PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE

LUCAS EMANUEL SILVA E OLIVEIRA

ASSEMBLING NATURAL LANGUAGE PROCESSING RESOURCES TO PERFORM THE SUMMARIZATION OF CLINICAL NARRATIVES

CURITIBA
2020

LUCAS EMANUEL SILVA E OLIVEIRA

ASSEMBLING NATURAL LANGUAGE PROCESSING RESOURCES TO PERFORM THE SUMMARIZATION OF CLINICAL NARRATIVES

Thesis presented to the Graduate Program in Health Technology of the Pontifical Catholic University of Paraná, as a partial requirement to obtain the title of Ph.D. in Health Technology.

Concentration Area: Health Informatics.

Supervisor: Prof. Dr. Claudia Maria Cabral Moro
Co-supervisor: Dr. Sheikh Sadid Al Hasan.

CURITIBA
2020

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Pamela Travassos de Freitas – CRB 9/1960

O48a
2020

Oliveira, Lucas Emanuel Silva e
Assembling natural language processing resources to perform the summarization of clinical narratives / Lucas Emanuel Silva e Oliveira ; supervisor: Claudia Maria Cabral Moro ; co-supervisor: Sheikh Sadid Al Hasan. – 2020.
165 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2020
Bibliografia: f. 144-165

1. Registros médicos. 2. Banco de dados. 3. Documentos médicos. 4. Doença Crônica. 5. Processamento de linguagem natural (Computação). 6. Registros eletrônicos de saúde. 7. Redes neurais (Computação). 8. Registros hospitalares. I. Moro, Claudia Maria Cabral. II. Hasan, Sheikh Sadid Al. III. Pontifícia Universidade Católica do Paraná. Pós-Graduação em Tecnologia em Saúde. IV. Título.

CDD 20. ed. – 651.504261

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Tecnologia em Saúde

STATEMENT OF THESIS APPROVAL No. 011

The Doctoral Thesis entitled "**ASSEMBLING NATURAL LANGUAGE PROCESSING RESOURCES TO PERFORM THE SUMMARIZATION OF CLINICAL NARRATIVES**" was presented during a virtual viva public session by the candidate **Lucas Emanuel Silva e Oliveira** on August 30, 2020. Therefore, it was judged to obtain the title of Doctor in Health Technology, and it was approved in its final form by the Graduate Program in Health Technology.

BOARD OF EXAMINERS:

Prof. Dr. Claudia Maria Cabral Moro Barra– Orientador e Presidente – PUCPR
Prof. Dr. Riccardo Bellazzi - University of Pavia
Prof. Dr. Noémie Elhadad - Columbia University
Prof. Dr. Roberto Flavio Silva Pecoits Filho - PUCPR
Prof. Dr. Deborah Ribeiro Carvalho - PUCPR

The original copy of this document with the signature of the PPGTS Coordinator, after the delivery of the final version of the work, is filed at the Program Secretariat.

Curitiba, October 30rd, 2020.

TERMO DE APROVAÇÃO DE TESE Nº 011

A Tese de Doutorado intitulada "**DESENVOLVIMENTO DE RECURSOS DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA SUMARIZAÇÃO DE NARRATIVAS CLÍNICAS**" defendida em sessão pública pelo(a) candidato(a) Lucas Emanuel Silva e Oliveira no dia 30 de agosto de 2020, foi julgada para a obtenção do título de Doutor em Tecnologia em Saúde, e aprovada em sua forma final, pelo Programa de Pós-Graduação em Tecnologia em Saúde.

BANCA EXAMINADORA:

Prof. Dr. Claudia Maria Cabral Moro Barra– Orientador e Presidente – PUCPR
Prof. Dr. Riccardo Bellazzi - University of Pavia
Prof. Dr. Noémie Elhadad - Columbia University
Prof. Dr. Roberto Flavio Silva Pecoits Filho – PUCPR
Prof. Dr. Deborah Ribeiro Carvalho - PUCPR

A via original deste documento encontra-se arquivada na Secretaria do Programa, contendo a assinatura do Coordenador após a entrega da versão corrigida do trabalho.

Curitiba, 30 de outubro de 2020.

Prof. Dr. Percy Nohama
Head of PPGTS PUCPR
Coordenador do PPGTS PUCPR

ABSTRACT

Introduction: The expansion of patient's information stored in the Electronic Health Records (EHR) is triggering a problem known as data overload, which can reduce the doctor-patient time and interfere with clinical decision making. This problem is more evident in patients with chronic diseases, who have frequent medical visits, and therefore a larger amount of data to deal with. In this context, Automatic Text Summarization (ATS) systems could process substantial volumes of data from the EHR, extract relevant information, and provide a patient's summary to aid the health practitioners. The lack of Natural Language Processing (NLP) resources and tools in the Portuguese language is a barrier to develop an ATS algorithm to deal with Brazilian Portuguese (pt-br) clinical texts. Therefore, to build a clinical summarization method, one could explore unsupervised and resource-free approaches, or develop NLP artifacts such as annotated corpora and Named Entity Recognition (NER) algorithms. **Objective:** Develop a method to provide summarized EHR information of chronic disease patients by assembling essential clinical NLP resources and exploring existing summarization strategies. **Method:** The first phase of the project aimed to build Information extraction tools, such as a NER algorithm, to identify medical terms within noisy clinical texts. State-of-the-art sequence labeling neural architectures were explored, and two semantically annotated corpora were developed, the SemClinBr and SummClinBr. The SemClinBr is a corpus focusing on the extraction of all the medical terms in clinical notes. The SummClinBr is centered on relevant information to support the development and evaluation of ATS. The annotations cover the longitudinal history of chronic kidney disease patients in concept and sentence-level. In the next phases, ATS methods which already been used for the English language and other domains were explored and eventually adapted and applied to the chronic disease summarization task. Three different methods were proposed. An unsupervised neural summarization model using a fine-tuned pre-trained BERT model was customized and evaluated. The developed corpora and NER algorithm supported both Sequence labeling and Dictionary-based ATS supervised approaches. To avoid redundancy in the generated summaries, a method based on Siamese Neural Networks to estimate textual semantic similarity was proposed. The experimental setup counted with the ROUGE automatic evaluation metric applied over a gold standard that reflects human judgments, which was complemented with an observational evaluation. **Results:** Regarding the completeness of information, the supervised approaches overcame the unsupervised in almost all the experiments, and the Sequence labeling and Dictionary-based alternated the best quantitative results. The results suggest that the use of supervised approaches and domain-specific resources could better supply the clinical area needs, as the intrinsic aspects related to the clinical data could be too complex to deal with statistical methods solely. However, the use of a hybrid strategy, which could explore the main strengths of each approach, together with visualization techniques, sounds promising. **Conclusion:** A set of crucial clinical NLP resources and tools were assembled, and a set of summarization approaches were explored to support an unprecedented clinical summarization attempt in pt-br EHR texts. The developed artifacts allowed the development of supervised methods and could structure background to boost not only the summarization field but multiple clinical NLP tasks as NER, Negation Detection, and Textual Semantic Similarity.

Keywords: Automatic Text Summarization. Clinical Summarization. Natural Language Processing. Machine Learning. Electronic Health Records. Chronic diseases.

RESUMO

Introdução: O aumento no volume de dados do paciente no Registro Eletrônico de Saúde (RES) está ocasionando uma sobrecarga de dados, que pode reduzir o tempo médico-paciente, bem como interferir na tomada de decisão clínica. Isto é um problema ainda mais evidente no contexto de pacientes com doenças crônicas, que devido às suas visitas médicas mais regulares, possuem uma coleção maior de dados. Neste cenário, sistemas de Sumarização Automática de Textos (SAT) podem processar grandes volumes de dados do RES, extraindo informações relevantes e gerando um sumário do paciente que auxilie os profissionais da saúde. A carência de recursos e ferramentas de Processamento de Linguagem Natural (PLN) para o idioma português é uma barreira no desenvolvimento de algoritmos de SAT para textos clínicos no idioma. Assim, para construir um método de sumarização clínica, pode-se explorar abordagens não-supervisionadas e independentes de recursos ou desenvolver artefatos de PLN, como corpora anotados e algoritmos de Reconhecimento de Entidade Nomeadas (REN). **Objetivo:** Desenvolver um método de sumarização de dados de pacientes com doença crônica reunindo recursos essenciais de PLN clínico e explorando estratégias existentes. **Método:** A primeira fase do projeto visou o desenvolvimento de ferramentas de extração de informação, como o REN, para identificar conceitos clínicos em meio ao ruidoso texto clínico. Arquiteturas de redes neurais foram exploradas, e dois corpora anotados semanticamente foram construídos, o SemClinBr e o SummClinBr. O SemClinBr é um corpus que foca na extração de todos termos médicos dos textos clínicos. Já o SummClinBr é centrado apenas nas informações relevantes, para auxiliar no desenvolvimento e avaliação de um sistema SAT. As anotações cobrem o histórico longitudinal de pacientes com doença renal crônica, a nível de conceito e sentenças. Nas fases subsequentes do projeto, métodos de SAT que já foram utilizados para o idioma inglês foram explorados e eventualmente adaptados à tarefa de sumarização clínica. Três diferentes métodos foram propostos. Uma arquitetura de redes neurais para sumarização não-supervisionada, baseada em um modelo pré-treinado do BERT foi customizado e avaliado. Os corpora desenvolvidos apoiaram o desenvolvimento de duas abordagens supervisionadas, uma baseada em rotulagem de sequências e outro em dicionários. Para evitar redundância nos sumários gerados, um método baseado em Redes Neurais Siamesas para estimar a similaridade semântica dos textos foi proposto. **Resultados:** Para avaliação dos resultados a métrica de avaliação automática ROUGE foi aplicada sobre um padrão-ouro, que representa sumários gerados por humanos. Além disso, uma análise observacional de aspectos de qualidade também foi realizada. Em relação a completude das informações, os métodos supervisionados superaram o não-supervisionado em quase todos experimentos, e os métodos baseados em rotulagem de sequências e dicionários alternaram os melhores resultados quantitativos. Os resultados sugerem que o uso de abordagens supervisionadas e recursos específicos de domínio podem melhor suprir as necessidades da área clínica, pois os aspectos intrínsecos relacionados aos dados clínicos podem ser muito complexos para lidar exclusivamente com métodos estatísticos. No entanto, deve-se considerar o uso de uma estratégia híbrida, que explore os principais pontos fortes de cada abordagem, juntamente a técnicas de visualização. **Conclusão:** Um conjunto de ferramentas e recursos de PLN clínico cruciais foram construídos e diferentes abordagens de sumarização foram exploradas para apoiar uma inédita tentativa de sumarização clínica de dados textuais do RES. Os artefatos desenvolvidos permitiram o desenvolvimento de métodos supervisionados e podem estruturar um alicerce de desenvolvimento não apenas para a sumarização clínica, mas para diversas tarefas de PLN, como o REN, Detecção de negação e Similaridade Semântica.

Palavras-chave: Sumarização Automática de Textos. Sumarização Clínica. Processamento de Linguagem Natural. Aprendizagem de Máquina. Registro Eletrônico de Saúde. Doenças crônicas.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, **Claudia Moro**, who, during these years, has always placed enormous trust in my work, and most important of all, has become a great friend.

I would like to thank my son, **Bruno**, who, even without knowing it, was my greatest source of inspiration and strength. You make me a better person every day, son, and I love you!

I also thank my wife, **Ana Carolina**, who besides being my biggest supporter, also spent a lot of time to make up for my absence. Without you, the completion of this work would not be possible.

Thank you **mom** and **dad** for all the teachings, support and love for more than three decades now.

I extend my thanks to all **family members** that helped during the last years.

Thanks to all my **friends and brothers of life**, who, as always, accompanied me during all the moments and provided me with countless moments of happiness.

To my **research colleagues**, professors and students, I would like to thank, either for the immense scientific contribution to this thesis or for the moments of fun. Thank you, Prof. Deborah, Yohan, João Vitor, Lucas, Lilian, Fernanda, and Elisa.

A special thanks to **Dr. Thyago Proença** and **Dr. Ellen Valente** for their precious time, and the valuable Nephrology lessons.

I would like to thank **Philips Healthcare** for funding this research and allowing me to have a privileged condition in the research scenario in Brazil. Count on me to continue supporting research in our country. A big thanks to my co-advisor, **Sadid Hasan**, for the valuable advice and encouragement.

And finally, I want to thank **PUCPR** for being, again, an important part of my professional qualification. Thank you for allowing me to be a professor at the institution, and to be able to give back all the lessons I’ve learned.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION AND ACRONYMS LIST

ATS	Automatic Text Summarization
AVS	After-visit summary
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bi-directional Long Short Term Memory network
BS	Biomedical Summarization
CKD	Chronic Kidney Disease
CRF	Conditional Random Fields
CS	Clinical Summarization
DCT	Document Creation Time
DUC	Document Understanding Conferences
EHR	Electronic Health Records
ESRD	End-stage renal disease
HIT	Health Information Technology
IAA	Inter annotator agreement
IE	Information Extraction
INDR	Indicator representation
IR	Information Retrieval
ISO	International Organization for Standardization
KDIGO	Kidney Disease: Improving Global Outcomes
KDOQI	Kidney Disease Outcomes Quality Initiative
LSA	Latent semantic analysis
ML	Machine Learning
MSE	Mean Squared Error
NCD	Noncommunicable diseases
NER	Named Entity Recognition
NFK	National Kidney Foundation
NLG	Natural Language Generation
NLP	Natural Language Processing
NN	Neural Network
PC	Pearson Correlation
pt-br	Brazilian Portuguese

RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisty Evaluation
SGR	UMLS semantic group
SNN	Siamise Neural Network
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
STY	UMLS semantic type
SVM	Support Vector Machines
TAC	Text Analysis Conference
TF-IDF	Term frequency-inverse document frequency
TR	Topic Representation
UMLS	Unified Medical Language System

SUMMARY

1 INTRODUCTION

The adoption of *Electronic Health Records* (EHR) provided medical practitioners with instant and reliable access to data about patient physiology, procedures, treatments, diagnoses, etc. While the EHR data are used to support and enhance healthcare, the EHR systems can interfere in clinical decision making with its inefficiency to present patients medical history in a comprehensive way, with effective cognitive support, presenting data in an optimal format when it is needed (KENEI et al., 2018; HIRSCH et al., 2014).

The workload of a physician was assessed by Arndt and colleagues (ARNDT et al., 2017), in their research, they concluded that a doctor spends more than half of his work time accessing the EHR. Another research (ALKUREISHI et al., 2016) shows that from 12% to 55% of a patient visit time, the physician utilizes the computer, evidencing the importance of a cohesive EHR to prevent a *decline in doctor-patient time*.

Healthcare providers are very reliant on the utilization of unstructured text to document patient history due to its meaningful and flexible essence. However, the health practitioner’s ability to retrieve information from the text for a clinical overview is severely affected when time pressure increases and documents get more extensive (SULTANUM et al., 2018).

The continuous growth of patient’s clinical information in EHR is causing a *data overload* problem to clinicians (FARRI et al., 2012; HALL; WALTON, 2004; VAN VLECK et al., 2007). Because they have to read and interpret information from such a massive amount of clinical notes to find the information they need. However, manual retrieval of information is time-consuming and can cause omission, communication, and safety errors (MCDONALD, 1976; MCDONALD et al., 2014; HOLDEN, 2011; LISSAUER et al., 1991).

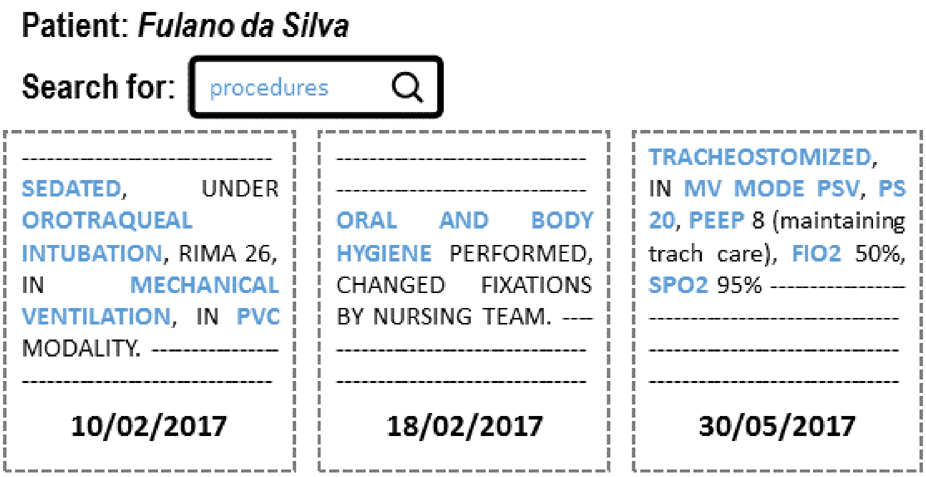
In particular, this can be a major problem when clinicians have to deal with *chronic conditions,* also known as *Noncommunicable diseases* (NCDs)*,* due to their permanent aspect, the patients accumulate a larger dataset based on their regular medical visits. Thus, it is often impossible to read all the notes that describe the history of their condition in every consultation (PIVOVAROV; ELHADAD, 2015). Researchers already investigated the particularities of chronic disease in healthcare, including the data needs (REICHERT et al., 2010; SAMAL et al., 2011) and system design (UNERTL et al., 2009) for chronic disease management, which are vital subjects given that the

NCDs are estimated to account for more than 70% of all global deaths and have a high financial impact on health systems worldwide (WHO, 2018; WHO 2020).

In this context, clinicians need to access relevant information quickly and efficiently. *Information Retrieval* (IR) and *Natural Language Processing* (NLP) systems can help them to manage the data overload problem (PLUYE et al., 2005), however, in many cases, they still have to view a lot of documents/pages to find relevant information (DAVIDOFF; MIGLUS, 2011).

To exemplify, an IR/NLP algorithm would be able to search for all procedures performed on a patient (Figure 1); however, the physician would still need to check all patient documents and irrelevant text to find useful information. In contrast, an *Automatic Text Summarization* (ATS) system could show all relevant and essential information about the patient, preventing the doctor from looking at several documents and unrelated text in search of information (Figure 2). In Figure 3, there is a diagram illustrating the longitudinal aspect of the EHR and a clinical overview (i.e., summary) of a patient, generated by an ATS system, presented in the form of text, graph, and diagram. Therefore, by presenting mainly relevant information, an ATS system could help the health professional to view the patient's general panorama, support the decision making, and save time.

Figure 1 - Example of an IR/NLP system searching for all patient’s procedures. All the documents are shown with the corresponding indication of a procedure within the text (blue color).



Patient: *Fulano da Silva*

Search for: procedures

----------------------------- SEDATED, UNDER OROTRAQUEAL INTUBATION, RIMA 26, IN MECHANICAL VENTILATION, IN PVC MODALITY. ---------------------------------------

10/02/2017

----------------------------- ----------------------------- ORAL AND BODY HYGIENE PERFORMED, CHANGED FIXATIONS BY NURSING TEAM. --- ----------------------------- -----------------------------

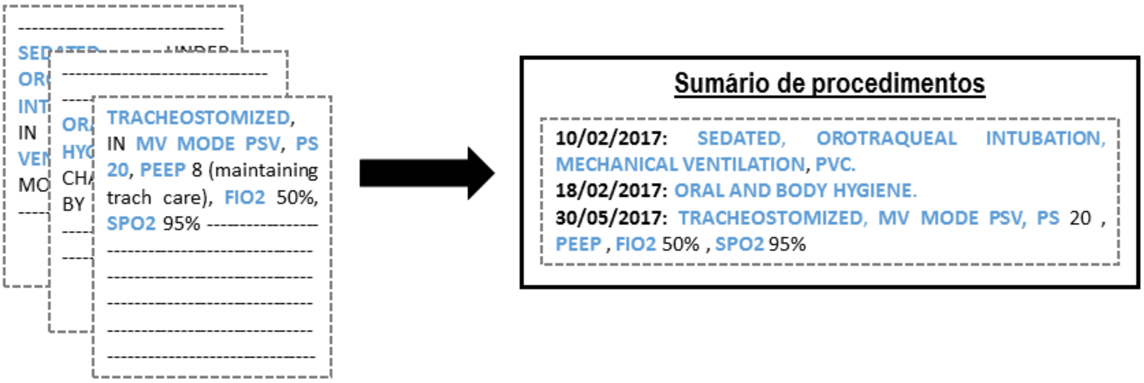
18/02/2017

TRACHEOSTOMIZED, IN MV MODE PSV, PS 20, PEEP 8 (maintaining trach care), FIO2 50%, SPO2 95% -------------- ----------------------------- ----------------------------- -----------------------------

30/05/2017

Source: the author.

Figure 2 - Example of an ATS system showing all essential information regarding the patient's procedures as a single summary.



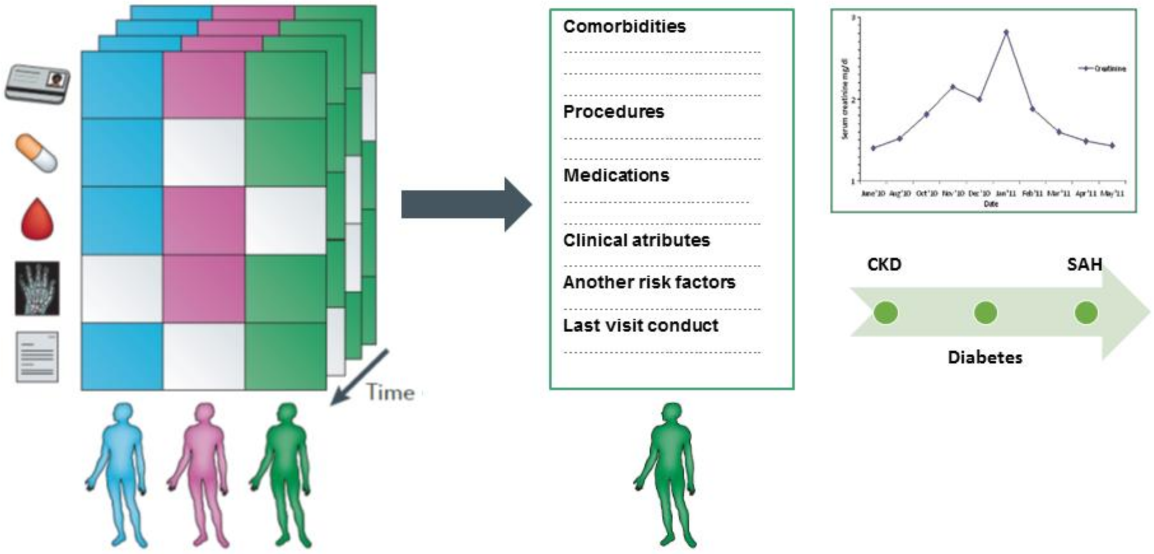
TRACHEOSTOMIZED, IN MV MODE PSV, PS 20, PEEP 8 (maintaining trach care), FIO2 50%, SPO2 95%

Sumário de procedimentos

10/02/2017: SEDATED, OROTRAQUEAL INTUBATION, MECHANICAL VENTILATION, PVC.
18/02/2017: ORAL AND BODY HYGIENE.
30/05/2017: TRACHEOSTOMIZED, MV MODE PSV, PS 20 , PEEP , FIO2 50% , SPO2 95%

Source: the author.

Figure 3 – Diagram representation of a patient's EHR longitudinal data being summarized and presented in three different formats (i.e., textual, graphical, and event timeline).



Source: adapted from Jensen, Jensen and Brunak, 2012.

The ATS research field can deal with *single* or *multiple* documents in order to create a comprehensible compendium of valuable information from the original document(s). An ATS algorithm is classified as *supervised* when it learns from labeled data to select essential information from a document to generate a summary. In opposition, an *unsupervised* system generates a summary without the need of training data, relying mainly on statistical methods to do so (MORADI; GHADIRI, 2018). There is a predominance of unsupervised approaches in the literature. The cost to obtain annotated data and lack of a reference standard to summarization annotation (due to immaturity in the research field) may be some reasons for that (MISHRA et al., 2014). Both approaches can make use of *Machine Learning* (ML) algorithms.

Researchers have developed and evaluated biomedical data summarizers aiming to reduce the amount of superfluous health-related text (LIU; FRIEDMAN, 2004; VAN VLECK; ELHADAD, 2010; WERE et al., 2010; SARKER; MOLLÁ, 2012; HIRSCH et al., 2014; GOLDSTEIN; SHAHAR, 2016; RAMANUJAM; KALIAPPAN, 2016; MORADI; SAMWALD, 2019; ROUANE; BELHADEF; BOUAKKAZ, 2019). They presented techniques to generate summaries with essential and relevant information from biomedical data, but most of them focused on literature summaries rather than the patient's health records summarization (MISHRA et al., 2014).

Concerning specifically the summarization of EHR data, Pivovarov and Elhadad (2015) performed a review focusing on patient longitudinal data summarization methods, where they highlighted a set of methodological challenges in the area and presented how the research community is trying to overcome these difficulties. Despite the increase of studies in this area, just a few attempts of chronic condition summarization were made (HSUEH et al., 2015; GOLDSTEIN; SHAHAR, 2016; ALEKSIĆ et al., 2017; VIANI et al., 2017; ACHARYA et al., 2018; LEVY-FIX et al., 2020), but most of them worked with structured data only. The language and resource dependency is a common characteristic in these studies as well.

Mishra et al. (2014) classified the summarization methods into four categories: statistical, NLP, ML, and hybrid technique. The hybrid approaches are predominant in state-of-the-art researches (SARKER; MOLLÁ, 2012; PLAZA et al., 2011; MORID et al., 2016), as they can apply the best technique to each problem to be solved. Another widely present feature is the use of *knowledge-rich approaches* that leverage a broad set of available tools and knowledge resources (VAN VLECK; ELHADAD, 2010; GOLDSTEIN; SHAHAR, 2016; PLAZA et al., 2011). Annotated corpora designed for ML methods are also common, even though constructed resources are often domain and language-dependent (KVIST et al., 2011; DEMNER-FUSHMAN et al., 2009).

Analyzing studies from the four categories, it is possible to notice that several of them employ primary NLP techniques as auxiliary processes like syntactic parsers to capture the structure of the text, POS-taggers to obtain morphological information and Named Entity Recognition (NER) algorithms to extract medical concepts from the text. Even these kinds of NLP resources are scarce in Brazilian Portuguese (pt-br) when dealing with the clinical domain. The limited number of NLP *resources and tools* regarding *lexical, syntactic,* and *semantic* information in languages other than English could be the cause that most of the work in *Biomedical Summarization* is developed for the English language (MISHRA et al., 2014; GAMBHIR; GUPTA, 2017).

Although many biomedical summarization methods are extremely dependent on a well-defined *annotated corpus,* for both *training* and *testing* purposes, it is difficult to find a generalizable biomedical summarization corpus or reference standard available for new researches, hampering the advance of ML approaches. (MISHRA et al., 2014; YAO et al., 2017).

The following are some examples that illustrate the lack of proper NLP resources and tools to support the summarization of EHR texts. The well-known and extensively used medical terminology mapping and extraction tools, like MetaMap (2010) or cTakes (2010), are not available for pt-br. At the beginning of this project, there was no POS-Tagging model for the clinical domain available for use, despite the mention of work in the literature (OLEYNIK et al., 2010). Moreover, just a few NER studies in pt-br are focused on texts from EHR (LOPES, TEIXEIRA, OLIVEIRA, 2019). In addition to the tools and resources, we were not able to find any work that summarized longitudinal and unstructured patient data in pt-br as well. It is worth to highlight that even in resource-rich languages as the English is difficult to find studies performing EHR summarization, evidencing the immaturity of the field.

Hence, to develop a summary of texts from the patient's medical record, two are the most viable paths: (i) explore completely unsupervised methods, based on statistics that do not require the tools and resources mentioned above, or (ii) build a minimal set of artifacts to support the task, including NLP tools and resources, such as annotated corpora or Information Extraction algorithms (e.g., NER).

Considering the biomedical summarization research gaps; the challenges associated to patient’s longitudinal data summarization, especially in a language with a lack of lexical and semantic resources; the lack of patient longitudinal data summarization in pt-br; and the necessity to reduce the data overload, increase the doctor-patient time and optimize the chronic condition patient care, we present below the hypothesis, objectives and research questions that will guide this study.

Therefore, the present research project has the hypothesis that: *it is possible to create a new Brazilian Portuguese clinical text summarization method focusing on chronic condition patients by developing vital NLP resources and adapting the existing extractive summarization approaches.*

1.1 RESEARCH OBJECTIVES

The current research is driven by a set of *research goals*. The initial goal is the main and broader one, and the subsequent are more specific and essential to fulfilling the primary goal.

- *Develop a method to provide summarized EHR information of chronic disease patients by assembling essential clinical NLP resources and exploring existing summarization strategies.*
    - *Explore and apply unsupervised, statistical, and low-resource summarization approaches.*
    - *Build a set of clinical NLP tools and resources to support the identification of relevant information within clinical text.*
    - *Explore and apply supervised summarization approaches.*

With the goals defined, the following *research questions* should be answered through experiments and further analysis:

- *How well unsupervised and low-resource summarization approaches could select the relevant information in the patient's EHR?*
- *How the lack of NLP tools and resources impacts the use of current unsupervised approaches?*
- *A Named Entity Recognition algorithm trained with an Information Extraction corpus of semantically annotated clinical texts could aid in chronic disease relevant information identification?*
- *Which types of data are relevant in a summary of a patient with chronic disease?*
- *A domain and task-specific gold standard (i.e., corpus for patient's longitudinal data summarization) could improve the chronic disease relevant information identification?*
- *How well supervised summarization approaches could select the relevant information in the patient's EHR?*

- *Deal with clinical summarization as a sequence labeling problem is feasible?*

1.2 RESEARCH SCOPE

Considering the task complexity and the considerable amount of gaps and challenges found in the Biomedical Summarization research field, especially in pt-br texts, we defined a set of boundaries to limit the project scope.

The summarization algorithms explored in this research were limited to some **unsupervised** strategies, approaches that make little or no use of **non-existent lexical and semantic resources** for pt-br, and methods that have already been used in **summarizing clinical data**. More details about the characteristics of these studies and their methods are described in section 2.5.2.

Due to the limited knowledge regarding patient data summarization, the focus of this research was on **extractive summarization**, which makes use of the original information in the documents to assemble the summary. This type of summarization is usually the first step so that we can subsequently have a higher level abstractive text. Furthermore, the study focused on textual outputs instead of applying visualization techniques.

We focused on Nephrology data, gathering clinical notes of **Chronic Kidney Disease** patients, which are a great example of a patient with regular medical visits, and above-average EHR size. Although we focused on the needs of this group of patients, at various points in this research, we considered the possibility of generalization for other chronic diseases, as in the construction of the SummClinBr corpus (i.e., definition of information categories). Moreover, our data comes from patients above stage 3 and below stage 5 of the disease. In this context, the patient was already referred from the primary care to the specialist, so it is not included in our scope to detect the signs and symptoms that lead to the disease, as the doctor already know about the current suspicious. Our research scope includes just the **selection of patient's relevant information aiming to decrease the time in which the medical practitioner searches information in the EHR**.

1.3 DOCUMENT STRUCTURE

The research project document is structured as follows. *Chapter 2* introduces the main topics, challenges, and approaches that are needed in order to understand the work and support the study development. First, we present the health-related topics, then we show the computational methods associated with this research, and finally, the application of those methods in the health domain. In the end, we present the summarization research trends and gaps, which delimits the scope of studies to be explored and applied in our task.

Chapter 3 contains the methodological aspects of how to perform Named Entity Recognition, aiming to classify medical concepts within the clinical text to support the summarization task. The chapter presents the following phases:

- The construction of a semantically annotated corpus focusing in pt-br clinical NLP tasks.
- Implementation of a NER algorithm.
- Utilization of a new state-of-the-art clinical POS-Tagger as a feature to NER.

Chapter 4 presents the methodological path through distinct text summarization approaches and methods in order to perform the clinical summarization, including:

- Unsupervised approaches exploration.
- The annotation of a summarization corpus based on chronic disease patients.
- Development of a semantic similarity algorithm.
- Supervised approaches exploration.

Chapter 5 presents the evaluation results, and *Chapter 6* shows the discussions about the thesis and revisits the research goals and questions. *Chapter 7* wraps-up this thesis with the concluding remarks, study limitations, research contributions and future work.

1.4 PUBLISHED WORK

This section lists the studies submitted and published during the period of this Ph.D. project. Most of them illustrate and describe parts of this thesis, sometimes including additional details and experiments that were out of the scope of this project, however, relevant to clinical NLP.

1. TRANSLATION OF UMLS ONTOLOGIES FROM EUROPEAN PORTUGUESE TO BRAZILIAN PORTUGUESE (OLIVEIRA et al., 2016).
2. METODOLOGIAS E FERRAMENTAS PARA ANOTAÇÃO DE NARRATIVAS CLÍNICAS (ANDRADE; OLIVEIRA; BARRA, 2016).
3. Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation (CLAVEAU; OLIVEIRA; BOUZILLÉ; CUGGIA; MORO; GRABAR, 2017).
4. A statistics and UMLS-based tool for assisted semantic annotation of Brazilian clinical documents (OLIVEIRA et al., 2017).
5. USE OF COMPUTATIONAL TOOLS AS SUPPORT TO THE CROSS-MAPPING METHOD BETWEEN CLINICAL TERMINOLOGIES (GOMES et al., 2019).
6. Incorporating multiple feature groups to a Siamese Neural Network for Semantic Textual Similarity task in Portuguese texts (DE SOUZA et al., 2019).
7. Named Entity Recognition for Clinical Portuguese Corpus with Conditional Random Fields and Semantic Groups (DE SOUZA et al., 2019).
8. Automatic Mapping Between Brazilian Portuguese Clinical Terms and International Classification for Nursing Practice (RONNAU et al., 2019).
9. Learning Portuguese Clinical Word Embeddings: a multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task (OLIVEIRA et al., 2019).
10. Exploiting Siamese Neural Networks on Short Text Similarity Tasks for Multiple Domains and Languages (DE SOUZA et al., 2019).
11. SemClinBr - a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks (OLIVEIRA et al., 2020 – in press).
12. Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts (OLIVEIRA et al., 2020).
13. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora (DALLOUX et al., 2020).

14. Towards a Portuguese Neural Language Model for Named Entity Recognition in Clinical Notes (SCHNEIDER et al., 2020 – in press).

1.5 PROJECT FUNDING AND PHD SPONSORSHIP

This thesis was funded by **Philips Healthcare**, through a partnership project with PUCPR called "*NLP for Portuguese Clinical Documents*", which aims to develop resources and algorithms for the identification and extraction of information from clinical texts in Portuguese. Among the sub-projects originating from this partnership is the doctoral research and sponsorship plan called "*Summarization of clinical data*", which focuses on the development of methods for summarizing data from the patient's EHR.

2 BACKGROUND AND RELATED WORK

This chapter introduces, based on literature, the fundamental concepts present in this research as *Electronic Health Records* data characterization; how healthcare providers deal with *Chronic Diseases* and their main concerns; *Automatic Text Summarization* techniques, categorization, and related work; an overview on *Biomedical Summarization* and *Clinical Summarization,* including research trends and gaps*;* and *NLP tools and resources* that could support the development of summarization algorithms.

In section 2.5.2, we exhibit the methodological challenges and approaches regarding Clinical Summarization and delimit the scope of studies that were explored in this thesis.

2.1 ELECTRONIC HEALTH RECORDS

The International Organization for Standardization (ISO) defines EHR as a *digital repository of patient data*, securely stored and exchanged, and with multiple authorized users' access. The EHR contains retrospective, current, and prospective information, and its main goal is to assist with continuing, efficient, and quality integrated healthcare (ISO, 2004).

In some cases, the EHR purpose of providing instant and comprehensible access to patient's health history is somehow opposed by the fact that clinicians still have to read the patient's full medical history, spending a considerable part of their time going through the patient's EHR in order to get a comprehensive overview of the patient, which is a time consuming and error-prone process. That is, although the EHR has complete information on patients' health, which supports healthcare improvement, it *lacks efficient information presentation tools*, which make the decision-making process difficult (THIESSARD et al., 2012; DABEK et al., 2017; KENEI et al., 2018).

Primarily, the EHR is designed to support medical operational needs, and it is employed for patient care planning, documenting the delivery of care, and assessing patient's outcomes. This information has different roles in the decision-making process, not only in patient care level but also in Hospital management (e.g., billing and payments) and health policy (HAYRINEN et al., 2008).

Later, many studies have found that EHR is an important resource for clinical informatics applications, and researchers started to make secondary use of it. The patient data in EHR has been used for tasks such as medical concept extraction, patient trajectory modeling, disease inference, and clinical decision support systems (SHICKEL et al., 2017). This EHR data, aggregated with other health-related data, can support the development of advanced data analytics also (YADAV et al., 2018).

The EHR holds a large amount of information, such as current and past diagnoses, active medications, lab and test results, radiology images, patient demographics, administrative and billing data, immunization dates, vital signs, medical histories, and progress notes. The EHR data is stored in a *structured* or *unstructured* way. Going further, Shortliffe and Barnett (2001) defined medical data as (a) *numerical and measurable*, like temperature, pulse, arterial pressure; (b) *images and graphics*; and (c) *narratives*. Where (a) is considered to be structured, and consequently a more accessible data to manipulate. While (c), due to its textual aspect, is defined as unstructured, known as a difficult type of data to understand and utilize automatically.

Generally, structured data follows a specified data model and value set, restricting the users to only be able to enter or choose pre-determined values, in opposition, unstructured data does not follow a pre-defined set of values, allowing users to enter instead narrative information about data using their own words.

2.1.1 Clinical Narratives

Data stored in the EHR as free text are defined as *clinical narratives* (or clinical notes), examples of this unstructured data are discharge summaries, nursing notes, and consultation notes (PACHECO, 2009).

The clinical narratives represent a great amount of valuable patient information (e.g., diagnosis, symptoms, treatment), according to Kong (2019) approximately 80% of EHR data remains unstructured. Sondhi (2012) claims that usually, the unstructured data are the most important part of EHR.

Despite its importance, deal (manually or computationally) with clinical narratives is challenging due to multiple issues. First, they frequently contain *contradictory information* because each statement indicates the problem understanding at the time it was written, so it is possible to find ambiguous information about a disease/treatment/hypothesis in different notes, for example (ROGERS et al., 2006).

Another issue is the *under-interpretation of information*, Rogers and colleagues (2006) describe the following case: by default, radiologists report only what they are sure on the image. So, a radiology report may have only “There is evidence of moderate osteoporosis in the bony spine”, leaving to the physician to complement with additional interpretations like “(but) there are no osteolytic lesions that might suggest to cancer has spread to the bone”.

Moreover, a similar situation is that *critical information is often left implicit*. For instance, when a drug needs cease due to its side effects (e.g., anemia due to chemotherapy), the link between a side effect and stopping the drug is rarely mentioned in the note (ROGERS et al., 2006).

Redundancy is a well-known problem in clinical narratives (COHEN et al., 2013; ZHANG et al., 2017), and most of it is caused by the copy/paste effect, where the EHR systems allow clinicians to copy and paste text from previous notes to start a new one. Altogether, this leads to extensive, repetitive, and potentially erroneous text with unimportant or obsolete information. The combination of redundant and new information also elevates the cognitive load when summarizing patient notes, particularly in cases where patients have complex conditions and medical histories (FARRI et al., 2012).

Moreover, it is also important to highlight the large amount of *misspelling* and the high usage of *acronyms* in clinical narratives, that conforming to Cohen and colleagues (2013) are main concerns when dealing with clinical free-text.

2.2 CHRONIC DISEASES

By definition, a chronic disease must have one or more of the following characteristics: be a *permanent* disease, leave residual *disability*, a *non-reversible* pathological alteration causes it, patient special training required for rehabilitation, or may be expected to demand a *long period of supervision*, observation, or care. Chronic diseases, also known as *Noncommunicable diseases* (NCDs), are a consequence of a combination of genetic, physiological, environmental, and behavior factors (WHO, 2014).

Some examples of NCDs are diabetes, cancer, cardiovascular disease (like heart attack and stroke), chronic respiratory diseases (such as asthma and chronic obstructive pulmonary disease), and chronic kidney disease (WHO, 2011).

The NCDs are **the major global health problem** and are causing a high number of premature deaths. In 2016, approximately 41 million deaths occurred due to NCDs, accounting for 71% of the overall total of 57 million deaths. (WHO, 2018). In Brazil, the NCDs represented 73% of all deaths, totaling 928 thousand deaths in 2016 (WHO, 2017).

In many countries, **half of the total population has a chronic disease**. This large number makes NCDs a critical social problem, and from the medical perspective, such prevalence classifies certain diseases as epidemic. The NCDs epidemic level leads to devastating consequences for the individuals, families, and communities involved, besides overloading the health systems (MALTA, 2017).

The World Health Organization (2011) estimates that during the 2011-2025 period the cumulative economic losses to low- and middle-income countries caused by NCDs will surpass 7 trillion dollars, almost 500 billion per year.

Investments on coverage improvement, *risk factors surveillance,* and NCDs *data quality* are essential to prevent and control NCDs, aiming to reduce disease incidence and prevalence; retard incapacities and complications; ease disease severity; and life extension with quality (MALTA, 2017).

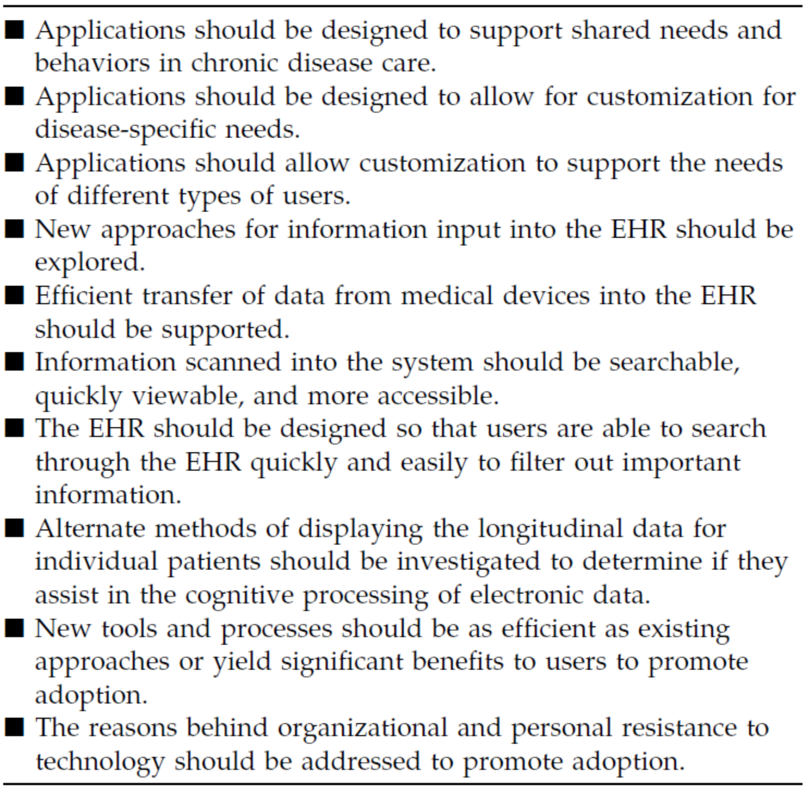
Chronic disease care providers need to seek information through a major collection of data, and the EHR must facilitate this process. Providers also demand to quickly synthesize information to produce a coherent scenario of patient status over time and then to define treatment procedures. Provide a longitudinal view of patient disease history is one of the requirements for Health Information Technology (HIT) systems for chronic disease care. The necessity of practical solutions is especially important in chronic disease care due to expanding patient populations and disease-related complexities (UNERTL et al., 2009).

One of the main inherent EHR features is to describe and document chronic conditions properly (TINETTI et al., 2012; NAVANEETHAN et al., 2013), but in addition to that, a great number of trending tools and summarization methods can be developed to support medical practitioners to deal with NCDs patients, aiming to improve the current NCDs epidemic status someway (ALEKSIĆ et al., 2017; MISHRA et al., 2014;

LAN et al., 2018; YADAV et al., 2018; YOO et al., 2012; WANG et al., 2018; ROGERS et al., 2006).

Moreover, Unertl and colleagues (2009) developed a set of guidelines for Health Information Technology design for chronic disease care (see Figure 4), which were defined during hours of direct observation of medical practitioners and patients during the care process.

Figure 4 - Guidelines for Health Information Technology design for Chronic Disease Care



- Applications should be designed to support shared needs and behaviors in chronic disease care.
- Applications should be designed to allow for customization for disease-specific needs.
- Applications should allow customization to support the needs of different types of users.
- New approaches for information input into the EHR should be explored.
- Efficient transfer of data from medical devices into the EHR should be supported.
- Information scanned into the system should be searchable, quickly viewable, and more accessible.
- The EHR should be designed so that users are able to search through the EHR quickly and easily to filter out important information.
- Alternate methods of displaying the longitudinal data for individual patients should be investigated to determine if they assist in the cognitive processing of electronic data.
- New tools and processes should be as efficient as existing approaches or yield significant benefits to users to promote adoption.
- The reasons behind organizational and personal resistance to technology should be addressed to promote adoption.

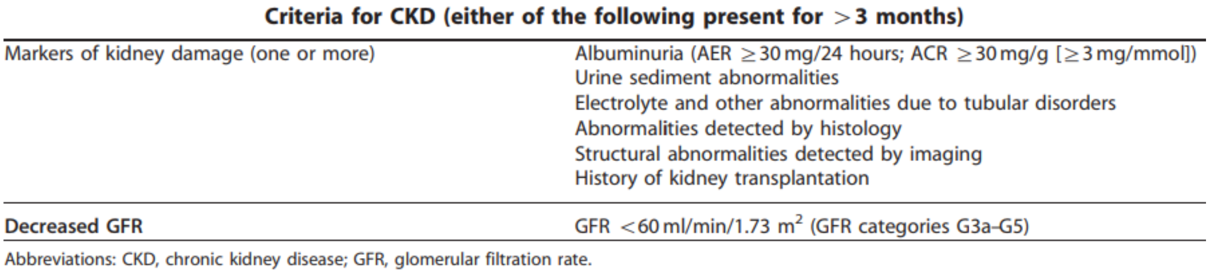
Source: Unertl et al., 2009.

2.2.1 Chronic Kidney Disease

The *Kidney Disease Outcomes Quality Initiative* (KDOQI) is a program from the National Kidney Foundation (NKF) that provides evidence-based guidelines for all stages of *Chronic Kidney Disease* (CKD) and related complications. The definition and classification of CKD proposed by KDOQI in 2002 were endorsed by other major institution called *Kidney Disease: Improving Global Outcomes* (KDIGO) in 2004. Since then, both institutions (KDOQI and KDIGO) collaborate with each other to update definitions and improve guidelines associated with CKD (LEVEY et al., 2010).

KDOQI and KDIGO agree that a CKD is defined as abnormalities of kidney structure or function, present for more than three months, with implications for health. Figure 5 shows the CKD criteria for abnormalities.

Figure 5 - KDIGO and KDOQI criteria for CKD abnormalities



Criteria for CKD (either of the following present for >3 months)

Markers of kidney damage (one or more)	Albuminuria (AER ≥ 30 mg/24 hours; ACR ≥ 30 mg/g [≥ 3 mg/mmol]) Urine sediment abnormalities Electrolyte and other abnormalities due to tubular disorders Abnormalities detected by histology Structural abnormalities detected by imaging History of kidney transplantation
Decreased GFR	GFR <60 ml/min/1.73 m^2 (GFR categories G3a–G5)

Abbreviations: CKD, chronic kidney disease; GFR, glomerular filtration rate.

Source: KDIGO, 2013.

