Extracting Clinical Knowledge from Electronic Medical Records

Manuel Lamy, Ruben Pereira, Joao C. Ferreira, Fernando Melo and Iria Velez

***Abstract* — As the adoption of Electronic Medical Records (EMRs) rises in the healthcare institutions, these resources' importance increases because of the clinical information they contain about patients. However, the unstructured information in the form of clinical narratives present in those records, makes it hard to extract and structure useful clinical knowledge. This unstructured information limits the potential of the EMRs, because the clinical information these records contain can be used to perform important tasks 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 tasks can only be done if the unstructured clinical information from the narratives is properly extracted, structured and transformed in clinical knowledge. Usually, this extraction is made manually by healthcare practitioners, which 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 knowledge from unstructured clinical information present in Portuguese EMRs, in an automated way, in order to help EMRs to fulfil their potential.**

***Keywords* — Information Extraction, Knowledge Extraction, Natural Language Processing, Text Mining**

I. Introduction

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 Electronic Medical Records (EMRs). 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 healthcare 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]. In United States, a study from 2012 [5] showed that 72% of office-based physicians used an EMR system.

In Europe, a survey validated by the European Commission to 1800 European hospitals, shows that the usage and deployment of eHealth applications in these healthcare institutions, such as Electronic Medical Records systems, has increased over the period of 2010-2013 [6] . In Portugal, statistics from 2014 [7] show that the amount of hospitals using EMRs rose from 42% in 2004 to 83% in 2014.

Despite their usage, EMRs usually contain unstructured clinical information in the form of narrative [8] written by the healthcare practitioners, concerning the patients. However, the amount of unstructured clinical information that is contained in the EMR presents a barrier to realizing the potential of EMRs [9]. 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" [10].

Healthcare institutions extract structured clinical information and clinical knowledge from the EMRs' clinical narratives "by employing of domain experts to manually curate such narratives" [9]. This practice is not efficient, is error-prone and consumes human resources that could be used for other tasks [11]. The desirable scenario is to be able to extract clinical knowledge from the unstructured clinical information present in EMRs in an automated and reliable way. This allows healthcare institutions to possess the clinical knowledge as fast and reliably possible, wasting the least amount of resources to obtain it. At the same time, the healthcare institutions can act and plan in a faster and more sustained style, based on the clinical knowledge obtained.

This research proposal aims to provide a system to extract clinical knowledge from the unstructured clinical information present in patients' EMRs. The EMRs are written in Portuguese language and were made available by a real Portuguese hospital. The knowledge extraction from the EMRs is performed in an automated and structured way, using Text Mining (TM) techniques, such as Natural Language Processing (NLP) and Information Extraction (IE), both subfields of TM.

The focus of this system is to output clinical knowledge in the form of relations between the different clinical specialties of the hospital and the occurrences of clinical entities in each one of those specialties. As an example, with this system the authors aim to find in an automated way, based solely on unstructured information from EMRs, which disease is more frequent in a given clinical specialty or which medications are more prescribed to a given diagnosis.

Manuscript received March 2018; revised April, 2018.

Manuel Lamy is a Master's student from Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon (e-mail: manuel_lamy@iscte-iul.pt).

Ruben Pereira and Joao C. Ferreira are with Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal (e-mail: ruben.filipe.pereira@iscte-iul.pt and joao.carlos.ferreira@iscte.pt)

Fernando Melo and Iria Velez are with Hospital Garcia da Horta (e-mail: ({fmelo,iria.velez}@hgo.min-saude.pt).

II. Data, Information and Knowledge in a clinical context

Before moving on, it's important to distinguish these three different concepts and their respective hierarchy, since they are frequently present in this research and are usually responsible for many misconceptions. Data consists in a collection of facts and statistics concerning an object or originated by an event. Information consists in processed data. This processing has the objective of increasing the usefulness of the data [12]. Finally, knowledge represents an understanding of certain information.

Based on these definitions and in the context of this article, clinical data of a patient EMR is all the raw data, such as the clinical narrative written by a healthcare practitioner originated in the occurrence of an event, such as a medical appointment between the patient and the healthcare practitioner. Still in this context, clinical information consists, for example, in the clinical terms found and extracted from the clinical data, such as medication or diseases. Finally, clinical knowledge consists in an understanding of that clinical information extracted such as the establishment of clinical relations between the patient diagnosis and the clinical terms found in his EMR.

