Argument-based natural language explanation generation and assessment in healthcare

Benjamin Molinet

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Génération et évaluation d'explications argumentatives en langage naturel appliquées au domaine médical

Benjamin MOLINET

Université Côte d'Azur, CNRS, Centre Inria d'Université Côte d'Azur, I3S
Équipe Wimmics

Présentée en vue de l'obtention du grade de docteur en Informatique **d'**Université Côte d'Azur

Dirigée par : Elena CABRIO, Professeure des universités, Université Côte d'Azur
Co-dirigée par : Serena VILLATA, Directrice de recherche, CNRS
Soutenue le : 18/12/2024

Devant le jury, composé de :
Catherine FARON ZUCKER, Professeure des universités, Université Côte d'Azur, Nice
Frédérique SEGOND, Directrice Inria Defence, HDR, Inria, Grenoble
Nathalie AUSSENAC-GILLES, Directrice de recherche, IRIT, Toulouse
Raphaël TRONCY, Maître de conférences, EURECOM, Biot

GÉNÉRATION ET ÉVALUATION D'EXPLICATIONS ARGUMENTATIVES EN LANGAGE NATUREL APPLIQUÉES AU DOMAINE MÉDICAL

Argument-based natural language explanation generation and assessment in healthcare

Benjamin MOLINET

Jury :

Président du jury

Catherine FARON ZUCKER,
Professeure des universités, Université Côte d'Azur, Nice

Rapporteurs

Frédérique SEGOND,
Directrice Inria Defence, HDR, Inria, Grenoble
Nathalie AUSSENAC-GILLES,
Directrice de recherche, IRIT, Toulouse

Examinateurs

Raphaël TRONCY,
Maître de conférences, EURECOM, Biot

Directeur de thèse

Elena CABRIO,
Professeure des universités, Université Côte d'Azur

Co-directeur de thèse

Serena VILLATA,
Directrice de recherche, CNRS

Université Côte d'Azur

Benjamin MOLINET

Génération et évaluation d'explications argumentatives en langage naturel appliquées au domaine médical

xii+129 p.

Ce document a été préparé avec LaTeX2e et la classe `these-ISSS` version v. 2.10.

Impression : output.tex – 9/2/2025 – 9:38

Révision pour la classe : these-ISSS.cls,v 2.10 2020/06/24 14:16:37 mpelleau

Génération et évaluation d'explications argumentatives en langage naturel appliquées au domaine médical

Résumé

L'Argument Mining, un domaine en pleine expansion du traitement automatique du langage naturel (TALN) et des modèles informatiques d'argumentation, vise à reconnaître automatiquement les structures d'argumentation (c'est-à-dire les composants et les relations) dans les ressources textuelles en langage naturel. Dans le domaine médical, l'Argument Mining s'est avérée bénéfique en fournissant des méthodes pour détecter automatiquement les structures argumentatives afin de soutenir la médecine fondée sur des preuves. L'importance de ces approches repose sur le fait que, malgré la précision des modèles neuronaux dans la prédiction de diagnostic médical, l'explication de leurs résultats reste problématique. Cette thèse aborde cette question ouverte et se concentre sur la génération et l'évaluation d'explications argumentatives en langage naturel pour les prédictions de diagnostic médicaux, afin d'aider les cliniciens dans la prise de décision et l'éducation. Tout d'abord, j'ai proposé un nouveau pipeline complet pour générer automatiquement des explications en langage naturel d'examens (QCM) médicaux sur les diagnostics en s'appuyant sur une ontologie médicale et des entités cliniques détectées à partir des textes d'examen. J'ai défini un système état de l'art de reconnaissance et de classification des entités nommées médicales (NERC) pour détecter les symptômes exprimés par les patients et les mesures médicales que j'aligne sur les termes de l'ontologie afin de justifier le diagnostic d'un cas clinique fourni aux étudiants en médecine. Le pipeline, appelé SYMEXP, permet à notre système de générer des explications argumentatives en langage naturel basées sur des templates afin de justifier pourquoi la bonne réponse est correcte et pourquoi les autres options proposées ne le sont pas. Deuxièmement, j'ai proposé un cadre d'évaluation des explications basées sur l'argumentation, appelé ABEXA, pour extraire automatiquement la structure argumentative d'un QCM médicale et mettre en évidence un ensemble de critères personnalisables pour caractériser l'explication clinique et l'argumentation du document. ABEXA aborde la question de l'évaluation des explications d'un point de vue argumentatif en définissant un ensemble de patterns sur un graphe argumentatif généré automatiquement. Troisièmement, j'ai contribué à la conception et au développement de la suite de logiciels ANTIDOTE, qui propose différents modules d'intelligence artificielle explicative guidée par l'argumentation pour la médecine. Notre système offre les fonctionnalités suivantes : analyse argumentative multilingue pour le domaine médical, explication, extraction et génération de diagnostics cliniques, modèles linguistiques multilingues pour le domaine médical, et le premier benchmark multilingue de QCM médicaux.

En conclusion, dans cette thèse, j'explore comment l'intelligence artificielle combinée à la théorie de l'argumentation pourrait conduire à des systèmes de soins et de santé plus transparents. Nous appliquons nos résultats au domaine critique de la médecine en montrant tout leur potentiel en termes de soutien à l'éducation, par exemple, des étudiants en médecine.

Mots-clés : Traitement Automatique du Langage Naturel, Extraction de structures argumentatives, Argumentation Explicative.

Argument-based natural language explanation generation and assessment in healthcare

Abstract

Argument(ation) mining, a rapidly growing area of Natural Language Processing (NLP) and computational models of argument, aims at the automatic recognition of argument structures (i.e., components and relations) in natural language textual resources. In the healthcare domain, argument mining has proven beneficial in providing methods to automatically detect argumentation structures to support Evidence-Based Medicine (EBM). The importance of these approaches relies on the fact that, despite the accuracy of neural models in medical diagnosis, explanation of their outcomes remains problematic. The thesis tackles this open issue and focuses on generation and assessment of natural language argumentative explanations for diagnosis predictions, supporting clinicians in decision making and learning. First, I proposed a novel complete pipeline to automatically generate natural language explanations of medical question answering exams for diagnoses relying on a medical ontology and clinical entities from exam texts. I defined a state-of-the-art medical named entity recognition and classification (NERC) system to detect layperson symptoms and medical findings that I align to ontology terms so as to justify a diagnosis of a clinical case provided to medical residents. NERC module allows our system, called SYMEXP, to generate template-based natural language argumentative explanations to justify why the correct answer is correct and why the other proposed options are not. Second, I proposed an argument-based explanation assessment framework, called ABEXA, to automatically extract the argumentation structure of a medical question answering document and highlight a set of customisable criteria to characterise the clinical explanation and the document argumentation. ABEXA tackles the issue of explanation assessment from the argumentation viewpoint by defining a set of graph rules over an automatically generated argumentation graph. Third, I contributed to the design and development of the ANTIDOTE software tool, proposing different modules for argumentation-driven explainable Artificial Intelligence for digital medicine. Our system offers the following functionalities: multilingual argumentative analysis for the medical domain, explanation, extraction and generation of clinical diagnoses, multilingual large language models for the medical domain, and the first multilingual benchmark for medical question-answering.

