





Knowledge Graphs Extracted from Medical Appointment Transcriptions: Results Generating Triples Relying on LLMs

Rafael Roque de Souza²^a, Thiago Luna Pinheiro²^b,
Julio Cesar Barbour Oliveira²^c and Julio Cesar dos Reis¹^d

¹*Precision Data, São Paulo, Brazil*

²*Institute of Computing, University of Campinas, Campinas, Brazil*

Keywords: Knowledge Graphs, RDF Triple Generation, eHealth, Telemedicine, Clinical Appointments, LLMs.

Abstract: Knowledge Graphs (KGs) represent computer-interpretable interactions between real-world entities. This can be valuable for representing medical data semantically. We address the challenge of automatically transforming transcribed medical conversations (clinical dialogues) into RDF triples to structure clinical information. In this article, we design and develop a software tool that simplifies clinical documentation. Our solution explores advanced techniques, such as the Fine-tuned GPT-NeoX 20B model, to extract and summarize crucial information from clinical dialogues. We designed the solution's architecture, supported by technologies such as Docker and MongoDB, to be durable and scalable. We achieve accurate medical entity detection from Portuguese-language textual data and identify semantic relationships in interactions between doctors and patients. By applying advanced Natural Language Processing techniques and Large Language Models (LLMs), our results improve the accuracy and relevance of RDF triples generated from clinical textual data.

1 INTRODUCTION


In today's medical landscape, patient records are a key source of information. They encompass diagnoses, medical histories, treatments, and other pertinent information. They are crucial to ensuring patient care quality and continuity. Many of these records are in unstructured formats, such as handwritten notes or transcripts of dialogues. The inherent complexity of medical terminology (Kormilitzin et al., 2021) and the lack of standardization in data formats make the analysis and interpretation of these records challenging (Wu et al., 2020; Honnibal and Johnson, 2015). This avoids the adequate use and analysis of healthcare data in the ecosystem.


Healthcare professionals, such as doctors, nurses, and specialists, constantly collect and update patient health information during appointments. Each interaction represents a chance to collect vital data, from symptoms to test results. Assimilating, interpreting, and synthesizing this information is crucial to deter-


mining the next steps in treatment (Velupillai et al., 2018).


In this context, healthcare professionals demand further help from technologies to facilitate and augment their experience. Given the growing amount of information and the need for standardization, professionals seek digital software tools that help them organize, structure, and visualize data. The advanced digitization of medical records presents challenges in extracting information from unstructured texts. Natural Processing Language (NLP) has emerged as a promising solution, but its application in clinical contexts still presents limitations, especially when identifying complex constructions (Sarzynska-Wawer et al., 2021).

Exploring technologies like NLP can revolutionize clinical data management. However, the path to its effective implementation presents open research challenges. The search for a solution that harmonizes efficiency, accuracy, and accessibility persists. The application of NLP in analyzing Electronic Health Records Electronic Health Record (EHR) in unstructured text formats opens the door to quantifying outcomes that traditionally require detailed abstraction of the records. To this end, Knowledge Graphs Knowledge Graphs (KGs) can be valuable to structure data

^a  <https://orcid.org/0000-0003-1492-5816>

^b  <https://orcid.org/0009-0000-5548-0150>

^c  <https://orcid.org/0000-0002-9990-9016>

^d  <https://orcid.org/0000-0002-9545-2098>

semantically (Kamdar and Dumontier, 2015; Kanza and Frey, 2019; Ruan et al., 2019). A KGs systematizes data resources and their interrelationships. The Resource Description Framework (RDF) serves as a standard for describing this semantically enriched data (Candan et al., 2001; Rossanez and dos Reis, 2019).

In this article, we address how structuring clinical data by combining NLP with KG to transform clinical dialogues into semantic representation via RDF triples. Our proposed methodology evaluates the clinical relevance of input textual data (transcribed from audio records in telemedicine). We explore advanced techniques for clinical data extraction and summarization, such as the Fine-tuned Generative Pre-trained Transformer (GPT)-NeoX 20B model. At the heart of this innovation, we originally designed and developed a software solution that simplifies clinical documentation and enhances medical decision-making.

Our experimental evaluation assesses automatic clinical text classification in identifying relevant clinical texts from the overall transcriptions. Our solution explored LLMs and few-shot learning for this purpose. In addition, we present our results of RDF triple extraction from textual data (relevant clinical texts). We found relevant findings exploring few-shot prompting for identifying RDF triples.

The remainder of this article is organized as follows: Section 2 introduces underlying concepts and presents the related work. Section 3 details our designed methodology. Section 4 presents key aspects of our original developed software tool for clinical data documentation. Section 5 describes our experimental evaluation and presents the achieved results, which are discussed in Section 6. Section 7 summarizes our findings and points out directions for future research.

2 BACKGROUND

KG structured human knowledge modeling the relationships between real-world entities (Ehrlinger and Wöß, 2016). They use the RDF triple representation for KG model. Triples, made up of subject, predicate, and object, constitute the fundamental structure of KG. This formal computational representation is essential for describing and understanding information about diseases. In the context of clinical data, proper data representation and integration is crucial. It allows healthcare practitioners and researchers to visualize interrelationships between concepts and findings (Auer et al., 2007), correlating their research with others. By observing these relationships, new hypothe-

ses can be formulated, advancing domain knowledge (Rossanez et al., 2020).

