Mining discharge letters for diagnoses validation and quality assessment

Pietro Barbieri,[*] Stefano Ballerio,[*] Dario Cerizza,[¥] Mauro Maistrello,[*] Anna Maria Paganoni[§]

Abstract

We present two projects where text mining techniques are applied to free text documents written by clinicians. In the first, text mining is applied to discharge letters related to patients with diagnoses of acute myocardial infarction (by ICD9CM coding). The aim is extracting information on diagnoses to validate them and to integrate administrative databases. In the second, text mining is applied to discharge letters related to patients that received a diagnosis of heart failure (by ICD9CM coding). The aim is assessing the presence of follow-up instructions of doctors to patients, as an aspect of information continuity and of the continuity and quality of care. Results show that text mining is a promising tool both for diagnoses validation and quality of care assessment.

Keywords

Text mining, diagnosis validation, quality of care, discharge letters

1. Introduction

In recent years, text mining has been applied to clinical documents to extract different kinds of information, such as information on diseases, adverse events, medication or pharmacotherapies, and DRG codes. In this paper we present two research projects in which text mining was applied to clinical documents with two different aims: validating diagnoses and assessing the quality of care.

In the first project, a text mining system was applied to discharge letters related to patients with Acute Myocardial Infarction (AMI). All letters were coded in administrative databases with an ICD9CM diagnosis of AMI and text mining was used to check if the letters contained any mentions of the diagnostic elements that by World Health Organization (WHO) standards should be present when a diagnosis of AMI is formulated. Once the system were fully developed, it could be used to extract information to integrate administrative databases, which contain little information on diagnoses and diagnostic accuracy.

In the second project, text mining was applied to discharge letters related to patients with Heart Failure (HF). All letters had an ICD9CM coded diagnosis of HF and text mining was used to check if the letters contained any follow-up instructions of doctors to patients. Joint Commission International standard 3.2 on Access to Care and Continuity of Care states in fact that discharge letters should contain these follow-up instructions, which contribute to information continuity from doctor to patient and thus to the continuity and quality of care. Once the system

[*] Azienda Ospedaliera di Melegnano.
[¥] CEFRIEL – Politecnico di Milano.
[§] Politecnico di Milano, Department of Mathematics «Francesco Brioschi».

were fully developed, it could be used to extract useful information to assess the quality of care from the point of view of information continuity.

2. Text mining for diagnoses validation: objectives and methodology

The first research project was conducted within a wider project whose general objectives were integrating and exploiting health system databases and building thematical records for acute coronary syndromes.[1] Since administrative databases of health care institutions contain little information on diagnoses and diagnostic accuracy, text mining could be used to extract such information from textual databases, so as to integrate administrative databases and build richer thematical databases on AMI.

The text mining system we wanted to develop should process discharge letters of patients admitted to hospital with a diagnosis of AMI and check if the letters contained any mentions of the three diagnostic elements that, by WHO standards, should be present when a diagnosis of AMI is formulated:[2] electrocardiographic evidence, myocardial markers from blood tests, and chest pain. Moreover, the system would check for mentions of secondary symptoms such as diaphoresis.

As a first step, a training set and two test sets of discharge letters were collected (see Table 1).

Set	Hospital	Years	Coded diagnosis
Training set, 188 letters	Uboldo Hospital (Azienda Ospedaliera di Melegnano)	2004-2008	AMI (by ICD9CM codes)
First test set, 150 letters	Uboldo Hospital	2004-2008	AMI (by ICD9CM codes)
Second test set, 100 letters	Uboldo and Vizzolo Predabissi Hospitals (Azienda Ospedaliera di Melegnano)	2008-2010	AMI (by ICD9CM codes)

Table 1 – The training set and the test sets (AMI)

As a second step, we analyzed the training set and the analysis showed that the system would have to face three main difficulties in processing natural language: first, negations should be detected as such, so as to avoid false positive answers (in statements such as *The patient reports feeling nausea but no chest pain*).[3] Second, and for the same reason, certain past facts should be recognized as such and distinguished from facts related to the present admission (in statements such as *In 2002, the patient was admitted to hospital because of strong chest pain*). Third, the language of the letters was not standard (the letters contained many syntactic irregularities, abbreviations, and typos).

The third step was searching for a software tool to process the letters. We finally chose GATE (General Architecture for Text Engineering) [4] because, once provided with the necessary linguistic resources, it could grant the flexibility that was necessary to deal with the non standard language of the letters. Besides, it is a free and robust application.

As a fourth step, we extracted from the training set the linguistic resources that were necessary to process the letters with GATE. First, a lexicon of 1914 word forms was extracted. The

[1] *Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction*. The project was financially supported by Regione Lombardia and the partners in this section were Azienda Ospedaliera di Melegnano, Politecnico di Milano, and CEFRIEL.
[2] At least two out of three of them should be present.
[3] Obviously, the letters were written in Italian. Here and elsewhere, we have translated for better clarity.
[4] The GATE project is based at the University of Sheffield. GATE is freely available at http://gate.ac.uk/.

lexicon contained technical terms such as *necrosis* or *troponin* and common words such as *not* or *year*. Second, a set of 27 JAPE rules (i.e. decision rules that are coded in a specific GATE formalism) was developed to detect and categorize complex expressions such as *electrocardiographic evidence seems to indicate anterior septal ischemia*).

