ISCTE IUL

University Institute of Lisbon

Department of Information Science and Technology

Extracting Clinical Knowledge from Electronic Medical Records

Manuel Maria Vilela Pestana de Moura Lamy

A Thesis presented in partial fulfillment of the Requirements for the Degree of

Master in Telecommunications and Computer Engineering

Supervisor
Prof. Dr. Rúben Filipe de Sousa Pereira, Assistant Professor
ISCTE-IUL

Co-Supervisor
Prof. Dr. João Carlos Amaro Ferreira, Assistant Professor
ISCTE-IUL

September 2018

Acknowledgments

I would like to thank my supervisor Professor Rúben Pereira and co-supervisor Professor João Carlos Ferreira for their guidance during the whole process of this Master Thesis.

I would also like to thank the hospital and their staff for collaborating with me. Their help, availability and collaboration were essential to this research.

Finally, I would also like to thank my family and friends for their amazing support during the whole time.

Resumo

Com a adopção cada vez maior das instituições de saúde face aos Processos Clínicos Electrónicos (PCE), estes documentos ganham cada vez mais importância em contexto clínico, devido a toda a informação clínica que contêm relativamente aos pacientes. No entanto, a informação não estruturada na forma de narrativas clínicas presente nestes documentos electrónicos, faz com que seja difícil extrair e estruturar deles conhecimento clínico. Esta informação não estruturada limita o potencial dos PCE, uma vez que essa mesma informação, caso seja extraída e estruturada devidamente, pode servir para que as instituições de saúde possam efectuar actividades importantes com maior eficiência e sucesso, como por exemplo actividades de pesquisa, sumarização, apoio à decisão, análises estatísticas, suporte a decisões de gestão e de investigação. Este tipo de actividades apenas podem ser feitas com sucesso caso a informação clínica não estruturada presente nos PCE seja devidamente extraída, estruturada e processada em conhecimento clínico. Habitualmente, esta extração é realizada manualmente pelos profissionais médicos, o que não é eficiente e é susceptível a erros. Esta dissertação pretende então propôr uma solução para este problema, ao utilizar técnicas de Tradução Automática (TA) da língua portuguesa para a língua inglesa, Processamento de Linguagem Natural (PLN) e Extração de Informação (EI). O objectivo é desenvolver um sistema protótipo de módulos em série que utilize estas técnicas, possibilitando a extração de conhecimento clínico, de uma forma automática, de informação clínica não estruturada presente nos PCE de um hospital português. O principal objetivo é ajudar os PCE a atingirem todo o seu potencial em termos de conhecimento clínico que contêm e consequentemente ajudar o hospital português em questão envolvido nesta dissertação, demonstrando também que este sistema protótipo e esta abordagem podem potencialmente ser aplicados a outros hospitais, mesmo que não sejam de língua portuguesa.

Palavras-chave

Extração de Conhecimento, Extração de Informação, Mineração de Texto, Processamento de Linguagem Natural, Tradução Automática

Abstract

As the adoption of Electronic Medical Records (EMRs) rises in the healthcare institutions, these resources' importance increases due to all clinical information they contain about patients. However, the unstructured information in the form of clinical narratives present in these 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 essential 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 appropriately extracted, structured and processed in clinical knowledge. Usually, this information extraction and structuration in clinical knowledge is performed manually by healthcare practitioners, which is not efficient and is error-prone. This research aims to propose a solution to this problem, by using Machine Translation (MT) from the Portuguese language to the English language, Natural Language Processing (NLP) and Information Extraction (IE) techniques. With the help of these techniques, the goal is to develop a prototype pipeline modular system that can extract clinical knowledge from unstructured clinical information contained in Portuguese EMRs, in an automated way, in order to help EMRs to fulfil their potential and consequently help the Portuguese hospital involved in this research. This research also intends to show that this generic prototype system and approach can potentially be applied to other hospitals, even if they don't use the Portuguese language.

Keywords

Information Extraction, Knowledge Extraction, Machine Translation, Natural Language Processing, Text Mining

CONTENTS

List of tables

List of figures

Acronyms

AES	Advanced Encryption Standard
ASCKE	Automated System for Clinical Knowledge Extraction
BioTeKS	Biological Text Knowledge Services
CDO	Clinical Delivery Organization
CPE	Collection Processing Engine
CREOLE	Collection of Reusable Objects for Language Engineering
cTAKES	clinical Text Analysis and Knowledge Extraction System
CVD	CAS Visual Debugger
EMR	Electronic Medical Record
GDM	Gate Document Manager
HITEx	Health Information Text Extraction
IE	Information Extraction
MC	Multiple Sclerosis
ML	Machine Learning
MMTx	Meta Map Transfer
MT	Machine Translation
NER	Named Entity Recognition
NLP	Natural Language Processing
PAD	Peripheral Arterial Disease
PCE	Patient Controlled Encryption
POS	Part-of-speech
UMLS	Unified Medical Language System

1. 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 the 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 EMRs systems, has increased over the period of 2010-2013 [6] . In Portugal, statistics from 2014 [7] show that the number of hospitals using EMRs rose from 42% in 2004 to 83% in 2014.

1.1 Data, Information and Knowledge in a Clinical Context

Before moving on, it's important to distinguish these three different concepts and their hierarchy, since they are frequently present in this research and are usually responsible for some misconceptions. Data consists of a collection of facts and statistics concerning an object or originated by an event. Information consists of processed data. This processing has the objective of increasing the usefulness of the data [8]. Finally, knowledge represents an understanding of specific information.

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

Another example of clinical knowledge could be the discovery of which medications are more prescribed in a given clinical speciality (e.g. pulmonology), based on the clinical information extracted from the pulmonology speciality EMRs' clinical data, written in the form of a narrative by the healthcare practitioners. The hierarchy of these three concepts is depicted in Figure I.

"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." [9]. Now that these three concepts are clarified, it's possible to have a better understanding of the following chapters of this research.

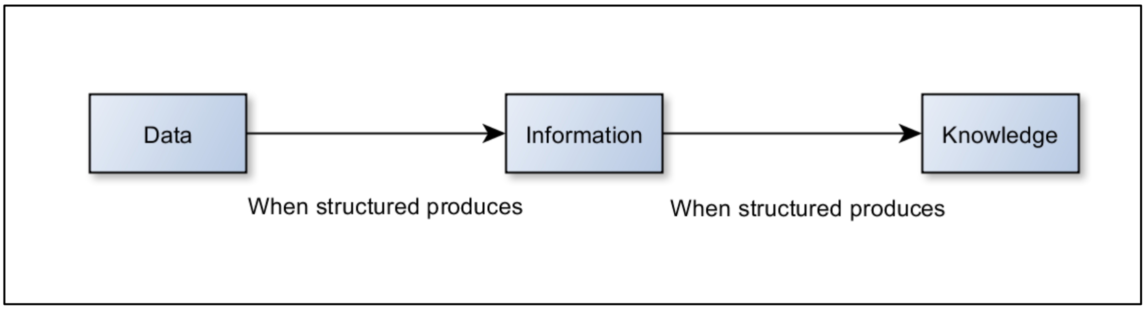


Figure I – Hierarchy of data, information and knowledge

1.2 Motivation

EMRs usually contain unstructured clinical information in the form of narrative [10] written by the healthcare practitioners, concerning the patients. However, the amount of unstructured clinical information that is contained in the EMRs presents a barrier to realize their potential [11]. This free-text form used by healthcare practitioners is advantageous to "demonstrate concepts and events but is difficult for searching, summarization, provide decision support or perform statistical analyses" [12].

Healthcare institutions extract structured clinical information and knowledge from the EMRs' clinical narratives "by employing domain experts to manually curate such narratives" [11]. This practice is not efficient, is error-prone and consumes human resources that could be used for other tasks [13].

The desirable scenario is to be able to extract clinical knowledge from the unstructured clinical information present in EMRs using a system, performing that extraction in an automated, fast and reliable way, as depicted in Figure II. This would allow 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 could act and plan in a faster and more sustained style, based on the faster clinical knowledge obtained.

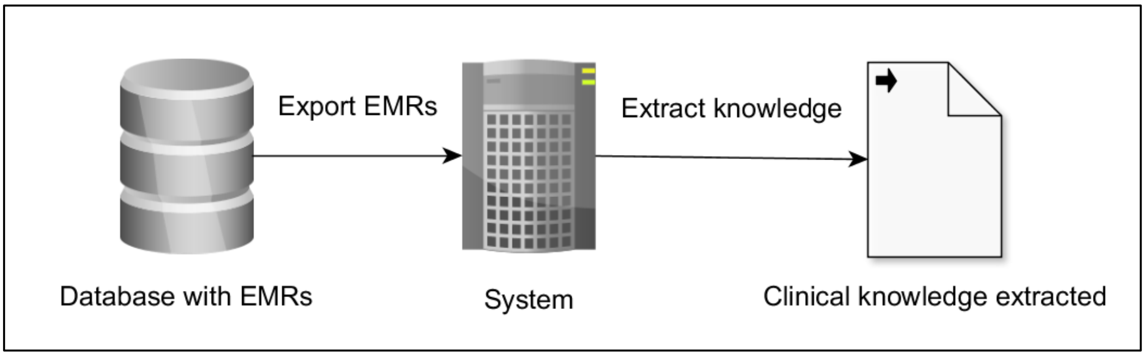


Figure II – Desired scenario in terms of clinical knowledge extraction

1.3 Objectives

This research proposal aims to build a prototype system called ASCKE (Automated System for Clinical Knowledge Extraction) capable of extracting clinical knowledge, in an automated way, from the unstructured clinical information present in patients' EMRs. EMRs written in the Portuguese language were made available by a Portuguese hospital in order to test the system. The knowledge extraction from the EMRs in this research is performed using Machine Translation (MT) and Text Mining (TM) techniques, such as Natural Language Processing (NLP) and Information Extraction (IE), both subfields of TM.

More specifically, concerning the clinical knowledge extraction, the focus of the ASCKE system is to output clinical knowledge in the form of relations between the different clinical specialities of an hospital and the occurrences of clinical entities in each one of those specialities. As an example, ASCKE should be capable of finding, in an automated way and based solely on unstructured information from EMRs, which disease is more frequent in a given clinical speciality or which medications are more prescribed to a given diagnosis, besides several other findings.

For a better understanding, a high-level example of the ASCKE application is depicted in Figure III. As seen in Figure III, concerning the pulmonology clinical speciality, the ASCKE prototype system should be capable of exporting from an hospital database the pulmonology EMRs and extract clinical knowledge from them, such as which medications or symptoms were more identified in the patients' EMRs. This knowledge extraction could be scheduled and the type of knowledge specifically configured for each clinical speciality, depending on the needs.

This research also aims to show that this prototype system and approach could potentially be applied in any other hospital, even if they don't use the Portuguese language, as long a translation with good performance from the original EMRs language to the English language is possible.

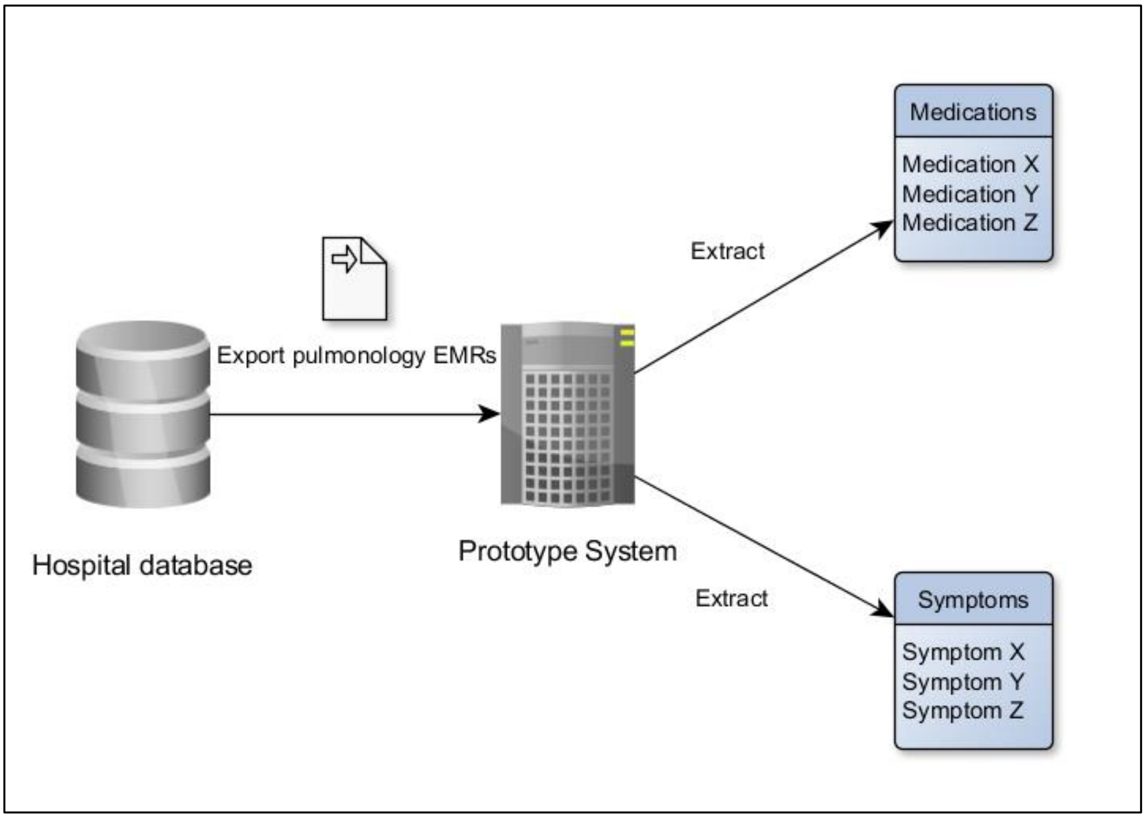


Figure III – Example of ASKCE application concerning clinical knowledge extraction

1.4 Research questions

This research intends to propose an answer to the research questions shown in Table I.