An example of clinical knowledge could be the discovery of which medications are more prescribed in a given clinical specialty, based on the clinical information extracted, in this case all the medications that were extracted from the EMRs' clinical data, written in the form of narratives.

"In the hierarchy of data, information and knowledge, computations with elaborate algorithms play a major role in the initial processing of data to information, but computations with good reference databases become more important in the following processing to compile knowledge." [13]. Now that these three concepts are clarified, it's possible to have a better understanding of the following sections of this work.

III. Literature Review

This section aims to give an overview of what has already been made in the field of clinical knowledge extraction and to understand the positioning of this work. There are already several case studies that were capable of extracting clinical knowledge from unstructured information present in EMRs.

A research conducted by the Faculty of Medicine, University of São Paulo, in 2007, proposed a pipeline system capable of extracting clinical knowledge from clinical reports, by coupling Machine Translation (MT) and a NLP system together [14]. However, this research was limited to chest x-rays reports only. To add to that, this study is from 2007 and since then the MT and NLP systems were improved. Nonetheless, this research showed that is indeed possible to achieve success, by coupling MT and NLP together to extract clinical knowledge.

A different research conducted in Nashville, by the Tennessee Valley Healthcare System, in 2011, showed a solution capable of extracting clinical knowledge from EMRs, that allowed an automatic identification of postoperative complications within EMRs, using NLP techniques to process the unstructured information [15].

Still in 2011, the Mayo Clinic, the Children's Hospital Boston and the Harvard Medical School worked together in a solution that allowed the discover of relations between prescribed drugs and the side effects, just from the EMRs' clinical narratives [16]. EMRs were solely from psychiatry and psychology patients and the system was able to extract side effect and causative drug pairs with a good performance, using a NLP system in conjunction with Machine Learning (ML) techniques and pattern matching rules.

Extracted clinical knowledge from EMRs can also serve for classification systems, as shown in 2013 by the Massachusetts General Hospital, Harvard Medical School and the Harvard School of Public Health. These institutions developed a system together, that was capable of extracting clinical knowledge from EMRs, in order to successfully classify in an automated way the respective patients as having Crohn's disease and ulcerative colitis, based solely in the unstructured information of EMRs [8].

Still in 2013 and in the domain of classification systems, the National Taiwan University and the King's College London coupled together to develop a system able to identify smoking status in EMRs of patients with mental disorders [17].

In 2016, the Mayo Clinic proposed a system capable of extracting clinical knowledge of unstructured clinical notes, that allowed the automatic identification of the presence or not of Peripheral Arterial Disease (PAD) in the respective patients' EMRs, using NLP and IE [18].

A research conducted by the Columbia University Medical Center in 2017 proposed a solution capable of early recognition of Multiple Sclerosis (MC) by applying NLP techniques [19]. This early identification, before the official recognition by the healthcare providers, can potentially reduce the time to diagnosis.

More researches were made in the clinical knowledge extraction area, but only the most recent and relevant ones, concerning this research, were enunciated. Despite having reasonable performances, all of these studies focus on very specific domains. Our goal in this work is to build a system capable of extracting clinical knowledge in a broader spectrum, by obtaining clinical knowledge from EMRs that belong to different clinical specialties of a hospital. Therefore, the authors aim to relate the different clinical specialties with the occurrences of different clinical entities in those same specialties, such as diseases, symptoms, medications, procedures and anatomical regions.

IV. Electronic medical records

One big issue in clinical data external access is privacy. Our approach for privacy is based on the fact that health information can be used for research, but names can't be released. To overcome this problem, the authors developed tools for Patient Controlled Encryption (PCE) as an approach to handle privacy issues of patients. To provide external data access, every personal information of the patients (e.g. name, address, e-mail) and doctors is encrypted and an identifier is created to each doctor and each patient. Correlation of the patient identifier to patient information or doctor identifier to doctor information is possible only based on the knowledge of a private key.