Finally, the system was applied to the letters of the test sets.

3. Text mining for diagnoses validation: results and discussion

The answers of the system, which produces a set of annotations on the processed letters (see Figure 1), were validated by a human reader.

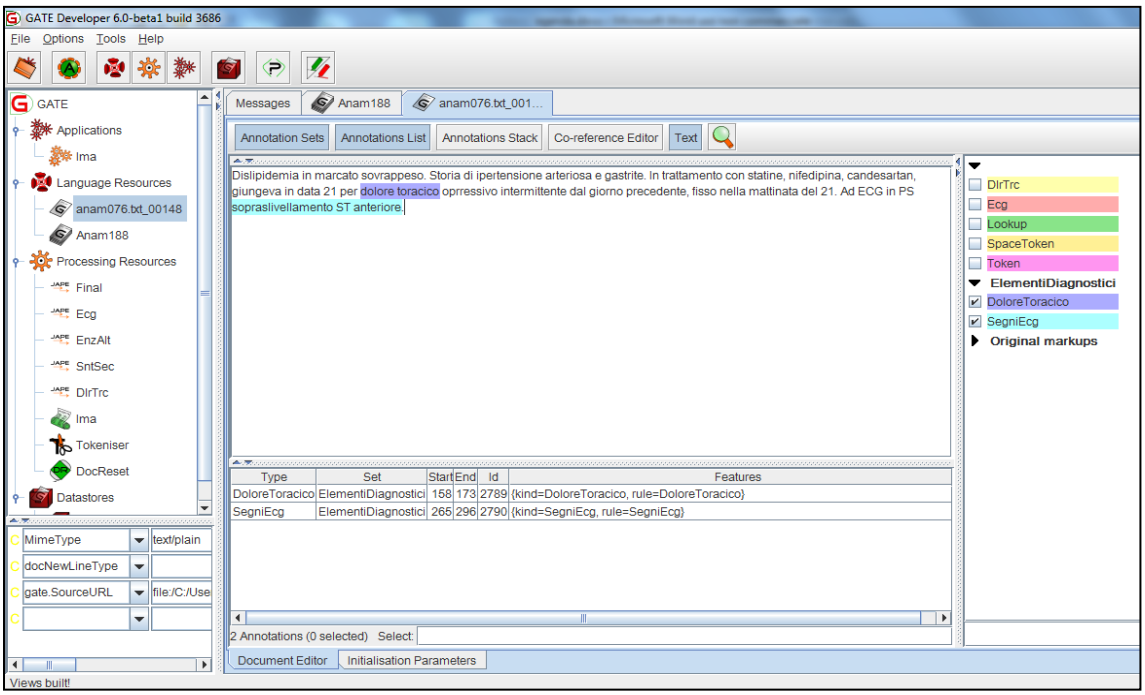

Figure 1 – A discharge letter annotated by the system

Table 2 shows the results of the validation of the answers for the first test set.

First test set	Validation by mention		Validation by letter	
Diagnostic element	Precision	Recall	Precision	Recall
Electrocardiographic evidence	0.913	0.896	0.959	0.959
Myocardial markers	1	0.972	1	0.992
Chest pain	0.907	0.948	0.983	0.958
Secondary symptoms	1	0.974	1	0.971

Table 2 – Validation, first test set (AMI)

Validation was performed by mention and by letter, because the same letter can contain more than one mention of the same diagnostic element. The doctor, for example, might state that myocardial markers from blood tests suggest that acute myocardial infarction has occurred and then report the creatine phosphokinase value and the troponin value from blood tests. In this case, if the system should recognize two out of the three mentions of the «myocardial markers» element, validation by mention would record two true positive answers and one false negative

answer, while validation by letter would only record one true positive answer (because the system correctly reports that the letter under scrutiny contains the «myocardial markers» element). The first validation allows a more precise evaluation of the system; the second is more similar to the way the system might be actually used to support diagnoses validation.

On the whole, results are quite satisfactory. One reason is redundancy: letters often contain more than one mention of the same element and this helps the system achieve good recall (in the validation by letter). A second reason is that letters contain very few deceptive statements: doctors rarely write about chest pain episodes related to previous admissions, for example, and they never report myocardial markers values related to previous infarctions; this helps the system achieve good precision. A third reason is that the letters of the test set came from the same hospital as the letters of the training set. This means that the authors were the same, so that the letters of the two sets were very similar from a linguistic point of view. This third reason is why we chose to validate the system on a second test set, which contained letters written by other doctors too. Table 3 shows the results of this second validation.