In the biomedical field, KG have been gaining prominence. Recent initiatives propose innovative approaches for classification and search strategies, from user interaction to machine learning. For example, studies have converted neuroscience information into RDF format (Lam et al., 2007), whereas others have developed frameworks that integrate information from multiple domains (Rossanez et al., 2020).

In medicine, clinical transcriptions play a crucial role in documenting clinical information. The digital revolution has intensified this relevance, converting clinical dialogues into structured data. This transformation enhances evidence-based decisions and improves the continuity of patient care. However, the medical language, full of jargon and specific terminology, poses challenges to the analysis of these transcripts (Exner and Nugues, 2012). Large Language Models (LLM) have emerged as a response to these challenges, improving natural language analysis. LLM have been essential in creating KG in the biomedical field, combining efficiency and precision (Lam et al., 2007; Exner and Nugues, 2012; Manning et al., 2014).

The extraction of RDF triples from texts has become a central issue. In existing studies, techniques such as Semantic Role Labeling (SRL) are applied to map entities and determine (Exner and Nugues, 2012) relationships. However, creating KG from scientific literature poses challenges. The literature has a particular and diverse writing style characterized by long sentences, abbreviations, and technical terms. NLP tools must be adequately trained for this specific lexicon. In this sense, the automatic generation of KG from scientific literature proves challenging.

Building a KG for all diseases is challenging. For this reason, DEKGB (Sheng et al., 2019) proposed an efficient and extensible framework to build KG for specific diseases based on doctors' knowledge. They described the process by extending an existing health KG to include a new disease.

In this work, we originally explore the potential of LLMs in generating RDF triples from medical consultation transcripts. We present results that highlight the effectiveness of our approach and establish a novel findings for the construction of KG in the medical domain.

3 METHODOLOGY

This section describes the conceptual methodology, as illustrated in Figure 1. We detail the conducted

research into developing our AIRDoc system. This research comprised several stages, from data collection and processing to implementing advanced NLP techniques. We employed state-of-the-art computational models like Bidirectional Encoder Representations from Transformers (BERT) and Fine-tuned GPT-NeoX to refine our analysis.

In this research, we obtained results and insights that led us to develop the AIRdoc system (cf. Section 4). This aims to transform medical transcriptions into organized and intuitive KGs. AIRdoc's main strength is its ability to improve accuracy in analyzing medical dialogues. It is positioned as a relevant software tool for healthcare professionals, academics, and researchers, supporting decision-making and enhancing excellence in medical care.

The following presents the methodology used to build KG from medical consultation transcripts.

3.1 Stage 1: Video/Audio Data Acquisition

We propose an innovative approach to transform medical transcriptions into KG using the representation of RDF triples by exploring video and audio input in the solution. Based on this data acquisition, we use advanced NLP techniques and machine learning models focused on accurately detecting entities and identifying semantic relationships in interactions between doctor and patient. We then represent these relationships in a KG, providing a hierarchical and organized view of clinical information.

3.2 Stage 2: Speech Recognition

In the digital age, converting speech into text has become essential for various applications, from virtual assistants to medical transcriptions. In this stage, we prioritize the implementation of advanced Speech-to-Text tools. We calibrate these technological tools to ensure the highest accuracy in audio transcription, especially in medical contexts where clarity and precision are crucial. These technologies have made it possible to efficiently capture verbal interactions and convert them into textual records, ready for analysis and storage.

3.3 Stage 3: Clinical Text Classification

Analyzing medical dialogues requires a meticulous approach due to the complexity and specificity of medical terminology. Therefore, at this stage, we explore and experiment with models to help us classify clinical texts. On this basis, we use the BERT model

(Devlin et al., 2018), one of the most advanced NLP architectures. We train BERT to identify and classify entities in dialogues, such as symptoms, medications, and diagnoses. With this model, we extracted valuable information from the clinical conversations, enhancing our understanding of medical interactions.

3.4 Stage 4: Data Extraction and Structuring

In this stage, we focus on extracting relevant information from the data collected. The extraction process transforms raw transcripts into structured information, such as medications, symptoms, diseases, and summarizations.

3.4.1 NER

In the Named Entity Recognition (NER) technique, extracting and categorizing specific information in medical transcriptions is fundamental. This technique identifies and categorizes entities in texts, such as names, places, and temporal expressions. In the medical field, NER highlights terms such as medicines, diseases, and medical procedures (Neumann et al., 2019).

In the medical context, the importance of NER is amplified due to the complexity and specificity of the terminology used. The correct identification of medical terms, considering Portuguese's linguistic nuances, is crucial to ensure the accuracy and relevance of the information extracted. To achieve this purpose, we used pre-trained models in the Portuguese language, refining them with specific medical data. This provides greater sensitivity to the clinical context. With the help of NER, we extracted crucial information from the transcripts, such as symptoms mentioned by the patients, prescribed medications, and medical histories. This information forms the basis for subsequent analysis and drawing up KG, essential for a comprehensive understanding of doctor-patient interactions.