CKD includes conditions that damage the kidneys and reduce their ability to work normally and keep the patient healthy. The patient can develop complications like anemia, weak bones, nerve damage, and high blood pressure due to the high level of wastes not filtered by the kidney. A set of comorbidities are associated with CKD, and those problems may appear slowly over a period of time. CKD may be caused by high blood pressure, diabetes, and other diseases. Rapid detection and treatment commonly stop the progression of the disease. When the disease advances, it can be irreversible (end-stage renal disease), leading to kidney failure, which requires dialysis or kidney transplant (renal replacement therapy) to save the patient (INKER et al., 2014; MARINHO et al., 2017).

In developed countries, the **end-stage renal disease (ESRD) represents a major cost source to healthcare systems**. Dialysis programs had an annual growth between 6% and 12% over the last two decades, and continue to grow. Over 2 million people now require renal replacement therapy to survive worldwide (COUSER et al., 2011).

In Brazil, the prevalence and incidence of CKD increased 2-3 times during the period between 2000 and 2012 (PEREIRA et al., 2016). Silva and colleagues (2016) shown that the Brazilian Healthcare System (Sistema Único de Saúde – SUS) financed 84% of patients in renal replacement therapy, including dialysis procedures, totaling R$ 2 billion in investments. Besides that, 90% of kidney transplantations were paid by

SUS, an expense of R$139.6 million. Another aspect regarding the relation between SUS and chronic kidney disease patients is the SUS fragmented characteristic, which causes a **lack of communication between the primary care and the hospital**, generating a great gap of information when the patient arrives at the specialist.

Moved by the alarming outcomes and high costs associated with kidney failure, the health community increased its attention to the prevalence, prevention, and consequences of earlier and milder forms of renal impairment, in an attempt to reduce deaths and costs associated to CKD (COUSER et al., 2011).

Health Information Technology has an important role in this scenario, as effective use of EHR could support the early identification of CKD patients and improve the quality of care delivered to them. Moreover, systems that facilitate patient management and enhance the physician's time with the patient could help to diminish the disease progress by improving healthcare (NAVANEETHAN et al., 2013).

In our context of CKD patients (see section 4.3 for more details regarding the used data), the importance of computational tools that could reduce the time in which the medical staff access the EHR, and increase their time with the patient is evident. According to the involved doctors of the patients seen, about 80% have the costs paid by SUS and 20% for supplementary health (private health insurance). And while the average time of care for the supplementary health patient is 60 minutes, for the **SUS patient, the consultation time is only 5 minutes**. Therefore, any time that can be saved for this patient could result in a better outcome.

2.3 AUTOMATIC TEXT SUMMARIZATION

In this chapter, we will introduce and formalize important concepts in *Automatic Text Summarization* (ATS); distinguish approaches and classifications; present the steps needed to perform ATS; and then describe the ATS common evaluation methods and metrics.

Conforming to Radev and colleagues (2002), the objective of a text summary is to present informative contents of a document (or a set of them) in a compressed way. A summary built by an ATS system should contain the most relevant information in the original document, and at the same time, it should occupy less space, that is, a text summarizer should provide the meaning of an actual document in fewer words and sentences (GAMBHIR; GUPTA, 2017; SARWADNYA; SONAWANE, 2018).

In ATS research field, we can differentiate the algorithms by input, output, content, among other classifications, all of them presented below.

2.3.1 Single vs. Multi-document

The ATS research started dealing with *single-document summarization*, in which the algorithm receives only one document as input in order to generate a summary. The algorithms often deal with a news story, scientific article, or a lecture (YAO et al., 2017; MORADI; GHADIRI, 2018).

As research progressed, *multi-document summarization* appeared, using multiple documents as input to the algorithm and then clustering these documents to generate a single summary (e.g., cluster news articles on the same event to produce a short summary about the event) (YAO et al., 2017).

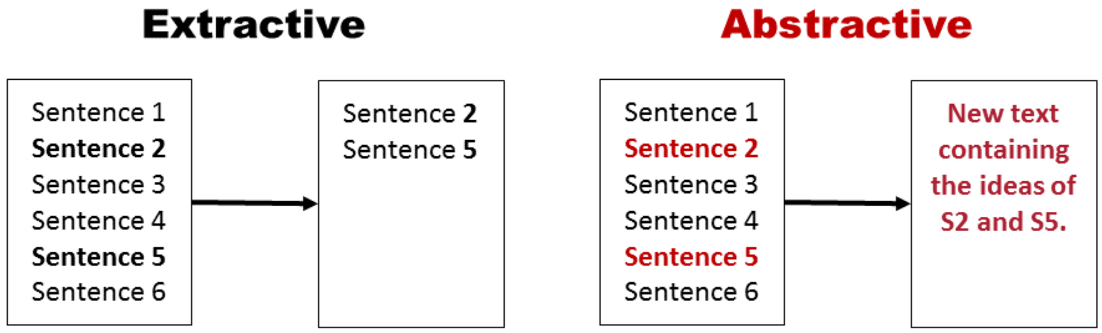
Gambhir and Gupta (2017) state that it is more difficult to summarize multiple documents than a single document due to some known issues like *redundancy*, *co-reference resolution,* and *temporal reasoning,* which are much greater problems when dealing with a set of documents instead of one document only.

2.3.2 Extractive vs. Abstractive

Text summarization methods are classified into extractive and abstractive approaches. *Extractive summaries* are a result of the concatenation of several sentences (or paragraphs and phrases) from the original document(s) being summarized. The summary size depends on the compression ratio. Saliency scores are assigned to each sentence, and then the top-scored sentences are selected to create the summary (NENKOVA; MCKEOWN, 2011; GAMBHIR; GUPTA, 2017; MORADI; GHADIRI, 2018).

On the other hand, *abstractive summaries* describe relevant contents from the original document(s) using different words, that is, an abstract is a summary containing the main ideas and concepts from the original document, but re-interpreted and presented in a different form. An abstractive summarizer uses NLP techniques to process the text, to infer a new version (GAMBHIR; GUPTA, 2017; YAO et al., 2017; MORADI; GHADIRI, 2018). Figure 6 illustrates both approaches.

Figure 6 – Extractive vs. Abstractive summarization.



Source: the author.

2.3.3 Informative vs. Indicative

Summaries can be distinguished by their content also, classified as informative or indicative. *Informative summaries* can replace the original document(s) in order to make the reader understand its content, that is, the reader does not need to access the original input for understanding (NENKOVA; MCKEOWN, 2011; MISHRA et al., 2014). In the clinical context, an informative summary can be used independently of the complete patient record.

Differently, *indicative summaries* give the users an idea of the original content, but they still need to access the input document(s) for understanding (MISHRA et al., 2014). It is like highlighting important parts of the text to the reader (e.g., show important diagnoses and laboratory tests made to the physician).

2.3.4 Generic vs. User-oriented

An ATS system can generate two types of summaries: generic or user-oriented (as known as query-focused or topic-focused). *Generic summaries* represent much of the work done in ATS so far. They simply take a document (or a set of them) to produce a summary, making assumptions about the audience and summary's goal (NENKOVA; MCKEOWN, 2011; MISHRA et al., 2014).

User-oriented summaries show different information depending on user interaction with the system, that is, a specific user query defines the information to be shown (NENKOVA; MCKEOWN, 2011; YAO et al., 2017).

2.3.5 Supervised vs. Unsupervised

Regarding the need of training data, we classify summarizers as supervised or unsupervised. *Supervised approaches* learn from labeled data to select relevant content from documents. Usually, a supervised summarization system is defined as a binary classification problem, with sentences belonging to the summary labeled as positive, otherwise, negative. Then for performing sentence classification, some popular classification methods are used, such as Support Vector Machines (SVM) and Neural Networks (NN). A large amount of annotated data is needed for this kind of approach (GAMBHIR; GUPTA, 2017; MORADI; GHADIRI, 2018). The unsupervised approaches are the majority in the literature, probably because of the cost to obtain annotated data and lack of a reference standard to summarization annotation (MISHRA et al., 2014).

When there is no need for annotated data, we categorize the summarizer as *unsupervised*. In that case, algorithms make use of statistical calculations and/or heuristic rules to extract relevant sentences from the text. Such systems are suitable for any new data added, without any modifications, but usually have inferior results if compared to supervised approaches (GAMBHIR; GUPTA, 2017; MORADI; GHADIRI, 2018).

2.3.6 Summarization Pipeline

In this chapter, we introduce the common summarization pipeline and possible approaches, providing an idea of the steps necessary to create a summary and which methods are available. Some comprehensive summarization surveys guided this review in terms of organization and categorization – see them for an in-depth review of works (NENKOVA; MCKEOWN, 2011; NENKOVA; MCKEOWN, 2012; GAMBHIR; GUPTA, 2017; YAO et al., 2017). Moreover, a newly published survey is also available (EL-KASSAS et al., 2020), with some updates regarding new studies. Also, it contains a complete categorization of summarization methods already developed, which were not explored in this thesis.

Extractive summarization approaches dominated the last decade earlier research, and almost all the systems share some common components/tasks. Nenkova and McKeown (2012), in their survey on extractive summarization techniques, suggest the following typical steps in ATS systems: (1) ***intermediate representation***, (2) ***sentence scoring***, and (3) ***sentence selection***. Yao and colleagues (2017), as their survey scope includes Abstractive summarization, indicate an additional fourth step called (4) ***sentence reformulation***.

In order to build a good summary, algorithms need an ***intermediate representation*** of the input text, containing only the key aspects of the text. Popular summarization methods rely on *Topic representation* (TR) approaches, which convert the text to topics presented in the original document.

We can find different levels of complexity in TR techniques, approaches like *frequency*, *TF-IDF* and *topic-word* comprise a simple table of words and their respective calculated weights; *lexical chains* approaches use a knowledge database like UMLS to find concepts that are semantically related to words, and then give weight to the concepts; in *latent semantic analysis* (LSA) word co-occurrence patterns are spotted and interpreted as weighted topics; *Bayesian topic models* in which the input is transformed in a set of topics, and each topic has his own table of word probabilities.

Differently than TR approaches, *Indicator representation* (INDR) approaches do not represent text as topics. However, they build a representation that can directly rank sentences by importance (i.e., list of importance indicators as sentence length, location in the document, presence of certain information, etc.).

Graph-based methods represent the document as a network of inter-related sentences in which they derive an indicator of sentence importance from sentence centrality in the graph, while other approaches use a variety of indicators and combine them *heuristically or with Machine Learning* to define the sentences to be outputted.

With an intermediate representation derived, the ***sentence scoring*** step takes place, where the sentences are assigned with a score indicating their importance, and the ones with higher scores tend to be selected in the summarization process. In *TR* approaches, the score is often associated with sentences that better represent the most important topics or how well it combines different topics information. For *INDR* methods, each sentence weight is defined by combining different indicators (commonly using ML). One of the most known *graph-based* algorithms, the LexRank (ERKAN; RADEV, 2004), applies stochastic techniques to the graph representation to calculate the weight of each sentence.

The next step to build a summary is ***sentence selection***, where the summarizer selects the best combination of important sentences to form a summary. The simplest approach is to select sentences with higher scores directly. However, many issues make this step more complex to solve, especially in multi-document summarization. One of these issues is the *redundancy* because a good summary should not contain repeated information.

One popular approach for sentence selection is *maximum marginal relevance* (MMR), where an iterative greedy procedure selects the sentences by recomputing the importance scores using a linear combination between the original weight and its similarity with already selected sentences, that is, sentences that are equivalent to already chosen sentences are discarded.

In *global optimization* approaches, a set of constraints are defined in order to select the optimal collection of sentences. These constraints try to maximize informativeness, minimize repetition, and sometimes maximize coherence of the summary, all of this following the required summary length.

When it comes to Abstractive summarization, one additional step called ***sentence reformulation*** is incorporated into the summarizers. Here, the extracted sentences are reformulated using *rule-based* approaches or more sophisticated methods like *paraphrasing* and *sentence fusion* to improve summary informativeness and compactness. Due to the immaturity of the current Natural Language Generation (NLG) research field, some of these steps may damage summary readability.

In multi-document summarization, an issue that influences the summary quality is the *sentence ordering*. One classic reordering approach is to infer order from a *weighted sentence graph*. Execute a *chronological ordering algorithm*, sorting sentences based on timestamp and position is another option.

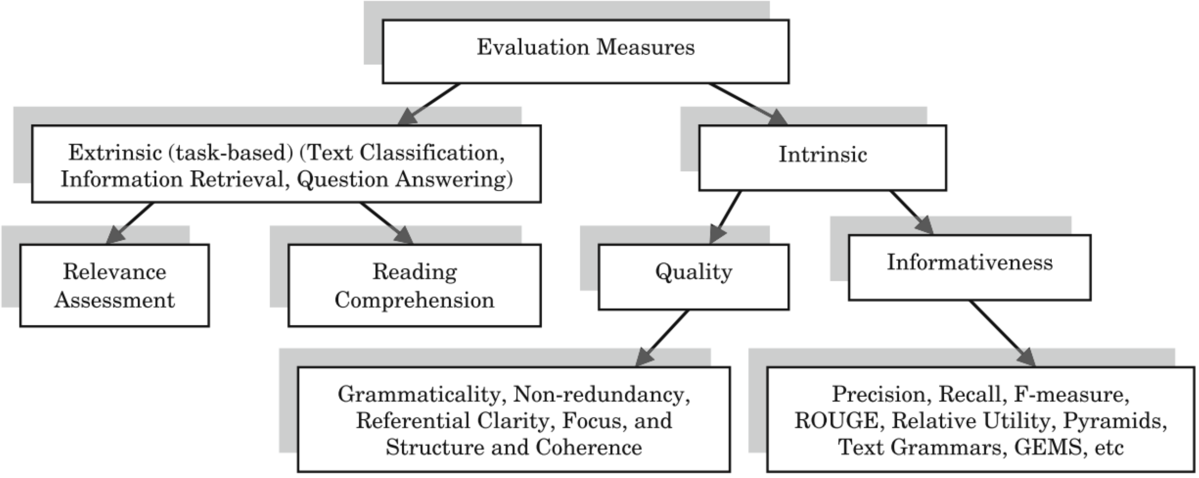
2.3.7 Evaluation Methods and Metrics

Evaluation is essential for ATS development. Although, it still remains uncertain which evaluation criteria should be used for assessing summarization systems due to the subjectivity to define what makes a summary good. Given this subjectivity and the diversity of evaluation settings and types, evaluating a summary turns into a very challenging problem. Lloret et al. (2018) made a full overview about this matter. They present all the evaluation metrics, methods, and datasets available, point out the strengths and weaknesses of each method, and discuss the major challenges to define a good evaluation setup.

Manual ATS evaluation is laborious and time-consuming. Therefore, a set of evaluation methods have been proposed to automate the evaluation process (fully or partially) and fulfill the need for fast and consistent summary evaluation (GAMBHIR; GUPTA, 2017; YAO et al., 2017).

The two ways of assessing ATS performance are *intrinsic* and *extrinsic* evaluation. Figure 7 shows the complete taxonomy of summary evaluation measures included in these two evaluation categories.

Figure 7 - Taxonomy of summary evaluation measures


Evaluation Measures
Extrinsic (task-based) (Text Classification, Information Retrieval, Question Answering)
Intrinsic
Relevance Assessment
Reading Comprehension
Quality
Informativeness
Grammaticality, Non-redundancy, Referential Clarity, Focus, and Structure and Coherence
Precision, Recall, F-measure, ROUGE, Relative Utility, Pyramids, Text Grammars, GEMS, etc

Source: Gambhir; Gupta, 2017.

Extrinsic evaluation measures a summary quality based on how it impacts other tasks (e.g., Information Retrieval, Question Answering, Decision Support). If a summary provides help with other tasks, it is defined as a good and effective summary, which in the clinical domain, for instance, could indicate if a summary can improve patient outcome, can save the health practitioners time or improve decision making (some of these are very subjective to assess). Thus, effectiveness measurement depends on the summarizer domain and purpose. To the best of our knowledge, there is no standard evaluation to perform summarization of patient data, for example.

For instance, we can evaluate extrinsically by comparing how fast and accurately a clinician can identify eligible patients for a clinical trial, with and without the use of a summary.

It is possible to address extrinsic evaluation using *questionnaire-based* and *self-report* approaches (e.g., Hunter et al., 2012), but the evaluation often relies on some specific performance measure or achievement collected *in loco* during the system's use (e.g., decision-making accuracy, impact on patient outcome, usability, accomplishment time, success rate) (GATT; KRAMER, 2018).

According to Gambhir and Gupta (2017), extrinsic evaluations are divided into two categories: *relevance assessment* and *reading comprehension*. The first one uses a set of methods to assess a topic's relevance present in the summary of the original document. The second use multiple-choice tests after reading the summary.

The ***intrinsic evaluation*** assesses the "goodness" of a summary output according to certain criteria, like readability, comprehensiveness, accuracy, and relevancy. Output summaries can be rated by users or compared with a gold standard, typically hand-crafted by humans (MISHRA et al., 2014).

The most common way to evaluate summaries generated by an automatic summarizer is to compare them against summaries generated manually. In such an evaluation method, the similarity between the content of the system and model summaries are estimated. The performance is better when there is a great overlap of content between the system and model summary. Nevertheless, to obtain manually made summaries is a time-consuming and challenging task because they have to be provided by human experts. Furthermore, human-generated model summaries are highly subjective (MORADI; GHADIRI, 2018).

The two main aspects in which the summaries are evaluated are *quality* and *informativeness*. Several methods cover the *informativeness* evaluation, such as Precision/Recall/F-Measure, ROUGE, Relative Utility, and Pyramid (GAMBHIR; GUPTA, 2017).

Most summarizers select relevant sentences to create an extractive summary. In such summaries, the well-known Information Retrieval metrics like *Precision, Recall and F-Measure* can be used to evaluate a summary, as a human can build a gold standard of representative sentences and then the system can automatically compare with the computer-generated ones. Yet, Nenkova and McKeown (2011) describe a set of issues with these measures like *human variation*, *sentence granularity,* and *semantic equivalence*.

Currently, ROUGE (Recall-Oriented Understudy for Gisty Evaluation) (LIN, 2004) is the most used and the standard for summarization automatic evaluation. The ROUGE metrics compare n-grams between the evaluated summary and one or several human-made reference summaries. ROUGE method has five different measures, including ROUGE-N (n-grams), ROUGE-L (the longest common sequence), ROUGE-W (weighted longest common sequence), ROUGE-S (Skip-Bigram) and ROUGE-SU (skip-bigrams and uni-grams).

Based on the aforementioned ROUGE metrics, the system calculates both Precision and Recall, where the Precision is the percentage of n-grams in the generated summary that also occur in the gold-standard summary, and the Recall is the percentage of n-grams in the gold-standard that also occur in the generated summary.

ROUGE-1, ROUGE-2, and ROUGE-SU metrics have been extensively used in NLP studies and shared-tasks for automatic summary evaluation. ROUGE-1 and ROUGE-2 compute the number of unigrams and bigrams that are shared by the generated and the gold-standard summaries, respectively. ROUGE-SU4 estimates the overlap of skip-bigrams between the generated and the gold-standard summaries, allowing a skip distance of four.

Some researchers like to supplement the use of ROUGE with a manual evaluation such as Pyramid method (NENKOVA; PASSONNEAU, 2004), at least in a small subset of test data (NENKOVA; MCKEOWN, 2011). Lloret et al. (2018) suggest that the Pyramid scheme offers a way to reduce disagreement between humans when selecting relevant sentences. Gatt and Krahmer (2018) recommended using multiple methods to evaluate a summary, and when possible, report not only their results but also the correlation between them.

It is possible to evaluate *quality* besides informativeness, in that case, linguistic aspects are considered. For instance, in DUC[1] (Document Understanding Conferences) and TAC[2] (Text Analysis Conference) conferences, questions based on linguistic quality and related to non-redundancy, focus, grammaticality, referential clarity and structure are applied to human experts to manually assign a score to each aspect. Another possibility to measure quality is to analyze readability factors (PITLER; NENKOVA, 2008) (e.g., vocabulary, syntax, discourse) so that a correlation can be calculated between these factors and already obtained human ratings. Other possible quality evaluation methods are local coherence, centering theory, syntactic and semantic models, and grammaticality of a grammar.

[1] https://www-nlpir.nist.gov/projects/duc/

[2] https://tac.nist.gov/

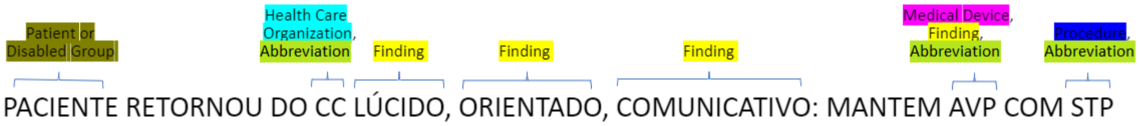
2.4 NATURAL LANGUAGE PROCESSING TOOLS AND RESOURCES TO SUPPORT TEXT SUMMARIZATION

Considering the typical steps of an ATS algorithm pipeline, several approaches make use of additional NLP tools and resources to assist the summarization, mainly in the intermediate representation step, where the main aspects of the text are extracted and used to build a new representation.

Information Extraction (IE) is the task of extracting structured information from unstructured data by applying NLP techniques. The **Named Entity Recognition** (NER) is among the sub-tasks of IE, it extracts and identifies entities (e.g., medical concepts, persons, organizations) in the middle of the text (see Figure 8). It could be used as a first step, not only for an ATS algorithm, but for many NLP tasks as topic modeling, question answering, information retrieval, co-reference resolution, sentiment analysis, and others (YADAV; BETHARD, 2018). The NER is considered a sequence labeling problem, as the entity type of a concept is defined by its context and by the entity type of its neighbors (THIYAGU; MANJULA; SHRIDHAR, 2019).

The main approaches to build a NER system are (i) rule-based methods that utilize hand-crafted rules and gazetteers, (ii) supervised machine learning classifiers (e.g., conditional random fields, support vector machines) based on an annotated corpus and lexical, syntactic, and semantic features, (iii) neural networks (e.g., recurrent neural network, convolutional neural network), which usually require more annotated data than traditional classifiers, however, needs minimal feature engineering, and (iv) unsupervised approaches, which do not need labeled data. It is worth noting that the systems with top performances are mostly based on supervised approaches (WU et al., 2017; THOMAS; SANGEETHA, 2020; LI et al., 2020).

Figure 8 – Named Entity Recognition algorithm applied to a clinical text to identify the categories of each concept found within the text.


Patient or Disabled Group
Health Care Organization, Abbreviation
Finding
Finding
Finding
Medical Device, Finding, Abbreviation
Procedure, Abbreviation
PACIENTE RETORNOU DO CC LÚCIDO, ORIENTADO, COMUNICATIVO: MANTEM AVP COM STP

Source: the author.

Before the advent of modern neural networks architectures (as the Recurrent Neural Networks), the studies that applied the **Conditional Random Fields (CRF)** classifier (LAFFERTY; MCCALLUM; PEREIRA, 2001) were the top performers regarding the NER task. The CRF is a framework for building probabilistic models to segment and label sequence data.

The use of neural network approaches is a trend regarding sequence-labeling tasks, in which the **Recurrent Neural Network (RNN)** is often used, due to its ability to maintain a memory based on past information, enabling the model to predict the output considering long-distance features (HUANG; XU; YU, 2015). Furthermore, RNNs only use the previous context, while bidirectional (Bi) RNNs also explore the future context, processing the data in both directions with separated RNNs (Graves et al., 2013), improving the performance of labeling tasks. Long short-term memory networks (LSTM) are RNNs that contain memory cells, which improve the model's ability to find long-range dependencies within the text (HUANG; XU; YU, 2015). Thus, **Bi-LSTM networks**, which are RNNs with a bidirectional aspect, are widely used for sequence labeling tasks.

Akbik and colleagues (AKBIK; BLYTHE; VOLLGRAF, 2018) produced state-of-the-art results for many sequence labeling tasks in multiple languages. They used Huang et al.'s (2015) architecture, a Bi-LSTM with a CRF layer (**BiLSTM-CRF**), incorporated with a novel character contextual embeddings, instead of distributional word-level vector representations (e.g., word2vec).

Vaswani introduced a novel architecture called **Transformer**, which used feed-forward networks and attention mechanisms (VASWANI, et al., 2017) to ease some of the RNNs issues. Then, researchers from Google developed an unsupervised learning architecture on top of the Transformer architecture called **BERT (Bidirectional Encoder Representations from Transformers)** that outperformed almost all existing NLP models in a wide range of tasks (DEVLIN et al., 2019). They also published several pre-trained models that could be used for transfer learning on a multitude of different domains and tasks, significantly reducing the need for high-end computational resources in the development of NLP methods when using these pre-trained models.

To input an annotated corpus into a NER model, the data is transformed following a tagging schema, such as IO, IOB2, IOBES, etc. The main need for using the schemes is to characterize the taxonomy of words and separate occurrences followed by the same semantics by two different entities, in other words, tell the algorithm where the concepts begin and end within the text. Table 1 presents a sentence tagged using three different schemas.

Table 1 - Tagging schemes for sequence labeling. Lucas is tagged as PERSON, shopping mall as LOCATION, and Curitiba as LOCATION.

Schema	Lucas	went	to	the	shopping	mall	in	Curitiba	.
IO	I-PER	O	O	O	I-LOC	I-LOC	O	I-LOC	O
IOB2	B-PER	O	O	O	B-LOC	I-LOC	O	B-LOC	O
IOBES	S-PER	O	O	O	B-LOC	E-LOC	O	S-LOC	O

Source: the author.

Stanislawek et al. (2019) analyzed recent NER models and pointed out their weak and strong aspects, which should be taken into account when developing a solution. Huggard et al. (2019) discussed the feature importance in a biomedical NER. They experimented on various word representations features (i.e., **Word Embeddings**), and additional orthographic, lexical and syntactic features, in the end, the Word Embeddings, trained on a biomedical corpus, performed best and were more significate to the task.

Another important feature to be considered when performing NER is the **Part-of-speech tagging** (POS-Tagging), which gives the words morphological meaning by labeling them with tags such as nouns, verbs, adjectives. As the NER systems, the POS-Tagging algorithms also require supervised sequence labeling methods, supported by a domain-specific annotated corpus. A recent study (GÜNGÖR; GÜNGÖR; ÜSKÜDARLI, 2019), discussed the effect of morphological features in NER considering morphologically rich languages. The experiments have shown that augmenting morphological features increases the NER performance.

Regarding the Biomedical domain, the NER is a fundamental task, since it extracts specific entities of interest, aiming to support both clinical practice (e.g., decision support systems, hospital processes optimization) and medical research (e.g., clinical trials, pharmacovigilance, molecular biology). It is not trivial to adapt existing general domain NER systems to biomedical, due to several issues concerning the biomedical content (e.g., non-standard jargon, speculative language), and because of that, the biomedical and clinical NER are still a very challenging task (THIYAGU; MANJULA; SHRIDHAR, 2019; WU et al., 2017).

The clinical NLP community developed and made available various tools to process the clinical narratives, such as MetaMap (ARONSON; LANG, 2010), cTakes (SAVOVA et al., 2010) and CLAMP (SOYSAL et al., 2018). All of these tools provide, among other features, the possibility to perform NER in clinical texts. Unfortunately, they support almost exclusively English language texts. In pt-br, there are just a few attempts to perform clinical NER (SANTOS et al., 2019; LOPES; TEIXEIRA; OLIVEIRA, 2019; LOPES; TEIXEIRA; OLIVEIRA, 2020), in which very specific NER corpus was used (e.g., clinical narratives from Neurology).

In fact, the necessity of a **gold-standard corpus** is one of the major bottlenecks regarding the clinical NLP algorithms, such as NER, POS-Tagging, and even the summarization itself. Because, in addition to the expense and issues associated with annotation projects (XIA; YETISGEN-YILDIZ, 2012), there is still a prevalence of biomedical literature corpora over clinical data (ROBERTS et al., 2009; XIA; YETISGEN-YILDIZ, 2012; BRETONNEL COHEN; DEMNER-FUSHMAN, 2014). While the biomedical usually is composed of open scientific data like scientific papers and gene data, the clinical utilizes patient's data from the EHR, which require ethical committee approval and an anonymization process in order to use and release the data. To the best of our knowledge, until the beginning of this research project, there was no pt-br clinical gold-standard corpus available for both NER and summarization tasks. Regarding the POS-Tagging, our research group developed an annotated corpus for pt-br (PETERS et al., 2010); however, no trained model was available.

2.5 BIOMEDICAL AND CLINICAL SUMMARIZATION

When the input text comes from a particular domain, has a known and specified structure, or certain unique properties, the summarization algorithms can benefit from these characteristics to improve the important information identification (NENKOVA; MCKEOWN, 2011). Hence, to perform ***biomedical summarization***, special attention to biomedical data particularities is needed in order to create optimal summaries. The main divergence between generic ATS and biomedical summarization is in the nature of data, unlike in ATS, biomedical data (mostly the EHR data) is a mixture of unstructured plain text and semi-structured data (DEVARAKONDA et al., 2014).

For the sake of this research project scope and objectives, we focused on patient's data summarization (for now on, called *clinical summarization* or *EHR summarization*). When convenient, studies related to biomedical literature summarization are cited as well. In the following sections of this chapter, we introduce important notions of clinical summarization (including a conceptual clinical summarization framework proposed by Feblowitz et al.); present an overview of biomedical summarization works that are relevant to this study; describe the main methodological challenges in each dimension of the task; and discuss the current research trends and gaps

Feblowitz and colleagues (2011) defined ***clinical summarization*** as "*the act of collecting, distilling, and synthesizing patient information for the purpose of facilitating any of a wide range of clinical tasks*". Moreover, they divide clinical summarization into three interrelated categories: source-, time-, and concept-oriented views. The ***source-oriented view*** (the most common one) organizes the information according to its origin (i.e., information grouped into categories such as medications, laboratory results, and imaging). In ***time-oriented view,*** the information is displayed chronologically (e.g., as a timeline), presenting a sequence of events or details of the care plan. And finally, the ***concept-oriented view***, where the information is assembled around clinical concepts (e.g., medical problems, organs), usually this view requires significant use of medical expertise or knowledge databases.

Another aspect discussed is about the *extractive summaries*, which in the clinical domain are considered ***patient-state independent*** or *knowledge-poor* because they do not need any knowledge-based interpretation or advanced clinical expertise of the patient's state. In contrast, *abstractive summaries* require this kind of knowledge and are considered to be ***patient-state dependent*** or *knowledge-rich*.

2.5.1 AORTIS Model

The AORTIS model (Feblowitz et al., 2011) was developed based on the idea that the creation of clinical summaries can be modeled with five steps: *aggregation*; *organization*; *reduction* and *transformation*; *interpretation*; and *synthesis*. And any of these steps could be executed by a computer or a clinician, sequentially, or not even applied to a particular summarization scenario (see their study for examples).

Aggregation is the collection of clinical data from different available sources, including numerical, structured, and unstructured data. In ***organization*** step, the data are structured (e.g., grouping, sorting) without altering the data values itself. At this point, the data needs to be *condensed* to reduce information overload and ease comprehension. This is possible in two possible ways.

The first condensation possibility is ***reduction***, which is the process of selecting salient information to decrease the amount of data (e.g., get most recent lab results, minimum/maximum values, notes of a certain range of time or specific category). The second is ***transformation,*** which facilitates understanding by altering data density or data view (e.g., graphical or textual display of lab results, metaphor graphics, timelines).

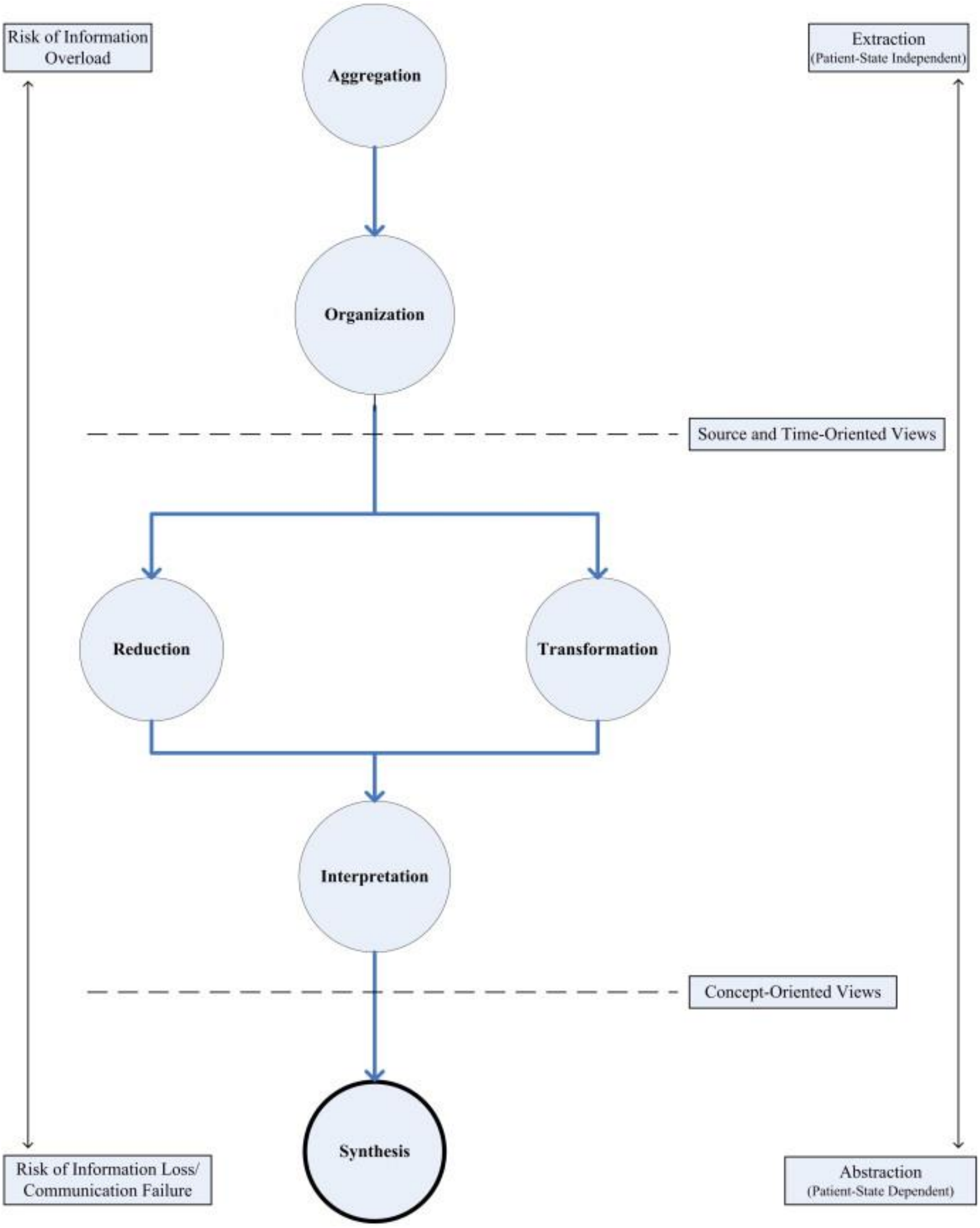
Interpretation is the ability to analyze clinical data using general medical knowledge. One example would be showing abnormal lab results in the patient's summary by consulting a knowledge base containing these abnormality ranges.

Synthesis would be the last and more sophisticated step, where two or more data items are combined with knowledge-based interpretation of patient state in order to give meaning or suggest action. For instance, an abnormal lab result is interpreted and synthesized to produce a warning like "In response to elevated *lab-result* last month, a *medication* was initiated and *lab-result* return to normal levels".

As shown in Figure 9, the *risks and benefits* change through the framework stages. For instance, after aggregation, there is a maximal risk of data overload because all available data is present, while on the following stages, when data is removed or updated, the risk of overload is reduced, however, the risk of information loss, erroneous interpretation and communication failure increases.

It is important to note that the initial stages of *aggregation* and *organization* do not need any clinical knowledge. In opposition, *interpretation* and *synthesis* require knowledge and clinical understanding of the patient state. The middle stages (*reduction* and *transformation*) may need a limited general knowledge base.

Figure 9 - The AORTIS model


Risk of Information Overload
Aggregation
Extraction (Patient-State Independent)
Organization
Source and Time-Oriented Views
Reduction
Transformation
Interpretation
Concept-Oriented Views
Risk of Information Loss/ Communication Failure
Synthesis
Abstraction (Patient-State Dependent)

Source: Feblowitz et al., 2011.

Some studies already take advantage of using the AORTIS model. Laxmisan and colleagues (2012) compared clinical summarization capabilities of various EHR systems under the stages of AORTIS model. Aleksić et al. (2017) developed an extension of existing data structures of an EHR system and introduced a summarization method for chronic disease tracking. Pivovarov and Elhadad (2015) presented challenges in summarization within the context of AORTIS framework. Devarakonda et al. (2014) argued that they did not use AORTIS because its abstraction works well only in a single-patient problem, which was not the case in their study.

2.5.2 Methodological Challenges and Approaches

Pivovarov and Elhadad (2015) reviewed summarization methods for EHR data, focusing on patient longitudinal data. Besides the discussion about current work, they present a set of methodological challenges in clinical summarization (Table 2). They contextualize the discussion around the summarization stages defined in AORTIS conceptual model. We start this section with an overview of their findings and, when necessary, present complementary and updated information. After that, we expand the current challenges by exploring a set of topics and studies which are relevant to this project.

Table 2 - Pivovarov and Elhaddad (2015) challenges in clinical summarization and the related AORTIS stages

Challenge	AORTIS steps
Identifying and aggregating similar information	Aggregation
Organizing and reasoning over temporal events	Organization; Interpretation
Accounting for and interpreting missing data	Organization; Interpretation

Reducing information to only the most salient	Reduction;
	Transformation;
	Synthesis

Using existing clinical knowledge	Interpretation; Synthesis
Deploying summarization tools into the clinic	-

Source: adapted from Pivovarov and Elhadad, 2015.

Identifying and aggregating similar information: one key concept around successfully aggregating data from EHR is *similarity recognition*. The UMLS is a clinical knowledge database that provides help with the similarity issue, because it has words grouped into concepts, and the words that share lexical similarity are mapped to the same concept. Particularly English-language studies, where UMLS has greater vocabulary coverage and mapping tools available (e.g., MetaMap, cTakes, MedLEE) than other languages, the normalization of words to concepts is a well-known solution (ZHANG et al., 2011; DEVARAKONDA et al., 2014; DI EUGENIO et al., 2014; HIRSH et al., 2014; JOHNSON et al., 2015; KIM; LEE, 2017). In fact, some studies with other languages still use UMLS, even with smaller coverage and more straightforward methods doing the mapping, like Viani et al., 2017, that developed a summarizer for Italian medical reports.

Mapping words from the text to concepts is very challenging in pt-br, as we do not have any tool available in the clinical domain, relying only on more simple approaches like exact-matching, regular expressions, minimum edit distance, and hand-crafted rules. It is possible to find concepts in the text without mapping them to terminology like in Chalapathy et al. (2016), where they perform clinical concept extraction using a recurrent neural network or using general-purpose methods for term extraction in texts (DA SILVA CONRADO et al., 2014). The main disadvantage of doing it is to abdicate all clinical knowledge available in terminologies such as concept hierarchy, relationships, synonyms, descriptions, and threshold values.

Due to the formulaic aspect and the copy-and-paste habit, clinical reports very often contain redundant spans of text. The *redundancy* problem is even more significant when dealing with patient longitudinal information (i.e., multi-document summarization). Possible options from ATS research to resolve the redundancy are: identify similarity (word- and statement-level) to find redundant spans of text and remove them, like the general CopyFind tool (HIRSCHTICK, 2006); and Maximal Marginal Relevance (MMR) approach (CARBONELL; GOLDSTEIN, 1998).

Sometimes even different concepts have a certain level of semantic similarity (e.g., epilepsy [C0014544] and seizure [C0036572] UMLS concepts). In well-defined domains, it is possible to use a *knowledge-based* approach, relying on semantic relations and concept definitions available in ontologies like UMLS and SNOMED-CT (PIVOVAROV; ELHADAD, 2012). Examining the similarity between concepts' definitions is another possibility (more details in Pedersen et al., 2007; Pesquita et al., 2009 and Zhang et al., 2017).

The *knowledge-free* methods consider the hypothesis that concepts that appear in similar contexts are similar, to compare them in a *vector space* (i.e., word embeddings), which is an excellent option when there is few ontological knowledge available. This approach has been used multiple times in biomedical NLP (PEDERSEN et al., 2007; PIVOVAROV; ELHADAD, 2012; KOOPMAN et al., 2012; DE VINE et al., 2014; WU et al., 2015; TULKENS et al., 2016; YU et al., 2016; ELEKES et al., 2018).

Organizing and reasoning over temporal events: to perform time-dependent clinical summarization is essential to automatically *identify, extract*, *order*, *reason* over temporal clinical events. Sultanum et al. (2018) state that temporal awareness is fundamental for clinical understanding and decision-making, particularly when creating tools for clinical text overview. Especially in *chronic disease management*, Samal and colleagues (2011) concluded that incorporating time-oriented views should reduce costs, improve efficiency and quality of care. However precise patient's temporal representation remains a challenging task (SUN et al., 2013b; SULTANUM et al., 2018), even with several initiatives towards the progress of clinical temporal reasoning (SUN et al., 2013a; BETHARD et al., 2015; BETHARD et al., 2016; BETHARD et al., 2017).

It is worth noting that different clinical notes have distinct temporal relationships (e.g., nursing notes often contain information regarding one specific moment in patient's history, while a discharge summary describes all inpatient hospital stay and future directions of care). This may explain the fact that most multi-document summarization studies work with homogeneous documents. (ACHARYA et al., 2018)

Styler and colleagues (2014) pointed four challenges with extracting temporal information in clinical data: (a) diversity of time expressions; (b) complexity of determining temporal relations among events; (c) the difficulty of handling the temporal granularity of an event; and (d) general NLP issues.

Besides the extraction, *temporal ordering* is challenging as well, mainly because of imprecise wording and the necessity of clinical knowledge to check how long a condition will persist (e.g., chronic vs. transitory condition). Many studies have to deal with these issues in order to build time-dependent summaries/histories (HIRSH et al., 2014; DABEK et al., 2017; VIANI et al., 2017; LONG; YUAN, 2017; GUO et al., 2018).

Accounting for and interpreting missing data: besides the large amount of data present in EHR, lots of information are missing, because normally the clinical documentation only occurs when the patient is seen by a health practitioner. Thus, everything that happens between medical visits remains out of the records. Moreover, in the Brazilian health system context, the lack of communication between primary care and specialists increases the gap of missing data. Despite the importance of interpreting and determining missing data saliency, there is no research in clinical summarization focused on this issue, while in the Statistics research area, missing data is an active topic (e.g., DING; LI, 2018).

In the clinical domain some attempts to infer health status and trend lines using missing data (WEBER; KOHANE, 2013; POH; DE LUSIGNAN, 2011); and estimate duration of events were performed (e.g., Van Vleck; Elhadad, 2010 and Perotte; Hripcsak, 2013 – the second one classified ICD-9 codes into chronic and acute conditions also).

Reducing information to only the most salient: finding salient information has been deeply studied in ATS. The algorithms usually rely on: term-frequency; TF-IDF; document, syntax or discourse structure; heuristic rules; centroid clusters; graph-based centrality; probabilistic models; hand-crafted features for machine learning; and more recently, word vectors and neural networks (for a more detailed review of techniques refer to NENKOVA; MCKEOWN, 2011; NENKOVA; MCKEOWN, 2012; PIVOVAROV; ELHADAD, 2015; GAMBHIR; GUPTA, 2017; YAO et al., 2017; EL-KASSAS et al., 2020).

How to translate these approaches to the clinical domain and define which one suits better to each task remains unclear since clinical data is very unconventional, and some of the approaches have not been used in the clinical summarization yet. In summary, connect EHR’s textual observations to high-level clinical abstractions is one of the biggest challenges in Health Informatics research. However, despite being a substantial challenge, it is easy to affirm that all salient elements in the patient record are associated with *patient health status and how it changes through time.*

Using existing clinical knowledge: the use of current health terminologies and ontologies provides handy resources to solve *similarity*, *temporality,* and *salience* challenges in the clinical summarization area (i.e., a significant part of the previously discussed challenges). Usually, these clinical knowledge representations are used to generate rules and heuristics in order to capture information, and they are a current trend in clinical summarization.

It is important to highlight that sometimes it is not possible to use these repositories due to language restrictions. For example, SNOMED-CT still does not have a Portuguese version. The UMLS 2018AA[3] release has 70.68% of concepts in English, against only 2.5% in Portuguese. To overcome the lack of semantic/lexical resources, it is possible to focus on resource-lean and language-independent methods (e.g., Moen et al., 2016).

Deploying summarization tools into the clinic: it is a consensus that clinical summarization research needs to go towards an extrinsic evaluation in order to assess the summarization effectiveness in downstream tasks (MISHRA et al., 2014; PIVOVAROV; ELHADAD, 2015; MOEN et al., 2016).

Mishra and colleagues (2014) suggest that the scarcity of studies performing extrinsic evaluation is a reflex of the immaturity of biomedical summarization research. None of the selected studies have been deployed and assessed in patient care settings or in actual research applications. Hence, so much more attention is needed to studies focusing on summarizer’s deployment in real patient care settings and evaluating the impact of it on decision-making and patient outcome.

Going abstractive: Lexmisan and colleagues (2012) compared the clinical summarization ability of twelve EHR systems and concluded that most of them use less sophisticated summarization techniques such as *aggregation* and *organization* rather than an *interpretation* of information using clinical rules and the *synthesis* of recommendations for further action. They claim that when a summarizer goes further into advanced stages of clinical summarization, it will provide more significant help in clinical decision support and patient outcome prediction.

However, it is well-known in ATS research field that abstractive summarization is generally much more difficult than extractive, including sophisticated techniques for meaning representation, content organization, surface realization, etc. (YAO et al., 2017).

Abstractive approaches can have an extra barrier when it comes to pt-br language. Because to create new texts, the summarization algorithm relies on NLP techniques (e.g., POS-Tagger), which are dependent on linguistic resources (e.g., morphosyntactic corpora), which are scarce specifically in the biomedical domain. But despite that, it is possible to find initiatives towards Natural Language Generation

[3] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

(NLG) in pt-br (DE NOVAIS; PARABONI, 2013; OLIVEIRA; SRIPADA, 2014), although not evaluated with biomedical text generation.

Comparing baseline scores: there is no evidence on which method is better to perform biomedical summarization. And one of the reasons is because it is very difficult to compare different studies due to the lack of widely used standard evaluation methods; and a generalizable biomedical summarization corpus with the potential of being used by the research community (MISHRA et al., 2014).

The shared-tasks/challenges environment can provide the possibility for researchers to compare their algorithms with other baselines. But regarding the biomedical summarization field, there were only a few initiatives (e.g., TAC 2014 Biomedical Summarization Track[4], BioASQ Biomedical Semantic QA tasks[5]), unlike other areas of biomedical NLP. Further research is needed to allow publicly available summarization corpora to assist the development of summarization tools.

Summarizing chronic disease information: dealing with longitudinal clinical data is a great challenge itself, but when we focus on chronic conditions, a new level of complexity is revealed. Mainly because chronic disease patients have a lot more data than regular patients, leading to information overload, which increases some issues like redundancy. Furthermore, chronic conditions remain during longer spans of time, and often include patients with multiple comorbidities associated, requiring additional sophisticated techniques to handle the data, especially the temporal aspects (GOLDSTEIN; SHAHAR, 2016; ALEKSIĆ et al., 2017). Previous work already investigated the particularities of chronic diseases in health care. Table 3 lists studies performing an analysis of chronic disease data.

