التشخيص الطبي التلقائي والتنبؤ بالتصنيف الدولي للأمراض باستخدام الأنطولوجيا الدلالية ومعالجة اللغة الطبيعية

د. أيمن عيسى*

(الإيداع: 18 تشرين الأول 2020، القبول: 30 تشرين الثاني 2020)

الملخص:

إن التشخيص الطبي التلقائي مجال البحث النشط في صناعات الرعاية الصحية. يزود المرضى والمهنيين الصحيين برمز التصنيف الدولي للأمراض (ICD) الذي يتم فيه تصنيف الأمراض إلى مجموعة متنوعة من العلامات والأعراض والنتائج غير الطبيعية والشكاوى والظروف الاجتماعية والأسباب الخارجية للإصابة أو المرض ، وأكثر الأمور تحديًا هي كيفية العثور على رمز مرض التصنيف الدولي للأمراض المناسب للمرضى من قراءة وتحليل شكاواهم ، في هذه الورقة قمنا بتطوير نظام التشخيص الطبي التلقائي للعثور على أنسب كود التصنيف الدولي للأمراض لمرض المريض عن طريق استثمار تكنولوجيا الويب الدلالي ومعالجة اللغة الطبيعية لفهم شكوى المريض وإيجاد رمز التصنيف الدولي للأمراض ، في هذا البحث استخدمنا أساسًا أنطولوجيا التصنيف الدولي للأمراض ICD ، وأنطولوجيا الأعراض التي تحتوي على وصف الأمراض وأعراض هذه الأمراض.

الكلمات المفتاحية: الويب الدلالي، نظام التشخيص؛ التكنولوجيا والطب؛ مرض؛ الأعراض، علم الوجود التصنيف الدولي للأمراض، علم الوجود الأعراض.

*مدرس في كلية الهندسة المعلوماتية في جامعة الرشيد الخاصة

Automatic Medical diagnosis and International Classification of Disease code ICD prediction Using Semantic Ontologies and Natural Language Processing

Dr.Ayman Issa*

(Received: 18 October 2020, Accepted: 30 November 2020)

Abstract:

Automatic Medical diagnosis has been the area of active research in healthcare industries. It provides patients and health Professional with International Classification of Disease code (ICD) in which diseases are classified to a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease, and the most challenge matter is how to find the right ICD disease code to the patients from reading and analyzing their complaints ,In this paper we have developed Automatic Medical diagnosis System for finding the most suitable ICD code for the patient disease by investment the technology of semantic web and natural language processing in manipulating and understating the patient complaint and finding the ICD code for it , in this research we used mainly ICD ontology , Symptom ontology which contains the diseases and symptoms description for all diseases.

Keywords: semantic web, Diagnosis System; technology and medicine; Disease; Symptoms, ICD Ontology, Symptom Ontology.

***Instructor in the College of Informatics Engineering at Al-Rasheed Private University**

1. Introduction

Automatic medical diagnosis, and related medical ontologies have recently of vital value in medical sector, Ontology[1][3] encompasses a representation, formal naming, and definition of the categories, properties, and relations between the concepts, data, and entities that substantiate one, or all domains, as shown in figure (1).

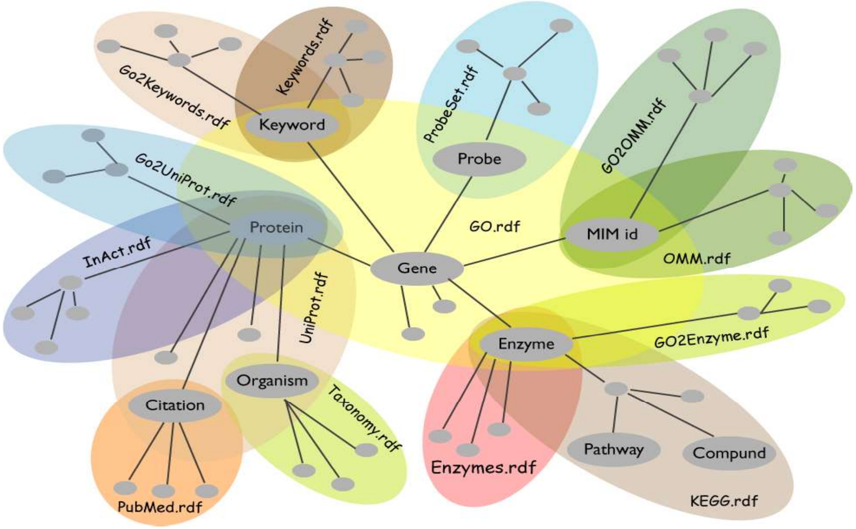


Figure (1) :ontologies[2]