3.4.2 Summarization

The summarization technique aims to condense information from long texts, keeping only the most relevant content. In our context, this technique becomes fundamental due to the intrinsic complexity of clinical texts. The aim is to uncomplicate medical terminology, facilitate the identification and extraction of relationships between entities, optimize data processing, and, at the same time, preserve the informative core of the text.

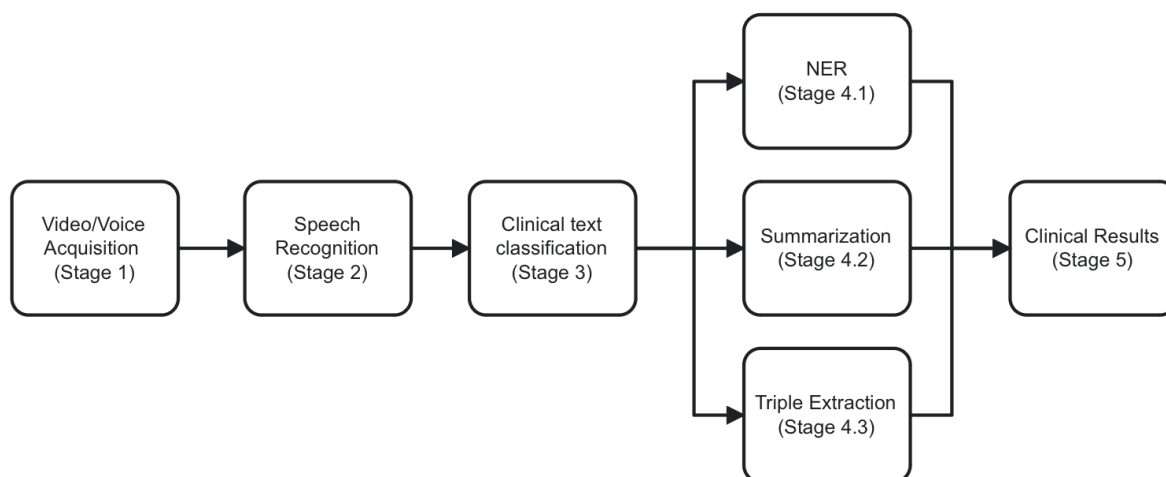


Figure 1: Developed Methodology.

The nature of the input provided to our models significantly can impact the triples’ quality. Our experimental study proposes two main approaches to feeding the models: the first directly uses the content of the MTS-Dialog dataset¹, which transcribes dialogues between doctors and patients. The second processes and summarizes this input, eliminating colloquialisms, redundancies, and other typical speech features, making the text more concise. The latter approach arose from the idea that summarization could improve triple extraction from unstructured text sentences.

3.4.3 RDF Triple Extraction

The task of extracting semantic relationships from texts is complex and requires precision. Before proceeding with the triple extraction, the texts underwent meticulous pre-processing to ensure their quality and uniformity. We use advanced models, such as mREBEL and GPT-NeoX Fine-tuned, specifically for this purpose. These models were trained to identify and extract semantic relationships from natural language texts, culminating in generating RDF triples. These triples offer structured representations of the information, simplifying data integration, queries, and analysis.

3.5 Step 5: Clinical Results

The clinical outcome stage is fundamental in analyzing medical transcripts, focusing on evaluating and interpreting generated data in the clinical context. After extraction and processing, the RDF triples are thoroughly analyzed to discern their clinical significance,

¹<https://github.com/abachaa/MTS-Dialog>

relating identified entities, such as symptoms and diagnoses, to established clinical patterns. The accuracy and relevance of the information extracted is vital; to ensure its reliability, the results are validated by physician experts, who identify and correct possible inconsistencies.

4 AIRDoc: AN AI-AUGMENTED SOFTWARE TOOL FOR CLINICAL DOCUMENTATION

Architeturual Aspects. The AIRdoc software tool was designed based on a methodology that employs a sophisticated conceptual architecture to improve interaction in telemedicine and the EHR. This architecture integrates advances in NLP, database management, KGs, and state-of-the-art language models. This might satisfy the intricate demands of the medical domain. Figure 2 presents our proposed architecture. This figure presents the front end, delineated by dashed blue lines; the back-end modules are enclosed within dashed red lines.

Frontend. In the context of our research into converting medical transcripts into KGs, AIRdoc’s frontend interface, as depicted in Figure 3, stands out as a crucial element. In our proposed design, the defined interfaces act as a visual window for healthcare professionals interact with patients. In addition, we devised interfaces for facilitating exploration and interaction with the generated KGs. The design strives for simplicity and minimalist aesthetics, prioritizing clarity in the presentation of data. In the solution for KG interaction, elements in the graph, such as nodes and connections, have different colors, making it more in-

