Medical Informatics Europe 2002
G Surján et al. (eds)
IOS Press 2002

Automatic Diagnosis Classification of Patient Discharge Letters

Dr. Nikitas N. KARANIKOLAS[1], Prof. Christos SKOURLAS[2]

System Head, Areteion University Hospital, University of Athens,
76 Vas. Sofias, 115 28, Greece, nnk@aretaieio.uoa.gr
Dept. of Informatics, Technological – Educational Institute of Athens,
Ag. Spyridonos, 112 10, Greece, cskourlas@hol.gr

Abstract. CAIRN (Computer Assisted Medical Information Resource Navigation) is a prototyping System that allows flexible medical data storage and retrieval supporting medical informatics research. In this paper methods that automate the selection of ICD-9 diagnosis (International Classification of Diseases and Diagnoses, 9th Revision) are investigated. We present the Text Data Mining module extension of CAIRN and its application in order to organize in a systematic way uncontrolled terms, to propose relationships between uncontrolled terms and finally aid the diagnosis classification.

1. Introduction

1.1. Motivation

The experience of last six years from the application of a new Hospital Information System (HIS) in the Areteion University Hospital [1], in Athens, has revealed major difficulties. A major one is that physicians feel quite unhappy because they have the duty to select the correct, for each patient, ICD-9 diagnosis, looking at thousands of possible ones.

Patient discharge letters, written by the same doctors, can be thought as a narrative that describes in more detail the ICD-9 diagnosis that the physicians are looking for. The main target of our pilot study is to investigate methods that offer to the physicians the opportunity to automatically or semi-automatically match a patient discharge letter against the ICD-9 list and retrieve the most likely ICD-9 diagnosis for each patient case.

1.2. A technical discussion of the problems

CAIRN (Computer Assisted medical Information Resources Navigation) is a medical information retrieval system that allows physicians to store full text medical information from any resource, organize and retrieve it, using queries in Natural language [2]. Moreover CAIRN is a prototyping System supporting medical informatics research. Another example of a medical information and retrieval system used for reasons apart from the well known ones incorporated into the classical Hospital Information Systems is presented in [3]. Information Retrieval systems (IR) [4] has been mainly used for searching and matching a user's query against a collection of documents. In most cases the submitted queries vary in size between one word to few sentences. On the other hand the size of the documents can vary from a few sentences to some pages.

In our study, we try to use our CAIRN system to inverse this situation: The size of queries, usually, varies from twenty to seventy lines while the documents (ICD terms) that the system tries to match to the queries does not exceed ten-words length. Hence, the traditional IR process (of retrieving a ranked list of documents in response to a statement (query) expressing our information needs) was applied as a first step to automatically focus on some ICD-9 diagnoses. The process is based on information mainly included in the discharge letters.

Hence, our source of interest consists of the discharge letters (see Fig. 1) and the International Classification of Diseases and Diagnoses, 9th Revision (ICD-9) list. ICD-9, in the CAIRN system, has to be used for the classification of morbidity and mortality information for statistical purposes, and for the indexing of hospital records by disease and operations, for data storage and retrieval.

Apart from the IR and Document Management possibilities of CAIRN system [2, 5] we decided to extend it, trying to incorporate into our system concepts related to Knowledge Discovery in Databases and Data Mining [6, 7].

2. Background - Text Data Mining

Helma [8] discusses what is the Data Mining task, in general, the Representation Languages, the Data Mining Models and presents some of the search strategies employed for finding models. Hearst [9] discusses what Text Data Mining is, and its relationship with IR and Computational Linguistics. According to Hearst the Exploratory Data Analysis is what Text Data Mining must constitute. He also gives examples of using Text to Form Hypotheses about Disease and Uncover Social Impact. Amini [10] proposes a method that is an instance of the Classification EM algorithm and relies on a "discriminant" approach instead of using density estimation, as proposed by other classical methods of unsupervised and semi-supervised learning. This method is used for automatic text summarization in the context of sentence extraction based summarization and is a binary classification model that has the advantage of taking into account the coherence of the whole set of relevant sentences for the summaries.