ID	Research Question
RQ1	Is it possible to extract reliable and structured clinical information from unstructured clinical information contained in Portuguese EMRs?
RQ2	Is the coupling of MT, NLP and IE a valid approach to extract clinical information and ultimately extract clinical knowledge with reasonable performance, from different languages than English?
RQ3	Is it possible to successfully extract useful clinical knowledge from the EMRs of an hospital?

Table I – Identified research questions

With the creation of a prototype system that integrates an NLP system in conjunction with MT, both applied to the EMRs of a Portuguese hospital, this research aims to propose answers to these questions.

1.5 Dissertation structure

In Chapter 2, the author reviews state of the art concerning the most similar studies that he could find, concerning this research, in order to justify this research positioning, objectives and motivation. In Chapter 3, the author explains the initial approach to the ASCKE prototype system developed in this research and how he coordinated with the hospital in order to test the system in an appropriate way. In Chapter 4, a description about the tools used in the ASCKE prototype system is given. In Chapter 5, the architecture of the ASCKE prototype system developed in this research is explained in detail. In Chapter 6, the author shows the ASCKE system evaluation results and discusses them. Finally, in Chapter 7, the conclusions about this research are presented and the possible future work is addressed too.

2. State of the Art

This chapter explains what has been made in the past concerning the extraction of clinical knowledge in different healthcare environments. Different existing NLP systems and some of their respective case studies are enunciated too, in order to justify the choice of the NLP system used in this research. In this chapter, it's also possible to understand where this research is positioned in relation to the actual state of the art of this area and justify its objectives and motivations.

There are many existent NLP systems and case studies until the moment this work is being done. However, only the most relevant systems and case studies are addressed, in the biomedical domain, considering the objectives of this research.

2.1 Biomedical NLP

NLP is a research field dedicated to enable computers with the right knowledge for understanding natural language, ultimately to facilitate the different types of natural language interaction between humans and computers [14].

NLP can be applied to natural language expressed in the form of voice or text. In order to be applied, NLP uses knowledge concerning lexicons, syntaxes, the semantics of the language being processed, as well as specific domain knowledge. Typical tasks of NLP are named entity recognition, information retrieval, information extraction and automatic summarization.

NLP is usually applied in different stages, concerning the processing of the text. Firstly, it splits the text into sentences, using punctuation marks or other elements as a splitting reference. After that, each sentence is split in tokens. Each token can correspond to a word or a punctuation mark. Table II shows an example of a sentence split in tokens.

The	patient	has	a	normal	respiratory	effort	.

Table II – Sentence split into tokens

Following that, a stemmer is usually used in order to transform inflectional and derivationally forms of a word to its most common base form. For example, in the quote shown in Table II, the stemming process would replace the word "has" with the word "have". A Part-Of-Speech Tagger (POS Tagger) is also usually used, in order to assign parts of speech to each token, like nouns, verbs, adjectives and punctuation marks. An example of the application of a POS Tagger is shown in Table III.

Article	Noun	Verb	Article	Adjective	Adjective	Noun	Punctuation
The	patient	has	a	normal	respiratory	effort	.

Table III – Tokens tagged with their respective part of speech

The stages used and the order in that they are applied depend entirely of the NLP system being used and his architecture. Nonetheless, these modules, such as the sentence splitter, tokenizer, stemmer and POS Tagger, are present in almost every NLP system.

In the biomedical domain, the utilization of EMRs and other clinical electronic resources is growing fast with the "parallel growth of narrative data in electronic form, along with the needs for improved quality of care and reduced medical errors" [12]. These factors are creating a consequent grow of NLP applied in the biomedical domain, also known as Biomedical NLP. Concerning EMRs, one of the main goals that biomedical NLP aims to achieve is the extraction of the patients' structured clinical information in an automated way from the narrative texts of EMRs. In the next section, an overview of existing biomedical NLP systems is presented, in order to understand which biomedical NLP systems exist and how and where they are being applied.

2.2 Overview of existent biomedical NLP systems

In this section, examples of already existent biomedical NLP systems and some of their case studies are enunciated. The main goal is to show which NLP systems are available and justify the choice of one of them in order to be used in this research.

2.2.1 GATE

This open-source NLP system called GATE (General Architecture for Text Engineering) was created in 1996 and is based in three main modules:

- GDM (Gate Document Manager), based on the TIPSTER document manager.
- CREOLE(Collection of Reusable Objects for Language Engineering), responsible for analyzing the text. This module is responsible for performing common NLP tasks, such as tokenizing, parsing and part-of-speech tagging.

- CGI, a graphical tool that encapsulates the GDM and CREOLE modules.

This system aims to combine already existing language engineering modules in order to extract structured information from unstructured text. More information about this system can be found here [15].

In 2005, a group of researchers from the University of Pittsburgh developed a pipeline-based system to extract structured information from the narrative texts present in surgical pathology reports, using GATE as their NLP system. These reports contained important information such as cancer type, location, pathological stage, values of prognostic attributes, tumour size and weight [16].

2.2.2 HITEx

In 2006, a group of researchers at the Brigham and Women's Hospital and Harvard Medical School developed an open-source NLP system called HITEx(Health Information Text Extraction). This system uses 11 modules from the previously described GATE system and the rest of the modules are developed solely for HITEx.

This group applied this system to extract diagnosis, co-morbidity and smoking status concerning asthma research from textual data contained in discharge summaries and EMRs [17].

2.2.3 MMTx

MMTx (Meta Map Transfer) is an open-source NLP system created in 2001 by the United States National Library of Medicine. This system allows the discovery of clinical terms and concepts from the UMLS(Unified Medical Language System)[18] module Metathesaurus in arbitrary text. Metathesaurus is part of the UMLS ontology and contains plenty of biomedical terms and concepts based in controlled vocabularies and classification systems.

This system processes the text using a series of modules. First of all, the text is parsed into sentences, paragraphs, phrases, lexical elements and tokens. After that, candidate concepts from the UMLS Metathesaurus are evaluated against the parsed content. Finally, the best candidate concepts are mapped in a way to best cover the text. More information about this system can be found in the work of Aronson[19].

In 2005, this NLP system was applied to the extraction of medical problems from the narrative texts of EMRs. The objective was the maintenance of the electronic problem lists associated with the patients, making them more complete and updated [12]. Other applications of this system are the extraction of structured information from surgical pathology reports [20] and the retrieval of cardiac clinical findings in echocardiogram reports [21].

2.2.4 MedLEE

MedLEE is an NLP system created in 1994 with the capability of identifying clinical information in text and mapping that information into a structured model that incorporates the clinical terms [22].

This system was used already to extract structured clinical information from chest x-ray reports written in Brazilian Portuguese, as shown in this research [23]. Since the MedLEE system was developed for the English language only, this study used MT techniques in order to translate the clinical texts first from Brazilian Portuguese language to the English language and only after extract the clinical information with MedLEE.

Other applications of this NLP system are the identification of findings suspicious for breast cancer in mammogram reports [24] and more recently the extraction of signs and symptoms of multiple sclerosis disease from EMRs [25].

2.2.5 BioTeKS

IBM (International Business Machines) developed a system called BioTeKS (Biological Text Knowledge Services), used for “text analysis, mining, and information retrieval in the biomedical domain”[26]. This system was developed in collaboration with the University of Colorado. BioTeKS mechanism relies on understanding the semantic context of the text being analyzed first and only after applying the extraction of information and other NLP tasks such as summarization. This system has already identified clinical terms with success in medical records as can be verified in this research [27].

2.2.6 MedEx

MedEx is another NLP system capable of extracting clinical information from clinical texts. This system was developed initially just based on discharge summaries and was then improved in order to extract information from clinic visit notes too. This system obtained a good performance extracting not only medication information “but also signature information, such as strength, route and frequency” [28]. This system can “map medication text into structured representation using a sequential semantic tagger and a chart parser" [28]. This system was already used to extract medication information [29] and drug-dose information [30] from clinical texts.

2.2.7 cTAKES

An open-source NLP system called cTAKES (clinical Text Analysis and Knowledge Extraction System) was developed in 2010 by the Mayo Clinic College of Medicine in Rochester, Minnesota. This system was developed with the goal of performing "information extraction from electronic medical records' clinical free-text"[11]. This system combines rule-based and machine learning techniques.

The strategy this system uses is based in the modular processing of the data. The cTAKES system organizes itself in different modules, such as: sentence boundary detector; tokenizer; normalizer; part-of-speech (POS) tagger, shallow parser and named entity recognition annotator, including status and negation annotators [11].

This system was already used with success to identify the patients smoking status from clinical texts [31], apply summarization [32], confirm cases of hepatic decompensation in radiology reports [33] and extract clinical information concerning Crohn's disease and ulcerative colitis from EMRs [10].

2.3 Comparison between systems

This research aims to use and configure an existent NLP system in order to extract structured clinical information from EMRs, that will allow the knowledge extraction right after. In order to do that, the author had to choose the system that fits better the objectives, in order to be integrated in the ASCKE system. In Table IV, an overview of the NLP systems referenced in the previous section is shown with some of their characteristics and information.

System	Application Domain	Creation Date	Successful applications	Programming language
MedLEE	Chest x-ray reports, mammogram reports and EMRs	1994	[22][23][24][25]	Prolog
GATE	Surgical pathology reports	1996	[16][34]	Java
MMTx	Surgical pathology reports, EMRs and echocardiogram reports	2001	[20][21][35]	Java
BioTeKS	EMRs	2003	[27]	Java, C++
HITEx	Discharge summaries and EMRs	2006	[17][36][37]	Java
MedEx	Discharge summaries and clinic visit notes	2010	[28][29][30]	Java, Python
cTAKES	Radiology reports and EMRs	2010	[10][31][32][33]	Java

Table IV – Overview of existent NLP systems

Even though there are some more NLP systems available besides the ones referenced above, these ones are the most widely used in the biomedical domain [12].

From the referenced systems, cTAKES, developed by Apache, “aims to provide best-of-breed NLP modules to the community and facilitates the translation of research into practice” [38]. To add to that, this system is open-source, what makes it available to being adapted to specific scenarios. These reasons, allied to the fact that this system performed well in its case studies described above concerning structured clinical information extraction, made the author choose cTAKES as the NLP system to integrate in the ASCKE prototype system built in this research.

2.4 Clinical knowledge extraction case studies

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 chest x-rays reports written in Brazilian Portuguese, by coupling Machine Translation (MT) and an

NLP system together [39]. 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 validates that is indeed possible to achieve success by coupling MT and NLP together in order to extract clinical knowledge from clinical reports successfully.

Later on, research conducted in 2008 by the Partners HealthCare System, Brigham & Women's Hospital, Harvard Medical School and Columbia University, proposed a solution capable of extracting clinical knowledge from the patients' discharge summaries. Firstly, the authors used an NLP system in order to extract diseases and drugs contained in the discharge summaries. Following that, the authors established associations between the extracted diseases and drugs using co-occurrence statistics, obtaining valuable clinical knowledge in an automated way [40].

In 2011, the Mayo Clinic, the Children's Hospital Boston and the Harvard Medical School worked together in a solution that allowed the discovery of relations between prescribed drugs and the side effects, just from the EMRs' clinical narratives [41]. 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 an 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 [10].

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 [34].

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 [42]. Later on, 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 [25]. This early identification, before the official recognition by the healthcare providers, can potentially reduce the time to diagnosis. An overview of all these researches can be seen in Table V.

Country	Date	Data type	NLP system	Uses MT	Research
Brazil	2007	Chest x-ray reports	MedLEE	Yes	[39]
USA	2008	Discharge Summaries	MedLEE	No	[40]
USA	2011	Electronic Medical Records	cTAKES	No	[41]
USA	2013	Electronic Medical Records	cTAKES	No	[10]
Taiwan/UK	2013	Electronic Health Records	GATE	No	[34]
USA	2016	Electronic Medical Records	cTAKES	No	[42]
USA	2017	Electronic Health Records	MedLEE	No	[25]

Table V – Overview of clinical knowledge extraction case studies

More researches were made in the clinical knowledge extraction area, but only the most recent and relevant ones, concerning this research, are enunciated in this chapter. Despite having reasonable performances, all of these studies focus on particular domains and the major part of them are applied in clinical documents written natively in the English language. This research aims to build a system capable of extracting clinical knowledge in a broader spectrum, by obtaining clinical knowledge from EMRs that belong to different clinical specialities and domains of an hospital.

This research also aims to perform that extraction in EMRs written in the Portuguese language. Therefore, the author aims to establish associations between the different clinical specialities and the occurrences of the extracted clinical terms in those same specialities, such as diseases, symptoms, medications, procedures and anatomical regions.

3. ASCKE Requisites and Proposed System Architecture

This chapter's main purpose is to explain and justify the author's first approach to build a prototype system capable of clinical knowledge extraction from clinical documents of any hospital, that he called Automated System for Clinical Knowledge Extraction (ASCKE).

Since this research is done and tested in a partnership with a real Portuguese hospital, the author had to understand which were their needs and objectives in terms of clinical knowledge extraction. This information is important in order to build an adequate prototype system that could fulfill this hospital goals and work for any other hospital too, since almost all the hospitals follow the same architecture and logic in terms of the clinical resources' persistence, management and structure.

Therefore, this chapter starts by explaining the initial hospital scenario that the author was confronted with. Following that, requisites were defined together with the hospital concerning the prototype system ASCKE and they are explained too. Based on the hospital requisites, a use case diagram was created in order to show the prototype system functionalities and the correspondent actors. This diagram is also explained and depicted in this chapter.

