Expert Systems with Applications 42 (2015) 2949–2958

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Computer aided classification of diagnostic terms in spanish

Alicia Pérez [a], Koldo Gojenola [a], Arantza Casillas [b,*], Maite Oronoz [c], Arantza Díaz de Ilarraza [c]

[a] IXA Taldea, Dep. Languages and Computer Systems, Technical School of Engineering Bilbao, UPV-EHU, Spain
[b] IXA Taldea, Dep. Electricity and Electronics, Faculty of Science and Technology, UPV-EHU, Spain
[c] IXA Taldea, Dep. Languages and Computer Systems, Faculty of Computer Science, UPV-EHU, Spain

ARTICLE INFO

Article history:
Available online 28 November 2014

Keywords:
Classification of Medical Records
Natural language processing
Finite-State Transducers
Applications in medicine

ABSTRACT

The goal of this paper is to classify Medical Records (MRs) by their diagnostic terms (DTs) according to the International Classification of Diseases Clinical Modification (ICD-9-CM). The challenge we face is twofold: (i) to treat the natural and non-standard language in which doctors express their diagnostics and (ii) to perform a large-scale classification problem.

We propose the use of Finite-State Transducers (FSTs) that, for their underlying topology, constrain the allowed input DT string while synchronously produce the output ICD-9-CM class. It is outstanding their versatility to efficiently implement soft-matching operations between terms expressed in natural language to standard terms and, hence, to the final ICD-9-CM code. The FSTs were built up from a corpora and standard resources such as the ICD-9-CM and SNOMED CT amongst others. Our corpora count on a big-data comprising more than 20,000 DTs from MRs from the Basque Hospital System so as to model natural language in this domain. An F1-measure of 91.2 was achieved on a test-set of 2850 randomly selected DTs, and a random 5-fold cross validation on a training set served to double-check the stability of the provided results. Real MRs were of much help to adapt the system to natural language. Misspellings, colloquial and specific language and abbreviations made the classification process difficult. The FSTs were proven efficient in this large-scale classification task. Moreover, the composition operation for FSTs made it easy the addition of new features to the system.

1. Introduction

The aim of this paper is to explore computer aided approaches to classify Medical Records (MRs) written in Spanish according to their diagnostic term (DT) following the World Health Organization's 9th Revision of the International Classification of Diseases Clinical Modification (denoted as ICD-9-CM from now onwards). So far, it is common practice in the hospitals attached to the Spanish Ministry of Health, Social Services and Equality to classify the MRs manually, as it is the case of the Galdakao-Usansolo Hospital (GUH). Nevertheless, GUH is concerned with the automation of the Clinical Documentation Service and is pioneering this area in the community.

There is an increasing interest in the evolution of the automatic or semi-automatic classification of MRs, amongst others, due to economic factors. According to Farkas and Szarvas (2008), the approximate cost of ICD-9-CM coding clinical records and correcting related errors is estimated to be about $25 billion per year in the US. As another example, at the New York Presbiterian Hospital the amount of data in electronic health records has been increasing at an exponential rate (Holmes et al., 2011). According to the Clinical Documentation Service from the GUH, automatically classifying 1% of the MRs would have an outstanding impact in terms of person-months work. Nevertheless, it is the high precision above the speed that is crucial in this context. The ultimate goal is to alleviate the cumbersome workload by the Clinical Documentation Service but never at the expense of losing accuracy. Besides, computer aided classification of MRs represents a worthwhile support for information retrieval, since the documentation services count on electronic documents stored over years that include valuable lexical resources.

In this work we promote the use of Finite-State Transducers (FSTs) as an asset to overcome the aforementioned challenges. FSTs serve as acceptors of an input language and as generators of an output language, being the two operations synchronous. FSTs have proven outstanding models to implement searching operations

* Corresponding author at: Dep. Electricity and Electronics, Faculty of Science and Technology, UPV-EHU, Barrio Sarriena s/n, 48940 Leioa (Bizkaia), Spain. Tel.: +34 946015321.

E-mail addresses: alicia.perez@ehu.es (A. Pérez), koldo.gojenola@ehu.es (K. Gojenola), arantza.casillas@ehu.es (A. Casillas), maite.oronoz@ehu.es (M. Oronoz), a.diazdeilarraza@ehu.es (A. Díaz de Ilarraza).

http://dx.doi.org/10.1016/j.eswa.2014.11.035
0957-4174/

with low computational cost, also allowing versatile composition operations with other FSTs (Mohri, 1997).