Discharge Diagnosis
Primary malignant neoplasm of the liver
Admitting Diagnosis
Liver cancer
Past history and Presentation
71 years old male patient with a history of a AAA (Abdominal Aortic Aneurysm) repair 13 months ago presents with RUQ (Right Upper Quadrum) pain and a palpable mass of two months duration. An abdominal CT scan showed a 12 cm tumor in the right lobe of the liver and a smaller one in segment V. He is admitted for surgical treatment.
Progress notes
The pt. underwent a full pre-op evaluation including vascular consultation and cardiac and pulmonary evaluation. An MRI showed a multifocal liver tumor that proved to be by a FNA (Fine Needle Aspiration) a moderately differentiated primary hepatocellular carcinoma. An upper and lower GI (Gastro Intenstinal) endoscopy, as well as a chest CT scan were negative. On 31/8/00 he underwent an extended right hepatectomy and was transferred to the ICU (Intensive Care Unit). His post-op course was complicated by pulmonary and liver failure. His bilirubin reached 12.7mg/dl and his AST, ALT and γGT were markedly increased. His renal function also deteriorated with a peak creatinine level of 2.9 mg/dl. Eventually his condition improved and was transferred to the ward where he was treated with diuretics and TPN (Total Parenteral Nutrition). On POD (Post Operative Date) #8 he developed wound infection and the wound was opened and drained. He also received appropriate antibiotics. He developed also decubitus ulcers, which were surgically debrided. On POD#15 his condition improved, was started on oral feedings and was ambulated. His albs showed improvement. Eventually his ascites diminished and his liver and renal function improved. He was discharged with the following labs: bil=4.9 mg/dl, alb=2.3g/dl, cr=1.75 mg/dl, γGT=136, WBC=6600 and Ht=35.3%
Discharge orders
High protein diet
Return for a follow up visit in 4 months

Figure 1: patient discharge letter example

Our study has focused on the extraction of a model (Rule or Tree) that is able to characterize some of a document's sentences as more salient for classification. Hence our model is a Text Data Mining approach.

3. Materials and Methods

3.1. Patient discharge letter

Let us consider the case of *patient discharge letters* (see figure 1). Patients' exit from the Hospital is always accompanied by this obligatory document which is written and signed by the doctor attested to the patient. These letters form a potential source of information for extracting the ICD codes related with the diagnosis.

3.2. Structure and Handling

All discharge letters share a common structure (form). They include information organized in text mode having the following subheadings:

3.2.1. Final diagnosis

A key text field for our study. It is written as a plain text (in the worst case) or is written (exclusively) *using some uncontrolled terms* summarizing the present situation of the patient and the final diagnosis. Uncontrolled Terms (UTs), for librarians, are not a form of plain text but phrases like keywords. The difference is that keywords have to belongs to a specific authority list (or in a thesaurus) e.g. ICD-9. Hence, you can form a list of UTs and using various methods to form an authority list of keywords in the future. For example two such UTs in Greek and their translations are:

- κακοήθη νεοπλάσματα ήπατος, πρωτοπαθές (malignant neoplasm of the liver, primary)
- κακοήθη νεοπλάσματα αριστερής κολικής καμπής (malignant neoplasm of left colic flexure)

CAIRN offers us many possibilities for *extracting and proposing, even from the plain text, uncontrolled terms*. The methods that CAIRN uses to identify and propose possible UTs is based on [11]. We used these terms as the source of a first method of attack and some simple statistics were extracted related to the terms used by specific doctors, who wrote and signed such letters. Results were given to the doctors and discussed and were eventually integrated in a list for all the doctors in the Second Surgery Clinic of the Areteion University Hospital. The rationale of using this method was our attempt to organize in a more systematic way all these UTs, eliminate spelling errors, and incorporate a subsystem of handling keywords in the future CAIRN system.