Important remarks can be extracted from these studies. Just to name a few: the *Guidelines for HIT Design for Chronic Disease Care* defined by Unertl et al. (2009) should be taken into account when developing tools to assist chronic disease care; Reichert and colleagues (2010) make clear that the main sources of information for chronic disease conditions are the clinical narratives, laboratory and radiology results.

As there is no definition in the research literature of which method is most suitable for clinical summarization, and there is also no methodological basis for comparing existing methods, since each has been tested in different protocols and

[4] https://tac.nist.gov//2014/BiomedSumm/

[5] http://bioasq.org/participate/challenges

databases, we have chosen to select studies who have already summarized patients with chronic disease and some studies that, because they are not dependent on resources that do not exist for the pt-br, can be experienced in our context.

Table 3 - A summary of studies analyzing chronic disease data.

Author(s), year	Condition(s)	Study overview

Unertl et al., 2009	Multiple sclerosis;
	Cystic fibrosis;
	Diabetes mellitus
		The objective of this study was to understand the work practices, workflow, and information flow in chronic disease care, to support the development of Health Informatics Tools (HIT). The result was a set of *Guidelines for HIT Design for Chronic Disease Care* (previously presented in Figure 4).

Reichert et al., 2010	Chronic Kidney Disease	Physicians were asked to produce summaries of CKD patients by accessing all their EHR data. The goal of the study was to provide evidence on how to create a good summary by analyzing the accessed EHR areas and how the doctors navigated through the longitudinal patient record. *Clinical narratives* took half of the time to create the summaries, and considerable time was used in *laboratory and radiology findings*, evidencing the importance of these sections.

Samal et al., 2011	Chronic Kidney Disease;
	Cancer;
	Chronic Coronary Artery Disease
		They present five hypothetical clinical scenarios to illustrate data needs in chronic disease management. For each scenario, they propose *improvements in temporal data views*. Besides that, they suggest *visualization techniques for numerical data* and *disease-oriented summaries for non-numerical data*.

Poh; De Lusignan, 2011	Chronic Kidney Disease	They lead a pilot study aiming to demonstrate *biometric methods of data visualization* to display trends in renal function in individual patients and indicate if a new value is part of normal fluctuation or an abnormal finding
Clarke et al., 2018	Diabetes mellitus; Various acute diseases	The main goal of their study was to explore the information needs of patients in an after-visit summary (AVS). They recruited patients with chronic and acute conditions. The results of the study can be used to support the development of an AVS that improves patient *adherence* and *understanding* of the treatment plan.

Source: the author

In order to **delimit the scope of summarization studies to be explored in this thesis**, we present a relation of previous summarization attempts that could be applied to our task. Table 4 describes anterior attempts to build a clinical summarizer for chronic condition patients. In Table 5, we indicate summarization studies that we think can be replicated/adapted in order to perform chronic disease summarization, even having a different focus.

Several methods were excluded from this research scope due to some limitations related to the pt-br language NLP resources or major differences regarding the data (e.g., use of structured data), objectives, and summarization characteristics (e.g., topic-based output). For example, various studies have a dependency on an ontology mapper as MetaMap (PIVOVAROV, 2015; MORADI; GHADIRI, 2018; LIANG; TSOU; PODDAR, 2019; MACAVANEY et al., 2019; WENG; CHUNG; TONG, 2020).

Table 4 - Studies developing clinical summarizers for chronic condition patients.

Author(s), year	Condition(s)	Project outline
Hsueh et al., 2015	Diabetes mellitus	Unstructured data such as diagnosis codes, medications, and lab results were transformed into relevant explanatory variables for multivariate predictive modeling. The objective is to identify risk factors for classifying patients at risk of worsening disease progression (i.e., automatic risk factors summarization).
Goldstein; Shahar, 2015-2016	Diabetes mellitus; Cardiothoracic post-surgery patients	The authors automatically transformed structured time-based data into a free-text summary. They opted to use only structured data because of their temporal-abstraction method demand this kind of data. The patient textual data was used only for evaluation purposes.
Pivovarov, 2015	Chronic Kidney Disease*	The author of this thesis developed a set of algorithms to solve some challenges in EHR summarization. They used hybrid data-driven and knowledge-based approaches to deal with redundancy in clinical narratives, a data-driven approach to identify and mitigate biases in laboratory testing patterns, and a probabilistic modeling approach to automatically summarize patient records. *Used CKD to perform a part of the experiments.
Aleksić et al., 2017	Various chronic diseases	This study proposes a data structure extension in the EHR database to allow summarization. They used the structured data (e.g., diagnosis codes, medication prescriptions, and visits information) from EHR relational database to *aggregate* and display information using a time-oriented layout.
Viani et al., 2017	Inherited and acquired heart diseases	They developed a language-specific clinical summarizer for Italian clinical notes. To overcome the lack of lexical and semantic resources, they used a Machine Learning supervised approach that learned from annotated data. All extracted events from the text were *aggregated* and *organized* into a timeline. No *reduction*, *transformation* or *synthesis* steps were defined. Nevertheless, the authors claim that "*the proposed approach has the potential to enhance the process of reviewing patient clinical histories, reducing the time needed to access large amounts of data*"
Acharya et al., 2018	Chronic heart failure	The study integrates physician (free-text) and nursing (concept codes) notes to generate summaries of what happened to a patient in the hospital. They use MedLEE to extract concepts and SimpleNLG (GATT; REITER, 2009) to generate the text.
Levy-Fix et al., 2020*	HIV	This study proposes an unsupervised phenotyping approach that could learn patient's problems from both structured and unstructured data, and then produce a visual problem-oriented summary concerning the history of the patient. They applied to HIV patients' data. *Published recently, few weeks before the ending of this project

Source: the author.

Table 5 - Summarization studies that can possibly be used for chronic disease summarization.

Author(s), year	Why is it relevant?	Brief overview
Moen et al., 2016	They worked with texts outside the English scope and patients with health-related issues (which includes chronic conditions). Moreover, it is a completely unsupervised and knowledge-free study	The authors designed, implemented, and evaluated new unsupervised methods for summarizing multiple clinical notes. They compared the new methods with some known baselines.
Litvak et al., 2013	Besides the needed adaptations to deal with clinical text, multi-document, and need for a clinical POS-Tagger, this study is language-independent, hence there is no reason to assume that such a method might not be useful also for pt-br. Moreover, there is no need for external knowledge resources.	They introduce DegExt, a graph-based language-independent keyphrase extractor, which extends other previous studies. Their method is unsupervised and extractive. They evaluated using English and Hebrew journalistic texts.
Sarkar et al., 2011	A supervised alternative for medical text summarization. Use a straightforward method based on cue-phrases and gold standard phrases defined by experts.	They treat a document as a set of sentences, which the learning algorithm must learn to classify as positive or negative examples of sentences, and an ML algorithm learns with some pre-annotated summaries. To complement the algorithm, a set of important cue-phases is used to find relevant statements. They evaluated using a set of medical articles.
Miller, 2019 and Moradi; Samwald, 2019	These are totally unsupervised methods that were tested on two different domains with promising results. Moreover, it uses the BERT architecture, which is the current state-of-the-art model for many NLP tasks.	Both studies transform texts using a BERT model and then apply a clustering algorithm to select candidate sentences. One study summarizes lecture notes and the other work with biomedical texts.

Source: the author.

2.5.3 Research Trends and Gaps

After reading the most recent literature reviews on Biomedical Summarization (MISHRA et al., 2014), Clinical Summarization (PIVOVAROV; ELHADAD, 2015), Automatic Text Summarization (GAMBHIR; GUPTA, 2017; YAO et al., 2017); and performing an additional review on Clinical Summarization to gather studies out of the time range of Pivovarov and Elhadad work (i.e., studies published from 2015 to 2018); it became clear what are the actual research gaps and trends in the biomedical/clinical summarization field (compiled in Table 6 and Table 7 respectively).

Table 6 - Compilation of summarization research gaps.

Gap	Description
Lack of summarization corpora and annotation standards	The shortage of annotated corpora and reference standards publicly available for text summarization is an obstacle to research progress, as several approaches (especially the Machine Learning ones) depend on it for training and testing purposes, and generating these resources is costly and time-consuming. The available standard datasets for document summarization tasks are often small-scaled, containing a short amount of topics. The scarcity of data is more apparent in domains other than news (e.g., biomedical) and languages other than English. Consequently, the research field lacks domain- and language-focused studies.
Lack of methods using languages other than English	For obvious reasons, this issue is directly related to the previous one, given the lack of corpora in other languages than English. But another major cause is related to the limitations of lexical and semantic tools in other languages, which not only impairs Machine Learning approaches (as in the first case), but also methods based on Natural Language Processing for example.
Lack of visual output	Most of biomedical summarization work generated textual summaries, with very few studies including visual outputs (e.g., statistics, tables, graphics, visual-rating scales). Currently, there is no consensus on which information should be displayed textually and which can be made more accessible in a graphical way. To solve this, more interaction between the summarization and visualization communities is needed, besides more initiatives developing such visual summaries.
Lack of extrinsic evaluation	Most of biomedical summarization studies are focusing on intrinsic evaluation rather than extrinsic, that is, a few studies assessed their summarization methods in a real-world environment to measure the real impact of their system in downstream biomedical tasks like decision support. This may indicate an immature level of biomedical summarization research.
Lack of quality aspect in the intrinsic evaluation	Currently, the ROUGE metrics (categorized as an *informativeness* measure) are the most used intrinsic automatic evaluation, probably because it is a fast and cheap method. Since ROUGE is designed mostly for detecting information coverage rather than quality factors, it would be valuable to assess quality elements such as grammaticality, coherence, and clarity.
Underexplored Neural Networks-based summarization methods	Methods based on modern neural network architectures have been achieving state-of-the-art results on multiple NLP tasks, and the advent of large pre-trained models based on Transformers architecture popularized the use of these resources, which are still underused in the context of clinical summarization.

Source: the author.

Table 7 - Compilation of summarization research trends**.**

Trend	Description
Multi-document summarization	With the exponential growth of both literature and EHR data is important to summarize information regarding multiple documents, especially because relevant information is very often distributed among several documents, such as published clinical studies and clinical narratives in a patient's EHR.
Abstractive summarization	The extractive approaches still lead the summarization initiatives in biomedical summarization, but there is an increasing interest in abstractive methods. This could be due to the natural evolution of the field, going towards a summary more focused on the patient state, giving more meaning to information.
Knowledge-rich methods	Contributing to this trend is the fact that a large number of knowledge resources (e.g., UMLS, SNOMED-CT, PubMed) and NLP tools (MetaMap, cTAKES, MedLEE) are available, and can assist algorithms in giving meaning to medical concepts.
Hybrid methods	Mixing statistical, NLP, and Machine Learning approaches trending in biomedical summarization. As a summarizer algorithm can be divided into many steps (like the AORTIS model with aggregation; organization; reduction and transformation; interpretation; and synthesis) it is convenient to use the best approach possible to each step, or even combine extractive and abstractive techniques in order to produce a good summary.
Intrinsic evaluation	Even if the recommendations of the scientific community point to extrinsic evaluation, the intrinsic assessment still dominates the literature.
User-oriented summarization	In biomedical summarization, they are far away from being dominant, but there is a growing interest in developing summaries that involve user interactions. Because in some cases, the users have different requirements, consequently a certain level of personalization is desirable.

Source: the author.

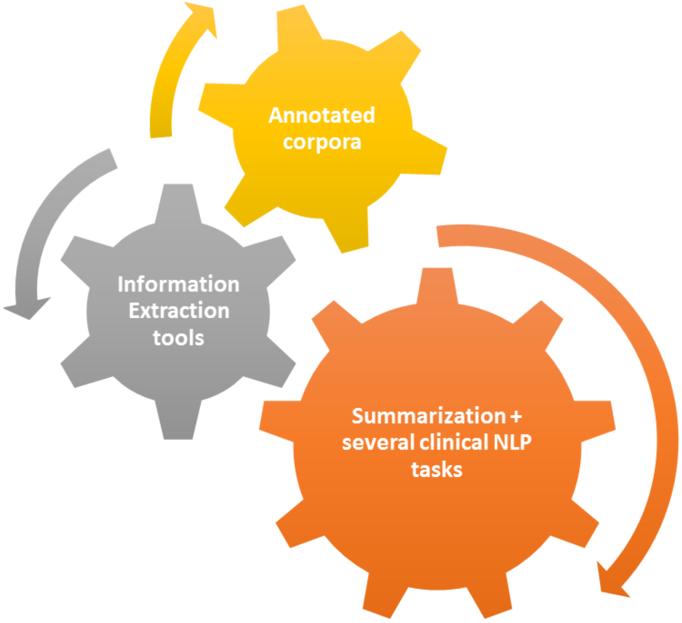
3 INFORMATION EXTRACTION ON CLINICAL TEXTS

In this chapter, we present the complete development of a Named Entity Recognition algorithm for pt-br clinical narratives, since the construction of the SemClinBr, a semantically annotated corpus for clinical NLP tasks, to the training of multiple methods to identify and classify medical concepts within the text. It is worth mentioning the definition of a new POS-Tagging environment used to improve NER performance.

The main motivation to build a NER algorithm, and the additional resources, is to support clinical summarization algorithms, including in the intermediate representation step, because when the clinical concepts in the text are known, it is easier to define the terms' importance, and consequently, to select the most important sentences. Furthermore, it is significant that EHR-related systems could go beyond simply showing patient's stored data but could process, transform, interpret and infer information from it, and knowing the clinical concepts within unstructured data is vital to do so.

Most NER methods are based on supervised learning, which led us to develop a gold-standard corpus to train and test our algorithm. Additionally, the lack of resources of this kind for pt-br instigated us to build the SemClinBr corpus and share it with the research community aiming to boost the biomedical NLP area for the Portuguese language. Figure 10 represents the dependency of the clinical summarization and other NLP tasks on the produced resources and tools.

Figure 10 - Representation of the need for NLP resources and tools to move the gear of clinical summarization and other tasks.


Annotated corpora
Information Extraction tools
Summarization + several clinical NLP tasks

Source: the author.

3.1 SEMCLINBR - A MULTI-INSTITUTIONAL AND MULTI-SPECIALTY SEMANTICALLY ANNOTATED CORPUS FOR PORTUGUESE CLINICAL NLP TASKS

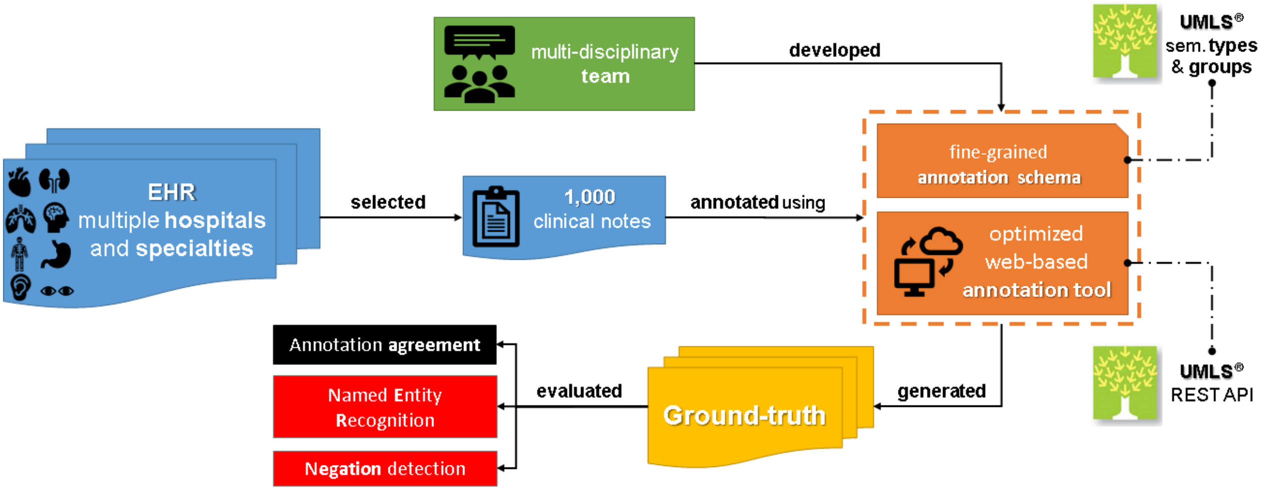
In this section, we detail the database used in this project and explain all steps regarding the annotation process, including the definition of an annotation schema, the development of a web-based annotation tool, and the team involved in the annotation. An overview of the whole process is shown in Figure 11, in blue, the selection of a thousand clinical notes from multiple hospitals and medical specialties. A multi-disciplinary team developed the elements in orange, which represent the fine-grained annotation schema following the UMLS semantic types, and the web-based annotation tool featuring the UMLS REST API. These resources supported the generation of the ground-truth (i.e., gold standard), which was evaluated intrinsically (i.e., inter annotation agreement) and extrinsically in two different NLP tasks (i.e., Named Entity Recognition and Negation Detection).

It is worth mentioning that the motivation behind the SemClinBr is three-fold: (i) to support the development of a NER algorithm to be used in a summarization method, (ii) to evaluate a semantic search algorithm developed by Philips Healthcare, the sponsors of this research project, and (iii) share a multi-purpose corpus with the research community. More details about the construction of the corpus can be found in the original article (OLIVEIRA et al., 2020).

3.1.1 Data acquisition

The acquired data comes from two distinct de-identified and approved by Ethical Committee databases (Certificate of presentation for Ethical Appreciation number 51376015.4.0000.0020), the first is a corpus of 2,094,929 EHR records from a group of three hospitals over a period of five years (hereafter, group corpus), and the second is a dataset composed by 5,617 entries from one single hospital (henceforth, single corpus). Table 8 shows the information present in group corpus, comprising structured (e.g., gender, medical specialty, inclusion date) and unstructured data, in a free-text format representing the sections of a clinical narrative (e.g., main complaint, history of the disease, past history, family history). In addition to the multi-institutional characteristic of the corpus, it covers several medical specialties (e.g., cardiology, nephrology, and endocrinology) and document types (e.g., discharge summaries, nursing notes, admission notes, and ambulatory notes).

Figure 11 – A wide perspective of SemClinBr development.



Source: Oliveira et al., 2020.

The single corpus has only discharge summaries originating from the cardiology sector exclusively. The data configuration has structured data (i.e., gender, birth date, begin date, end date, and icd-10 code) and a free-text data field, concerning the discharge summary.

Both corpora are characterized by texts containing writing problems, some already characteristic of clinical texts (DALIANIS, 2018), such as redundancy, elevated use of acronyms and medical jargon, misspellings, negation and uncertainty, erroneous or lack of punctuation, and fragmented sentences. Moreover, our datasets have additional issues. For instance, in the group corpus, the clinical note is supposed to be written using the following fields: main-complaint, history-of-disease, past-history, family-history, physical-examination, main-diagnosis-hypothesis, initial-plan, and observations. But most of the clinicians put all the note text into the history-of-disease field, with the others often remaining empty, making it difficult to search for specific information in the narrative (e.g., look for family history). Additionally, some texts are entirely written in upper case letters and interfering directly in some text processing, such as finding abbreviations and identifying proper nouns.

Table 8 - Group corpus data fields and the respective data types of each column.

Field	Data type
occurrence-id	Number
patient-id	Number
gender	Text
birth-date	Date
inclusion-date	Date
discharge-date	Date
discharge-type	Text
discharge-reason	Text
icd-10	Text
medical-specialty	Text
care-reason	Text
main-complaint	Text
history-of-disease	Text
past-history	Text
family-history	Text
physical-examination	Text
main-diagnosis-hypothesis	Text
initial-plan	Text
observations	Text

Source: the author.

In Table 9, we present some text examples. The first sample is a cardiology discharge summary with a complete history of care, provided by regular/long sentences with no apparent format standardization. The second example is an ambulatory note describing a patient visit to the Nephrology department; the text includes concise sentences written in uppercase letters, a high frequency of acronyms, and lacking punctuation. The last text is a nursing note, which describes the monitoring of the nursing team to the patient.

Table 9 - Examples of clinical narratives included in our corpus, with different types (e.g., discharge summary, ambulatory note, nursing note) and medical specialties. The second column show the original pt-br text and in the next column the translated version of it (maybe some acronyms translation does not make sense in English).

Type/Specialty	Original note	Translated note
Discharge summary Cardiology	Paciente ex-tabagista, vem à emergência com quadro de dispnéia progressiva, ortopnéia, dispnéia paroxística noturna, edema de membros inferiores, turgência jugular. Diagnóstico de insuficiência cardíaca, com classe funcional IV (NYHA) na chegada. Sem história de dor torácica. ECG da chegada sem alterações. Marcadores de necrose miocárdica normais. Manejado para insuficiência cardíaca com boa resposta clínica. Ecocardiograma demonstrando dilatação de cavidades (AE = 5,3 cm, DDVE = 7,0, DSVE = 5,8), disfunção sistólica (FEVE = 35%) por hipocinesia difusa, septo e parede posterior de 0,9 cm, insuficiência mitral e tricúspide leves e PSAP = 52 mmHg. Realizado investigação etiológica com sorologia negativa para Chagas, cintilografia demonstrando necrose apical, sem condições de discriminar isquemia. Optado então pela realização de cateterismo cardíaco, que revelou artéria circunflexa dominante e livre de lesões significativas; artéria coronária direita livre com sinais de aterosclerose, mas sem lesões significativas; artéria descendente anterior de pequeno calibre, com lesão de cerca de 60% no terço proximal e lesão crítica no terço médio. Após revisão do filme, observou-se tratar de lesão de difícil manejo percutâneo, devido à sua extensão e ao pequeno calibre da artéria descrita. Após discussão do caso, optou-se por manejo clínico devido ao fato do paciente não apresentar angina, ter respondido com sucesso à terapêutica instituída e não apresentar evidência clara de benefício atual com procedimento de revascularização. Impressão de que a lesão em DAE não explicaria a hipocinesia difusa apresentada pelo paciente, devendo ser portanto doença aterosclerótica coexistindo em um coração com miocardiopatia dilatada. Realizado ainda espirometria que evidenciou distúrbio obstrutivo	Ex-smoker patient comes to the emergency room with progressive dyspnea, orthopnea, paroxysmal nocturnal dyspnea, lower limb edema, and jugular turgence. Heart failure diagnosis, with functional class IV (NYHA) upon arrival. No history of chest pain. ECG on arrival without change. Normal myocardial necrosis markers. Managed for heart failure with good clinical response. Echocardiogram showing cavity dilatation (LA = 5.3 cm, LVDD = 7.0, LVSD = 5.8), systolic dysfunction (LVEF = 35%) due to diffuse hypokinesia, a 0.9 cm septum and posterior wall, mild mitral and tricuspid regurgitation, and APSP = 52 mmHg. Etiological investigation with Chagas negative serology, scintigraphy showing apical necrosis, unable to discriminate ischemia. Then opted for catheterization which revealed a dominant circumflex artery free of significant lesions; free right coronary artery with signs of atherosclerosis but no significant lesions; small anterior descending artery with a lesion of about 60% in the proximal third and critical injury in the middle third. After review of the film, it was observed that it was a difficult percutaneous management injury owing to its extension and the small caliber of the described artery. After discussion of the case, we opted for clinical management because the patient did not have angina, successfully responded to the therapy instituted, and did not present clear evidence of being benefitted by the revascularization procedure. The impression that the lesion in LAD would not explain the diffuse hypokinesia presented by the patient; therefore, atherosclerotic disease coexisting in a heart with dilated cardiomyopathy. Accomplished yet spirometry that showed moderate

	moderado. DCE estimada em 57 ml/min. Paciente recebe alta em bom estado geral, afebril, eupnéico, em otimização do tratamento para ICC (já em uso de betabloqueador, IECA e espironolactona), com plano de ajustes de doses a nível ambulatorial. OBS: peso na alta: 76 Kg.	obstructive disorder. DCE estimated at 57 ml / min. Patient is discharged in good general condition, afebrile, eupneic condition, optimizing treatment for CHF (already using beta-blocker, ACEI and spironolactone), with outpatient dose adjustment plan. OBS: weight in the high: 76 Kg
Ambulatory note Nephrology	NEFROPATIA DIABETICA EM TTO CONSERVADOR CANDIDATA A TX RENAL PREEMPTIVO LIBERADA PELA URO E ANESTESIO CANDIDATA A TX RENAL PREEMPTIVO ASSINTOMÁTICA, EXCETO PELOS SINAIS E SINTOMAS ASSOCIADOS A NEUROPATIA PERIFERICA (DIABETICA / UREMIA) SEM SINTOMAS URINARIOS AO EXAME PA 150/100 P 108 T 36 DIURESE FRR NORMAL HIPOCORADA + CPP LIVRES PC RITMO REGULAR, TAQUICARDICO ABD RHA+, PLANO, FLÁCIDO, CIC CX CST MMII PULSOS PRESENTES E SIMETRICOS	DIABETIC NEPHROPATHY IN CONSERVATIVE TREATMENT PREEMPTIVE KIDNEY TRANSPLANT CANDIDATE RELEASED BY UROLOGY AND ANESTHESIOLOGY PREEMPTIVE KIDNEY TRANSPLANT CANDIDATE ASYMPTOMATIC, EXCEPT FOR SIGNS AND SYMPTOMS ASSOCIATED WITH PERIPHERAL NEUROPATHY (DIABETIC / UREMIA) NO URINARY SYMPTOMS ON EXAMINATION BP 150/100 HR 108 T 36 DIURESE RR NORMAL PALLOR + FREE LF CS REGULAR RHYTHM, TACHYCARDIC ABDOMEN RHA +, FLAT, FLACCID, CIC CX CST LLLL PRESENT AND SYMMETRICAL PULSES
Nursing note Not defined	Pcte com RNM de crânio agendada para hoje às 23:00h. Por volta das 21:00h pcte apresentou quadro de confusão mental , seguida de crise convulsiva generalizada , prontamente atendido na sala de poli , com MCC + oximetria digital de pulso + PNI contínuos . Instalado O2 , medicado CPM e mantido em observação no leito . Hidantalizado pela R1 Vital Brasil da neurocirurgia , procedimento realizado sem intercorrências . Pcte bastante sonolento , mantido em sala de poli e suspenso RNM por hora . Diurese espontânea , com controle através de uropen . SSVV às 05:45h PA = 133/74mmhg , FC = 114bpm, SpO2 = 93% . Conforme orientação da neurocirurgia , mantém observação na sala de poli sob cuidados intensivos de enfermagem . CHOQUE NAO ESPECIFICADO	Patient Skull MRI scheduled today at 23:00. At around 21:00, the patient presented with mental confusion, followed by generalized seizure, promptly treated in the multiple trauma room, with MCC + digital pulse oximetry + continuous NIBP. Installed O2, medicated as prescribed and kept under observation in bed. Hidrantalized by R1 Vital Brasil of neurosurgery, procedure performed without complications. Very sleepy patient kept in emergency room and suspended MRI for hour. Spontaneous diuresis, with uropen control. VVSS at 05:45 h BP = 133 / 74 mmhg, HR = 114 bpm, PsO2 = 93%. As directed by neurosurgery, maintains observation in the emergency room under intensive nursing care. SHOCK NOT SPECIFIED

Source: Oliveira et al. (2020).

We selected 1,000 texts from both corpora to be annotated (Table 10 shows the number of documents per specialty). The average character token size was 148, and the average sentence size was approximately ten tokens. Note that several documents are assigned as "Not defined". Nevertheless, by looking at these documents, we can conclude that these patients are (a) under the care of multiple medical specialties (e.g., patient with multiple trauma, in Intensive Care Unit) or (b) in the middle of a diagnostic investigation. Moreover, we grouped the specialties with less than ten documents as "Others" (e.g., urology, oncology, gynecology, rheumatology, proctology).

Table 10 - The medical specialties of the selected clinical narratives order by their frequency in our corpus. Medical specialties with less than ten occurrences were grouped into "Others" category.

Specialty	**Number**
Cardiology	260
Nephrology	157
Orthopedics	126
Not defined	122
Surgery (general)	61
Neurology	45
Neurosurgery	32
Dermatology	23
Ophthalmology	22
Endocrinology	19
Gastroenterology	16
Otolaryngology	14
Pneumology	11
Others	92

Source: Oliveira et al. (2020).

3.1.2 Annotation guidelines

The annotation guidelines serve as a guide to annotators, providing details on how to annotate each concept and listing a set of useful examples. They are crucial to maintain the homogeneity during the entire annotation process and consequently ensure the gold standard quality.

The **UMLS semantic types**[6] were selected as the annotation tags for the project, i.e., the clinical concepts within the text would follow the UMLS semantic categories (e.g., Disease or Syndrome, Sign or Symptom, Therapeutic or Preventive Procedure, Laboratory or Test Result). To illustrate, Table 11 presents annotated text samples with their respective semantic types and groups.

[6] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

Table 11 – Text samples containing the most used semantic types (STY) and their corresponding semantic groups (SGR). The third column shows the original examples, and the fourth shows the translated versions. The underlined passages indicate annotated concepts.

SGR	STY	Original examples	Translated examples
Anatomy	Body Location or Region	MEIA TALA GESSADA EM MIE apresenta edema em região craniana ABDÔMEN PLANO E FLÁCIDO	Half-length plaster cast in LLL presents edema in the cranial region FLAT AND FLACID ABDOMEN
Anatomy	Body Part, Organ, or Organ Component	acesso venoso central em jugular D ACESSO VENOSO PERIFERICO EM BRAÇO DIREITO	right jugular central venous access RIGHT ARM PERIPHERAL VENOSOUS ACCESS
Chemicals & Drugs	Organic Chemical	Fez uso de atenolol por 3 anos cefaléia em regiao parietal bilateral que melhora com dipirona	used atenolol for 3 years headache in bilateral parietal region improved with dipyrone
Chemicals & Drugs	Pharmacologic Substance	asmatica em uso de salbutamol e budesonida	asthmatic person using salbutamol and budesonide
Concepts & Ideas	Temporal Concept	POI DE LAVAGEM + CURETA DE TECIDO NECRÓTICO Paciente em Pré-operatório de FX fêmur	WASHING IP + NECROTIC TISSUE CURETAGE Preoperative patient of femur fracture
Devices	Drug Delivery Device	cloreto de potassio a 42 ml/h em bomba de infusão	potassium chloride at 42 ml/h in infusion pump
Devices	Medical Device	AVP em MSE com soroterapia em curso SVD com diurese efetiva	PVA in LUL with ongoing serotherapy DBP with effective diuresis
Disorders	Disease or Syndrome	REFERE HIPERTENSÃO E DIABETES EM USO DE INSULINA. SINDROME DE GUILLAIN BARRE.	REFERS HYPERTENSION AND DIABETES IN INSULIN USE GUILLAIN BARRE SYNDROME
Disorders	Finding	RETORNOU DO CC LÚCIDO, ORIENTADO, COMUNICATIVO consciente, comunicativo, pupilas isocóricas fotoreagentes	RETURNED LUCID FROM SC CONSCIOUS, COMMUNICATIVE conscious, communicative, photoreagent isochoric pupils
Disorders	Injury or Poisoning	TRAUMA CRÂNIOCERVICAL APÓS QUEDA FRATURAS MULTIPLAS DA COLUNA TORACICA.	SKULL-CERVICAL TRAUMA AFTER FALL MULTIPLE FRACTURES IN THORACIC COLUMN
Disorders	Sign or Symptom	relata cefaléia SINAIS VITAIS ESTAVÉIS, REFERE ALGIA	reports headache STABLE VITAL SIGNS, REFERS PAIN
Living Beings	Patient or Disabled Group	paciente nega queixas, nega dor, dispnéia. Pcte com cultura de Secreção Tibial	patient denies complaints, denies pain, dyspnea. Ptt with Tibial Secretion culture
Living Beings	Professional or Occupational Group	Orientada a equipe de enfermagem que o mesmo esta em jejum segundo a farmacêutica e o médico	Advised nursing staff that the patient is fasting according to the pharmacist and the doctor
Organizations	Health Care Related Organization	CONFORME ROTINA DA UTI RETORNOU DO CC ÀS 14:30HRS	AS ICU ROUTINE RETURNED FROM SC AT 2:30pm
Phenomena	Laboratory or Test Result	Glicose 335; LDH 223; Teste rápido para HIV negativo	Glucose 335; LDH 223; HIV negative rapid test
Physiology	Clinical Attribute	PA = 130/70 PESO 67,4	BP = 130/70 WEIGHT 67.4
Procedures	Diagnostic Procedure	AUSCULTA PULMONAR; MV +, RONCOS DIFUSOS EM BASES Monitorização cardíaca contínua, PAM e oximetria digital.	PULMONARY AUSCULTATION; VM +, DIFFUSED WHEEZES IN BASES Continuous cardiac monitoring, MAP, and digital oximetry.
Procedures	Health Care Activity	EM ACOMPANHAMENTO NA ENDOCRINO DO HC Internamento em janeiro por taquicardia atrial com aberrância	FOLLOW-UP ON ENDOCRINOLOGY AT HC Admission in January for aberrant atrial tachycardia
Procedures	Therapeutic or Preventive Procedure	SVD COM 100 ML DEBITO SEM GRUMOS IRC EM DIALISE	DC WITH 100 ML DEBIT WITHOUT GROUNDS CRF IN DIALYSIS
N/A	Abbreviation	CONFORME ROTINA DA UTI[Unidade de Terapia Intensiva] MEIA TALA GESSADA EM MIE[Membro inferior esquerdo]	AS ICU[Intensive Care Unit] ROUTINE Half-length plaster cast in LLL [Lower left limb]
N/A	Negation	Paciente eupnéico e afebril Paciente nega algia SEM IRRADIAÇAO	Eupneic and feverless patient Patient denies pain NO IRRADIATION

Source: Oliveira et al. (2020).

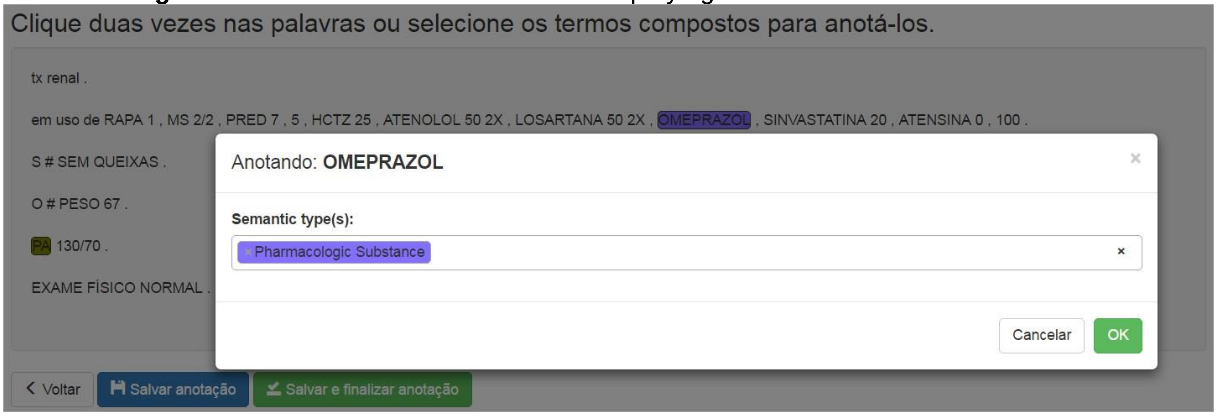
The use of UMLS semantic types allows the automatic reduction of corpus granularity if needed, by making the correlation to their corresponding semantic groups[7]. Moreover, by opting for UMLS semantic types, we can explore UMLS Metathesaurus as an additional guide for annotators and also make use of the UMLS API[8] to support our annotation tool (more details in next section).

In addition to UMLS types, two more types were added in SemClinBr's tagset, the "Negation" and "Abbreviation" tags. The first one aims to identify negation cues associated with clinical concepts. The "Abbreviation" type was incorporated to help us in the process of abbreviation disambiguation and expansion.

3.1.3 Annotation tool

To overcome some of the issues and difficulties regarding the clinical annotation (XIA; YETISGEN-YILDIZ, 2012) and based on a review on annotation methods and tools (ANDRADE; OLIVEIRA; MORO, 2016), we decided to build a new annotation tool (see Figure 12 for a screenshot of the tool).

Figure 12 - Annotation tool screenshot displaying the annotation of a medication.



Clique duas vezes nas palavras ou selecione os termos compostos para anotá-los.

tx renal .

em uso de RAPA 1 , MS 2/2 , PRED 7 , 5 , HCTZ 25 , ATENOLOL 50 2X , LOSARTANA 50 2X , OMEPRAZOL , SINVASTATINA 20 , ATENSINA 0 , 100 .

S # SEM QUEIXAS .

O # PESO 67 .

PA 130/70 .

EXAME FÍSICO NORMAL .

Anotando: **OMEPRAZOL**

Semantic type(s):

Pharmacologic Substance

Source: Oliveira et al. (2017).

The idea was to provide a real-time annotation environment, shared by the entire annotation team, in which they could work anywhere and anytime (i.e., web-based application). Since the participants had very different and irregular time schedules, it was important that the project manager could supervise and assign

[7] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

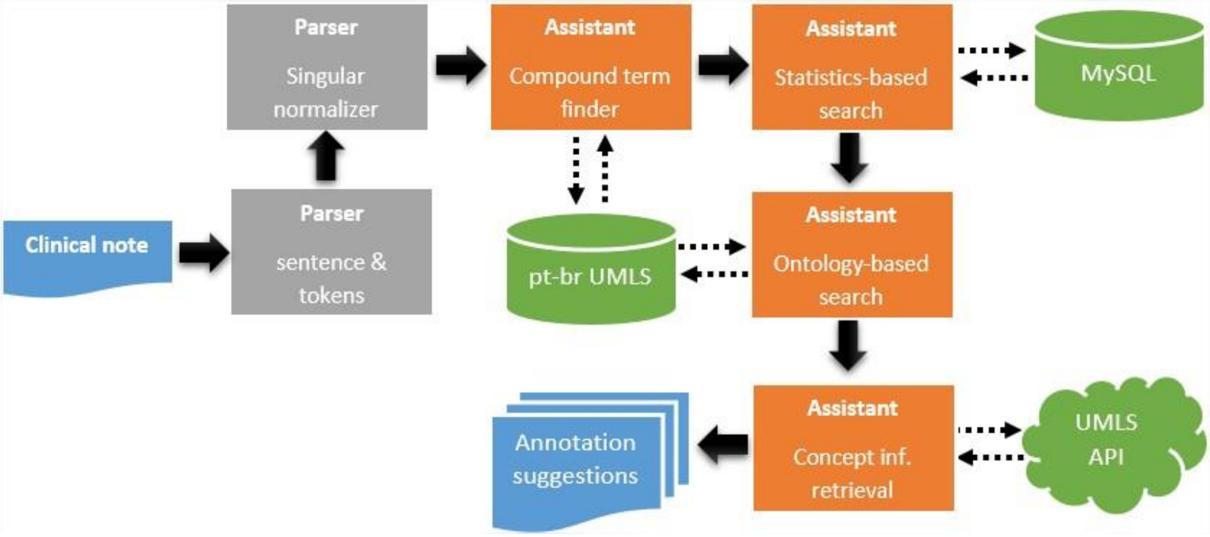
[8] https://documentation.uts.nlm.nih.gov/rest/home.html

remaining work with no need for an in-person meeting. Furthermore, as we utilized a UMLS semantic type scheme, it would be desirable to exploit the UMLS API and other local resources (e.g., clinical dictionaries, previous annotations) in order to make annotation suggestions to the user without pre-annotating it.

Finally, we needed a tool that fitted exactly into our annotation workflow, with the raw data input into our environment, and a gold-standard output at the end of the process, dispensing the use of external applications. Our tool workflow was composed of six main modules:

- Importation: import data files into the system
- Review: manually remove PHI information that the anonymization algorithm failed to catch
- Assignment: allocate text to annotators
- Annotation: allow labeling of the clinical concepts within the text with one or multiple semantic types, supported by the Annotation Assistant feature
- Adjudication: resolve double-annotation divergences and creation of the gold standard
- Exportation: exports the gold standard as JSON or XML

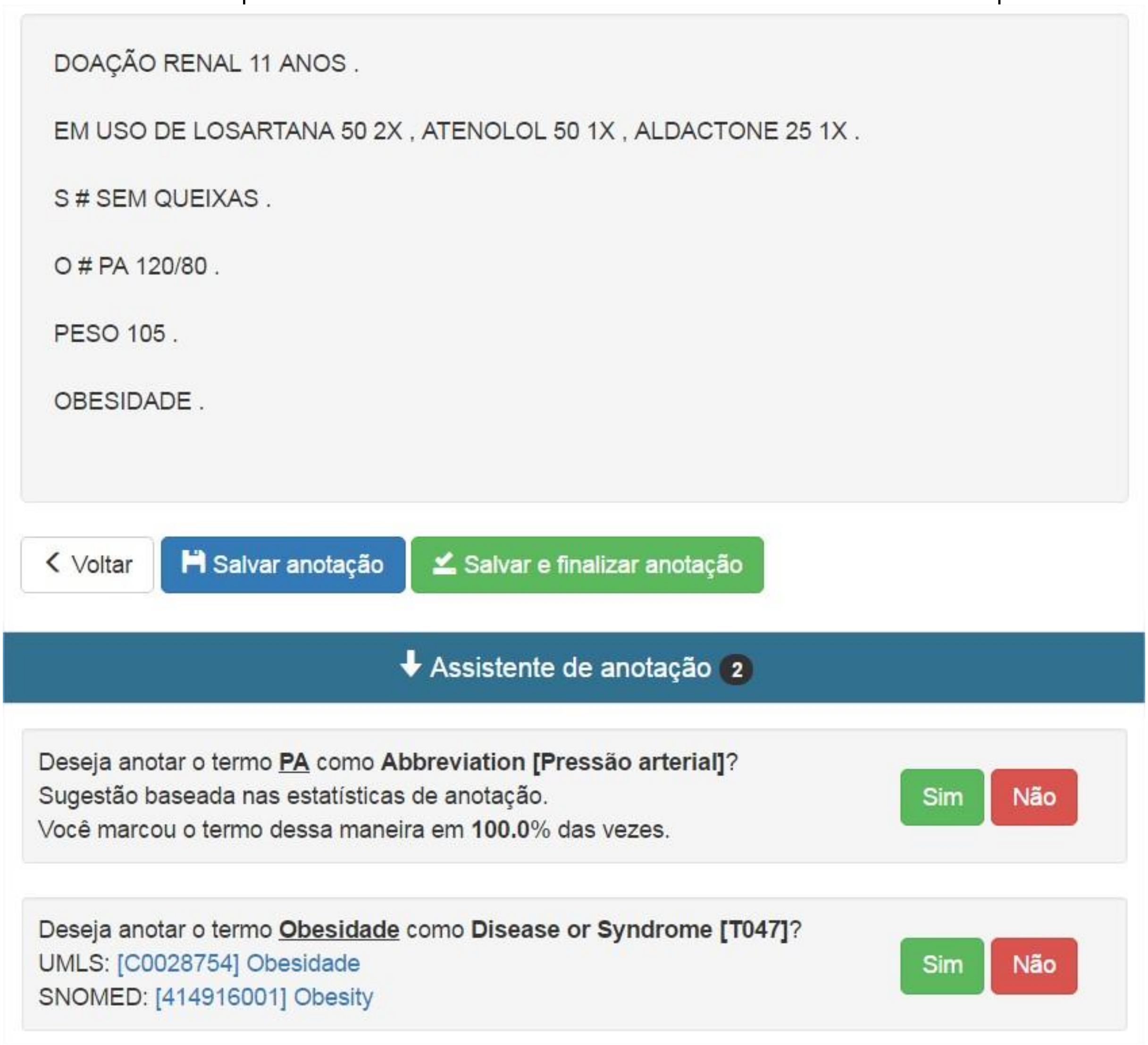
Figure 13 - Annotation assistant overview. The text is normalized, multiword expressions are found. The suggestions are given based on previous annotations and UMLS concepts match.



Source: Oliveira et al. (2017).

The annotation assistant component was developed to prevent annotators from labeling all the text from scratch by giving them suggestions of possible annotations based on (a) previously made annotations and (b) UMLS API exact-match and minor edit-distance lookup. Figure 13 shows the steps of the annotation assistant module, and Figure 14, a screenshot exemplifying its use.

Figure 14 – Annotation tool screenshot showing the annotation assistant. The first suggestion is based on previous annotations and the second based on the UMLS API lookup.



DOAÇÃO RENAL 11 ANOS .

EM USO DE LOSARTANA 50 2X , ATENOLOL 50 1X , ALDACTONE 25 1X .

S # SEM QUEIXAS .

O # PA 120/80 .

PESO 105 .

OBESIDADE .

Deseja anotar o termo PA como **Abbreviation [Pressão arterial]**?
Sugestão baseada nas estatísticas de anotação.
Você marcou o termo dessa maneira em **100.0**% das vezes.

Deseja anotar o termo **Obesidade** como **Disease or Syndrome [T047]**?
UMLS: [C0028754] Obesidade
SNOMED: [414916001] Obesity

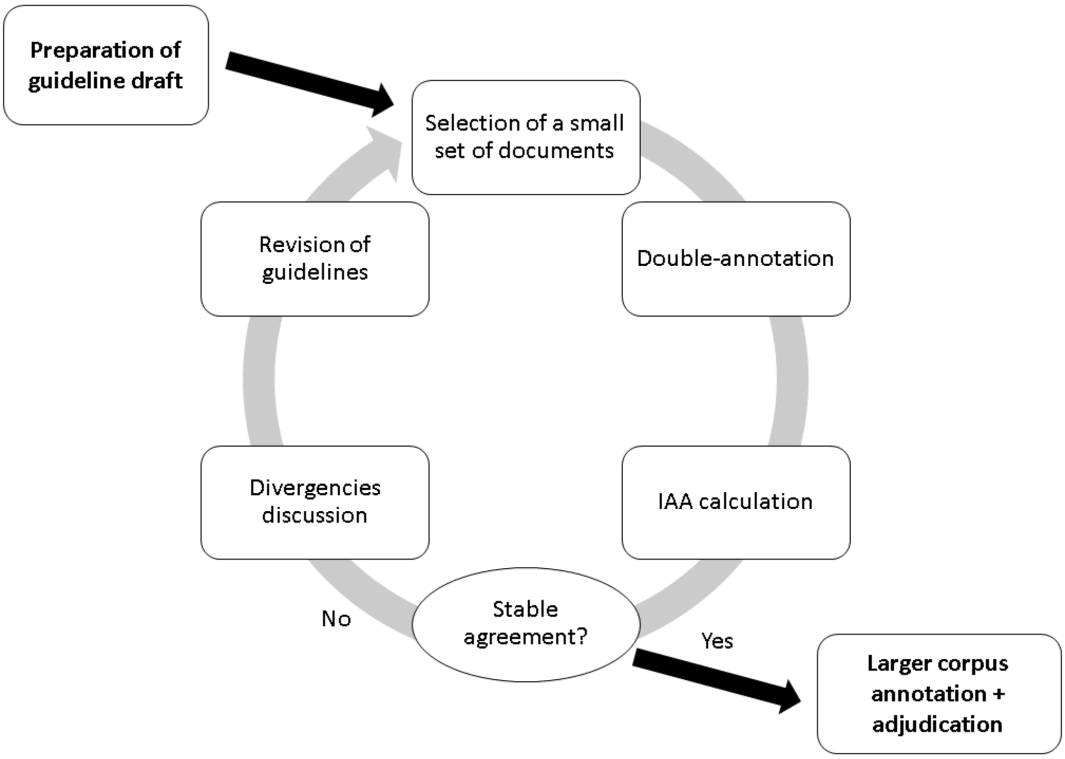
Source: Oliveira et al. (2017).

