as ordinary text. In information extraction task, computers try to conceptualize the facts using the text with some kind of reverse engineering.

Information extraction from text is a very complicated task. On one hand various styles of writing and unintelligible grammatical rules make the text analysis very difficult and on the other hand understanding the text, by machine or human, required some background knowledge about meaning and roles of all words, are used in the text. There are a lot of works in this area since second half of twentieth century and there are some significant results, for example search engines performance is advanced extremely, however there is a long way to go to extract information from handwritten text by computers as well as they extract data from databases. [3]

One of the most important steps in information extraction from text is developing ontologies. Ontology is a representation vocabulary, specialized to some domain or subject matter. In each ontology, significant terms are detected and classified, synonyms terms are determined and probably some rules and restrictions are defined. Nowadays several ontologies are developed on various domains such as Electronics, Chemistry, religion science, and medicine and et cetera. Ontologies describe the world and share knowledge. Information extraction systems utilize ontologies to get background information about the text. [4]

1.3 Problem Statement

ICD code assignment is a tedious, laborious and highly expensive and forfeitable procedure. Many health care organizations pursue the goal of developing automated procedures. Nowadays, in most countries interpolation of ICD-10 in patients' files is compulsory. Usually, physician record his diagnosis after admission, death and before discharging the patient and then, a medical documents specialist codes the diagnosis

and add it to patient file. In some countries hospitals are not forced to use computer to store the records and diagnosis as well as codes can be filled in paper documents nevertheless since several years ago, normally before vogue of ICD-10 codes, many hospitals have been using computers to store diagnosis and prescribed medicine records.

Medical records, which have been recorded before ICD-10, are not eligible to be used in statistical surveys. Coding all bygone records is not frugal therefore just some cases are coded after official inquiries. On the other hand coding is forfeitable task because it is done by a feasible human and usually these errors could not be detected because nobody double checks the assigned codes.

This paper is going to introduce a machine learning method to explore available medical records including physician diagnosis, prescribed medicines and assigned ICD-10 codes and then be able to assign ICD-10 code to bygone records, which do not have, as well as double check ICD-10 codes of other records and warning the user of probable mistakes.

In current method, all outputs are in form of probabilistic data. [5] There have been several efforts to do assign ICD codes using medical records since 1996. Nonetheless, this is the first encounter with the problem from probabilistic data point of view. In earlier systems, taken decision to assigned codes was deterministic therefore failure in choosing the correct answer would cause of total failure for the case, nevertheless current system suggests multiple codes for a case with different degree of probability and system fault will cause a comparative error.

2 Related Works

Ontologies in biomedical domain are foundation of expert systems for health care as well as information extraction from medical records. There have been several efforts to develop ontologies in biomedical domain however, most of available ontologies are specialized and do not cover whole medical domain. [6] Bioontology is an online service that allows users to access several ontologies in biomedical domain. This portal is running by NCBO, one of seven research centers of National Institutes of Health of United States, working on biomedical computing. [7]

For diagnosis of diseases by computers, there have been a lot of efforts over recent decades. MYCIN was the first expert system for diagnosis blood infections with asking some yes/no questions with accuracy of 69 percent. MYCYN was developed at Stanford University in 1970s. [8]

HELP presented in 1996 as a system for extracting information from patients' medical records and coding those, according ICD9- coding system using natural language processing techniques and Bayesian networks. [9] At 2008 MIDAS was developed to face same challenge using a linguistic processor. [10] In 2010 a new method was presented by Waraporn using machine learning-based multi-label classification. [11]

3 Methodology

3.1 Terms Valuation

A lot of medical files, including physician's diagnosis, prescribed medicines and their assigned ICD-10 codes are available to the system. Medical documents are divided according to the specialty such as psychiatry, pediatrics, oncology, endocrinology, orthopedics etc. and searching codes will be limited to the category.

At the first step, important terms should be recognized and valued. In information retrieval point of view, in some cases terms include just one word and in other cases include two or more. Multiple words can easily be found because they frequently appear together.

Processing all terms be time consuming, therefore, eliminating the valueless term is the first priority. First, standard stop words such as "is", "the" and "about" are removed and then, remaining recognized terms, in disease description or prescribed medicine, are valued. Criteria of valuation of terms are their specialization. A general term or medicine, e.g. Exhaustion or Ibuprofen, is not valuable because it appears in different types of disease and cannot lead the system to limited ICD-10 codes. We define value of w term as:

$$V_w = \log \frac{M}{C_t}$$

M is number of all codes in category, have been used in the hospital.

C_t is the number of ICD-codes, in category, that have at least one record including w term.

In this phase system regards to first and second part of codes therefore all subtype disease are seen as one code.

All detected terms are searched in medical Bioontology and synonyms are assimilated, prescribed medicines are also compared with pharmacopeia and similar drugs, especially different brands names of same generic drug, are replaced in this stage. A commercial database from pharmacy software has been permitted to be used in this research to find similar drugs and brand names. As physician's diagnoses are in form of short writing, system will find many term out of ontologies and pharmacopeia. At this point machine leaning must be supervised. A medical documents specialist assign these terms to ontologies and pharmacopeia or just remove them if believe they are worthless. Terms with higher value (V) would have a higher priority to be supervised and terms with very low value of V are removed without supervision. V value is computed again for assimilated terms.

3.2 Relationship between Terms and ICD-10 codes

Valueless terms will not associate in this phase. To remove the valueless terms we need a threshold value for V . If it is too low, near zero, system will be more accurate but there will be a lot of term to be processed. We have found $V = 0.5$ as an appreciate threshold value but is depends on types and number of medical records.

