Extracting Clinical Information from Electronic Medical Records

Abstract. As the adoption of Electronic Medical Records (EMRs) rises in the healthcare institutions, these resources are each day more important because of the clinical data they contain about patients. However, the unstructured textual data in the form of narrative present in those records, makes it hard to extract and structure useful clinical information. This unstructured text limits the potential of the EMRs, because the clinical data these records contain, can be used to perform important operations inside healthcare institutions such as searching, summarization, decision support and statistical analysis, as well as be used to support management decisions or serve for research. These operations can only be done if the clinical data from the narratives is properly extracted and structured. Usually this extraction is made manually by healthcare practitioners, what is not efficient and is error-prone. This research uses Natural Language Processing (NLP) and Information Extraction (IE) techniques in order to develop a pipeline system that can extract clinical information directly from unstructured texts present in Portuguese EMRs, in an automated way, in order to help EMRs to fulfil their potential.

Keywords: Electronic Medical Records • Information Extraction • Machine Translation • Natural Language Processing • Text Mining

1 Introduction

Electronic Medical Records (EMRs) are computerized medical systems that collect, store and display a specific patient clinical information [1]. These records are used "by healthcare practitioners to document, monitor, and manage health care delivery within a care delivery organization (CDO). The data in the EMR is the legal record of what happened to the patient during their encounter at the CDO and is owned by the CDO"

[2]. Many types of clinical information are stored in EMRs, such as x-rays, prescriptions, physician's notes, diagnostic images and other types of medical documentation [3]. EMRs became one of the most important new technologies in healthcare [4].

Hospitals play a central role in the healthcare domain and in any society. These healthcare institutions produce large amounts of digital information, mainly with the broad utilization of EMRs. In United States, a study [5] from 2012 showed that 72% of office-based physicians used an EMR system. In Portugal, statistics from 2014 [6] show that the amount of hospitals using EMRs rose from 42% in 2004 to 83% in 2014.

EMRs usually contain unstructured clinical text in the form of narrative [7] written by the healthcare practitioners, concerning the patients. However, the amount of unstructured textual data that is contained in the EMR presents a barrier to realizing the potential of EMRs [8]. This free-text form used by healthcare practitioners is advantageous to "demonstrate concepts and events, but is difficult for searching, summarization, decision support or statistical analysis" [9].

Healthcare practitioners extract clinical information from the unstructured text of EMRs "by employing of domain experts to manually curate such narratives"[8]. This practice is not efficient, is error-prone and consumes human resources that could be used for other tasks [10]. The desirable scenario is to be able to extract clinical information from the unstructured texts in an automated and reliable way. The proposed system in this research aims to provide a solution to extract clinical information from the clinical narratives of Portuguese EMRs in an automated and structured way, in order to facilitate the day-to-day activities of healthcare practitioners.

2 Information Extraction of Electronic Medical Records

Natural Language Processing (NLP) is a "theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [11]. Information Extraction (IE), considered a sub-field of NLP, is the task of retrieving certain types of information from unstructured natural language texts by processing them in an automatic way [12]. IE differs from Information Retrieval (IR), since the goal of IR is retrieving a subset of documents relevant to a query and the goal of IE is to extract information from the documents themselves [13]. NLP, expressed by the means of IE, can then be used to extract clinical information from the unstructured clinical narrative texts present in EMRs.

3 Importance and motivation of extracting clinical information

The narrative parts present in the EMRs have many purposes, such as give a description about the patient clinical situation, be used for management decisions, support research and serve legal issues. The clinical data present in these narratives, if extracted and structured, can be used to perform operations like search, summarize or statistical analysis too. The major part of the time, a manual extraction of clinical information is done, typically by health professionals [14]. However, when the extraction, standardization

and structuring of clinical data is performed manually by the healthcare practitioners, text ambiguity and personal differences can lead to inconsistencies [15]. Replacing the narratives by structured data would also not be ideal, since significant information is lost because of limiting the expressive power of narratives [16].

The motivation of this research is to show that is possible to extract clinical information in an automated way from EMRs written in the Portuguese language, with precision and not losing the expressive power of the narrative. The clinical information that the authors propose to extract from EMRs is concerned only with the patients' diseases, symptoms and clinical procedures.

This helps reducing the amount of time and resources used by the hospital for manual analysis of the EMRs, by automating the process of extracting clinical information, using a NLP system and IE techniques. To add to that, by extracting clinical information in an automated way, operations like searching, summarizing and doing statistical analyses of the clinical information extracted, can be done faster and efficiently by CDOs too, since once the data is extracted it can easily be structured in a database and worked on as clinical information at ease.

4 Review of literature and positioning of this work

There are already several NLP systems capable of extracting clinical information from EMRs, with the major part of them only working for the English language. A review of some of these systems is depicted in this study [9]. A review of the literature in the NLP domain revealed that there are not many studies and systems that focus in extracting clinical information from EMRs written in the Portuguese language.