3.1.4 Annotation process

The SemClinBr was an end-to-end double annotation project, in which all the texts were annotated by two different annotators, and the differences were resolved by a third experienced annotator (i.e., adjudicator), resulting in the gold standard (i.e., ground-truth). It is worth mentioning that the adjudicator cannot remove annotations made by both annotators, and neither create new annotations, hampering a gold standard made with the opinion of a single person. Pairing annotators to perform a double annotation of a document prevents bias caused by possible mannerisms and recurrent errors of a single annotator. Moreover, it is possible to check the annotation quality by measuring the agreement between both annotators.

A training phase was provided to the annotators to acquaint them with the annotation tool and familiarize them with some of the difficulties in the process. Then, similarly to Roberts and colleagues (ROBERTS et al., 2009), an iterative process was started to enhance the guidelines, and check the consistency of annotation between annotators and provide feedback on their initial work. If, in three consecutive rounds, the agreement stayed stable (no significate reduction or improvement), there was no room for guideline adaptation, and the final annotation process could be initiated. A flowchart of the process is given in Figure 15.

Figure 15 – The iterative process started with the first guideline draft, then a small number of documents were double-annotated, and their inter-annotator agreement calculated. If the agreement maintained stable, then the guideline was considered good enough to proceed with the gold-standard production. If not, the annotation differences were discussed, the guidelines updated, and the process reinitiated.

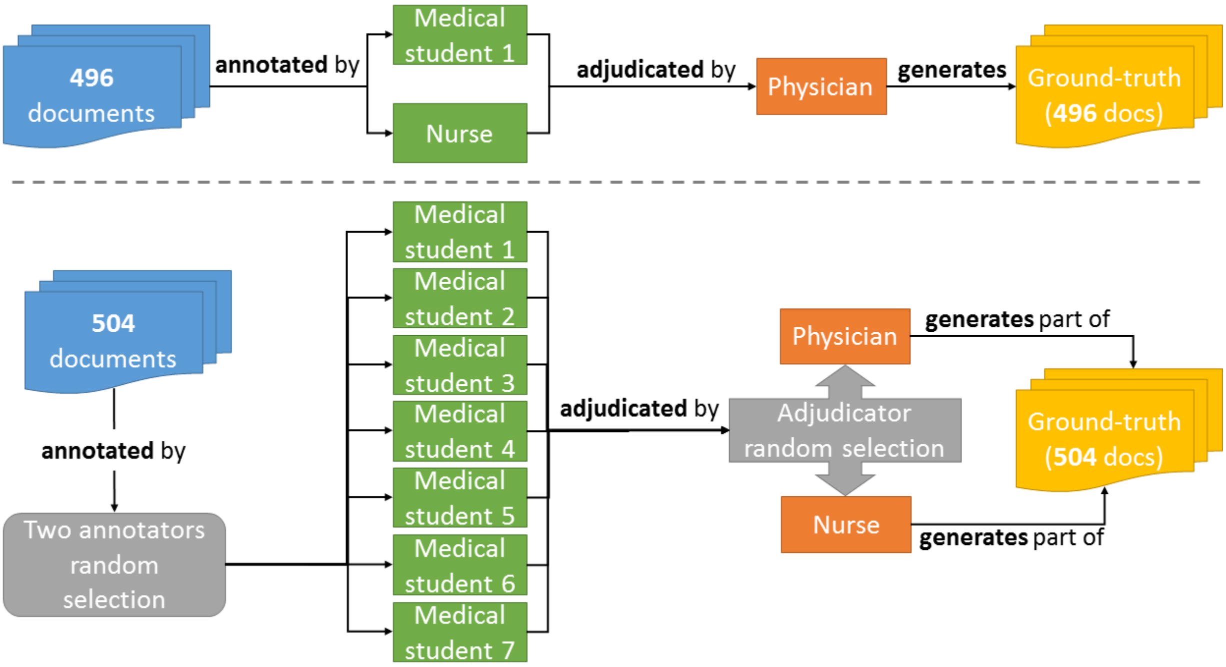

Preparation of guideline draft
Selection of a small set of documents
Double-annotation
IAA calculation
Stable agreement?
Yes
Larger corpus annotation + adjudication
No
Divergencies discussion
Revision of guidelines

Source: Oliveira et al. (2020)

After the guideline maturing process, we started the final development stage of the gold standard, which posteriorly was divided into ground-truth phases 1 and 2. We decided to recruit annotators with different profiles and levels of expertise to give us different points of view during the guideline definition process, and to determine if there were differences in annotation performance between annotators with different profiles.

Ground-truth phase 1 counted with a team of three persons: (1) a physician with experience in documenting patient care and participation in a previous clinical text annotation project; (2) an experienced nurse; and (3) a medical student who already had ambulatory and EHR's writing experience. The nurse and the medical student were responsible for the double-annotation of the text, and the physician was responsible for adjudicating them. When the process was almost 50% complete (with 496 documents annotated and adjudicated), we managed to engage more people to assist in finishing the task (in what we called **ground-truth phase 2**). An extra team of 6 medical students, with the same background as the first one, were recruited (Figure 16 illustrates these phases). We held a meeting in which we presented the actual guidelines document and trained them on using the annotation tool.

Figure 16 – The annotation process was divided into ground-truth phases 1 and 2, which are located above and below the dashed line, respectively. The elements in green represent the annotators and in orange the adjudicators.



Source: Oliveira et al. (2020)

In phase 2, we had two adjudicators, the physician, and the nurse. We added the nurse as an adjudicator as we needed one extra adjudicator during this phase, and the nurse had more hospital experience than medical student 1. Then, we had a homogeneous group of seven medical students annotating the texts. The physician, the nurse, and medical student 1 supervised the first set of annotations of all the students. The number of documents to be annotated were divided equally between the annotators and adjudicators, and the selection of double-annotators for each document was made randomly, as was done for the adjudicators. It is worth noting that besides the people mentioned above, who worked directly with annotation and adjudication, we had a team of Health Informatics researchers who participated in supporting the annotation project with other activities, including annotation tool development, guidelines discussion, and annotation feedback.

3.1.5 Corpus reliability and segmentation

Taking advantage of the fact that we had double-annotated the entire collection of documents, we calculated the inter-annotator agreement (IAA) of all the data using the *observed agreement* metric, which takes into account the matches between the annotators and the divergences (non-matches), as shown in Equation 1. For the *strict* version of IAA, a situation was considered a match when the two annotators label the same textual span with an equal semantic type. All other cases were calculated as a non-match. We reported the *lenient* version of IAA as well, which considers partial matches, that is, the annotations that have overlaps in the selected textual spans (with the same semantic type); these are counted as a half-match in the formula. The third version of IAA, called *flexible*, was calculated. We transformed the annotated semantic type to its corresponding semantic group (e.g., “Sign or Symptom,” “Finding,” and “Disease or Syndrome” semantic types are converted to the “Disorder” semantic group). Then, we performed a comparison to determine whether the semantic groups are equal (the textual span needs to be the same). Finally, the fourth version of IAA was *relaxed*, i.e., we considered partial textual spans (overlaps) and semantic groups at the same time.

$$IAA = \frac{matches}{matches + non\_matches} \tag{1}$$

Boisen et al. (2000) recommend that only documents with an acceptable level of agreement should be included in the gold standard, and we followed their recommendation. However, because of the scarcity of this kind of data in pt-br bio-NLP research, and as the limited amount of annotated data is often a bottleneck in ML

(DOMINGOS, 2012), we did not exclude documents from our corpus, but opted to segment it in two, namely gold and platinum. This division was made based on the IAA values of each annotated document, where documents with an IAA greater than 0.67 belong to the gold standard and all the other ones to platinum. We picked the 0.67 threshold because it is the one Artstein and Poesio (2008) discussed to be a tolerable value. Additionally, we think the 0.8 threshold is rigorous, considering the complexity of our task and the number of persons involved in it. The task complexity is explained by the heterogeneity of the data, which are obtained from multiple institutions, various medical specialties, and different types of clinical narratives. The study that comes closest in data diverseness to ours is Patel et al. (2018), with the exception that their data come from a single institution. Moreover, despite the large amount of data they used, there are differences between their study and ours; for example, they used a coarse-grained annotation scheme by grouping the semantic types, which made the labeling less prone to errors. Moreover, we believe that a significant portion of errors that caused disagreements came from repeated mistakes on the part of one annotator in the pair, and owing to this, the error could be easily corrected by the adjudicator, as the examples in the following sections reveal.

3.2 NAMED ENTITY RECOGNITION FOR CLINICAL TEXTS

This section displays three distinct NER methods focusing on clinical texts assessed in this project and trained with SemClinBr corpus. The first one was based on a Condition Random Fields classifier (CRF), the second uses a BiLSTM-CRF architecture in combination with contextual character embeddings, and the third explored the use of BERT contextual embedding fine-tuned with clinical narratives texts. The evaluation serves as SemClinBr's extrinsic evaluation and, at the same time, to measure the performance of a tool that should be applied to the summarization task.

3.2.1 CRF method

The SemClinBr corpus was transformed into IOB2 labeling format, as shown in the example below, which shows the sentence "*Cerebral infarction due to thrombosis of pre-cerebral arteries"* in IOB2 format.

Cerebral [**B-Disease or Syndrome**]

infarction [**I-Disease or Syndrome**]

due [O]

to [O]

thrombosis [**B-Disease or Syndrome**]

of [**I-Disease or Syndrome**]

pre-cerebral [**I-Disease or Syndrome**]

arteries [**I-Disease or Syndrome**]

A set of morphologic and orthographic features were extracted from the text. The features followed Al-Hegami et al. (2017) biomedical NER model, and are presented in Table 12. Due to the importance of POS-tagging in feature-based NER approaches (GÜNGÖR; GÜNGÖR; ÜSKÜDARLI, 2019), we decided to improve the extraction of this feature, that has been extracted by a POS-Tagger trained over journalistic texts and presenting regular results in clinical texts labeling (more details in section 3.3).

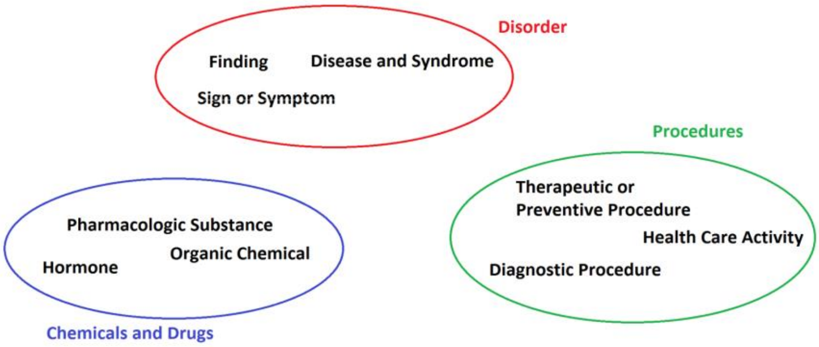
Table 12 - The relation of all the features extracted from the text and inputted to the CRF classifier.

Nº	Feature	Nº	Feature
1	Word is lowercase	9	Word is uppercase
2	Word has more than two consecutive consonants	10	Maximum number of consecutive consonants
3	Word begins with a capital letter	11	Word is at the beginning of the sentence
4	All words in sentence are uppercase	12	All words in sentence are lowercase
5	Number of letters in the word	13	Word has accent
6	Number of word vowels	14	Word has only numbers.
7	Word has no vowels	15	Part-of-speech Tag
8	Maximum number of consecutive vowels	16	Word is at the end of the sentence

Source: de Souza et al. (2019).

The experimental setup included the classification of the most annotated UMLS semantic types, and three major semantic groups (i.e., Disorder, Procedures and Chemicals and Drugs), as shown in Figure 17. A 10-fold cross-validation evaluation model was applied over the corpus, and Precision, Recall, and F1-scores for each semantic types and groups were assessed. The passive-aggressive CRF classifier variant was used.

Figure 17 – The semantic groups used in the experimental setup and their respective semantic types insides the ellipses.

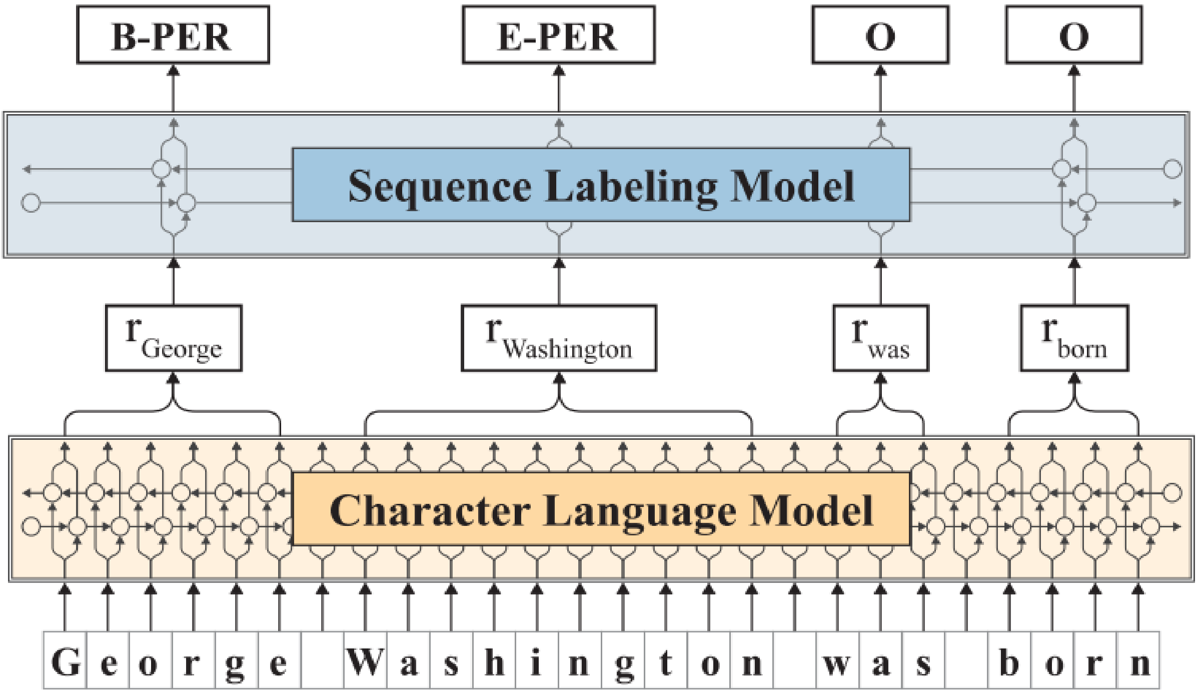


Source: de Souza et al. (2019).

3.2.2 BiLSTM-CRF + Character Embeddings method

The second method relies on a sequence labeling environment proposed by Akbik et al. (2018). Their study exploited a BiLSTM-CRF neural network architecture developed by Huang et al. (2015), input with novel contextual character-level word embeddings, as displayed in Figure 18.

Figure 18 – An overview of the proposed architecture. A pre-trained bidirectional character language model receives a sentence as a character sequence and passes a contextual embedding into the BiLSTM-CRF sequence labeler, which outputs the resulting labels.



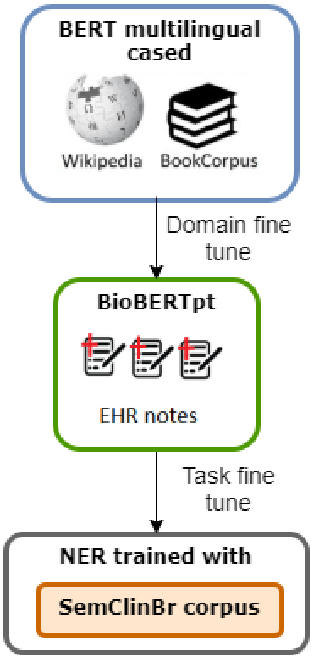
Source: Akbik et al. (2018).

To generate the contextual embeddings with the Character Language Model, we used a pre-trained model, the WikiNER, a dataset automatically generated from Wikipedia articles in Portuguese (NOTHMAN et al., 2013). The experimental setup included the same semantic types and groups from the previous method and applied a 10-fold cross-validation evaluation to assess Precision, Recall, and F1-score.

3.2.3 Clinical fine-tuned BERT method (BioBERTpt)

The last NER method also relies on the architecture of neural networks and contextual embeddings; however, using the BERT implementation developed by Hugging Face and a pre-trained model fine-tuned for the clinical domain. To create a clinical model (named BioBERTpt), we used the original database acquired for SemClinBr, with more than two million EHR notes (Figure 19 shows the method overview).

Figure 19 - An overview of BioBERTpt development process and NER experiments. The green border is the new contextual embeddings created by fine-tuning the existing BERT multilingual cased over 2 million clinical notes and evaluated in the NER task using the SemClinBr corpus.



Source: Adapted from Schneider et al. (2020).

We split the notes into sentences and tokenize them with *BertWordPieceTokenizer*. We fine-tuned the BERT multilingual cased model[9] with the clinical data using five epochs, a batch size of 4, learning rate of 2e-5, and 512 as block size, and "*mlm*" value for the Masked Language Model. Then, we experimented with the fine-tuned model into the NER task. The evaluation method used was a holdout with a corpus split of 60% for training, 20% for validation, and 20% for test. We used the Hugging Face API that provides the *BertForTokenClassification* class, which adds a token-level classifier, a linear layer that uses the last hidden state of the sequence. We calculated the Precision, Recall, and F1-score for each semantic type and micro-averaged the global scores. More details will be published in an in-press paper (SCHNEIDER et al., 2020).

3.3 DEFINING A STATE-OF-THE-ART POS-TAGGING ENVIRONMENT FOR BRAZILIAN PORTUGUESE CLINICAL TEXTS

In this section, we present the assembling of a clinical POS-Tagging environment aiming to improve the quality of NER's feature extraction and support the development of NLP and ML methods used to build information extraction and clinical decision support systems.

3.3.1 Data acquisition

A POS-Tagging annotated corpus, developed by Peters et al. (2010), was used to train and evaluate the proposed model. The corpus is composed of a fraction of randomly selected data from two different collections of discharge summaries, totaling 123,018 tokens, 5,964 sentences, and 692 notes annotated (Table 13 shows annotated examples). To annotate the texts, they employed an active learning approach supported by an ensemble of general-domain POS-taggers. Additional experiments with biomedical, journalistic, and multiple-domain corpora are described in de Oliveira et al. (2020), as well as the tagset normalization process over all corpora.

[9] https://github.com/google-research/bert/blob/master/multilingual.md

3.3.2 BiLSTM-CRF + Character Embeddings architecture

Due to the lack of studies focusing on clinical POS-Tagging for pt-br based on a modern neural network model, we decided to utilize one of the supervised approaches applied to the NER task, consisting in a BiLSTM-CRF architecture in conjunction with contextual character embeddings, which achieved state-of-the-art results for many sequence labeling texts in multiple languages (AKBIK; BLYTHE; VOLGRAF, 2018).

Concerning the contextual character embeddings employed in Akbik's study, it is particularly important in our POS-Tagging task, because it can capture word meaning in context and build different embeddings for words with multiple meanings depending on their usage (i.e., polysemous words). Moreover, as they model words and context as sequences of characters, they can better manipulate rare and misspelled words, and model subword structures such as prefixes and endings. That is, the contextual character embeddings cover some of the intrinsic issues in clinical texts, which count with an elevated number of orthographic errors and high use of words containing similar prefixes and endings (especially diseases and medications).

Table 13 - Samples of discharge summaries annotated with morphological tags. The actual words are before the underline, and the tags are after, in a WORD_TAG schema. The first column shows the original pt-br text with the respective POS-tags and, in the next column, the translated version of the text. "N" stands for NOUN, "V" for VERB, "PREP" for PREPOSITION, "ADJ" for ADJECTIVE, "ART" for ARTICLE, "CJ" for CONJUNCTION, and "NUM" for NUMBER.

Original sample with POS-tags	Translated sample
Paciente_N foi_V a_PREP indução_N com_PREP ocitocina_N 6_NUM horas_N após_PREP último_ADJ misoprostol_N	Patient was induced with oxytocin 6 hours after last misoprostol
História_N de_PREP dispnéia_N a_PREP os_ART moderados_ADJ esforços_N e_CJ angina_N noturna_ADJ	History of dyspnea on moderate exertion and nocturnal angina
Recebe_V alta_N em_PREP bom_ADJ estado_N geral_ADJ ferida_N operatória_ADJ seca_ADJ com_PREP boa_ADJ cicatrização_N	Discharged in good general condition dry surgical wound with good healing

Source: adapted from de Oliveira et al. (2020).

The Character Language Model used a pre-trained model, the WikiNER (NOTHMAN et al., 2013). The training parameters included a learning rate of 0.1, a mini_batch_size of 32, and the maximum epochs defined as 100. The evaluation metrics calculated were Accuracy and F1-score. The corpus was split into three parts, 60% for training, 20% for testing, and 20% for validation.

4 CHRONIC DISEASE SUMMARIZATION

In this chapter, we present the incremental development of a chronic disease summarization system, where we experimented with summarization baseline algorithms, unsupervised and supervised approaches, built a new clinical summarization corpus, and developed a semantic similarity algorithm.

It is worth remembering that due to the gaps, issues, and challenges associated with the summarization research field, and the scope of this thesis, we firstly prioritized unsupervised methods that do not need any resource or tool currently unavailable for pt-br. Methods that already been used in the biomedical summarization context; or, known to be language- or domain-independent, were prioritized as well, including the supervised approaches.

We did not experience with the majority of the studies that already attempted to produce a summary of chronic disease patients (presented in section 2.5.2). This is due to the fact that most of the studies used only structured data as input to their summarizer (HSUEH et al., 2015; GOLDSTEIN; SHAHAR, 2016; ALEKSIĆ et al., 2017), while our main focus relies on unstructured data. Although Acharya et al. (2018) used both types of data, their method depends on tools to map words to ontological concepts and produce natural language, which is an issue for us due to language limitations. Pivovarov (2015) used SNOMED-CT (which are not translated to pt-br yet) to solve redundancy problems, focused on phenotyping, and to assess the summarizer.

4.1 PRELIMINARY EXPERIMENTS – UNSUPERVISED SUMMARIZATION BASELINES

This section presents the first summarization algorithms experiments on the patient’s data. The idea was to investigate the peculiarities of some **unsupervised baseline summarization approaches** when applied on longitudinal EHR information, by simply analyzing the resulting outputs, and with that, try to identify text aspects that could guide the development of the main summarization system.

As input to the algorithms, we used the EHR entries of a chronic kidney disease patient composed of 15 ambulatory notes, written during almost five years. And to better understand the summarization outputs, we produced two human-made summaries to represent the history of this patient (shown in Table 14). The first, an abstractive one, in which the human read all the notes and produced a text containing what is considered essential to know about the patient, with the orientation to make it as brief as possible. The second summary was made by manually extracting important sentences from the notes to form a new text, again with an orientation to make as short as possible, and with a 15 sentences limit. More details on the creation of these human-made summaries are given in the section corresponding to the creation of the summarization corpus (section 4.3).

Table 14 – Human-made extractive and abstractive summaries based on EHR entries of a chronic kidney disease patient.

Abstractive	Extractive
PACIENTE EM ACOMPANHAMENTO NO AMBULATORIO DE NEFROLOGIA POR DRC POR NEFROESCLEROSE HIPERTENSIVA. ECOGRAFIA MOSTRA CISTO SIMPLES NO RIM D, REDUCAO VOLUMETRICA DE AMBOS OS RINS. NEGA DM, NEGA IAM/AVC. HAS EM USO DE ATENSINA 100MG, ANLODIPINO, HIDROCLOROTIAZIDA.	HAS HÁ MAIS DE 10 ANOS. NEGA DM. DRC ESTAGIO IV - POSSIVEL NEFROPATIA HIPERTENSIVA. CR: 2.1 TFG 17 ML/MIN COM ALBUMINA DE 429MG E CALCIO IONICO DE 1.44. US DE AP. URINARIO: (_DATA_). RIM DIREITO COM COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE REDUZIDOS. NAO HA EVIDENCIA DE DILATACAO DO SISTEMA. COLETOR, AUSENCIA DE IMAGEM DE LITIASE. RIM MEDINDO 8,6X3,8X4,2. RIM ESQUERDO COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE REDUZIDOS. NAO HA EVIDENCIA DE DILACATAO. Em uso: ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA. DRC POR NEFROESCLEROSE HIPERTENSIVA BENIGNA COM TFG ESTIMADA EM 24.

Source: the author.

We experimented the patient’s data with five different baseline methods, which are: LexRank (ERKAN; RADEV, 2004), TextRank (MIHALCEA; TARAU, 2004), LSA (STEINBERGER; JEŽEK, 2004), SumBasic (VANDERWENDE et al., 2007) and KLSum (HAGHIGHI; VANDERWENDE, 2009). The algorithms were parameterized to produce summaries with 10 sentences. The output summaries of the methods are presented in Table 15.

Table 15 - Output summaries for fifteen ambulatory notes from a chronic kidney disease patient. The overlap information between these and the human-made summaries are highlighted in bold.

Algorithm	Output summary
LexRank	**Hipertensa com lesão renal** retorna com exames. BEG,bradipsiquica,sem edema,PS:110/70mmHg. Comorbidade: **HAS há 10 anos**, dislipidemia, Normotensa sem uso de medicações, porém com história prévia de HAS> vitD 51,7 Ca 1,28 Na 141 S# Paciente sem queixa no momento da consulta.Sem alterações urinárias ou edema em MMII Exames Lab (11/07/2016): vit D: 27,8; Fósforo urnário: urina de 24h (0,6) e urina isolada (60,3) MUC# **ANLODIPINO 5 mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA** Prescrevo Vit D Mantenho medicacoes Retorno em 4 meses com novos exames refere esporadicamente pouco edema em MMII Refere bom controle de PA em domicilio. #nega alergias #Em uso: **ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA** #A: **DRC Estagio IV - ClCr: 17 (CKD-EPI) POR HAS**
TextRank	Solicito dosagem vitamina D Retorno em 3 meses com exames Hx: filha com nefrolítiase, pai, falecido com 60 anos, de morte súbita, mãe falecida com 100 anos. Exames 20/02 Gasometria: ph 7,33; pCO2 50,7; pO2 21,5; HCO3 26,2 BE -0,3; Ca 1.33; reticulócitos 1,1% Hemograma: Hb 13,5 Ht 41,3%; leucócitos 7640, plaq 245000 **Creatinina 2,4**; Glicose 77; Ureia 123, Mg 2; Albumina 4,1; FA 110; triglicerídeos 167; HDL 38; Na 139; K 5,1; colesterol total 175; fósforo 4,2; LDL 104, Fe 68; ferritina 52,5; PTH 248,9; V B12 265; ácido fólico >20; Vit D 38,9; Parcial de urina normal; índice de saturação de transferrina 18,4% Normotensa sem uso de medicações, porém com **história prévia de HAS**> vitD 51,7 Ca 1,28 Na 141 S# Paciente sem queixa no momento da consulta.Sem alterações urinárias ou edema em MMII Exames Lab (11/07/2016): vit D: 27,8; Fósforo urnário: urina de 24h (0,6) e urina isolada (60,3) #Em uso: **ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA** Prescrevo Vit D Mantenho medicacoes Retorno em 4 meses com novos exames #Em uso: **ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA** Mantenho medicacoes Retorno em 4 meses com novos exames #nega alergias #Em uso: **ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA**
LSA	Ecografia:cisto simples no rim D/redução volumetrica de ambos os rins/sedimento mamário movel no interior da bexiga.Hb:13,7/**creatinina:2**/glicose:80/ureia:111/sorologias negativas/ **clearance decreatinina :33,2ml/m** Paciente em acompanhamento no ambulatório de nefrologia por **DRC devida a nefroesclerose hipertensiva**. Solicito dosagem vitamina D Retorno em 3 meses com exames Paciente portadora de **DRC secundaria a hipertensão**, retorna a consulta para reavaliação Exames 20/02 Gasometria: ph 7,33; pCO2 50,7; pO2 21,5; HCO3 26,2 BE -0,3; Ca 1.33; reticulócitos 1,1% Hemograma: Hb 13,5 Ht 41,3%; leucócitos 7640, plaq 245000 **Creatinina 2,4**; Glicose 77; Ureia 123, Mg 2; Albumina 4,1; FA 110; triglicerídeos 167; HDL 38; Na 139; K 5,1; colesterol total 175; fósforo 4,2; LDL 104, Fe 68; ferritina 52,5; PTH 248,9; V B12 265; ácido fólico >20; Vit D 38,9; Parcial de urina normal; índice de saturação de transferrina 18,4%

	Normotensa sem uso de medicações, porém com **história prévia de HAS**> vitD 51,7 Ca 1,28 Na 141 S# Paciente sem queixa no momento da consulta.Sem alterações urinárias ou edema em MMII Exames Lab (11/07/2016): vit D: 27,8; Fósforo urnário: urina de 24h (0,6) e urina isolada (60,3) Prescrevo Vit D Mantenho medicacoes Retorno em 4 meses com novos exames paciente refere dor lombar bilateral, que iniciou apos subir no telhado, com melhora progressiva. #nega alergias #Em uso: **ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA**
SumBasic	Sem queixas. **creat 2,1** ureia 96 NOME_PACIENTE, 69 anos MUC# **anlodipina5 mg, sinvastatina 20 mg, furosemida 40mg**. S# Paciente sem queixa no momento da consulta.Sem alterações urinárias ou edema em MMII Exames Lab (11/07/2016): vit D: 27,8; Fósforo urnário: urina de 24h (0,6) e urina isolada (60,3) P# **Losartana 25mg 1cp/ dia.** Refere bom controle de PA em domicilio. #nega alergias Mantenho medicacoes Retorno em 4 meses com novos exames #A: **DRC Estagio IV - ClCr: 17 (CKD-EPI) POR HAS**
KLSum	Ecografia:cisto simples no rim D/redução volumetrica de ambos os rins/sedimento mamário movel no interior da bexiga.Hb:13,7/**creatinina:2**/glicose:80/ureia:111/sorologias negativas/ clearance decreatinina :33,2ml/m BEG,bradipsiquica,sem edema,PS:110/70mmHg. ** restante do atendimento encontra-se no prontuário físico do paciente** sem sinais ou sintomas de uremia Nega queixas. BCRNF sem sopros MV presente, sem RA ABD: flácido, plano, indolor, RHA +, sem visceromegalias MMII: pulsos presentes, sem edema Exames 20/02 Gasometria: ph 7,33; pCO2 50,7; pO2 21,5; HCO3 26,2 BE -0,3; Ca 1.33; reticulócitos 1,1% Hemograma: Hb 13,5 Ht 41,3%; leucócitos 7640, plaq 245000 **Creatinina 2,4**; Glicose 77; Ureia 123, Mg 2; Albumina 4,1; FA 110; triglicerídeos 167; HDL 38; Na 139; K 5,1; colesterol total 175; fósforo 4,2; LDL 104, Fe 68; ferritina 52,5; PTH 248,9; V B12 265; ácido fólico >20; Vit D 38,9; Parcial de urina normal; índice de saturação de transferrina 18,4% PTHi 239 urina I --> Hem 2 /ml Leuc 1/ml nega queixas ou alterações urinarias, febre, dispneia. #nega alergias #Em uso: **ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA**
Frequency-based	PAROU HÁ 6 ANOS. NOME_PACIENTE, 69 ANOS #**HAS HÁ MAIS DE 10 ANOS**. **RIM MEDINDO 8,6X3,8X4,2** **RIM ESQUERDO COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE REDUZIDOS**. **NEGA DM** #**DRC ESTAGIO IV - POSSIVEL NEFROPATIA HIPERTENSIVA** #EX TABAGISTA POR 25 ANOS 4 MAÇOS MES. RIM MEDINDO 7,9X 3 X 3,2 BEXIGA COM FORMATO HABITUAL, PAREDES LISAS E SEM CONTEUDO ANOMALO.

Source: the author.

Additionally we developed a straight-forward frequency-based method, which counts the words found in patient’s EHR, and then score each sentence based on its words. More frequented words, produce higher scores. To avoid over-scoring for longer sentences, each sentence received an average score of all its words. A redundancy checker based on Jaccard Distance was added to the pipeline, to avoid multiple equal sentences to be selected. The sentences with higher scores are selected to form the final summary.

It is worth mentioning that we used the Natural Language Toolkit’s (NLTK) word and sentence tokenizer, which, due to the known orthographic issues in clinical texts (e.g., lack of punctuation, fragmented sentences), did not work properly, sometimes combining multiple sentences into one, and then, generating summaries larger than the ten sentences limit.

It is possible to check that most of the information in the summaries were not considered relevant by humans (except on frequency-based-method), and most of the overlap information is from medications and lab results, probably due to the redundancy of this information in multiple texts and the elevated number of related terms in the same sentence. The repetition of medication prescriptions is another point to be highlighted.

Another important aspect to highlight is that SumBasic and Frequency-based algorithms ended up selecting shorter sentences than the other methods, which in general, selected longer sentences. The final size of the summary could be a challenge in the task of chronic-disease summarization since there may be patients with a large amount of data in the EHR, however, few relevant data associated with their condition. In other cases, on the other hand, the patient may have few visits to the doctor; however, it may contain a dense history of risk factors important to his condition. Therefore, the output summary size is not always associated with the patient’s EHR size. In addition, due to the large variation in the size of the sentences in the medical record, the resulting summary can vary greatly in size, as can be seen in the summary generated by the SumBasic and Frequency-based methods, both based on frequency of terms.

The non-linearity or lack of a “timestamp” of the events presented in the summaries could be a problem because, for the health professional, it is important to understand the order of events. For example, to identify adverse effects to medications, to understand whether the condition has improved or worsened according to symptoms and test results, among other needs. Moreover, it is not easy to find, for example, comorbidities or symptoms of the patient because the text is not organized or labeled by categories. Thus, it is difficult to trust in a summary with such a level of repeated, unordered, and unidentified information, leading us to explore other approaches, including the supervised ones and trying to solve the aforementioned issues.

4.2 UNSUPERVISED APPROACHES

In a previous section (i.e., Methodological Challenges and Approaches), we listed studies that could possibly be used for chronic disease summarization, including mainly unsupervised approaches (see Table 4 and 5). This section depicts the methodological challenges in applying these studies to our task and details the adoption of a contextual embeddings approach to perform the unsupervised summarization.

The methods presented by Moen et al. (2016) were very appealing because they do not need any external knowledge resource, except word embeddings, that could be easily generated by currently available tools. Moreover, they already validated their methods in a chronic-condition environment. The only drawback is that the principal methods need an original discharge summary of the patient to score the sentences in the care episodes, which is unavailable in our dataset, which precluded the application of the method.

Like the previous study, Litvak et al.'s (2013) work is knowledge-free. Their summarization algorithm is language-independent as well (evaluated in two different languages). One challenge is to adapt a single-document algorithm to deal with the redundant aspect of a patient's longitudinal data. Furthermore, the method is somewhat similar to the graph-based TextRank, which been used in the previous section, with poor results.

4.2.1 BERT-based extractive summarization + biomedical fine-tuned model

This section describes a BERT-based method for extractive summarization applied over the chronic-disease patient's data, shows its advantages over the previously mentioned unsupervised studies, and presents the experimental setup defined for evaluation.

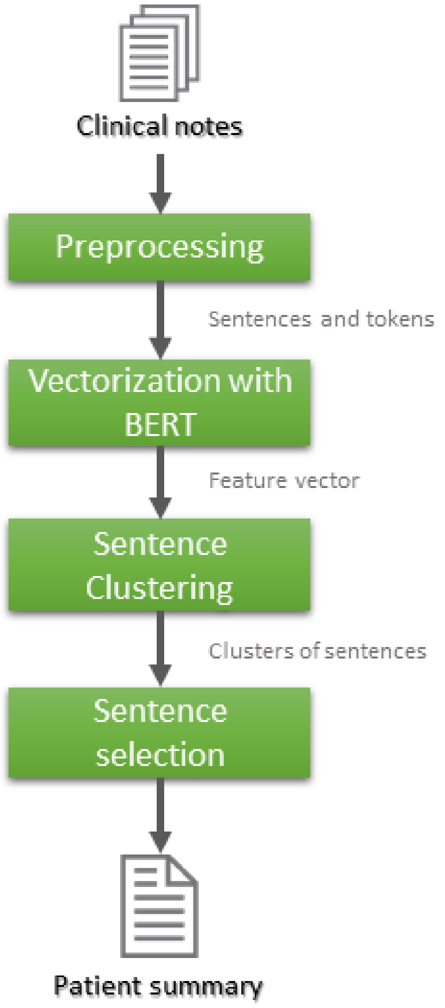
The application of BERT to the summarization task could be justified in many ways, as the BERT model achieved state-of-the-art results for several NLP tasks, including in the clinical domain. We can take advantage of huge pre-trained models available for various languages and tasks. Moreover, as one of the main challenges of clinical summarization is the redundancy, is expected that the word embeddings produced by BERT could deal with it, as we can throw the clinical concepts into a vector space and follow the intuition that concepts that appear in similar contexts are similar in some way.

We explored the work of Miller (2019), which is very similar to Moradi and Samwald (2019) as both apply the sentences to BERT, and then apply clustering techniques to define the candidate sentences. The great advantage of this method over those previously listed is that the method is completely unsupervised, and it does not require any additional tools or semantic resources. The only resource needed would be a pre-trained BERT model for the desired language, which we already have available for use in pt-br (SOUZA; NOGUEIRA; LOTUFO, 2019). In addition, we also have an improved BERT model for use in the biomedical domain, which was described in section 3.2.3. That is, although the method was developed to summarize lecture notes in English, it can be easily adapted to other languages and domains by making use of the corresponding pre-trained model.

Another advantage of Miller's method is that it is possible to create dynamic summaries, with variable sizes instead of fixed-length ones. Since the summary output size is uncertain in our task due to different levels of complexities and events associated with a patient, it is essential that we could change the output size dynamically.

The method executes three steps of AORTIS model: **aggregation**, **organization**, and **reduction** (Figure 20 shows a simplified diagram of the method). It receives the text as input and runs the following steps: (i) tokenize the sentences, (ii) input the sentences into a BERT implementation to generate the corresponding vector space representations of each sentence, (iii) cluster the embeddings with K-Means algorithm, (iv) select the embedded sentences that are closest to the centroid as candidates to compose the final summary, and (v) remove candidate sentences with additional context aspect (e.g., sentences beginning with conjunctions), and remove small and large sentences.

Figure 20 - An overview of the unsupervised BERT-based summarization method phases.



Source: the author.

We needed to adapt some method steps to fit our problem, including the sentence tokenizer, which we replaced by a custom algorithm as the lack of punctuation, incorrect uppercasing, and fragmented sentences in the clinical environment were causing several tokenization problems.

The algorithm was adapted to utilize the BioBERTpt, the BERT model fine-tuned for pt-br biomedical texts presented in a previous section. Thereby, the BioBERTpt model generates a more cohesive vector representation as it was trained in a domain with a larger vocabulary overlap with the clinical domain and in a more representative context. The last removal step was eliminated because the clinical narratives contain minimal size sentences that are very relevant (e.g., “#HAS”, which means “Patient with systemic arterial hypertension”).

Table 16 - Summaries generated by both original and adapted BERT-based summarizer (MILLER, 2019). The overlap information between these and the human-made summaries are highlighted in bold.

Version	Output summary
Original	**Hipertensa com lesão renal** retorna com exames. **Creatinina = 2,2** **CLCr CKD = 22 ml/min** PTH = 199 Solicito dosagem vitamina D Retorno em 3 meses com exames ** restante do atendimento encontra-se no prontuário físico do paciente** NOME_PACIENTE, 67 ANOS, **QP: DRC** EM ACOMPANHAMENTO NO AMBULATORIO HA 1 ANO POR **DRC**. NAO HA EVIDENCIA DE DILATACAO DO SISTEMA COLETOR, AUSENCIA DE IMAGEM DE LITIASE. **CR 1,8 TFG (CKD-EPI) 28,2** UR 73 O# FC 70 PA 110/70 BCRNF SEM SOPROS MV PRESENTE BILATERALMENTE SEM RA MMII SEM EDEMA #A: **HAS** **DRC Estagio IV - ClCr: 28,2 (CKD-EPI)** #P: Prescrevo Vit D Mantenho medicacoes Retorno em 4 meses com novos exames NOME_RESIDENTE NOME_MEDICO NOME_PACIENTE, 71 ANOS **#HAS HÁ MAIS DE 10 ANOS**.
Adapted	**Hipertensa com lesão renal** retorna com exames Conduta:mantida medicação/Urolo0gia/retorno com exames EXAMES: **CR: 2.5 U: 126 CLEARENCE: 25 ML/MIN NEGA DM** NAO HA EVIDENCIA DE DILATACAO DO SISTEMA COLETOR, AUSENCIA DE IMAGEM DE LITIASE AUSENMCIA DE LITIASE RIM MEDINDO 7,9X 3 X 3,2 HD: MMII SEM EDEMA **#HAS HÁ MAIS DE 10 ANOS** O# FC 62 PA 120/74. **DRC ESTAGIO IV - POSSIVEL NEFROPATIA HIPERTENSIVA.**

Source: the author.

In Table 16, we present two versions of generated summaries, one using the original method and another with the adapted version of it. The patient used in this example is the same presented in Table 14, in the previous section. Note that due to the erroneous sentence tokenizer, the original method generated much longer summaries.

To quantitatively evaluate the model, we applied the ROUGE metrics over the resulting summaries of ten different patients, and utilizing three distinct summary sizes, containing 10, 15, and 20 sentences. The gold standard used for comparison is detailed in section 4.3.

4.3 SUMMCLINBR – A CORPUS FOR PATIENT'S LONGITUDINAL DATA SUMMARIZATION

In this section, we detail the development steps of a corpus for a patient's longitudinal data summarization (i.e., SummClinBr), focusing on chronic disease. The corpus could be used for training of supervised summarization algorithms, and for testing in both supervised and unsupervised approaches.

4.3.1 Data acquisition

The data was retrieved from the group corpus (the same used in SemClinBr), and filtered by medical students, following instructions of two nephrologists. Although we believe in the hypothesis that the annotation methodology used in the construction of this corpus can be generalized, the initial focus was on Chronic Kidney Disease (CKD). We selected only patients who have not yet undergone hemodialysis, for two reasons: (i) because we have limited access to additional patient's data, related to hemodialysis (which are not available in the group corpus), and (ii) because we believe that they have many specificities of CKD, which cannot be generalized to chronic diseases of other specialties (e.g., cardiovascular diseases and diabetes).

In addition, we will focus on outpatient data with stage 3B and 4 CKD patients, as the hospitals that provided the data serve only patients from stage 3B onwards, who are referred from the Basic Health Unit (i.e., primary care). Table 17 shows the partial history of a patient (3 clinical notes) since his referral from the Basic Health Unit (primary care), investigation of the condition and diagnosis of chronic kidney disease, and follow-up consultation. It is possible to verify that the **clinical notes vary widely in size, format, completeness and phase of care**. These aspects could differ depending on the doctor who is attending and the stage of patient care (e.g., first visit, diagnosis, follow-up).

Table 17 - Patient’s history through a set of three clinical notes.

Visit outline	Clinical note
The patient’s first visit referred from the Basic Health Unit. Investigation in course.	Paciente encaminhado da unidade basica devido inapetencia e fraqueza, com historico de anemia. Em exame anterior observado sangue oculto nas fezes. Conduta: realizado pedido de colonoscopia e retorno com exames.
Patient’s second visit. The main diagnosis given. Other investigations in course.	DRC DE ETIOLOGIA INDEFINIDA PARCIAL NORMAL ECO COM CISTO RENAL SIMPLES A DIREITA CREATININA DE 2.5 COM CLEARENCE ESTIMADO DE 24 ML/MIN PA: 110/70 K: 5.8 ATR? HIPOALDO? SOLICITO RENINA ALDOSTERONA E GASO VENOSA RETORNO COM EXAMES EM 3 MESES
Patient’s fifth visit. Follow-up consultation.	NOME_PACIENTE, 71 ANOS #DRC ESTAGIO 4 DE CAUSA INDETERMINADA #SEM COMORBIDADES #NEGA TABAGISMO #NEGA ETILISMO #S: SEM QUEIXAS NO MOMENTO. RETORNA COM EXAMES #O: BEG, LOTE, HIPOCORADO, HIDRATADO PA:140/80 FC:83 BCRNF RCR 2T MV+ SRA SEM EDEMA DE MMII EXAMES: US 04/06/13: RIM D: 9 X 4 X 3 RIM E: 9 X 5X 3 CISTO RENAL SIMPLES À DIREITA LAB:28/05/15: MICROALBUMINURIA: 19.2 Cr: 2.9 Ur: 59 Hb: 11,3 Ht: 34,1 Na: 139 K: 6,2 P: 3,8 PTH: 123,7 FA: 85 VIT. B12: 161 CALCIO IONICO: 1,21 CKDEPI: 21ML/MIN #A: DOENÇA RENAL CRÔNICA #P: PRESCREVO CITONEURIN 5000 1CP VO 12/12 POR 3 MESES, SOLICITO AVALIAÇÃO COM A NUTRIÇÃO PARA AVALIAR INGESTA DE POTÁSSIO, SOLICITO EXAMES LABORATORIAIS, RETORNO EM 6 MESES

Source: the author.

At some point in the process, if the use of the above criteria limits the amount of data too much, we allow the selection of transplanted patients and with other related diseases as Diabetic Nephropathy. As the objective is to develop a summarization algorithm to capture longitudinal information, we selected patients with a minimum of five clinical notes, and more than half of notes should be associated with the Nephrology department. In Table 18, the numerical information related to the collected data is displayed.

Table 18 – SummClinBr corpus numerical overview.

Information	Number
Number of Patients	41
Number of Clinical notes	471
Average notes per patient	~11
Average character size per note	1907
Average token size per note	198

Source: the author.

4.3.2 Annotation guidelines

The generation of a corpus for a summarization task could be performed in different ways, such as sentence salience binary annotation (i.e., sentence should be selected or not), human-generated summaries (often used for abstractive approaches) or even annotation of important events/concepts within the text (summarization as a sequence labeling task).

Due to the incremental facet of this study, in which several summarization approaches could be tried, it was essential to build a gold standard that could be explored by different methods, both for training and for testing the algorithms. Therefore, three levels of annotation were generated: (i) **concept-level annotation** of events and problems related to the chronic disease, following a set of categories established by specialists, (ii) **sentence-level annotation** where the sentences considered relevant to the patient’s final summary were labeled (extractive summary), and (iii) the writing of an **abstractive summary**, in which the annotator describes with his own words the relevant information of the patient.

The SemClinBr corpus utilized the set of UMLS semantics types as the annotation tags; however, for this endeavor, we opted for a new group of tags that better represented the information related to chronic diseases and that did not have a high level of granularity such as the UMLS semantic types. Together with two nephrologists, we examined a set of real clinical cases and their EHR entries. In this analysis, it was very clear the overlap that chronic kidney disease has with other chronic conditions, such as cardiovascular diseases and diabetes. We follow the intuition that chronic diseases from other medical specialties (e.g., cardiology, endocrinology) have a large set of data in common with chronic kidney disease. Based on this, we defined a series of data types that would be relevant in a summary of the chronic patient's condition. In Table 19, we present the SummClinBr’s tagset for **concept-level annotation**, which is available through the Annotation Guidelines document[10] as well.

It is important to highlight that for this concept-level annotation, the guidelines guide the annotator to mark all occurrences of the tags in any of the patient's documents. That is, the concept-level gold standard helps us to understand what are the important concepts in the text and what are their writing standards, but it does not define what data should be shown in the final summary. Because if we show all the annotated data, we would probably have an output with overload and duplication of information.

