Machine learning and features selection for semi-automatic ICD-9-CM encoding

Julia Medori
CENTAL
Université catholique de Louvain
Place Blaise Pascal, 1
1348 Louvain-la-neuve
julia.medori@uclouvain.be

Cédrick Fairon
CENTAL
Université catholique de Louvain
Place Blaise Pascal, 1
1348 Louvain-la-neuve
cedrick.fairon@uclouvain.be

Abstract

This paper describes the architecture of an encoding system which aim is to be implemented as a coding help at the *Cliniques universtaires Saint-Luc*, a hospital in Brussels. This paper focuses on machine learning methods, more specifically, on the appropriate set of attributes to be chosen in order to optimize the results of these methods. A series of four experiments was conducted on a baseline method: Naïve Bayes with varying sets of attributes. These experiments showed that a first step consisting in the extraction of information to be coded (such as diseases, procedures, aggravating factors, etc.) is essential. It also demonstrated the importance of stemming features. Restraining the classes to categories resulted in a recall of 81.1 %.

1 Introduction

This paper describes a series of experiments carried out within the framework of the CAPADIS project.[1] This project is the product of a collaboration between the UCL (Université catholique de Louvain, Belgium) and the Cliniques universitaires Saint-Luc. Saint-Luc is one of the major hospitals in Belgium. Each year, a team of file clerks processes more than 85,000 patient discharge summaries and assigns to each of them classification codes taken from the ICD-9-CM (International Classification of Diseases – Ninth Revision – Clinical modification) (PMIC, 2005).

The encoding of clinical notes (or patient discharge summaries) into nomenclatures such as the International Classification of Diseases (ICD) is a time-consuming, yet necessary task in hospitals. This essential process aims at evaluating the costs and budget in each medical unit. In Belgium, these data are sent to the National Health Department so as to compute part of the hospital's funding.

Our aim is to help coders with their ever-growing workload. More and more patients' stays need to be encoded while the number of coders remains the same. Our goal is therefore to develop an semi-automatic encoding system where the role of the coders would be to check and complete the codes provided by the system.

This paper focuses on machine learning methods as automatic encoding techniques. More specifically, it focuses on the appropriate set of attributes to be chosen in order to optimize the results of these methods.

It will therefore present the structure of the system and compare the results of different inputs to the machine learning approach. Section 2 gives a more detailed description of the objectives of this project. Section 3 gives an overview of the architecture of the system: first, the extraction part will be described, and then, the automatic encoding stage will be discussed. Section 4 will focus on the machine learning experiments that

[1] http://www.iwoib.irisnet.be/PRFB/t10/t10_medori_fr.html

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 84–89,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

were conducted. The results will be presented and discussed in sections 5 and 6.

2 Objectives

Since the early 1990s and the rise of the computational linguistics field, many scientists have looked into the possible automation of the encoding process (Ananiadou and McNaught, 2006; Ceusters et al., 1994; Deville etal., 1996; Friedman et al., 2004; Sager et al., 1995; Zweigenbaum et al., 1995). Two different approaches distinguish themselves from one another: a symbolic approach as in (Pereira et al., 2006) and a statistical one. Both methods scored highly in the "Computational Medicine Challenge" (CMC) organized by the "National Library of Medicine" in 2007 (Pestian et al., 2007): among the best three systems, two combined a statistic and a symbolic approach and only one relies only on a symbolic approach. Most systems participating took a hybrid approach as in (Farkas and Szarvas, 2008).

During ACL 2007, Aronson (2007) presented within the framework of the same challenge, four different approaches, symbolic, statistical and hybrid. His conclusion was that combining different methods and approaches performed better and were more stable than their contributing methods. Pakhomov (2006) describes Autocoder, an automatic encoding system implemented at Mayo Clinic that combines example-based rules and a machine learning module using Naïve Bayes.

Within the scope of this challenge, only a limited number of codes were involved.

The objective of our work is to build such a tool to help the team of coders from the *Cliniques Universitaires Saint-Luc*. Three facts are noteworthy: the clinical notes we work on are written in French; they originate from all medical units; and all the codes from the ICD are used in the process (around 15,000). Most studies are limited on at least one of these criteria: most systems are developed on English as more language resources are available, and they often focus on specific types of notes, e.g. the CMC focused on radiology reports.