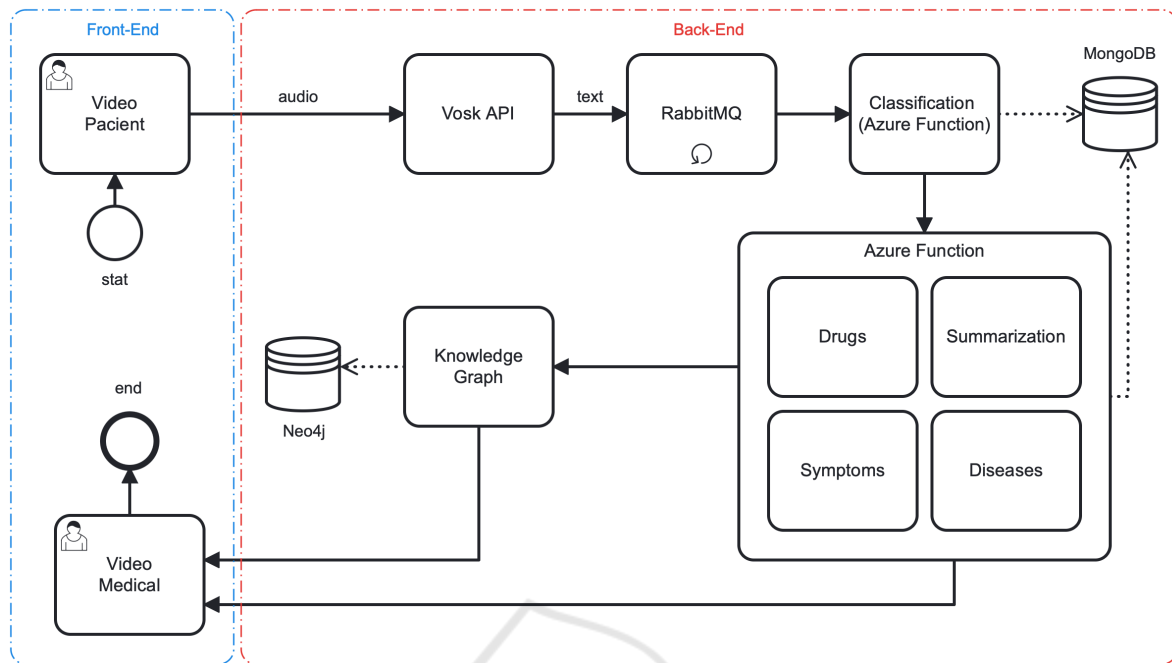


Figure 2: AIRDoc’s Software Architecture for Interactive Construction of Knowledge Graphs.

tuitive to identify and understand the various entities and their interrelationships from the suggested RDF triples. This enhances the user experience and aligns with AIRdoc’s aim of making medical information more intuitive and easily accessible.

Backend. The conceptual architecture of the AIRdoc application has been designed, encompassing everything from creating video rooms to advanced voice capture and real-time transcription techniques. It is based on database systems such as MongoDB² and Neo4j³, ensuring efficiency and robustness in data management. In our implementation, we use Docker containers, integrated with RabbitMQ⁴, an advanced messaging system. This shows AIRdoc’s commitment to providing a cohesive, scalable, high-performance software environment.

Although AIRdoc’s development began with Java and Spark, there has been a migration to Python and FastAPI, reflecting an ongoing drive for greater efficiency and adaptability. This change has enhanced integration with external platforms, such as NLPcloud⁵ and the Vosk⁶ API for speech recognition, expanding its NLP capabilities. Features such as NER, classification and summarization of dialogues, and the generation of RDF semantic triples attests to AIRdoc’s

versatility.

Incorporating the KG generator module and the connection to the Neo4j graph database for storage purposes underlines AIRdoc’s innovative approach. These components provide a detailed visual representation of the interactions between data and ensure a comprehensive and integrated analysis. In our development decisions, we explored the Azure Function Serverless environment, which offers serverless and highly scalable execution based on the introduction of activity logs.

5 EXPERIMENTAL EVALUATION

5.1 Experimental Protocol

Text generation models are machine learning systems that generate coherent and contextually relevant text sequences. Recent developments in NLP have established these models as essential tools in tasks involving language understanding and generation.

Model Selection. We began our experimentation by carefully evaluating available state-of-the-art machine learning models. After careful analysis, we chose *Fine-tuned GPT-NeoX* for assessing the classification task (accessed via the *endpoint* of *Text Generation*⁷ on the NLPcloud platform). For the triple ex-