In conclusion, in this thesis, I explore how artificial intelligence combined with the argumentation theory could lead to more transparent healthcare systems. We apply our results to the critical domain of medicine showing all their potential in terms of support for education, for example, of clinical residents.

Keywords: Natural Language Processing, Argumentation Mining, Explanatory Argumentation.

Remerciements

My thesis. Although I have devoted most of my time to my thesis over the last three years, it seems that this trip is soon coming to an end. I would therefore like to thank all the colleagues and friends who have supported me all the way.

First of all, I would like to thank my supervisors Elena and Serena for your advice and scientific opinions, without which this thesis would not have been possible. Thank you for your time, your support, your optimism, your patience and your perseverance towards me.

Many thanks to my rapporteurs Frédérique and Nathalie for your involvement in reading and correcting my manuscript.

Thank Rodrigo, Bernardo, Marcin, Peter, Andrea and other international partners on the ANTIDOTE project for their scientific collaboration, as well as for our exchanges and meals shared in an incomparably good mood. I would also like to thank Olivier and the doctors who took time out of their work day to help me with my experiments.

I would like to thank all the members of WIMMICS and SPARKS for the intellectual exchange and assistance they have given me. I have always felt welcome and I have enjoyed various collaborations. A big thank to Fabien for creating and maintaining motivating and friendly atmosphere within the team. I would also like to thank Marco and Franck for their help with logistics and my experiments, as well as Lionel, Christine and Delphine for their assistance and patience with the administration.

Special thanks to my friends Arnaud, Clément, Pierre, Rémi 1 and Rémi 2, who always listened to my ideas, complaints, doubts and problems. They helped me stay on track during this trip, and we were always able to laugh together about our problems over a good cup of coffee. I hope we can continue to do this in the future.

Finally, I would also like to express my deepest and most sincere thanks to my family. Thanks to my parents, who have supported me every step of the way and encouraged me to pursue my ambitions, even if it meant missing out on some family occasions. Thank to you, Ekaterina, for your unconditional love and support, for putting up with me in my moments of doubt, fear and stress, for giving me all that time and attention. You've helped me mentally and physically through these trials and I'm infinitely grateful to you. Thank you for everything Ekaterina.

FUNDING : This work has been supported by the CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002.

Table of Contents

List of Abbreviations

Acronyms

AA	**A**bstract **A**rgument
ABEXA	**A**rgument-**B**ased **EX**planation **A**ssessmen
ACTA	**A**rgumentative **C**linical **T**rial **A**nalysis
AF	**A**rgumentative **F**rameworks
AM	**A**rgument(ation) **M**ining
ANTIDOTE	**A**rgume**N**ta**TI**on-**D**riven explainable artificial intelligence f**O**r digi**T**al m**E**dicine
ASL	**A**rgument **S**tructure **L**earning
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
BIO	**B**eginning, **I**nside **O**utside
BoW	**B**ag-**o**f-**W**ords
CBOW	**C**ontinuous **B**ag-**o**f-**W**ords
CoT	**C**hain **o**f **T**hought
CRF	**C**onditional **R**andom **F**ield
CUI	**C**oncept **U**nique **I**dentifiers
DL	**D**eep **L**earning
EA	**E**xplanatory **A**rgumentation
EBM	**E**vidence-**B**ased **M**edicine
EHR	**E**lectronic **H**ealth **R**ecord
EMR	**E**lectronic **M**edical **R**ecord
GloVe	**Glo**bal **Ve**ctors
GPT	**G**enerative **P**retrained **T**ransformer
GPU	**G**raphics **P**rocessing **U**nit
GRU	**G**ated **R**ecurrent **U**nit
HPO	**H**uman **P**henotype **O**ntology
ICD	**I**nternational **C**lassification of **D**iseases
IE	**I**nformation **E**xtraction
IO	**I**nput-**O**utpu
KNN	**K**-**N**earest **N**eighbor
LIME	**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
LiR	**Li**near **R**egression
LLM	**L**arge **L**anguage **M**odel
LM	**L**anguage **M**odel
LOINC	**L**ogical **O**bservation **I**dentifiers **N**ames and **C**ode
LoR	**Lo**gic **R**egression
LSTM	**L**ong **S**hort-**T**erm **M**emory

MaxL	**Max**imum **L**ikelihood
MCMLE	**M**ainland **C**hina **M**edical **L**icensing **E**xamination
MeSH	**Me**dical **S**ubject **H**eadings
MIR	**M**edical **I**ntern **R**esident
ML	**M**achine **L**earning
MedMT5	**Med**ical **M**ultitask **T5**
NER	**N**amed **E**ntity **R**ecognition
NLG	**N**atural **L**anguage **G**eneration
NLI	**N**atural **L**anguage **I**nference
NLP	**N**atural **L**anguage **P**rocessing
PICO	**P**atient, **I**ntervention, **C**omparison, **O**utcome
PLM	**P**re-trained **L**anguage **M**odels
PoS	**P**art-**o**f-**S**peech
QA	**Q**uestion **A**nswering
RAM	**R**andom-**A**ccess **M**emory
RCT	**R**andomized **C**linical **T**rial
RLHF	**R**einforcement **L**earning from **H**uman **F**eedback
RQ	**R**esearch **Q**uestions
RQE	**R**ecognizing **Q**uestion **E**ntailment
RNN	**R**ecurrent **N**eural **N**etworks
SAEI	**A**ndalusian **S**ociety of **I**nfectious **D**iseases
SBMI	**S**chool of **Bio**m**e**dical **I**nformatics
SC	**S**elf **C**onsistency
SHAP	**SH**apley **A**dditive ex**P**lanations
SNOMED CT	**S**ystematized **N**omenclature of **Med**icine - **C**linical **T**erms
SOTA	**S**tate-**O**f-**T**he-**A**rt
SVCCA	**S**ingular **V**ector **C**anonical **C**orrelation **A**nalysis
SVM	**S**upport **V**ector **M**achines
SYMEXP	**SYM**ptomaticaly **EXP**lanation
TCAV	**T**esting with **C**oncept **A**ctivation **V**ector
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requenc
t-SNE	**t**-distributed **S**tochastic **N**eighbor **E**mbedding
T5	**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer
UMLS	**M**edical **L**anguage **S**ystem
USMLE	**U**nited **S**tates **M**edical **L**icensing **E**xamination
TWMLE	**T**ai**W**an **M**edical **L**icensing **E**xamination
XAI	E**x**plainable **A**rtificial **I**ntelligence

CHAPTER 1

Introduction

This chapter outlines the motivations behind the work presented in this thesis. It highlights the importance of explanations in sensitive domains such as medicine, and justifies the need for automatically generating natural language explanations. Further, it highlights the necessity of characterizing explanations from an argumentation viewpoint, particularly for educational purposes. It then presents the definition and development of tools for explanatory argumentation and their adaptation to the medical domain. Finally, the precise research questions I answered in this thesis are formulated, and an overview of the manuscript structure is provided.

