Automatic extraction of numerical values from unstructured data in EHRs

Elise BIGEARD[a], Vianney JOUHET[b], Fleur MOUGIN[b], Frantz THIESSARD[b],
Natalia GRABAR[a1]
[a] *STL, CNRS UMR 8163, Université Lille 3, France*
[b] *ERIAS, INSERM U897, ISPED, Université Bordeaux, Bordeaux, France*

Abstract. Clinical data recorded in modern EHRs are very rich, although their secondary use research and medical decision may be complicated (*eg*, missing and incorrect data, data spread over several clinical databases, information available only within unstructured narrative documents). We propose to address the issue related to the processing of narrative documents in order to detect and extract numerical values and to associate them with the corresponding concepts (or themes) and units. We propose to use a CRF supervised categorisation for the detection of segments (themes, numerical sequences and units) and a rules-based system for the association of these segments among them in order to build semantically meaningful sequences. The average results obtained are competitive (0.96 precision, 0.78 recall, and 0.86 F-measure) and we plan to use the system with larger clinical data.

Keywords. Natural Language Processing, Text Mining, Software Design, Information Storage and retrieval, France

Introduction

The recent (r)evolution of the medical domain leads to an increasing amount of clinical data stored in digital format, as part of the Electronic Health Records (EHRs). This provides a unique opportunity for the healthcare professionals and researchers to interact with the patient data, using them in the decision-making process [1] and for research purposes [2]. Nevertheless, several challenges remain before such utilization of data becomes a reality. For instance, structured and unstructured data, which co-exist in modern hospitals, are spread over several databases which requires to extract, transform and merge these data together within common data repository [3-4]. Another aspect is that data are duplicated across various patient documents, and it is necessary to deduplicate them at different levels *(e.g.,* documents, events, concepts) before an efficient use of them is possible. Yet another aspect, widely stressed due to the frequent reuse of data, is the quality of clinical data [5-7]. Indeed, some values may be erroneous, missing or contradictory *(e.g.,* size and weight of patients, daily intake of prescribed medication, date of the previous visit to the hospital). Finally, when data are available in unstructured texts only, specific methods and resources must be used to access the necessary elements [8-9]. We propose to address the following problem: the

[1] Corresponding Author: natalia.grabar@univ-lille3.fr

information to be extracted is available in narrative documents only, hence its detection and extraction require text mining methods and resources. More specifically, we focus on extraction of numerical values associated with a given set of concepts *(eg,* size of nodules, size and weight of patients, laboratory results).

The presentation of our work is organized in the following way: we first present the Material used, and the Methods we designed. Then, the Results obtained and their Discussion follow. We conclude with some direction for the future work.

Material and Methods

The material used consists of several sets written in French:

- a set of 82 themes (or concepts) manually collected for which numerical values are to be extracted, such as *size, IMS, nodule, albumin, factor, ASAT, Karnofski, urée, Hg, TNM, Marge*. Each theme is associated with its semantic type defined specifically for this study: *size, IMS, Karnofski* describe patient state; *albumin, ASAT, urea, factor* are laboratory results; *nodule* is related to the tumour size; TNM and *Marge* are associated with the evaluation of tumours. These themes may also be associated with their normalized forms: *factor (facteur)* is in fact *factor V (facteur V), Karnofski* is *Karnofsky index (index de Karnofsky);*
- a set of 117 measure units (called units), generally used in relation to the themes. We can find for instance: *days (jours), %, kPa, kUI/l, m2, µmol/mmol Cr, sec, kgs.* These units can also be associated with their normalized forms;
- the main material is a set of clinical anonymized discharge summaries from electronic health records of a digestive system oncology department. In these documents, the segments that contain themes, numerical values and units were annotated by a medical doctor, such as in these examples (themes with simple, numerical values with double, units with dotted underline):

  Bon appétit mais pas de prise pondérale toujours très probablement en rapport avec un diabète très mal équilibré (hémoglobine glyquée encore supérieure à 10 % : indication à insulinothérapie +++)
  Taille : 173 cm ; Poids : 61 kg ; IMC : 20,38

For the detection and extraction of numerical values, we apply a five-step method:

(1) The *pre-processing* of data is responsible for the tokenizing of the corpus and its POS-tagging with TreeTagger [10] that allows to associate words to linguistic categories such as Noun, Verb, Adjective, etc.;