A NLP system capable of extracting clinical terms from discharge records written in European Portuguese is MedInX, developed in the Institute of Electronics and Telematics Engineering of Aveiro in 2011[10]. Despite of having a good performance, this is a proprietary system and it was only tested concerning the extraction of clinical terms related with hypertension. There are also studies made for the Brazilian variant of the Portuguese language, concerning IE from clinical data.

A study conducted by the Faculty of Medicine, University of São Paulo, in 2007, proposed a pipeline system capable of extracting clinical terms from clinical reports, by coupling Machine Translation (MT) and a NLP system together [17]. However, despite having a good performance, this study only aimed to extract twenty-two different types of clinical terms and was limited to chest x-rays reports. To add to that, this research is from 2007 and since then the MT and NLP systems were improved. Nonetheless, this study showed that is indeed possible to achieve success, by coupling MT and NLP together to extract clinical terms correctly.

In this research, the authors pretend to use a similar approach to the problem by using Machine Translation(MT) first, in order to translate the clinical texts from Portuguese to English and only after perform the clinical information extraction from EMRs using NLP and IE techniques. This approach is justified by the fact of almost all the dictionaries and ontologies being much more mature for the English language, allied to the predominance of the English language in the biomedical field. To add to that, the

major part of MT and NLP systems are built and optimized to work with the English language and are much more improved than ten years ago. The authors also pretend to use only open-source software during the whole process and analyze the findings, in order to build a solution immediately available to everyone.

5 Data description

The hospital made available 34295 EMRs in an Excel file. The EMRs were collected with permission and previously de-identified by the hospital itself. Of these EMRs, 11358 are from ambulatory care, while the others 22937 are from inpatient care. The EMRs from ambulatory care are from different specialties of the hospital, such as gastroenterology, hematology, immunohemotherapy, infectology, nephrology, neurology, medical oncology, pediatrics, pediatrics hematology, pediatrics infectology, pulmonology, rheumatology, pain unit, urology and oncology. The three specialties more represented by the EMRs can be seen in Table 1.

Table 1. Most represented specialties

Specialty	Number of EMRs
Medical Oncology	3199
Pain Unit	2058
Hematology	1810

Each EMR is composed by different fields, such as a sequence number, number of clinical episode, specialty, specialty code, diagnosis code, diagnosis description, date and a clinical narrative text. Table 2 presents the top 5 of the most frequent diagnoses. Table 3 presents the obtained statistical results regarding clinical narratives.

Table 2. Diagnosis count

Diagnosis	Occurrences
Tumors(neoplasms)	3972
Blood and Hematopoietic Organs diseases	2188
Osteoarticular System diseases	2023
Infectious and Parasitic diseases	1494
Nervous System diseases	455

Table 3. Statistical analysis of EMRs' clinical narratives

Criteria / Type of Care	Ambulatory	Inpatient
Total number of words	465862	3627576
Mean number of words per narrative	41	158
Most used clinical word	Transfusion	Pain

6 Methodology applied

In order to extract clinical information from the Portuguese EMRs, a whole system based in open-source modules is coupled together. Firstly, the authors need a translator able to translate the EMRs' clinical narratives from Portuguese to the English language, with a reasonable performance. Secondly, the authors need an open-source NLP system responsible by applying IE techniques and perform the clinical information extraction from the EMRs' narratives. To explain the methodology applied, a top-down approach is used. First, a high-level overview of the whole system, followed then by a more refined explanation of each sub-system. A high-level depiction of the pipeline system the authors aimed to build in this research, is shown in Fig. 1.

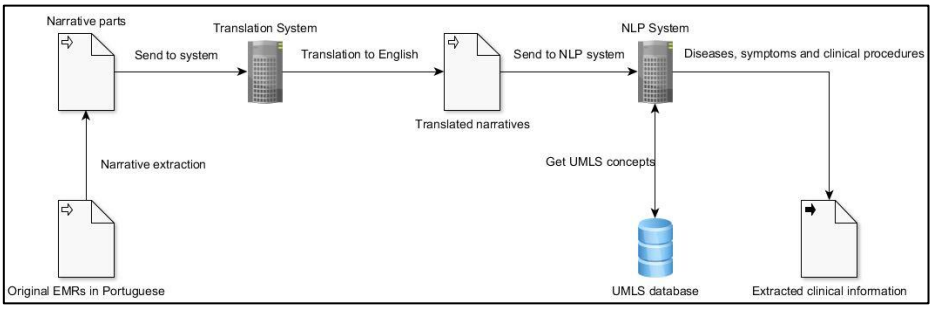

Fig. 1. High-level view of the whole pipeline system

As can be seen in Fig. 1, the whole process initiates by extracting the narratives from the Excel files. After that, the authors sent the narratives to the open-source translation system, in order to translate each one of them to the English language. After all the narratives being translated, the authors fed the NLP system with each one of them, in order to extract diseases, symptoms and clinical procedures. The NLP system uses a database filled with clinical terms and concepts from the Unified Medical Language System (UMLS) in order to be able to identify and extract the clinical entities found in the narratives. UMLS is a repository of biomedical vocabularies developed by the US National Library of Medicine, containing more than 2.2 million concepts and 8.2 million concept names, some of them in different languages than English [18]. Finally, the NLP system extracts diseases, symptoms and clinical procedures, from the EMRs' narratives, outputting a file with all of the extractions concerning those domains.