Finally, the initial approach envisioned by the author for the ASCKE prototype system, based on the requisites given by the hospital, is explained and justified. The final purpose of this chapter is to support and understand the author's line of thought and decision-making during the beginning of all the process, when confronted with the hospital scenario, in order to build the prototype system ASCKE, capable of extracting clinical knowledge from clinical documents.

3.1 Work scenario

Since this research is done in a partnership with a Portuguese hospital, several meetings were held with the hospital staff in order to understand their needs in terms of clinical knowledge extraction. The hospital promptly expressed their interest in a solution to extract clinical knowledge, in an automated way, from the clinical narratives that are produced everyday in the hospital written by the doctors. A solution like this, as already explained in chapter 1, would help the hospital in tasks such as searching, summarization, decision support, and statistical analysis, as well as be used to support management decisions or serve for research. All of these tasks would be performed faster and sometimes better, by extracting clinical knowledge in an automated way.

As discussed with the hospital, one of the major sources of clinical narratives are the patients' EMRs, so the author and the hospital together concluded that these clinical documents that contain all the patient information, were great resources to work within this research. A more detailed explanation of how the EMRs are created and structured in this hospital is given in the next section.

3.2 Electronic Medical Records

The author had access to 5255 authentic EMRs from the hospital database, exported to an Excel file. As explained in chapter 1, EMRs contain all the patient clinical information that results from a medical appointment with a doctor. The EMRs are created in the hospital database following the activities depicted in Figure IV.

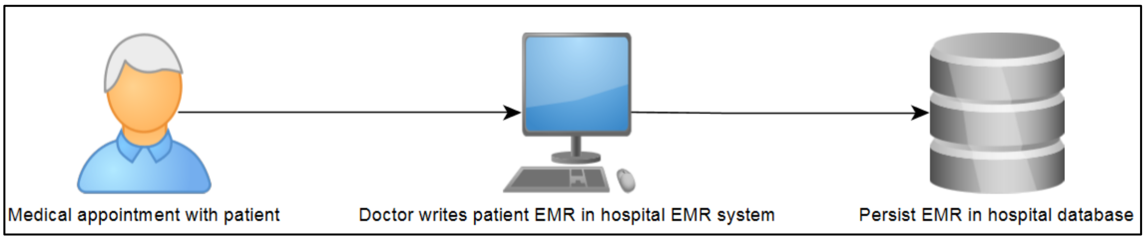
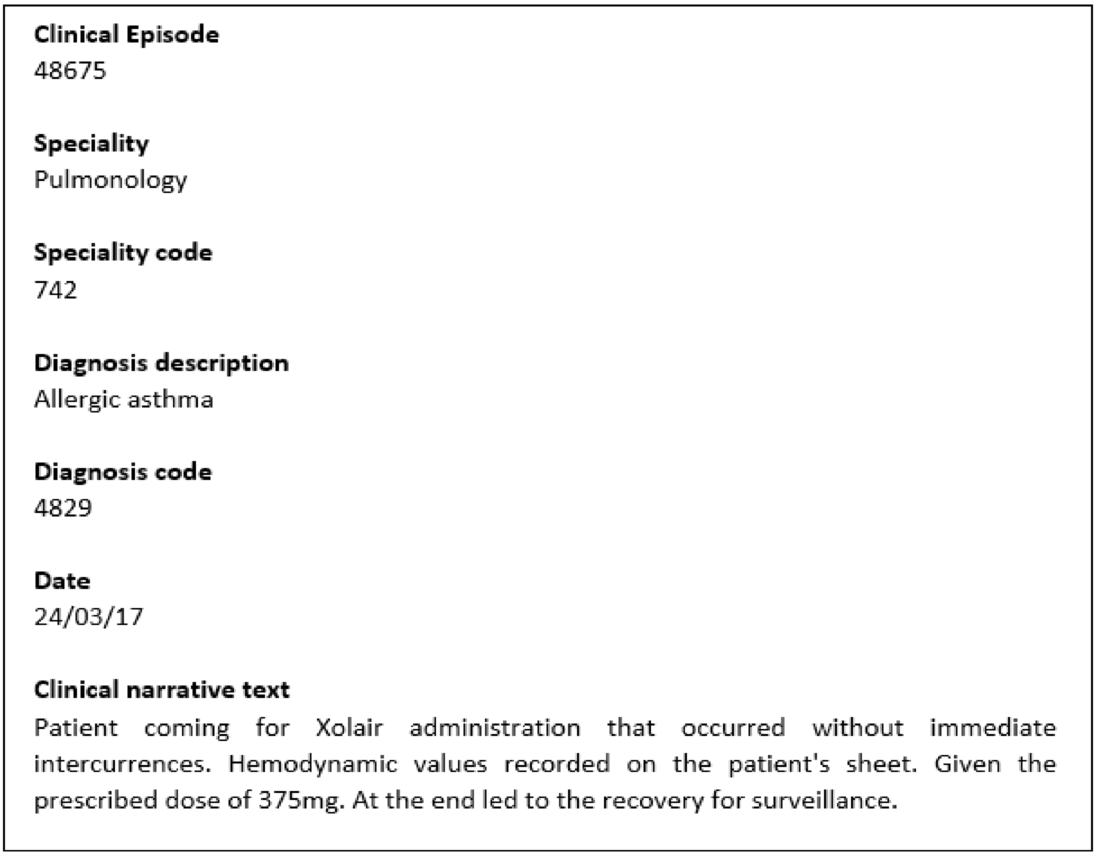

Medical appointment with patient
Doctor writes patient EMR in hospital EMR system
Persist EMR in hospital database

Figure IV – Activities that lead to the EMRs creation and persistence in the hospital database

As shown in Figure IV the whole process starts with a medical appointment between the doctor and the patient. While the medical appointment is happening or at the end of it, the doctor typically writes the EMR concerning the appointment and the patient clinical information, directly in a form that belongs to the hospital dedicated EMR system. Each EMR is composed of different fields, such as a sequence number, number of the clinical episode, specialty, specialty code, diagnosis code, diagnosis description, date and a clinical narrative text containing the patient clinical information. All these fields are filled by the doctor that conducts the medical appointment. An example of an EMR filled by a doctor at the end of a medical appointment is shown in Figure V. Originally this EMR is written in the Portuguese language since it is a Portuguese hospital but for the purpose of this example the EMR is translated to the English language.

Finally, when the medical appointment ends the EMR gets persisted in the hospital database. Once they are persisted, the EMRs can easily be exported from the database, as it was done in an Excel format in this research.

The Excel format was chosen because it’s easy to manipulate the information with it, directly in the file or using a programming language that contains libraries capable of it. In that Excel file exported from the hospital database, each row corresponds to an EMR, with all the information showed in Figure V present in each one of the EMR columns.



Clinical Episode
48675

Speciality
Pulmonology

Speciality code
742

Diagnosis description
Allergic asthma

Diagnosis code
4829

Date
24/03/17

Clinical narrative text
Patient coming for Xolair administration that occurred without immediate intercurrences. Hemodynamic values recorded on the patient's sheet. Given the prescribed dose of 375mg. At the end led to the recovery for surveillance.

Figure V – Example of an EMR filled by a doctor at the end of a medical appointment

Privacy measures concerning patients’ and doctors’ data were taken too, by removing all the patients’ and doctors’ identification from the EMRs. The EMRs obtained from the hospital are all from 2017 and ambulatory care. Ambulatory care refers to all medical services that are performed on an outpatient basis, without the need for admission to an hospital or other facility [44]. These medical services can be a diagnosis, observation, treatment and rehabilitation. The EMRs are from different specialties of the hospital, such as gastroenterology, hematology, nephrology, oncology, pediatrics, pediatric hematology, pulmonology, rheumatology, urology and oncology.

Now that exists a better notion of how EMRs are created and structured in this hospital, the next section explains the system requisites defined by the author and the hospital together, concerning the clinical knowledge extraction from the EMRs.

3.3 Hospital requisites towards ASCKE

The meetings held with the hospital served as well for the author to identify requisites, together with the hospital, in order to build the ASCKE prototype system and obtain results accordingly with those requisites. The author also aimed to build ASCKE in a generic way that could be applied to any other hospital. After all the meetings with the hospital, Table VI shows which system requisites were agreed together, concerning the clinical knowledge extraction from the EMRs. Three requisites were then defined for the ASCKE system. Considering the requisites, the author could test the ASCKE system in this particular hospital and validate if the system worked or not.

[1]	The system should be capable of extracting the most represented clinical specialties and diagnosis in the EMRs.
[2]	The system should be capable of extracting the most identified clinical terms in the EMRs, by each clinical specialty or diagnosis, such as medications, diseases, signs/symptoms, anatomical regions and clinical procedures.
[3]	The system should be capable of extracting those most identified clinical terms in the EMRs and show their incidence by different time periods.

Table VI – Prototype system defined requisites concerning clinical knowledge extraction

3.4 ASCKE functionalities

Considering the requisites defined in the previous section, a use case diagram was created in order to structure the ASCKE system functionalities and understand how users interact with the ASCKE. In Figure VI is depicted ASCKE's use case diagram.

As depicted in Figure VI, the hospital staff can configure the system. As an example, this configuration could be the scheduling of clinical knowledge extraction from a specific clinical specialty, in order to show clinical knowledge results periodically from that clinical speciality. Other configuration could be to define how those results are shown. To add to that and based on the defined ASCKE system requisites, the hospital staff can check different types of clinical knowledge, as depicted in Figure VI.

All of those functionalities concerning clinical knowledge extraction require the ASCKE system to perform an extraction of clinical information first, using an NLP and translation system. Now that ASCKE system functionalities are clear, the next section explains the initial approach to the creation of the ASCKE system.

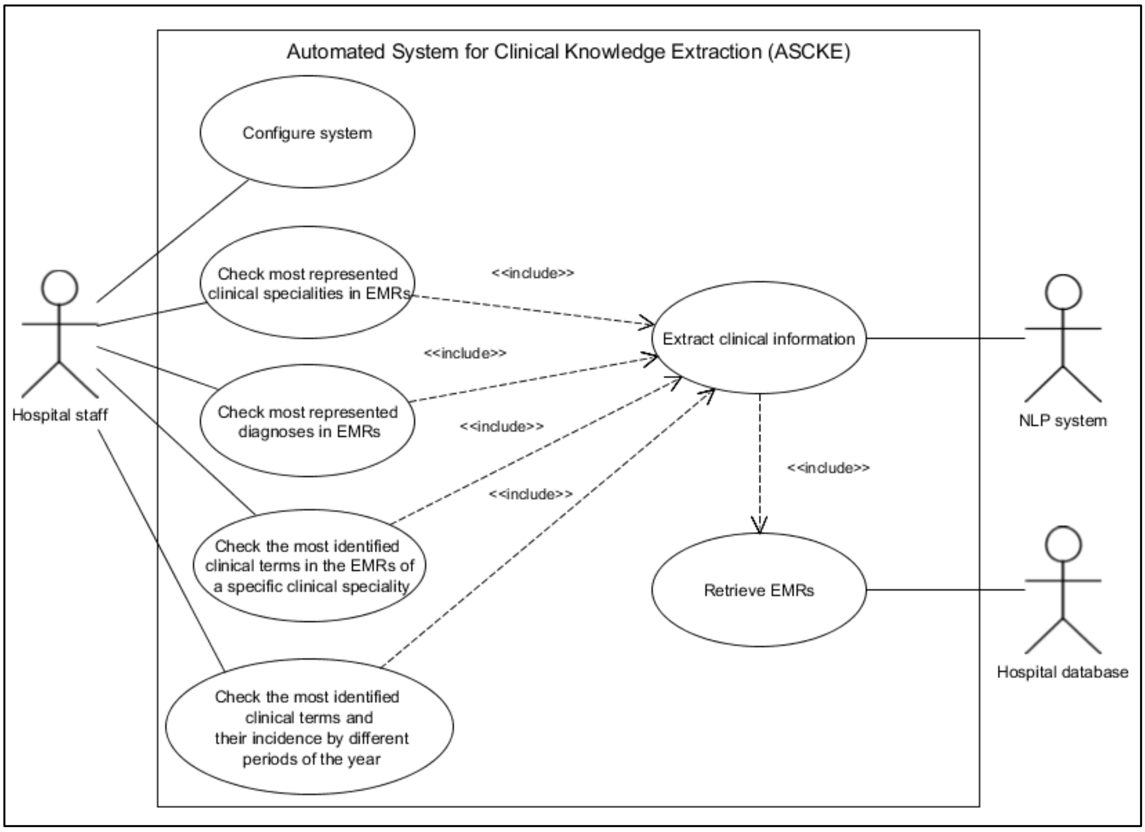

Automated System for Clinical Knowledge Extraction (ASCKE)
Configure system
Check most represented clinical specialities in EMRs
<<include>>
Extract clinical information
<<include>>
Check most represented diagnoses in EMRs
<<include>>
<<include>>
Check the most identified clinical terms in the EMRs of a specific clinical speciality
<<include>>
Retrieve EMRs
Check the most identified clinical terms and their incidence by different periods of the year
Hospital staff
NLP system
Hospital database

Figure VI – ASCKE system use case diagram

3.5 ASCKE initial approach

The author started by defining the architecture of the ASCKE system to be developed. In this research, the author knew that an NLP system had to be used in ASCKE in order to extract the clinical information from the EMRs, with the ultimate goal of extracting clinical knowledge right after. However, the EMRs used in this research are written in the Portuguese language since these documents belong to a Portuguese hospital. At the time this research is being written, there is no clinical NLP system built to work with the Portuguese language, because there are no complete Portuguese medical ontologies available and enough clinical resources in this language. This problem is transversal to many other different languages than Portuguese.