²www.mongodb.com

³<https://neo4j.com>

⁴www.rabbitmq.com

⁵<https://nlpcloud.com>

⁶<https://alphacephei.com/vosk/>

⁷<https://docs.nlpcloud.com/#generation>

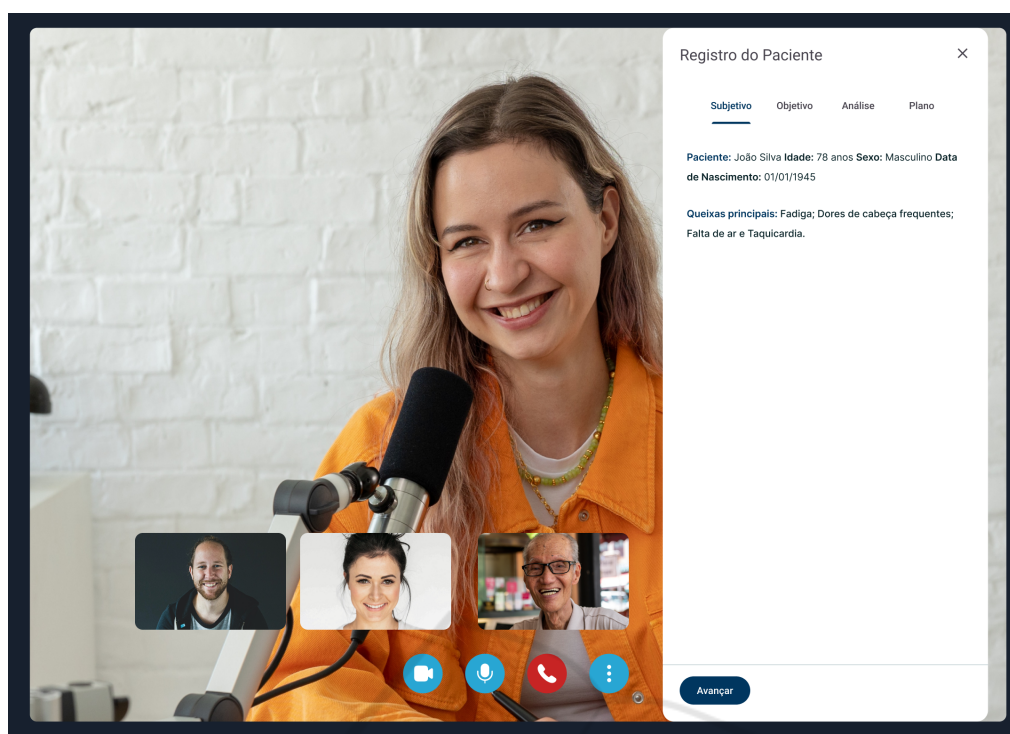


Figure 3: AIRDoc's Videocall interface for clinical documentation.

traction, we compared the effectiveness of *Fine-tuned GPT-NeoX* considering also the mREBEL⁸ and the BLOOM⁹ model, both from the Hugging Face platform¹⁰. Our decision was based on the effectiveness, flexibility, and accessibility of the models.

Definition of Prompts. We clearly defined how to direct the models for the generation task. This targeting, done through a *prompt*, is essential. We adopted three strategies to create these *prompts*:

1. **Few-shot prompting.** We provided the model with examples of inputs and their expected outputs. In our context, we gave fifteen clinical sentences and the corresponding triples extracted from each sentence. These examples are visually presented in Figure 4. Some cases show the object acting as the subject, and in others, the predicate is inferred from the context of the sentence.
2. **Simple Instruction.** We provided the model with clear instructions using natural language. We asked it to extract RDF semantic triples and present the results in JSON format;
3. **Combination of Both.** We combined the two previous strategies, starting with a natural instruc-

tion followed by examples in the style of *few-shot learning*.

Parameters. After defining the prompts, we set the generation parameters. Two essential parameters are *temperature* and *top - p*. We set *temperature* to control the randomness of the output: 0 for Fine-tuned GPT-NeoX and 0.1 for BLOOM. We aimed to obtain consistent outputs. For *top - p*, we varied its values between 0.5 and 1 for GPT and between 0.1, 0.5, and 0.9 for BLOOM.

Execution and Evaluation. We ran the models and compared the outcome generated with the expected results to assess the effectiveness of the models and the chosen approach.

Interaction. Based on the results, we reviewed the setup and made adjustments as necessary. We repeated the process until we achieved the desired results.

Metrics. We use four main metrics to evaluate effectiveness: accuracy, *recall*, precision and *F1-score*. Accuracy gives us an overview of the model's output quality. The *recall* indicates the proportion of positive ("clinical") cases correctly classified concerning the total predicted. Accuracy shows the proportion of patients correctly classified as positive. The *F1-score* represents a harmonic mean between accuracy and *recall*, balancing both metrics.

⁸<https://huggingface.co/Babelscape/mrebel-large>

⁹<https://huggingface.co/bigscience/bloom>

¹⁰<https://huggingface.co>

<p>Input: O paciente disse que estava sentindo febre e dores no corpo.</p> <p>Output: [{"sujeito":"paciente","predicado":"sintoma","objeto":"febre"}, {"sujeito":"paciente","predicado":"sintoma","objeto":"dores no corpo"}]</p>
<p>Input: O médico e o paciente estão conversando sobre uma alergia do paciente à aspirina com revestimento entérico, que é conhecida como Ciprofloxacina.</p> <p>Output: [{"sujeito":"paciente","predicado":"alergia","objeto":"Ciprofloxacina"}]</p>
<p>Input: O paciente fuma 3 maços por semana.</p> <p>Output: [{"sujeito":"paciente","predicado":"hábito","objeto":"fumar"}, {"sujeito":"fumar","predicado":"frequência","objeto":"3 maços por semana"}]</p>
<p>Input: O paciente informou o médico de que os seus parentes tiveram doença arterial coronariana e pressão arterial elevada, mas não cancro.</p> <p>Output: [{"sujeito":"paciente","predicado":"histórico familiar","objeto":"doença arterial coronariana"}, {"sujeito":"paciente","predicado":"histórico familiar","objeto":"pressão arterial elevada"}]</p>
<p>Input: O paciente revela que quando toma Clonidine, desenvolve uma erupção cutânea forte, e quando toma Medifast, fica muito cansado.</p> <p>Output: [{"sujeito":"paciente","predicado":"toma","objeto":"Clonidine"}, {"sujeito":"Clonidine","predicado":"efeito colateral","objeto":"erupção cutânea forte"}, {"sujeito":"paciente","predicado":"toma","objeto":"Medifast"}, {"sujeito":"Medifast","predicado":"efeito colateral","objeto":"cansaço"}]</p>

Figure 4: Examples of *few-shot prompting* where objects can act as subjects or predicates inferred from context.

$$Ac = \frac{VP + VN}{P + N} = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

$$Recall = \frac{VP}{P} = \frac{VP}{VP + FN} \quad (2)$$

$$Prec = \frac{VP}{VP + FP} \quad (3)$$

$$F_1 = \frac{2 * Prec * Recall}{Prec + Recall} \quad (4)$$

The equations from 1 to 4 detail the computation of the metrics. Equation 1 describes Accuracy, followed by *Recall* (Equation 2), *Precision* (Equation 3) and *F1-score* (Equation 4).