Still, regarding the concept-level annotation, two additional attributes should be defined at each labeled concept. The first one is the **polarity**. In case a concept is negated in the text (e.g., *patient denies diabetes*), the annotator should mark it with the negative polarity. The second attribute is the **relevancy** that has two possible values: essential and important. A concept is marked as essential when the annotated concept is essential and mandatory to appear in a possible final summary (e.g., the patient progressed to CKD stage 4). The important label is used when the concept is important, but if the summary final size is already fulfilled, the concept can be ignored (e.g., *weight 75kg*).

The **sentence-level annotation** generated human-made extractive summaries for each patient. The annotator selected the sentences in the patient’s EHR that would compose a final chronic-disease summary. The annotators were oriented to select a maximum of 15 sentences per patient, which is almost the average clinical note size,

[10] http://bit.ly/2WO0DSN

and as discussed with the nephrologists, could probably represent all the relevant information regarding the patient's condition.

Table 19 - SummClinBr corpus tagset overview. Examples of each tag annotation are shown in the third column. The underline represents the labeled span.

Tag	Description	Examples
Comorbidity	Comorbidities related to the patient’s chronic disease. The appearance of new comorbidities can mean the worsening of the condition.	*HAS compensada* *Paciente com Diabetes descompensada* *Retinopatia* *Hipotireoidismo*
Exam results	Exam results related to the patient’s chronic disease and main comorbidities. It is important for the follow-up and to avoid duplicated exams.	*CR 1,55* *ureia: 122* *CKD-EPI 32 mL/min*
Medication	Medications in use by the patient. Important for follow-up.	*LEVOTIROXINA 100, ANLO 20, CONCARDIO 2,5mg* *PROPANOLOL, LOSARTANA* *Tratamento endovenoso com ATB*
Procedure	Important procedures associated with chronic disease and its comorbidities.	*Tx renal* *Quimioterapia* *Cateterismo* *Hemodiálise*
Clinical attribute	Clinical attributes for general monitoring of patient health.	*Peso 75* *PA 110/70*
Risk factor	Risk factors that may contribute to worsening the condition. Risk factors that have already been considered in any of the previous categories (e.g., comorbidities) should not be marked in this category.	*Pai com doença cardíaca* *Família com histórico de diabetes* *Metástase*
Last visit conduct	The conduct given by the physician to the patient in the last visit.	*Suspendo LEVOTIROXINA* *Encaminho para cardio* *Solicito ecocardio*

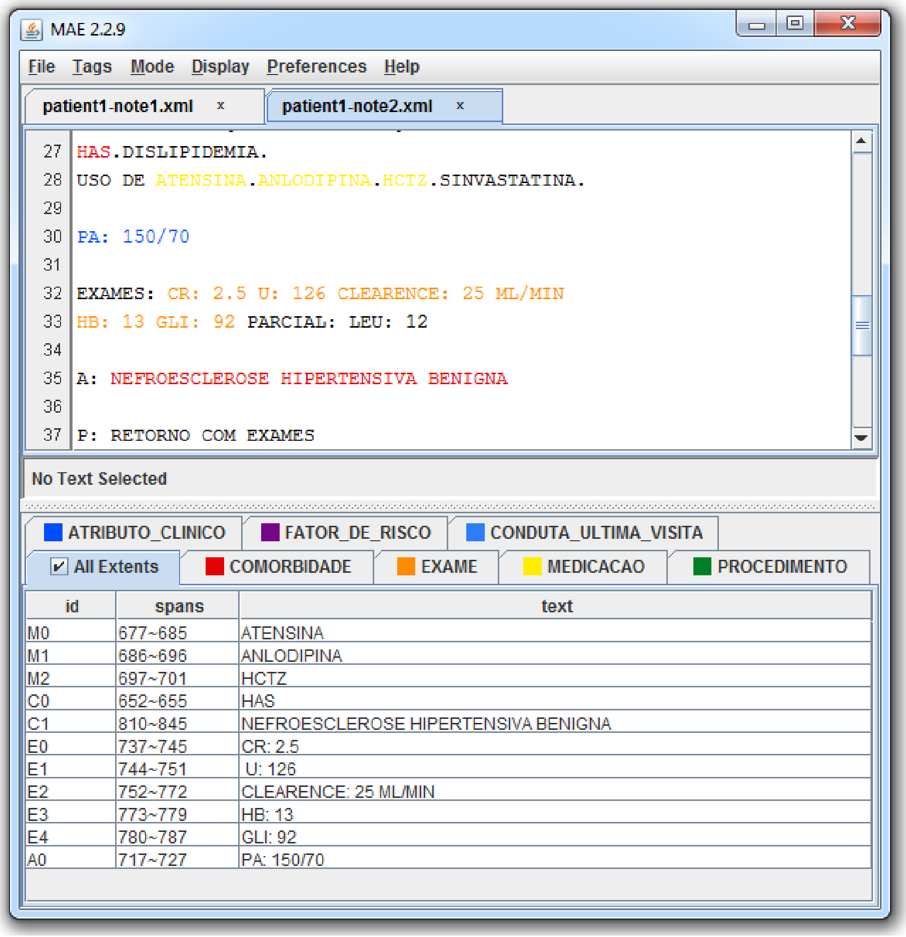
Source: the author.

Based on the extractive summary, the annotators were oriented to write an **abstractive summary** about the patient, however, with their own words. They were free to define the format and style of the text. The only orientation was to write the summary as concisely as possible. An example of the human-made summaries is shown in Table 14.

4.3.3 Annotation process

The annotation was initially performed by two medical students with ambulatory experience. They participated in an iterative training phase aiming to improve the guidelines, and check the consistency of annotation between annotators and provide feedback on their initial work (similar to the process shown in Figure 9). They used the Multi-Purpose Annotation Environment 2 – MAE2 (RIM, 2016) for the concept-level and sentence-level annotation, Figure 21 shows an example of concept-level annotation. The guideline improvements were discussed between the author of the thesis, the students, and two nephrologists.

Figure 21 – The MAE2 annotation tool been used to perform the concept-level annotation of SummClinBr.



```
HAS.DISLIPIDEMIA.
USO DE ATENSINA.ANLODIPINA.HCTZ.SINVASTATINA.

PA: 150/70

EXAMES: CR: 2.5 U: 126 CLEARENCE: 25 ML/MIN
HB: 13 GLI: 92 PARCIAL: LEU: 12

A: NEFROESCLEROSE HIPERTENSIVA BENIGNA

P: RETORNO COM EXAMES
```

No Text Selected

id	spans	text
M0	677~685	ATENSINA
M1	686~696	ANLODIPINA
M2	697~701	HCTZ
C0	652~655	HAS
C1	810~845	NEFROESCLEROSE HIPERTENSIVA BENIGNA
E0	737~745	CR: 2.5
E1	744~751	U: 126
E2	752~772	CLEARENCE: 25 ML/MIN
E3	773~779	HB: 13
E4	780~787	GLI: 92
A0	717~727	PA: 150/70

Source: the author.

Once the training phase was over, the annotators started to annotate the texts corresponding to the 41 patients of SummClinBr corpus. We intended to perform double-annotation of all the documents; however, one of the annotators left the project, and therefore we performed single-annotation. We assumed that the corpus could be reliable even with single-annotation since, in the training phase, the annotators achieved almost 0.90 of inter-annotator agreement in concept-level annotation. This corpus was used to train one of the supervised summarization methods and used to evaluate all the developed methods.

4.4 SEMANTIC SIMILARITY ESTIMATION

This section covers the development of a text similarity algorithm aiming to measure the semantic proximity of two short sentences. This calculation could impact a summarization system by avoiding the selection of redundant sentences, which is a very common issue in clinical texts (DALIANIS, 2018).

4.4.1 Data acquisition

To assess the developed algorithm, a gold standard containing a pair of texts and their respective semantic correlation value is needed. We obtained the datasets from the following semantic similarity shared-tasks: (i) ASSIN 2 (Second Semantic Similarity and Textual Inference Evaluation) and (ii) 2019 n2c2 Shared-Task - Track 1: n2c2/OHNLP Track on Clinical Semantic Textual Similarity. Both datasets are composed of pairs of sentences and their respective similarity value, defined by humans, where five stands for equivalent sentence and one a completely dissimilar sentence (in ASSIN the value is one, but on n2c2 the minimum value is zero).

The main differences between the databases rely on the language and domain of the texts. While the first use general texts in Portuguese, the second has data in English from the clinical area. The size of the dataset is presented in Table 20, and examples of texts in Table 21.

Table 20 – The size of both corpora (number of sentences).

	n2c2	**ASSIN 2**
Training	1,642	7,000
Test	412	2,448
Total	2,054	9,448

Source: the author.

4.4.2 Siamese Neural Network architecture

The Siamise Neural Network (SNN) is an architecture composed of two or more identical subnetworks, which parallelly process entities and share the parameters between the layers. Figure 22 outlines the SNN architecture.

The SNN achieved good results in image similarity studies and has not been properly tested in the context of textual similarity. Moreover, it has low susceptibility to overfitting and fewer data requirements, leading us to develop our text semantic similarity module based on this architecture.

We adapted the SNN proposed by Mueller and Thyagarajan (2016) to receive three new lexical features. Three dense layers with 100 units were added after the lexical similarity feature extractor (Dense 2, 3 and 4), and a final dense layer with 50 units to infer from the values of the Word Embeddings and the three lexical features layers. The features extracted were similarity metrics such as Cosine similarity, Dice coefficient, and Jaccard index to represent the number of common tokens of both sentences. Furthermore, the sigmoid activation function was changed to a ReLU function because instead of values between 0 and 1 produced by the sigmoid function, we needed values ranging between 0 or 1 to 5. The adapted model is shown in Figure 23, and the hyperparameters are listed below.

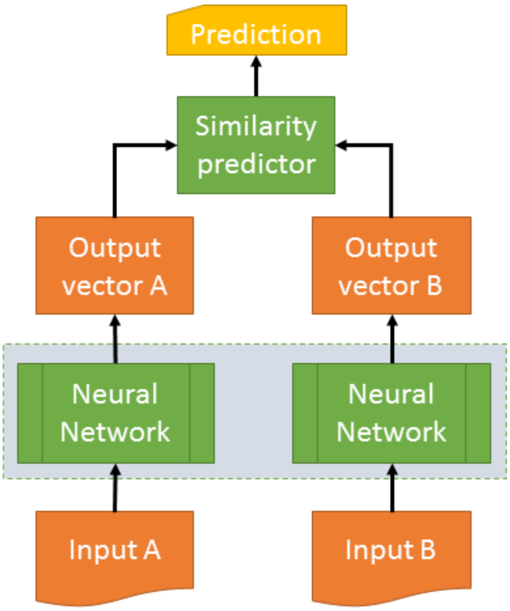
- Embedding Layer Dimension size: 100
- Number LSTM cells: 300
- Number Dense 1, 2, 3 & 4 units: 100
- Number Dense 5 units: 50
- Drop Rate LSTM: 0.17
- Drop Rate Denses: 0.25
- Activation Function: ReLU
- Epochs: 100

Table 21 – Text pairs from ASSIN 2 and n2c2 with their respective descriptions and similarity scores.

Dataset	Description, sentence pair and score
ASSIN	The two sentences are totally unrelated. • *Não tem água sendo bebida por um gato* • *Um caminhão está descendo rapidamente um morro* Score: **1.0**
ASSIN	The two sentences are not equivalent, but share some details. • *Um cachorro preto escuro e um cachorro castanho claro estão brincando no quintal* • *A luz no quintal está pregando truques no escuro no cachorro castanho e preto* Score: **3.1**
ASSIN	The two sentences are completely equivalent. • *Um homem de terno está de pé em frente a um microfone e cantando* • *O homem de terno está de pé de frente para um microfone e cantando* Score: **5.0**
n2c2	The two sentences are totally unrelated. • *Patient has not been the victim of other types of abuse* • *Information collected has not been verified by the patient* Score: **0.0**
n2c2	The two sentences are not equivalent, but share some details. • *Strattera 40 mg capsule 1 capsule by mouth one time daily* • *Arimidex 1 mg tablet 1 tablet by mouth one time daily* Score: **2.5**
n2c2	The two sentences are completely equivalent. • *I have reviewed her note in the electronic medical record* • *Collateral history was reviewed from the electronic medical record* Score: **5.0**

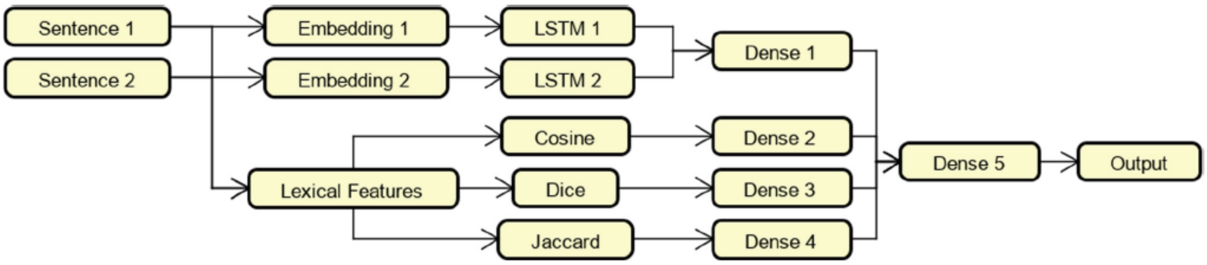
Source: adapted from de Souza et al. (2019)

Figure 22 – The Siamese Neural Network architecture. Shaded boxes represent shared parameters that are updated simultaneously.



Source: de Souza et al. (2020).

Figure 23 - Adapted Siamese Neural Network incorporated with lexical similarity features, embedding layers, and adjusted dense layers.



Source: de Souza et al. (2019).

The Embeddings layers count with a set of pre-trained word2vec models. For the ASSIN 2 dataset, we used the 300-sized CBOW model developed by Hartmann et al. (2017). For the English language, we used the 300-sized Google News vector.

4.4.3 Experimental setup

Additionally to the SNN model, we added to the experiments a Logistic Regression algorithm fed by the same lexical features. Both models (SNN and Logistic Regression), were experimented in a 10-fold cross-validation protocol and trained over both datasets (ASSIN 2 and n2c2). The SNN model was evaluated with and without the use of the lexical features to check if they impacted the results. We calculated the Pearson correlation (PC) and Mean Squared Error (MSE) metrics between the predicted values and the gold-standard values. The Pearson correlation measures how linearly the values are related, and the Mean Squared Error (MSE) estimates the average squared difference between the values, giving the error of the prediction.

4.5 SUPERVISED APPROACHES

In this section, we explore supervised summarization approaches, trying to leverage both sequence labeling and dictionary-based approaches supported by the SummClinBr corpus.

4.5.1 Sequence labeling models

This section covers a method that treats the summarization task as a sequence labeling problem, following the idea of Viani et al. (2017), which annotated clinical events to train an ML algorithm to extract concepts from the text.

The justification behind this choice is that since we are looking for a summary that reduces information overload and consequently the health practitioner’s time dealing with EHR, the use of an algorithm to extract clinical concepts in the middle of the text (i.e., NER), has the natural ability to know in which places there are clinical concepts, and therefore, being able to exclude any sentence without occurrences of these concepts. Also, the simple fact of moving the clinical concepts from the text and to another form of presentation can already bring significant results. For example, Viani et al.'s work, in which no event related to the patient is removed during the process, however, by showing them without the complementary part of the text, in a timeline, considerably reduces data overload.

The method executes three steps of AORTIS model: **aggregation**, **organization**, and **reduction**. The list of sub-steps are listed below:

1) Selects the patient’s longitudinal data
    a. Retrieves all the unstructured text from the clinical notes
2) Inputs data to NER algorithm (could be trained with SemClinBr or SummClinBr)
    a. Extracts multiple features from the text
    b. Classifies the concepts within the text
    c. Outputs the labeled texts

3) Orders the sentences from the most recent to the older ones
4) Generates summary composed by multiple reduced documents
    a. Removes sentences with no tags labeled (Reduction step 1)
5) Generates summary composed by a single document with *N* sentences
    a. Selects the first *N* sentences, and removes the others (Reduction step 2)
    b. Estimates the semantic similarity between the selected sentences
    c. Removes redundant sentences (if removed, returns to step 5a)
    d. Put each sentence in the respective section (i.e., SummClinBr categories)

Regarding the method, a few remarks need to be made. Since both SemClinBr and SummClinBr do not include temporal relations annotation, we simplified the date associated with each clinical concept found, then, to every concept, was assigned the document creation time. Another important aspect is that it is possible to generate two types of summary. The first is composed of all the original documents with less information each (because sentences with no labels were removed). It is possible to apply a tool such as *matplotlib* to generate a timeline, as in Viani et al. (2017). The other summary type uses a very straightforward logic, by selecting the *N* more recent sentences to generate the final summary. This was made by observing the SummClinBr sentence-level gold standard, which shows that the majority of relevant information is contained in the last note. The semantic similarity calculation used the algorithm presented in section 4.4.

The experimental setup to evaluate the method and its variations was threefold: (i) measures the ability of NER to extract clinical concepts from text (step 4 summary), (ii) estimate the *informativeness* of the single-document summaries (step 5 summary), and (iii) analyze the overall quality of the multi-document summaries (step 5 summary).

For the **first experiment**, we trained a CRF classifier in a 10-fold cross-validation protocol with the SummClinBr corpus. The Precision, Recall, F1-score metrics were calculated. Similar experiments were already made in section 3.2 to evaluate the NER with SemClinBr.

The **second experiment**, in addition to measuring the level of information present in the final summary, will also show us whether the use of SemClinBr, a multi-purpose information extraction corpus, can support the task of summarizing chronic diseases, as we can compare the results with SummClinBr, a corpus made specifically for this task. We calculated the ROUGE scores for ten different patients and used three distinct *N* values: 10, 15, and 20 sentences.

In the **third experiment**, we analyzed qualitatively (i.e., no metrics used, just observations) the outputs generated in step 5. We checked the redundancy of information, the structure and coherence of the text, and other quality aspects.

4.5.2 Dictionary-based model

Inspired by the idea of Sarkar et al. (2011), which defined a set of medical cue-phrases to improve the sentence scoring in their biomedical summarization system, we developed a summarization method that exploits a clinical dictionary to score and select the sentences for the summary. Machine learning-based extraction approaches often do not recognize all the entities from text, so we wanted to check if a more straight-forward and specialized concept-matching method could improve the results. The dictionary (i.e., lookup table) is composed of the terms annotated concepts in SummClinBr.

As the specialists stated that all the seven categories of SummClinBr' tagset are considered relevant, we defined an implicit output template for the summary, formed by Comorbidities, Exam results, Medications, Procedures, Clinical attributes, Risk factors, and Last visit conduct. Each category has its own slots of sentences, variable according to the final number of sentences (*N*). In another consultation with the nephrologists, they affirmed that some categories are more relevant than others, then, we considered Comorbidities, Exam results and Medications as **essential** information, and the others as **important**. Therefore, the essential tags have two times the number of slots than the important ones. For instance, if we want a summary with *N*=10 sentences, the Comorbidities, Exam results, Medications tags will have two sentence slots each, and the remaining categories one sentence each.

The method execute four steps of AORTIS model: **aggregation**, **organization**, **reduction** and **transformation**, and is composed by the following sub-steps

1) Calculate the template slots per tag based on *N* parameter
2) Selects the patient's longitudinal data
    a. Retrieves all the unstructured text from the clinical notes

3) Inputs to a fuzzy-match algorithm
    a. Loads the dictionary of terms (built from SummClinBr annotations)
    b. Normalizes the input text and dictionary terms
    c. Labels the concepts found within the text
    d. Outputs the tagged text
4) Orders the sentences from the most recent to the older ones
5) Generates summary composed by multiple reduced documents
    a. Removes sentences with no tags labeled (Reduction step 1)
6) Scores remaining sentences (each tag found +1 score)
7) Fills the template slots
    a. Inserts higher score sentences on each tag slot
    b. Removes other sentences (Reduction step 2)
    c. Estimates the semantic similarity between the selected sentences
    d. Removes redundant sentences (if removed, returns to step 7a)
    e. If some category slot remains incomplete, add extra slots to essential categories
8) Generates a template-based summary composed by *N* sentences distributed in the template slots

Regarding step 3, we applied both exact and fuzzy-match approaches to find the dictionary terms within the text, using a maximum edit distance of 1. It is important to mention that the dictionary is loaded only with terms prevenient from patients other than the one being tested at the time, to test the representativeness of the list.

In the normalization step (3b), all the text is lowercased, and the numbers, which are common in exam results and medications, are replaced with a proxy word "_NUMBER_". With that, it is possible to use regular expressions to find patterns that were labeled with different values (e.g., Cr: 2.9 and Cr: 2.4, both are transformed into Cr: _NUMBER_). The semantic similarity calculation used the algorithm presented in section 4.4.

After preliminary testing, some clear issues were found regarding the Exam results, Medications, and Last visit conduct sections. For the nephrologists, it is important to know about the last **medication** prescriptions only, and with the current method, sentences with several medication prescriptions reach a very high score, regardless of their temporal location, often causing the selection of sentences containing old prescriptions. To solve that, the process of assembling the medication section was updated, giving more priority to newer labeled sentences than higher scoring ones. Therefore, the medications contained in the patient's last note take precedence over any others found. The same modification was made to the Last visit conduct section, since only the last conduct matters.

The **exam results section** suffered from the same "*overscoring*" problem since it is common to find a sequence of exams results in a sentence. However, in addition to the possibility of using the same solution as the medication section, which prioritizes the most recent occurrences, we can also obtain all exam occurrences, group them, and plot these data over a timeline, providing the doctor with a longitudinal view of the evolution of these parameters, and consequently the patient's condition. To group the exam results written in different ways (e.g., CR: 2.4, creatinine: 2.4), we used a set of regular expressions developed by Silva (2018), which covers several exam names variations.

The experimental setup is composed of two experiments, aiming to (i) calculate the *informativeness* of the single-document summaries (step 5 summary), and (ii) analyze the overall quality of the multi-document summaries (step 8 summary).

The **first experiment** calculated the ROUGE scores for ten different patients and used three distinct *N* values: 10, 15, and 20 sentences.

In the **second experiment**, we analyzed the outputs generated in step 8. We checked the redundancy of information, the structure and coherence of the text, and other quality aspects.

5 RESULTS

This chapter shows the evaluation results of this thesis, including the reliability assessment of both built corpora (i.e., SemClinBr and SummClinBr), the performance measurement of the developed NLP tools (i.e., NER, POS-Tagger, and Text Semantic Similarity algorithms), and the EHR summarization outcomes. The last chapter describes the association between the results and research objectives.

5.1 SEMCLINBR

This section compiles quantitative and qualitative results regarding SemClinBr development. We detail the IAA information and analyze the errors found during the annotation.

5.1.1 Corpus statistics

The corpus development involved eight annotators, two adjudicators, and four Health Informatics researchers, totaling a team of 14 people. Our corpus comprehended 100 UMLS semantic types (STY) representing the entities, two extra semantic types typifying Abbreviations and Negations. The annotation process was 100% double-annotated and adjudicated, and lasted 14 months, resulting in a corpus composed of 1,000 documents (148,033 tokens), with 65,129 entities. In Table 22, we present the number of annotations of the most frequent semantic types.

5.1.2 Inter annotator agreement

The average agreement between all the 1,000 double-annotated documents in the corpus using four different IAA versions (i.e., strict, lenient, flexible, and relaxed) were calculated. We achieved an average **strict IAA of ~0.71** and **~0.92 for relaxed version**. In Table 23, we detail the average IAA values for the entire corpus, and Figure 24 presents, in the form of a heatmap, the average agreement considering the most frequent semantic types.

Table 22 - The number of annotated entities per semantic type and their corresponding semantic groups (first column).

SGR	STY	# Entities
Anatomy	Body Location or Region	1,452
Anatomy	Body Part, Organ, or Organ Component	1,373
Chemicals & Drugs	Organic Chemical	2,000
Chemicals & Drugs	Pharmacologic Substance	3,013
Concepts & Ideas	Quantitative Concept	3,953
Concepts & Ideas	Qualitative Concept	500
Concepts & Ideas	Temporal Concept	1,663
Devices	Medical Device	1,617
Disorders	Disease or Syndrome	2,650
Disorders	Finding	6,867
Disorders	Injury or Poisoning	521
Disorders	Sign or Symptom	4,707
Living Beings	Patient or Disabled Group	844
Living Beings	Professional or Occupational Group	720
Organizations	Health Care Related Organization	639
Phenomena	Laboratory or Test Result	3,079
Physiology	Clinical Attribute	1,128
Procedures	Diagnostic Procedure	2,012
Procedures	Health Care Activity	2,763
Procedures	Therapeutic or Preventive Procedure	4,791
N/A	Abbreviation	12,629
N/A	Negation	2,676

Source: Oliveira et al. (2020).

Table 23 – Average IAA values for the entire corpus.

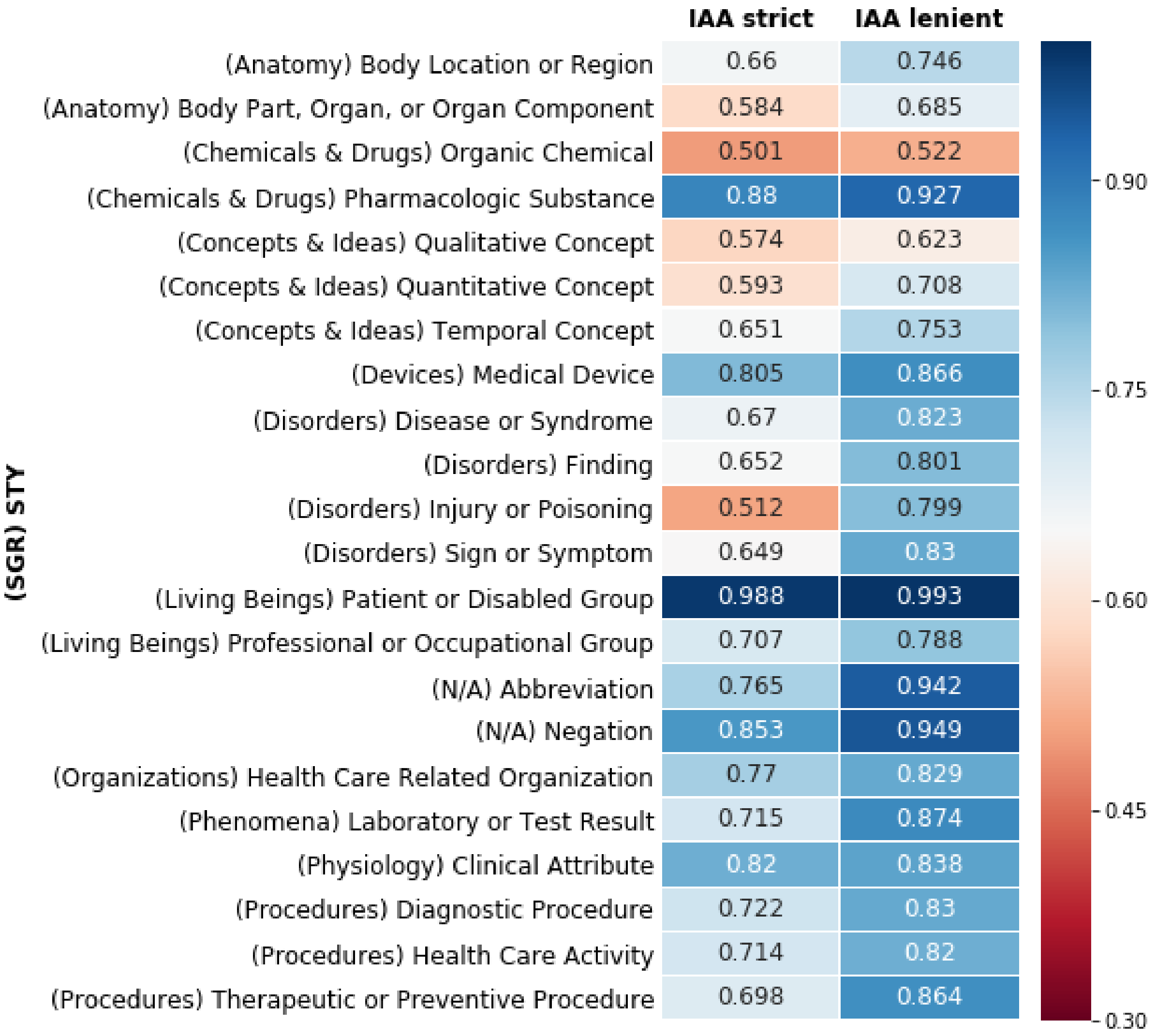
IAA type	IAA
Strict (full span + STY match)	0.708
Lenient (partial span + STY match)	0.834
Flexible (full span + SGR match)	0.774
Relaxed (partial span + SGR match)	0.921

Source: Oliveira et al. (2020).

Taking into account the challenges and particularities of our corpus (discussed in chapter 6), we compared our results with previous initiatives and compiled the IAA values of each corpus (see Table 24). The IAA percentage difference for entity annotation are ranging from 2.8% and 18.3% using the strict match, and 3.6% to 23.2% for the lenient match. Except for MiPACQ, all the other corpora had better IAA values in the strict match, probably due to the issues mentioned in the discussion section.

However, when we compare the lenient match scores, our performance is better than all other corpora except IxaMed-GS that had the best results of all of them. This led us to believe that our annotators had more trouble in defining the correct text spans than the ones in other projects because when we use a partial span match approach our results improved by 16.9% (from 0.71 to 0.83), and the other corpora improvement ranged from 3.8% and 8.6%. We are not sure if the cause of this is the lack of proper guideline definition, annotators experience, or even the document types we used (examples of annotation span issues are detailed in error analysis section).

Figure 24 - Average IAA scores for the most frequent semantic types and their corresponding semantic groups (in parenthesis). The heat map indicates in blue the highest values and in red the lower ones.



(SGR) STY	IAA strict	IAA lenient
(Anatomy) Body Location or Region	0.66	0.746
(Anatomy) Body Part, Organ, or Organ Component	0.584	0.685
(Chemicals & Drugs) Organic Chemical	0.501	0.522
(Chemicals & Drugs) Pharmacologic Substance	0.88	0.927
(Concepts & Ideas) Qualitative Concept	0.574	0.623
(Concepts & Ideas) Quantitative Concept	0.593	0.708
(Concepts & Ideas) Temporal Concept	0.651	0.753
(Devices) Medical Device	0.805	0.866
(Disorders) Disease or Syndrome	0.67	0.823
(Disorders) Finding	0.652	0.801
(Disorders) Injury or Poisoning	0.512	0.799
(Disorders) Sign or Symptom	0.649	0.83
(Living Beings) Patient or Disabled Group	0.988	0.993
(Living Beings) Professional or Occupational Group	0.707	0.788
(N/A) Abbreviation	0.765	0.942
(N/A) Negation	0.853	0.949
(Organizations) Health Care Related Organization	0.77	0.829
(Phenomena) Laboratory or Test Result	0.715	0.874
(Physiology) Clinical Attribute	0.82	0.838
(Procedures) Diagnostic Procedure	0.722	0.83
(Procedures) Health Care Activity	0.714	0.82
(Procedures) Therapeutic or Preventive Procedure	0.698	0.864

Source: Oliveira et al. (2020).

If we use the flexible and relaxed match instead of strict and lenient for entity annotation, our result goes from 0.71 to 0.77, and 0.83 to 0.92 respectively, and in that setting, we think we have a fairer evaluation, because the corpus granularity and complexity are more similar to the other works, and then, our results come closest to the others. Compared to the CLEF corpus, we achieved the same IAA for strict vs. flexible and increased by 13% the results for lenient vs. relaxed. MERLOT and MedAlert had slightly better results (2.6% and 3.9%, respectively). IxaMed-GS still have a 9.1% advantage for strict vs. flexible, maybe because it is the most specific and least in-depth corpus compared to others, but even so, it has a 2.2% disadvantage for lenient vs. relaxed. And finally, we exceeded MiPACQ's results by 10.4% and 18.5% using flexible and relaxed approaches.

Table 24 - Comparison between similar clinical annotation projects. In parentheses, the percentage difference in performance comparing our corpus to other clinical annotation projects. Note that the IAA values for Flexible and Relaxed match are a copy of Strict and Lenient scores because the other authors did not calculate these metrics, and we wanted to know the percentage difference between their values and ours.

Corpus	Strict	Lenient	Flexible	Relaxed
CLEF Roberts et al. (2009)	0.77 (8.5%)	0.80 (-3.6%)	0.77 (0%)	0.80 (-13.0%)
IxaMed-GS Oronoz et al. (2015)	0.84 (18.3%)	0.90 (8.4%)	0.84 (9.1%)	0.90 (-2.2%)
MERLOT Campillos et al. (2018)	0.79 (11.2%)	-	0.79 (2.6%)	-
MedAlert Ferreira et al. (2009-2010)	0.80 (12.6%)	-	0.80 (3.9%)	-
MiPACQ Albright et al. (2013)	0.69 (-2.8%)	0.75 (-9.6%)	0.69 (-10.4%)	0.75 (-18.5%)

Source: adapted from Oliveira et al. (2020).

5.1.3 Error analysis

The error (or disagreement) analysis showed the most common mistakes that have impacted the agreement results, and it was performed by the Health Informatics team continuously during the annotation process so that the annotators would be given feedback on their work. As performing a full error analysis for the entire corpus would be highly time-consuming, we only analyzed the part of the documents in which agreement had not reached the 0.67 IAA threshold. Moreover, the adjudicators were already aware of the persistent errors. As expected, a large number of errors occurred at the beginning of the annotation phases (i.e., ground-truth phases 1 and 2), because despite the training, the annotators were still getting used to the annotation process and using the guidelines document. Another common aspect of most of the disagreements is that they are not conceptual, that is, the disagreement does not originate from the semantic value given to the clinical entity, but rather from the different word span selection (term boundaries) generally associated with omission or inclusion of non-essential modifiers and verbs to a term (e.g., "*o tratamento*" vs "*tratamento*" labeled as "Therapeutic or Preventive Procedure" – "*the treatment*" vs "*treatment*").

The STYs high granularity caused two types of annotation divergences. The first one was concerning the annotation using different STYs with close semantic meaning because they are directly related to the UMLS hierarchy. One of the most occurring errors of this type is related to "Finding" and "Sign or Symptom", even with the simplification that we stated in our Guidelines that says: annotators should always give preference to disease/disorders and lab results STY's over the "Finding" STY. Only results of physical examination considered to be normal should be marked as "Finding". The abnormal ones should be labeled as "Sign or Symptom". Another example of this kind of error is when the annotators should decide between "Medical Device" and "Drug Delivery Device" like with the "infusion pump" device. The second type of error associated with the high granularity occurred because some uncommon concepts could be labeled with some infrequent STYs not remembered by the other annotator (e.g., "Element, Ion, or Isotope", "Age Group", "Machine Activity").

Erroneous decomposition of multiword expressions occurred even with many examples explicitly described in the guidelines. This error occurred when one annotator thought a compound term should be labeled as a single annotation and the other annotator as two or more different terms (annotations). There was no unique rule to follow in this case, as it depends on the context. Perhaps this is the reason for this type of error. For instance, the term "*Acesso venoso central direito*" ("right central venous access") needs to be decomposed as "*right*" (Spatial concept) and "*central venous access*" (Medical Device), but some annotators simply annotated all the term as "Medical Device". And other terms do not need to be decomposed as "*DRC estágio V*" ("Chronic kidney disease stage 5") that must be annotated as "Disease or Syndrome".

We found some errors caused by the ambiguity of certain words that could be misinterpreted in its sense, and this happened mainly in abbreviations. For instance, "*AC*" could be "*ausculta cardíaca*", "*anticorpo*" or "*ácido*" – "cardiac auscultation", "antibody" or "acid"). The term "*EM*" that could be "*Enfarte do miocárdio*", "*Esclerose múltipla*" or "*Estenose mitral*" – "Myocardial infarction", "Multiple sclerosis" or "Mitral stenosis”. We found simple omission errors of some concepts during the analysis as well.

In summary, the STYs performance (Figure 24) reflects the complexity of each STY, for example, the “Pharmacologic Substance” is composed mainly of single-word terms, and “Patient or Disable Group” has just a few terms encompassed by it, explaining their high IAA scores. Unlike “Finding” and “Sign or Symptom” for instance, that have a high frequency and very similar interpretations.

5.2 NAMED ENTITY RECOGNITION FOR CLINICAL TEXTS

The evaluation results for the three experimented NER methods, trained over the SemClinBr corpus, are presented in this section. They are all compiled in the same table to facilitate the comparison of values (see Table 25). As shown, for all the semantic groups, the BioBERTpt model achieved the best results. When we augment the annotation granularity (i.e., use the semantic types), most of the better result comes from the BiLSTM-CRF model. It is worth mentioning that we just used exact matches to calculate the metrics, even knowing that several times the algorithm classified the concept right, but selected a partial span of text.

Table 25 - NER F1-scores over **SemClinBr** corpus in a 10-fold cross-validation protocol of three different models.

Semantic type or group	CRF	BiLSTM-CRF	BioBERTpt
STY | Diagnostic Procedure	0.66	**0.74**	0.56
STY | Disease or Syndrome	0.55	0.56	**0.58**
STY | Finding	0.67	**0.72**	0.52
STY | Pharmacologic Substance	0.46	0.44	**0.78**
STY | Sign or Symptom	0.45	**0.62**	0.54
STY | Therapeutic or Preventive Procedure	**0.61**	0.42	0.46
SGR | Chemicals & Drugs	0.42	0.83	**0.90**
SGR | Disorders	0.70	0.69	**0.78**
SGR | Procedures	0.65	0.57	**0.69**

Source: the author.

As a part of the summarization sequence labeling models (section 4.5.1), a CRF NER model was trained using the SummClinBr corpus in order to extract relevant concepts from the text. The results are detailed in Table 26.

Table 26 – NER F1-scores over **SummClinBr** corpus in a 10-fold cross-validation protocol of a CRF model.

Category / tag	CRF
Comorbidity	0.70
Medication	0.77
Exam result	0.64
Procedure	0.28
Clinical attribute	0.58
Risk factor	0.38
Last visit conduct	0.43

Source: the author.

5.3 POS-TAGGING FOR CLINICAL TEXTS

The developed clinical POS-Tagger achieved 92.39% of accuracy, and 96.05% of F1-score overall, averaging all the POS-tags. Table 27 shows the accuracy scores for each of the tags separately. The three least frequent tags in the corpus were the ones with the lowest accuracy (PRP, NNP, and RB), probably due to little data for training.

Table 27 – Accuracy scores (%) by POS-tags.

POS-tag	Accuracy
CC – Coordinating conjunctions	99.84
CD – Cardinal numbers	95.21
DT – Determiners	91.51
IN – Prepositions	96.53
JJ – Adjectives	85.25
NN – Nouns	90.85
NNP – Proper nouns	58.05
PRP – Personal nouns	52.00
RB – Adverbs	76.41
VB - Verbs	92.47

Source: the author.

5.4 SEMANTIC SIMILARITY

The results of the textual semantic similarity algorithm are presented in Table 28. For both datasets, the best results (larger PC and minor MSE) were achieved with the SNN model using the lexical features. It is possible to check that the SNN model is better than the baseline algorithm even with no additional features to support the classification, and the results are even better by training with the features.

Table 28 - Pearson Correlation (PC) and Mean Squared Errors (MSE) of the SNN model (with and without the lexical features) and the Logistic Regression baseline. Best values per dataset in bold.

	SNN no feat.		Logistic with feat.		SNN with feat.	
Dataset	**PC**	**MSE**	**PC**	**MSE**	**PC**	**MSE**
n2c2	0.58	1.75	0.53	1.84	**0.64**	**1.58**
ASSIN 2	0.64	0.78	0.57	0.75	**0.69**	**0.74**

Source: the author.

We listed below a set of sentences that the semantic similarity algorithm considered to be redundant in the summarization task. Note that most of them have some lexical overlap. The algorithm did not find any sentence with the same meaning but expressed in different words.

- *HAS COMPENSADO → HAS COMPENSADA*
- *#MUC:EM USO DE: FUROSEMIDA 40MG/D → FUROSEMIDA 40MG/D*
- *HAS → TEM HAS*
- *INICIO SINVASTATINA 20MG/DIA → SINVASTATINA 20MG*
- *ECO COM CISTO RENAL SIMPLES A DIREITA → CISTO RENAL SIMPLES A DIR*

5.5 CHRONIC DISEASE SUMMARIZATION

The evaluation protocol of the proposed clinical summarizers covers the information overlap between the gold-standard (i.e., SummClinBr sentence-level annotation) and the generated summaries, by applying the ROUGE-N and ROUGE-SU4 metrics (the most used in summarization tasks). Additional observations regarding the readability, coherence, and redundancy were described as well. After showing the individual performance of each method, we present a section where we compare the average ROUGE results.

For a better understanding of the results, in Table 29 and 30, we show some statistics and additional information regarding the ten patients used for evaluation. As shown, the amount of notes is very unbalanced between the patients, ranging from 5 to 30. However, having fewer notes is not necessarily a signal of less information, the patient 4, for instance, has only five notes, yet, the tokens per note value for this patient is the higher between all patients.

Table 29 - Number of clinical notes, sentences, and tokens per patient.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
# notes	15	12	7	5	25	26	28	20	30	11
# sentences	367	216	247	251	443	473	698	668	381	310
# sentences per note	24,5	18,0	35,3	50,2	17,7	18,2	24,9	33,4	12,7	28,2
# tokens	2337	1210	1594	1670	3222	2720	5348	4998	2117	1853
# tokens per note	155,8	100,8	227,7	334,0	128,9	104,6	191,0	249,9	70,6	168,5

Source: the author.

Another interesting aspect of these patients is concerning medical specialty distribution. Each note has a specialty assigned, and most of the patients have multiple specialties associated. In addition, several documents were marked as "Not informed", making automatic recognition of specialties difficult.

Additionally, in the next section, we calculated the ROUGE metrics by comparing the gold-standard summaries with the last clinical note of the evaluated patients, because most of the sentences selected for the final summary in the gold standard are from the last note. With that, we have a baseline result to compare with the proposed methods, as there was no other study performing clinical summarization in pt-br. Moreover, the best-performing baseline algorithm (i.e., frequency-based), presented in section 4.1 is evaluated as well.

5.5.1 Last note and Frequency-based baselines

Table 31 presents the ROUGE-N and ROUGE-SU4 metrics calculated over the gold-standard and the last notes of each patient. It is important to note that, although most of the data originated from the patient's last note, the results for each patient varied considerably. For patients 1, 4, and 6, the Recall values were very high, indicating that most of the information needed in the summary was, in fact, present in the last note. For the rest of the patients, the information overlap was low, which reinforces the idea that the quality, format, cohesion, and especially, the completeness of the data contained in the clinical note vary greatly according to several variables, such as the clinical case, responsible physician, type of encounter, etc. As expected, the Precision values were low in all cases, as the last note will always have more data than the final summary.

Table 30 - Number of clinical notes of each medical specialty and the main diagnoses per patient.

	Specialties	Main Diagnoses
P1	Not informed: 10 **Nephrology: 4** Urology: 1	**CKD 4** Hypertension
P2	Not informed: 8 **Nephrology: 3** General surgery: 1	**CKD 4**
P3	Not informed: 4 **Nephrology: 3**	**CKD 3** DM Peripheral neuropathy Diabetic retinopathy
P4	Not informed: 4 **Nephrology: 1**	**CKD 4** Hypertension
P5	Vascular surgery: 12 Cardiology: 5 **Nephrology: 4** Not informed: 4	**CKD 4** Coronary heart disease Hypertension Hypothyroidism
P6	Not informed: 9 Ophthalmology: 9 **Nephrology: 4** General surgery: 2 Anesthesiology: 1 Cardiology: 1	**Diabetic Nephropathy** DM Hypertension Coronary heart disease Hypothyroidism Dyslipidemia Glaucoma Gallstones
P7	Cardiology: 9 General surgery: 7 Not informed: 7 Ophthalmology: 2 **Nephrology: 1** Vascular surgery: 1 Cardiac surgery: 1	**CKD 4** Hypertension Atherosclerosis Coronary heart disease
P8	Cardiology: 10 Gastroenterology: 4 Not informed: 4 **Nephrology: 2**	**CKD** Congestive heart failure Chagas disease Hypothyroidism Asthma
P9	**Nephrology: 24** General surgery: 5 Anesthesiology: 1	**CKD 5 - transplanted** Hypertension Hypothyroidism Dyslipidemia
P10	Not informed: 5 General surgery: 2 **Nephrology: 2** Vascular surgery: 1 Otorhinolaryngology: 1	**CKD 4** Hypertension DM Glaucoma

Source: the author.

Table 31 – ROUGE scores for each patient's last note. The Precision (P), Recall (R), and F1 (F) scores were calculated for each metric (ROUGE-SU4, ROUGE-1, and ROUGE-2).

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
ROUGE-SU4	P	0,435	0,194	0,245	0,264	0,081	0,276	0,052	0,108	0,277	0,190
	R	0,832	0,343	0,439	0,843	0,005	0,713	0,053	0,084	0,199	0,190
	F	0,572	0,247	0,315	0,402	0,010	0,398	0,052	0,094	0,232	0,190
ROUGE-1	P	0,500	0,274	0,345	0,318	0,308	0,332	0,221	0,361	0,473	0,397
	R	0,945	0,475	0,607	1,000	0,025	0,842	0,223	0,282	0,344	0,397
	F	0,654	0,347	0,440	0,483	0,047	0,476	0,222	0,317	0,398	0,397
ROUGE-2	P	0,463	0,210	0,286	0,294	0,000	0,300	0,029	0,099	0,283	0,186
	R	0,880	0,367	0,506	0,929	0,000	0,766	0,029	0,077	0,205	0,186
	F	0,607	0,267	0,365	0,446	0,000	0,431	0,029	0,087	0,237	0,186

Source: the author.

In Table 32, the patients 5, 7, and 10, which had low ROUGE results when compared to the last note, have their gold standard extractive summaries, and last notes are shown. For patient 5, the last note is very succinct and superficial, indicating a visit just to renew the medication prescriptions. The note says "*patient asymptomatic, made routine exams, renewed the prescriptions and scheduled the return to April*", and this explains the low ROUGE score. Patient 7's last note is regarding consultation with the ophthalmological team, and for that reason, the subject of this text is out of the scope of the gold standard. And for patient 10, the last note is quite complete concerning the patient's main condition (chronic kidney disease). But why the low ROUGE score, even for recall? This case highlights one of the limitations of the ROUGE metric, which uses only a lexical match instead of a semantic match, causing texts written with different words, but with the same semantic meaning, to be penalized. One example is when we check the following sentences, the first from the gold standard and the second from the last note.

- *HMA: DM há 30 anos e faz uso de insulina NPH (de manhã 26U, antes do almoço 26U, antes de deitar 12U) e regular (de manhã e antes do jantar 4U ou 6U), com controle de contagem de glicemia diariamente*
- *DM HA 30 ANOS, INSULINODEPENDENTE*

Both sentences refer to Diabetes diagnostic and insulin treatment but written in different ways. The first one is more detailed than the second. Another example from the same patient is shown below. The text presents the Chronic Kidney Disease diagnostic in distinct ways.