1.1 Motivations

Medical decision-making is a process that requires doctors to apply complex reasoning, taking into account potential symptoms, medical history and patient's laboratory test results in order to formulate a diagnosis, decide on a treatment or any type of medical procedure. In addition, clinical doctors must also be able to justify their choices to patients, medical residents or other clinicians. Decision explanation is essential as it helps in strengthening trustworthiness in diagnosis or treatment, allowing other people to understand reasoning behind medical deliberation. Explanations are even more important in high-stakes situations where the justification for a treatment or diagnosis is as important as the decision itself.

Given the complexity of medical data and due to advancements in Artificial Intelligence (AI), thechnological solutions like IBM Watson for Oncology [184] or AI-based image analysis tools such as DeepMind's retinal disease detection system [56] have become essential to assist doctors in a range of medical tasks, such as automatically predicting diagnoses, surgical robots, and medical image analysis. Systems based on Artificial Intelligence methods, more specifically deep learning, have demonstrated their performance by ingesting vast quantities of data in order to produce predictions comparable to those made by humans in certain tasks [66, 79]. These systems have great potential to improve Evidence-Based Medicine (EBM) by serving as decision-making support for clinicians, providing rapid and effective predictions over huge quantity of data. However, while AI offers a significant increase in performance, the process by which these predictions are made often remains hidden.

Despite the advantages of AI in the medical domain, the majority of systems are based on neural models considered as "black boxes" [78] where the way the model reaches a prediction is opaque to users. While clinicians may be able to obtain high-quality predictions about a patient's diagnosis or treatment, they have no opportunity to find out the reasons behind such predictions. This problem is of main importance, especially when the system proposes false or biased decisions, in sensitive use case scenarios like medicine and law. This lack of transparency is an open challenge in the field of Artificial Intelligence, being even more significant when it comes to the medical field where trustworthiness and comprehensibility are critical for decision making. Without a clear explanation based on verified facts or evidence, these neural approaches are troublesome for the medical field.

To address these limitations, a lot of interest is focused on the field called eXplainable Artificial Intelligence (XAI) [134, 197, 10], which has the aim of making predictions of AI systems transparent and understandable to humans. The need for XAI in healthcare is even more pressing as clinicians need to be able to understand the results of these approaches in order to profitably employ them in their diagnoses. One research line of XAI consists in generating natural language explanations that can be directly interpreted by clinicians or patients. The generation of these natural language explanations relies, in particular in domain-specific use cases, on the employment of specific (medical) knowledge, to ground the explanations on reliable evidence and to make the reasoning behind the

decision more transparent.

While XAI proposes solutions to improve the transparency and interpretability of AI systems, it often focuses on formally explaining machine learning models and understanding how they employ features to make predictions. While some of these approaches are human-interpretable, they still require additional effort and a solid understanding of machine and deep learning methods to fully grasp the results [197]. Looking at XAI systems that are directly interpretable by humans and in natural language, we find few approaches that allow us to justify predictions through natural language explanations [223, 54], and even fewer that are applied to the medical field [40, 118]. Argumentation offers an opportinity to generate structured natural language explanations increasing human understanding [169, 50]. In the medical domain, clinical decisions are often built up with structured arguments, where evidence is presented, weighed, and justified to support a diagnosis or treatment choice. Although many XAI approaches have been applied to the medical domain, argumentation-based explanation generation remains underexplored in healthcare applications. For instance, a medical case typically includes clinical information about a patient's health, such as symptoms and test results, which guide a clinician's decision on diagnosis or treatment. An approach that focuses solely on extracting this information cannot justify the clinician's decision at a further step. Moreover, providing a natural language explanation for a correct diagnosis or treatment requires understanding of the argumentative structure to capture the reasoning behind the doctor's decision. Addressing this gap between approaches for argument mining and those for the detection and extraction of expert knowledge towards the generation of coherent and well-grounded natural language explanations can help clinicians to make informed decisions.

This thesis has been pursued in the context of the European ANTIDOTE project [1], whose goal is to meet the challenge of providing a unified computational framework for joint learning of clinical predictions and associated argumentative justifications, promoting natural interaction with clinicians through explanatory dialogues. More precisely, we aim at investigating joint learning of predictions such as diagnoses and their supporting evidence across a range of clinical scenarios with varying levels of complexity. This includes settings where clinical cases contain the majority of necessary evidence, as well as those where evidence is derived from external knowledge sources. Moreover, we seek to connect these predictions and their corresponding evidence into logically consistent explanations following well-established theoretical frameworks in argumentation theory. In addition, we need to explore how these explanations can be integrated into explanatory dialogues, where users with different expertise, such as medical students or practitioners, engage in a task-oriented dialogue aimed at assessing understanding or guiding clinical decision-making. Lastly, the project aims to establish new benchmarks for generating high-quality explanatory arguments, with a focus on tailoring explanations to different audiences by adapting explanation modes to meet user-specific needs.

This PhD thesis aims to address several challenges raised in the ANTIDOTE project by exploring novel approaches for generating natural language explanations that rely on expert

1. https://univ-cotedazur.eu/antidote

knowledge automatically extracted from medical documents and aligned with specialized ontologies. Additionally, it seeks to assess the argumentation of these medical explanations.

1.2 Research Questions

In this section, I detail the research questions I answered in this theis, the methodology I adopted to answer the research questions, and the scientific publications resulted from each research question.

The generation of natural language explanations justifying medical decisions is a critical issue. My first research line has been therefore to explore the generation of argument-based natural language explanations, so that explanatons are grounded both on facts and hypotheses extracted from medical documents and in accordance with external reliable sources of knowledge. More precisely, I answered the following research question (RQ) :

RQ1 : *How can we generate structured natural language explanations for medical diagnoses that integrate expert knowledge and align with it ? This question breaks down into the following sub-questions :*

- *How can we automatically identify and interpret patient's information expressed in layperson terms to justify medical diagnoses ?*
- *Does incorporating contextual information from a clinical case improve the alignment of extracted patient's data with structured medical knowledge sources like ontologies ?*
- *How can we generate argumentative explanations in natural language that comply with reliable medical knowledge ?*

I answer this research question in Chapter 3 by focusing on clinical examination cases for medical students, questioning them about patient diagnoses. I identified, with the assistance of a medical expert, the main features that enable diagnoses to be justified, and proposed a novel method for identifying them despite the fact that they are described by the patient in layperson terms. A new dataset of 314 clinical cases annotated with labels from a medical vocabulary and a medical findings conversion database have been built to train these models and interpret patient observations.