5.2 Doctor-Patient Dialogues

We used the MTS-Dialog dataset¹¹, translated to Brazilian Portuguese, as the basis for our experiments. This is composed of dialogues between doctors and patients. This dataset exposed the challenges associated with the absence of semantic information in clinical conversations.

5.3 Results of the Clinical Text Classification

Initially, we analyzed the results from the classification stage of 200 doctor-patient dialogues mentioned. We considered small dialogues with a maximum of

¹¹<https://github.com/abachaa/MTS-Dialog>

400 characters. In our evaluation, we assessed both their original and summarized versions. We evaluated the classification model's performance by comparing its predictions with the classifications in our previously established Gold Standard dataset.

Table 1 shows the classification of the 200 dialogues in their original version, without summarization, compared to the expected result (Gold Standard). It classifies the texts into "clinical" and "non-clinical", showing the totals for each category. The analysis shows that the model correctly classified 160 dialogues as "clinical" and 17 as "non-clinical". However, a discrepancy was identified in 23 dialogues, with 9 cases of false positives and 14 false negatives. We defined the Positive class for "clinical" texts and the Negative class for "non-clinical" texts.

Table 1: We compared the gold standard classification with the one predicted by our model in 200 dialogues between doctors and patients. We carried out this classification using the dialogues in their original form.

		Prediction (Original Version)		
		clinical	non-clinical	total
Gold Standard	clinical	160	14	174
	non-clinical	9	17	26
	total	169	31	200

Table 2 shows the results of the evaluation metrics for classifying the dialogues in their original version. According to Equation 1, the model correctly classified 177 of the 200 dialogues evaluated, resulting in an accuracy of 88.5%. The *Recall* was 92.5%, according to Equation 2. The *Precision*, determined by

Equation 3, was 94.7%. Finally, the F_1 score, calculated by Equation 4, reached 93.6%.

Table 2: We calculated the evaluation metrics: Accuracy, Recall, Precision, and F_1 -score for the 200 dialogues between doctors and patients. We used the dialogues in their original form to carry out this classification.

Ac	Recall	Prec	F_1
88.5%	92.5%	94.7%	93.6%

Results Based on Summarized Texts. Table 3 presents the analysis focusing on the dialogues after summarization. The aim is to assess the model's ability to distinguish between categories, even with more summarized information. Of the 200 summarized dialogues, the model categorized 174 as Positive (P) and 26 as Negative (N). This analysis resulted in 158 cases of True Positive (VP), 18 of True Negative (VN), 16 of False Negative (FN), and 8 of False Positive (FP).

Table 3: We compared the gold standard classification with the one predicted by our model in the 200 dialogues between doctors and patients. We used the **dialogues in their summarized form** for this classification analysis.

		Prediction (Summarized Version)		
		clinical	non-clinical	total
Gold Standard)	clinical	158	16	174
	non-clinical	8	18	26
	total	166	34	200

Table 4 shows the values of the evaluation metrics for classifying the dialogues relying on summarized texts as input for the classification. The model resulted in an accuracy of 88.0%. The *Recall* was 91.3%, according to Equation 2. The Precision, determined by Equation 3 was 95.2%. Finally, the F_1 score, calculated by Equation 4, reached 93.2%.

Table 4: We calculated the evaluation metrics: Accuracy, Recall, Precision and F_1 -score for the 200 dialogues between doctors and patients. We used the dialogues in their **summarized form** for this classification.

Ac	Recall	Prec	F_1
88.0%	91.3%	95.2%	93.2%

5.4 Results of the RDF Triples Extraction

We present the results of the complete execution of the *pipeline*. We used a subset of 20 doctor-patient dialogues in this experiment, randomly selected from the 200 dialogues from the previous evaluation. We start by discussing the results of classifying the clin-

ical aspects in this subset and then make a qualitative analysis of the triples generated in the extraction phase, highlighting the best effectiveness of each model, among other aspects.

Table 5 compares the classification predicted by our model with the expected classification in our Gold Standard. Table 5 shows the results for both strategies: the one that uses the original dialogue and the one that uses its summarized version. In the analysis, we observed 13 and 14 cases of True Positive for the original and summarized versions, respectively, and 3 and 4 instances of True Negative. As for the predicted classifications, the original version had 1 case of False Positive and 3 of False Negative. The summarized version presented 0 and 2 instances of these errors, respectively.