Semantic web technology has been utilized to determine the best code of the disease using International Classification of Disease standard code ICD [4]. That these codes have the description of the user disease, and the major thing that we have based on it in the investment of this technology is the ontologies , especially the medical ontologies , we have used the symptoms ontology[5],diseases ontology[6],that The symptom ontology was designed around the guiding concept of a symptom being: "A perceived change in function, sensation or appearance reported by a patient indicative of a disease". Understanding the close relationship of Signs and Symptoms, where Signs are the objective observation of an illness, the Symptom Ontology will work to broaden its scope to capture and document in a more robust manor these two sets of terms. Understanding that at times, the same term may be both a Sign and a Symptom.

The International Classification of Diseases ICD is the global health information standard for mortality and morbidity statistics. ICD is increasingly used in clinical care and research to define diseases and study disease patterns, as well as manage health care, monitor outcomes and allocate resources. About 70% of the world’s health expenditures (USD $ 3.5 billion) are allocated using ICD for reimbursement and resource allocation ICD has been translated into

43 languages. The 11th revision process is underway and the final ICD-11 will be released in 2018[7].

2. Aim

The aim of the study was to find the most correct ICD10 code and full description of the disease from analyzing the patient input complaint.

3. Methodology

Our proposed algorithm contains two main Stages (Training Stage, Testing Stage), as shown in the figure (2); the input of each Stage is the patient compliant, and the output of each Stage is the ICD code and the description of the patient disease. The proposed method is described figure 2:

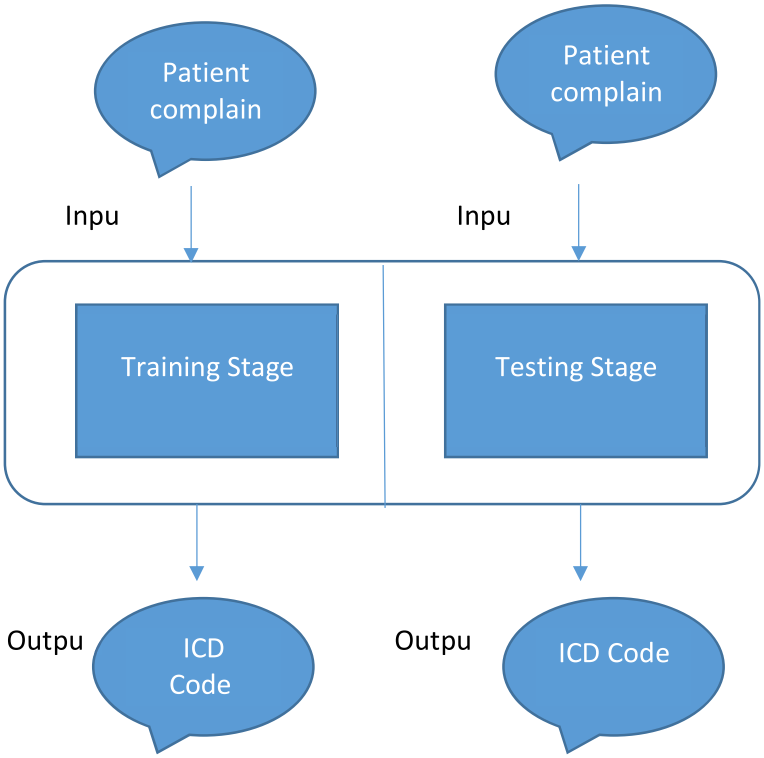


Figure (2): Proposed System

3-1. Methodology Definitions:

There are the following definitions, which are used, in the developed algorithm.

- One gram set: this set of elements and each element has one word, these sets are generated from the complaints descriptions by using natural language processing tools, that every element in the compliant set is a stemmed (stemming means return the word to it root, plural to singular), and non-stop word. (The elements of these sets only nouns, which are, extracted from complaints descriptions using natural language processing tools.

