Automatic Classification of Medical Reports, the CIREA project

Elisabeth METAIS(1), Didier NAKACHE(1) (2) (3)
and Jean-François TIMSIT (4)

(1) CEDRIC / CNAM: 292 rue Saint Martin - 75003 Paris, France
(2) CRAMIF: 17 / 19 rue de Flandre - 75019 Paris, France
(3) Oxymel: Le Campus - Bat A - 6, rue Jean-Pierre Timbaud
78180 Montigny le Bretonneux, France
(4) Outcomerea: 46, rue de la côte des chênes – 93110 Rosny sous bois, France

Abstract: Choosing a patient's reasons for staying in hospital amongst the 52,000 pathology codes listed in the ICD-10 (International Classification of Diseases) requires that the practitioner spends a large amount of time keyboarding and searching, which may discourage him. However these codes are mandatory in many countries when the patient leaves the hospital, for biostatistical and administrative studies. The aim of the CIREA project is to propose an automatic ICD coding approach by mining textual medical reports. For that purpose we have proposed new algorithms such the EDA desuffixer, the CLO3 classification algorithm and the K-measure indicator.

Key-Words: NLP, Textmining, classification, ICD, desuffixation, F-measure

1 Introduction

In France and in many countries around the world, it is a legal obligation to supply ICD10 codes whenever a patient's treatment reaches its end. Some hospitals employ whole time members of the medical profession to fulfill this task. The purpose is triple: (1) basis for appropriations assigning, considering these codes represent the unit's activity, (2) authorizations for nationwide and international epidemiological studies, thanks to a systematic and standardized coding of pathologies, and (3) factor for calculating indicators of the quality of treatments, such as the number of nosocomial infections2, which justifies that the quality of coding itself has been proposed as a quality indicator [Foqual 2000].

This paper presents the CIREA project. In section 2, we present the context of the project. Section 3 is the state of the art. Finally, after presentation of general principle of CIREA (section 4), we present in details and step by step what we realized (section 5) and our future works (section 6).

2 Context of the Project

2.1 The Rhea project

CIREA is a part of the Rhea project, which is partially financed by French minister of research.

The aim of the Rhea project is to develop a decision support tool based on data warehouse architecture. The warehouse is fed by a selective extraction from the local database of each Intensive Care Unit. A biostatistical exploitation of the warehouse enlightens correlations between data and enables specifying prediction tools on the evolving of the patient's gravity, including iatrogenic and nosocomial events, with the view to reduce their incidence and improve health care. The project includes now a network of 30 hospitals [5].

2.2 The ICD-10

Medical reports are coded according to the International Classification of Diseases, 10th revision (ICD10) built by WHO (World Health Organization). It is a hierarchical classification, each lever giving more precise pathologies than the previous one; as we can see on figure 1. We can find 21 codes at the first level (the more general level), 268 codes at the second one, 2019 codes at the third one, 15000 ones at the Forth one and 54000 codes at the fifth one (the more detailed one).

The works presented in this paper aim to automatically find in this classification the codes corresponding to the diagnostics of the patient by text mining the medical report written by the doctor.

A00-B99	Infectious and parasitic diseases
A00	Intestinal infectious diseases
A00.0	Cholera due to Vibrio
	cholerae 01, biovar cholerae,

[1] ICD : International Classification of Diseases

[2] Nosocomial Infections: Infections acquired within the hospital.

	Classical cholera
A00.1	Cholera due to Vibrio cholerae 01, biovar eltor, Cholera eltor
A00.9	Cholera, unspecified
C00-D48	Neoplasm
D50-D89	Disorders of the blood … involving the immune mechanism
E00-E90	Endocrine, nutritional and metabolic diseases
F00-F99	Mental and behavioural disorders
G00-G99	Diseases of the nervous system
H00-H59	Disease of Ear and Mastoid Process
…	…

Fig. 1. Except from the first, second and third levels of the ICD

3 Related Works

Several algorithms for classification exist, but SVM [12] is one of the most powerful algorithms in textual classification [3]. The main idea is to find a hyperplane which as well as possible separates data and with separation (or strokes: outdistance separating border from nearest example) is as large as possible. Knn is based on similarity search with other reports. In classification of medical report, we assume that two similar reports should lead to similar ICD codes. Neural networks, in particular multi-layers perceptrons, give good results in textual data analysis. LLSF (Linear Least Squares Fit) is a classification method based on terms frequency appearance and linear combinations developed by [14] whose first application was hospital reports classification. [2] has implemented this technique for mammography reports automatic classification. Results obtained show an accuracy of 83.4% + - 5.3% and a recall of 35.4% + - 5.6%. Method principle consists in analyzing appearance frequency of a term according to diagnosis. Bayesian networks are a purely statistical application, based on conditional probabilities.