Based on the results in Table 5, we calculated the Accuracy, *Recall*, Precision, and F_1 -score metrics for both strategies. Table 6 shows the results for the strategy that uses dialogues in their original form. In this context, the accuracy was 80.0%. The *Recall* reached 81.3%, while the Precision was 92.9%. The F_1 -score registered 86.7%.

On the other hand, Table 7 presents the results relying on the summarized version of the dialogs. The Accuracy and *Recall* reached 90.0% and 87.5%, respectively. Accuracy reached 100%, and the F_1 -score was 93.3%.

In the qualitative analysis of the RDF triples generated (relying on these 20 randomly selected dialogs), we observed the key effectiveness of the three selected models.

- **mREBEL**: Among the models evaluated for triple extraction, mREBEL performed worst in extracting RDF triples for original and summarized inputs. The triples generated by this model were often incoherent or did not capture the essential information. Thus, mREBEL could not produce RDF triples that adequately reflected the patient's clinical condition.
- **BLOOM**: The BLOOM model, when tested with the *few-shot prompting* strategy and variation of the *top-p* parameter, showed superior performance to mREBEL. This produced quality triples in several assessments. However, the triples lacked objectivity and standardization in some situations, especially when the original dialogue was used as input. The extracted triples were long-winded in certain cases, using extensive dialogue segments instead of keywords or concepts. The "object" field was often affected by this problem and was not even generated occasionally.
- **Fine-tuned GPT-NeoX 20B**. This model of the GPT family stood out in our evaluation, outper-

Table 5: Comparing the classification of our gold standard with that classified by our model in 20 randomly selected dialogues between doctors and patients. We performed this classification using the dialogues in their original and summarized form.

		Classification (Original Version)			Classification (Summarized Version)		
		clinical	non-clinical	total	clinical	non-clinical	total
Gold Standard	clinical	13	3	16	14	2	16
	non-clinical	1	3	4	0	4	4
	total	14	6	20	14	6	20

Table 6: Evaluation Metrics: Accuracy, Recall, Precision and F_1 -score based on the 20 randomly selected dialogues in their original form.

Ac	Recall	Prec	F_1
80.0%	81.3%	92.9%	86.7%

Table 7: Results for the Evaluation Metrics: Accuracy, Recall, Precision and F_1 -score using the **summarized 20 randomly selected dialogues**.

Ac	Recall	Prec	F_1
90.0%	87.5%	100.0%	93.3%

forming the others regardless of the *prompt* strategy applied. The *prompt* based on simple natural language instructions had remarkable results but also showed a few inconsistencies, especially in the output formatting. In contrast, the strategies of *few-shot prompting* and the combination of *few-shot prompting* with simple instruction produced consistent, high-quality results.

The triples generated were relevant and crucial to understanding the patient’s clinical condition. The model performed similarly for original and summarized inputs. In addition, the best effectiveness was observed when the *top - p* parameter was set to 0.5.

Figure 5 shows the triples extracted by each model with an input example in its original and summarized version. These examples were extracted relying on the BLOOM model configured with a *temperature* parameter of 0.1 and *top - p* of 0.9. On the other hand, the Fine-tuned GPT-NeoX model was adjusted with the *temperature* and *top - p* parameters set to 0.0 and 0.5, respectively. Both models employed the *few-shot prompting* strategy for these examples.

With the right approach, we found that text generation models are highly effective for extracting semantic triples from clinical doctor-patient dialogues in Portuguese Language. The success depends on the right combination of model selection, prompt definition, and parameter tuning.

6 DISCUSSION

In the dynamic digital health scenario, telemedicine is emerging as a key solution, offering accessible medical care and overcoming geographical obstacles. In this scenario, technological tools such as AIRdoc are being developed to enrich the experience of health-care professionals and patients. However, such innovations must undergo careful evaluation to confirm their effectiveness and relevance.

In this study, we explored the ability of sophisticated NLP techniques to structure clinical data, focusing on generating RDF triples from clinical dialogues. The main motivation for this research was the demand for a more effective and accessible representation of clinical information. Our study not only defined a new standard for the representation of clinical information but also, indicated a potential impact on healthcare, optimizing patient care and clinical decision-making.

We explored and investigated the use of the mREBEL, BLOOM and GPT-NeoX models adjusted to deal with the complexity of semantic relationships in RDF triple extraction task. A qualitative analysis showed that while mREBEL underperformed, the adjusted GPT-NeoX was the most promising model from our findings. BLOOM, despite outperforming mREBEL, still lacked objectivity.

In our experiments, we observed that the summarization technique proved to be valuable for improving the quality of the generated RDF triples, especially for dense clinical texts. The ability to synthesize information while maintaining its essence was decisive for the success of the process. The similarity of the results between the original and summarized texts attested to the effectiveness of summarization for RDF triple extraction.