(2) An *additional annotation of the documents with the semantic resources* is done in order to tag the themes and units already recorded in these resources. Towards this end, a simple string to string comparison is performed. At this step, we also detect all the sequences containing only numbers. A first challenge is to remove from these sequences those that may correspond to phone numbers, various clinical identifiers, and address-related numbers. As a matter of fact, these can be filtered out easily, except for the zip-codes which we prefer to keep at this step (they will be filtered automatically at the next step);

(3) The *supervised machine learning* is responsible for the detection and extraction of numerical values possibly related to the processed themes. This step is done with the Wapiti CRF implementation [11]. Wapiti allows to consider properties associated with a given token (or word) as well as properties related to the

context of this token. We perform an automatic categorization with the following features associated with the processed token and to the neighbours within the three-word window on the right and on the left of the token: form of the word as it occurs in the text, POS-tag and lemma of the word as they are computed by TreeTagger, and length of the word. Four categories are to be detected: *T* theme, *N* numerical value, *U* unit, and *O* out positions (all the rest of the document content dimmed irrelevant). This step is performed according to two sub-steps: learning of the model (done on 50% of data) and test of the model (done on the remaining 50% of data). The previously tagged elements (themes, units, and number sequences) can be modified by the model (begin and end-offset change, removal, addition);

(4) The *association of the extracted elements* (*T* theme, *N* numerical value, *U* unit) among them is then performed in order to build coherent and semantically meaningful sequences. This step is done with a rule-based approach, in which we apply a set of patterns that permit to associate these elements with each other. The most typical pattern is TNU (like in *Taille : 173 cm* (*Size : 173 cms*)), but several other patterns are possible. For instance, additional text can be inserted in any position (*hémoglobine glyquée encore supérieure à 10 %* (*glycated haemoglobin still higher than 10 %*)); the unit can be missing (*IMC : 20,38*); the numerical values can form an interval (*cette image tissulaire mesure 2 à 3 cm* (*this tissue image measures 2 to 3 cms*)); a given element can contain more than one token (*hémoglobine glyquée encore supérieure à 10 %* or *tumeur PT 3 N 1*); the position of elements can be inverted (*21000 GB*); several numerical values and units can be associated with a given theme (*tumeur PT 3 N 1* (*tumor PT 3 N 1*)), etc. The most common and general pattern is *TN*U?*, where the theme T is mandatory, the numerical value N occurs at least once but it may occur several times, the unit U is optional as several themes are measured without units (*eg*, various scores and indexes). This pattern covers the major part of cases;

(5) The last step of the method is the evaluation against the reference data provided by the medical doctor. The evaluation is done according to three evaluation measures: recall or sensitivity that indicates the completeness of the data extracted, precision or specificity that indicates the correctness of the data extracted, and F-measure that is the harmonic mean of precision and recall.

Results

Table 1. Performance of the extraction of numerical values.

	Precision	Recall	F-measure
Theme	0.94	0.68	0.79
Numerical	0.98	0.83	0.90
Unit	0.94	0.63	0.76
Out positions	0.99	1.00	1.00
Average	0.96	0.78	0.86

The proposed system permits to extract 210 themes, 215 numerical values and 106 units. The themes are the core elements, for which at least one numerical value is associated, but for which the units are optional. In Table 1, we indicate the performance on the test set. We can observe that precision is very competitive (0.94-0.99), while recall is less competitive, especially with themes and units. As result, F-measure values obtained are between 0.76 and 1.00. The best values are observed with out *O* positions (non-relevant text), mainly because this corresponds to a large amount of data, within which the elements that are not correctly categorized remain quite few. The average values of the results are high: 0.96 precision, 0.78 recall, and 0.86 F-measure.

Discussion

The proposed method can extract information on numerical values from narrative documents with high performance: 0.96 precision, 0.78 recall, and 0.86 F-measure. The supervised CRF model, that relies on properties inherent to the processed tokens and to their close context (three-word windows), provides a suitable model for the detection of the syntagmatic chains such as sequences composed of themes, numerical values and measure units. As the evaluations were performed on a set of real clinical documents, we expect we can use this system for a routine information extraction from medical records in oncology domain as well as in other domains. This possibility is suitable in several situations, such as comparison with values recorded as structured data, especially when these values are conflicting; extraction of values that are rarely recorded in the structured data (*eg*, number and size of nodules, TNM), which may also help to enrich the structured data; pre-filling of forms when preparing specific (*eg*, in oncology) multidisciplinary meetings and when stating on patient condition; providing data for tumor response follow-up in oncology [12]; supporting the decision-making process and diagnosis (numerical values may complete or be indicative of disorders in situations where the terminology-based coding of medical records is not very efficient).