3.2.2. Initial diagnosis

This text field is similar to the previous one and is handled using the same way (method). The essential difference with the final diagnosis field is the fact that terms (or the plain text) in the initial diagnosis give the rationale (reasons) for the patients' entry into the Hospital while the terms in the final diagnosis summarize and "illustrate" the present situation. For example two such terms in Greek and their translations are:

- CA-ήπατος (liver cancer)
- Αδενο-CA σιγμοειδούς (adenocarcinoma of sigmoid)

The ability of CAIRN to extract and propose, even from the plain text, UTs can be further verified with methods that measure the percent of existence of n-grams and words in text corpora [12]. In order to study the distribution of UTs we selected, randomly, a small collection of hundred and ninety two (192) patient discharge letters (in the following we call this collection as "training set"). In the following tables we present the distribution of UTs in relation to the ICD-9 codes that characterize the documents (discharge letters) of collection.

Table 1: Greek UT <*ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ> and its relationship with ICD-9 codes

ICD-9 code	ICD-9 diagnosis	%
155	Malignant neoplasm of liver and intrahepatic bile ducts	10
155.0	Malignant neoplasm of Liver, primary	20
155.2	Malignant neoplasm of Liver, not specified as primary or secondary	50
157.8	Malignant neoplasm of pancreas, Other specified sites of pancreas	10
197.7	Secondary malignant neoplasm of Liver, specified as secondary	10

Table 2: Greek UT <*ΟΖΩΔ* ΒΡΟΓΧΟΚΗΛΗ> and its relationship with ICD-9 codes

ICD-9 code	ICD-9 diagnosis	%
241	Nontoxic nodular goiter	33.3
241.1	Nontoxic multinodular goiter	33.3
242.2	Toxic multinodular goiter	16.7
246	Other disorders of thyroid	16.7

3.2.3. Description of the case

In the following tables some of the UTs, extracted from the description of the case part of the discharge letters, are given. Their relationship with ICD-9 code(s) is also presented.

Table 3: Greek UT <ΧΑΜΗΛΗ ΠΡΟΣΘΙΑ ΕΚΤΟΜΗ> and its relationship with ICD-9 codes

ICD-9 code	ICD-9 diagnosis	%
153.3	Malignant neoplasm of colon, Sigmoid colon	25
154	Malignant neoplasm of rectum, rectosigmoid junction, and anus	25
154.1	Malignant neoplasm of Rectum	50

Table 4: Greek UT <ΔΙΑΤΑΣΗ * ΧΟΛΗΦΟΡΩΝ> and its relationship with ICD-9 codes

ICD-9 code	ICD-9 diagnosis	%
0010	Obstructive Jaundice	26.7
155	Malignant neoplasm of liver and intrahepatic bile ducts	6.7
156.2	Malignant neoplasm of gallbladder and extrahepatic bile ducts, Ampulla of Vater	6.7
157	Malignant neoplasm of pancreas	13.3
157.0	Malignant neoplasm of pancreas, Head of pancreas	13.3
442	Other aneurysm	6.7
574	Cholelithiasis	6.7
574.3	Calculus of bile duct with acute cholecystitis	6.7
575.0	Acute cholecystitis	6.7
576.9	Unspecified disorder of biliary tract	6.7

3.3. Result, restrictions and possible improvements

The existence, of only a small subset, of UTs for each ICD-9 code triggered us to investigate the possibility to have a model that could automatically assign ICD-9 codes to patient discharge letters according to the existence of UTs. Hence, we submitted all the automatically extracted UTs to human experts (physicians) and asked them to select the appropriate ones to characterize patient discharge letters. This task resulted to the creation

of an authority list. Then, a vector was automatically created for each document of the training set and each item of these vectors has two possible values (true, false) that represents the existence or not of the corresponding UT in the document. The last item of each vector has the ICD-9 code that characterizes the patient discharge letter (the class of document). Having a set of labeled data we applied the Data Mining algorithm that produced a set of a few classification rules of the form:

$$(A_{\lambda 1} = v_{\lambda 1}) \wedge (A_{\lambda 2} = v_{\lambda 2}) \wedge \ldots \wedge (A_{\lambda j} = v_{\lambda j}) \supset (A_{m+1} = B) \quad (1)$$

where $1 \leq \lambda_1 < \lambda_2 < \ldots < \lambda_j \leq m$ for each attribute A,

m is the number of UTs in the authority list,

$v_i \in$ {true, false} and

$B \in$ {b | b is any valid ICD-9 code}

Our training collection is constituted of 192 patient discharge letters. There are 96 distinguished ICD-9 codes that classify these documents. The selected training set is small but permitted us to have an estimation of the functionality and a measure of the acceptance of our method by the physicians. The training set documents are selected from the second surgery clinic of ARETEION University Hospital.