To address the alignment with external knowledge, I adopted the HPO ontology [110] to retrieve medical knowledge validated by experts and I aligned the concepts automatically extracted from clinical cases with the ontology conceptss, by comparing the impact of the context of the latter on the alignment performance.

Finally, to generate reliable explanations, I worked out templates with a clinician to avoid any hallucination resulting from the adoption of Large Language Models for the generation of the natural language explanations. I decided to integrate three approaches to justify the diagnosis together, highlighting why a certain diagnosis is correct, why the other options are not, and underlying the relevant information for decision-making that is missing from the clinical case. The generated explanations are based on the clinical case's references, on the ontology's aligned concepts explicitly associated with the clinical case's entities. The explanations are also based on statistical information retrieved from the ontology.

Related publications :

— **Molinet, B.**, Marro, S., Cabrio, E., & Villata, S. (2024). Explanatory argumentation in natural language for correct and incorrect medical diagnoses. Journal of Biomedical Semantics, 15(1), 8.

— Marro, S., **Molinet, B.**, Cabrio, E., & Villata, S. (2023, February). Natural language explanatory arguments for correct and incorrect diagnoses of clinical cases. In ICAART 2023-15th International Conference on Agents and Artificial Intelligence (Vol. 1, pp. 438-449).

After the generation of argument-based natural language explanations, particularly in the medical field, it is necessary to evaluate these explanations, with a special interest on educational purposes. For residents, developing a proper argumentation in their medical explanations is essential to improve the quality of the interaction with the patient. In this thesis, I focused on characterising the argument-based features of natural language explanations to automatically assess them. The answered the following research question :

RQ2 : *Can argumentation be applied to identify specific features of medical explanations ?*

To answer this question, I carried out an empirical analysis of clinical cases annotated with argument components (premises, claims) and relations (supports, attacks) and proposed a set of criteria to characterise the argumentation of medical explanations. In Chapter 4, I present a novel pipeline to automatically extract the argumentation structure of medical documents to retrieve the components and relations between arguments components. The resulting argumentation graph is then analysed to detect previously introduced patterns and to provide the writer of the explanation with the characteristics of the latter, enabling her to improve it.

Related publication :

— **Molinet, B.**, Villata, S., & Cabrio, E. (2022, July). Assessing Argument-based Natural Language Explanations in Medical Text. In SAC 2025-ACM SIGAPP Symposium on Applied Computing (*under review*).

Finally, the development of new specialised tools for explanatory argumentation, particularly adapted to critical fields such as medicine, is necessary. This allows the development of more complex underlying pipelines adapted to more specific tasks such as the generation of explanations based on natural language. I answered this research question :

RQ3 : *How to design and develop new tools for the task of argumentation analysis and natural language argument-based explanation generation for medical applications ?*

I answered this research question by designing and developing some tools which are described in Chapter 5. More specifically, we have developed a tool to automatically detect argumentative structures in medical texts and improved its robustness, incresing also its modularity and accessibility. We have also focused on the development of multilingual and multitask LLMs pre-trained on medical documents, showing improved performance on

under-represented languages for the medical domain. Finally, in the context of the ANTIDOTE project, we have proposed a software suite to improve explanatory argumentation, particularly in terms of explanation generation.

Related publications :

- Cardellino, C., Collias, T., **Molinet, B.**, Hain, E., Sun, W., Agerri, R., ... & Cabrio, E. (2024, October). ANTIDOTE : ArgumeNtaTIon-Driven explainable Artificial Intelligence fOr digiTal mEdicine. In ECAI-24-Demos Proceedings-27th European Conference on Artificial Intelligence.
- **Molinet, B.**, Marro, S., Cabrio, E., Villata, S., & Mayer, T. (2022, July). ACTA 2.0 : A modular architecture for multi-layer argumentative analysis of clinical trials. In IJCAI 2022-Thirty-First International Joint Conference on Artificial Intelligence.
- García-Ferrero, I., Agerri, R., Salazar, A. A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., **Molinet, B**, ... & Zaninello, A. (2024, May). MedMT5 : An Open-Source Multilingual Text-to-Text LLM for the Medical Domain. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 11165-11177)

1.3 Structure of the thesis

The thesis is organized as follows :

Chapter 2 describes the preliminaries, which are used throughout the thesis. It gives an overview of the background and methods for representing language numerically, as well as classical transformer-based approaches. It also covers the fundamentals of Argument Mining for extracting argumentation from text, as well as an overview of explanatory Artificial Intelligence, specifically in natural language. Finally, the main knowledge sources specialized in the medical field are described.

Chapter 3 presents a novel automatic pipeline for generating template-based explanations for medical diagnoses from medical exams. It addresses the automatic extraction of key concepts for disease diagnosis and the conversion of health measurements into a standardized vocabulary. A dataset of 314 clinical cases annotated with medical concepts from ontology is built, together with a conversion database of 100 findings and their boundaries, developed and validated by a clinician. The explanations are based on concepts detected and converted in line with a specialized ontology (i.e., HPO) automatically aligned using a context-sensitive neural method showing superior performance to non-contextual approaches. Natural language explanation generation is tackled with templates, automatically populated with the detected and aligned concepts according to argumentation patterns justifying why a certain diagnosis is correct and the other options are not, highlighting missing information to improve the diagnosis explanation.

Chapter 4 explores how medical natural language explanations can be assessed from an argumentative viewpoint. It analyses the Casimedicos dataset composed of 553 question

answering clinical cases and introduces a set of criteria to characterize natural language explanations from an argumentation perspective. A proposed architecture addresses an end-to-end argument mining pipeline to identify argument components and relations, and then, analyzes the resulting argumentation graphs to identify patterns which characterize explanations from an argumentation viewpoint.

Chapter 5 introduces tools for Argumentation-Driven Explainable AI for Medicine. These tools have been developed or improved, such as the ACTA tool, to increase including performance, distribution and the addition of features for evidence-based medicine. The chapter also introduces the Medical Multilingual T5 (MedMT5) language model, which allows a multitude of natural language processing medical tasks to be performed in four languages (i.e., English, French, Italian, Spanish). Finally, the ANTIDOTE tool suite is presented, offering a number of tools for Argumentation-Driven Explainable AI for Medicine.

Chapter 6 concludes the thesis by summarizing the main contributions. Perspectives for future research directions and applications are discussed, as well as potential plans to improve existing work.

CHAPTER 2

Background

This chapter summarizes the preliminaries used throughout the thesis. First, I present the key concepts which establish the basis of my research work. Next, the ways of representing language are discussed. They lay the foundation for tools available for carrying out Natural Language Processing (NLP) tasks. Then, an overview of the Argument Mining task is given, discussing its main applications. Then, I set the context for Explanatory Argumentation by zooming in on its formulation in natural language applied to the medical field. Finally, I present existing resources and knowledge bases in the medical domain.