- Tow gram set :this set of elements and each element has two words, and these words are related to each other by a relation , the relations that we have used to generate these elements (every element in set is a pair of words which are related to each other by a relation) are nn realation, dobj relation, conj_and relation, amod relation :

Table Namber(1) : nouns relations

Relation Symbol	Relation Meaning	Example

NN	noun compound modifier	"Oil price futures"
		NN(futures, oil)
		NN(futures, price)

dobj	direct object	"She gave me a raise" dobj(gave, raise)
conj_and	And between nouns	computer and building products conj_and(computer, building)
amod	adjectival modifier	"Sam eats red meat" amod(meat, red)

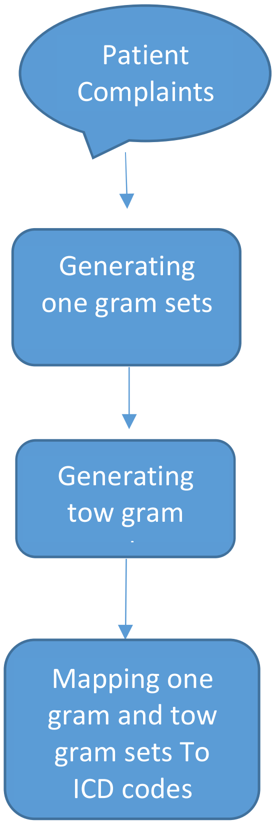
- Sets intersection : these intersection is calculated between the training sets and the testing sets , and this intersection is calculated as the following formula :

Intersection Score= (count of identical elements of the two sets) / (size of the bigger set)

This formula is used on wide rang in calculation intersection between sets in information theory.

3-2.Methodology Workflow

In This Section, the Workflow of the proposed algorithm is described in details, for the two stages of the algorithm. The following figure shows the steps of the training Stage:



Training Stage: it contains the two steps :

○ Generating one Gram sets as the following:

Input data records are parsed and by using Stanford natural language processing tools, **nouns** are extracted and every complaint is converted to one gram set of stemmed and non-stop words nouns. (One gram which means that the main element of the set is one word only).

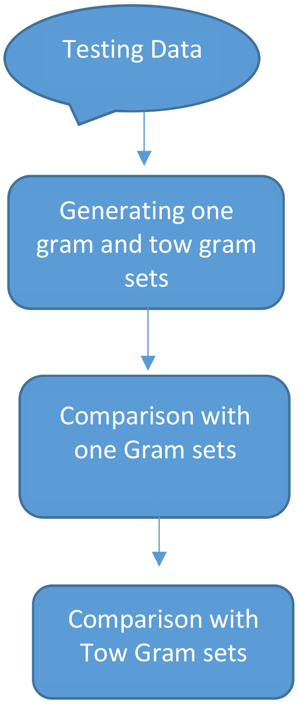
○ Generating Tow Gram sets as the following :

Input data records are parsed and by using Stanford natural language processing tools, **nouns with relations** are extracted and every complaint is converted to tow gram set of stemmed and non-stop words nouns. (Tow gram which means that the main element of the set is tow words connected by relation, we have mentioned the relations types above).

- Mapping one gram and Tow Gram sets to ICD codes as the following :

Generated sets are mapped to the principle and secondary diagnosis ICD codes, that every complaint set is mapped to one principle diagnosis ICD code and a list of secondary diagnosis ICD codes.

Testing Stage: The following figure shows the steps of the **Testing** Stage:


Testing Data
Generating one gram and tow gram sets
Comparison with one Gram sets
Comparison with Tow Gram sets

- Generating one gram and tow gram sets: In this step, the testing records are manipulated and converted to One gram and tow gram sets.
- Comparison with one-Gram sets: In this step of testing Stage, every complaint one-gram set is interested with the training one-gram sets, and we take the sets with its ICD diagnosis codes, which are most relevant to the input test set.

Because we are using set intersection, we will get a list of intersected sets with the input set, so we choose the most relevant ones according to intersection score with the input set.

- Comparison with tow Gram sets : In this step of testing phase every complaint tow gram set is interested with the training tow gram sets (August, July complaints one Gram sets) and we take the sets with its ICD diagnosis codes which are most relevant to the input test set.

Because we are using set intersection, we will get a list of intersected sets with the input set, so we choose the most relevant ones according to intersection score with the input set.

By the end of the testing phase we will get for every input complaint record, we will get a list of principles and secondary diagnosis codes, which are, represent the most relevant sets of the training data sets.

The developed algorithm basing on fuzzy sets intersections ,which give us for an input complaint a list of principles and secondary diagnosis ICD codes ,which are arranged according to the intersection score(in other word it is the certainty factor that the diagnosis ICD code is the most suitable for the complaint).