The employed models proved robust in classification, achieving an accuracy of over 88% for both original and summarized dialogues. A high recall rate suggested success in identifying clinical texts. The transition to the NeoX model highlights the ongoing commitment to improving the platform’s efficiency. This choice reinforces the need to continually adapt

	<i>Input</i>	<i>mREBEL</i>	<i>BLOOM</i>	<i>Fine-tuned GP-NeoX</i>
Original Version	Médico: Você já fez algum procedimento cirúrgico?, Paciente: Sim, na verdade eu fiz uma histerectomia em março de noventa e nove., Médico: Entendi.	{'head': 'nove', 'head_type': 'concept', 'type': 'follows', 'tail': 'noventa', 'tail_type': 'concept'}}	{('sujeito': "paciente", "predicado": "cirurgia", "objeto": "histerectomia"), ('sujeito': "histerectomia", "predicado": "ano", "objeto": "março de noventa e nove")}	{('sujeito': "paciente", "predicado": "cirurgia", "objeto": "histerectomia")}
Summarized version	O médico perguntou à paciente se tinha feito alguma cirurgia e a paciente respondeu que tinha feito uma histerectomia em Março de 1999, ao que o médico respondeu que compreendia.	{'head': 'histerectomia', 'head_type': 'concept', 'type': 'subclass of', 'tail': 'cirurgia', 'tail_type': 'concept'}}	{('sujeito': "paciente", "predicado": "cirurgia", "objeto": "histerectomia"), ('sujeito': "histerectomia", "predicado": "data", "objeto": "Março de 1999")}	{('sujeito': "paciente", "predicado": "cirurgia", "objeto": "histerectomia"), ('sujeito': "histerectomia", "predicado": "data", "objeto": "Março de 1999")}

Figure 5: Examples for comparison among the models used for the task of extracting RDF triples (text in Portuguese Language).

and evolve the software tool and the underlying techniques adopted. We recognize areas for improvement in our study. The models' ability to distinguish between affirmations, negations, and mentions is challenging, given the possibility of misinterpretations in clinical contexts.

In the technological panorama, it is essential to highlight the role of AIRdoc technology. Even with its focus on specific NLP models and techniques, the AIRdoc solution marks a breakthrough in applying these techniques for clinical data in the Portuguese language. Integrating advanced NLP models into a user-friendly interactive solution such as AIRdoc can advance the current status regarding healthcare professionals' access to and interpretation of clinical information.

In summary, this study shed light on the potential and challenges of NLP techniques in structuring clinical data. The insights gained can reshape the representation of clinical information, benefiting professionals and patients. It is crucial to understand that the findings are initial and future research is demanded to further validate and improve the techniques presented.

7 CONCLUSION

This study explored the confluence of advanced NLP and KG techniques to revolutionize clinical information representation, interpretation, and organization. We developed an innovative method applied to a software tool that transforms transcripts from clinical dialogues into RDF triples, marking a significant transformation in clinical data representation. We highlighted the implementation of a summarization stage, which proved crucial in highlighting and condensing essential clinical aspects of the RDF extraction task. We successfully assessed the effectiveness of the proposed approach for RDF triple generation based on LLMs and few-shot learning. Results suggested a

promising future for the management and exploration of semantically structured clinical records. Our found obstacles reinforce the relevance of future research to improve the techniques presented. Future studies involve integrating more clinical data sources, improving summarization accuracy, and expanding the approach to specific clinical domains.

ACKNOWLEDGEMENTS

This work was supported by the São Paulo Research Foundation (FAPESP) (Grant #2022/13201-3)¹².

REFERENCES

- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.
- Candan, K. S., Liu, H., and Suvarna, R. (2001). Resource description framework: metadata and its applications. *Acm Sigkdd Explorations Newsletter*, 3(1):6–19.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTICS (Posters, Demos, SuCCESS)*, 48(1-4):2.
- Exner, P. and Nugues, P. (2012). Entity extraction: From unstructured text to dbpedia rdf triples. In *WoLE@ ISWC*, pages 58–69.
- Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the conference on empirical methods in NLP*, pages 1373–1378.
- Kamdar, M. R. and Dumontier, M. (2015). An ebola virus-centered knowledge base. *Database*, 2015:bav049.

¹²The opinions expressed in this work do not necessarily reflect those of the funding agencies.

- Kanza, S. and Frey, J. G. (2019). A new wave of innovation in semantic web tools for drug discovery. *Expert Opinion on Drug Discovery*, 14(5):433–444.
- Kormilitzin, A., Vaci, N., Liu, Q., and Nevado-Holgado, A. (2021). Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086.
- Lam, H. Y., Marengo, L., Clark, T., Gao, Y., Kinoshita, J., Shepherd, G., Miller, P., Wu, E., Wong, G. T., Liu, N., et al. (2007). Alzpharm: integration of neurodegeneration data using rdf. *BMC bioinformatics*, 8:1–12.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.
- Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.
- Rossanez, A. and dos Reis, J. C. (2019). Generating knowledge graphs from scientific literature of degenerative diseases. In *SEFDA@ ISWC*, pages 12–23.
- Rossanez, A., Dos Reis, J. C., Torres, R. d. S., and de Ribaupierre, H. (2020). Kgen: a knowledge graph generator from biomedical scientific literature. *BMC medical informatics and decision making*, 20(4):1–24.
- Ruan, T., Huang, Y., Liu, X., Xia, Y., and Gao, J. (2019). Qanalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. *BMC medical informatics and decision making*, 19:1–13.
- Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., and Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.
- Sheng, M., Shao, Y., Zhang, Y., Li, C., Xing, C., Zhang, H., Wang, J., and Gao, F. (2019). Dekgb: an extensible framework for health knowledge graph. In *International Conference on Smart Health*, pages 27–38. Springer.
- Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., et al. (2018). Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.
- Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.