7 Machine Translation and Natural Language Processing

Machine Translation is considered a sub-field of computational linguistics that consists in using software to translate text or speech from one language to a different one. The authors are still studying which open-source translator fits best this research needs and ambitions.

From the translators already considered, an open-source translator called OpenNMT [19], developed in Harvard University in 2017 and with major source contributions by

a proprietary translator, seems the strongest possibility. OpenNMT uses neural machine translation, a recently proposed approach to MT, that unlike statistical and rule-based MT, aims to build a single neural network that can be tuned and trained to achieve maximum translation performance [20]. Given the available EMRs database to train the system with, OpenNMT seems a solid solution to perform the translation. The MT process is a critical part of our research, since clinical data can be lost in the translation process. We are working on this process to obtain the best performance possible in the translation from Portuguese to English language.

A research was already conducted to verify which open-source NLP system should be used. From several options, an open-source NLP system developed in the Mayo Clinic College of Medicine in Rochester called cTAKES [8], is our final decision to use in this research. This system is currently maintained by the Apache Software Foundation. This system was already used with success to identify the patients smoking status from clinical texts [21], apply summarization [22], confirm cases of hepatic decompensation in radiology reports [23] and extract clinical information concerning Crohn's disease and ulcerative colitis from EMRs [7].

8 Results

This section presents the results of some initial experimental tests performed, using part of the EMRs and processing each one individually. No metric-based evaluations were made yet. For these tests, the translator system used was OpenNMT and it was not pretrained at all. The NLP system used was cTAKES. The authors only wished to verify the behavior and performance of the systems coupled together, without making any kind of tuning and configurations yet. An example of a translated clinical narrative already processed by cTAKES is shown in Fig. 2. Fig. 2 has some handmade annotations in order to facilitate the explanation of the figure. In the right side of Fig.2, one can see the clinical narrative translated to English that is processed by cTAKES while the left side presents the analysis results shown by the system.

Considering Fig. 2 and concerning clinical entities, by default cTAKES is able to identify several clinical mentions in the text, as can be seen in the annotation 1) in Fig. 2. In the annotation 2) it can be seen that is possible to select a group of mentions and iterate by each one of them, as is being done in annotation 3) right below. By selecting each one of the mentions, its position appears immediately highlighted in the clinical narrative, as shown by the symptom annotation 4) of "fever", in Fig. 2.

This simple and straightforward user interface can be really useful to verify exactly which kind of clinical information is being extracted from the narrative. The cTAKES system can also output the processing results in different formats, such as XMI, XML, HTML, plain text or directly to a database. This allows a simple structuration of the data automatically extracted, preparing this data to be easily used as clinical information, what opens a wide range of possibilities that can hugely benefit CDOs and their healthcare practitioners in their day-to-day activities.

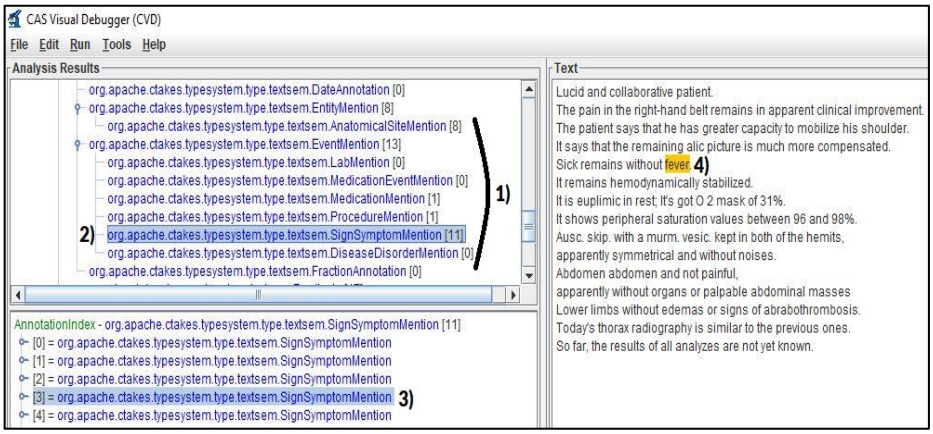

Fig. 2. - Example of cTAKES processing results

9 Conclusions

The authors were already capable of doing some experimental tests of the pipeline system aiming to build with open-source software and real EMRs from the hospital. Those tests already extracted clinical information from the EMRs narratives with some degree of success, depending of the EMR narrative being tested. A really valuable factor in this research is having thousands of real EMRs, from a hospital itself, to work with in order to tune the systems used.