3 System description

The system is divided into two units: an extraction unit which aims at marking up information considered as relevant in the encoding process, and an encoding unit which, from extracted information generates a list of codes.

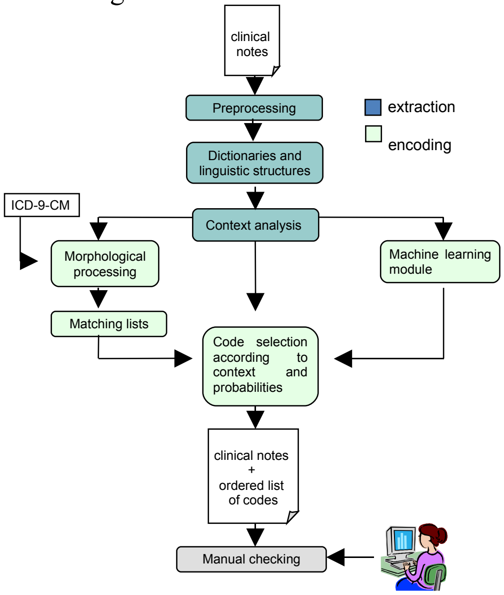

Figure 1. System structure

Extraction: The system aims at reproducing the work of human coders. Coders first read the text, extract all the pieces of information that have to be encoded, and 'translate' information into codes of the ICD-9-CM. The idea behind our tool is to recreate this process. The main source of information coders use are the patient discharge summaries written by doctors summarizing all that happened during the patient's stay: diagnoses, procedures, as well as the aggravating factors, the patient's medical history, etc. These files are electronic documents written in free text with no specific structure.

We developed a tool which aims at extracting the necessary information from these texts: terms referring to diseases but also anatomical terms, the degree of seriousness or probability of a disease, aggravating factors such as smoking, allergies, or other types of information that may influence the choice of a code.

There are many ways of referring to the same diagnosis or procedure, we therefore needed to build specialized dictionaries that would comprise as many of these wordings as possible. The dictionaries of diseases and procedures were mainly built automatically using the UMLS and the classifications in French it comprises. Other specialized dictionaries (anatomical terms, medical departments, medications, etc.) were developed from existing lists. These then were gradually completed manually as the development of the extraction tool went on.

However, the plain detection of terms is not sufficient. It is important to detect in which context these terms occur. For instance, a diagnosis that is negated will not be encoded. The identification of contexts required the use of finite-state automata and transducers. These transducers are represented by graphs that describe the linguistic structures indicating specific contexts. These graphs were hand crafted using the UNITEX software tool[2] (Paumier, 2003). An example of a graph matching fractures and sprains is presented in figure 2.[3] Each path of the graph describes a recognized linguistic structure.

Graphs were also used to broaden the scope of the terms detected by dictionaries. For instance, not only do diseases need to be extracted but, to code, one also needs to know which part of the body is affected.

Certain types of diagnoses also have to be described via graphs such as smoking as there are many ways in which to say that someone smokes or not. Ex: “he smokes 3 cigarettes a day.” “He used to smoke.” “Occasionally smokes.” “Heavy smoker.” “Does not smoke.”

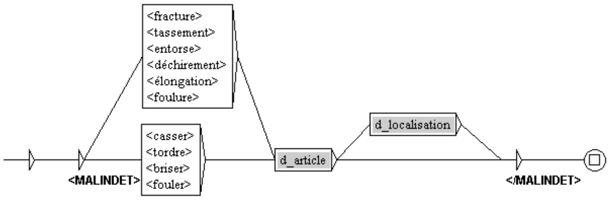

Figure 2. Example of a UNITEX graph matching patterns such as fractures and sprains.

Our aim was to develop a wide-coverage system. We therefore focused mainly on the General Internal Medicine service in order to develop the grammars and dictionaries. It is a very diverse department where physicians have to face all kinds of diseases.

The graphs and dictionaries on which is based our extraction system were built during the first phase of the project. A more detailed description and evaluation of the extraction process can be found in (Medori, 2008).