There are four categories of tools for searching ICD-10 codes:

- Tools based on navigation in a hierarchical taxonomy. They allow to each level to visualize the lower levels but require a good knowledge of the hierarchy;
- Tools based on a lexical search. They allow search for a term in the text of the diagnosis. They are easy to develop and implement, but their results remain limited. Their main interest is the possibility for persons that are not familiar with the ICD-10 to ignore the hierarchical structure of the classification and to encode a term without knowing in which chapter it must be searched;
- Tools that use search procedures for documents;
- Finally, knowledge based tools that take into account the context and guide the user towards specific codes. CIREA belongs to this last category.

Menelas [15,16] is a project of classification that identifies ICD codes but only for coronary pathologies. This project uses ontology of 1800 primary concepts organized in a tree with 300 relations. Blanquet created a automated extraction of ICD10 codes from medical reports [1], which runs as a routine. His approach is simple and efficient (98.2 % of good affectations with an average of 4.7 presented codes) but it uses a specific thesaurus and concerns only hematology. It uses XML labels and asks the user to navigate between different possibilities to find the good resulting codes. It is not transposable to intensive care units where patients have often multiple pathologies. In [13] Wilcox made a comparison of automated ICD-9 classification of medical reports and compares different approaches: classification from words, concepts … using a TF/IDF algorithm. His main contribution is that he could demonstrate the importance of 'no-concept' which improve 3% his results. 'No concept' consists in finding a specific method to detect negation of concepts. For each concept, a separate concept representing it in negated form was included. For example, the concept “pneumonia” would be converted to two concepts, “pneumonia” and “no_pneumonia” for use in the vector representation. In [4] Le Moigno used a combination of syntactic analysis and distance function. Results are pertinent but the scale of evaluation is too small (11 medical reports analyzed).

None of these projects is generic, and none of them takes into account the services in which patients often have multiple disease and in which the system must not only identify the possible codes but also use some rules allowing to classify them. The research works performed until now allows using automatic clustering from natural language but in very limited domains (hematology, radiology, cardiology), often using a specific thesaurus.

4 General Principal

The general principle of CIREA is described on figure 2. It consist in implementing a learning algorithm from medical reports with ICD code. When we present a new report, the classification algorithm will use this learning to propose a code. It will use dictionaries and ontologies.

To reach these objectives, we need to develop several steps:

- Collecting medical reports;
- Creating database;

- Preparing reports;
- Desuffixation algorithm;
- Classification algorithms;
- Evaluation.

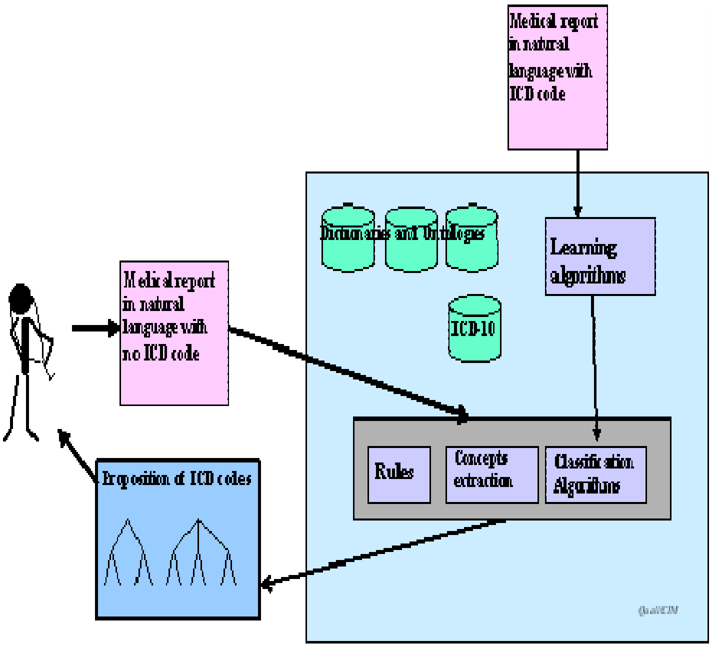
Medical report in natural language with ICD code
Dictionaries and Ontologies
Learning algorithms
ICD-10
Medical report in natural language with no ICD code
Rules
Concepts extraction
Classification Algorithms
Proposition of ICD codes