By translating the narrative parts of EMRs given by the Portuguese hospital, not all terms and concepts were correctly translated to English. Not an unexpected finding since the MT system was not trained yet. This is an issue to improve, by training the translating system and adding some rules to it as needed. However, grounded on the performed tests, the authors conclude that the major part of each narrative was well translated. Ensuring a good translation from Portuguese to English language is a crucial step to use a NLP system and clinical knowledge base already well tuned and configured to English language and consequently guarantee the success of this research.
Improving the MT process and its performance is then our immediate objective in the near future, in order to be able to build a pipeline system able to extract diseases, symptoms and clinical procedures from the EMRs narratives in an automated way.

References

1. A. Boonstra and M. Broekhuis, “Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions,” *BMC Health Serv. Res.*, vol. 10, no. 1, p. 231, Dec. 2010.
2. D. Garets and M. Davis, “Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference A HIMSS Analytics TM White Paper Executive Summary,” 2006.
3. D. B. Meinert and D. Peterson, “Perceived importance of EMR functions and physician characteristics,” *J. Syst. Inf. Technol.*, vol. 11, no. 1, pp. 57–70, 2009.

4. E. C. Murphy, F. L. Ferris, W. R. O'Donnell, and W. R. O'Donnell, "An electronic medical records system for clinical research and the EMR EDC interface.," *Invest. Ophthalmol. Vis. Sci.*, vol. 48, no. 10, pp. 4383–9, Oct. 2007.
5. C.-J. Hsiao and E. Hing, "Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2012.," *NCHS Data Brief*, pp. 1–8, 2012.
6. Instituto Nacional de Estatistica, "Statistics Portugal," 2014. [Online]. Available: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE. [Accessed: 03-Feb-2018].
7. A. N. Ananthakrishnan *et al.*, "Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing," *Inflamm. Bowel Dis.*, vol. 19, no. 7, pp. 1411–1420, Jun. 2013.
8. G. K. Savova *et al.*, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 5, pp. 507–513, Sep. 2010.
9. S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research," *IMIA Yearb. Med. Informatics Methods Inf Med*, vol. 47, no. 1, pp. 128–44, 2008.
10. L. da S. Ferreira, "Medical Information Extraction in European Portuguese," p. 262, 2011.
11. E. D.Liddy, "Natural Language Processing," *Encycl. Libr. Inf. Sci.*, 2001.
12. D. C. Wimalasuriya and D. Dejing Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *J. Inf. Sci.*, vol. 36, no. 3, pp. 306–323, Jun. 2010.
13. "GATE.ac.uk - ie/index.html." [Online]. Available: https://gate.ac.uk/ie/. [Accessed: 03-Feb-2018].
14. A. A. Thomas *et al.*, "Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results," *World J. Urol.*, vol. 32, no. 1, pp. 99–103, Feb. 2014.
15. H. Suominen, *-Machine Learning and Clinical Text. Supporting Health Information flow*, no. 125. 2009.
16. C. Lovis, R. H. Baud, and P. Planche, "Power of expression in the electronic patient record: Structured data or narrative text?," *International Journal of Medical Informatics*, vol. 58–59. Elsevier, pp. 101–110, 01-Sep-2000.
17. C. Castilla, "Instrumento de Investigação Clínico-Epidemiológica em Cardiologia Fundamentado no Processamento de Linguagem Natural," p. 112, 2007.
18. R. Kleinsorge, C. Tilley, and J. Willis, "Unified Medical Language System (UMLS)," *Encycl. Libr. Inf. Sci.*, pp. 369–378, 2002.
19. G. Klein, Y. Kim, Y. Deng, J. Crego, J. Senellart, and A. M. Rush, "OpenNMT: Open-source Toolkit for Neural Machine Translation," 2017.
20. D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Sep. 2014.
21. G. K. Savova, P. V Ogren, P. H. Duffy, J. D. Buntrock, and C. G. Chute, "Mayo Clinic NLP System for Patient Smoking Status Identification," *J. Am. Med. Informatics Assoc.*, vol. 15, no. 1, pp. 25–28, 2008.
22. S. Sohn and G. K. Savova, "Mayo clinic smoking status classification system: extensions and improvements.," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2009, pp. 619–23, 2009.
23. V. Garla *et al.*, "The Yale cTAKES extensions for document classification: Architecture and application," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 5, pp. 614–620, Sep. 2011.