PCE uses standards encryption approach based on: 1) a RSA algorithm to handle key transfer process based on a digital signature [20]; 2) an encryption algorithm, for

example the AES [21], which takes the generated key by RSA algorithm and performs encryption for that file field or XML extracted; 3) a decryption algorithm that takes encrypted information and produces the decrypted file. This procedure allows to export EMRs without personal information available to an Excel file. The authors have access to 5255 EMRs. The EMRs were collected with permission by the hospital itself and are all from ambulatory care. The EMRs are from different specialties of the hospital, such as gastroenterology, hematology, nephrology, medical oncology, pediatrics, pediatrics hematology, pulmonology, rheumatology, urology and oncology. The three clinical specialties more represented by the EMRs can be seen in Table I.

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 II presents the top five of the most frequent patients' diagnoses. Plus, the authors also present general statistic results regarding the clinical narratives in Table III. About 215540 words were processed and in average each narrative has 41 words.

TABLE I
MOST REPRESENTED CLINICAL SPECIALTIES

Specialty	Number of EMRs
Medical Oncology	3150
Rheumatology	619
Pneumology	529

TABLE II
DIAGNOSIS COUNT

Diagnosis	Occurrences
Tumors(neoplasms)	3748
Rheumatoid arthritis	421
Digestive system disease	251
Blood disease	182
Digestive system disease	165

TABLE III
STATISTICAL ANALYSIS OF EMRS' CLINICAL NARRATIVES

Criteria / Type of care	Ambulatory
Total number of words	215540
Mean number of words per narrative	41

V. SYSTEM DESCRIPTION

In order to extract clinical knowledge from the Portuguese EMRs, a pipeline system of different modules was built. A high-level overview of the whole system is depicted in Fig. 1. To begin with, the hospital made available all the 5255 original EMRs in an Excel file, as explained in the previous section.

The first step was the pre-processing of the data. Healthcare practitioners typically use many clinical abbreviations and acronyms, what presents a big challenge to the processing and translation of the original text. In order to solve that problem, one of the pre-processing activities was writing all the abbreviations and acronyms present in the EMRs in their full form. The other pre-processing activity was the correction of orthographic errors. All of this pre-processing was done directly in Excel using macros and regular expressions.

After the pre-processing, a translation of the EMRs from Portuguese to English language is necessary, since the NLP system used in this work is built for the English language. Since the translation domain is out of the scope of this research, the authors used one of the best available translators as a black-box, the Google Translate. With a good pre-processing of the EMRs, this translator is able to translate with a good performance, with the result in English language losing very little expressivity and information when compared with the original Portuguese text.

In order to translate the EMRs, the authors extracted all of EMRs from the Excel file and translated them using the Google Translation API [21], in order to translate the 5255 EMRs available. All of this data manipulation and calls to the Google Translation API were made using Python programming language.

Having the translated EMRs ready and saved in files, they are ready to be sent to the NLP system, in order to extract all structured clinical information possible, such as diseases, medications, symptoms, signs, anatomical regions and clinical procedures. The NLP system used, called cTAKES [9], can go to a directory and process all of the files inside, creating an output file by each input file, with all the structured clinical information extracted for the respective EMR.

In order to be able to identify and extract the clinical entities found in the narratives, the NLP system uses a database filled with clinical terms and concepts from the Unified Medical Language System (UMLS). 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 [22].

With all the information extracted and saved in files, we can then extract clinical knowledge. In order to do that, the authors persisted all the clinical information in a logic way in a SQLite database, allowing to relate all domains at ease.