However, many open-source clinical NLP systems exist to work specifically with the English language, having great performances in clinical information extraction tasks, as already discussed in chapter 2. To add to that, the English language has plenty of resources that the major part of other languages doesn’t have, such as really complete clinical vocabularies and medical ontologies that are crucial to perform information extraction in this domain [43]. Considering all these facts, the author decided to perform a translation of the EMRs from the Portuguese language to the English language in ASCKE first, before extracting the clinical information from them, since the translation results in the English language lose very little expressivity and information when compared with the original Portuguese text.

This step enables the utilization of open-source clinical NLP systems that work specifically to the English language, which gives benefits by allowing the utilization of the complete clinical vocabularies and medical ontologies already existent for the English language. Another reason to support this translation step is that the author wants to build ASCKE in order for it to work in any hospital, despite their language. The translation step enables just that, depending if a translation from the original language to the English language with a good performance is possible. Considering all this line of thought, the ASCKE high-level architecture envisioned by the author, in an initial stage of this research, is depicted in Figure VII.

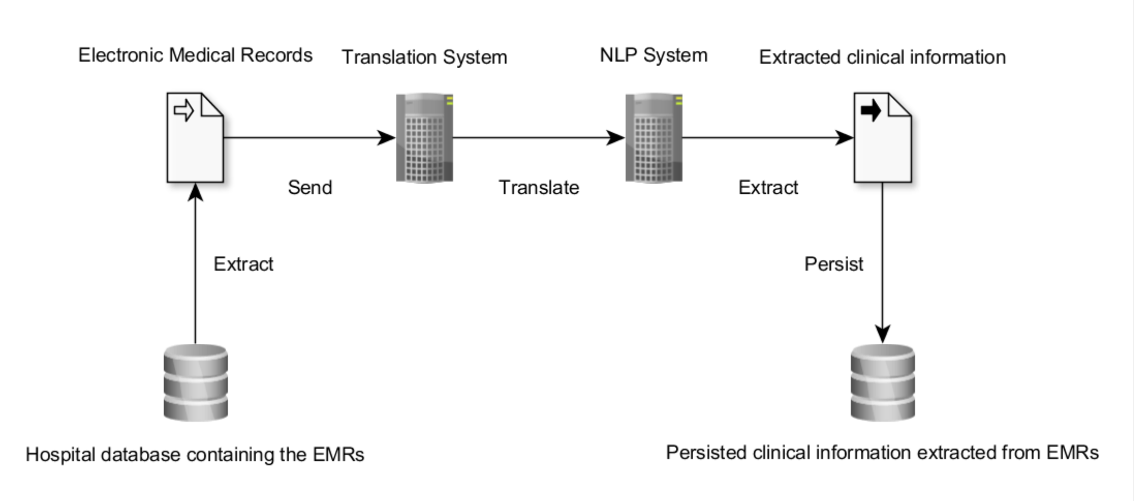

Electronic Medical Records
Translation System
NLP System
Extracted clinical information
Send
Translate
Extract
Extract
Persist
Hospital database containing the EMRs
Persisted clinical information extracted from EMRs

Figure VII – ASCKE initial high-level architecture

As shown in Figure VII, the first task of the ASCKE system should be the extraction of the EMRs from the hospital database. This extraction can be made in any format since different parsers can be integrated in ASCKE, in order to convert the EMRs in an appropriate format to be sent to the MT and NLP system. After that, considering the reasons already explained in this section, the EMRs should be sent to a translation system in order to translate them from the original language to the English language. After the translation is performed, the EMRs should be processed right after by the English language specialized clinical NLP system, in order to extract all the clinical information from the EMRs, in a structured format. Finally, all that clinical information extracted should be persisted in a database, in order to extract clinical knowledge right after. In chapter 5 a more detailed architecture of the ASCKE system is given.

4. Tools used in ASCKE

This chapter describes the tools used in the ASCKE system developed in this research in order to approach the problem. Firstly, an explanation of the used translation system is given. After that, this chapter provides a detailed description of the NLP system chosen to integrate ASCKE in this research. Finally, a description of the clinical base of knowledge used by the NLP system, in order to successfully identify clinical terms in the clinical narratives, is given too.

4.1 Google Translate

A translation of the EMRs from Portuguese to the English language is needed, in order to use a specialized open-source clinical NLP system that is built to work with the English language by default. As explained in section 3.1, another reason for the translation is that the English language has plenty of resources that the major part of other languages doesn't have, such as really complete clinical vocabularies and medical ontologies that are crucial to perform information extraction [43].

The author used one of the best available translators, the Google Translate. With a good pre-processing of the EMRs, this translator is able to achieve a great performance. To add to that, the result in the English language loses very little expressivity and information when compared with the original Portuguese text.

The authors extracted all of EMRs from the Excel file and translated them using the Google Translation API[45], in order to translate the 5255 EMRs available. All the data manipulation and calls to the Google Translation API were performed using Python.

Having all translated EMRs extracted and saved in text files, they were then ready to be sent to the NLP system, in order to extract structured clinical information, such as diseases, medications, symptoms, signs, anatomical regions and clinical procedures. All of this extraction process is explained in more detail in chapter 5.

4.2 cTAKES

The cTAKES tool is an open-source clinical NLP system implemented in Java. The NLP component of cTAKES, used in this work, consists of "a modular system of pipelined components combining rule-based and machine learning techniques aiming at information extraction from the clinical narrative" [11]. As explained in chapter 2, from the several NLP systems studied by the author, cTAKES is the one with the best performance among them all for the English language. To add to that, this system is open-source, what makes it available to being adapted to specific scenarios. These reasons made the author choose this NLP system to be integrated in ASCKE. This chapter aims then to explain how cTAKES works in order to extract clinical information from clinical texts. It starts by explaining his design and architecture, followed by an explanation of the clinical base of knowledge that cTAKES uses, in order to discover and classify clinical terms present in text.

4.2.1 Design and architecture

The cTAKES system is composed of different components that are involved in the processing of clinical narratives. Each component contributes with a specific operation made to the text being processed. A schema of all the components involved in the text processing is shown in Figure VIII:

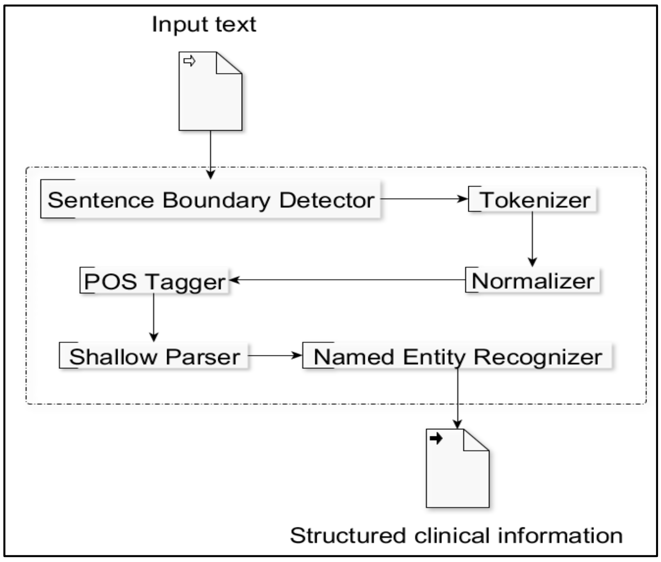


Figure VIII – Schema of cTAKES components involved in the text processing

First of all, the input text (clinical narrative) is exposed to the component Sentence Boundary Detector that splits the narrative text into sentences. After that, the component Tokenizer splits the quote in different tokens (words and punctuation marks). As an example, let’s consider the quote present in Table VII, already split in tokens.

Patient	has	strong	abdominal	pain	but	no	blurred	vision	.

Table VII – Sentence split in tokens in cTAKES

After that, the Normalizer component is applied. This component replaces words to their most common base forms, by removing prefixes and suffixes from them, for example. This operation is also known as stemming. The changes made to the sentence after the application of the Normalizer component can be seen in Table VIII, marked bold.

Patient	**have**	strong	**abdomen**	pain	but	no	**blur**	vision	.

Table VIII – Normalized tokens after stemming operations in cTAKES

After the normalization of the sentence, the POS tagger component is applied. This component tags each token of the sentence with a part of speech correspondent to that token. The result of the application of this component can be seen in Table IX.

Noun	Verb	Adjective	Noun	Noun	Conjunction	Adverb	Noun	Noun	
Patient	**have**	strong	**abdomen**	pain	but	no	**blur**	vision	.

Table IX – Tokens tagged with part of speech in cTAKES

After this stage, the Shallow Parser component is applied. In the context of a sentence, the Shallow Parser takes all the tagged tokens and tries to link them together in higher logical units, like a noun or verb groups, using a medical ontology to do so. Table X shows the application of this component.

Noun	Verb	Noun Group			Conjunction	Noun Group			
Patient	has	strong	abdominal	pain	but	no	blurred	vision	.

Table X – Output by Shallow Parser in cTAKES

Following the Shallow Parser comes the Named Entity Recognition component. This component uses a dictionary look-up algorithm in order to discover clinical information within the sentence. A specific dictionary with clinical terms and their relationships can be configured in cTAKES in order to find the clinical terms in the text. The clinical terms found can be diseases, signs/symptoms, parts of the body, procedures and medications. In order to find the clinical terms, this component takes all entities identified in the text and performs a dictionary look-up, in order to map each named entity to a concept.

The Named Entity Recognition component can also detect if a clinical term is negated or has a specific status associated with it. Using the example output of the Shallow Parser from Table X, the Named Entity Recognition component would output the following final result shown in Table XI.

Clinical term	Classification	Negation status
Strong abdominal pain	Sign/Symptom	Not negated
Blurred vision	Sign/Symptom	Negated
Abdominal	Anatomical region	Not negated

Table XI – Output of the Named Entity Recognizer in cTAKES

As it's possible to observe, in this example cTAKES was able to successfully identify two signs/symptoms and one anatomical region of the patient and identify if they are negated or not. This is how cTAKES internally makes use of his NLP components to process text in order to extract structured clinical information.

4.2.2 cTAKES modes

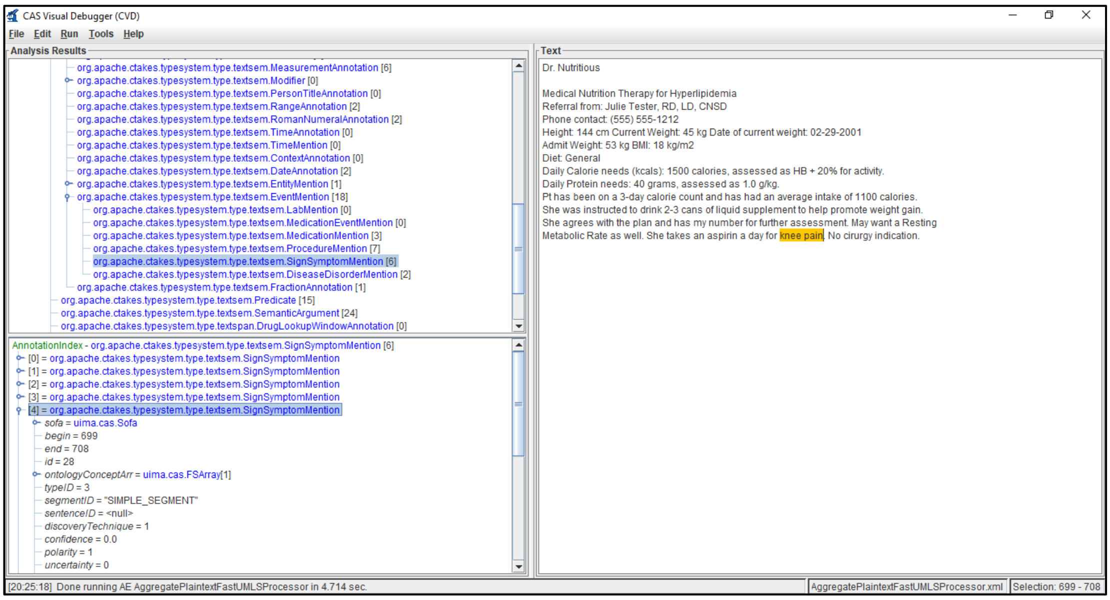
The cTAKES system has two different modes in which it can operate. Each mode gives different options and interfaces to the users. In this section, the two cTAKES modes are explained in more detail.

One of the cTAKES modes is called CAS Visual Debugger (CVD). This mode allows to process a clinical text and immediately see the result of the clinical entities found in the text. This mode can process the text altogether, finding the most basic concepts related to natural language, to the most complex clinical terms and relations between them. It also displays the findings in a friendly and comprehensible user interface.

An example of this mode working can be seen in Figure IX. As can be seen in Figure IX, the right side of the interface is where the input text is supposed to be inserted in order to be processed. The left side shows the results of the NLP processing of the text in a well-defined tree of clinical entities found for the respective text being processed.

It's possible to observe in Figure IX that concerning clinical terms, cTAKES can find laboratory mentions, medications, procedures, symptoms, anatomic regions, diseases and disorders in the text. We can iterate each finding in the lower left screen. Still, in the same Figure IX, symptoms are being checked, and as we iterate through them, they get highlighted in the text itself. As an example, observable in Figure IX, the fourth symptom selected in the list at the left corresponds to the highlighted symptom in the text "knee pain" at the right.