Encoding: As was said above, two main approaches to the encoding problem coexist: the symbolic approach and the statistical approach. Both have their benefits and drawbacks. The symbolic approach is a time-consuming approach as it involves describing linguistic rules linking text to diseases. The statistical approach has the advantage of being fast to compute but the need for a large amount of data often hampers the use of these methods. However, both methods give reliable results, and a combination of both is the option generally favored. In our context, we chose to combine both approaches as a large corpus of clinical notes is at our disposal.

Saint-Luc provided us with a corpus of 166,670 clinical notes. The codes that were assigned to them by the coders were also provided. This corpus gives us the chance to develop and test statistical methods in a ‘real life’ experiment.[4]

However, whatever the results, we will need to combine these methods with linguistic rules. There are two main reasons for this : in the near future, we will have to face the problem of having to switch to another classification. The change to ICD-10-CM is planned for 2015. Therefore, at that time, we will not have enough learning data to be able to generate the list of codes in a statistical manner. The second reason is that there are codes that are seldom assigned and for which we will not have enough occurrences in our corpus to be able to extract them statistically.

This paper focuses on the statistical tests that were conducted on our corpus. An insight into a symbolic method using the matching of morphemes can be found in (Medori, 2008).

[2] http://www-igm.univ-mlv.fr/~unitex/

[3] The grey boxes indicate calls to other graphs. Here, *d_localisation* is a graph matching anatomical terms.

[4] In this paper, the experiments were conducted on a smaller corpus. At a later stage, the methods chosen for the final system will need to be trained on the full corpus.

4 Experiment

As a first encoding experiment, we chose to focus on a baseline machine learning method: Naïve Bayes. This method has often been used and proves to be robust.

To conduct this experiment, we used Weka, a data mining software[5] developed at the University of Waikato. For more information on this tool, see (Witten, 2004).

In order to test this method we took a sub-set of 19,994 discharge summaries from the General Internal Medicine department. In order to test how necessary the extraction step is, we chose the texts from the department on which the development of the extraction rules were based.

These notes were assigned 102,855 codes which makes up 4,039 distinct codes.

This corpus was then divided into two subsets: 90% of the 19,994 patient discharge summaries were used as the training corpus and 10% as the test set.

As with any machine learning method, enough data for each class is needed in the training set in order to be able to classify correctly. Therefore, we built a classifier for each code that was manually assigned at least 6 times in our corpus. This resulted in 1,497 classifiers, which means that we did not have enough data to be able to assign 2,542 codes which make up 5% of all the assigned codes.

Four experiments were conducted:

Experiment 1. In our first experiment, the selected attributes were the terms that were highlighted as diagnoses by the extraction step. The diagnoses identified in a negative context were removed from the features list. These resulting list of extracted terms went through a normalization process: accents and stop words were removed; words were decapitalized.

Experiment 2. The second experiment aimed at proving the relevance of the stemming of these terms. The attributes in this experiment were therefore the terms that were extracted, then normalized and stemmed using Snowball Stemmer[6] which is an implementation of the Porter algorithm.

Experiment 3. In this third experiment, we wanted to check the relevance of the extraction process (see experiments 1 and 2). Therefore, the attributes comprised all the words contained in the clinical notes apart from stop words. The words were stemmed in the same way as the extracted terms in experiment 2.

Experiment 4. In all the previous experiments, the classes to be assigned consisted in codes. In this experiment, classes are reduced to categories of codes: represented by the first three digits of a code. The attributes are the same as in experiment 1: extracted terms (non-stemmed). As the system is designed as a coding help i.e. its aim is to generate a list of suggested codes, and not as a fully automated encoding system, one could imagine listing categories of codes instead of codes themselves and then let the coders look up in the hierarchy for the appropriate code within the selected category.

At the end of each experiment, we end up with a list of the 1,497 codes from ICD-9-CM ordered by their Naïve Bayes score for each letter.