Now system determines dependency of ICD-10 codes and terms. W_i is a valuable term and C_j is an ICD-10 code and D_{wicj} will be dependency of W_i and C_j .

$$D_{wicj} = \frac{\text{Number of records assigned to } C_j \text{ code containing } W_i \text{ term}}{\text{Number of records containing } W_i \text{ term}}$$

Some term may be correlated each others. For example injecting medicines are usually prescribed with a syringe. Terms, which are telling to choose an ICD-10 code, should be independent; correlation between terms will increase chance of selection of an ICD-10 code mendaciously. In our system two terms are independent if and only if:

$$M_{wiwj} = \pm 0.25 \left(\frac{M_{wi} * M_{wj}}{M} \right)$$

M_{wiwj} is number of records containing W_i and W_j , M_{wi} is number of records containing W_i and M is number of all records. If two terms are not independent in one ICD-10 code, system keeps the term with greater value of D and remove another.

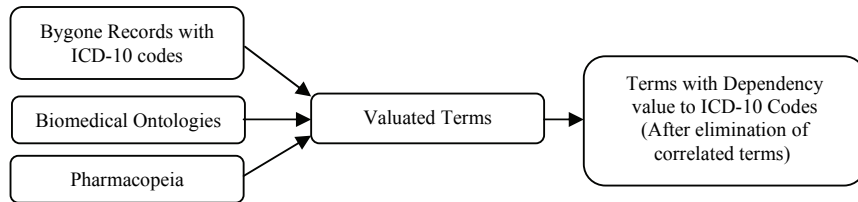


Fig. 1. Finding valuable terms and determine their dependency to ICD-10 codes

3.2 Identifying probable ICD-10 codes for an specific record

To identify some ICD-10 codes for a specific record, system extract all terms of record and compare them with terms synonym list, has been built using biomedical ontologies, pharmacopeia and supervised mapped terms. The record will be examined with an ICD-10 code, if they have at least one term in common. If W_1, W_2, \dots, W_n are common terms and C_m , ICD-10 code would be:

$$P_2 = D_{W1cm} + D_{W2cm} - (D_{W1cm} * D_{W2cm})$$

$$P_n = P_{n-1} + D_{Wncm} - (P_{n-1} * D_{Wncm})$$

P_{cm} is computed using above recursive formula. After determination of P for all ICD-10 codes, one or several most probable ICD-10 codes and their probabilities are presented to user.

4 Conclusion and Future Works

In this paper, we presented a method which investigates patients' medical records including diagnosis, which are written by a physician, at the time of admission and discharge from the hospital or death as well as assigned ICD-10 codes. Thenceforth system would be able to assign ICD-10 code to bygone files, which do not have. Furthermore, system is able to double check assigned codes and detects probable mistakes of the coder person. All responses are in form of probabilistic data.

Current method has been examined with limited number of artificial data. It should be evaluated with enough number of actual medical files. In next steps, more complex forms of medical records can be regarded such as memoir of the psychiatric patients. Furthermore, a probable data processing system can be developed for answering queries from probabilistic results of current system and interpret them to be exploitable in statistical surveys.

5 Acknowledgement

This project is sponsored by the Ministry of Science, Technology and Innovation Malaysia under E-Science grant 01-01-06-SF0539.

References

1. "History of the development of the ICD," <http://www.who.int/classifications/icd/en/HistoryOfICD.pdf>.
2. C. W Ab, S. B Bcd, and G. A. De Moor, "Ontology-Based Integration of Medical Coding Systems and Electronic Patient Records," *Relation* 10, no. 1.73 (2008): 4077.
3. Philipp Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications* (Springer US, 2009).
4. B. Chandrasekaran, John R. Josephson, and V. Richard Benjamins, "What Are Ontologies, and Why Do We Need Them?" *IEEE Intelligent Systems* 14, no. 1 (1999): 20-26.
5. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," *The VLDB Journal* 16, no. 4 (2007): 523-544.
6. F. Pincioli and D. M Pisanelli, "The unexpected high practical value of medical ontologies," *Computers in Biology and Medicine* 36, no. 7-8 (2006): 669-673.

7. "BIO Portal" <http://www.bioontology.org>
8. E. H Shortliffe et al., "Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system," *Computers and Biomedical Research* 8, no. 4 (1975): 303–320.
9. M. L. Gundersen et al., "Development and evaluation of a computerized admission diagnoses encoding system.," *Computers and biomedical research, an international journal* 29, no. 5 (1996): 351.
10. A. Sotelsek-margalef, J. Villena-román, and I. Madrid, "MIDAS: An Information-Extraction Approach to Medical Text Classification MIDAS: Un enfoque de extracción de información," *Relation* 10, no. 1.43 (2009): 7517.
11. P. Waraporn, P. Meesad, and G. Clayton, "Ontology-supported processing of clinical text using medical knowledge integration for multi-label classification of diagnosis coding," *Arxiv preprint arXiv:1004.1230* (2010).



Naomie Salim has contributed her service to Universiti Teknologi Malaysia for more than 20 years. Her contribution in the Databases and Information Retrieval field is illustrated by the numerous researches. She has also

continuously serves as an administrator at the university, postgraduate coordinator, head of department and now, Professor Dr. Naomie Salim is the Deputy Dean of Research and Postgraduate Studies of Faculty of Computer Science and Information Systems.



Pouya Foudeh was born in Isfahan, Iran in 1976. He received his B.Sc. and M.Sc. in Computer Engineering from Islamic Azad University, Iran in 2001 and 2005. He has served as a system analyst in Azadi Psychiatric Hospital and the General Manager in Abarkakia LTD. He has also served as an academic staff and visiting lecturer in Islamic Azad University from 2006 to 2009. Pouya Foudeh is currently a Ph.D student in Universiti Teknologi Malaysia working in Probabilistic Data and Information Retrieval area.