- *#HMA: PACIENTE RENAL CRONICA HÁ APROXIMADAMENTE 4 ANOS*

- *# DRC ESTADIO 4*

Table 32 - Gold-standard vs. Last note of the patient with some of the worst overlap (i.e., lower ROUGE scores)

Patient	Extractive gold standard sentences	Last note
P5	Acompanhamento desde 11/2010 devido a nefropatia crônica (sem etiologia definida) HAS há 8 anos (controlada) DAC: cateterismo em 2002 Em uso de: AAS 100mg, Losartana 50mg/dia, Sinvastatina 1cp 1x/d, Cilostozol 100mg 12/12h, Carvedilol 6,25mg 12/12h, Puran 50mCg 1x/d, ANLODIPINO 5Mg 12/12h, HCTZ 25mg Ecografia (06/17): Cisto renal simples a direita. Artéria renal direita: pérvia, sem sinais de estenoses hemodinamicamente signficativas, não há sinais de aumeno significativo da resistencia vascular renal Artéria renal esquerda: pérvia, sem sinais de estenoses hemodinamicamente significativas. não há sinais de aumeno significativo da resistencia vascular renal, porém rim apresentando diminuição da vascularização e perda da diferenciação cortico medular. Sinais de nefropatia crônica Mantendo função renal (CKD 25) microalb: 18,7mg DOENCA RENAL CRONICA (IV) DE ETIOLOGIA INDETERMINADA, POSSIVELMENTE ASSOCIADA A DOENÇA ISQUEMICA RENAL Dça renovascular? Evidência de aterosclerose	PACIETNE ASSINTOMATICA EXAMES DE ROTINA E REFAZER A RECEITA MEDICA RETORNO EM ABRIL
P7	HAS EM TTO. DRC EC V EM TRATAMENTO CLÍNICO DAC conhecida / IAM em 2004-RVM e plastia de valva mitral. 03 pontes: SAF_CD / SAF-CX / MIE-DA	R1 NOME_MEDICO PARA R2 NOME_MEDICO EQUIPE DA RETINA # 02/2017 >> OACR EM OD # NEGA CX OU TRAUMA OCULAR PREVIOS # HMF NEGATIVA PARA DÇAS OCULARES # COMORBIDADES: HAS EM TTO.

	MUC:EM USO DE: FUROSEMIDA 40MG/D, SINVASTATINA 20MG 12/12H, CARVEDILOL 25MG 12/12H, AAS 100MG/D, HIDRALAZINA 50mg 12/12h, MONORDIL 20mg 12/12h, VASTAREL 35MG 12/12H REALIZADO FISTULA NO INÍCIO DE SETEMBRO (2015) >> FECHADA TFG CKD-EPI 19,0 Mantido sem espirono devido ao risco de hipercalemia devido a IRC DRC - ClCR 12,2 - DOENÇA RENAL POLICISTICA EM ACOMPANHAMENTO COM A NEFRO, ESTA AGUARDANDO DIALISE PERITONEAL, SEM SINTOMAS DE EMERGENCIA DIALÍTICA, REFERE DISPNEIA E DOR TORACICA AOS MEDIOS ESFORÇOS QUE MELHORAM COM REPOUSO E COM USO DE ISORDIL, DURAÇÃO MENOR QUE 20 MIN HIPERPARA SECUNDÁRIO AVC ISQ AOS 40 ANOS	IAM + REVASCULARIZAÇÃO MIOCARDIO + ICC ISQUEMICA DOENÇA RENAL CRÔNICA DOENÇA ATEROSCLERÓTICA #OCT (25/05/17) ATROFIA DE RETINA NEUROSSENSORIAL OD PACIENTE RETORNA COM AF SOLICITADA EM ULTIMA CONSULTA, sem novas queixas. AF(16/6/17): OD: D.O. COM ATROFIA EPR PERIPAPILAR, TORTUOSIDADE VASCULAR COM AFINAMENTO VASCULAR E HIPOPERFUSÃO EM PP ACOMETENDO MACULA. OE: D.O. SEM LATERAÇÕES TORTUOSIDADE VASCULAR. #AV: CD3M 20/20 #BIO: CAF, CALMA, AMPLA, CORNEA CLARA, CATARATA AO #BIOF OD: TORTUOSIDADE VASCULAR, ATROFIA DE EPR EM PP. #PIO: 12/11 #CD: RETORNO EM 3 MESES, ORIENTO PROGNOSTICO VISUAL DECLARAÇÃO DA ACUIDADE VISUAL DO DIA DE HOJE COM CID H34.1
P10	HMA: DM há 30 anos e faz uso de insulina NPH (de manhã 26U, antes do almoço 26U, antes de deitar 12U) e regular (de manhã e antes do jantar 4U ou 6U), com controle de contagem de glicemia diariamente Relata hipertensão e faz uso de inalapril 10mg, atenolol 50mg, hidroclorotiazida 25mg. Doença renal crônica, secundária a nefropatia diabética. #DRC classe IV - CKD EPI 20 ATUALMENTE REFERE DIMINUIÇÃO NA FREQUÊNCIA URINÁRIA, NEGA DISÚRIA. ECODOPPLER DE ARTERIAS RENAIS 27/08/2015 ARTERIA RENAL DIREITA: ALTERAÇÃO HEMODINAMICAMENTE SIGNIFICATIVA (PLACA DE ATEROMA CALCIFICADA), NÃO HÁ	NOME_PACIENTE, 75a # HAS HA 30 ANOS # DM HA 30 ANOS, INSULINODEPENDENTE # REFERE POLIPOS INTESTINAIS SOB INVESTIGAÇÃO # GLAUCOMA # DRC ESTADIO 4 #NEGA TABAGISMO E ETILISMO MUC: ATENOLOL 50MG 12/12 , ANLODIPINO 5MG 1x/dia AAS 100MG/DIA, APRESOLINA 25MG 1-0-1 SINVASTATINA 40MG/DIA, FUROSEMIDA 40MG /DIA , INSULINA NPH 18-25-0 LINAGLIPTINA 5MG/DIA #HMA: PACIENTE RENAL CRONICA HÁ APROXIMADAMENTE 4 ANOS , EM EXAME DE TRIAGEM DOS VASOS RENAIS POR DOPPLER FOI COSTATADA OBSTRUÇÃO DE ARTERIA RENAL ESQUERDA (11/2015). FOI ENCAMINHADA

	SINAIS DE AUMENT SIGNIFICATIVO DA RESISTÊNCIA VASCULAR RENAL (PARENQUIMA) ARTERIA RENAL ESQUERDA: ALTERAÇÃO HEMODINAMICAMENTE SIGNIFICATIVA NA ARTERIA RENAL (PLACA DE ATEROMA CALCIFICADA), SINAIS DE AUMENTO DA RESISTENCIA VASCULAR RENAL (PARENQUIMA) FOI REALIZADO CATETERISMO E COLOCAÇÃO DE STENT EM A RENAL ESQUEDA trás doppler com artérias renais com indices ao/ renais < que 0,35 . FUROSEMIDA 40MG /DIA #HMA: PACIENTE RENAL CRONICA HÁ APROXIMADAMENTE 4 ANOS , EM EXAME DE TRIAGEM DOS VASOS RENAIS POR DOPPLER FOI COSTATADA]	PARA AGIOPLASTIA COM COLOCAÇÃO DE STENT DE ARTERIA RENAL #S: QUEIXA-SE DE EDEMA EM MEMBROS INFERIORES. SEM OUTRAS QUEIXAS #O: BEG, LOTE, CORADA, HIDRATADA, EUPNEICA pa=130x70 HEMODINAMICAMENTE ESTÁVEL MMII: EDEMA 1+/4+ MIE EXAMES LAB 08/07 CR 2,0 HBA1C 8,8 VITAMINA D 24,7 HIPERGLICEMIA PELA MANHÃ #A: HAS DM NÃO CONTROLADA - ACOMPANHA COM PESQUISA CLÍNICA DOENÇA RENAL CRONICA EST IV #P: RETORNO EM 2 MESES COM EXAMES. ORIENTAÇÕES. TRAZER CONTROLE DE DEXTRO NA PRÓXIMA CONSULTA REPOSIÇÃO VITAMINA D 50000 4 SEMANAS NEFRO _NOMEMEDICO_

Source: the author.

Table 33 - ROUGE scores for each patient's Frequency-based summary with 15 sentences. The Precision (P), Recall (R), and F1 (F) scores were calculated for each metric (ROUGE-SU4, ROUGE-1, and ROUGE-2).

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
ROUGE-SU4	P	0,168	0,092	0,116	0,065	0,060	0,034	0,101	0,085	0,076	0,104
	R	0,105	0,089	0,057	0,023	0,031	0,013	0,051	0,024	0,055	0,031
	F	0,129	0,090	0,077	0,034	0,041	0,018	0,068	0,037	0,064	0,048
ROUGE-1	P	0,362	0,254	0,256	0,268	0,241	0,189	0,319	0,348	0,272	0,316
	R	0,229	0,246	0,131	0,097	0,127	0,074	0,165	0,103	0,195	0,098
	F	0,281	0,250	0,173	0,143	0,167	0,106	0,218	0,158	0,227	0,149
ROUGE-2	P	0,176	0,103	0,095	0,050	0,024	0,000	0,085	0,067	0,044	0,089
	R	0,111	0,100	0,048	0,018	0,013	0,000	0,043	0,019	0,032	0,027
	F	0,136	0,102	0,064	0,026	0,017	0,000	0,057	0,030	0,037	0,042

Source: the author.

The Frequency-based baseline results are shown in Table 33, presenting a low-performance in most of the patients. The first aspect of analyzing is redundancy detection, which worked only at a lexical level, permitting various similar sentences in the final summary. As known, the frequency should not be the only aspect to rely on, as some repetitive sentences could not have real importance on the patient's context (e.g., "*PACIENTE SEM QUEIXAS*" → "*Patient with no complaints*"). A patient with comorbidity as Glaucoma that had regular visits to the ophthalmologist, for instance, could impact the final summary, as redundant data from this medical specialty could achieve high scores in the Frequency-based algorithm, a consequently been selected.

5.5.2 Unsupervised BERT-based summarization

The same ROUGE metrics and patients used in the last note baseline were used to evaluate the BERT-based summarizer. Additionally, we did tests with different sizes of summaries, including 10, 15, and 20 sentences. Table 34 shows the scores. For all the metrics, in almost all the patients, the Recall increased as we added sentences to the final summary, which is expected, but at the same time shows that the summarizer is selecting sentences with relevant information. Regarding the Precision metrics, the behavior was not as predictable as the Recall, as it would be expected that the Precision would decrease whenever we add more sentences. However, there was a large variation of the summary that obtained better precision, that of 10, 15, or 20 sentences. Patients 1, 5, and 8 obtained the best average results in the BERT method.

In Table 35, the summaries with the highest and lowest average F1-scores are shown with their respective gold standards. Regarding **redundancy**, both summaries have similar sentences that could be removed from the summaries as they carry similar information. The exam results, in addition to being duplicated in some cases, are not time-identified, making the information unreliable for the medical team, as it may be an exam of a long time ago. See the following examples with their respective translations to English.

- Example 1
  - TFG ESTIMADA EM 24 MMII → *GRF ESTIMATED IN 24 MMII*
  - TFG (CKD-EPI) 26,1 → *GFR (CKD-EPI) 26,1*
- Example 2
  - Cr 2,5 (previa 2,4) = CKD EPI 38ml/min → *Cr 2,5 (previous 2,4) = CKD EPI 38ml/min*
  - EPI 25,1(DRC ESTÁGIO IV); → *EPI 25,1(CKD STAGE IV);*

- #DRC classe IV - etiologia desconhecida → *CKD class IV - unknown etiology*

Table 34 - ROUGE scores for each patient's BERT summary. The Precision (P), Recall (R), and F1 (F) scores were calculated for each metric (ROUGE-SU4, ROUGE-1, and ROUGE-2). Three different output sizes (number of sentences) were assessed (N10, N15, N20).

			P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
N10	**ROUGE-SU4**	**P**	0,278	0,046	0,101	0,240	0,269	0,084	0,104	0,259	0,246	0,263
		R	0,171	0,040	0,115	0,122	0,180	0,083	0,078	0,191	0,152	0,130
		F	0,212	0,043	0,107	0,162	0,216	0,083	0,089	0,220	0,188	0,174
	ROUGE-1	**P**	0,471	0,245	0,305	0,441	0,434	0,298	0,305	0,491	0,488	0,457
		R	0,294	0,213	0,345	0,230	0,293	0,295	0,230	0,365	0,305	0,228
		F	0,362	0,228	0,324	0,302	0,350	0,296	0,262	0,419	0,375	0,304
	ROUGE-2	**P**	0,328	0,000	0,106	0,259	0,276	0,043	0,096	0,217	0,215	0,275
		R	0,204	0,000	0,120	0,134	0,186	0,043	0,072	0,161	0,134	0,137
		F	0,251	0,000	0,113	0,176	0,222	0,043	0,083	0,185	0,165	0,182
N15	**ROUGE-SU4**	**P**	0,286	0,061	0,122	0,148	0,230	0,074	0,055	0,230	0,186	0,204
		R	0,281	0,071	0,223	0,115	0,187	0,101	0,062	0,250	0,146	0,145
		F	0,283	0,066	0,158	0,129	0,206	0,085	0,059	0,240	0,164	0,170
	ROUGE-1	**P**	0,495	0,268	0,258	0,341	0,398	0,233	0,224	0,379	0,386	0,402
		R	0,486	0,311	0,464	0,265	0,325	0,316	0,252	0,410	0,305	0,288
		F	0,491	0,288	0,332	0,299	0,358	0,268	0,237	0,394	0,341	0,335
	ROUGE-2	**P**	0,302	0,014	0,120	0,161	0,244	0,039	0,013	0,214	0,200	0,214
		R	0,296	0,017	0,217	0,125	0,199	0,053	0,014	0,232	0,157	0,153
		F	0,299	0,015	0,155	0,141	0,219	0,045	0,014	0,223	0,176	0,178
N20	**ROUGE-SU4**	**P**	0,313	0,035	0,078	0,198	0,235	0,082	0,082	0,216	0,101	0,181
		R	0,375	0,083	0,168	0,248	0,235	0,112	0,121	0,295	0,081	0,149
		F	0,341	0,050	0,107	0,220	0,235	0,095	0,098	0,249	0,090	0,163
	ROUGE-1	**P**	0,469	0,158	0,225	0,355	0,414	0,295	0,235	0,382	0,320	0,362
		R	0,560	0,361	0,476	0,442	0,414	0,400	0,345	0,519	0,258	0,299
		F	0,510	0,220	0,305	0,394	0,414	0,339	0,280	0,440	0,286	0,327
	ROUGE-2	**P**	0,326	0,007	0,073	0,214	0,237	0,055	0,064	0,227	0,098	0,179
		R	0,389	0,017	0,157	0,268	0,237	0,074	0,094	0,310	0,079	0,148
		F	0,354	0,010	0,100	0,238	0,237	0,063	0,076	0,262	0,087	0,162

Source: the author.

In general terms, the **readability** of the BERT-based summaries has some issues, as it does not follow a specific format or order, impacting the comprehension of the patient's history and current condition.

Another aspect that could be seen is that although the two patients (with lowest and highest scores) had similar clinical cases (CKD IV), equivalent amount of data (see

Table 29), and homogeneous medical specialties distribution (see Table 30), both obtained very different quantitative results.

Table 35 – Examples of BERT-based summaries with high and low F1 score, compared with their respective gold standards. In bold semantically similar information.

Patient	Extractive gold standard summary	BERT N20 summary
P1 highest F1	**HAS HÁ MAIS DE 10 ANOS** **NEGA DM** **DRC ESTAGIO IV - POSSIVEL NEFROPATIA HIPERTENSIVA** **CR: 2.1 TFG 17 ML/MIN** COM ALBUMINA DE 429MG E CALCIO IONICO DE 1.44 US DE AP. **URINARIO: (23/05/14)** RIM DIREITO COM COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE REDUZIDOS NAO HA EVIDENCIA DE DILATACAO DO SISTEMA COLETOR, AUSENCIA DE IMAGEM DE LITIASE **RIM MEDINDO 8,6X3,8X4,2** **RIM ESQUERDO COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE REDUZIDOS.** **NAO HA EVIDENCIA DE DILACATAO.** **Em uso: ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA**, LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA **DRC POR NEFROESCLEROSE HIPERTENSIVA BENIGNA COM TFG ESTIMADA EM 24**	**Hipertensa** com lesão renal retorna com exames Conduta:mantida medicação/Urolo0gia/retorno com exames) HMP: **NEGA DM** PARCIAL DE URINA: SEM ALTERACOES (DENSIIDADE, COR, ASPECTO, PH = OK **URINARIO: (23/05/14) RIM MEDINDO 8,6X3,8X4,2 RIM ESQUERDO COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE REDUZIDOS NAO HA EVIDENCIA DE DILACATAO** BEXIGA COM FORMATO HABITUAL, PAREDES LISAS E SEM CONTEUDO ANOMALO Nega queixas PTH 167,1 **#HAS HÁ MAIS DE 10 ANOS** #G6 P5 A1 **#MUC: ANLODIPINO 5 MG/DIA, FUROSEMIDA 40 MG/DIA, SINVASTATINA 20 MG/DIA #A: DRC POR NEFROESCLEROSE HIPERTENSIVA BENIGNA COM TFG ESTIMADA EM 24 MMII** NOME_PACIENTE, 69 ANOS **CR 1,9 TFG (CKD-EPI) 26,1** (exame previo Cr 1,8 com CKD-EPI 28) O# FC 62 PA 120/70
P2 lowest F1	**DRC IV**, em acompanhamento ambulatorial desde 2013 **DRC de etiologia desconhecida** **Nega DM e HAS** **MUC: nao faz uso de medicações contínuas** **MUC: furosemida 40 mg/dia** **Cr 3,0 CKD 19,7** US 04/06/13 - cisto renal simples a direita anemia normo normo com vitamina b12 baixa hipercalemia - último exame: 5,8 acidose metabólica - bic: 24	dspl emxaes: **Cr 2,5 (previa 2,4) = CKD EPI 38ml/min CREATININA 2,5**; UREIA 53; **EPI 25,1(DRC ESTÁGIO IV)**;hb 10,5 hcm 32%, leucocitos 7280; HDL 38; LDL 125; TRIGLICERIDEOS 130; CALCIO 8,8; FOSFORO 3,6; PTH 86,1 PA: 110/70 ATR **sem outras comorbidades - nao utilixa medicação nenhum**a P: 3,8 #P: PRESCREVO CITONEURIN 5000 1CP VO 12/12 POR 3 MESES, SOLICITO AVALIAÇÃO COM A NUTRIÇÃO PARA AVALIAR INGESTA DE POTÁSSIO, SOLICITO EXAMES LABORATORIAIS, RETORNO EM 6 MESES #S: Retorno em 4 meses FC: 64 #**DRC classe IV - etiologia desconhecida** Sem queixas NOME_PACIENTE , 73 anos, aposentado (auxiliar de trasnporte) Ao exame : BEG, corado e hidratado, eupneico BCRNFSS 2T Abdome sp Nefro NOME_MEDICO/ #Na ultima consulta foi prescrito **furosemida 40mg/dia**+ albendazol(eosinofilia) - Nega edema

Source: the author.

5.5.3 Supervised Sequence labeling summarization

The Sequence labeling method results followed the expected behavior in both Recall and Precision. Overall, the Recall levels increased as we increase the number of sentences (see Table 36), and the Precision decreased as we increased the number of sentences. The best average results were achieved when summarizing patients 1, 6, 8, and 10.

Table 36 - ROUGE scores for each patient's Sequence labeling summary. The Precision (P), Recall (R), and F1 (F) scores were calculated for each metric (ROUGE-SU4, ROUGE-1, and ROUGE-2). Three different output sizes (number of sentences) were assessed (N10, N15, N20).

			P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
N10	ROUGE-SU4	P	0,288	0,052	0,124	0,168	0,249	0,184	0,181	0,316	0,123	0,434
		R	0,342	0,117	0,170	0,234	0,188	0,280	0,273	0,384	0,136	0,453
		F	0,312	0,072	0,144	0,196	0,214	0,222	0,218	0,346	0,129	0,443
	ROUGE-1	P	0,426	0,179	0,263	0,282	0,361	0,343	0,279	0,476	0,270	0,547
		R	0,505	0,393	0,357	0,389	0,274	0,516	0,417	0,577	0,297	0,571
		F	0,462	0,246	0,303	0,327	0,312	0,412	0,334	0,522	0,283	0,559
	ROUGE-2	P	0,289	0,045	0,115	0,181	0,254	0,190	0,179	0,303	0,100	0,440
		R	0,343	0,100	0,157	0,250	0,192	0,287	0,268	0,368	0,110	0,459
		F	0,314	0,062	0,133	0,210	0,219	0,229	0,214	0,332	0,105	0,449
N15	ROUGE-SU4	P	0,179	0,056	0,075	0,107	0,161	0,190	0,132	0,204	0,095	0,303
		R	0,350	0,171	0,195	0,251	0,210	0,395	0,278	0,401	0,156	0,472
		F	0,237	0,085	0,109	0,150	0,182	0,256	0,179	0,271	0,118	0,369
	ROUGE-1	P	0,276	0,182	0,174	0,199	0,260	0,313	0,208	0,332	0,208	0,411
		R	0,532	0,541	0,440	0,460	0,338	0,642	0,432	0,647	0,336	0,636
		F	0,364	0,273	0,249	0,278	0,294	0,421	0,280	0,439	0,257	0,499
	ROUGE-2	P	0,182	0,056	0,075	0,112	0,158	0,206	0,128	0,195	0,083	0,313
		R	0,352	0,167	0,193	0,259	0,205	0,426	0,268	0,381	0,134	0,486
		F	0,240	0,083	0,108	0,156	0,178	0,278	0,174	0,258	0,102	0,381
N20	ROUGE-SU4	P	0,179	0,056	0,075	0,107	0,161	0,190	0,132	0,184	0,101	0,303
		R	0,350	0,171	0,195	0,251	0,210	0,395	0,278	0,407	0,177	0,472
		F	0,237	0,085	0,109	0,150	0,182	0,256	0,179	0,254	0,128	0,369
	ROUGE-1	P	0,276	0,182	0,174	0,199	0,260	0,313	0,208	0,299	0,220	0,411
		R	0,532	0,541	0,440	0,460	0,338	0,642	0,432	0,654	0,383	0,636
		F	0,364	0,273	0,249	0,278	0,294	0,421	0,280	0,410	0,279	0,499
	ROUGE-2	P	0,182	0,056	0,075	0,112	0,158	0,206	0,128	0,174	0,081	0,313
		R	0,352	0,167	0,193	0,259	0,205	0,426	0,268	0,381	0,142	0,486
		F	0,240	0,083	0,108	0,156	0,178	0,278	0,174	0,238	0,103	0,381

Source: the author.

It is worth mentioning that we are presenting the results when training the NER algorithm with the SummClinBr corpus only, as the results with the SemClinBr were lowest, even when correlating the semantic types and groups to the SummClinBr categories (e.g., SemClinBr “Chemicals & Drugs” semantic group → SummClinBr “Medication” category). In terms of Recall, the results were similar; however, the Precision went down with SemClinBr, as it identifies all clinical concepts, not only the relevant to chronic kidney disease patients.

Table 37 presents examples of summaries with high and low F1-scores. Concerning the **redundancy**, multiple sentences are identifying the main diagnosis (i.e., CKD), written in distinct lexical formats, which were not distinguished by the Semantic Similarity algorithm. As the following examples that have the exact same meaning.

- DRC estadio 4 - etiologia indeterminada → *CKD stage 4 - undetermined etiology*
- #DRC ESTAGIO 4 DE CAUSA INDETERMINADA → *CKD STAGE 4 OF UNDETERMINED CAUSE*

Table 37 - Examples of Sequence labeling summaries with high and low F1 score, compared with their respective gold standards. In bold semantically similar information.

Patient	Extractive gold standard summary	Sequence labeling N15 summary
P10 highest F1	**HMA: DM há 30 anos e faz uso de insulina NPH (de manhã 26U, antes do almoço 26U, antes de deitar 12U) e regular (de manhã e antes do jantar 4U ou 6U), com controle de contagem de glicemia diariamente** **Relata hipertensão e faz uso de inalapril 10mg, atenolol 50mg, hidroclorotiazida 25mg.** **Doença renal crônica, secundária a nefropatia diabética.** **#DRC classe IV - CKD EPI 20** ATUALMENTE REFERE DIMINUIÇÃO NA FREQUÊNCIA URINÁRIA, NEGA DISÚRIA. ECODOPPLER DE ARTERIAS RENAIS 27/08/2015 ARTERIA RENAL DIREITA: ALTERAÇÃO HEMODINAMICAMENTE SIGNIFICATIVA (PLACA DE ATEROMA CALCIFICADA), NÃO HÁ SINAIS DE AUMENT SIGNIFICATIVO DA RESISTÊNCIA VASCULAR RENAL (PARENQUIMA) **ARTERIA RENAL ESQUERDA: ALTERAÇÃO HEMODINAMICAMENTE SIGNIFICATIVA NA ARTERIA RENAL** (PLACA DE ATEROMA CALCIFICADA), SINAIS DE AUMENTO DA	[COMORBIDADES] **HMA: DM há 30 anos e faz uso de insulina NPH (de manhã 26U, antes do almoço 26U, antes de deitar 12U) e regular (de manhã e antes do jantar 4U ou 6U), com controle de contagem de glicemia diariamente** **#HMA: PACIENTE RENAL CRONICA HÁ APROXIMADAMENTE 4 ANOS , EM EXAME DE TRIAGEM DOS VASOS RENAIS POR DOPPLER FOI COSTATADA OBSTRUÇÃO DE ARTERIA RENAL ESQUERDA (11/2015)** HMF: Mãe diabética, faleceu com 80 anos de cardiopatia **Doença renal crônica, secundária a nefropatia diabética** dm e **has** **#DRC classe IV - CKD EPI 20** [EXAMES] CA 1,22 FÓSFORO 4,8 FA 76 #Exames: Hb 14,0/ Leuc 7810/ Plaq 228.000/ Creat 2,3/ Ur 106/ K 5,2/ Vit D 38/ EAS sem alterações/ Hb glicada 6,65%

	RESISTENCIA VASCULAR RENAL (PARENQUIMA) FOI REALIZADO CATETERISMO E COLOCAÇÃO DE STENT EM A RENAL ESQUEDA trás doppler com artérias renais com indices ao/ renais < que 0,35 . FUROSEMIDA 40MG /DIA , **#HMA: PACIENTE RENAL CRONICA HÁ APROXIMADAMENTE 4 ANOS , EM EXAME DE TRIAGEM DOS VASOS RENAIS POR DOPPLER FOI COSTATADA OBSTRUÇÃO DE ARTERIA RENAL ESQUERDA (11/2015)**	Relação albumina/ creatinina 61,7 mg/g (microalbuminúria) [MEDICAÇÕES] EM USO DE **INSULINA**, LINAGLIPTINA, **ENALAPRIL**, **HCTZ**, **ATENOLOL**, SINVASTATINA, AAS, OMEPRAZOL **Relata hipertensão e faz uso de inalapril 10mg, atenolol 50mg, hidroclorotiazida 25mg** [PROCEDIMENTOS] po atp arteria renal direita 25/11/15 [ATRIBUTOS CLINICOS] HMP: Em 2013 disúria, dor na bexiga com tratamento eficaz, mas não lembra os medicamentos TRAÇOS DE PROTEÍNA, 1 ERITRÓCITO, E LEUCÓCITO, AUSENCIA DE CILINDROS, CRISTAIS E FILAMENTO DE MUCO
P2 lowest F1	**DRC IV**, em acompanhamento ambulatorial desde 2013 **DRC de etiologia desconhecida** Nega DM e HAS MUC: nao faz uso de medicações contínuas **MUC: furosemida 40 mg/dia** **Cr 3,0 CKD 19,7** **US 04/06/13 - cisto renal simples a direita** anemia normo normo com **vitamina b12 baixa** hipercalemia - último exame: 5,8 acidose metabólica - bic: 24	[COMORBIDADES] **DRC estadio 4 - etiologia indeterminada** acidose metabólica - bic: 24 **#DRC ESTAGIO 4 DE CAUSA INDETERMINADA** US: rins diminuidos, boa diferenciação cortico-medular [MEDICAÇÕES] #Na ultima consulta foi prescrito **furosemida 40mg/dia**+ albendazol(eosinofilia) # Refere que fez uso de sulfato ferroso e vit b por 3 meses, supsendou pois nao sabia se deveria continuar # Ultima consulta em 09/12 : preecrito sulfato ferroso 40mg/d + complexo b 1 cp/d Foi receitado vitamina e diurético (sic) - não sabe informar quais [EXAMES] **CREATININA 2,5**; UREIA 53; **EPI 25,1**(**DRC ESTÁGIO IV**);hb 10,5 hcm 32%, leucocitos 7280; HDL 38; LDL 125; TRIGLICERIDEOS 130; CALCIO 8,8; FOSFORO 3,6; PTH 86,1 # Labs 24/02/2017: hb 11,7 vcm 97 leuco 190310 eosinofilos 32%* plaq 210.000; ferrtina 37, **vit b12 183**, saturacao transferrina 37 Ferritina 17 Indice de Sat 22,6% Hb 10,4 K 5,6 Lab: **Creat 3,1 >> 3,1 >> 2,8** **CKD: 19,1 - DRC estágio 4** [ATRIBUTOS CLINICOS] Paciente encaminhado da unidade basica devido inapetencia e fraqueza, com historico de anemia

		ECO COM CISTO RENAL SIMPLES A DIREITA

Source: the author.

Even on the lowest scoring summaries, the **readability** of the Sequence Labeling approach is better than the BERT method. The created sections, representing the annotation categories, make it easier to find information and offers a more pleasant reading. The coverage of terms found by the NER algorithm could be an issue, since some categories did not have any sentence selected, even with information present in the original texts. However, even in the gold standard, we do not find all categories.

Again, patient 2 had the worst quantitative results. The patient with the best performance (i.e., patient 10) has a similar amount of information; however, he is a patient with a greater variety of associated comorbidities and clinical notes from more medical specialties.

5.5.4 Supervised Dictionary-based summarization

Concerning the Precision and Recall scores, in general, the Recall results increased with more sentences, and the Precision decreased (see Table 38). The method achieved the best average results on patients 5, 6, 7, and 10.

It is important to remind that the results of this approach were penalized in ROUGE scores since the Exam category was not included textually in the final summary but plotted as a line graph. In this scenario, it was not possible for the summarizer to select sentences with exam results, and consequently making it impossible to achieve the best performance.

Table 39 displays the summaries of patients 6 and 8, which achieved the higher and lowest F1-scores, respectively. Figure 25 and 26 show the line graphs plotted with the exams found within the clinical notes, representing the exam results section. Both patients presented some issues related to **redundancy**, mainly concerning the hypertension diagnosis (i.e., HAS).

The source-oriented view, where each category is displayed in its own section, improves the **readability**, as happened with the Sequence Labeling approach. The insertion of exams plotted as graphs instead of text prevented the problem of showing not up to date results and could give a better overview of the patient and a more accurate perspective on the progression of the disease.

Table 38 – ROUGE scores for each patient's Dictionary-based summary. The Precision (P), Recall (R), and F1 (F) scores were calculated for each metric (ROUGE-SU4, ROUGE-1, and ROUGE-2). Three different output sizes (number of sentences) were assessed (N10, N15, N20).

			P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
N10	ROUGE-SU4	P	0,183	0,108	0,204	0,135	0,328	0,518	0,247	0,081	0,078	0,298
		R	0,185	0,160	0,176	0,110	0,192	0,305	0,308	0,030	0,074	0,074
		F	0,184	0,129	0,189	0,121	0,243	0,384	0,274	0,044	0,076	0,119
	ROUGE-1	P	0,282	0,236	0,397	0,323	0,398	0,737	0,364	0,317	0,195	0,542
		R	0,284	0,344	0,345	0,265	0,236	0,442	0,453	0,122	0,188	0,141
		F	0,283	0,280	0,369	0,291	0,296	0,553	0,404	0,176	0,191	0,224
	ROUGE-2	P	0,174	0,114	0,250	0,163	0,337	0,607	0,244	0,085	0,049	0,319
		R	0,176	0,167	0,217	0,134	0,199	0,362	0,304	0,032	0,047	0,082
		F	0,175	0,135	0,232	0,147	0,250	0,453	0,271	0,047	0,048	0,130
N15	ROUGE-SU4	P	0,119	0,089	0,153	0,118	0,231	0,275	0,212	0,086	0,136	0,385
		R	0,191	0,166	0,189	0,121	0,222	0,352	0,381	0,076	0,164	0,243
		F	0,147	0,116	0,169	0,119	0,227	0,309	0,272	0,081	0,149	0,298
	ROUGE-1	P	0,202	0,207	0,320	0,310	0,364	0,446	0,319	0,281	0,294	0,607
		R	0,321	0,377	0,393	0,319	0,350	0,568	0,568	0,250	0,352	0,386
		F	0,248	0,267	0,353	0,314	0,357	0,500	0,408	0,264	0,320	0,472
	ROUGE-2	P	0,110	0,091	0,176	0,130	0,233	0,325	0,211	0,065	0,132	0,405
		R	0,176	0,167	0,217	0,134	0,224	0,415	0,377	0,058	0,157	0,257
		F	0,136	0,118	0,195	0,132	0,229	0,364	0,270	0,061	0,143	0,314
N20	ROUGE-SU4	P	0,116	0,089	0,149	0,115	0,199	0,249	0,199	0,212	0,128	0,345
		R	0,193	0,166	0,191	0,122	0,224	0,354	0,389	0,255	0,170	0,248
		F	0,145	0,116	0,167	0,119	0,211	0,292	0,263	0,232	0,146	0,289
	ROUGE-1	P	0,200	0,207	0,318	0,300	0,318	0,410	0,309	0,332	0,284	0,564
		R	0,330	0,377	0,405	0,319	0,357	0,579	0,597	0,397	0,375	0,408
		F	0,249	0,267	0,356	0,309	0,336	0,480	0,407	0,362	0,323	0,473
	ROUGE-2	P	0,106	0,091	0,170	0,126	0,200	0,293	0,194	0,199	0,119	0,356
		R	0,176	0,167	0,217	0,134	0,224	0,415	0,377	0,239	0,157	0,257
		F	0,132	0,118	0,190	0,130	0,211	0,344	0,256	0,217	0,136	0,298

Source: the author.

It is worth to mention some issues found when analyzing the produced summaries. For both patients in Table 39, the main diagnosis (i.e., CKD) was not shown, the existence of various comorbidities associated with the patients (as shown in Table 30) may be the cause of it.

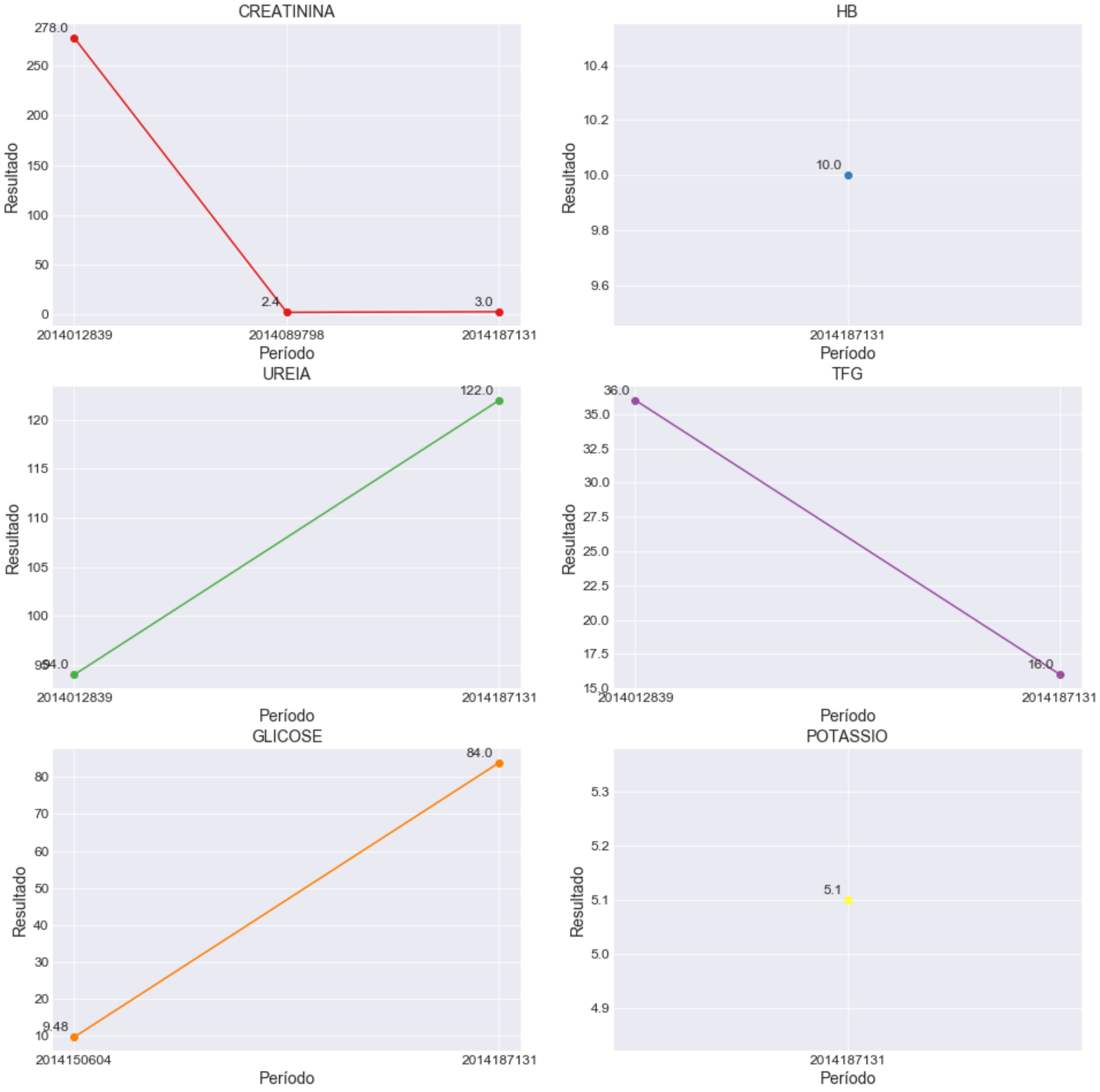
Table 39 - Examples of Dictionary-based summaries with high and low F1 score, compared with their respective gold standards. In bold semantically similar information.

Patient	Extractive gold standard summary	Dictionary-based N15 summary
P6 highest F1	Nefropatia Diabética **DM** (há 16 anos) **HAS (há cerca de 10 anos)** **DAC - ATC em junho 2017** Enalapril 10 mg 1x/dia (uso irregular) **Furosemida 40 mg/dia** **Carvedilol 6,25 2 cp cedo e 2 à noite** **Insulina NPH (35 - 0 - 20)** 3. DRC Estágio IV - Cr 3,0, TFG 16 (CKD-EPI) Paciente com queixa de dor precordial e dispnéia aos esforços **Sulfato ferroso 300 mg 12/12h** **Carbonato de Cálcio 2cp após o almoço e 1 após o jantar** Hiperfosfatemia Cisto simples no rim esquerdo	[COMORBIDADES] #**DM** , **HAS** E HIPOTIREOIDISMO paciente com colelitiase, com **dm** e **has** - **HAS há 05a** **HAS** [MEDICAÇÕES] Suspendo **sulfato ferroso** **Carbonato de Cálcio 2cp após o almoço e 1 após o jantar** **Sulfato ferroso 300 mg 12/12h** Levotiroxina 100 mcg **Insulina NPH (35 - 0 - 20)** **Carvedilol 6,25 2 cp cedo e 2 à noite** **Furosemida 40 mg/dia** [PROCEDIMENTOS] **# DAC - ATC em junho 2017** [ATRIBUTOS CLINICOS] **Diabética** e hipertensa com tendencia para hipercalemia NEGA FEBRE ASSOCIADA [CONDUTA ULTIMA VISITA] Suspendo sulfato ferroso - Retorno em 3 meses com exames laboratoriais
P8 lowest F1	**HA** há >10 a HIPOTIREOIDISMO ASMA - ACOMPANHA COM PNEUMOLOGISTA MP paciente colocou marca-passo em 2009, fez 2 cateterismos. DRC POR SÍNDROME CARDIO RENAL TIPO II COM CLEARENCE DE CREATININA ESTIMADO DE 19 E MEDIDO DE 20 ML/MIN Hiperparatireoidismo **ICC** POR MIOCARDIOPATIA DILATADA POR DOENÇA DE CHAGAS - SEM DOENÇA CORONARIANA - CINEANGIOCORONARIOGRAFIA (29/03/2016) QUE NÃO EVIDENCIOU LESÕES OBSTRUTIVAS EM USO PRÉVIO DE: LEVOTIROXINA 75MCG/DIA, **LOSARTANA 50MG/D**, **CARVEDILOL 25MG 12/12 HORAS**, **FUROSEMIDA 40MG 1X/DIA**, **SINVASTATINA 20MG/DIA**, PANTOPRAZOL, ALENIA 2X/DIA, ETODOLACO 400MG 12/12H (PRESCRITO PELO ORTOPEDISTA). REFERE DIMINUIÇÃO DO VOLUME URINÁRIO APÓS INCIAR USO DE RETEMIC, PAROU A MEDICAÇÃO POR CONTA, MAS O SINTOMA PERMANECEU	[COMORBIDADES] **HAS** **#HAS** **ICC com marca-passo há 6 anos** HMF: mãe com ICC, HAS e diabetes, faleceu com 63 anos de infarto [MEDICAÇÕES] **Furosemida 40mg 1/2-0-0** Sulfato Ferroso 300mg 1-1-1 **Sinvastatina 20mg 0-0-1** **Carvedilol 25mg 1-0-1** **Losartana 50mg 1/2-0-1/2** **Ajusto dose da Losartana 50mg 1-0-1** [PROCEDIMENTOS] conduta: solicitado US de partes moles para investigar area do implante do marca-passo [ATRIBUTOS CLINICOS] DOR EM QUEIMAÇÃO EM EPIGÁSTRO E NO LOCAL DO MARCAPASSO, QUE IRRADIA PARA OS BRAÇOS E QUE CESSAM COM A PAUSA NO ESFORÇO Hiperuricemia [CONDUTA ULTIMA VISITA]

	QUEIXA-SE DE GANHO DE PESO, MAS NÃO SOUBBE PRECISAR QUANTO, PESAVA 75 KG. REFERE URGÊINCONTINÊNCIA, INCONTINÊNCIA AOS EFORÇOS, HÁ 2 ANOS. QUEIXA-SE DE INCHAÇO EM MMII	RETORNO EM 3 MESES COM EXAMES - SOLICITO NOVOS EXAMES E RETORNO EM 4 MESES

Source: the author.

Figure 25 - Exam results plotted into a line graph (Patient 6 partial example).

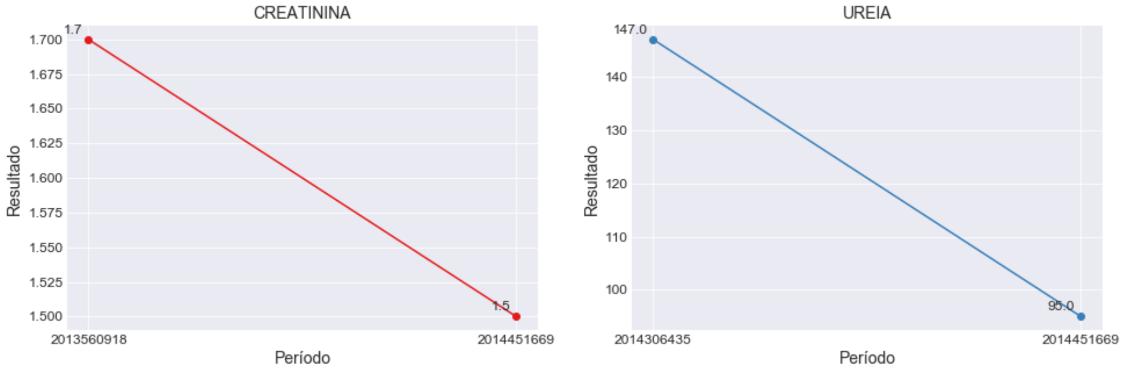


Source: the author.

Regarding the medication section, some issues should be indicated. The lack of available slots due to fragmented prescriptions (one medication per sentence) could be a problem, for instance, the patient 6 has multiple medications prescribed, but some of them were not in the final summary (e.g., "*Enalapril 10 mg 1x/dia*") because the medication section was already full.

Another issue occurred in the patient 6 summary, which shows both the suspension and the prescription of medication. “Suspendo sulfato ferroso” (*Suspending ferrous sulphate*) and “Sulfato ferroso 300 mg 12/12h” (*Ferrous sulfate 300 mg 12/12h*) were shown. In that case, is not possible to know which information is correct. Another similar issue appeared for patient 8. A dosage adjustment and the original dosage prescription are shown in the summary: “Losartana 50mg 1/2-0-1/2” (*Losartan 50mg 1/2-0-1/2*) and “Ajusto dose da Losartana 50mg 1-0-1” (*Losartan 50mg 1-0-1 dose adjustment*).

Figure 26 – Exam results plotted into a line graph (Patient 8 partial example).



Source: the author.

5.5.5 Comparative analysis of results

The ROUGE scores of the evaluated patients were averaged and presented in Table 40, grouped by summary size (i.e., number of sentences) and ROUGE metric. The unsupervised BERT-based summarization achieved the lowest scores between the three approaches. The Sequence labeling method reached the best scores mostly on small summaries (i.e., ten sentences), on the other hand, the Dictionary-based approach obtained the best results when it generated larger summaries (i.e., twenty sentences).

When averaging the F1-scores per patient, it is possible to check which method performed better for each patient and summary size (see Table 41). Although in the average of all metrics, the BERT method had the worst performance, including in larger summaries, when analyzing the patients individually it is possible to verify that 4 out of 10 patients (e.g., P1, P4, P5, and P8) had the best F1 with the BERT method for summaries with 20 sentences. These numbers suggest an irregularity in the method, which has very poor results for some patients (e.g., P2) and good results for others (e.g., P1). The method with most of the best F1 scores is the Dictionary-based, with half of every top-performing score for all summary sizes.

Table 40 - Average ROUGE scores for BERT-based (BERT), Sequence labeling (NER), and Dictionary-based (DICT) approach. The Precision (P), Recall (R), and F1 (F) scores were calculated for each metric (ROUGE-SU4, ROUGE-1, and ROUGE-2). Three different output sizes (number of sentences) were assessed (N10, N15, N20). In bold, the best values for each metric.

			BERT	NER	DICT
N10	ROUGE-SU4	P	0,189	0,212	**0,218**
		R	0,126	**0,258**	0,162
		F	0,149	**0,230**	0,176
	ROUGE-1	P	**0,393**	0,343	0,379
		R	0,280	**0,430**	0,282
		F	0,322	**0,376**	0,307
	ROUGE-2	P	0,182	0,210	**0,234**
		R	0,119	**0,253**	0,172
		F	0,142	**0,227**	0,189
N15	ROUGE-SU4	P	0,160	0,150	**0,180**
		R	0,158	**0,288**	0,210
		F	0,156	**0,196**	0,189
	ROUGE-1	P	**0,338**	0,256	0,335
		R	0,342	**0,500**	0,388
		F	0,334	0,335	**0,350**
	ROUGE-2	P	0,152	0,151	**0,188**
		R	0,146	**0,287**	0,218
		F	0,146	**0,196**	**0,196**
N20	ROUGE-SU4	P	0,152	0,149	**0,180**
		R	0,187	**0,290**	0,231
		F	0,165	0,195	**0,198**
	ROUGE-1	P	0,322	0,254	**0,324**
		R	0,407	**0,506**	0,414
		F	0,352	0,335	**0,356**
	ROUGE-2	P	0,148	0,148	**0,185**
		R	0,177	**0,288**	0,236
		F	0,159	0,194	**0,203**

Source: the author

The last note baseline scores, presented in Table 31, if compared with the three developed methods shows that 50% of the patients (P1, P2, P3, P4, and P9) achieved the highest ROUGE averaged F1 score using the last note. Four of these patients are among those with the least clinical notes (P1, P2, P3, P4), and the other one (P9) is the one with most clinical notes, however, one of the lowest values for "tokens per note", indicating very short texts. Moreover, P9 is the only patient ever transplanted from our test set, and one of the few within SummClinBr, that is, the important data for a summary of this type of patient changes considerably when compared to pre-dialysis and pre-transplant patients. Maybe this explains the below-average performance for this patient. In this context, the numbers indicate that the methods work best in cases with more information available (higher number of clinical notes) and that belong to the target group of our annotation (DRC 3 and 4 patients).