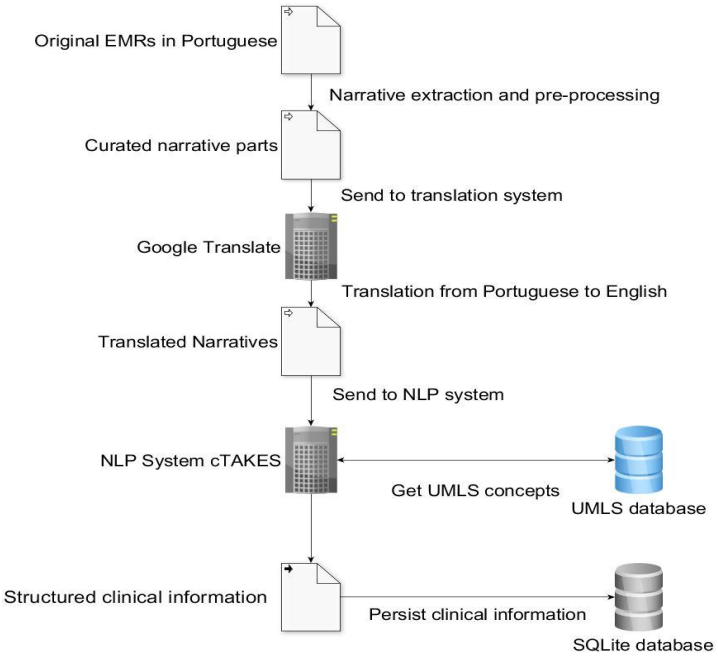

Fig. 1. High-level view of the whole system

VI. Clinical information extraction

In order to extract structured clinical information from the EMRs, the authors had to choose a clinical NLP system to use in this research. An investigation was 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 [9], was the final decision to use in this research. This system is currently maintained by the Apache Software Foundation and was already used with success to identify patients' smoking status from clinical texts [23], apply summarization [24], confirm cases of hepatic decompensation in radiology reports [25] and extract clinical information concerning Crohn's disease and ulcerative colitis from EMRs [8].

The cTAKES is an open-source NLP system implemented in Java and consists in "a modular system of pipelined components combining rule-based and machine learning techniques aiming at information extraction from the clinical narrative" [9]. This system is then composed by different components that are involved in the processing of clinical narratives. Each component contributes with a specific operation made to the text being processed. The components present in cTAKES can be seen in Fig. 2. This system can process files individually or a directory full of files at once.

When presented with a text file, cTAKES starts by splitting the text in sentences, with the Sentence Boundary Detector component. After that, it splits each sentence in tokens and normalizes each token (word) to its most common base form, by removing prefixes and suffixes. This is done by the Tokenizer and Normalizer components respectively. The following component, the POS Tagger, tags each token of the sentence with a part of speech correspondent to that token, identifying if each token is a noun, verb, adverb, adjective, article or conjunction. The Shallow Parser component takes all the tagged tokens and tries to link them together in higher logical units, like noun or verb groups.

Finally, the component Named Entity Recognizer is used. This component uses a dictionary look-up algorithm in order to discover clinical information within the sentence. A customized dictionary with clinical terms and their

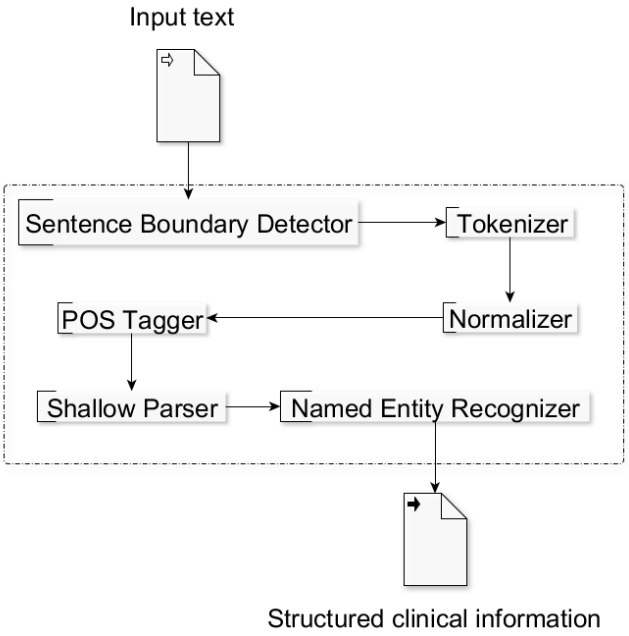

Fig. 2. cTAKES NLP components and flow

relationships can be configured within cTAKES, in order to find the clinical terms in the text. In this research, the authors used the SNOMED-CT dictionary, that contains three types of components: concepts, descriptions and relationships [26]. This dictionary is one of the biggest sources of clinical knowledge available, with over than 300.000 active clinical terms in the English language, being a crucial tool to identify the clinical information in text [27]. There is still no official version of this dictionary to the Portuguese language, what reinforces the authors' decision to translate EMRs to the English language first.