Fig. 2. General principle of CIREA

5 CIREA Step by Step

The general principle of CIREA is described on figure 2. It consist in implementing a learning algorithm from medical reports with ICD code. When we present a new report, the classification algorithm will use this learning to propose a code. It will use dictionaries and ontologies.

5.1 Collecting and cleaning of medical reports

We collected 30 000 medical reports in 15 hospitals from different geographical places. The main problem was to go in each hospital and try to find how to extract medical reports from the information system.

5.2 Creating the database

We collected several dictionaries:

- Common words (550 000 words);
- Medical concepts (from MeSH and other sources), vocabulary, prefix and suffix;
- ICD 10 classification (54 000 diagnostics);
- Medical acronyms;
- List of stop words.

5.3 Preparing reports

Medical reports are very few structured and where given in several formats (text, specific, MS-Word ….). We did integrate all reports in a unique table under this light structure:

- Reasons of hospitalization;
- Examination;
- Evolution;
- Conclusion.

>From this point, we had a complete database and could begin to perform algorithms.

5.4 Desuffixation and enrichment algorithms

5.4.1. Desuffixation

The first problem we had was: ”how to process words?”. For example, how can we know that ‘neural’ ‘neuron’ neurons’… represents the same thing? We could choose between finding each word in our large database of common and medical words or trying to use an other method like desuffixation or stemming. We have chosen the desuffixation one to avoid the storage of a large set of terms.

Traditional desuffixer algorithms are usually very good for natural language. One of the best algorithms is ‘Porter algorithm’. It has been adapted to French language (Carry), but not to medical language.

So, we have proposed a new algorithm called EDA which is a desuffixer algorithm devoted to the medical language [8]. This algorithm runs in three steps. The first step of this algorithm consists in harmonization:

- All words are in small letters;
- Accentuations are removed;
- Double letters are suppressed;
- Harmonization of special characters ('cœur' becomes 'coeur‘);

Identical sounds are changed in same letter (like ‘y’ and ‘i’, ‘qu’ and ‘k’ and ‘c’).

Initial word	Applied rules	Resulting word
INTESTIN	None	INTESTIN
INTESTINS	1	INTESTIN
INTESTINES	1 and 2	INTESTIN
INTESTINAL	5	INTESTIN
INTESTINAUX	3 and 6	INTESTIN
INTESTINALES	1, 2, and 5	INTESTIN
INTESTINALE	2 and 5	INTESTIN

Fig. 3. Example of result from EDA desuffixer algorithm

The second step consists in applying successively 37 rules (Example: suppress final letter if ‘s’). We can find in figure 3 an example of result.

5.4.2 Enrichment by analyzing medical suffixes

The last step is more semantic. Prefix and suffix of medical terms can carry a lot of semantics. Thus our idea is to enrich report text by adding words issued from the analysis of prefixes and suffixes. For example, every time a word begins by 'cardio….' We add term 'heart' to medical report.

5.4.3 Measures of the gain

We did experiment EDA using Naïve Bayes [9]:

$$p(c_j / d) = \frac{p(c_j) * p(d / c_j)}{p(d)} \quad \text{with} \quad p(d) = \sum_{j=1}^{|C|} p(c_j) p(d / c_j)$$

The result shows (see figure 4) that we improve the classification of 5.5%.

Desuffixer	Precision	Recall	F measure
None	0.715	0.671	0.692
Carry (French 'Porter' algorithm)	0.747	0.700	0.723
EDA	0.772	0.724	**0.747**

Fig. 4. Results of classification with Desuffixer

5.5 Classification algorithms

We have proposed and implemented the CLO3 algorithm for multi label classification. One of the problems was that for one medical report –on the contrary of usual classification applications- we have several diagnostics to identify. As we can see on figure 5 there are between 1 and 32 diagnostics by patient.