The increasing use of AI in healthcare has raised interest in eXplainable AI systems, particularly in generating natural language explanations. Such systems aim to enhance the interpretability and transparency of AI-generated predictions by providing explanations that are understandable to both medical professionals and patients. A promising approach to enhancing the clarity of these explanations is to employ argumentation, where the inferential reasoning and justification are the main components.

In this chapter, I explore the fundamental aspects necessary to understand the mechanisms behind argument-based natural language explanation in healthcare. This chapter bridges the fields of NLP, argumentation theory, and medical knowledge, focusing on how they contribute to the creation and assessment of explanations. The interdisciplinary nature of this work reflects the complex challenges of generating arguments that are not only accurate, but also relevant and comprehensible within a medical context.

The chapter is organised as follows. Section 2.1 discusses natural language representations, an essential basis for NLP models. Section 2.2 presents argument mining, an essential computational task for identifying arguments from textual data. In Section 2.3, I focus on explanatory argumentation, which combines principles of argumentation theory and philosophy of explanations to produce meaningful justifications in the healthcare context. Finally, Section 2.4 reviews medical knowledge bases to ensure that the explanations generated are sound and reliable.

2.1 Natural Language Representation

Natural language is a human mechanism of communication that can be processed thanks to human cognitive functions. Making natural language interpretable and compatible with machines is a major challenge in computer science and linguistics [99]. While many machine learning algorithms perform well on mathematical tasks and make good predictions on unseen data based on training, applying these methods to language requires a mathematical representation of the latter. Inherent to NLP, the evolution of language representation for machines has been dominated by rule-based systems, rapidly replaced by statistical representations based on machine and deep learning. This numerical representation of language forms the foundation of all the NLP tasks [93]. The most fundamental tasks of prediction, inference and generation are grounded in mathematical representations of language. However, converting text into numbers (i.e., especially vectors) is not an easy task and depends on many factors such as vocabulary, grammar [44], or context [159]. Context is even more important as it enables words to be disambiguated according to the situation in which they are used. If this task is sometimes difficult for a human, it is even more so for a machine, which is why it remains an open question in NLP. The NLP community has nonetheless proposed a range of solutions that use or do not use context to represent the language, and has gradually developed different architectures that are suitable for certain NLP tasks. This section focuses on the use of context to represent language and then zooms in on the major architectures based on transformers, the most used architecture to represent and work on language.

2.1.1 Context-free Representations

Early approaches to natural language representation treat words independently of their context, assigning each word a fixed numerical representation. This section discusses several context-free methods, including One-Hot encoding, frequency-based methods, and neural embeddings.

One-Hot Encoding. One-Hot Encoding is a fist approach for converting words into numerical form, where each word in the vocabulary is represented by a vector of zeros with a single one at the position corresponding to that word's index. This approach results in a high-dimensional and sparse vector space, with dimensionality equal to the size of the vocabulary. While straightforward to implement, One-Hot Encoding does not capture any semantic relationships between words and all words are equidistant in this representation, failing to reflect similarities or differences in meaning. For instance, the words "cat" and "dog" are represented as entirely distinct, despite their related meanings (e.g., animals). Moreover, high dimensionality poses computational challenges, especially when dealing with large vocabularies common in natural language processing tasks.

Frequency-Based Methods. Frequency-based methods represent text by quantifying the occurrence of words within documents, capturing basic statistical properties of a language. A fundamental approach is the Bag-of-Words (BoW) model, which represents a document as a vector of word frequencies, ignoring grammar and word order while preserving the multiplicity of words [83]. Another widely used frequency-based technique is Term Frequency-Inverse Document Frequency (TF-IDF), which weighs the importance of a word in a document relative to its frequency across a corpus [186]. An extension of these methods involves *n-grams*, which are contiguous sequences of n words used to capture local context [180]. N-gram models estimate the probability of a word based on the occurrence frequencies of its preceding n-1 words, thus incorporating some sequential information. While unigram models (n=1) consider individual word frequencies such as BoW or TF-IDF, bigram (n=2) and trigram (n=3) models account for short-range dependencies between words. These models have been fundamental in statistical language modeling and have improved tasks like speech recognition and text prediction. However, n-gram models suffer from data sparsity issues as n increases and still does not capture long-range dependencies or deeper semantic relationships. While these frequency-based methods are simple and effective for tasks like information retrieval and text classification, they treat each word or n-gram as an independent unit without considering broader context. Moreover, the resulting high-dimensional and sparse representations pose computational challenges.

Neural Representations. Neural embeddings address the limitations of one-hot encoding and frequency-based methods by capturing semantic relationships between words in a continuous vector space. This approach was pioneered by Bengio et al. [23], who introduced a neural probabilistic language model (LM) that learns word embeddings jointly with a statistical LM. A significant advancement came with the introduction of Word2Vec by Mikolov et al. [133]. Word2Vec includes two architectures : Continuous Bag-of-Words (CBOW) and Skip-Gram, illustrated in Figure 2.1.

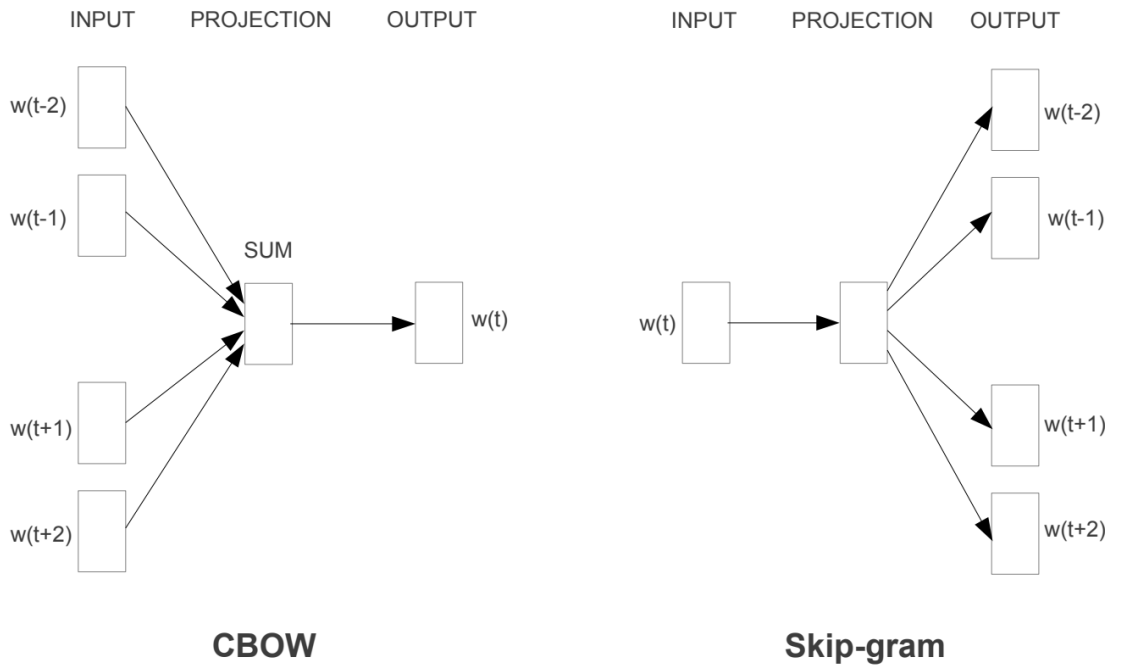

Figure 2.1 – Word2vec CBOW and Skip-Gram representation. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

