AUTOMATIC ANALYSIS OF ELECTRONIC DISCHARGE LETTERS AS A MEANS TO EVALUATE THE CONTINUITY OF INFORMATION AND OF PATIENT CARE

Stefano Ballerio

Dipartimento di Ingegneria Gestionale, Politecnico di Milano
Milano, Italy
e-mail: stefano.ballerio@polimi.it

Abstract
Joint Commission International standard 3.2 on Access to Care and Continuity of Care states that discharge letters should contain information about follow-up instructions of doctors to patients. We developed a text mining system to analyze a collection of 413 discharge letters of heart failure patients and checked their compliance with standard 3.2. We built a domain-specific ontology and a thesaurus and mined the collection with CASOS AutoMap. After validation, the system sensitivity was 0.484; specificity was 0.834; positive predictive value was 0.555; negative predictive value was 0.790. Improving these results requires more powerful natural language processing tools, but text mining seems a promising way to evaluate the continuity of information and of care.

Keywords: text mining; continuity of care; discharge letters.

1. Introduction

The constant growth of published biomedical research and of archives of documents generated by clinical practice creates the need for tools that help researchers and practitioners of the biomedical field to cope with this large amount of information effectively and efficiently. One such tool is text mining [1], which aims to identify and explore new, interesting and useful patterns in large collections of texts [2]. A text mining operation usually includes this sequence of steps: defining the objectives and heuristics of the research; collecting the texts to be examined; preprocessing each text of the collection; mining the texts; and evaluating the results. In the biomedical field text mining has been applied to classify scientific literature; to develop annotated databases of abstracts and articles; for hypothesis generation in complementary structures in disjoint literatures and to analyze scientific literature in general [3]; and to examine large collections of clinical documents such as X-ray reports [4] [5] and discharge letters [6] [8].

We analyzed a collection of discharge letters of heart failure patients. Joint Commission International standard 3.2 on Access to Care and Continuity of Care states that discharge letters should contain information about follow-up instructions of doctors to patients. Our research focused on this point, so that information was viewed as a factor for the continuity of care. We aimed to experiment with new indicators to evaluate the continuity of information as a factor for the continuity of care and to develop tools to automatically process and analyze electronic documents.

2. Methodology and results

Our collection was composed of 413 discharge letters from the cardiology department of Uboldo Hospital (Azienda Ospedaliera di Melegnano). The letters were related to heart failure patients and had been written between 2004 and 2008. They made up a corpus of 46,061 words, with an average of 112 words per letter. The lexicon included 5,014 different words.

These data were produced by CASOS AutoMap, a software for text mining and semantic network analysis that was developed by Kathleen M. Carley at the Center for Computational Analysis of Social Organizational Systems (CASOS) of Carnegie Mellon University. We chose AutoMap for our analysis because its preprocessing and mining utilities can be applied to texts written in Italian, it is relatively easy to use and the license is free for research purposes [9].

As we were to assess the presence or absence of follow-up instructions in the letters, we decided that our first step would be to develop a simple ontology of follow-up instructions, which would be composed of the different categories of instructions that doctors can give heart failure patients. As a second step, we would link a thesaurus to the ontology, which would include the linguistic expressions that might instantiate the categories of the ontology. As a third step, we would preprocess each text by applying these two knowledge bases; and as a final step, we would mine the preprocessed collection. This way the analysis would not be entirely open. Instead, it would be driven by the knowledge bases we would develop. This choice is usually more effective when you are representing domain-specific knowledge [10]; it also allowed us to prevent the very specific bits of information we were searching for from being lost in the general noise.

To build the ontology of follow-up instructions, we used the *Guidelines on Heart Failure* of the Associazione Nazionale Medici Cardiologi Ospedalieri, the Società Italiana di Cardiologia and the Associazione Nazionale Medici Cardiologi Extra-ospedalieri [11]. From these guidelines we extracted twelve categories of follow-up instructions that doctors can give heart failure patients when patients are discharged: *attività* (activity), *riposo* (rest), *dieta* (diet), *alcol* (alcohol), *fumo* (smoke), *perdita di peso* (weight loss), *controllo del peso* (weight monitoring), *controllo della diuresi* (diuresis monitoring), *altitudine* (altitude), *temperature* (temperature), *umidità* (humidity), *varie* (miscellaneous).