This mode is useful for demonstrating the processing results made by cTAKES in an user interface. However, in this mode, one can only process a piece of text individually. As can be seen in the lower left side of Figure IX, the processing of this little piece of text took 4.714 seconds. The desirable scenario would be to process many EMRs at once and in less time, which is possible to perform with the other mode of cTAKES that is explained right after.



```
org.apache.ctakes.typesystem.type.textsem.MeasurementAnnotation [6]
org.apache.ctakes.typesystem.type.textsem.Modifier [0]
org.apache.ctakes.typesystem.type.textsem.PersonTitleAnnotation [0]
org.apache.ctakes.typesystem.type.textsem.RangeAnnotation [2]
org.apache.ctakes.typesystem.type.textsem.RomanNumeralAnnotation [2]
org.apache.ctakes.typesystem.type.textsem.TimeAnnotation [0]
org.apache.ctakes.typesystem.type.textsem.TimeMention [0]
org.apache.ctakes.typesystem.type.textsem.ContextAnnotation [0]
org.apache.ctakes.typesystem.type.textsem.DateAnnotation [2]
org.apache.ctakes.typesystem.type.textsem.EntityMention [1]
org.apache.ctakes.typesystem.type.textsem.EventMention [18]
    org.apache.ctakes.typesystem.type.textsem.LabMention [0]
    org.apache.ctakes.typesystem.type.textsem.MedicationEventMention [0]
    org.apache.ctakes.typesystem.type.textsem.MedicationMention [3]
    org.apache.ctakes.typesystem.type.textsem.ProcedureMention [7]
    org.apache.ctakes.typesystem.type.textsem.SignSymptomMention [6]
    org.apache.ctakes.typesystem.type.textsem.DiseaseDisorderMention [2]
org.apache.ctakes.typesystem.type.textsem.FractionAnnotation [1]
org.apache.ctakes.typesystem.type.textsem.Predicate [15]
org.apache.ctakes.typesystem.type.textsem.SemanticArgument [24]
org.apache.ctakes.typesystem.type.textspan.DrugLookupWindowAnnotation [0]
```

```
AnnotationIndex - org.apache.ctakes.typesystem.type.textsem.SignSymptomMention [6]
[0] = org.apache.ctakes.typesystem.type.textsem.SignSymptomMention
[1] = org.apache.ctakes.typesystem.type.textsem.SignSymptomMention
[2] = org.apache.ctakes.typesystem.type.textsem.SignSymptomMention
[3] = org.apache.ctakes.typesystem.type.textsem.SignSymptomMention
[4] = org.apache.ctakes.typesystem.type.textsem.SignSymptomMention
    sofa = uima.cas.Sofa
    begin = 699
    end = 708
    id = 28
    ontologyConceptArr = uima.cas.FSArray[1]
    typeID = 3
    segmentID = "SIMPLE_SEGMENT"
    sentenceID = <null>
    discoveryTechnique = 1
    confidence = 0.0
    polarity = 1
    uncertainty = 0
```

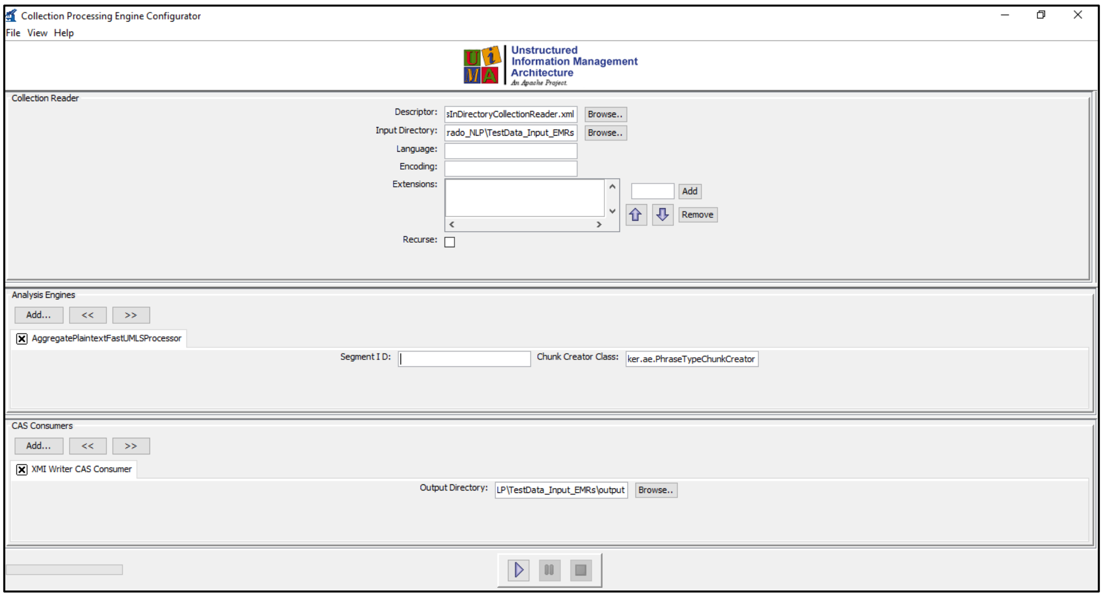
Dr. Nutritious

Medical Nutrition Therapy for Hyperlipidemia
Referral from: Julie Tester, RD, LD, CNSD
Phone contact: (555) 555-1212
Height: 144 cm Current Weight: 45 kg Date of current weight: 02-29-2001
Admit Weight: 53 kg BMI: 18 kg/m2
Diet: General
Daily Calorie needs (kcals): 1500 calories, assessed as HB + 20% for activity.
Daily Protein needs: 40 grams, assessed as 1.0 g/kg.
Pt has been on a 3-day calorie count and has had an average intake of 1100 calories.
She was instructed to drink 2-3 cans of liquid supplement to help promote weight gain.
She agrees with the plan and has my number for further assessment. May want a Resting
Metabolic Rate as well. She takes an aspirin a day for knee pain. No cirurgy indication.

```
[20:25:18]  Done running AE AggregatePlaintextFastUMLSProcessor in 4.714 sec.
AggregatePlaintextFastUMLSProcessor.xml
Selection: 699 - 708
```

Figure IX - cTAKES CVD mode processing an EMR narrative text

The automated processing of many EMRs at once is possible using another mode of cTAKES called Collection Processing Engine (CPE). In Figure X, it's possible to see an example of this mode's user interface.


Collection Processing Engine Configurator
File View Help
Unstructured Information Management Architecture
An Apache Project
Browse..
Add
Remove
Add...
<<
>>
Browse..

Collection Reader

Descriptor: sInDirectoryCollectionReader.xml

Input Directory: rado_NLP\TestData_Input_EMRs

Language:

Encoding:

Extensions:

Recurse: □

Analysis Engines

☒ AggregatePlaintextFastUMLSProcessor

Segment I D:

Chunk Creator Class: ker.ae.PhraseTypeChunkCreator

CAS Consumers

☒ XMI Writer CAS Consumer

Output Directory: LP\TestData_Input_EMRs\output

Figure X – cTAKES CPE mode

This mode can be split into three main modules. Firstly, this mode allows defining an input that can be a full directory or a database, in order to get a lot of files or data at once in order to be processed. Secondly, it allows defining the analysis engine to be used in order to process the text. It's possible to define a simple engine that only splits the sentences in tokens and do some basic POS tagging, or it's possible to define most complex engines, that can extract clinical terms and associations from the text too.

Lastly, it's possible to define a consumer. A consumer consists of a writer that defines the format and content of the files that will be outputted by cTAKES after the processing takes place, with all the output extracted information. The output formats can be XML, XMI (XML Metadata Interchange) or HTML, for example. The consumer chosen in this research is the XML Writer, since the XML format is simple, can be directly persisted in a database and is easy to process the information with. Since a lot of EMRs will have to be processed at the same time in the shortest amount of time possible, this CPE mode of cTAKES is the one that will be used in this research.

4.3 Clinical base of knowledge

Clinical IE and NLP systems are capable of extracting clinical information from clinical resources. In order to do that task, these systems need to build up associations and representations of the clinical terms found in the text. This construction of associations and representations is only possible if these systems have access to a clinical base of knowledge, allowing them to interpret the clinical terms in terms of meaning (e.g. disease, medication) and establish relations between them.

Therefore, it’s crucial to have a clinical base of knowledge where biomedical ontologies, terminologies, lexicons and a controlled vocabulary are present and continuously updated, in order to allow the NLP and IE systems to perform their operations. An ontology is “a representation of entities and their relationships in a particular domain” [46] and it’s one of the most important resources while performing knowledge extraction in any domain. In Figure XI is depicted an example of a little part of a biomedical ontology, showing some clinical terms and the links between them.

A clinical base of knowledge like the one described is used in this research and is explained in more detail in the next section.

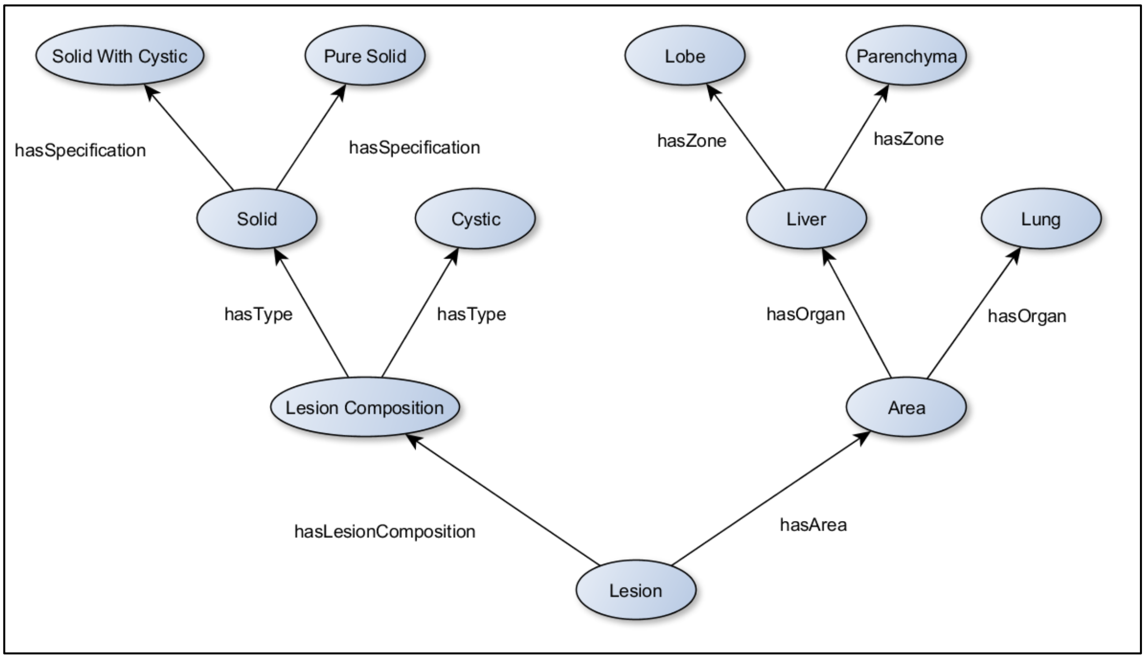


Figure XI – Medical ontology example

4.3.1 Unified Medical Language System

One of the most complete clinical knowledge environments in the biomedical domain is the Unified Medical Language System (UMLS)[47]. The UMLS is a system built with the purpose of giving support in the development of systems that help the healthcare domain, by retrieving clinical information from plenty of different trustable sources. This system possesses many controlled biomedical vocabularies and ontologies in the biomedical domain and can be considered as a centralized system of biomedical knowledge.

These biomedical vocabularies and ontologies exist in different languages in UMLS. However, the vocabularies and ontologies concerning the Portuguese are really poor and lack a lot of biomedical information, what prevents its direct utilization in EMRs written in Portuguese language, like the ones used in this research.

The UMLS system has three main knowledge sources that can be used:

- Metathesaurus: contains plenty of biomedical concepts based in controlled vocabularies and classification systems.
- Semantic Network: contains plenty of semantic types and their relationships. This network allows the connection of the Metathesaurus' concepts with their semantic counterparts.
- SPECIALIST Lexicon: contains orthographic, syntactic and morphological information in the biomedical domain.

The main module of UMLS is the Metathesaurus. In this module, each term has a unique identifier (SUI) that is mapped to a concept identifier (CUI). Each concept can have several different terms associated, but each term can only have one concept associated. This strategy allows the normalization of different terms that express the same concept. This is useful for retrieving clinical information from narrative texts, where some concepts can be expressed by many different terms.

4.3.2 Clinical base of knowledge used in cTAKES

The UMLS system is the base of knowledge used by the NLP system cTAKES used in this research, in order to identify the clinical terms, concepts and the relations between them, present in the EMRs. In order to be able to identify and extract the clinical terms found in the clinical narratives, the NLP system cTAKES uses a dictionary filled with clinical terms and concepts from the Unified Medical Language System (UMLS). The dictionary that cTAKES uses is configurable and can be fully personalized by the user.

The cTAKES system uses by default a dictionary that is a subset of UMLS in order to map the identified clinical terms to concepts. This dictionary includes SNOMED CT [48] and RxNORM [49] concepts. SNOMED CT and RxNORM have recognized collections of clinical terminology and vocabulary. Since these two collections are widely used in this type of researches and cover a significant part of clinical terminology and vocabulary, they are used in this research too.

5. ASCKE Development

This chapter gives a detailed explanation about the development of the prototype system ASCKE. Firstly, it is explained how the data received from the hospital was prepared and pre-processed. Following that, an explanation concerning the MT and NLP components of the ASCKE system is given. Finally, it's described at the end of this chapter how precisely the author used the NLP system to extract clinical information from the EMRs and finally extract clinical knowledge. Each component and activity are explained in this chapter accordingly with the order defined in Figure XII that depicts the ASCKE system architecture.