The selected training set documents are all cases of the same clinic (from the second surgery clinic of Areteion Univ. Hospital). In case that we want to introduce our method to another clinic we have to repeat our data mining method, since other clinics should use another subset of ICD-9 list. The idea to apply our method to a whole hospital seem strange and should be impossible since it will increase the size of attributes (UTs) vector and will increase dramatically the response time.

There are some other attributes that could be used in order to improve the classification ability of the classification rules. For example available information such as the age, sex and the duration of nursing could extend the vector that characterize patient discharge letters and could produce better classification rules. Negatives (e.g. "no malignant neoplasm of liver found") could also be dealt by searching the actual text of a discharge letter and rejecting a proposed ICD-9 code if the text contains a linguistic negation of any UT that is member of the classification rule that proposed this ICD-9 code.

So far, we have two methods to help users select the correct ICD-9 code for a patient discharge letter: The traditional IR solution and the more sophisticated data mined classification rules. We plan to seek the possibility to combine these methods. One possible way could be to present to the user the ICD-9 codes selected by the classification rules as the more salient ones and the ICD-9 codes selected by the IR methodology as alternative solutions. In this way we can bypass extreme cases where the ICD-9 code that fits to the current document had not been used in the training set and consequently could not been proposed by the classification rules.

4. Discussion and Conclusions

We have suggested methods for automatic diagnosis classification of patient discharge letters based on information retrieval and knowledge discovery in databases. The measurement of redundancy and entropy of UTs in correlation with ICD-9 diagnosis can also be used as a tool that offers to the physicians the opportunity to automatically match a patient discharge letter against the ICD-9 list.

Apart from the above mentioned and discussed general structure of a discharge letter there are also some other fields (of the discharge letters) that are case dependent (e.g. Laboratory exercises, Operations' related fields). We shall focus on results related to this

analysis in the near future. There is also a plan to extent the existing HIS to "reflect" the new understanding of information, and use CAIRN in analyzing and storing more cases from other hospital's units.

References.

[1] Papoutsis J: Hospital Information Systems: The case of Areteion University Hospital. PhD Dissertation, Athens University, Greece, 1997.
[2] Karanikolas NN and Skourlas C: Computer Assisted Information Resources Navigation. Medical Informatics & the Internet in Medicine 25: 133-146, 2000.
[3] Dodra W et. al.: Archimed: A medical information and retrieval system. Methods of Information in Medicine 38: 16-24, 1999.
[4] Kowalski G: Information Retrieval Systems. Theory and Implementation, 1st edn (Printed in the USA; Kluwer Academic Publishers), ISBN 0-7923-9899-8, 1997.
[5] Deal DC: Techniques of Document Management: A review of Text Retrieval and related technologies. Journal of Documentation 57, 192-217, 2001.
[6] Miguel ER: Data Mining Patient Records, http://www.cs.uiowa.edu/~mruiz/papers/medinfo/paper/HEALTH-DATAMINING.html
[7] Prather JC et. al.: Medical Data Mining: Knowledge discovery in a clinical data warehouse. Proceedings of the AMIA Annual Fall Symposium, AMIA'1997, pp. 101-105, 1997.
[8] Helma C, Gottmann E and Kramer S: Knowledge Discovery and Data Mining in Toxicology. Statistical Methods in Medical Research 9, 329-58, 2000.
[9] Hearst MA: Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99, 1999.
[10] Amini MR and Gallinari P: Automatic Text Summarization Using Unsupervised and Semi-supervised Learning. Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD'2001, pp. 16-28, 2001.
[11] Crow D and Smith B: DB_Habits: Comparing Minimal Knowledge and Knowledge-Based Approaches to Pattern Recognition in the Domain of User-Computer Interactions. Technical Report 91.21, University of Leeds, UK, July 1991.
[12] Yannakoudakis EJ, Tsomokos I and Hutton PJ: n-Grams and their implications to Natural Language Understanding. Pattern Recognition 23, 509-528, 1990.