Table 41 – Average F1-scores (ROUGE-1, ROUGE-2, and ROUGE-SU4) per patient for BERT-based (BERT), Sequence labeling (NER) and Dictionary-based (DICT) approaches. Three different output sizes (number of sentences) were assessed (N10, N15, N20). In green, the best values per patient for each summary size.

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
N10	**BERT**	0,275	0,090	0,181	0,214	0,263	0,141	0,145	0,275	0,243	0,220
	NER	0,363	0,127	0,193	0,244	0,248	0,288	0,255	0,400	0,172	0,484
	DICT	0,214	0,181	0,264	0,187	0,263	0,463	0,316	0,089	0,105	0,158
N15	**BERT**	0,358	0,123	0,215	0,190	0,261	0,133	0,103	0,286	0,227	0,228
	NER	0,280	0,147	0,155	0,195	0,218	0,318	0,211	0,322	0,159	0,416
	DICT	0,177	0,167	0,239	0,189	0,271	0,391	0,317	0,135	0,204	0,361
N20	**BERT**	0,402	0,093	0,171	0,284	0,296	0,166	0,151	0,317	0,154	0,217
	NER	0,280	0,147	0,155	0,195	0,218	0,318	0,211	0,301	0,170	0,416
	DICT	0,175	0,167	0,238	0,186	0,253	0,372	0,309	0,270	0,202	0,353

Comparing the developed methods results with the Frequency-based baseline (Table 33), it is possible to check that all three methods performed better that the baseline if we average the F1-scores.

Regarding the summary size, the BERT-based method achieved most of its best results in larger summaries (20 sentences), the Sequence Labeling produced most of the best F1 scores in minor summaries (10 sentences), and the Dictionary-based approach reached its bests performances in medium-sized summaries (15 sentences), followed closely of 10 sentences summaries.

6 DISCUSSION

The methodology and results associated with the products of this thesis are discussed in this section. First, we revisit the research objectives and questions, and then we focus on the findings related to the research items by interpreting the results and commenting on the implications of them.

6.1 REVISITING RESEARCH GOALS AND QUESTIONS

This section has the intention to correlate the research goals with the developed methods and obtained results. Moreover, we describe how this thesis tried to respond to the research questions. This section can be used as a summary of all the achievements and findings of this thesis, and serve as a guide through the methodological course of research.

6.1.1 Research goals achievement

Below, we unravel all the **research goals** and summarize the methodological steps to accomplish each one of them.

Research goal 1
Develop a method to provide summarized EHR information of chronic disease patients by assembling essential clinical NLP resources and exploring existing supervised and unsupervised strategies.

The main objective of this work was to build an EHR summarizer focused on chronic disease patients, which have above average amounts of data and need to be continuously monitored by the health team, who should not be impacted by the patient's data overload and should be able to have a general overview of their condition, to prevent disease progression.

In this study, three different summarization methods are presented (sections 4.2.1, 4.5.1, and 4.5.2), all of them are adaptations or inspired by existing algorithms, which were chosen and delimited in section 2.5.2. The evaluation protocol included the comparison of produced summaries against human-generated summaries (i.e., gold standard). The results were measured quantitatively, using ROUGE metrics, and qualitatively by analyzing quality aspects as summary readability and redundancy (section 5.5). The main issues, considerations, and further analysis were discussed in the following sections.

Research goal 2

Explore and apply unsupervised, statistical, and low-resource summarization approaches.

The exploration of unsupervised and low-resource approaches has been proposed due to the lack of NLP resources and tools for the clinical context in the pt-br language. The good news is that most of the summarization methods are unsupervised; however, this is due precisely to the immaturity of the research area, which does not have an abundant amount of resources as a gold standard, besides that there are few studies applying summarization techniques in the clinical domain, including for the English language.

Initially, a preliminary experiment was executed, where a set of classic unsupervised summarization methods (e.g., TextRank, LSA, SumBasic) and an additional Frequency-based baseline were applied to a series of clinical notes of a CKD patient (section 4.1). Then, the algorithms included in the research scope (listed in section 2.5.2), were analyzed, and a BERT-based approach was chosen as the most adequate to experiment. Section 4.2.1 describes the method and the adaptations made in order to perform clinical summarization. The quantitative and qualitative evaluation results are presented in section 5.5.2. Additional remarks are made in the Discussion section.

Research goal 3

Build a set of clinical NLP tools and resources to support the identification of relevant information within the clinical text.

The structuring of background to support the clinical NLP in pt-br is vital to enable the training of several high-performance supervised methods, including information extraction and summarization algorithms, in addition to serving as a starting point for new research in the area. Moreover, a gold standard can serve as a common environment for evaluating methods, thus allowing us to compare different methods.

The identification of clinical concepts within the patients' texts is essential to understand the events that occurred during patient care and then to filter the most relevant data for summarization. With that in mind, a huge annotation effort was made to build the SemClinBr, a corpus annotated with clinical concepts to support the development and evaluation of clinical NLP tasks (section 3.1).

To extract the concepts from the clinical texts, three NER models were proposed, trained with the SemClinBr corpus. The first model uses a CRF classifier (section 3.2.1), the second utilizes a BiLSTM-CRF neural network and character embeddings (section 3.2.2), and the third relies on a BERT architecture, fine-tuned for the clinical domain (section 3.2.3).

As an attempt to improve the results of the NER algorithm, we proposed the use of a new POS-Tagging environment for clinical texts (section 3.3), since the POS-tag is one of the most important attributes for classifying concepts in a NER task, and until then there was no POS-Tagger for clinical texts in pt-br available for use. A BiLSTM-CRF architecture with contextual embeddings, trained in a clinical corpus annotated with morphological tags, was used for this.

Additionally, an algorithm for textual semantic similarity was proposed, aiming to measure the similarity between sentences, and with that, assist the summarization methods to remove redundant sentences from the final summary, which is one of the main issues in summarization systems. The model used a Siamese Neural Network architecture and lexical features to calculate the similarity score.

The evaluation of SemClinBr, the NER models, the POS-tagging, and the semantic similarity algorithm were presented in sections 5.1, 5.2, 5.3, and 5.4, respectively.

Research goal 4

Explore and apply supervised summarization approaches.

The exploration and application of supervised approaches have been proposed, and as in many NLP tasks, they produced the best summary results. Furthermore, some of the methods included in the scope of the thesis (section 2.5.2) use supervised approaches.

The first proposed method aims to treat the summarization task as a sequence labeling problem. The use of a NER algorithm trained with SemClinBr, which contains all types of clinical concepts and events, could be an option; however, it would require the use of an additional unsupervised statistical method, to select the concepts considered important to the patient and consequently to the summary. In this context, a second gold standard was developed, but this time focused exclusively on the clinical summarization task.

The corpus, called SummClinBr, is composed of clinical notes from patients with chronic kidney disease and involved three levels of annotation: (a) concept-level, (b) sentence-level (i.e., extractive summary) and (c) abstractive summary writing. An annotation guideline was developed, and the following set of annotation categories were defined: Comorbidity, Exam results, Medication, Procedure, Clinical attribute, Risk factor, and Last visit conduct (see section 4.3 for details).

Once the SummClinBr corpus was built, the first supervised method could be prepared (section 4.5.1). A NER algorithm was trained with SummClinBr data and was used to identify sentences that contained the patient’s relevant information. Summary reduction steps, sentence order, and redundancy identification were added to the method as well.

The second supervised method, called Dictionary-based (section 4.5.2), relied on a dictionary of terms, functioning as a lookup table using fuzzy-match logic. The dictionary was built with the terms annotated in SummClinBr. As the first method, additional reduction, ordering, and redundancy detection steps were added. In this method, we experimented with plotting the results of the exams instead of presenting them textually.

Both methods were evaluated quantitatively, using the SummClinBr gold standard to calculate ROUGE metrics, and qualitatively, where we superficially analyzed some quality aspects in the summaries.

6.1.2 Research questions response

This section revisits the presented research questions and based on the research and results of this study, aims to answer them.

Research question 1

How well unsupervised and low-resource summarization approaches could select the relevant information in the patient's EHR?

The preliminary experiments, made with a set of unsupervised summarization baselines in section 4.1, shown that the application of these methods in the clinical domain with no adaptation, results in a poor selection of relevant information (although we have not done a quantitative assessment). The baselines covered graph-based, semantic-based, and statistical-based approaches.

Regarding the adapted BERT-based method (section 4.2.1), which uses a neural network to produce a set of embeddings and then cluster them into vector space, if we check the ROUGE-1 metric, which considers unigrams overlap, the average Recall was 0.407. This indicates that about 40% of the gold-standard words were selected. Further experiments with additional clinical features and summarization methods should be executed as an attempt to improve the results.

Research question 2

How the lack of NLP tools and resources impacts the use of current unsupervised approaches?

The negative impact of the lack of resources and tools is substantial, as most unsupervised biomedical summarization methods use medical knowledge bases (such as UMLS and SNOMED-CT) and clinical data mapping tools (such as MetaMap and cTakes) to perform the summarization (as shown in section 2.5.2).

The more resources of clinical NLP for pt-br we have available, the greater the possibility of extracting significant information from the text and, consequently, the better the results.

Research question 3

A Named Entity Recognition algorithm trained with an Information Extraction corpus of semantically annotated clinical texts could aid in chronic disease relevant information identification?

Yes, but with some limitations. As described in section 5.5.3, the exclusive use of the SemClinBr (an Information Extraction corpus) for the selection of relevant data had inferior results comparing with the SummClinBr (a task-specific corpus), and should not be indicated to this task. However, the identification of clinical concepts in the text can be the starting point for the construction of a terminological mapping tool (e.g., MetaMap), which is widely used in the context of biomedical summarization. Moreover, it can be used to support the extraction of additional clinical features. For instance, we could adapt the BERT-based approach to work just with the clinical concepts and cleaning the rest of the text. Furthermore, a corpus containing additional temporality annotation can even be used to extract clinical concepts and build a patient's timeline, which can be considered as a time-oriented summary.

Research question 4

Which types of data are relevant in a summary of a patient with a chronic disease?

As discussed with the specialists (section 4.3), a summary aiming the follow-up consultations of a chronic disease patient should have at least the seven categories defined for SemClinBr corpus, which are: Comorbidities, Exam results, Medications in use, Procedures, Clinical attributes, Risk factors, and Last visit conduct. Information specific to each disease may be necessary, for example, for dialysis patients, who require even more specialized monitoring, depending on additional data.

Research question 5

A domain and task-specific gold standard (i.e., corpus for patient’s longitudinal data summarization) could improve the chronic disease relevant information identification?

Yes, as presented in the results sections of the two supervised approaches (sections 5.5.3 and 5.5.4), which made use of the SummClinBr corpus. In addition, a corpus with these characteristics can still be used to improve several other methods and be used to evaluate and benchmark summarization systems.

Research question 6

How well supervised summarization approaches could select the relevant information in the patient’s EHR?

The proposed supervised methods performed better than the unsupervised one; however, further improvements still need to be made. Regarding the sequence labeling method (section 4.5.1), if we check the ROUGE-1 metric, which considers unigrams overlap, the average Recall was 0.506. This indicates that about 50% of the gold-standard words were selected.

Research question 7

Deal with clinical summarization as a sequence labeling problem is feasible?

Yes, but additional experiments with extra features and classification techniques must be done. The selection of this type of approach is justified because, in the clinical domain, sentences are often short, and the meaning depends heavily on the context, which is taken into account when applying sequence labeling algorithms. However, when dealing with such a specialized domain (chronic disease summarization), the need for more specialized data is evident, and it makes the generalization of the method not trivial.

6.2 SEMCLINBR

In line with previous studies (XIA; YETISGEN-YILDIZ, 2012) that discussed the high complexity of clinical annotation, our results confirm the difficulty to achieve a perfect annotation agreement, even with a proper annotation environment, composed by a comprehensive set of guidelines, a customized annotation tool, clinically trained annotators, and constant reliability analysis.

Even reaching agreement levels similar to those of other studies, as shown in section 5.1.2, we understand that the fact that our annotation process has been 100% carried out in double annotation, it has a greater guarantee of quality, because we have a clearer view of the whole corpus. As our own project shows, below average agreement values emerge during certain project phases (e.g., because of a new annotation team or guidelines changes).

Our annotation guideline included all the UMLS semantic types, which are more error-prone than using more coarse-grained categories as other studies (FERREIRA et al., 2010; ALBRIGHT et al., 2013; CAMPILLOS et al., 2018), particularly if we consider semantic types in the same branch of the UMLS hierarchy tree (e.g., “Sign or

Symptom” and “Finding”). Moreover, the annotation of documents from multiple institutions brings to light some additional issues, such as dealing with diverse text formats due to specific institutional workflows and new sets of local abbreviations. To the best of our knowledge, no other clinical annotation study has covered documents from multiple institutions.

Furthermore, the multiple medical specialties and document types annotated added an extra layer of difficulty as well. The first scenario brings difficulty because the chance to find new challenging, ambiguous, and exception cases during annotation is higher if compared with a single medical specialty scope. For instance, the CLEF corpus (ROBERTS et al., 2009) covered just patients with neoplasms. The document type diversity could also affect the annotation, as the clinical notes are produced in different moments during the care workflow and are written by different professionals (physician, nurse, medical student, or intern); this can sometimes cause interpretation problems due to their different perspectives and brings, even more, writing styles to the table.

Regarding the developed annotation tool, the annotation team claimed that annotation suggestions based on previously labeled terms and UMLS API saved them a substantial amount of time. The UMLS API suggestion feature prevented the annotators from searching concepts on the UMLS Metathesaurus browser, which was an issue in a previous study (DELEGER et al., 2012), which indicated that the access to an online browser slowed down the annotation process.

Additional annotation types such as “Negation” and “Abbreviation” gave SemClinBr an even greater coverage of clinical NLP tasks, making it possible to tackle tasks such as negation detection and abbreviation disambiguation.

The recruitment of medical students to perform the annotation made it possible to finish the annotation much faster than if we had relied exclusively on doctors and other health professionals, due to their availability. In some sense, this statement reinforces Xia and Yetisgen-Yildiz (2012) claim that medical training is not the only factor to consider for biomedical annotation tasks.

6.3 SEQUENCE LABELING FOR CLINICAL TEXTS

The results obtained for the three different NER models are aligned with other NER studies concerning the incremental improvements of CRF, BiLSTM-CRF and

BioBERTpt models. For instance, in Lopes, Teixeira and Oliveira (2019) study, the BiLSTM-CRF architecture outperformed the CRF as well, posteriorly they applied an in-domain embedding and had another performance improvement.

A NER evaluation protocol considering partial matches should be made in our work, as in several cases the sequence labeler correctly found concepts within the text, but considered partial spans of text. Especially in our case, in which we are working with sentence-level summarization, the algorithm just needs to inform that the concept exists, the correct text span does not change the final result. Certainly, the results would have a considerable increase in performance, considering partial matches.

When we reduced the granularity of the NER annotations (using semantic groups instead of types) or specialized the application domain (using SummClinBr instead of SemClinBr), it becomes evident that the NER performed better, especially in neural network approaches that benefit from the increase in data when we transformed semantic types into groups. Regarding the POS-Tagging, the three tags with the least annotations in the gold-standard were the ones with the worst performance.

The availability of a clinical NER model is the first step towards the development of a terminological mapping tool, such as MetaMap and cTakes. Tools of this type are widely used in biomedical summarization (as shown in section 2.5) and can open doors to experiments on several published methods.

6.4 TEXTUAL SEMANTIC SIMILARITY FOR CLINICAL TEXTS

The textual semantic similarity system was developed and submitted to both (i) ASSIN 2 (Second Semantic Similarity and Textual Inference Evaluation) and (ii) 2019 n2c2 Shared-Task - Track 1: n2c2/OHNLP Track on Clinical Semantic Textual Similarity. The exclusive use of lexical features (which are independent of language or domain) is due to the fact that both challenges are in different languages and domains. This choice ended up having a negative impact on the results if we compare to better classified works in the shared-tasks (e.g., RODRIGUES; COUTO; RODRIGUES, 2019).

The adaptation of the model to work with contextual embeddings (as the one described in section 3.2.3) instead of distributional embeddings, should generate performance improvements since the two best-performing systems in the ASSIN 2 shared-task applied BERT in their methods (RODRIGUES; COUTO; RODRIGUES, 2019; RODRIGUES et al., 2019). The addition of specific features from the clinical domain should also help in improving the results.

6.5 CHRONIC DISEASE SUMMARIZATION

The implications of performing clinical text summarization can be seen when comparing the results of the methods and when viewing the patient's histories (section 5.5). What we could verify is that doctors try to insert as much information as possible into the patient's clinical note, putting together a kind of summary about the patient's condition, however, for various reasons related to the **hospital workflow** (e.g., lack of time, a visit related to a medical specialty other than the main disease), **stage of care** (e.g., first visit, investigation visit, follow-up visit) and **stage of disease** (e.g., chronic kidney disease level 3, transplanted patient), in some cases this does not occur. In cases where this premise is true, most of the relevant information about the patient is already described in his latest clinical note, making the challenge of summarizing data more focused on a single-document strategy, trying to remove unimportant information from the last note, as studies like (LIANG; TSOU; PODDAR, 2019).

It is possible to check the negative impact when the last note did not have this summary aspect when we presented the last note baseline results (section 5.5.1). The application of document clustering methods aiming to find specific patient and document characteristics could assist the summarization algorithm in understanding which clinical notes have the most relevant information, and then, try to extract sentences primarily from these documents. These particularities of clinical domain difficult the use of biomedical summarization methods, that often work with homogenous and syntactically perfect texts.

The negation detection, which is a frequent focus of several clinical information extraction studies, did not need to be applied in our study, since even the denied diagnoses were important to the summary (e.g., patient denies Diabetes). If we needed to use a specific visualization technique or if we used high levels of the AORTIS model to infer information, then it may be necessary to identify the negation. The good news is that both SemClinBr and SummClinBr already annotated data regarding negation.

The SummClinBr corpus could assist in both extractive and abstractive summarization task since it counts with extractive sentence-level annotation and human-generated abstractive summaries regarding the patient condition. Additionally, the concept-level annotation focus on labeling important concepts within the text. The summarization results, presented in section 5.5, evidenced that sometimes the sentence-level annotation does not count with all the seven available categories, which can be a little contradictory since the specialists defined all the seven categories as relevant information.

The results build on existing evidence that straight-forward techniques as string-matching could work better than machine learning-based approaches. The dictionary-based approach was the one with higher F1 scores in our evaluation. Another example of that is Lee et al. (2018) study, in which the string matching approach was more effective for identifying "Test," "Medication," and "Location" information than the Machine Learning method.

Not enough evidence was found, both in our work and in the literature, concerning the ideal size of a clinical summary of the patient. Since our results varied according to the method applied (see section 5.5.5), to alleviate the problem of data overload in a single category, we could choose to limit the size of the summary by the number of tokens and not by the sentences. An example of this is when all important medications are described in a single sentence, and in other cases, each medication occupies a sentence, leading to a lack of medication slots in the final summary.

The evaluation protocol results corroborate with previous studies, such as Lloret and colleagues (LLORET; PLAZA; AKER, 2018) that describe the subjective nature of creating gold-standard summaries and how the automatic evaluation methods, as ROUGE, could penalize semantically equivalent sentences. It is possible to find some initiatives to ease this evaluation issue that could be applied (NG; ABRECHT, 2015; SUN, NENKOVA, 2019). Moreover, as the human evaluation feedback on the plotted exam results was positive, the ROUGE scores penalized the lack of sentences corresponding to exams. Due to the lack of deeper manual evaluation, the results cannot confirm the real quality of the generated summaries regarding the text fluency, adequate length, and most specific quality measures.

The results suggest that the use of supervised approaches and domain-specific resources could improve the outcomes in the research area, as the intrinsic aspects related to the clinical data could be too complex to deal with statistical methods solely, although other studies have obtained good results combining unsupervised methods with advanced visualization techniques (LEVY-FIX, 2020). Furthermore, the use of a hybrid strategy that could explore the main strengths of each approach is promising.

It is still difficult to establish whether Precision or Recall is more important for our task, because while we want to avoid overloading information, that is, presenting as little data as possible to relieve the doctor time, we also want to obtain as much relevant information as possible, because a summary that contains insufficient information is unlikely to be of much use.

Regarding the summary format, the observational analyzes performed on the generated summaries (section 5.5) might suggest that textual extractive summarization solely could not be the best choice for clinical summarization, where the texts are fragmented, syntactically incorrect, and sometimes not fluent. Then, just selecting these sentences and showing them together could heavily impact on the quality of the summary. It is beyond the scope of this study to experiment with other visualization techniques, but we hypothesize that complementing the summary with other graphical options such as problem-oriented visualizations (LEVY-FIX et al., 2020), time-oriented views (VIANI et al., 2017) and medical trajectory graphs (ZAMORA; GAVALDA, 2019) may be the ideal choice for the clinical context. For instance, a graphical view of the exams could improve the understanding of disease progression and improve the avoidance of duplicated exams. Moreover, a timeline view could better show the emergence of new comorbidities. And finally, the use of abstractive summarization could be an option once we achieve optimal extractive performance, as one could change the way information is presented textually. The application of a NER to present indicative summaries could be very helpful to clinicians as well.

7 CONCLUSION AND FUTURE WORK

This concluding chapter aims to recapitulate the produced research work, highlight the main contributions of this thesis, describe the study limitations, and recommend research directions for future work.

7.1 CONCLUSION

The negative effects of the data overload in the patient's EHR bring concerns to health practitioners with regard to the decrease in doctor-patient time, and consequently, in the quality of care. In chronic disease patients, who have a condition of continuous treatment, which causes frequent medical visits, the problem of data overload is even more evident. In an attempt to attenuate this situation, we present in this thesis a series of methods for the **extraction of relevant data from the clinical notes of patients with chronic disease**, aiming to condense these data into a unique textual summary containing essential and important information regarding patient’s condition and treatment.

The lack of clinical NLP resources and tools available in the pt-br language has caused a large gap in the construction of computer systems that can effectively improve the quality of patient care, such as decision support systems and clinical research, including the task of clinical summarization. Therefore, we developed in this thesis a **set of resources and tools to support the most diverse tasks of NLP for the clinical domain and pt-br language**. A set of NER models, a POS-Tagging algorithm, and a Textual Semantic Similarity estimator have been assembled. Moreover, two corpora were built, the SemClinBr, which focuses on information extraction tasks, and the SummClinBr for patient's longitudinal data summarization.

We examined the application of different unsupervised and supervised summarization approaches. In the literature, there is **no definition of the best summarization methods for chronic disease summarization**, or even for biomedical summarization in general. Of the existing methods, each one was tested on a different dataset, with distinct experimental protocols, making it impossible to compare them. Because of this, we delimited the scope of studies by focusing on the ones that already performed experiments on chronic disease patients and methods which were resource-lean. Our evaluation protocol differed for each developed resource (e.g., NER, Text Similarity, summarizers); however, overall **emphasized intrinsic quantitative assessment**. Due to the limitations in the automatic evaluation metrics for summarization (i.e., ROUGE), we added an observational evaluation stage, aiming at simple quality aspects of the generated summaries.

The experimental results gave us the understanding that even working in such a specific domain (e.g., chronic kidney disease patients) the **data heterogeneity remains an issue**, as patients' data have (i) multiple writing styles, (ii) substantial size diversity, (iii) texts from diverse medical specialties, (iv) distinct phases of care and (v) different stages of the disease.

The evaluation suggests that **supervised approaches are able to reason over relevant information better than unsupervised**, as their performance was more regular over multiple summary sizes, however, both approaches should be further explored, especially aiming to a hybrid system. Additionally, a **single textual extractive summary of the patient may be insufficient** for the task, as the clinical notes issues could constrain the ability to provide a superior overview of the patient's condition, which could be better accomplished by applying additional visualization techniques and different summary perspectives. Furthermore, the dictionary-based performance evidence **that straight-forward techniques can still provide competitive performances** compared to more complex and modern techniques.

In low resource languages, plenty of extra work is needed in order to achieve **high-level and patient-state dependent summaries**, considering the AORTIS model (section 2.5.1). To reach this level of summarization is required resources to produce abstractive summarization in the target language and access to medical knowledge bases to aid the system in interpreting and synthesizing clinical relationships.

In conclusion, it is possible to build a clinical text summarizer focused on patients with chronic disease by developing valuable NLP resources and adapting existing extractive methods. However, further improvements to these intermediate resources (e.g., NER, gold-standards), production of even more complex resources (i.e., clinical terminology mapper), and additional adaptations to existing summarization methods (such as the application of different forms of presentation of the information) are still needed, so that we could reach summaries with a high level of abstraction.

7.2 CONTRIBUTIONS

This thesis contributes to the clinical NLP research by exploring some of its critical tasks that could increase, directly or indirectly, the quality of patient care. It is worth highlighting that the production and availability of resources and evidence of this study should contribute to the development and evaluation of several methods to come, including the clinical summarization ones. The complete set of contributions are the following:

- **Developed and evaluated a set of models for summarizing data from patients with chronic disease**. To the best of our knowledge, this was the first published attempt to perform the summarization of EHR textual data in pt-br, which produced valuable perceptions on different summarization approaches and provided insights on the summary output design.
- **Generated a longitudinal clinical summarization corpus**. The SummClinBr[11] has the ability to support the development of novel supervised approaches, and to provide researchers a gold standard that enables a quantitative benchmark between multiple summarization solutions, which is a gap not only in pt-br summarization but in English also. The different levels of annotation should allow its use on various distinct approaches.
- **Provided a complete framework for extracting clinical concepts from texts**. An extensive and meticulous annotation process resulted in SemClinBr[12], a corpus focused on information extraction that should support several clinical NLP tasks. We used the corpus to train three different NER models, which can extract clinical concepts with distinct levels of granularity within clinical notes.
- **Defined a state-of-the-art clinical POS-tagging environment**. We applied a neural network architecture known to produce high-performance results in sequence labeling tasks to a clinical morphosyntactic gold-

[11] https://github.com/HAILab-PUCPR/SummClinBr

[12] https://github.com/HAILab-PUCPR/SemClinBr

standard, creating a new state-of-the-art model[13] for pt-br clinical texts. This tool can assist in the improvement of NER algorithms, rule-based systems, and even abstractive summarization, where we need to understand the morphosyntactic role of words in the text.

- **Experimented a neural network-based summarization method**. One of the research gaps identified in section 2.5.3 was the lack of studies utilizing such methods. We not only used a method based on neural networks, but we also explored a Transformers architecture called BERT, which has the best results in several NLP tasks. Moreover, we investigated how to improve the results by fine-tuning a pre-trained model for the clinical context.
- **Identified a set of relevant information to a chronic disease patient**. We defined with a team of specialists which information is relevant to chronic disease patients by analyzing the clinical notes of a set of patients with chronic kidney disease and multiple comorbidities associated with cardiovascular diseases and diabetes.

7.3 LIMITATIONS

This study has a number of limitations, and we would like to recognize them below:

- **Evaluation protocol and benchmarking**. The performance of the summarization models was assessed by both quantitative and qualitative evaluations. The application of ROUGE to calculate the *informativeness* of the summary has some known limitations, as it considers just a lexical overlap between the generated summary and the human-generated summary, penalizing sentences that present different lexical formats but similar semantic meaning. At the same time, our quality evaluation was observational only, with no systematic method, which could not capture all the important aspects of the summary as factual correctness (ZHANG et al., 2019) or faithfulness, which would be essential for clinical decision.

[13] https://github.com/HAILab-PUCPR/portuguese-clinical-pos-tagger

Moreover, we could not assess how the summarization impacts the doctor-patient time.

- **Generalizability**. Our goal was to perform the summarization of chronic disease patients. To do that, we focused on patients from Nephrology with chronic kidney disease, known for the multimorbidity aspect and the overlap with Endocrinology (i.e., diabetes) and Cardiology (i.e., cardiovascular diseases). Although the definition of the relevant information categories took into account the three mentioned specialties, the evaluation of models in other medical specialties was not carried out. Furthermore, a more in-depth analysis is needed in order to define common aspects between chronic diseases, such as disease progression, patient care process in general, and the Brazilian health system workflow for each one of them.
- **Heuristics**. At certain points in this research, some heuristic-based thresholds for parameterizing the methods were defined. In the dictionary-based method, for example, we defined that the maximum edit distance to consider the match would be 1. The number of sentences that the summarizer should generate was also defined through intuition and conversation with the experts to know, on average, how many sentences would be necessary to describe the patient's condition. Because of this, we limit the tests to 10, 15, and 20 sentences. The redundancy resolution step has a similarity calculation algorithm that goes from 0 to 1. We defined that if the sentences obtain a similarity value less than or equal to 0.25, the sentences would be considered redundant. The modification of these thresholds should impact on the results.
- **Experimental scope**. At the beginning of this research project, a survey was carried out, and a set of candidate summary methods to be tried out was delimited (section 2.5.2). Because of this, a wide range of summarization methods stayed out of the experiments and should be tested in the clinical domain, especially considering all the resources developed in this thesis (SemClinBr, SummClinBr, POS-Tagger, NER), which allow the applicability of methods that were not possible at the beginning of the project. In this context, it is not yet possible to state which approaches and methods are best for summarizing data from patients with chronic disease.

- **Auxiliary resources impact**. As we exploit auxiliary resources as POS-Tagging and NER, a deeper evaluation is needed to measure the impact of errors coming from these tools in our summarization algorithms. Because sometimes the main cause of low performance could be the intermediary method and not the summarization itself.

7.4 FUTURE WORK

In this section, we recommend several research directions that could be explored in further studies. In addition, we also highlight some work in progress been developed by our research group, called **HAILab** (Health Artificial Intelligence Lab), which has a direct impact on future work related to this thesis

- **Perform patient and document clustering**. Our methods are based on the location of clinical concepts in the text and subsequent scoring and ordering of sentences. As evidenced, the clinical notes are very heterogeneous, varying in size, data completeness, writing style, medical specialty, phase of care, and stage of the disease. We hypothesize that if we could cluster all these aspects, we could know which document is more likely to have relevant information. For instance, if the last clinical note of a patient has a certain size and medical specialty, and it is originated from a follow-up consultation, this could indicate that this note has a good amount of relevant information.
- **Explore different summary visualizations and orientation**. The textual extractive summary generated in this thesis could be used in conjunction with other types of visualization and orientation aiming to improve the physician's understanding of the patient's whole picture and consequently saving his time. Time-oriented views could represent more realistic the continuous aspect of chronic-diseases, and visualization techniques that could help the medical team to visualize possible disease progression would be perfect.
- **Investigate how NER results are impacting the summarization method**. An optimal NER performance could improve the final summarization results.

Moreover, a high-performance NER would allow the application of more sophisticated approaches such as abstractive summarization.

- **Increase the summarization approaches scope**. A great number of summarization methods could be applied to the clinical summarization task, once we have some additional resources as POS-Taggers, NER, and terminology mapping tools. Further studies should analyze all those untested methods and apply them to this task. Moreover, one should focus on the high-level steps of AORTIS model (section 2.5.1), as this work was limited to the four first phases of it. The interpretation and synthesis steps should considerably improve the quality and usefulness of the clinical summaries.
- **Propose a clinical summarization evaluation method**. The summarization evaluation is a very challenging and subjective task. The most common approaches that focus on the content overlap and automatic evaluation (e.g., ROUGE), and the ones focused on quality aspects are often manual and time-consuming. An evaluation method that takes into account the specific aspects of the clinical area and that could quickly show quality indicators of the summary could help us to better understand the results of our work, and would also allow a fairer comparison between different methods.
- **Improve the textual semantic similarity method**. One of the major issues in clinical texts is the redundancy, which causes the selection of sentences with similar meaning in the summary. As our textual semantic similarity method was originally developed to work on multiple domains and languages, we just extracted lexical features. In order to improve the results, clinical features should be extracted, and the BioBERTpt should replace the current word embeddings.
- **Develop a terminology mapping tool**. Such a tool is already under development in our research group (HAILab) and can assist in solving numerous clinical NLP tasks, including summarizing patient data, in which several studies in the English language have explored similar tools such as MetaMap and cTakes.

- **Annotate and reason temporal information**. This is another example of a project being developed at HAILab. A corpus called TempClinBr was built, and algorithms are being developed to extract concepts and identify them temporally within the patient's history. Such a resource is extremely important when assembling a time-oriented summary (i.e., patient timeline).
- **Clinical dependency parsing**. The HAILab team is syntactically annotating the SemClinBr texts, in order to generate a corpus to train a clinical dependency parsing. This functionality can be used in several NLP tasks, such as abstractive summarization, entity linking, and negation detection. Furthermore, with the ability to syntactically parse the clinical notes, it is possible to focus on other extractive summarization levels (ZHOU; WEI; ZHOU, 2020), not only the sentence-level.

REFERENCES

ACHARYA, S.; DI EUGENIO, B.; BOYD, A.; CAMERON, R.; DUNN LOPEZ, K.; MARTYN-NEMETH, P.; DICKENS, C.; ARDATI, A. Towards Generating Personalized Hospitalization Summaries. 2018. **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018. p. 74–82. DOI 10.18653/v1/N18-4011. Available at: http://aclweb.org/anthology/N18-4011.

AKBIK, A.; BLYTHE, D.; VOLLGRAF, R. Contextual string embeddings for sequence labeling. 2018. **Proceedings of the 27th International Conference on Computational Linguistics** [...]. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 1638–1649. Available at: http://aclweb.org/anthology/C18-1139.

ALBRIGHT, D.; LANFRANCHI, A.; FREDRIKSEN, A.; STYLER, W. F.; WARNER, C.; HWANG, J. D.; CHOI, J. D.; DLIGACH, D.; NIELSEN, R. D.; MARTIN, J.; WARD, W.; PALMER, M.; SAVOVA, G. K. Towards comprehensive syntactic and semantic annotations of the clinical narrative. **Journal of the American Medical Informatics Association**, vol. 20, no. 5, p. 922–930, Sep. 2013. DOI 10.1136/amiajnl-2012-001317. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2012-001317.

ALEKSIĆ, D.; RAJKOVIĆ, P.; VUČKOVIĆ, D.; JANKOVIĆ, D.; MILENKOVIĆ, A. Data summarization method for chronic disease tracking. **Journal of Biomedical Informatics**, vol. 69, p. 188–202, May 2017. DOI 10.1016/j.jbi.2017.04.012. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1532046417300825.

AL-HEGAMI, A. S.; FAREA OTHMAN, A. M.; BAGASH, F. T. A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set. **International Journal of Computer Science and Network Security**, vol. 17, no. 1, p. 170–176, 2017. Available at: http://paper.ijcsns.org/07_book/201701/20170126.pdf.

AL-HEGAMI, A. S.; FAREA OTHMAN, A. M.; BAGASH, F. T. A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set. **INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY**, vol. 17, no. 1, 2017.

ALKUREISHI, M. A.; LEE, W. W.; LYONS, M.; PRESS, V. G.; IMAM, S.; NKANSAH-AMANKRA, A.; WERNER, D.; ARORA, V. M. Impact of Electronic Medical Record Use on the Patient–Doctor Relationship and Communication: A Systematic Review. **Journal of General Internal Medicine**, vol. 31, no. 5, p. 548–560, 19 May 2016. DOI 10.1007/s11606-015-3582-1. Available at: http://link.springer.com/10.1007/s11606-015-3582-1.

ANDRADE, G. H. B.; OLIVEIRA, L. E. S. e; MORO, C. M. C. METODOLOGIAS E FERRAMENTAS PARA ANOTAÇÃO DE NARRATIVAS CLÍNICAS. 2016. **CBIS 2016 - XV Congresso Brasileiro de Informática em Saúde** [...]. Goiânia: [*s. n.*], 2016. p. 1031–1040.

ARNDT, B. G.; BEASLEY, J. W.; WATKINSON, M. D.; TEMTE, J. L.; TUAN, W.-J.; SINSKY, C. A.; GILCHRIST, V. J. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. **The Annals of Family Medicine**, vol. 15, no. 5, p. 419–426, 11 Sep. 2017. DOI 10.1370/afm.2121. Available at: http://www.annfammed.org/lookup/doi/10.1370/afm.2121.

ARONSON, A. R.; LANG, F.-M. An overview of MetaMap: historical perspective and recent advances. **Journal of the American Medical Informatics Association**, vol. 17, no. 3, p. 229–236, 1 May 2010. DOI 10.1136/jamia.2009.002733. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.2009.002733.

ARTSTEIN, R.; POESIO, M. Inter-Coder Agreement for Computational Linguistics. **Computational Linguistics**, vol. 34, no. 4, p. 555–596, Dec. 2008. DOI 10.1162/coli.07-034-R2. Available at: http://www.mitpressjournals.org/doi/10.1162/coli.07-034-R2.

BETHARD, S.; CHEN, W. Te; PUSTEJOVSKY, J.; SAVOVA, G.; DERCZYNSKI, L.; VERHAGEN, M. SemEval-2016 task 12: Clinical TempEval. 2016. **SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings** [...]. [*S. l.*: *s. n.*], 2016. https://doi.org/10.18653/v1/s17-2093.

BETHARD, S.; DERCZYNSKI, L.; SAVOVA, G.; PUSTEJOVSKY, J.; VERHAGEN, M. SemEval-2015 Task 6: Clinical TempEval. 2015. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015. p. 806–814. DOI 10.18653/v1/S15-2136. Available at: http://aclweb.org/anthology/S15-2136.

BETHARD, S.; SAVOVA, G.; PALMER, M.; PUSTEJOVSKY, J. SemEval-2017 Task 12: Clinical TempEval. Aug. 2017. **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)** [...]. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017. p. 565–572. DOI 10.18653/v1/S17-2093. Available at: https://www.aclweb.org/anthology/S17-2093.

BOISEN, S.; CRYSTAL, M. R.; SCHWARTZ, R.; STONE, R.; WEISCHEDEL, R. Annotating Resources for Information Extraction. 2000. **Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)** [...]. Athens: [*s. n.*], 2000. p. 1211--1214. Available at: http://www.lrec-conf.org/proceedings/lrec2000/pdf/263.pdf.

BRETONNEL COHEN, K.; DEMNER-FUSHMAN, D. **Biomedical Natural Language Processing**. Amsterdam: John Benjamins Publishing Company, 2014. vol. 11, (Natural Language Processing). DOI 10.1075/nlp.11. Available at: http://www.jbe-platform.com/content/books/9789027271068.

CAMPILLOS, L.; DELÉGER, L.; GROUIN, C.; HAMON, T.; LIGOZAT, A.-L.; NÉVÉOL, A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). **Language Resources and Evaluation**, vol. 52, no. 2, p. 571–601, 15 Jun. 2018. DOI

10.1007/s10579-017-9382-y. Available at: http://link.springer.com/10.1007/s10579-017-9382-y.

CARBONELL, J.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. 1998. **Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98** [...]. New York, New York, USA: ACM Press, 1998. p. 335–336. DOI 10.1145/290941.291025. Available at: http://portal.acm.org/citation.cfm?doid=290941.291025.

CHALAPATHY, R.; ZARE BORZESHI, E.; PICCARDI, M. Bidirectional LSTM-CRF for Clinical Concept Extraction. Dec. 2016. **Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)** [...]. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016. p. 7–12. Available at: https://www.aclweb.org/anthology/W16-4202.

CLARKE, M. A.; MOORE, J. L.; STEEGE, L. M.; KOOPMAN, R. J.; BELDEN, J. L.; CANFIELD, S. M.; KIM, M. S. Toward a patient-centered ambulatory after-visit summary: Identifying primary care patients' information needs. **Informatics for Health and Social Care**, vol. 43, no. 3, p. 248–263, 3 Jul. 2018. DOI 10.1080/17538157.2017.1297305. Available at: https://www.tandfonline.com/doi/full/10.1080/17538157.2017.1297305.

COUSER, W. G.; REMUZZI, G.; MENDIS, S.; TONELLI, M. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. **Kidney International**, vol. 80, no. 12, p. 1258–1270, Dec. 2011. DOI 10.1038/ki.2011.368. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0085253815550047.

DA SILVA CONRADO, M.; FELIPPO, A. Di; SALGUEIRO PARDO, T. A.; REZENDE, S. O. A survey of automatic term extraction for Brazilian Portuguese. **Journal of the Brazilian Computer Society**, vol. 20, no. 1, p. 12, 30 Dec. 2014. DOI 10.1186/1678-4804-20-12. Available at: https://journal-bcs.springeropen.com/articles/10.1186/1678-4804-20-12.

DABEK, F.; JIMENEZ, E.; CABAN, J. J. A timeline-based framework for aggregating and summarizing electronic health records. Oct. 2017. **2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)** [...]. [*S. l.*]: IEEE, Oct. 2017. p. 55–61. DOI 10.1109/VAHC.2017.8387501. Available at: https://ieeexplore.ieee.org/document/8387501/.

DALIANIS, H. Characteristics of Patient Records and Clinical Corpora. **Clinical Text Mining**. Cham: Springer International Publishing, 2018. p. 21–34. DOI 10.1007/978-3-319-78503-5_4. Available at: http://link.springer.com/10.1007/978-3-319-78503-5_4.

DAVIDOFF, F.; MIGLUS, J. Delivering Clinical Evidence Where It's Needed. **JAMA**, vol. 305, no. 18, p. 1906, 11 May 2011. DOI 10.1001/jama.2011.619. Available at: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2011.619.

DE NOVAIS, E. M.; PARABONI, I. Portuguese text generation using factored language models. **Journal of the Brazilian Computer Society**, vol. 19, no. 2, p. 135–146, 24

Jun. 2013. DOI 10.1007/s13173-012-0095-1. Available at: https://journal-bcs.springeropen.com/articles/10.1007/s13173-012-0095-1.

DE OLIVEIRA, L. F. A.; E OLIVEIRA, L. E. S.; GUMIEL, Y. B.; CARVALHO, D. R.; MORO, C. M. C. Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts. **Research on Biomedical Engineering**, vol. 36, no. 3, p. 267–276, 19 Sep. 2020. DOI 10.1007/s42600-020-00067-7. Available at: http://link.springer.com/10.1007/s42600-020-00067-7.

DE SOUZA, JOÃO VITOR ANDRIOLI GUMIEL, Y. B.; OLIVEIRA, LUCAS EMANUEL SILVA MORO, C. M. C. Named Entity Recognition for Clinical Portuguese Corpus with Conditional Random Fields and Semantic Groups. 2019. **Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde** [...]. [*S. l.*: *s. n.*], 2019. Available at: https://portaldeconteudo.sbc.org.br/index.php/sbcas/article/view/6269.

DE SOUZA, J. V. A.; E OLIVEIRA, L. E. S.; GUMIEL, Y. B.; CARVALHO, D. R.; MORO, C. M. C. Exploiting siamese neural networks on short text similarity tasks for multiple domains and languages. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, vol. 12037 LNAI, p. 357–367, 2020. https://doi.org/10.1007/978-3-030-41505-1_34.

DE VINE, L.; ZUCCON, G.; KOOPMAN, B.; SITBON, L.; BRUZA, P. Medical semantic similarity with a neural language model. **Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CIKM '14**, , p. 1819–1822, 2014. DOI 10.1145/2661829.2661974. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84937604534&partnerID=40&md5=ea7e4efe39692579f3877685fb789651.

DEMNER-FUSHMAN, D.; CHAPMAN, W. W.; MCDONALD, C. J. What can natural language processing do for clinical decision support ? **Journal of Biomedical Informatics**, vol. 42, no. 5, p. 760–772, 2009. DOI 10.1016/j.jbi.2009.08.007. Available at: http://dx.doi.org/10.1016/j.jbi.2009.08.007.

DEVARAKONDA, M.; DONGYANG ZHANG; CHING-HUEI TSOU; BORNEA, M. Problem-oriented patient record summary: An early report on a Watson application. Oct. 2014. **2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)** [...]. [*S. l.*]: IEEE, Oct. 2014. p. 281–286. DOI 10.1109/HealthCom.2014.7001855. Available at: http://ieeexplore.ieee.org/document/7001855/.

DI EUGENIO, B.; BOYD, A. D.; LUGARESI, C.; BALASUBRAMANIAN, A.; KEENAN, G. M.; BURTON, M.; MACIEIRA, T. G. R.; LOPEZ, K. D.; FRIEDMAN, C.; LI, J. PatientNarr: Towards generating patient-centric summaries of hospital stays. **INLG 2014 - Proceedings of the Eighth International Natural Language Generation Conference**, vol. 2, no. June, p. 6, 2014. Available at: http://www.aclweb.org/anthology/W14-4402.

DING, P.; LI, F. Causal inference: A missing data perspective. **Statistical Science**, vol. 33, no. 2, 2018. https://doi.org/10.1214/18-STS645.DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, vol. 55, no.

10, p. 78, 1 Oct. 2012. DOI 10.1145/2347736.2347755. Available at: http://dl.acm.org/citation.cfm?doid=2347736.2347755.

ELEKES, Á.; ENGLHARDT, A.; SCHÄLER, M.; BÖHM, K. Toward meaningful notions of similarity in NLP embedding models. **International Journal on Digital Libraries**, 20 Apr. 2018. DOI 10.1007/s00799-018-0237-y. Available at: https://doi.org/10.1007/s00799-018-0237-y.

EL-KASSAS, W. S.; SALAMA, C. R.; RAFEA, A. A.; MOHAMED, H. K. Automatic Text Summarization: A Comprehensive Survey. **Expert Systems with Applications**, , p. 113679, Jul. 2020. DOI 10.1016/j.eswa.2020.113679. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0957417420305030.

ERKAN, G.; RADEV, D. R. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. **Journal of Artificial Intelligence Research**, vol. 22, p. 457–479, 1 Dec. 2004. DOI 10.1613/jair.1523. Available at: https://jair.org/index.php/jair/article/view/10396.

FARRI, O.; PIECKIEWICZ, D. S.; RAHMAN, A. S.; ADAM, T. J.; PAKHOMOV, S. V; MELTON, G. B. A qualitative analysis of EHR clinical document synthesis by clinicians. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, vol. 2012, no. 11, p. 1211–20, 2012. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540510/pdf/amia_2012_symp_1211.pdf.

FEBLOWITZ, J. C.; WRIGHT, A.; SINGH, H.; SAMAL, L.; SITTIG, D. F. Summarization of clinical information: A conceptual model. **Journal of Biomedical Informatics**, vol. 44, no. 4, p. 688–699, Aug. 2011. DOI 10.1016/j.jbi.2011.03.008. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1532046411000591.

FERREIRA, L.; OLIVEIRA, C. T.; TEIXEIRA, A.; CUNHA, J. P. da S. Extracção de Informação de Relatórios Médicos. **Linguamática**, vol. 1, no. Maio, p. 89–102, 2009.

FERREIRA, L.; TEIXEIRA, A.; CUNHA, J. P. da S. Information Extraction from Portuguese Hospital Discharge Letters. 2010. **VI Jornadas en Technologia del Habla and II Iberian SL Tech Workshop** [...]. [*S. l.*: *s. n.*], 2010. p. 39–42.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. **Artificial Intelligence Review**, vol. 47, no. 1, p. 1–66, 29 Jan. 2017. DOI 10.1007/s10462-016-9475-9. Available at: http://link.springer.com/10.1007/s10462-016-9475-9.

GATT, A.; KRAHMER, E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, vol. 61, p. 65–170, 27 Jan. 2018. DOI 10.1613/jair.5477. Available at: https://jair.org/index.php/jair/article/view/11173.