As we have observed, the semantic tagging done before the automatic categorization step can be adjusted by the model, and the status of sequences can be changed: remaining phone numbers, zip codes etc. can be filtered out. Notice that the semantic tagging step can be omitted and the system can be applied to the texts only annotated with linguistic information (POS-tags and lemmas). The results remain stable. Similarly, the CRF contextual tagging of words permits to detect elements with misspellings that cannot be tagged with the available resources. For instance, in this sequence *lhémoglobine est à 13,9 g/ dl* (*thehemoglobin is 13.9 g/ dl*), *lhémoglobine* is tagged although it is misspelled: the determinant *l* (for *le (the)*) is stuck with the noun *hémoglobine* (*hemoglobin*).

Concerning the labeling of the themes and units, the main difficulty is due to the instability of the reference data. For instance, when tagging a given theme, the medical doctor may consider more or less large sequence: include, or not, adjectives and determinants, according to the contextual semantics and meaning. This situation provides unstable reference data that can become problematic when the automatic categorization does the generalization over such data and creates the model. Nevertheless, the core noun element is still detected and extracted, while its syntactic extensions may be missing. Another difficult situation is related to creation of links between the elements. For instance, the correct detection of intervals remain complicated. In a sequence like *2 lésions de CHC du segment IV B de 23 et 19 mm* (*2 HCC lesions of 23 and 19 mm in the segment IV B*), only the last numerical value and its unit *19 mm* are detected and extracted. More sophisticated rules must be designed to make the correct detection possible. We have also tried to detect these semantic relations with the Wapiti supervised categorization, but the results were less efficient.

Future work

In future, we plan to improve the detection of some themes but we aim mainly the improvement of the relations built between the three elements of interest (themes, numerical values and units). We plan to test the system on a larger set of clinical data. More specifically, we would like to test its efficiency for preparing data and forms for the multidisciplinary meetings (eg, in oncology) and for the decision-making process. Combination of terminology-based indexing with numerical values is another perspective we would like to test: we assume these two types of information may be complementary and help the patient diagnosis or the coding of medical records. We also expect that the models learned are quite general, because they exploit the linguistic regularities of clinical data: we would like to test these models on other themes non studied in our work and on clinical data from other medical specialties.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) and the DGA, under grant Tecsan (ANR-11-TECS-012).

References

[1] J Wyatt. Medical informatics, artefacts or science? Methods Inf Med. 1996 sept;35(3):197–200.

[2] M Wiesenauer, C Johner, R Röhrig. Secondary use of clinical data in healthcare providers - an overview on research, regulatory and ethical requirements. Stud Health Technol Inform. 2012;180:614-618.

[3] R Verma, J Harper. *Life cycle of a data warehousing project in healthcare*. J Healthc Inf Manag. 2001 Summer;**15**(2):107-17.

[4] E Zapletal, N Rodon, N Grabar, P Degoulet, Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform - Medinfo 160(Pt 1) 2010:193-197.

[5] A Boussadi, C Bousquet, B Sabatier, I Colombet, P Degoulet, Specification of business rules for the development of hospital alarm system: application to the pharmaceutical validation. Stud Health Technol Inform 2008:145-50.

[6] K Kerr, T Norris, R Stockdale, Data Quality Information and Decision Making: A Healthcare Case Study. Australasian Conference on Information Systems 2007.

[7] S Mattke, AM Epstein, S Leatherman. *The OECD health care quality indicators project: history and background.* Int J Qual Health Care **18**(1), 2006:1-4.

[8] U Hahn, J Wermter, R Blasczyk, PA Horn. *Text mining: powering the database revolution*. Nature **448**(7150) 2007:130-131.

[9] D Rebholz-Schulmann, H Kirsch, F Cauto. *Facts from Texts - Is Text Mining ready to deliver?* PLoS Biology **3**(2) 2005:e65.

[10] H Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing 1995:44-49.

[11] T Lavergne, O Cappé, F Yvon, Practical Very Large Scale CRFs. Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL) 2010:504-513.

[12] EA Eisenhauer, P Therasse, J Bogaerts, LH Schwartz, D Sargent, R Ford, J Dancey, et al. *New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (version 1.1)*. European Journal of Cancer (Oxford, England: 1990) **45(2):** 228-247.