The clinical terms found can be diseases, medications, signs, symptoms, anatomical regions and procedures. In order to find the clinical terms, this component takes all nouns and noun groups found in the text being processed and searches for them in the dictionary. The Named Entity Recognizer component can also detect if a clinical term is negated or has a specific status associated with it. Having all these components applied, the structured clinical

TABLE IV
Most extracted clinical entities by each clinical specialty

Most Identified	Medical Oncology	Number	%	Rheumatology	Number	%	Pneumology	Number	%
Diseases by specialty	*Neoplasm*	*2974*	*73*	*Rheumatoid Arthritis*	*422*	*27.5*	*Asthma*	*47*	*37.9*
	Gastroesophageal reflux disease	96	*2.3*	*Spondylitis*	*169*	*11*	*Respiratory tract infections*	*32*	*25.8*
	Neutropenia	*89*	*2.2*	*Spondylarthritis*	*50*	*3.3*	*Pneumonia*	*20*	*16.1*
Medications by specialty	*Bevacizumab*	*193*	*4.5*	*Infliximab*	*261*	*6.4*	*Carboplatin*	*134*	*13.7*
	Capecitabine	*160*	*3.9*	*Tocilizumab*	*232*	*5.7*	*Vinorelbine*	*70*	*7.1*
	Metoclopramide	*132*	*3.3*	*Methotrexate*	*154*	*3.8*	*Erlotinib*	*57*	*5.8*
Signs/Symptoms by specialty	*Pain*	*354*	*17.7*	*Pain*	*196*	*9.9*	*Chest Pain*	*49*	*11.8*
	Nausea	*156*	*7.8*	*Arthralgia*	*194*	*9.8*	*Tremor*	*35*	*8.5*
	Poor venous access	*103*	*5.3*	*Joint swelling*	*185*	*9.4*	*Severe asthma*	*31*	*7.5*
Anatomical Regions by specialty	*Skin*	*210*	*15.9*	*Joints*	*195*	*18.3*	*Oral cavity*	*34*	*17.5*
	Breast	*126*	*9.6*	*Vertebral column*	*65*	*8.8*	*Respiratory system*	*29*	*14.9*
	Oral Cavity	*111*	*8.4*	*Hand*	*34*	*4.6*	*Veins*	*14*	*7.2*
Clinical Procedures by specialty	*Administration Procedure*	*415*	*33.4*	*Administration procedure*	*301*	*21.7*	*Administration procedure*	*64*	*44.1*
	Analysis of substances	*131*	*10.6*	*Weighing patient*	*244*	*17.6*	*Chemotherapy cycle*	*20*	*13.8*
	Chemotherapy cycle	*93*	*7.5*	*Joint examination*	*195*	*14*	*Analysis of substances*	*12*	*8.3*

TABLE V
MOST IDENTIFIED SIGNS/SYMPTOMS AND MEDICATIONS BY DIAGNOSIS

Most Identified	Digestive system disease	Number	%	Neoplasm	Number	%	Rheumatoid Arthritis	Number	%
Signs/Symptoms by diagnosis	*Digestion problems*	*111*	*36*	*Pain*	*412*	*10.7*	*Pain*	*126*	*17.1*
	Abdominal pain	*42*	*13.6*	*Nausea*	*294*	*7.6*	*Arthralgia*	*109*	*14.9*
	Colic	*35*	*11.4*	*Tremor*	*102*	*2.6*	*Joint swelling*	*108*	*14.7*
Medications by diagnosis	*Ranitidine*	*52*	*19.1*	*Bevacizumab*	*203*	*4.9*	*Tocilizumab*	*231*	*8.6*
	Infliximab	*48*	*17.7*	*Carboplatin*	*196*	*4.8*	*Infliximab*	*108*	*4*
	Azathioprine	*16*	*5.9*	*Capecitabine*	*160*	*3.9*	*Prednisolone*	*91*	*3.3*

information is finally extracted by cTAKES and can be presented in many different file formats.

The authors chose to output the structured clinical information in files with a XMI (XML Metadata Interchange) format. The authors chose this format since it's small in size and easy to read and process the information. The processing of the XMI file and persistence in database of the clinical information extracted, is made with Python programming language too, using the SQLite library. In the next section, the knowledge extraction results are depicted.

VII. CLINICAL KNOWLEDGE EXTRACTION

After having all the extracted clinical information structured and persisted in a database, it's possible to extract clinical knowledge, using SQL and Python to query and manipulate the data.