GATT, A.; REITER, E. SimpleNLG : A realisation engine for practical applications. **Proceedings of the 12th European Workshop on Natural Language Generation**, no. March, p. 90–93, 2009. Available at: https://aclanthology.coli.uni-saarland.de/papers/W09-0613/w09-0613.

GOLDSTEIN, A.; SHAHAR, Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. **Journal of Biomedical Informatics**, vol. 61, p. 159–175, Jun. 2016. DOI 10.1016/j.jbi.2016.03.022. Available at: http://dx.doi.org/10.1016/j.jbi.2016.03.022.

GOLDSTEIN, A.; SHAHAR, Y. Generation of Natural-Language Textual Summaries from Longitudinal Clinical Records. **Studies in health technology and informatics**, vol. 216, p. 594–8, 2015. DOI 10.3233/978-1-61499-564-7-594. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26262120.

GOLDSTEIN, A.; SHAHAR, Y.; ORENBUCH, E.; COHEN, M. J. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. **Artificial Intelligence in Medicine**, vol. 82, p. 20–33, Oct. 2017. DOI 10.1016/j.artmed.2017.09.001. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0933365716304651.

GÜNGÖR, O.; GÜNGÖR, T.; ÜSKÜDARLI, S. The effect of morphology in named entity recognition with sequence tagging. **Natural Language Engineering**, vol. 25, no. 1, p. 147–169, 27 Jan. 2019. DOI 10.1017/S1351324918000281. Available at: https://www.cambridge.org/core/product/identifier/S1351324918000281/type/journal_article.

GUO, S.; XU, K.; ZHAO, R.; GOTZ, D.; ZHA, H.; CAO, N. EventThread: Visual Summarization and Stage Analysis of Event Sequence Data. **IEEE Transactions on Visualization and Computer Graphics**, vol. 24, no. 1, p. 56–65, Jan. 2018. DOI 10.1109/TVCG.2017.2745320. Available at: http://ieeexplore.ieee.org/document/8017612/.

HAGHIGHI, A.; VANDERWENDE, L. Exploring content models for multi-document summarization. 2009. **Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09** [...]. Morristown, NJ, USA: Association for Computational Linguistics, 2009. p. 362. DOI 10.3115/1620754.1620807. Available at: http://portal.acm.org/citation.cfm?doid=1620754.1620807.

HALL, A.; WALTON, G. Information overload within the health care system: a literature review. **Health Information & Libraries Journal**, vol. 21, no. 2, p. 102–108, 10 Jun. 2004. DOI 10.1111/j.1471-1842.2004.00506.x. Available at: http://doi.wiley.com/10.1111/j.1471-1842.2004.00506.x.

HARTMANN, N. S.; FONSECA, E.; SHULBY, C. D.; TREVISO, M. V; RODRIGUES, J. S.; ALUÍSIO, S. M. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. 2017. **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology** [...]. Uberlândia, MG, Brazil: Sociedade Brasileira de Computação, 2017. p. 122–131. Available at: https://export.arxiv.org/pdf/1708.06025.

HAYRINEN, K.; SARANTO, K.; NYKANEN, P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. **International Journal of Medical Informatics**, vol. 77, no. 5, p. 291–304, May 2008. DOI 10.1016/j.ijmedinf.2007.09.001. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1386505607001682.

HIRSCH, J. S.; TANENBAUM, J. S.; LIPSKY GORMAN, S.; LIU, C.; SCHMITZ, E.; HASHORVA, D.; ERVITS, A.; VAWDREY, D.; STURM, M.; ELHADAD, N. HARVEST, a longitudinal patient record summarizer. **Journal of the American Medical Informatics Association**, vol. 22, no. 2, p. 263–74, 28 Oct. 2014. DOI 10.1136/amiajnl-2014-002945. Available at: http://jamia.oxfordjournals.org/content/22/2/263.abstract.

HIRSCHTICK, R. E. A piece of my mind. Copy-and-paste. **JAMA : the journal of the American Medical Association**, vol. 295, no. 20, 2006. https://doi.org/10.1001/jama.295.20.2335.

HOLDEN, R. J. Cognitive performance-altering effects of electronic medical records: an application of the human factors paradigm for patient safety. **Cognition, Technology & Work**, vol. 13, no. 1, p. 11–29, 25 Mar. 2011. DOI 10.1007/s10111-010-0141-8. Available at: http://link.springer.com/10.1007/s10111-010-0141-8.

HUANG, Z.; XU, W.; YU, K. Bidirectional LSTM-CRF Models for Sequence Tagging. **CoRR**, 2015. Available at: http://arxiv.org/abs/1508.01991.

HUGGARD, H.; ZHANG, A.; ZHANG, E.; KOH, Y. S. Feature Importance for Biomedical Named Entity Recognition. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [*S. l.*: *s. n.*], 2019. vol. 11919 LNAI, p. 406–417. DOI 10.1007/978-3-030-35288-2_33. Available at: http://link.springer.com/10.1007/978-3-030-35288-2_33.

HUNTER, J.; FREER, Y.; GATT, A.; REITER, E.; SRIPADA, S.; SYKES, C. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. **Artificial Intelligence in Medicine**, vol. 56, no. 3, p. 157–172, Nov. 2012. DOI 10.1016/j.artmed.2012.09.002. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0933365712001170.

INKER, L. A.; ASTOR, B. C.; FOX, C. H.; ISAKOVA, T.; LASH, J. P.; PERALTA, C. A.; KURELLA TAMURA, M.; FELDMAN, H. I. KDOQI US Commentary on the 2012 KDIGO Clinical Practice Guideline for the Evaluation and Management of CKD. **American Journal of Kidney Diseases**, vol. 63, no. 5, p. 713–735, May 2014. DOI 10.1053/j.ajkd.2014.01.416. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0272638614004910.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/DTR 20514. Health informatics - Electronic Health Record – Definition, Scope, and Context. 2004.

JENSEN, P. B.; JENSEN, L. J.; BRUNAK, S. Mining electronic health records: towards better research applications and clinical care. **Nature Reviews Genetics**, vol. 13, no.

6, p. 395–405, 2 Jun. 2012. DOI 10.1038/nrg3208. Available at: http://www.nature.com/articles/nrg3208.

JOHNSON, E.; BAUGHMAN, W. C.; OZSOYOGLU, G. A distributional approach to summarization of radiology reports. Nov. 2015. **2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)** [...]. [*S. l.*]: IEEE, Nov. 2015. p. 973–974. DOI 10.1109/BIBM.2015.7359815. Available at: http://ieeexplore.ieee.org/document/7359815/.

KDIGO. Summary of Recommendation Statements. **Kidney International Supplements**, vol. 3, no. 3, p. 263–265, Nov. 2013. DOI 10.1038/kisup.2013.31. Available at: http://linkinghub.elsevier.com/retrieve/pii/S215717161531145X.

KENEI, J.; OPIYO, T. O. E.; OBOKO, R.; MOSO, J. Clinical Documents Summarization using Text Visualization Technique. **International Journal of Computer and Information Technology**, vol. 07, no. 04, p. 139–156, 2018.

KIM, G.-W.; LEE, D.-H. Personalised health document summarisation exploiting Unified Medical Language System and topic-based clustering for mobile healthcare. **Journal of Information Science**, 9 Aug. 2017. DOI 10.1177/0165551517722983. Available at: http://journals.sagepub.com/doi/10.1177/0165551517722983.

KINOSHITA, J.; EDUARDO, C.; MENEZES, D. De. CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus. 2006. **Fifth International Conference on Language Resources and Evaluation - LREC 2006** [...]. [*S. l.*: *s. n.*], 2006. p. 2190–2193.

KONG, H. J. Managing unstructured big data in healthcare system. **Healthcare Informatics Research**, vol. 25, no. 1, 2019. https://doi.org/10.4258/hir.2019.25.1.1.

KOOPMAN, B.; ZUCCON, G.; BRUZA, P.; SITBON, L.; LAWLEY, M. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. 2012. **Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12** [...]. New York, New York, USA: ACM Press, 2012. p. 2439. DOI 10.1145/2396761.2398661. Available at: http://dl.acm.org/citation.cfm?doid=2396761.2398661.

KVIST, M.; SKEPPSTEDT, M.; VELUPILLAI, S.; DALIANIS, H. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems - Future vision , a physician ' s perspective. 2011. **9th Scandinavian Conference on Health Informatics** [...]. [*S. l.*: *s. n.*], 2011.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 2001. **ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning** [...]. [*S. l.*: *s. n.*], 2001. p. 282–289.

LAN, K.; WANG, D.; FONG, S.; LIU, L.; WONG, K. K. L.; DEY, N. A Survey of Data Mining and Deep Learning in Bioinformatics. **Journal of Medical Systems**, vol. 42, no. 8, p. 139, 28 Aug. 2018. DOI 10.1007/s10916-018-1003-9. Available at: http://link.springer.com/10.1007/s10916-018-1003-9.

LAST, M.; LITVAK, M. Language-independent Techniques for Automated Text Summarization. **Web Intelligence and Security - Advances in Data and Text Mining Techniques for Detecting and Preventing Terrorist Activities on the Web**, vol. 27, p. 207–237, 2010. Available at: http://dblp.uni-trier.de/db/series/natosec/natosec27.html#Last10.

LAXMISAN, A.; MCCOY, A. B.; WRIGHT, A.; SITTIG, D. F. Clinical Summarization Capabilities of Commercially-available and Internally-developed Electronic Health Records. **Applied Clinical Informatics**, vol. 3, no. 1, p. 80–93, 22 Feb. 2012. DOI 10.4338/ACI-2011-11-RA-0066. Available at: http://www.schattauer.de/index.php?id=1214&doi=10.4338/ACI-2011-11-RA-0066.

LEVEY, A. S.; DE JONG, P. E.; CORESH, J.; NAHAS, M. El; ASTOR, B. C.; MATSUSHITA, K.; GANSEVOORT, R. T.; KASISKE, B. L.; ECKARDT, K. The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. **Kidney International**, vol. 80, no. 1, p. 17–28, Jul. 2011. DOI 10.1038/ki.2010.483. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0085253815549247.

LEVY-FIX, G.; ZUCKER, J.; STOJANOVIC, K.; ELHADAD, N. Towards Patient Record Summarization Through Joint Phenotype Learning in HIV Patients. 9 Mar. 2020. Available at: http://arxiv.org/abs/2003.11474.

LI, J.; SUN, A.; HAN, J.; LI, C. A Survey on Deep Learning for Named Entity Recognition. **IEEE Transactions on Knowledge and Data Engineering**, , p. 1–1, 2020. DOI 10.1109/TKDE.2020.2981314. Available at: https://ieeexplore.ieee.org/document/9039685/.

LIANG, J.; TSOU, C.-H.; PODDAR, A. A Novel System for Extractive Clinical Note Summarization using. 2019. **Proceedings of the 2nd Clinical Natural Language Processing Workshop** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. p. 46–54. DOI 10.18653/v1/W19-1906. Available at: http://aclweb.org/anthology/W19-1906.

LIN, C. Y. Rouge: A package for automatic evaluation of summaries. **Proceedings of the workshop on text summarization branches out (WAS 2004)**, no. 1, p. 25–26, 2004. Available at: http://www.aclweb.org/anthology/W04-1013.

LISSAUER, T.; PATERSON, C. M.; SIMONS, A.; BEARD, R. W. Evaluation of computer generated neonatal discharge summaries. **Archives of Disease in Childhood**, vol. 66, no. 4 Spec No, p. 433–436, 1 Apr. 1991. DOI 10.1136/adc.66.4_Spec_No.433. Available at: http://adc.bmj.com/cgi/doi/10.1136/adc.66.4_Spec_No.433.

LITVAK, M.; LAST, M.; KANDEL, A. DegExt: a language-independent keyphrase extractor. **Journal of Ambient Intelligence and Humanized Computing**, vol. 4, no.

3, p. 377–387, 29 Jun. 2013. DOI 10.1007/s12652-012-0109-z. Available at: http://link.springer.com/10.1007/s12652-012-0109-z.

LIU, H.; FRIEDMAN, C. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. **Studies in health technology and informatics**, vol. 107, no. Pt 1, p. 639–43, 2004. DOI 10.3233/978-1-60750-949-3-639. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15360891.

LLORET, E.; PLAZA, L.; AKER, A. The challenging task of summary evaluation: an overview. **Language Resources and Evaluation**, vol. 52, no. 1, p. 101–148, 2 Mar. 2018. DOI 10.1007/s10579-017-9399-2. Available at: http://link.springer.com/10.1007/s10579-017-9399-2.

LONG, J.; YUAN, M. J. A novel clinical decision support algorithm for constructing complete medication histories. **Computer Methods and Programs in Biomedicine**, vol. 145, p. 127–133, Jul. 2017. DOI 10.1016/j.cmpb.2017.04.004. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0169260716308902.

LOPES, F.; TEIXEIRA, C.; GONÇALO OLIVEIRA, H. Contributions to Clinical Named Entity Recognition in Portuguese. 2019. **Proceedings of the 18th BioNLP Workshop and Shared Task** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. p. 223–233. DOI 10.18653/v1/W19-5024. Available at: https://www.aclweb.org/anthology/W19-5024.

LOPES, F.; TEIXEIRA, C.; GONÇALO OLIVEIRA, H. Named Entity Recognition in Portuguese Neurology Text Using CRF. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [*S. l.*: *s. n.*], 2019. vol. 11804 LNAI, p. 336–348. DOI 10.1007/978-3-030-30241-2_29. Available at: http://link.springer.com/10.1007/978-3-030-30241-2_29.

LOPES, F.; TEIXEIRA, C.; GONÇALO OLIVEIRA, H. Comparing Different Methods for Named Entity Recognition in Portuguese Neurology Text. **Journal of Medical Systems**, vol. 44, no. 4, 2020. https://doi.org/10.1007/s10916-020-1542-8.

MACAVANEY, S.; SOTUDEH, S.; COHAN, A.; GOHARIAN, N.; TALATI, I.; FILICE, R. W. Ontology-Aware Clinical Abstractive Summarization. 18 Jul. 2019. **Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval** [...]. New York, NY, USA: ACM, 18 Jul. 2019. p. 1013–1016. DOI 10.1145/3331184.3331319. Available at: https://dl.acm.org/doi/10.1145/3331184.3331319.

MALTA, D. C.; BERNAL, R. T. I.; LIMA, M. G.; ARAÚJO, S. S. C. de; SILVA, M. M. A. da; FREITAS, M. I. de F.; BARROS, M. B. de A. Noncommunicable diseases and the use of health services: analysis of the National Health Survey in Brazil. **Revista de Saúde Pública**, vol. 51, no. suppl 1, p. 1–10, 2017. DOI 10.1590/s1518-8787.2017051000090. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102017000200306&lng=en&tlng=en.

MARINHO, A. W. G. B.; PENHA, A. da P.; SILVA, M. T.; GALVÃO, T. F. Prevalência de doença renal crônica em adultos no Brasil: revisão sistemática da literatura. **Cadernos Saúde Coletiva**, vol. 25, no. 3, p. 379–388, 2017. DOI 10.1590/1414-462x201700030134. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-462X2017000300379&lng=pt&tlng=pt.

MCDONALD, C. J. Protocol-Based Computer Reminders, the Quality of Care and the Non-Perfectibility of Man. **New England Journal of Medicine**, vol. 295, no. 24, p. 1351–1355, 9 Dec. 1976. DOI 10.1056/NEJM197612092952405. Available at: http://www.nejm.org/doi/abs/10.1056/NEJM197612092952405.

MCDONALD, C. J.; CALLAGHAN, F. M.; WEISSMAN, A.; GOODWIN, R. M.; MUNDKUR, M.; KUHN, T. Use of Internist's Free Time by Ambulatory Care Electronic Medical Record Systems. **JAMA Internal Medicine**, vol. 174, no. 11, p. 1860, 1 Nov. 2014. DOI 10.1001/jamainternmed.2014.4506. Available at: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2014.4506.

MIHALCEA, R.; TARAU, P. TextRank: Bringing Order into Texts. 2004. **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing** [...]. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411. Available at: https://www.aclweb.org/anthology/W04-3252.

MILLER, D. Leveraging BERT for Extractive Text Summarization on Lectures. 7 Jun. 2019. Available at: http://arxiv.org/abs/1906.04165.

MISHRA, R.; BIAN, J.; FISZMAN, M.; WEIR, C. R.; JONNALAGADDA, S.; MOSTAFA, J.; DEL FIOL, G. Text summarization in the biomedical domain: A systematic review of recent research. **Journal of Biomedical Informatics**, vol. 52, no. 2, p. 457–467, Dec. 2014. DOI 10.1016/j.jbi.2014.06.009. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1532046414001476.

MOEN, H.; PELTONEN, L.-M.; HEIMONEN, J.; AIROLA, A.; PAHIKKALA, T.; SALAKOSKI, T.; SALANTERÄ, S. Comparison of automatic summarisation methods for clinical free text notes. **Artificial Intelligence in Medicine**, vol. 67, p. 25–37, 2016. DOI 10.1016/j.artmed.2016.01.003. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0933365716000051.

MORADI, M.; GHADIRI, N. Different approaches for identifying important concepts in probabilistic biomedical text summarization. **Artificial Intelligence in Medicine**, vol. 84, p. 101–116, Jan. 2018. DOI 10.1016/j.artmed.2017.11.004. Available at: http://dx.doi.org/10.1016/j.artmed.2017.11.004.

MORADI, M.; SAMWALD, M. Clustering of Deep Contextualized Representations for Summarization of Biomedical Texts. , p. 1–5, 6 Aug. 2019. Available at: http://arxiv.org/abs/1908.02286.

MORID, M. A.; FISZMAN, M.; RAJA, K.; JONNALAGADDA, S. R.; DEL FIOL, G. Classification of clinically useful sentences in clinical evidence resources. **Journal of**

Biomedical Informatics, vol. 60, p. 14–22, 2016. DOI 10.1016/j.jbi.2016.01.003. Available at: http://dx.doi.org/10.1016/j.jbi.2016.01.003.

MUELLER, J.; THYAGARAJAN, A. Siamese recurrent architectures for learning sentence similarity. 2016. **30th AAAI Conference on Artificial Intelligence, AAAI 2016** [...]. Phoenix, Arizona: AAAI Press, 2016. p. 2786–2792.

NAVANEETHAN, S. D.; JOLLY, S. E.; SHARP, J.; JAIN, A.; SCHOLD, J. D.; JR, M. J. S.; JR, J. V. N. Electronic health records: a new tool to combat chronic kidney disease? **Clinical Nephrology**, vol. 79, no. 03, p. 175–183, 1 Mar. 2013. DOI 10.5414/CN107757. Available at: http://www.dustri.com/article_response_page.html?artId=10331&doi=10.5414/CN107757&L=0.

NENKOVA, A.; MCKEOWN, K. Automatic Summarization. **Foundations and Trends® in Information Retrieval**, vol. 5, no. 2, p. 103–233, 2011. DOI 10.1561/1500000015. Available at: http://www.nowpublishers.com/article/Details/INR-015.

NENKOVA, A.; MCKEOWN, K. A SURVEY OF TEXT SUMMARIZATION TECHNIQUES. **Mining Text Data**. [*S. l.*: *s. n.*], 2012. vol. 9781461432, p. 1–522. https://doi.org/10.1007/978-1-4614-3223-4.

NENKOVA, A.; PASSONNEAU, R. Evaluating content selection in summarization: The pyramid method. **Proceedings of HLT-NAACL**, vol. 2004, p. 145–152, 2004. Available at: https://aclanthology.coli.uni-saarland.de/papers/N04-1019/n04-1019.

NG, J.-P.; ABRECHT, V. Better Summarization Evaluation with Word Embeddings for ROUGE. 2015. **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015. p. 1925–1930. DOI 10.18653/v1/D15-1222. Available at: http://aclweb.org/anthology/D15-1222.

NOTHMAN, J.; RINGLAND, N.; RADFORD, W.; MURPHY, T.; CURRAN, J. R. Learning multilingual named entity recognition from Wikipedia. **Artificial Intelligence**, vol. 194, p. 151–175, Jan. 2013. DOI 10.1016/j.artint.2012.03.006. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0004370212000276.

OLEYNIK, M.; NOHAMA, P.; CANCIAN, P. S.; SCHULZ, S. Performance analysis of a POS tagger applied to discharge summaries in portuguese. **Studies in Health Technology and Informatics**, vol. 160, no. PART 1, p. 959–963, 2010. https://doi.org/10.3233/978-1-60750-588-4-959.

OLIVEIRA, L. E. S.; GEBELUCA, C. P.; SILVA, A. M. P.; MORO, C. M. C.; HASAN, S. A.; FARRI, O. A statistics and UMLS-based tool for assisted semantic annotation of Brazilian clinical documents. 2017. **2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)** [...]. [*S. l.*]: IEEE, 2017. p. 1072–1078. DOI 10.1109/BIBM.2017.8217805. Available at: http://ieeexplore.ieee.org/document/8217805/.

OLIVEIRA, L. E. S.; GEBELUCA, C. P.; SILVA, A. M. P.; MORO, C. M. C.; HASAN, S. A.; FARRI, O. A statistics and UMLS-based tool for assisted semantic annotation of Brazilian clinical documents. 2017. **2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)** [...]. [*S. l.*]: IEEE, 2017. p. 1072–1078. DOI 10.1109/BIBM.2017.8217805. Available at: http://ieeexplore.ieee.org/document/8217805/.

OLIVEIRA, L. E. S. e; GUMIEL, Y. B.; DOS SANTOS, A. B. V.; CINTHO, L. M. M.; CARVALHO, D. R.; HASAN, S. A.; MORO, C. M. C. Learning Portuguese Clinical Word Embeddings: A Multi-Specialty and Multi-Institutional Corpus of Clinical Narratives Supporting a Downstream Biomedical Task. **Studies in health technology and informatics**, vol. 264, p. 123–127, 21 Aug. 2019. DOI 10.3233/SHTI190196. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31437898.

OLIVEIRA, LUCAS EMANUEL SILVA HASAN, S. Al; FARRI, O.; MORO, C. M. C. TRANSLATION OF UMLS ONTOLOGIES FROM EUROPEAN PORTUGUESE TO BRAZILIAN PORTUGUESE. 2016. **CBIS 2016 - XV Congresso Brasileiro de Informática em Saúde** [...]. [*S. l.*: *s. n.*], 2016. p. 373–379.

OLIVEIRA, R. de; SRIPADA, S. Adapting SimpleNLG for Brazilian Portuguese realisation. **Proceedings of the 8th International Natural Language Generation Conference**, no. June, p. 93–94, 2014. Available at: http://www.aclweb.org/anthology/W14-4412.

ORONOZ, M.; GOJENOLA, K.; PÉREZ, A.; DE ILARRAZA, A. D.; CASILLAS, A. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. **Journal of Biomedical Informatics**, vol. 56, p. 318–332, Aug. 2015. DOI 10.1016/j.jbi.2015.06.016. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1532046415001264.

PACHECO, E. J. MorphoMap : Mapeamento automático de narrativas clínicas para uma terminologia médica. **Repositório Institucional da Universidade Tecnológica Federal do Paraná**, 2009.

PATEL, P.; DAVEY, D.; PANCHAL, V.; PATHAK, P. Annotation of a Large Clinical Entity Corpus. 2018. **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing** [...]. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 2033–2042. Available at: https://www.aclweb.org/anthology/D18-1228.

PEDERSEN, T.; PAKHOMOV, S. V. S.; PATWARDHAN, S.; CHUTE, C. G. Measures of semantic similarity and relatedness in the biomedical domain. **Journal of Biomedical Informatics**, vol. 40, no. 3, p. 288–299, 2007. DOI 10.1016/j.jbi.2006.06.004. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16875881.

PEREIRA, E. R. S.; PEREIRA, A. de C.; ANDRADE, G. B. de; NAGHETTINI, A. V.; PINTO, F. K. M. S.; BATISTA, S. R.; MARQUES, S. M. Prevalence of chronic renal disease in adults attended by the family health strategy. **Jornal Brasileiro de**

Nefrologia, vol. 38, no. 1, p. 22–30, 2016. DOI 10.5935/0101-2800.20160005. Available at: http://www.gnresearch.org/doi/10.5935/0101-2800.20160005.

PEROTTE, A.; HRIPCSAK, G. Temporal Properties of Diagnosis Code Time Series in Aggregate. **IEEE Journal of Biomedical and Health Informatics**, vol. 17, no. 2, p. 477–483, Mar. 2013. DOI 10.1109/JBHI.2013.2244610. Available at: http://ieeexplore.ieee.org/document/6471160/.

PESQUITA, C.; FARIA, D.; FALCÃO, A. O.; LORD, P.; COUTO, F. M. Semantic Similarity in Biomedical Ontologies. **PLoS Computational Biology**, vol. 5, no. 7, 31 Jul. 2009. DOI 10.1371/journal.pcbi.1000443. Available at: http://dx.plos.org/10.1371/journal.pcbi.1000443.

PETERS, A. C.; OLEYNIK, M.; PACHECO, E. J.; MORO, C. M. C.; SCHULZ, S.; NOHAMA, P. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. **Anais do XII Congresso Brasileiro de Informática em Saúde**, no. October 2010, 2010. https://doi.org/10.13140/RG.2.1.4412.7441.

PITLER, E.; NENKOVA, A. Revisiting readability: A unified framework for predicting text quality. 2008. **EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing** [...]. [*S. l.*: *s. n.*], 2008. p. 186–195. Available at: https://dl.acm.org/citation.cfm?id=1613742.

PIVOVAROV, R. **Electronic Health Record Summarization over Heterogeneous and Irregularly Sampled Clinical Data**. 2016. Columbia University, 2016. Available at: https://academiccommons.columbia.edu/doi/10.7916/D89W0F6V.

PIVOVAROV, R.; ELHADAD, N. Automated methods for the summarization of electronic health records. **Journal of the American Medical Informatics Association**, vol. 22, no. 5, p. 938–947, Sep. 2015. DOI 10.1093/jamia/ocv032. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocv032.

PIVOVAROV, R.; ELHADAD, N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. **Journal of Biomedical Informatics**, vol. 45, no. 3, p. 471–481, 2012. DOI 10.1016/j.jbi.2012.01.002. Available at: http://dx.doi.org/10.1016/j.jbi.2012.01.002.

PLAZA, L.; DÍAZ, A.; GERVÁS, P. A semantic graph-based approach to biomedical summarisation. **Artificial Intelligence in Medicine**, vol. 53, no. 1, p. 1–14, Sep. 2011. DOI 10.1016/j.artmed.2011.06.005. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0933365711000753.

PLUYE, P.; GRAD, R. M.; DUNIKOWSKI, L. G.; STEPHENSON, R. Impact of clinical information-retrieval technology on physicians: A literature review of quantitative, qualitative and mixed methods studies. **International Journal of Medical Informatics**, vol. 74, no. 9, p. 745–768, Sep. 2005. DOI 10.1016/j.ijmedinf.2005.05.004. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1386505605000572.

POH, N.; DE LUSIGNAN, S. Data-modelling and visualisation in chronic kidney disease (CKD): a step towards personalised medicine. **Informatics in primary care**, vol. 19, no. 2, p. 57–63, 2011. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22417815.

QUIMBAYA, A. P.; MÚNERA, A. S.; RIVERA, R. A. G.; RODRÍGUEZ, J. C. D.; VELANDIA, O. M. M.; PEÑA, A. A. G.; LABBÉ, C. Named Entity Recognition over Electronic Health Records Through a Combined Dictionary-based Approach. **Procedia Computer Science**, vol. 100, p. 55–61, 2016. https://doi.org/10.1016/j.procs.2016.09.123.

RADEV, D.; HOVY, E.; MCKEOWN, K. Introduction to the special issue on summarization. **Computational linguistics**, vol. 28, no. 4, p. 399–408, 2002. DOI 10.1016/j.jbi.2011.03.008. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671927.

RAGHAVAN, P.; FOSLER-LUSSIER, E.; ELHADAD, N.; LAI, A. M. Cross-narrative Temporal Ordering of Medical Events. 2014. **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. p. 998–1008. DOI 10.3115/v1/P14-1094. Available at: http://aclweb.org/anthology/P14-1094.

RAMANUJAM, N.; KALIAPPAN, M. An Automatic Multidocument Text Summarization Approach Based on Naïve Bayesian Classifier Using Timestamp Strategy. **The Scientific World Journal**, vol. 2016, p. 1–10, 2016. DOI 10.1155/2016/1784827. Available at: http://www.hindawi.com/journals/tswj/2016/1784827/.

REICHERT, D.; KAUFMAN, D.; BLOXHAM, B.; CHASE, H.; ELHADAD, N. Cognitive analysis of the summarization of longitudinal patient records. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, vol. 2010, p. 667–71, 13 Nov. 2010. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21347062.

RIM, K. MAE2 : Portable Annotation Tool for General Natural Language Use. **Proceedings of the LREC 2016Workshop “12th Joint ACL - ISOWorkshop on Interoperable Semantic Annotation" (ISA-12)**, , p. 75–80, 2016. Available at: https://github.com/keighrim/mae-annotation.

ROBERTS, A.; GAIZAUSKAS, R.; HEPPLE, M.; DEMETRIOU, G.; GUO, Y.; ROBERTS, I.; SETZER, A. Building a semantically annotated corpus of clinical texts. **Journal of Biomedical Informatics**, vol. 42, no. 5, p. 950–966, Oct. 2009. DOI 10.1016/j.jbi.2008.12.013. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1532046409000069.

RODRIGUES, R. C.; DA SILVA, J. R.; DE CASTRO, P. V. Q.; DA SILVA, N. F. F.; DA SILVA SOARES, A. Multilingual transformer ensembles for portuguese natural language tasks. **CEUR Workshop Proceedings**, vol. 2583, p. 27–38, 2020. .

RODRIGUES, R.; COUTO, P.; RODRIGUES, I. IPR: The semantic textual similarity and recognizing textual entailment systems. **CEUR Workshop Proceedings**, vol. 2583, p. 39–47, 2020.

ROGERS, J.; PULESTON, C.; RECTOR, A. The CLEF Chronicle: Patient Histories Derived from Electronic Health Records. 2006. **22nd International Conference on Data Engineering Workshops (ICDEW'06)** [...]. [*S. l.*]: IEEE, 2006. p. x109–x109. DOI 10.1109/ICDEW.2006.144. Available at: http://ieeexplore.ieee.org/document/1623902/.

ROUANE, O.; BELHADEF, H.; BOUAKKAZ, M. Word Embedding-Based Biomedical Text Summarization. **IRICT 2019: Emerging Trends in Intelligent Computing and Informatics**. [*S. l.*]: Springer, Cham, 2020. vol. 1073, p. 288–297. DOI 10.1007/978-3-030-33582-3_28. Available at: http://link.springer.com/10.1007/978-3-030-33582-3_28.

SAMAL, L.; WRIGHT, A.; WONG, B. T.; LINDER, J. A.; BATES, D. W. Leveraging electronic health records to support chronic disease management: the need for temporal data views. **Informatics in primary care**, vol. 19, no. 2, p. 65–74, 2011. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22417816.

SANTOS, J.; TERRA, J.; CONSOLI, B.; VIEIRA, R. Multidomain contextual embeddings for named entity recognition. **CEUR Workshop Proceedings**, vol. 2421, no. Lm, p. 434–441, 2019.

SARKAR, K. Using Domain Knowledge for Text Summarization in Medical Domain. **International Journal of Recent Trends in Engineering (ACEEE)**, vol. 1, no. 1, p. 6, 2009.

SARKAR, K.; NASIPURI, M.; GHOSE, S. Using Machine Learning for Medical Document Summarization. **International Journal of Database Theory and Application International Journal of Database Theory and Application**, vol. 4, no. 1, p. 31–48, 2011.

SARKER, A.; MOLLA, D. Extractive summarization of medical documents using domain knowledge and corpus statistics. **Australasian Medical Journal**, vol. 5, no. 9, p. 478–481, 1 Oct. 2012. DOI 10.4066/AMJ.2012.1361. Available at: http://www.amj.net.au/index.php?journal=AMJ&page=article&op=viewFile&path%5B%5D=1361&path%5B%5D=976.

SARWADNYA, V. V; SONAWANE, S. S. Extractive Summarizer Construction Techniques : A Survey. **International Journal of Scientific Research in Computer Science, Engineering and Information Technology**, vol. 3, no. 3, p. 2058–2066, 2018. Available at: http://ijsrcseit.com/CSEIT183152.

SAVOVA, G. K.; MASANZ, J. J.; OGREN, P. V.; ZHENG, J.; SOHN, S.; KIPPER-SCHULER, K. C.; CHUTE, C. G. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. **Journal of the American Medical Informatics Association**, vol. 17, no. 5, p. 507–513, Sep.

2010. DOI 10.1136/jamia.2009.001560. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.2009.001560.

SCHNEIDER, E. T. R.; SOUZA, J. V. A. de; KNAFOU, J.; OLIVEIRA, L. E. S. e; COPARA, J.; GUMIEL, Y. B.; OLIVEIRA, L. F. A. de; PARAISO, E. C.; TEODORO, D.; MORO, C. **Publicly Available Portuguese Neural Language Model for Clinical Named Entity Recognition**. 2020. (work in progress)

SHICKEL, B.; TIGHE, P. J.; BIHORAC, A.; RASHIDI, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. **IEEE Journal of Biomedical and Health Informatics**, vol. 22, no. 5, p. 1589–1604, 2018. https://doi.org/10.1109/JBHI.2017.2767063.

SHORTLIFFE, E. H.; BARNETT, G. O. Medical Data: Their Acquisition, Storage, and Use. **Medical Informatics: Computer Applications in Health Care**. USA: Addison-Wesley Longman Publishing Co., Inc., 1990. p. 37–69.

SILVA, A. M. P. da. **Processamento de linguagem natural na identificação de critérios de elegibilidade para pesquisa clínica**. 2018. Pontifícia Universidade Católica do Paraná, 2018.

SILVA, S. B.; CAULLIRAUX, H. M.; ARAÚJO, C. A. S.; ROCHA, E. Uma comparação dos custos do transplante renal em relação às diálises no Brasil. **Cadernos de Saúde Pública**, vol. 32, no. 6, p. 1–13, 2016. DOI 10.1590/0102-311x00013515. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2016000605005&lng=pt&tlng=pt.

SONDHI, P.; SUN, J.; TONG, H.; ZHAI, C. SympGraph: a framework for mining clinical notes through symptom relation graphs. 2012. **Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12** [...]. New York, New York, USA: ACM Press, 2012. p. 1167. DOI 10.1145/2339530.2339712. Available at: http://dl.acm.org/citation.cfm?doid=2339530.2339712.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese Named Entity Recognition using BERT-CRF. **arXiv preprint arXiv:1909.10649**, 2019. Available at: http://arxiv.org/abs/1909.10649.

SOUZA, J. V. A. De; OLIVEIRA, L. E. S. E.; GUMIEL, Y. B.; CARVALHO, D. R.; MORO, C. M. C. Incorporating multiple feature groups to a siamese neural network for semantic textual similarity task in portuguese texts. **CEUR Workshop Proceedings**, vol. 2583, p. 58–67, 2020.

SOYSAL, E.; WANG, J.; JIANG, M.; WU, Y.; PAKHOMOV, S.; LIU, H.; XU, H. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. **Journal of the American Medical Informatics Association**, vol. 25, no. 3, p. 331–336, 2018. https://doi.org/10.1093/jamia/ocx132.

STANISLAWEK, T.; WRÓBLEWSKA, A.; WÓJCICKA, A.; ZIEMBICKI, D.; BIECEK, P. Named Entity Recognition - Is There a Glass Ceiling? 2019. **Proceedings of the 23rd**

Conference on Computational Natural Language Learning (CoNLL) [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. p. 624–633. DOI 10.18653/v1/K19-1058. Available at: https://www.aclweb.org/anthology/K19-1058.

STEINBERGER, J.; JEŽEK, K. Using Latent Semantic Analysis in Text Summarization. 2004. **Proc. ISIM** [...]. [*S. l.*: *s. n.*], 2004. p. 93–100.

STYLER, W. F.; BETHARD, S.; FINAN, S.; PALMER, M.; PRADHAN, S.; DE GROEN, P. C.; ERICKSON, B.; MILLER, T.; LIN, C.; SAVOVA, G.; PUSTEJOVSKY, J. Temporal Annotation in the Clinical Domain. **Transactions of the Association for Computational Linguistics**, vol. 2, no. 1, p. 143–154, 2014. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29082229.

SULTANUM, N.; BRUDNO, M.; WIGDOR, D.; CHEVALIER, F. More Text Please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. 2018-April., 2018. **Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18** [...]. New York, New York, USA: ACM Press, 2018. vol. 2018-April, p. 1–13. DOI 10.1145/3173574.3173996. Available at: http://dl.acm.org/citation.cfm?doid=3173574.3173996.

SUN, S.; NENKOVA, A. The Feasibility of Embedding Based Automatic Evaluation for Single Document Summarization. 2019. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. p. 1216–1221. DOI 10.18653/v1/D19-1116. Available at: https://www.aclweb.org/anthology/D19-1116.

SUN, W.; RUMSHISKY, A.; UZUNER, O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. **Journal of the American Medical Informatics Association**, vol. 20, no. 5, p. 806–813, Sep. 2013. DOI 10.1136/amiajnl-2013-001628. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-001628.

SUN, W.; RUMSHISKY, A.; UZUNER, O. Temporal reasoning over clinical text: the state of the art. **Journal of the American Medical Informatics Association**, vol. 20, no. 5, p. 814–819, Sep. 2013. DOI 10.1136/amiajnl-2013-001760. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-001760.

THIESSARD, F.; MOUGIN, F.; DIALLO, G.; JOUHET, V.; COSSIN, S.; GARCELON, N.; CAMPILLO, B.; JOUINI, W.; GROSJEAN, J.; MASSARI, P.; GRIFFON, N.; DUPUCH, M.; TAYALATI, F.; DUGAS, E.; BALVET, A.; GRABAR, N.; PEREIRA, S.; FRANDJI, B.; DARMONI, S.; CUGGIA, M. RAVEL: Retrieval and visualization in ELectronic Health Records. **Studies in Health Technology and Informatics**, vol. 180, p. 194–198, 2012. DOI 10.3233/978-1-61499-101-4-194. Available at: http://ebooks.iospress.nl/publication/21731.

THOMAS, A.; SANGEETHA, S. **Deep Learning Architectures for Named Entity Recognition: A Survey**. [*S. l.*]: Springer Singapore, 2020. vol. 1082, . DOI

10.1007/978-981-15-1081-6_18. Available at: http://dx.doi.org/10.1007/978-981-15-1081-6_18.

THIYAGU, T. M.; MANJULA, D.; SHRIDHAR, S. Named Entity Recognition in Biomedical Domain: A Survey. **International Journal of Computer Applications**, vol. 181, no. 41, p. 30–37, 15 Feb. 2019. DOI 10.5120/ijca2019918469. Available at: http://www.ijcaonline.org/archives/volume181/number41/thiyagu-2019-ijca-918469.pdf.

TINETTI, M. E.; FRIED, T. R.; BOYD, C. M. Designing Health Care for the Most Common Chronic Condition—Multimorbidity. **JAMA**, vol. 307, no. 23, p. 2493–2494, 20 Jun. 2012. DOI 10.1001/jama.2012.5265. Available at: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2012.5265.

TULKENS, S.; SUSTER, S.; DAELEMANS, W. Using Distributed Representations to Disambiguate Biomedical and Clinical Concepts. 2016. **Proceedings of the 15th Workshop on Biomedical Natural Language Processing** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. p. 77–82. DOI 10.18653/v1/W16-2910. Available at: http://arxiv.org/abs/1608.05605.

UNERTL, K. M.; WEINGER, M. B.; JOHNSON, K. B.; LORENZI, N. M. Describing and Modeling Workflow and Information Flow in Chronic Disease Care. **Journal of the American Medical Informatics Association**, vol. 16, no. 6, p. 826–836, 1 Nov. 2009. DOI 10.1197/jamia.M3000. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M3000.

VAN VLECK, T. T.; ELHADAD, N. Corpus-Based Problem Selection for EHR Note Summarization. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, vol. 2010, p. 817–21, 13 Nov. 2010. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21347092.

VAN VLECK, T. T.; STEIN, D. M.; STETSON, P. D.; JOHNSON, S. B. Assessing data relevance for automated generation of a clinical summary. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, vol. 2007, p. 761–5, 11 Oct. 2007. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18693939.

VANDERWENDE, L.; SUZUKI, H.; BROCKETT, C.; NENKOVA, A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. **Information Processing & Management**, vol. 43, no. 6, p. 1606–1618, Nov. 2007. DOI 10.1016/j.ipm.2007.01.023. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0306457307000507.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention Is All You Need. 2017. **31st Conference on Neural Information Processing Systems (NIPS 2017)** [...]. Long Beach, CA, USA: [*s. n.*], 2017. Available at: http://arxiv.org/abs/1706.03762.

VIANI, N.; SACCHI, L.; TIBOLLO, V.; NAPOLITANO, C.; PRIORI, S. G.; BELLAZZI, R. Clinical timelines development from textual medical reports in Italian. Sep. 2017. **2017 IEEE 3rd International Forum on Research and Technologies for Society**

and Industry (RTSI) [...]. [*S. l.*]: IEEE, Sep. 2017. p. 1–5. DOI 10.1109/RTSI.2017.8065897. Available at: http://ieeexplore.ieee.org/document/8065897/.

WANG, Y.; WANG, L.; RASTEGAR-MOJARAD, M.; MOON, S.; SHEN, F.; AFZAL, N.; LIU, S.; ZENG, Y.; MEHRABI, S.; SOHN, S.; LIU, H. Clinical information extraction applications: A literature review. **Journal of Biomedical Informatics**, vol. 77, p. 34–49, Jan. 2018. DOI 10.1016/j.jbi.2017.11.011. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1532046417302563.

WEBER, G. M.; KOHANE, I. S. Extracting Physician Group Intelligence from Electronic Health Records to Support Evidence Based Medicine. **PLoS ONE**, vol. 8, no. 5, p. e64933, 29 May 2013. DOI 10.1371/journal.pone.0064933. Available at: http://dx.plos.org/10.1371/journal.pone.0064933.

WENG, W.-H.; CHUNG, Y.-A.; TONG, S. Clinical Text Summarization with Syntax-Based Negation and Semantic Concept Identification. no. 2, 29 Feb. 2020. Available at: http://arxiv.org/abs/2003.00353.

WERE, M. C.; SHEN, C.; BWANA, M.; EMENYONU, N.; MUSINGUZI, N.; NKUYAHAGA, F.; KEMBABAZI, A.; TIERNEY, W. M. Creation and Evaluation of EMR-based Paper Clinical Summaries to Support HIV-Care in Uganda, Africa. **International Journal of Medical Informatics**, vol. 79, no. 2, p. 1–13, 2010. https://doi.org/10.1016/j.ijmedinf.2009.11.006.Creation.

WORLD HEALTH ORGANIZATION (WHO). From burden to “best buys”: Reducing the economic impact of NCDs in low- and middle-income countries. Executive summary 2011. Geneva, 2011. Available at: http://www.who.int/nmh/publications/best_buys_summary/en/.

WORLD HEALTH ORGANIZATION (WHO). World health statistics 2018: monitoring health for the SDGs, sustainable development goals. Geneva, 2018. Available at: http://www.who.int/healthinfo/en/.

WORLD HEALTH ORGANIZATION (WHO). Noncommunicable Diseases Progress Monitor, 2017. Geneva, 2017. Available at: http://www.who.int/nmh/publications/ncd-progress-monitor-2017/en/.

WORLD HEALTH ORGANIZATION (WHO). Global status report on noncommunicable disease 2014. Geneva, 2014. Available at: http://www.who.int/nmh/publications/ncd-status-report-2014/en/.

WORLD HEALTH ORGANIZATION (WHO). World health statistics 2018: monitoring health for the SDGs, sustainable development goals. Geneva, 2018. Available at: http://www.who.int/healthinfo/en/.

WORLD HEALTH ORGANIZATION (WHO). Noncommunicable Diseases Progress Monitor, 2017. Geneva, 2017. Available at: http://www.who.int/nmh/publications/ncd-progress-monitor-2017/en/.

WU, Y.; JIANG, M.; XU, J.; ZHI, D.; XU, H. Clinical Named Entity Recognition Using Deep Learning Models. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, vol. 2017, p. 1812–1819, 2017.

WU, Y.; XU, J.; ZHANG, Y.; XU, H. Clinical Abbreviation Disambiguation Using Neural Word Embeddings. **Proceedings of BioNLP 15**, no. BioNLP, p. 171–176, 2015.

XIA, F.; YETISGEN-YILDIZ, M. Clinical Corpus Annotation: Challenges and Strategies. 2012. **Proceedings of Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) of the International Conference on Language Resources and Evaluation (LREC)** [...]. Istanbul: [*s. n.*], 2012. Available at: http://faculty.washington.edu/melihay/publications/LREC_BioTxtM_2012.pdf.

YADAV, P.; STEINBACH, M.; KUMAR, V.; SIMON, G. Mining Electronic Health Records (EHRs): A Survey. **ACM Computing Surveys**, vol. 50, no. 6, p. 1–40, 3 Jan. 2018. DOI 10.1145/3127881. Available at: http://dl.acm.org/citation.cfm?doid=3161158.3127881.

YADAV, V.; BETHARD, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. 2018. **Proceedings ofthe 27th International Conference on Computational Linguistics** [...]. Santa Fe, New Mexico, USA: [*s. n.*], 2018. p. 2145–2158.

YAO, J.; WAN, X.; XIAO, J. Recent advances in document summarization. **Knowledge and Information Systems**, vol. 53, no. 2, p. 297–336, 28 Nov. 2017. DOI 10.1007/s10115-017-1042-4. Available at: http://link.springer.com/10.1007/s10115-017-1042-4.

YOO, I.; ALAFAIREET, P.; MARINOV, M.; PENA-HERNANDEZ, K.; GOPIDI, R.; CHANG, J.-F.; HUA, L. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. **Journal of Medical Systems**, vol. 36, no. 4, p. 2431–2448, 3 Aug. 2012. DOI 10.1007/s10916-011-9710-5. Available at: http://link.springer.com/10.1007/s10916-011-9710-5.

YU, Z.; COHEN, T.; BERNSTAM, E. V; JOHNSON, T. R.; WALLACE, B. C. Retrofitting Word Vectors of MeSH Terms to Improve Semantic Similarity Measures. **Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis**, , p. 43–51, 2016. Available at: https://aclweb.org/anthology/W/W16/W16-6106.pdf.

ZAMORA, M.; GAVALDA, R. Interpretable Patient Trajectories from Temporally Annotated Health Records. 2019-June., Jun. 2019. **2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)** [...]. [*S. l.*]: IEEE, Jun. 2019. vol. 2019-June, p. 547–550. DOI 10.1109/CBMS.2019.00112. Available at: https://ieeexplore.ieee.org/document/8787507/.

ZHANG, R.; PAKHOMOV, S. V. S.; ARSONIADIS, E. G.; LEE, J. T.; WANG, Y.; MELTON, G. B. Detecting clinically relevant new information in clinical notes across specialties and settings. **BMC Medical Informatics and Decision Making**, vol. 17, no. S2, p. 68, 5 Jul. 2017. DOI 10.1186/s12911-017-0464-y. Available at: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0464-y.

ZHANG, Y.; MERCK, D.; TSAI, E. B.; MANNING, C. D.; LANGLOTZ, C. P. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. 6 Nov. 2019. Available at: http://arxiv.org/abs/1911.02541.

ZHOU, Q.; WEI, F.; ZHOU, M. At Which Level Should We Extract? An Empirical Study on Extractive Document Summarization. 6 Apr. 2020. Available at: http://arxiv.org/abs/2004.02664.