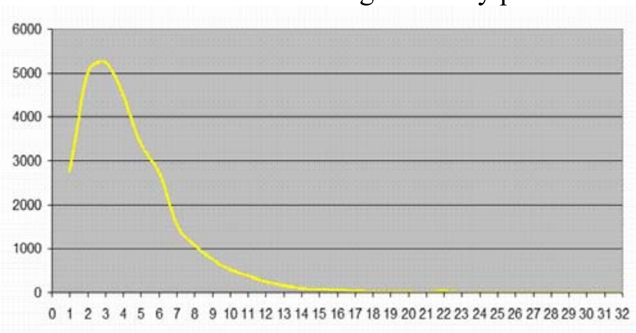

Fig. 5. Number of diagnostics for one patient

First, we consider that there exists a relationship between usage of terms and diagnostics. For example, we suppose that we can often see word 'coma' for diagnostics of 'coma'. We built a cross table with word (or concept), diagnostic, and count (see figure 6).

	Diagnostic 1	Diagnostic 2	….	sum
Concept 1 or word 1				
Concept 2 or word 2		Count of concept 2 found with diagnostic 2		
…				

Fig. 6. Cross table used by classification algorithm

Considering the previous relationship, we compute the weight of the concentration that we call brut_weight (kind of coefficient of variation):

Brut Weight = var(frequency) / avg(frequency)

But, we don't want a big result for rare case (for example, a couple found just one time), so we have to correct the weight like this:

Net Weight = Brut Weight frequency * count*

Finally, we 'normalize' weights by dividing each one by the average.

WeightA(couple)=NetWeight(couple)/avg(NetWieght (word))

We considered a second approach could be significant, inspired from probabilities. So we computed a second weight:

WeightB (couple) = N(couple) / N(term)

Finally, we had to find a combination between these two weights and we proposed as final result:

*weight_CLO3 = WeightA2 * WeightB5*

Computing classification with CLO3 algorithm improves result of 6.7%

5.6 Evaluation

The F-measure is the most usual indicator to evaluate the results of an automatic classification algorithm. However in this particular application we had two problems with evaluation:

How can we know if predicted result is correct?

How to represent the high level needs of doctors?

Our first question was: is coding reproducible. If yes, we could suppose that effective code is the right one. If no, we would need a manual evaluation to know if a proposed code not in the initial one is right or not. We analyzed 100 medical reports and gave each one to two doctors. Finally we compared initial codes to those proposed by two doctors [6]. We could measure that only 18% of codes are reproducible, as shown in figure 7.

Natural language processing produces many algorithms for classification, categorization and information retrieval. The performance of these algorithms is computed from several measures, like precision and recall. To make easier the reading of performance, [11] created a synthetic measure: the F-measure, which is a combination of these two indicators.

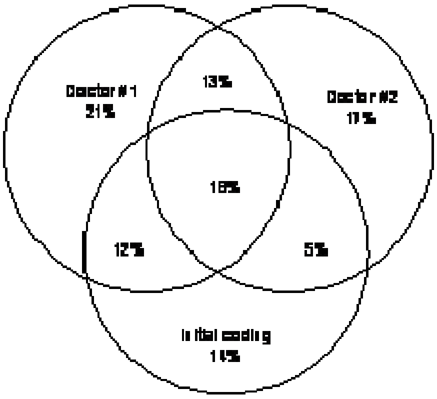

Fig. 7. Reproducibility of ICD codes from medical reports

Today, needs are diversified, problems are more complex, but we keep the same indicator since 25 years [10]. This usage is it still justified? Without renouncing existing scales, how to integer new needs? To represent a high level of exigencies (our second problem) we computed a new measure: K-measure [7].

$$K\text{-}Measure = (1+\beta^2)*(Precision*Recall)^{\alpha} / ((\beta^2*Precision)+Recall)$$

The α parameter allows to determine a high level of need. This is our case in medical domain. Doctors prefer that score strongly goes down when a result is bad. In our experimentation, we use α=1.5. Figure 8 shows the result of experimentation of k-measure with α=1.5.