Since the EMRs belong to ten different clinical specialties and a lot of diagnosis are present, the knowledge extraction is only exhibited concerning the three most represented specialties and diagnosis. However, the results could be easily obtained to the other specialties, since they are persisted in the database too.

All the extracted information for the different clinical specialties presented in the EMRs is identified in Table IV. Table IV presents the most identified medications, diseases, signs/symptoms, anatomical regions and clinical procedures by specialty. For example, one can see that in medical oncology the most identified disease is "Neoplasm" and in rheumatology is "Rheumatoid Arthritis" followed by "Spondylitis", what makes sense in the respective contexts. This is interesting knowledge that the authors intend to explore even more in the future, by integrating with physician's information and be able to extract even more clinical knowledge, like which medication was more prescribed by each physician, for example.

In table V, one can see the most identified signs/symptoms and medications, respectively, but this time for the three most represented diagnosis. One can then conclude that useful knowledge can be extracted and even used to characterize and discover patterns in the different domains of the hospital, such as clinical specialties and diagnosis.

VIII. SYSTEM EVALUATION

The evaluation of the system was performed based on standard metrics calculated manually for 400 of the 5255 EMRs. These standard metrics are precision, recall and F-measure. They are frequently used in the evaluation of IE systems [28]. The authors used these metrics to calculate the performance of the pipeline system in extracting clinical information in the Portuguese EMRs.

Precision can be calculated as defined in (1), as the ratio between the correctly identified terms and the total identified terms. This metric measures the number of correctly identified terms as a percentage of the total identified terms. Recall is calculated as defined in (2), as the ratio between the correctly identified terms and the total of terms that should have been correctly identified. Hence, this metric despises the wrongly identified terms.

In (3) one can see how the F-measure is calculated, by combining precision and recall, with β being the weight between precision and recall. In this research, the standard calculation of F-measure is used (also known as F_1 score), as can be seen in (4), by using a β set to 0.5. This value of β means that precision and recall are equally important.

The authors asked for help of healthcare practitioners from the hospital, who manually annotated the clinical entities present in 400 EMRs in order to establish a gold standard for this evaluation. The evaluation of the pipeline system built was made having in account the translation process, using Google Translator, as well as the information extraction process, using the NLP system cTAKES. The evaluation was made for all types of clinical entities in conjunction. It showed that our system has a precision of 0.75, a recall of 0.61 and a F_1 score of 0.67.

These results show that the system is viable to extract reliable clinical knowledge from Portuguese EMRs. The results obtained are not surprising, since Google Translator is one of the best translators available, which in conjunction with a good pre-processing of the data results in almost no loss of information in the process of translation. To add to that, the cTAKES system used in this research already has one of the greatest state-of-the-art results for the English language, with a precision of 0.8, recall of 0.65 and F_1 score of 0.72 [9].

The results obtained in this research are a little below of the cTAKES' results, since information is always lost in the process of translation, even if minimal. Some eventual errors in the pre-processing of the data can also explain the decrease of performance too.

$$Precision = \frac{Correctly\ identified\ terms}{Total\ identified\ terms} \quad (1)$$

$$Recall = \frac{Correctly\ identified\ terms}{Total\ correctly\ identified\ terms\ possible} \quad (2)$$

$$F - measure = \frac{(\beta^2 + 1)Precision * Recall}{(\beta^2 Recall) + Precision} \quad (3)$$

$$F_1\ score = \frac{Precision * Recall}{0.5 * (Precision + Recall)} \quad (4)$$

IX. Conclusions

This research shows that the pipeline system built by the authors, based in a translator and a NLP system, is capable of extracting useful and reliable clinical knowledge from EMRs written in Portuguese language, from a real hospital, what can be really important to support the said hospital in his tasks. It also shows an automated way of extracting the said clinical knowledge, without wasting human resources to manually review the EMRs.

One limitation of this research is the translation of the EMRs from Portuguese to the English language. Even if not much, performance is lost in this step, because not even all words or expressions get correctly translated. However, even knowing that the translation is out of the scope of this research, the authors consider that the translation occurred with good performance and not much information was lost in this step. The careful pre-processing made to the EMRs before translation was an important step to guarantee a good translation performance.