3-3.Methodology Implementation:

Here we will explain how we have implemented the developed system Stages in Details:

3-3-1.Training Stage Implementation:

It contains multi steps to make the implementation:

- Natural language processing step: In this step we manipulates the complaints text as we do the following(all these operations are done using Stanford Natural Language Processing tools [8]) :

- We segment each complaint text to its words.

- Remove stop words from complaint words list

- Stemming complaint words list (stemming is finding the root of the word, and convert plural to single [9])

- Get the nouns of the complaint words list; because knowledge is stored into nous not into verbs, (these nouns are the one gram set).

- Get the relations between nouns of the complaint words list; (these nouns are the tow gram set).

- Symptoms finding step : The input of this step is the nous of the complaints texts

  In this step we used the nous of the complaints texts and search for these nous into symptom ontology [10] (https://bioportal.bioontology.org/ontologies/SYMP) and get the symptom terms from this ontology for every noun into the complaint nouns list.

  The final output for this step is to build a set of symptom ontology terms for every complaint nouns list.

  We used the Ontology Lookup Service (OLS) https://www.ebi.ac.uk/ols/index which is a big repository of medical ontologies(200 medical ontology) , that we used this repository API to find the symptom terms of the complaint nouns , this terms contains the label and descriptions of symptoms and the synonyms of the symptom.

- ICD code searching step: The input of this step is the symptom ontology terms.

In this step, we find the ICD10 codes, descriptions, and synonyms for symptom terms, using ICD searcher service [11].

We use the Clinical Table Search Service https://clinicaltables.nlm.nih.gov/ to find the ICD10 code for every symptom term (which generated from complaint nouns), then we used ICD Access Management API https://icd.who.int/icdapi to get the full information about the ICD10 code for every disease.

- Weighting results step: The input of this step is a list of ICD10 diseases and the output is the most five relevant diseases to the complaint text words, and symptom terms sets.

We used the sets intersection between the ICD10 diseases and complaint text words, and the intersection between the ICD10 diseases and symptom terms of the complaint nouns, set a weight to every ICD10, and order them from the most relevant to the least relevant.

At the end of this step all one gram and tow gram sets are mapped to ICD10 codes.

3-3-2.Testing Stage Implementation: the implementation of this Stage includes Natural language processing and sets intersection to find the most suitable ICD code for the patient from his complaint.

In this Stage all Natural language processing operations are done as these in Training Stage (segmentation, stemming, finding nous, finding nous relations (table (1): nouns relations), by the ending of these operations, one gram and tow gram sets are generated.

Then the intersection between one gram and tow gram sets of the testing data are intersected with the one gram and tow gram sets of the training data. The **Intersection score** is done basing on its definition of **in Methodology Definitions** section. In addition, the most relevant sets of the training data are taken into consideration to use its ICD10 codes and descriptions to assign to the testing data one gram and tow gram sets.

For the developed system, we have user java programming language for building the developed system, and for the natural language processing operations we have used Stanford natural language processing tools

4. Data

The input data was about of 2000 records of patient's complaints. In addition, we have divided this data records between training and testing stages, 1000 record for the training stage, and 1000 record for the testing stage.

5. Testing and Results

We have trained the developed system on a set of complaints records about (2000 record), and then we tested the developed system on a small set of training data set and the results was very excellent, that the accuracy was about 90%, and the remaining 10% of accuracy was found because there were a few complaints which the system suggest a wrong ICD codes for it.

6. Conclusions

We have developed a system for ICD codes prediction, and this system was based on patients complaints, and this system need to continuous development by generating and feeding it with new knowledge for diseases diagnosis, that the system has its own database and it updating continuously by adding new complaints with its right ICD codes and descriptions. In addition, in the future we will expand and enrich the developed system by adding Expert rules for generating ICD codes.

7. References

1. https://en.wikipedia.org/wiki/Ontology
2. State of the Semantic Web www.w3.org
3. D. Allemang and J.A. Hendler, Semantic Web for the Working Ontologist - Effective Modeling in RDFS and OWL, Second Edition, Morgan Kaufmann, 2011. ISBN 978-0-12-385965-5.
4. https://www.who.int/classifications/icd/icdonlineversions/en/
5. https://www.ebi.ac.uk/ols/ontologies/symp
6. https://www.ebi.ac.uk/ols/ontologies/doid
7. https://www.who.int/news-room/detail/18-06-2018-who-releases-new-international-classification-of-diseases-(icd-11)
8. https://nlp.stanford.edu/software/
9. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
10. https://bioportal.bioontology.org/ontologies/SYMP
11. https://icdcodelookup.com/icd-10/codes