The measure that is most interesting here is the recall. The list of suggested codes needs to comprise most of the codes the coder will need so that he/she does not have to go elsewhere to find the appropriate code. Therefore, we kept three measures of recall. It is important to keep the list of codes to be presented to the user short and manageable. Larkey and Croft (1995) used the same measures and set the limit number of codes to 20. This choice is arbitrary but seems like a sensible limit. In Saint-Luc, the maximum number of codes a file clerk can assign to a patient discharge summary is 26 (the principal diagnosis is assigned the letter A and all the other codes are ordered according to the other letters of the alphabet). However, few reports are actually assigned 26 codes (15 out of 19,994). The average number of codes assigned by the file clerks in our set of 19,994 discharge summaries is 6.2.

The three measures of recall are **Recall10**, **Recall15** and **Recall20** which are the measures of micro averaged recall if we show the first 10, 15 and 20 most likely codes respectively.[7]

[5] http://www.cs.waikato.ac.nz/~ml/weka/

[6] http://snowball.tartarus.org/

[7] It should be noted that we keep in the list of suggested codes all the codes that tie last with the 10th, 15th and 20th position respectively.

5 Results

The results of the experiments described above are detailed in figure 3.

	Rec10	Rec15	Rec20
1(att: extracted terms)	50.4	56.4	60.5
2 (att: stemmed extracted terms)	56.1	64.1	69.1
3 (att: all words, stemmed)	39.1	40.3	41.4
4 (att: extracted terms classes : categories)	64.0	75.1	81.1

Figure 3. Recall for each experiment (in %)

Experiment 1. From the results of this baseline experiment, considering the extracted terms and retaining the 20 most likely codes according to the Naïve Bayesian statistics, more than 60% of the codes manually assigned to the test notes can be found in this list.

Experiment 2. The stemming of the extracted terms increased the recall by 8.6%.

Experiment 3. If considering all the words as attributes, the recall when retaining 20 possible candidates is around 40% while when attributes are selected through the extraction process, the recall increases to 69% which is an increase of about 28%. This result proves that the extraction process is an essential step in the system and clearly improves the performance of the statistical encoding unit.

Experiment 4. When classes are limited to categories, Recall20 jumps to 81.1% which is 20.6% more than in experiment 1 which was conducted with the same attributes but where classes were codes. This supports our idea that showing a list of categories instead of codes could be an interesting alternative for coders: they would be shown more codes while keeping the list manageable, and then could browse easily into the sub-structure of the classification.

6 Discussion

The choice of attributes is important when testing machine learning methods. In the framework of the development of an encoding system, we proved that a first step consisting in selecting the terms carrying the information that needs to be encoded is essential. We also showed that the use of a simple stemming algorithm clearly improves the performance of the method.

In the last experiment, classifying the clinical notes by categories of codes resulted in a recall of 81.1%. This reinforces our opinion that, to make sure that all the needed codes are present for the coder, we could list categories and let him/her browse through the codes from there.

It is important to put these results in light of where the codes originate. Most of the information that needs to be encoded is present in these clinical notes. However, even though efforts are made in order for this to change, many physicians still do not compile all the information into these notes. Coders therefore still have to look up into the whole patient record in order to find additional codes. The proportion of codes that cannot be inferred from the clinical notes can be very high. A study conducted by Sabine Regout, a patient discharge summary specialist in Saint-Luc, on 250 clinical notes from 25 medical units, showed that in most departments, 15 to 20% of the codes assigned by the clerks cannot be inferred from the notes. This proportion can increase up to 80% in some surgery departments. This evaluation proves that without a change of mind-set from the physicians, our system can only aim to be a coding help for file clerks. Analyzing all the different types of documents contained in patient records would be a difficult task as they comprise a variety of documents with different structures and formats, and some of them are hand-written documents. For our experiments, this also means that the maximal recall value we will be able to get is around 80%.

In these experiments, we were not able to check the inter-annotator agreement but we must keep in mind that, as in any classification task where humans set the gold-standard, one must expect some degree of errors and variation in the coding.

Another observation influences the maximal number of codes we will be able to retrieve is that we built classifiers for all the codes for which we had enough data. This lead to the building of 1497 classifiers. This represents 95% of all the codes assigned to our test notes. This decreases our maximal recall value by 5%.