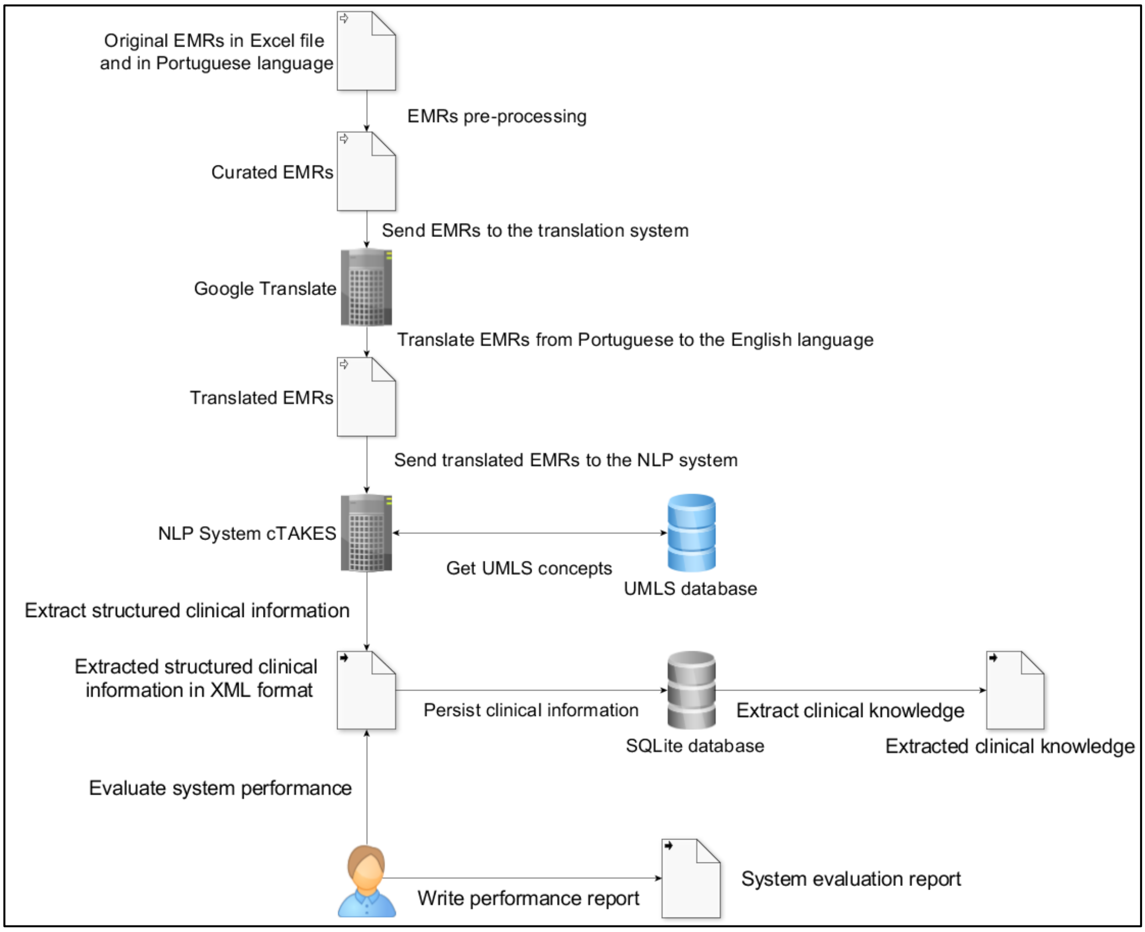

Original EMRs in Excel file and in Portuguese language
EMRs pre-processing
Curated EMRs
Send EMRs to the translation system
Google Translate
Translate EMRs from Portuguese to the English language
Translated EMRs
Send translated EMRs to the NLP system
NLP System cTAKES
Get UMLS concepts
UMLS database
Extract structured clinical information
Extracted structured clinical information in XML format
Persist clinical information
SQLite database
Extract clinical knowledge
Extracted clinical knowledge
Evaluate system performance
Write performance report
System evaluation report

Figure XII – ASCKE prototype system architecture

5.1 EMRs pre-processing

As shown in Figure XII, the first activity is the EMRs pre-processing. As already explained in section 3.2, each EMR corresponds to a line in the Excel file given by the hospital. Each column of the Excel file has specific information about the EMRs, and one of those columns corresponds to the EMRs' associated clinical narratives.

Healthcare practitioners typically use many clinical abbreviations and acronyms in the clinical narratives that they write in EMRs. This fact presents a severe challenge to the processing and translation of the original text, since abbreviations and acronyms typically don't get translated well. In order to overcome that problem, one of the pre-processing activities was identifying all the abbreviations and acronyms present in the EMRs and converting them to their full form. One healthcare practitioner from the hospital helped by clarifying the meaning of all the acronyms and abbreviatures present in the EMRs. The other pre-processing activities were the correction of orthographic errors and the removal of EMRs with empty narratives. All of this pre-processing was done directly in the Excel file using macros and regular expressions. After all these pre-processing activities the EMRs were then ready to be translated, as explained in section 5.2.

5.2 Send EMRs to the Translation System

As shown in Figure XII, the next activity consists in sending the EMR to the translation system. After the pre-processing, a translation of the EMRs from Portuguese to the English language was needed, in order to use the specialized open-source clinical NLP system cTAKES, that is built to work with the English language by default. This reason allied to several others, justify this need for translation, as already explained in detail in section 4.1.

Using a Python script developed by the author, all the information present in the EMRs was sent to the Google Translation API, in order to translate them. The translation results were treated as explained in section 5.3.

5.3 Translate EMRs from Portuguese to the English language

After translating each EMR, the developed Python script was also coded to write each one of the translation results obtained from the Google Translation API in new text files, so that each new text file corresponded to a given EMR translated to the English language. Having each translated EMR extracted and saved in a text file, these files were then 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. As shown in Figure XII and explained in section 5.4, sending the translated EMRs to the NLP system is then the next activity.

5.4 Send translated EMRs to the NLP system

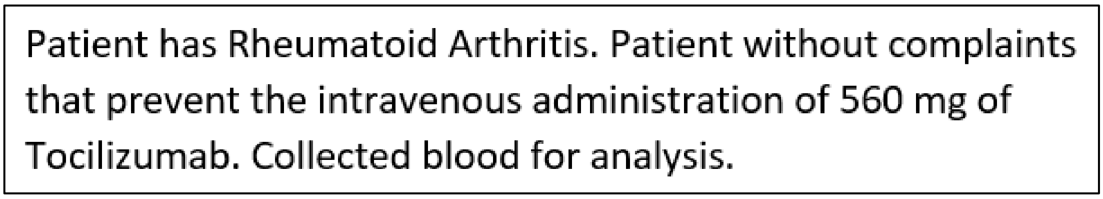
Having all text files created with the translated EMRs in a directory, it's possible to finally send them to the NLP system used in this research, cTAKES, in order to extract all the structured clinical information. The NLP system cTAKES can go to a directory and process all of the text files inside, creating an output file by each input file, with all the structured clinical information extracted for the respective EMR. In order to process the files, it's necessary to define an input directory in the CPE mode of cTAKES, choose the text processor, define the consumer and finally the output directory for the outputted files by cTAKES, as explained in section 4.2.2. As explained and justified in section 4.2.2 too, the consumer that outputs in an XML format is used in this research.

5.5 Processing the clinical narratives with cTAKES CPE module

In this activity, using the setup defined in section 5.4, cTAKES directly process each text file in the chosen input directory, outputting a new XML file for each respective input file, to the chosen output directory. Each outputted XML file contains all the clinical findings extracted from the respective input text file, in a structured way. In section 4.2.1 it was already explained how cTAKES internally processes the clinical narratives in order to extract structured clinical information. Therefore, in the next section, it's explained how exactly the outputted XML files are structured and processed in this research.

5.6 Extracted structured clinical information in XML format

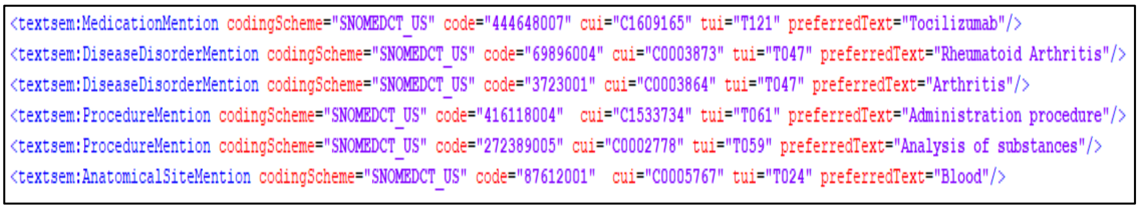
In this step, cTAKES already created all the XML files for each input clinical narrative. Each XML file has all the structured clinical information extracted for each clinical narrative from each EMR. As an example, let’s consider the text file presented in Figure XIII that corresponds to an EMR clinical narrative.



Patient has Rheumatoid Arthritis. Patient without complaints that prevent the intravenous administration of 560 mg of Tocilizumab. Collected blood for analysis.

Figure XIII – EMR clinical narrative example

After the cTAKES processing of this clinical narrative, an XML file is created with all the clinical information extracted. In Figure XIV one can partially see the XML file created for the clinical narrative of Figure XIII.



```
<textsem:MedicationMention codingScheme="SNOMEDCT_US" code="444648007" cui="C1609165" tui="T121" preferredText="Tocilizumab"/>
<textsem:DiseaseDisorderMention codingScheme="SNOMEDCT_US" code="69896004" cui="C0003873" tui="T047" preferredText="Rheumatoid Arthritis"/>
<textsem:DiseaseDisorderMention codingScheme="SNOMEDCT_US" code="3723001" cui="C0003864" tui="T047" preferredText="Arthritis"/>
<textsem:ProcedureMention codingScheme="SNOMEDCT_US" code="416118004"  cui="C1533734" tui="T061" preferredText="Administration procedure"/>
<textsem:ProcedureMention codingScheme="SNOMEDCT_US" code="272389005" cui="C0002778" tui="T059" preferredText="Analysis of substances"/>
<textsem:AnatomicalSiteMention codingScheme="SNOMEDCT_US" code="87612001"  cui="C0005767" tui="T024" preferredText="Blood"/>
```

Figure XIV – Part of the created output XML file with clinical information extracted

As it’s possible to observe in Figure XIV, the clinical information extracted from the clinical narrative of Figure XIII is contained in the XML file. The XML file shows that cTAKES was able to find a medication called “Tocilizumab”, a disease such as “Rheumatoid Arthritis”, procedures such as “Analysis of substances” and “Administration procedure” and finally “Blood” as a part of the human body. Despite not being shown in Figure XIV, the system could also retrieve as a measurement mention the amount of medication “560mg”. However, measurement findings are out of the scope for this research, so this finding was not considered. It’s also possible to see in Figure XIV that the SNOMED-CT collection of clinical terminologies and vocabularies was used to find these UMLS concepts.

With this example, it’s possible to conclude that cTAKES can successfully extract structured clinical information that exists in a given EMR and output an XML file with it. Therefore, the XML file contains all the extracted clinical information in a given EMR and more importantly, represents that information in a structured format. Henceforth, with all the extracted clinical information structured in XML files, it's possible to persist that information and finally extract clinical knowledge from it, as explained in the next section.

5.7 Database persistence

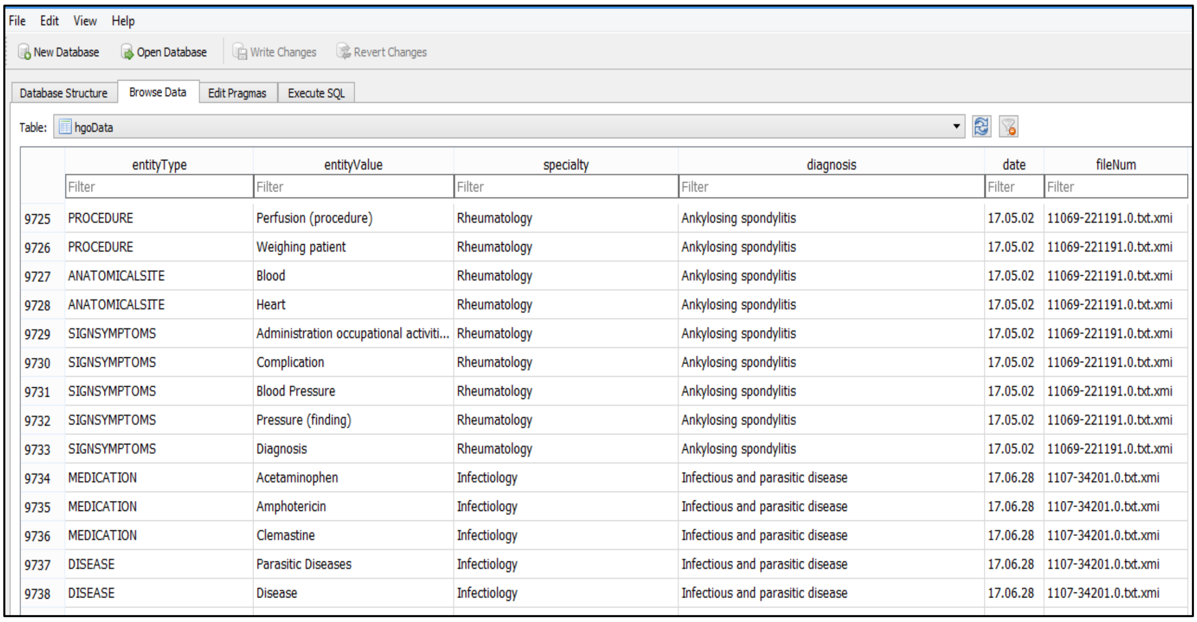
In this step, as shown in Figure XII, all the produced XML files are created and contained in a directory. The objective in this phase is to persist all the extracted clinical information contained in the XML files in a database, in order to extract clinical knowledge in a fast and automated way right after. Another Python script was created in order to process the XML files and persist their clinical information in a structured way in the database.

Since there are only 5255 EMRs available, the author considered that an SQLite database was appropriate to conduct this research, since the dataset is not too large. Since the complexity of the data is simple, the author opted for persisting all information in one single table.

A table was then created in SQLite with the following columns: entity type, entity value, speciality, diagnosis, date and file number. The entity type column can have the following values: medication, disease, anatomic region, sign/symptom or clinical procedure.