Algorithm	Precision	Recall	F-Measure	K-Measure
CLO3	0.804	0.733	0.767	**0.589**
Naïve Bayes	0.734	0.669	0.700	**0.490**
Difference	0.070	0.064	0.067	**0.099**

Fig. 8. Experimentation of F-measure and K-measure

It is interesting to notice that K-measure score is worst than F-measure. But the most interesting is that in our case, when F-measure increases 6.7%, satisfaction of user (computed by K-measure) increases of 9.9 %.

6 Conclusion and Future Works

This paper presents CIREA project which consists in an automated classification of diseases from textual medical reports. Our first results are satisfying the needs, but we will go further.

We plan to use linguistic treatment to explore training inductive mechanisms. We wish to exploit various sections of Hospitalization Reports to measure their contribution to diagnosis definition. We suppose that hospitalization report's sections have a variable contribution for diagnosis determination. For example, the conclusion zone should be more discriminating than others. But this postulate remains to be measured and evaluated. If it were proven, it should strongly optimize results. We plan to explore other ways, like:

- Extraction of concepts instead of words;
- Comparison between different algorithms for classification (Knn, SVM…);
- Usage of ontology;
- New algorithm (CLO4) for classification.

Acknowledgements:

This project has been partially founded by the French Institute “Ministere de l’education nationale de l’enseignement et de la recherché” in the framework of the RNTS RHEA project.

References:

[1] A. Blanquet, P. Zweigenbaum, “A lexical method for assisted extraction and coding of ICD-10 diagnoses from text patient discharge summaries”. In proceedings of TALN99.

[2] Burnside, Strasberg, and Rubin, "Automated Indexing of Mammography Reports Using Linear Least Squares Fit". Stanford Medical Informatics, Stanford, CA, 2000.

[3] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". Proceedings of the Tenth European Conference on Machine Learning (ECML'98), Springer Verlag, 137-142, 1998.

[4] S. Le Moigno, J. Charlet, D. Bourigault, M-C Jaulent, "Construction d’une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale", Actes des 13èmes journées francophones d'ingénierie des connaissances (IC 2002).

[5] E. Metais, D. Nakache et J.-F. Timsit, "Rhea: A Decision Support System for Intensive Care Units". In 4th International Multiconference on Computer Science and Information Technology CSIT, 5-7 April, Amman (Jordany), 2006.

[6] B. Misset, J.F. Timsit, E. Metais, D. Nakache, S. Dumont, D. Lassence A, D. M, G. Orgeas, "Reproductibilité des codages diagnostiques en réanimation Projet CIREA 1ere partie". In XXXIII conférence de la SRLF, 2005.

[7] D. Nakache, E. Metais and J. F. Timsit, "Evaluation for NLP" . In DEXA 05, Copenhague, Danmark, Aout, pp. 626-637, Springer Verlag, 2005.

[8] D. Nakache, E. Métais and A. Dierstein, "EDA : algorithme de désuffixation du langage médical". In EGC 2006, Lille, janvier 2006.

[9] K-M. Schneider, "On Word Frequency and Negative Evidence in Naive Bayes Text Classification", in proceedings of ESTAL2004, Adavances in Natural Language Processing, Alicante, Spain, LNAI Lecture Notes in Artificial Intelligence, Springer Verlag 2004, Volume 3230, pp 474-485.

[10] K. Sparck Jones, “Automatic language and information processing: Rethinking evaluation”. Natural Language Engineering, 7(1):29–46. 2001.

[11] K. Van Rijsbergen, “Information Retrieval”, (2nd Ed.) Butterworths, London. www.dcs.gla.ac.uk/Keith/Preface.html.1979.

[12] V.N. Vapnik, "The Nature of Statistical Learning Theory". Springer, 1995.

[13] A. Wilcox, G. Hripcsak; "Medical Text Representations for Inductive Learning" and Adam Wilcox, George Hripcsak, Carol Friedman: "Using Knowledge Sources to Improve Classification of Medical Text Reports", Held at KDD-2000, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2000, Boston, MA, USA.

[14] Y. Yang, C. G. Chute, "An example-based mapping method for text categorization and retrieval". ACM Transactions on Information Systems, 12(3), 252-277, 1994.

[15] P. Zweigenbaum and al., "MENELAS, The final report", Ménélas deliverable#17, Paris, 1995.

[16] P. Zweigenbaum, "Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances". Innovation Stratégique en Information de Santé, (2-3):27-47, 1999.