The codes that are seldom assigned will therefore never show up in our list of suggested codes. This is rather problematic and other nonstatistical methods will be needed to make up for this.

7 Future work

In the light of these results, the next step will be to conduct an experiment on categories as classes using stemmed extracted terms as features. This should improve further the 81.1% recall from the results of experiment 4.

These experiments were conducted in order to select the right features to be used as attributes for our machine learning module. We chose Naïve Bayes as a baseline method. However, other methods have been tested in previous works (Larkey and Croft, 1995) and have proved to give good results as well, such as k-nearest neighbors or Support Vector Machines.

We saw, at the end of section 6, that symbolic methods need to be developed in order to assist machine learning methods. Machine learning techniques have their limitations: they cannot assign codes for which they did not have enough data, and they cannot face the change to a new nomenclature. Therefore, in the near future we will have to develop a symbolic module comprising a series of linguistic rules in order to do the matching on all codes. A prototype based on the matching of morphemes has already been developed but will need to be experimented further.

The results of the experiments we conducted on a machine learning method were promising. Now, combining these two different approaches is the next challenging task in our project.

Acknowledgements

The CAPADIS project was funded by the government of the Brussels-Capital Region (Institute for the encouragement of Scientific Research and Innovation of Brussels) within the framework of the "Prospective research for Brussels 2006" program. I would also like to thank Benoît Debande, Claire Beguin, Sabine Regout and the Saint-Luc hospital team of coders.

References

Ananiadou S., McNaught J.: Introduction to Text Mining in Biology. In Ananiadou S., McNaught J. (eds.) Text Mining for Biology and Biomedicine, pp 1--12, Artech House Books (2006).

Aronson A. R.: MetaMap: Mapping Text to the UMLS Metathesaurus (2006).

Ceusters W., Michel C., Penson D., Mauclet E.: Semi-automated encoding of diagnoses and medical procedures combining ICD-9-CM with computational-linguistic tools. Ann Med Milit Belg;8(2):53—58 (1994).

Deville G., Herbigniaux E., Mousel P., Thienpont G., Wéry M.: ANTHEM: Advanced Natural Language Interface for Multilingual Text Generation in Healthcare (1996).

Farkas R., Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems', BMC Bioinformatics, 9 (2008).

Friedman C., Shagina L., Lussier Y.A., Hripcsak G.: Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392--402. Epub 2004 Jun 7 (2004).

Larkey L. S, Croft W. B. Automatic assignment of icd9 codes to discharge summaries. Technical report, University of Massachusetts at Amherst, Amherst, MA (1995).

Medori J. From Free Text to ICD: Development of a Coding Help, In: Proceedings of Louhi 08, Turku, 3-4 sept 2008 (2008).

Pakhomov S. V. S., Buntrock J. D., Chute C. G.: Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques (2006).

Paumier S. De la reconnaissance de formes linguistiques à l'analyse syntaxique. PhD thesis. Université de Marne-la-Vallée (2003).

Practice Management Information Corporation. ICD-9-CM Hospital Edition, International Classification of Diseases 9th Revision, Clinical Modification (Color-Coded, Volumes 1-3, Thumb-Indexed) (2005).

Pereira S., Névéol A., Massari P., Joubert M., Darmoni S.J. : Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. Proc. MIE. (2006).

Pestian J. P., Brew C., Matykiewicz P.M., Hovermale D.J., Johnson N., Cohen K.B., Duch W.: A shared task involving multi-label classification of clinical free text. Proceedings of ACL BioNLP; 2007 Jun; Prague (2007).

Sager N., Lyman M., Nhán N., Tick L.: Medical language processing: Applications to patient data representation and automatic encoding. Methods of Information in Medicine, (34):140 -- 146 (1995).

Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers. 2nd edition. 560 pp. ISBN 0-12-088407-0 (2005).

Zweigenbaum P. and Consortium MENELAS: MENELAS: coding and information retrieval from natural language patient discharge summaries. In Laires M. F., Ladeira M. J., Christensen J. P., (eds.), Advances in Health Telematics, pages 82-89. IOS Press, Amsterdam, 1995. MENELAS Final Edited Progress Report (1995).