The CBOW model predicts a target word based on its surrounding context words, effectively utilizing a sliding window over the text. The Skip-Gram model, on the other hand, predicts surrounding context words given a target word. Unlike traditional n-gram models that rely on frequency counts, Word2Vec learns embeddings through prediction tasks using neural networks, making it more scalable and capable of capturing long-range dependencies. These embeddings capture both syntactic and semantic relationships, exhibiting patterns that allow for vector operations reflecting linguistic analogies (e.g., vectors "king" - "man" + "woman" results in a vector close to "queen").

Making use of these concepts, Pennington et al. [157] developed GloVe (Global Vectors), which combines global word co-occurrence statistics with local context methods. GloVe constructs a word-word co-occurrence matrix from the corpus and factorizes it to produce word embeddings. This approach captures both global statistical information and local context, leading to improved performance on various linguistic tasks.

Further enhancements were introduced by Bojanowski et al. [26] with FastText, which incorporates subword information into word embeddings. FastText represents each word as a bag of character n-grams (e.g., "apple" will be represented together with "ap", "pp", "pl", "le"), allowing the model to generate embeddings for rare or misspelled words and to handle morphologically rich languages more effectively. This subword approach enriches the embeddings with morphological information, improving the model ability to capture word similarities based on shared substrings.

These neural representation methods significantly advance the numerical representation of language by producing dense, low-dimensional embeddings that capture meaningful

semantic and syntactic (role and meaning) relationship between words. However, even if they are trained with pieces of context, they still assign a single vector to each word regardless of its usage context, limiting their ability to represent words with multiple meanings. This limitation highlights the need for context-aware representations that adjust word embeddings based on their usage in different contexts.

2.1.2 Context-Aware Representations

The limitations of context-free models, which assign a fixed representation to each word and do not consider its usage in context, necessitate approaches that capture the dynamic nature of language. Words often have multiple meanings depending on their usage in different texts or discourses. Context-aware representations address this challenge by generating word embeddings that are sensitive to the surrounding words, effectively capturing contextual nuances inherent to natural language. These models enhance the ability of machines to interpret language accurately, improving performance across various NLP tasks such as machine translation, question answering, and sentiment analysis.

Recurrent Neural Networks. Recurrent Neural Networks (RNNs) are a class of neural networks designed to model sequential data by maintaining a hidden state that captures information about previous inputs [65]. Unlike feedforward neural networks (i.e., that processes the entire input data simultaneously), RNNs process input sequences one element at a time, allowing information to persist across time steps. This architecture makes them well-suited for processing natural language, where the meaning of a word often depends on the preceding words in a sentence.

However, training RNNs poses challenges due to the vanishing and exploding gradient problems, which make it difficult for the network to learn long-term dependencies [24]. The gradients used to update the network's weights can become exceedingly small or large as they are propagated back through many time steps, hindering effective learning.

Long Short-Term Memory Networks. To overcome the limitations of standard RNNs in capturing long-term dependencies, Hochreiter and Schmidhuber [89] introduced the Long Short-Term Memory (LSTM) network. LSTMs utilize a more complex architecture that includes memory cells and gating mechanisms to regulate the flow of information. Each LSTM cell contains input, output, and forget gates that control the cell state, allowing the network to learn when to remember or forget information. This structure enables LSTMs to maintain and update information over longer sequences, effectively mitigating the vanishing gradient problem [89]. LSTMs have been successfully applied to various NLP tasks, including language modeling [193], speech recognition [76], and machine translation [194].

Gated Recurrent Units. Gated Recurrent Units (GRUs), introduced by Cho et al.[43], provide an alternative gating mechanism to LSTMs while maintaining computational efficiency. GRUs combine the forget and input gates into a single update gate and merge the cell state and hidden state, resulting in a simpler architecture with fewer parameters. The GRU's gating mechanisms regulate the flow of information, allowing the network to capture dependencies over long sequences without the complexity of LSTMs [43]. GRUs have demonstrated comparable performance to LSTMs on various tasks such as machine

translation and speech recognition [46], often with faster convergence and reduced training time due to their simpler structure.

Attention Mechanisms. While LSTMs and GRUs improve the ability to capture long-term dependencies, they can still struggle with very long sequences due to the sequential nature of their architectures. Attention mechanisms, introduced by Bahdanau et al.[16], address this limitation by allowing the model to focus on specific parts of the input sequence when generating each part of the output sequence. In the context of machine translation, the attention mechanism computes a weighted sum of the encoder's hidden states, where the weights are learned to reflect the relevance of each input token to the current decoding step [16]. This approach enables the model to capture alignments between input and output sequences more effectively, improving translation quality and performance on other sequence-to-sequence tasks.

These attention mechanisms formed the basis for the development of Transformer architectures [202], which rely entirely on attention mechanisms and dispense with recurrence altogether. Transformers have further advanced context-aware representations by enabling models to capture dependencies regardless of sequence length with greater computational efficiency.

RNNs, LSTMs, GRUs, and attention mechanisms represent significant steps forward in context-aware language representation, enabling models to consider sequential and contextual nature of language. Their development has opened the way for more advanced architectures that enhance the capability of machines to understand and generate human language.

2.1.3 Transformer-Based Architectures and Models

The advent of Transformer-based architectures has enhanced natural language processing by enabling models to capture long-range dependencies more effectively than previous recurrent models [202]. These architectures rely entirely on attention mechanisms to process input sequences, allowing for greater parallelization and efficiency in training. The main variants of Transformer architectures can be categorized into encoder-only, encoder-decoder, and decoder-only models, each serving different types of NLP tasks. Figure 2.2 illustrates the Transformer architecture, highlighting the encoder (left part) and decoder (right part) components and the flow of information between them.