Second test set	Validation by mention		Validation by letter	
Diagnostic element	Precision	Recall	Precision	Recall
Electrocardiographic evidence	0.882	0.763	0.968	0.845
Myocardial markers	0.958	0.835	0.980	0.877
Chest pain	0.914	0.895	0.972	0.921
Secondary symptoms	0.968	0.909	1	0.923

Table 3 – Validation, second test set (AMI)

Precision and recall values are inferior, but satisfactory nevertheless.

4. Text mining to assess the quality of care: objectives and methodology

The second research project was conducted within a wider project whose general aim was developing new indicators to measure the continuity of care.[5] Joint Commission International standard 3.2 states that discharge letters should contain follow-up instructions of doctors to patients. These instructions contribute to information continuity and thus to the continuity of care. We wanted to use text mining to measure the presence of follow-up instructions in discharge letters of patients with HF and consequently to measure the continuity of information as an aspect of the quality of care.

By scientific societies guidelines, instructions of doctors to HF patients can belong to one of these categories: avoid efforts; perform moderate physical activity; observe a diet; avoid alcohol; avoid smoking; reduce weight; check weight; check diuresis; avoid places with extreme temperature, humidity and atmospheric pressure values.[6] The system should be able to recognize and distinguish all such instructions.

In this case too we began by collecting the training set and the test set. Table 4 shows the composition of both sets.

Set	Hospital	Years	Coded diagnosis
Training set, 200 letters	Uboldo Hospital	2004-2008	HF (by ICD9CM codes)
Test set, 213 letters	Uboldo Hospital	2004-2008	HF (by ICD9CM codes)

Table 4 – The training set and the test set (HF)

[5] *Experimenting new indicators to measure the continuity of care*. The project was financially supported by Regione Lombardia and it was conducted at Azienda Ospedaliera di Melegnano.

[6] Our categorization was based on guidelines by Associazione Nazionale Medici Cardiologi Ospedalieri, Società Italiana di Cardiologia, and Associazione Nazionale Cardiologi Extraospedalieri.

The analysis of the training set showed three main difficulties: first, instructions were very rare. This meant that little information would be available to build the necessary linguistic resources. Second, the system should be able to distinguish wether a certain fact was mentioned as something that had occurred during hospitalization (*during hospitalization, the patient observed a hypoglucidic diet*) or as something that was prescribed (*after discharge, the patient must observe a hypoglucidic diet*). Third, the language of the letters was not standard.

In this case too we chose to process the letters with GATE and we extracted from the training set the necessary linguistic resources: first, a lexicon of 215 word forms (both technical terms such as *hypoglucidic* and common words such as *recommend*); second, 16 JAPE rules to detect complex expressions (*must observe a hypoglucidic diet*).

The system was applied to the letters of the test set and a human reader validated its answers.

5. Text mining to assess the quality of care: results and discussion

Tables 5 and 6 show the results of the validation, by mention and by letter. The number of positive and negative answers, both true and false, has been added to precision and recall values.

Test set	Validation by mention					
Instruction type	TP	FP	TN	FN	Precision	Recall
Avoid efforts	24	0	192	0	1	1
Moderate activity	1	0	212	0	1	1
Keep on a diet	17	1	193	3	0.944	0.850
Avoid alcohol	5	0	209	0	1	1
Avoid smoking	2	0	210	1	1	0.667
Reduce weight	10	0	204	0	1	1
Check weight	15	0	199	0	1	1
Check diuresis	7	0	207	0	1	1
Temperature, humidity and atmospheric pressure	0	0	213	0	-	-

Table 5 – Validation by mention (HF)

Test set	Validation by letter					
Instruction type	TP	FP	TN	FN	Precision	Recall
Avoid efforts	21	0	192	0	1	1
Moderate activity	1	0	212	0	1	1
Keep on a diet	17	0	193	3	1	0.850
Avoid alcohol	4	0	209	0	1	1
Avoid smoking	2	0	210	1	1	0.667
Reduce weight	9	0	204	0	1	1
Check weight	14	0	199	0	1	1
Check diuresis	6	0	207	0	1	1
Temperature, humidity and atmospheric pressure	0	0	213	0	-	-

Table 6 – Validation by letter (HF)

Again, results are quite satisfactory, but deceptive statements were very few and the letters had been written by the same authors as the letters of the training set. Moreover, precision and recall values related to instruction types such as «avoid alcohol» are not very meaningful because of the very few occurrences of those instructions (perhaps, the low rate of certain instructions is the most relevant datum).

6. Conclusions and future work

Results show that text mining is a promising tool both for diagnoses validation and quality of care assessment. Three conditions appear to be a precise specification of what the system must detect, flexible software tools, and adequate linguistic resources. To further support these conclusions, the system we have presented should be applied to larger collections of letters (written by many different authors and with different coded diagnoses). This is part of the future work we are planning to do, together with developing more linguistic resources and trying new approaches, such as machine learning and statistical classificators, which require larger training sets.