The authors pretend, in the near future, to be able to extract more clinical knowledge that allows the establishment of even more patterns and relations than the ones done in this research. The hospital will also make available soon more 25000 EMRs in order to keep conducting this research, great part of them from inpatient care. This way the authors pretend to compare the differences in terms of clinical knowledge between the two types of patient care: ambulatory and inpatient care.

Finally, the authors pretend to extend this research to the healthcare practitioners, by associating the clinical knowledge extracted from EMRs with who wrote them. This way it will be possible to verify, for example, which medication is more prescribed or which procedure is more recommended by a given healthcare practitioner, and even relate that with specific periods of the year.

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 D. Mike, "Electronic Medical Records vs . Electronic Health Records : Yes , There Is a Difference By Dave Garets and Mike Davis Updated January 26 , 2006 HIMSS Analytics , LLC 230 E . Ohio St ., Suite 600 Chicago , IL 60611-3270 EMR vs . EHR : Definitions The marke," *Heal. (San Fr.*, pp. 1–14, 2006.

[3] D. Meinert and D. Peterson, "Anticipated Use of EMR Functions and Physician Characteristics," *IGI Glob.*, vol. 4, no. June, pp. 1–16, 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] E. Commission and J. R. C.-I. for P. T. Studies, "European Hospital Survey: Benchmarking Deployment of eHealth Services (2012-2013)," 2014.

[7] Instituto Nacional de Estatistica, "Statistics Portugal," 2014. [Online]. Available: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE. [Accessed: 03-Feb-2018].

[8] 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.

[9] 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.

[10] 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.

[11] L. da S. Ferreira, "Medical Information Extraction in European Portuguese," p. 262, 2011.

[12] R. L. Ackoff, "From data to wisdom," *J. Appl. Syst. Anal.*, vol. 16, no. 1, pp. 3–9, 1989.

[13] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: Back to metabolism in KEGG," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D199–D205, Jan. 2014.

[14] Zerbinatti L., "Extração de Conhecimento de Laudos de Radiologia Torácica Utilizando Técnicas de Processamento Estatístico de Linguagem Natural Extração de Conhecimento de Laudos de Radiologia Torácica Utilizando Técnicas de Processamento Estatístico de Linguagem Natural", Polytechnic School of the University of São Paulo, 2010.

[15] H. J. Murff *et al.*, "Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing," *JAMA*, vol. 306, no. 8, pp. 848–855, Aug. 2011.

[16] S. Sohn, J. P. A. Kocher, C. G. Chute, and G. K. Savova, "Drug side effect extraction from clinical narratives of psychiatry and psychology patients," *J. Am. Med. Informatics Assoc.*, vol. 18, no. SUPPL. 1, pp. 144–149, Dec. 2011.

[17] C. Y. Wu *et al.*, "Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register," *PLoS One*, vol. 8, no. 9, p. e74262, Sep. 2013.

[18] N. Afzal, S. Sohn, S. Abram, H. Liu, I. J. Kullo, and A. M. Arruda-Olson, "Identifying Peripheral Arterial Disease Cases Using Natural Language Processing of Clinical Notes.," *... IEEE-EMBS Int. Conf. Biomed. Heal. Informatics. IEEE-EMBS Int. Conf. Biomed. Heal. Informatics*, vol. 2016, pp. 126–131, Feb. 2016.

[19] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, p. 24, Dec. 2017.

[20] Cgi, "Public Key Encryption and Digital Signature : How do they work ?," *Reproduction*, 2004.

[21] R. Schaden, "AES - Advanced Encryption Standard," 2000.

[22] R. Kleinsorge, C. Tilley, and J. Willis, "Unified Medical Language System (UMLS)," *Encycl. Libr. Inf. Sci.*, pp. 369–378, 2002.

[23] 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.

[24] 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.

[25] 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.

[26] "Google Cloud Translation API Documentation | Translation API | Google Cloud Platform." [Online]. Available: https://cloud.google.com/translate/docs/. [Accessed: 04-Mar-2018].

[27] M. J. Silva, T. Chaves, and B. Simões, "An ontology-based approach for SNOMED CT translation," in *CEUR Workshop Proceedings*, 2015, vol. 1515.

[28] D. Maynard, W. Peters, and Y. Li, "Metrics for evaluation of ontology-based information extraction," *Int. World Wide Web Conf.*, vol. 179, pp. 1–8, 2006.