Encoder-Only. Encoder-only models utilize the Transformer encoder to generate contextualized representations of input sequences without a corresponding decoder component. A prominent example is BERT (Bidirectional Encoder Representations from Transformers) introduced by Devlin et al. [58], often used for its encoder part. BERT pretrains deep bidirectional representations by jointly conditioning on both left and right context using masked language modeling, where certain tokens are masked, and the model learns to predict them based on their surroundings. This approach allows BERT to capture nuanced meanings of words in different contexts. Encoder-only architectures excel in understanding and analyzing text, making them effective for tasks such as question-answering, natural language inference, and named entity recognition. Several domain-specific adaptations of BERT have been developed to enhance performance in specialized areas. For example,

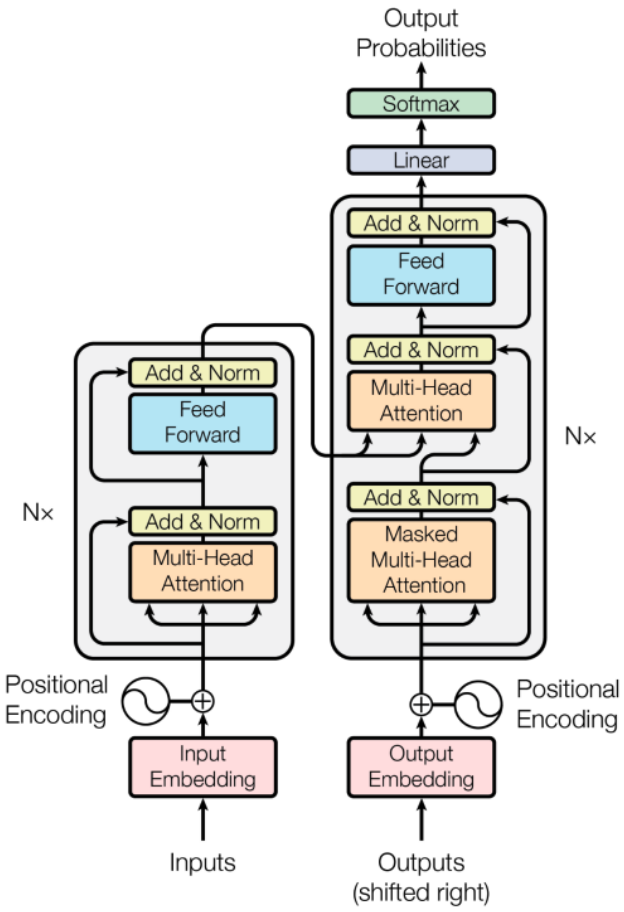

Figure 2.2 – The Transformer model architecture, consisting of stacked encoder and decoder layers. Figure adapted from Vaswani et al. [202].

SciBERT [20], which is pretrained from scratch on a large corpus of scientific publications, is tailored for scientific text. Similarly, BioBERT [117], initialized from BERT and further pretrained on biomedical text, is designed for biomedical applications, while ClinicalBERT [92], fine-tuned [1] on clinical narratives, is adapted for clinical settings.

Decoder-Only. Decoder-only models consist solely of the Transformer decoder component and are primarily designed for language generation tasks. The GPT (Generative Pretrained Transformer) series developed by OpenAI [164, 165, 30, 154], exemplifies this architecture. Each successive model in the GPT series has increased in size and capabilities, with GPT-3 [30] being a significant gap in terms of model parameters and performance. GPT models are pretrained on large-scale corpora using next-token prediction, learning to generate coherent and contextually appropriate text by predicting each word based on the preceding sequence. GPT-3, in particular, has demonstrated strong performance in few-shot learning settings, enabling it to perform tasks with minimal task-specific fine-tuning or even without any fine-tuning, relying instead on prompt engineering. Subsequent models, such as GPT-3.5 and GPT-4 [154], have incorporated techniques like instruction tuning and Reinforcement Learning from Human Feedback (RLHF) to better align generated text with user instructions and improve the quality and safety of the outputs. These models have demonstrated remarkable capabilities in generating human-like text and have been applied to tasks such as text completion, creative writing, and code generation.

Encoder-Decoder. Encoder-decoder models comprise an encoder that processes the input sequence and a decoder that generates an output sequence. The original Transformer model proposed by Vaswani et al.[202] is an encoder-decoder architecture. These models are well-suited for sequence-to-sequence tasks where the output is a transformation of the input, such as machine translation and text summarization. Another notable example is T5 (Text-to-Text Transfer Transformer) introduced by Raffel et al.[167], which frames every NLP task as a text-to-text problem. By pretraining on a diverse range of tasks, T5 demonstrates strong performance across various applications, highlighting the versatility of encoder-decoder architectures in both understanding and generating text.

These Transformer-based architectures have set new benchmarks in NLP, enabling models to handle complex language phenomena with greater efficacy. The choice of architecture depends on the specific requirements of the task at hand, with encoder-only models excelling in text understanding, encoder-decoder models in sequence transformation, and decoder-only models in text generation.

2.2 Argument Mining

Argumentation is the human process of presenting reasons, evidence, and logical analysis to support or refute a claim or position. It involves constructing arguments that are coherent, persuasive, and logically structured, with the aim of reaching a conclusion that others can accept based on the presented evidence and reasoning. Argumentation is inherent to various domains such as philosophy, linguistics and mathematics, therefore, it

1. The concept of fine-tuning in NLP refers to the process of taking a pre-trained LM and further training it on a specific task or dataset to improve its performance on that task.

recently gained attention of the Artificial Intelligence (AI) community, as argumentation reasoning or analysis might be facilitated by automatic processes.

Argumentation requires human cognitive properties such as argument identification, language comprehension (represented by semantics and pragmatics), reasoning skills and knowledge usage. Most of the contributions in computational argumentation carried out so far does not rely on real unstructured data (described in Section 2.4). Thus, to make computational approaches applicable to real-world texts, they need to be able to identify in unstructured data, the structure and process it. To do so, different argumentation theories propose schemes to model argumentation.

Regardless of various approaches to formalize argumentation structures, they share the definition of the argument components : claims (i.e., conclusions) and premises (i.e., evidence). Argumentation theories predominantly adopted are those of Toulmin [198], Walton [207] and Freeman [70, 210], proposing argumentation schemes of different granularity. An onverview of main Argumentation Frameworks proposed in the literature is presented in Section 2.3. Three of them consider a combination of a conclusion (claim) and some evidence (premises) as the main components of an argument unit [8]. I will ground on this view of argumentation as the way to model argumentation in a structured machine-interpretable format in this thesis.

Argument(ation) Mining (AM) stands at the junction of these two worlds of theory and application with the aims to extract natural language arguments and their relationship from text, with the ultimate goal of providing machine-processable structured data for computational models of argumentation [116, 32, 137]. Structured argumentation is usually a graph or a tree with extracted facts and hypotheses presented as nodes and argumentation relations presented as (headed) arcs between the nodes, demonstrating attacked or supported argument components from natural language texts.
The Argument Mining task breaks down into the following sub-tasks :

i) **Argument Components Extraction** identifies argument components such as *Claims* and *Premises* within the natural language text.

ii) **Relation Prediction** identifies relations between argumentation components such as *Support* or *Attack* relations.