Documents such as the *Guidelines* we referred to, of course, also provide some examples of the linguistic expressions that can instantiate the categories we were working on. We collected these examples and started building a thesaurus with them. Yet, these examples do not completely cover the linguistic usage of documents that are written in the course of daily clinical practice. This usage is in fact more informal and heterogeneous than the standard linguistic usage of scientific communication. When they write a discharge letter, doctors write more for patients than they write for the scientific community; moreover, these texts are produced in the midst of daily clinical practice and cannot be carefully written and edited. Typos are frequent, as well as abbreviations and telegraphic expressions, while verbs may not be conjugated. Consequently, we completed the thesaurus with words and expressions from a training set of 100 letters of our collection, which we extracted to better represent the real linguistic usage of doctors.

The result of this work was a thesaurus of follow-up instructions, which was linked to the ontology of follow-up instructions so that every category was associated to a set of linguistic expressions. The *controllo del peso* category, for example, was associated with this set of expressions: *controllo peso*, *controllare peso*, *monitoraggio peso*, *monitorare peso*, *autovalutazione peso*, *osservazione peso*, *verifica peso*, *verificare peso*, *controllo peso corporeo*, *controllare peso corporeo*, *monitoraggio peso corporeo*, *monitorare peso corporeo*, *autovalutazione peso corporeo*, *osservazione peso corporeo*, *verifica peso corporeo*, *verificare peso corporeo*.

Using the ontology and the thesaurus, we could then examine our collection. The texts were loaded into AutoMap and they were preprocessed. First, we removed symbols, numerals, and the stop words from a delete list we had prepared (words such as *il*, *con*, *del*, and so on). Then, we applied the thesaurus and the ontology, so that AutoMap could index every text of the collection in relation to the twelve categories of follow-up instructions. Finally, the preprocessed texts were mined. AutoMap calculated the frequency of each category and their combinations in the collection. Table 1 shows the most important data for each category.

Category	Instantiations	Positive texts	Percentage of positive texts
attività	1	1	0.24
riposo	22	22	5.33
dieta	24	22	5.33
perdita di peso	39	37	8.96
controllo del peso	12	11	2.66
alcol	3	3	0.73
fumo	7	7	1.69
other categories	0	0	0

Table 1: results of the mining.

Results were compared with a gold standard: all the texts of the collection were manually checked for all categories by one reviewer. Sensitivity of the system was 0.484; specificity was 0.834; positive predictive value was 0.555; and negative predictive value was 0.790.

3. Discussion and conclusions

The examination of false positives and false negatives reveals that sensitivity might quickly rise with few adjustments to the thesaurus. However, this requires natural language processing tools which AutoMap does not include. Consequently, we have now chosen to adopt a new application, NooJ by Max Silberztein, which has been developed for natural language processing [12]. This should allow us to enlarge and refine our knowledge bases. What has emerged, in fact, is that knowledge bases must be adequate for a rich, complex and heterogeneous linguistic usage. Given this condition, and because of the increasing availability of electronic textual data, text mining represents a promising way to evaluate the continuity of information and of care.

Acknowledgements

The research was conducted by the author with Dr. Pietro Barbieri and Dr. Mauro Maistrello of Uboldo Hospital; it was supported by Azienda Ospedaliera di Melegnano.

Notes and References

[1] COHEN, A.M. and HERSH, W.R. A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6 (1), 2005, p. 57-71.

[2] FELDMAN, R. and SANGER, J. *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*. New York : CUP, 2007.

[3] SWANSON, D.R. Complementary structures in disjoint science literatures. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Chicago IL: ACM Press, 1991, p. 280-289.

[4] FISZMAN, M. et al. Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports. Journal of the American Medical Informatics Association, 7 (6), 2000, p. 593-604.

[5] HRIPCSAK, G. et al. Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports. Radiology, 224 (1), 2002, p. 154-163.

[6] FORSTER, A. et al. Validation of a Discharge Summary Term Search Method to Detect Adverse Events. Journal of the American Medical Informatics Association, 12 (2), 2005, p. 200-206.

[7] MELTON, G. and HRIPCSAK, G. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. Journal of the American Medical Informatics Association, 12 (4), 2005, p. 448-457.

[8] MURFF, H. et al. Electronically Screening Discharge Summaries for Adverse Medical Events. Journal of the American Medical Informatics Association, 10 (4), 2003, p. 339-350.

[9] To learn more about AutoMap, visit http://www.casos.cs.cmu.edu.

[10] SPASIC, I. et al. Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics, 6 (3), 2005, p. 239-251.

[11] GAVAZZI, A. et al. Linee Guida sullo Scompenso Cardiaco. Associazione Nazionale Medici Cardiologi Ospedalieri, Società Italiana di Cardiologia e Associazione Nazionale Medici Cardiologi Extra-ospedalieri.

[12] To learn more about NooJ, visit http://www.nooj4nlp.net.

June 2009
Printed on demand
by *"Nuova Cultura"*
www.nuovacultura.it

Book orders: ordini@nuovacultura.it