2. Background and significance

2.1. Challenges

Computer aided classification of MRs by their DTs faces two main challenges:

1. The processing of natural language: doctors express DTs in natural language with their own style and different degree of precision, seldom as in the ICD-9-CM list (0.245% of the times in our experiments). The same DT in different MRs might appear with variations due to modifiers, abbreviations, acronyms, misspellings or different style. Amongst others, this is due to the fact that doctors devote their time to the patient exploratory and evaluation process rather than to the writing of the records. Hence, in spite of the fact that ICD-9-CM is taken as a reference, doctors rarely write the DTs in that way. Table 1(a) shows some examples from the ICD-9-CM list. Table 1(b) gives an example of DTs written by doctors in MRs and associated to the same codes. Note alternative writings for a DT with the same ICD-9-CM code. Often, for a given code, the alternatives differ much from the ICD-9-CM list (compare Table 1(a) and (b)).
2. Efficient methods for large scale classification: there is a need of managing big data; the system should be able to provide a class amongst a set of more than 14×10^3 different codes. From the machine learning standpoint this can be considered a tough classification problem. Note that, for a classification system to be inferred, wide variability of the training samples per class is required. What is more, for this task the system must ensure high precision.

2.2. Related work and motivation

Computer aided classification of MRs by their DTs can be seen as a syntactic pattern recognition task: it has to recognize unknown instances of expressions and assign them one element of a set of possible labels. This section delves into different ways in which this task can be approached.

The 2007 Computational Medicine Challenge (Pestian et al., 2007), the first shared task related to the main subject of this paper, was designed: (i) to facilitate advances in mining clinical free text and (ii) to create a publicly available gold standard that could serve as the seed for a larger, open source clinical corpus. This Challenge involved the classification of clinical free texts. One of the challenges involved the automatic assignment of ICD-9-CM codes in a limited domain devoted to radiology reports. Farkas and Szarvas (2008) addressed this shared task using machine learning approaches. Their results showed that handcrafted systems could be reproduced by replacing several laborious steps in their construction with machine learning models, reporting an F1-measure of 88.93. By contrast to Farkas and Szarvas (2008) we focus on the entire scope of the ICD-9-CM catalogue. That is, while they are dealing with 45 classes, we are dealing with more than 14×10^3 classes. This is why we promote alternative schemes efficient in terms of space and time, above all, a method that leverage its skills to cope with scarce samples.

Argaw, Hulth, and Megyesi (2007) showed that standard text categorization techniques using stemmed unigrams as the basis for learning can be applied directly to categorize medical reports, yielding a precision of 85.47% and a recall of 69.76%. These results were also on 45 classes. Once again, FSTs are considered to be variable-length n-gram models, thus, we do not restrict ourselves to unigrams and get better performance.

There is another outstanding work dealing with the classification of DTs restricted to the cardiology domain referred to as a large-scale classification process in Lita, Yu, Niculescu, and Bi (2008). This paper focuses on highly-frequent diagnostic codes so as to avoid data sparsity and achieve high performance with statistically motivated methods (Support Vector Machines and Bayesian Regression models). As far as our work is concerned, we consider both frequent and infrequent terms from all the medical subdomains. In 2013 Pereira, Rijoa, Silvaa, and Agostinho (2013) proposed an automatic process of classifying a subset of ICD-9-CM codes restricted to epileptic diagnoses. The system was based on processed Electronic Medical Records and made use of the K-Nearest Neighbor approach.

The studies mentioned above are limited to few ICD-9-CM codes. Next, focusing on works that deal with a significant portion of the ICD-9-CM codes, Medori and Fairon (2010) combine machine learning techniques with information extraction in order to classify medical notes written in French. In the evaluation phase they reported only the recall of the system, although they considered only codes that were manually assigned at least 6 times in the corpus. By contrast, we consider all the codes, including those that were not manually assigned. In a recent paper Perotte et al. (2014) used Support Vector Machines on a corpus of Electronic Medical Records achieving a F-measure of 39.5% in the task of predicting

Table 1
While for each ICD-9-CM code there is a single standard DT in the ICD-9-CM catalogue (see Table 1(a)), doctors make use of alternative DTs to refer to the same disease, that is, to the same ICD-9-CM code (see Table 1(b)).

(*a*) *Examples from the ICD-9-CM list taken as a standard*	
Standard DTs from ICD-9-CM list	ICD-9-CM
Neoplasia maligna de la próstata (Malignant neoplasm of prostate)	185
Neoplasia maligna de tráquea, bronquios y pulmón (Malignant neoplasm of trachea bronchus and lung)	162
Bronquios y pulmón, parte no especificada (Malignant neoplasm of bronchus and lung, unspecified)	162.9
Neoplasia maligna del páncreas (Malignant neoplasm of pancreas)	157
Páncreas, parte no especificada (Malignant neoplasm of pancreas, part unspecified)	157.9
(*b*) *Examples of DTs seen in real MRs and associated ICD-9-CM codes*	
DTs expressed by doctors in MRs	ICD-9-CM
Adenocarcinoma de prostata (adenocarcinoma of the prostate)	185
Adenocarcinomas próstata. (prostate adenocarcinoma)	185
Ca. prostata (prostate Ca.)	185
CÁNCER DE PRÓSTATA (PROSTATE CANCER)	185
adenocarcinoma de pulmon estadio IV (lung adenocarcinoma stage IV)	162.9
CA pulmón estadio 4 (Lung CA stage 4)	162.9
ADENOCARCINOMA PÁNCREAS (PANCREATIC ADENOCARCINOMA)	157.9