The entity value column corresponds literally to the entity value that was found in the XML. Speciality corresponds to the EMR clinical speciality. Diagnosis corresponds to the patient diagnosis since each EMR has a dedicated field that only has the patient diagnosis. The date corresponds to the date in which the EMR was created. The file number corresponds to the file that contained the curated EMR in which the clinical entity was found.

In Figure XV it’s possible to observe a snippet of some structured clinical information already persisted in the database, by using the Python script to process all the XML files information and fill the created columns. Now that all the structured clinical information is persisted, it’s possible to extract clinical knowledge as shown in Figure XII and explained in section 5.8.



	entityType	entityValue	specialty	diagnosis	date	fileNum
9725	PROCEDURE	Perfusion (procedure)	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9726	PROCEDURE	Weighing patient	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9727	ANATOMICALSITE	Blood	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9728	ANATOMICALSITE	Heart	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9729	SIGNSYMPTOMS	Administration occupational activiti...	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9730	SIGNSYMPTOMS	Complication	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9731	SIGNSYMPTOMS	Blood Pressure	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9732	SIGNSYMPTOMS	Pressure (finding)	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9733	SIGNSYMPTOMS	Diagnosis	Rheumatology	Ankylosing spondylitis	17.05.02	11069-221191.0.txt.xmi
9734	MEDICATION	Acetaminophen	Infectiology	Infectious and parasitic disease	17.06.28	1107-34201.0.txt.xmi
9735	MEDICATION	Amphotericin	Infectiology	Infectious and parasitic disease	17.06.28	1107-34201.0.txt.xmi
9736	MEDICATION	Clemastine	Infectiology	Infectious and parasitic disease	17.06.28	1107-34201.0.txt.xmi
9737	DISEASE	Parasitic Diseases	Infectiology	Infectious and parasitic disease	17.06.28	1107-34201.0.txt.xmi
9738	DISEASE	Disease	Infectiology	Infectious and parasitic disease	17.06.28	1107-34201.0.txt.xmi

Figure XV – Database screenshot partially showing the clinical information persisted

5.8 Extract Clinical Knowledge

With all the clinical information persisted and structured in the SQLite database, it’s finally possible to extract clinical knowledge, as shown in Figure XII. Having in mind the database structure presented in Figure XV, SQL was used to query all the clinical information and extract the clinical knowledge desired in this research. Clinical knowledge such as which diseases, medications, signs/symptoms, clinical procedures and anatomical regions are mostly identified in each clinical speciality can now be discovered and be of great use to the hospital, for example. All the clinical knowledge extracted with the ASCKE system developed in this research, is shown and explained in more detail in chapter 6.

6. ASCKE Evaluation

This chapter presents the results obtained with the application of the ASCKE prototype system, built in this research, to 5255 EMRs of a real Portuguese hospital. After having all the extracted clinical information structured and persisted in a database, it's possible to extract clinical knowledge using SQL to query the persisted information. That is the ultimate objective of this research: to validate if the ASCKE system built in this research is capable of extracting reliable clinical knowledge that can be useful for this Portuguese hospital and consequently have the potential to be applied in other hospitals too.

Using the ASCKE system to process the 5255 hospital EMRs, the following results were obtained. The three clinical specialities most represented in the EMRs can be seen in Table XII. Table XIII presents the top five of the most frequent patients' diagnoses found in the EMRs. Since a significant number of EMRs belong to oncology patients, even outside the oncology speciality, the most found diagnosis was "Tumours". The following speciality with more EMRs was rheumatology, what explains why the second most identified diagnosis was "Rheumatoid arthritis". These two tables, Table XII and Table XIII, meet the requisite (1) defined in section 3.3 in Table VI for the ASCKE system.

Speciality	Number of EMRs
Oncology	3150
Rheumatology	619
Pulmonology	529

Table XII – Most represented clinical specialties

Diagnosis	Occurrences
Tumours (neoplasms)	3448
Rheumatoid arthritis	421
Digestive system disease	251
Blood disease	182
Respiratory system disease	165

Table XIII – Top five diagnosis found

Since the EMRs belong to ten different clinical specialities and a lot of diagnoses are present, the knowledge extraction is only exhibited concerning the three most represented specialities and diagnosis. However, the results could be easily obtained to the other specialities too, since they are also persisted in the database. Different tables were created and are presented next in order to show more results obtained in this research concerning clinical knowledge extraction.

Considering now the requisite (2) of the ASCKE system defined in Table VI of section 3.3, the next tables present the clinical knowledge obtained by each speciality. Table XIV presents the most identified diseases by clinical speciality. One can see that in oncology the most identified disease is "Neoplasm" and in rheumatology is "Rheumatoid Arthritis" followed by "Spondylitis", what makes sense in the respective contexts. Table XV presents the most identified medications by clinical speciality. Table XVI presents the most identified signs/symptoms by clinical speciality. In this case, "Pain" is one of the most identified symptoms in the three clinical specialities. Table XVII presents the most identified anatomical regions by clinical speciality while Table XVIII presents the most identified clinical procedures by clinical speciality.

Oncology	Number	%	Rheumatology	Number	%	Pulmonology	Number	%
Neoplasm	2979	73	Rheumatoid Arthritis	678	38.5	Asthma	47	37.9
Gastroesophageal reflux disease	96	2.3	Spondylitis	169	17	Respiratory infections	32	25.8
Neutropenia	89	2.2	Spinal diseases	62	9.8	Pneumonia	20	16.1

Table XIV – Most identified diseases by speciality

Oncology	Number	%	Rheumatology	Number	%	Pulmonology	Number	%
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

Table XV – Most identified medications by speciality

Oncology	Number	%	Rheumatology	Number	%	Pulmonology	Number	%
Pain	354	20.7	Pain	196	9.9	Chest Pain	49	11.8
Nausea	156	9.8	Arthralgia	194	9.8	Tremor	35	8.5
Poor venous access	103	7.3	Joint swelling	185	9.4	Severe asthma	31	7.5

Table XVI – Most identified signs/symptoms by speciality

Oncology	Number	%	Rheumatology	Number	%	Pulmonology	Number	%
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

Table XVII – Most identified anatomical regions by speciality

Oncology	Number	%	Rheumatology	Number	%	Pulmonology	Number	%
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 XVIII – Most identified clinical procedures by speciality

Now concerning clinical knowledge obtained by each diagnosis, Table XIX presents the most identified medications by each diagnosis and Table XX presents the most identified signs/symptoms by each diagnosis. It’s possible to observe that the most identified sign/symptom in the “Digestive system disease” diagnosis is “Digestion problems” followed by “Abdominal pain”, while in the “Rheumatoid arthritis” diagnosis the most identified symptoms are “Pain” followed by “Arthralgia”. These findings make sense considering the diagnosis to which they are associated with.

Digestive system disease	Number	%	Neoplasm	Number	%	Rheumatoid Arthritis	Number	%
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

Table XIX – Most identified medications by diagnosis

Digestive system disease	Number	%	Neoplasm	Number	%	Rheumatoid Arthritis	Number	%
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

Table XX – Most identified signs/symptoms by diagnosis

Finally, considering the requisite (3) of the ASCKE system defined in Table VI of section 3.4, the next tables show the knowledge extracted by the ASCKE system, concerning the incidence of the extracted clinical terms in the different phases of the year. In order to do that, since the EMRs have a date field associated, the author decided to split the year in trimesters and verify the incidence of the most identified three diseases and signs/symptoms in each one of those trimesters. Table XXI presents the incidence of the most identified diseases and Table XXII the most identified signs/symptoms incidence, by each trimester of the year. Having in mind that the number of EMRs by speciality by each trimester is almost the same, it's possible to observe that a variance exists concerning the incidence of each clinical finding by each trimester of the year. This clinical knowledge can be useful for the hospital in terms of resources management for example.

Trimester	Oncology	#	%	Rheumatology	#	%	Pulmonology	#	%
From 01/01/2017 to 31/03/2017	Neoplasm	645	21.7	Rheumatoid Arthritis	140	20.6	Asthma	6	12.8
	Gastroesophageal reflux disease	24	25	Spondylitis	40	23.7	Respiratory infections	2	0.6
	Neutropenia	17	19.1	Spinal diseases	16	25.8	Pneumonia	8	40
From 01/04/2017 to 30/06/2017	Neoplasm	832	27.9	Rheumatoid Arthritis	149	22	Asthma	14	29.8
	Gastroesophageal reflux disease	23	24	Spondylitis	49	29	Respiratory infections	3	0.9
	Neutropenia	26	29.2	Spinal diseases	18	29	Pneumonia	3	15
From 01/07/2017 to 30/09/2017	Neoplasm	798	26.8	Rheumatoid Arthritis	185	27.4	Asthma	20	42.6
	Gastroesophageal reflux disease	21	21.8	Spondylitis	38	22.5	Respiratory infections	11	34.4
	Neutropenia	28	31.5	Spinal diseases	12	19.4	Pneumonia	1	5
From 01/10/2017 to 31/12/2017	Neoplasm	704	23.6	Rheumatoid Arthritis	204	30	Asthma	7	14.8
	Gastroesophageal reflux disease	28	29.1	Spondylitis	42	24.9	Respiratory infections	16	50
	Neutropenia	18	20.2	Spinal diseases	16	25.8	Pneumonia	8	40

Table XXI – Most identified diseases incidence by each trimester of the year

Trimester	Oncology	#	%	Rheumatology	#	%	Pulmonology	#	%
From 01/01/2017 to 31/03/2017	Pain	71	20	Pain	39	19.9	Chest Pain	8	16.3
	Nausea	24	15.4	Arthralgia	34	17.5	Tremor	5	14.3
	Poor venous access	16	15.5	Joint swelling	31	16.8	Severe asthma	4	12.9
From 01/04/2017 to 30/06/2017	Pain	122	34.4	Pain	67	34.2	Chest Pain	13	26.5
	Nausea	54	34.6	Arthralgia	67	34.5	Tremor	8	22.8
	Poor venous access	29	28.2	Joint swelling	62	33.5	Severe asthma	7	22.6
From 01/07/2017 to 30/09/2017	Pain	98	27.7	Pain	53	27	Chest Pain	17	34.7
	Nausea	49	31.4	Arthralgia	52	26.8	Tremor	15	42.9
	Poor venous access	43	41.7	Joint swelling	50	27	Severe asthma	14	45.2
From 01/10/2017 to 31/12/2017	Pain	63	17.9	Pain	37	18.9	Chest Pain	11	22.5
	Nausea	29	18.6	Arthralgia	41	21.2	Tremor	7	20
	Poor venous access	15	14.6	Joint swelling	42	22.7	Severe asthma	6	19.3

Table XXII – Most identified signs/symptoms incidence by each trimester of the year

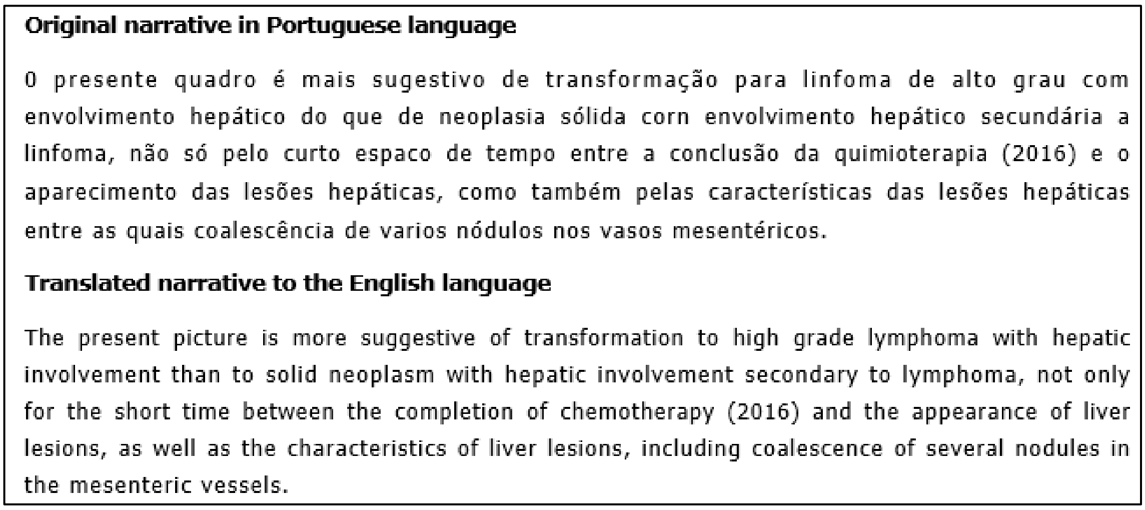
6.1 Individual components evaluation

The system evaluation is split into two parts in this research. Firstly, an evaluation of the translator used is conducted, in order to validate its performance in translating the EMR's clinical information from the Portuguese language to the English language. Secondly, an evaluation of the ASCKE system as a whole is performed, which means having the translator and the NLP system coupled and working together for the evaluation.

6.1.1 Translator Evaluation

Even though the translator is considered out of the scope of this research, an evaluation was made in order to guarantee that the Google Translate had a reasonable performance for this research. A reasonable performance means that not much clinical information is lost in the process of translation from the Portuguese language to the English language. In order to conduct this evaluation, 50 EMRs were translated and manually revised by the author in order to validate if not much clinical information was being lost in the process of translation.

The careful pre-processing of the clinical narratives used in this research, allied to the fact that the pair Portuguese-English language performs really well in terms of translation, made this translation step a success in terms of performance. From the 50 manually revised EMRs, 45 had no loss of clinical information, and the other 5 had only minor translation issues. In Figure XVI a correctly translated EMR is presented and in Figure XVII it's possible to observe one of the 5 EMRs that had some minor issues with the translation process.



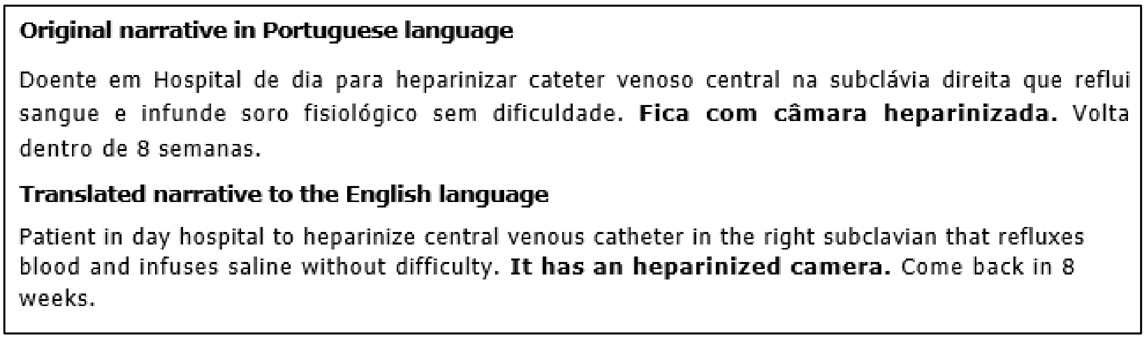
Original narrative in Portuguese language

O presente quadro é mais sugestivo de transformação para linfoma de alto grau com envolvimento hepático do que de neoplasia sólida corn envolvimento hepático secundária a linfoma, não só pelo curto espaco de tempo entre a conclusão da quimioterapia (2016) e o aparecimento das lesões hepáticas, como também pelas características das lesões hepáticas entre as quais coalescência de varios nódulos nos vasos mesentéricos.

Translated narrative to the English language

The present picture is more suggestive of transformation to high grade lymphoma with hepatic involvement than to solid neoplasm with hepatic involvement secondary to lymphoma, not only for the short time between the completion of chemotherapy (2016) and the appearance of liver lesions, as well as the characteristics of liver lesions, including coalescence of several nodules in the mesenteric vessels.

Figure XVI – Correctly translated EMR from the Portuguese language to the English language



Original narrative in Portuguese language

Doente em Hospital de dia para heparinizar cateter venoso central na subclávia direita que reflui sangue e infunde soro fisiológico sem dificuldade. **Fica com câmara heparinizada.** Volta dentro de 8 semanas.

Translated narrative to the English language

Patient in day hospital to heparinize central venous catheter in the right subclavian that refluxes blood and infuses saline without difficulty. **It has an heparinized camera.** Come back in 8 weeks.

Figure XVII – Partially incorrect translated EMR from Portuguese language to the English language

In Figure XVI the EMR got well translated from the Portuguese language to the English language with no loss of clinical information at all. In Figure XVII, a minor issue, marked bold in the text, occurred with the translation that led to the loss of some clinical information. The expression "heparinised camera" got literally translated from the Portuguese language, but in the English language this expression is different and shouldn't be literally translated. This can cause the NLP system not to identify this occurrence and clinical information to be lost. However, this kind of mistakes from the translator is rare from what the author tested using the Google Translate in this research.

6.1.2 Whole system evaluation

The evaluation of the ASCKE system was performed based on standard metrics calculated for 75 of the 5255 EMRs. These standard metrics are precision, recall and F_1 score. They are frequently used in the evaluation of IE systems [50]. The author used these metrics to calculate the performance of the ASCKE system in extracting clinical information from 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. The 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) it's possible to 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.

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

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

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

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

The author asked for the help of two healthcare practitioners from the hospital, who manually annotated the clinical terms present in 75 EMRs in order to establish a gold standard for this evaluation. The evaluation of the ASCKE 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. By using the gold standard established by the doctors, the author could obtain the correctly identified terms by the system, apply the metrics showed above and perform a system evaluation. The evaluation showed that the ASCKE system coupled together in this research has a precision of 0.75, recall of 0.61 and a F_1 score of 0.67.

6.2 Discussion

The results obtained show that the pipeline system coupled together in this research is viable to extract reliable clinical knowledge from Portuguese EMRs, using MT and NLP coupled together. 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 possibly has one of the greatest state of the art performance results for the English language, with a precision of 0.8, recall of 0.65 and F_1 score of 0.72 [11]. A direct comparison between the values of the metrics obtained in this research and the ones obtained in the state of the art by cTAKES can be observed in table XXIII.

The results obtained in this research by the ASCKE system are a just little below the cTAKES' state of the art 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 in performance too.

	ASCKE results	**cTAKES SotA results**
Precision	0.75	0.8
Recall	0.61	0.65
F_1 score	0.67	0.72

Table XXIII – Comparison between this research and the state of the art results

7. Conclusions and Future Work

This research shows that the ASCKE prototype system built by the author, based in an MT and an NLP system, is capable of extracting useful and reliable clinical knowledge from EMRs written in the Portuguese language from a real hospital. This research also shows that this system can be applied to any other hospital, even if the hospital doesn't use the Portuguese language, provided that a translation with a good performance is possible from the original language to the English language, as it's the case of this research. This extraction can be essential to support the hospital in his day-to-day activities and management tasks. It also shows an automated way of extracting clinical knowledge without wasting human resources to review the EMRs manually.

In summary, this research contributes by showing that an approach to clinical knowledge extraction using a system like ASCKE, that couples MT and NLP together, is valid. A system like this can be useful when working with clinical resources written in languages that don't have as much clinical resources as the English language and a translation from the original language to the English language can be achieved with a good performance.

This master thesis already originated the following publication in a journal [51] and a conference proceedings [52].

7.1 ASCKE limitations

One limitation of the ASCKE system is the translation of the EMRs from the Portuguese language to the English language. Some performance is always lost in this step because not even all clinical terms or expressions get correctly translated. This is a reason as for why the results of the ASCKE system were a little lower than state of the art results of the cTAKES system, that were obtained using clinical documents natively written in the English language. However, even knowing that the translation is out of the scope of this research, the author considers that the translation occurred with excellent performance and not much information was lost in this step, based on tests made by the author described in more detail in section 6.1.1.

The careful pre-processing made to the EMRs before the translation was an important step to guarantee a good translation performance. Nonetheless, the results are promising and give motivation to keep conducting this research and improving the system even more since they are close to the state of the art results.

7.2 Answers to Research Questions

Concerning questions (1) and (3), this research showed that it is indeed possible to extract clinical information and knowledge successfully from Portuguese EMRs. The results obtained in this research also answered question (2), by proving that the ASCKE system, by coupling MT and NLP together, can be used in order to perform those extractions successfully.

Therefore, different languages than English language have an opportunity to use this language plentiful biomedical resources available and extract clinical knowledge with a good performance, using a system like the ASCKE system built and described in this research. However, these other different languages than English are always dependent on the quality of translation available for them, as also shown in this research.

7.3 Future Work

The author also intends, 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 established in this research. The hospital will soon also make available more 25000 EMRs in order to keep conducting this research, with a significant part of them from inpatient care. This way the author pretends to compare the differences in terms of clinical knowledge between the two types of patient care: ambulatory and inpatient care.

Finally, the author pretends 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. Finally, the author aims to apply this methodology to other hospitals too and compare the results. There are then several possibilities to explore this research even further in the future, aiming always to improve the healthcare domain in every way possible.

Bibliography

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

[9] 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.

[10] A. N. Ananthakrishnan, T. Cai, G. Savova, S.-C. Cheng, P. Chen, R. G. Perez, V. S. Gainer, S. N. Murphy, P. Szolovits, Z. Xia, S. Shaw, S. Churchill, E. W. Karlson, I. Kohane, R. M. Plenge, and K. P. Liao, “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.

[11] G. K. Savova, J. J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, “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.

[12] 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.

[13] L. da S. Ferreira, “Medical Information Extraction in European Portuguese",

Universidade de Aveiro, 2011.

[14] H. Y. Feifan Liu , Chunhua Weng, “Natural Language Processing, Electronic Health Records, and Clinical Research,” 2012.

[15] H. Cunningham, H. Cunningham, K. Humphreys, K. Humphreys, R. Gaizauskas, R. Gaizauskas, Y. Wilks, and Y. Wilks, “GATE -- a TIPSTERbased General Architecture for Text Engineering,” in *TIPSTER Text Program (Phase III) 6 Month Workshop*, 1997.

[16] K. Liu, K. J. Mitchell, W. W. Chapman, and R. S. Crowley, “Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set.,” *AMIA Annu. Symp. Proc.*, vol. 11, no. Figure 1, pp. 460–4, 2005.

[17] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, “Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system,” *BMC Med. Inform. Decis. Mak.*, vol. 6, no. 1, p. 30, Dec. 2006.

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

[19] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” in *Proceedings of the AMIA Symposium*, 2001, p. 17.

[20] G. Schadow and C. J. McDonald, “Extracting structured information from free text pathology reports,” *AMIA Annu Symp Proc*, pp. 584–588, 2003.

[21] J. Chung and S. Murphy, “Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports.,” *AMIA Annu. Symp. Proc.*, pp. 131–5, 2005.

[22] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, “A general natural-language text processor for clinical radiology,” *J. Am. Med. Informatics Assoc.*, vol. 1, no. 2, pp. 161–174, 1994.

[23] A. C. Castilla, S. S. Furuie, and E. A. Mendonça, “Multilingual information retrieval in thoracic radiology: feasibility study.,” *Stud. Health Technol. Inform.*, vol. 129, no. Pt 1, pp. 387–91, 2007.

[24] N. L. Jain and C. Friedman, “Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports.,” *Proc. AMIA Annu. Fall Symp.*, pp. 829–833, 1997.

[25] 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.

[26] E. W. Brown, A. Dolbey, and L. Hunter, “TREC 2003 Genomics Track.”

[27] R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, A. Coden, J. Cooper, A. Inokuchi, B. Iyer, Y. Mass, H. Matsuzawa, and L. V. Subramaniam, “Text analytics for life science using the Unstructured Information Management Architecture,” *IBM Syst. J.*, vol. 43, no. 3, pp. 490–515, 2004.

[28] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, “MedEx: a medication information extraction system for clinical narratives,” *J. Am. Med. Informatics Assoc.*, vol. 17, no. 1, pp. 19–24, Jan. 2010.

[29] M. Jiang, Y. Wu, A. Shah, P. Priyanka, J. C. Denny, and H. Xu, "Extracting and standardizing medication information in clinical text - the MedEx-UIMA system.," *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2014, no. RxCUI 20610, pp. 37–42, 2014.

[30] H. Xu, M. Jiang, M. Oetjens, E. A. Bowton, A. H. Ramirez, J. M. Jeff, M. A. Basford, J. M. Pulley, J. D. Cowan, X. Wang, M. D. Ritchie, D. R. Masys, D. M. Roden, D. C. Crawford, and J. C. Denny, "Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin.," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 4, pp. 387–91, 2011.

[31] 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.

[32] 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.

[33] V. Garla, V. Lo Re, Z. Dorey-Stein, F. Kidwai, M. Scotch, J. Womack, A. Justice, and C. Brandt, "The Yale cTAKES extensions for document classification: Architecture and application," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 5, pp. 614–620, Sep. 2011.

[34] C. Y. Wu, C. K. Chang, D. Robson, R. Jackson, S. J. Chen, R. D. Hayes, and R. Stewart, "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.

[35] W. Pratt and M. Yetisgen-Yildiz, "A Study of Biomedical Concept Identification: MetaMap vs. People."

[36] K. P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng-treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, I. Kohane, E. W. Karlson, and R. M. Plenge, "Electronic medical records for discovery research in rheumatoid arthritis," *Arthritis Care Res. (Hoboken).*, vol. 62, no. 8, pp. 1120–1127, Mar. 2010.

[37] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 3, pp. 371–379, May 2009.

[38] J. Masanz, S. V Pakhomov, H. Xu, S. T. Wu, C. G. Chute, and H. Liu, "Open Source Clinical NLP - More than Any Single System.," *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2014, pp. 76–82, 2014.

[39] L. Zerbinatti, "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", Escola Politécnica da Universidade de São Paulo, 2010.

[40] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study," *J. Am. Med. Informatics Assoc.*, vol. 15, no. 1, pp. 87–98, Jan. 2008.

[41] 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.

[42] 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.

[43] A. Névéol, J. Grosjean, S. J. Darmoni, and P. Zweigenbaum, “Language Resources for French in the Biomedical Domain.”

[44] “Ambulatory Care Settings.” [Online]. Available: http://www.medpac.gov/-research-areas-/ambulatory-care-settings. [Accessed: 01-May-2018].

[45] Google, “Translation API Client Libraries | Translation API | Google Cloud,” *2018*, 2018. [Online]. Available: https://cloud.google.com/translate/docs/. [Accessed: 04-Mar-2018].

[46] K. Liu, W. R. Hogan, and R. S. Crowley, “Natural Language Processing methods and systems for biomedical ontology learning,” *Journal of Biomedical Informatics*, vol. 44, no. 1. NIH Public Access, pp. 163–179, Feb-2011.

[47] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, “The Unified Medical Language System (UMLS),” *Methods Inf. Med.*, vol. 32, no. 4, pp. 281–291, 1993.

[48] “SNOMED CT,” *International Health Terminology Standards Development Organisation*. [Online]. Available: https://www.nlm.nih.gov/healthit/snomedct/index.html. [Accessed: 02-May-2018].

[49] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, “RxNorm: Prescription for electronic drug information exchange,” *IT Professional*, vol. 7, no. 5. pp. 17–23, 2005.

[50] 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.

[51] M. Lamy, R. Pereira, J. Ferreira, Melo, F. & Velez, I. (2018). Extracting clinical knowledge from electronic medical records. IAENG International Journal of Computer Science. 45, 488-493

[52] M. Lamy, R. Pereira, J. Ferreira, (2018). Extracting Clinical Information from Electronic Medical Records. Proceedings ISAMI 2018, Toledo, Spain (accepted, awaiting publication)