Historically, an early contribution in this area is the concept of argumentation zoning [196], where sentences within a scientific paper are categorized according to their rhetorical purpose, such as referencing background literature or outlining the objectives of the research. Although this method does not explicitly focus on extracting argumentation structures, it is seen as a precursor to the development of Argument Mining methods. Limited by computational power and tools, first approaches subdivided the task of Argument Component Extraction into easier steps such as a) **Component Segmentation or Boundaries identification** to separate argumentation units from non-argumentation units, and b) **Argument Component Classification** to identify the label of the argumentation unit.

The Relations Prediction task was also subdivided into smaller sub-tasks such as c) **Relation Identification** to detect if a relation, regardless of the label, occurs between two components, and d) **Relation Classification** to predict the kind of relation previously detected. More recently, due to the huge advancements in NLP methods, many approaches started tackling Argument Component Extraction and Relation Prediction as an end-to-end pipeline. The introduction of the transformer architecture BERT [58] led the community to the direction of pre-trained Language Models (PLM) use and fine-tuning technics that

allow researchers to create highly performing models [64, 148]. The latest improvements of Language Models and creation of Large Language Models (LLM) significantly advanced Argument Mining, so that now AM is being explored as a Text-to-Text task [103]. This later approach tackle AM as one single step combining component extraction and relation prediction, generating the argumentation structure as text to be parsed to a graph or a tree.

The automatic detection and extraction of argumentation structures from texts resulted in new valuable applications. For example, the medical domain requires high precision regarding diagnosis or treatment decisions, therefore, in evidence-based medicine researchers explore how argumentation structures can assist in decision making [130]. Some other approaches investigate student argumentation in persuasive essays to achieve better understanding of argumentation structures [187]. Argumentation schemes are also leveraged for detection and reconstruction of implicit argument components [82, 9], as their usage hypothetically facilitates detection process and improves quality of restored components. Moreover, researchers working on implicit knowledge reconstruction in argumentative texts also employ argumentation structures for implicitness detection [19, 18]. With the development of LLMs, generative tasks have been proposed in argumentation, such as counter-argument generation [151].

2.3 Explanatory Argumentation

Explainable Artificial Intelligence aims to make AI systems more transparent by providing explanations for their behaviors [80], which is especially important in critical domains like medicine [10], law [11, 175], and politics [153], where trust and accountability are essential. Despite the performance gains of deep learning, these models are often viewed as "black boxes" [78], making it difficult to understand how they reach conclusions. This opacity raises ethical and legal concerns, as it can lead to biased decisions and hinder trust [11], particularly in sensitive areas like medical diagnosis and legal judgments. XAI addresses this challenge by enhancing interpretability, thus promoting responsible and reliable AI adoption.

2.3.1 Interpretability

To address the need of explainable computational models' behaviour, XAI aims to provide interpretable and robust explanations that clarify how decisions are made. Interpretability can be defined as the degree to which an observer can understand the cause of a decision [134] and the literature identified 3 types of models interpretability [197]. Some approaches and models are interpretable by design [142], such as Linear Models, Logistic Regression or K-Nearests Neighbors (KNN), meaning that their predictions can be explained by the model reasoning itself. However, more recent approaches are mainly based on neural models (i.e., blackbox by nature), thus, they require application of XAI methods to reach a certain level of interpretability. The two other major categories of approaches to tackle the lack of interpretability are perceptive interpretability and interpretability by mathematical structures [197]. Generated Mathematical Structure implies one more layer of cognitive processing to make the prediction interpretable, employing methods such as t-distributed stochastic neighbor embedding (t-SNE), Testing with Concept Activation

Vector (TCAV) [105] or correlation-based Singular Vector Canonical Correlation Analysis (SVCCA) [168]. On the other hand, perceptive interpretability is considered as human understandable as it is (i.e., without any further processing), and it is often presented with visualisation technics. This category is particularly interesting to democratise the use of AI while giving users explanations they can understand without any additional tools. Saliency represents perceptive interpretability that aims at assigning values to input components according to their contributions to the output. The most common saliancy frameworks are Local Interpretable Model-agnostic Explanations (LIME) [174], SHapley Additive exPlanations (SHAP) [122] and DeepLIFT [181]. Other methods of perceptive interpretability, often used in computer vision, named signal methods aim at tracking neuronal systems neurons activation to be able, for example, to higlight the impact of a component in a picture [192, 2]. Some approaches also use attention mecanism [94, 214]. Finally, a high level category identified in the literature as perceptive interpretability is verbal interpretability. In contrast with more abstract or mathematically complex forms of interpretability, verbal explanations can directly express causal relationships or logical statements in natural language. As verbal explanation are the easiest to human perception form of interpretability, they built a path to new research directions and models. Combining verbal interpretability to NLP techniques and argumentation structures, we can move to a new direction of research wich is Explanatory Argumentation.

2.3.2 Explanatory Argumentation

Explanatory Argumentation (EA) or Argumentative XAI refers to the process of providing reasoning with the primary goal to explain why certain facts, observations, or events occur. Explaining a fact or a decision in natural language is not easy, and structuring the argumentation contributes to the difficulty of automating this process. A well-structured explanation nevertheless makes it easier to convey information and makes the explanation more credible, which is necessary in critical fields such as medicine. Unlike persuasive argumentation, which aims at convincing an audience to adopt a particular stance or belief [116], EA focuses on offering the best possible explanation for observed phenomena. This involves constructing explanations that are logically sound, coherent with existing knowledge, and supported by evidence. Argumentation theory and methods is a good solution to help in formulating explanations over argumentation components. Argumentation therefore needs to be formalised and represented so that it can be treated automatically. To model argumentation, Argumentation Frameworks (AF) [62] are created aiming at understanding arguments, dialectical relations and semantics. An Argumentation Frameworks is a way for an agent to manage conflicting information and to draw consequences from it. For instance, Baroni et al. [17] proposed an overview of AF, while Cyras et al. [55] identified three categories of AF which we define in the following.

Abstract Arguments (AA). In the first category, arguments are treated as abstract entities without any internal structure. These frameworks focus on the relationships between arguments, primarily through attack [62] and, occasionally, support [63] relations. The semantics are defined in terms of extensions—sets of arguments that meet specific dialectical conditions, such as conflict-freeness.

Structured Arguments. The second category deals with structured arguments, where arguments are constructed from assumptions [27] and defeasible rules [138, 71]. Unlike abstract arguments, these frameworks consider the internal structure of arguments, focusing on how they are derived. While they still emphasize attacks between arguments, structured methods such as dialectical trees [71] are used to represent and evaluate the arguments.

Hybrid Models Using Abstract Argumentation. The third category combines the strengths of both structured reasoning and AA analysis. In this approach, structured arguments derived from specialized reasoning methods such as case-based reasoning [53, 51], abductive logic programming [205], logical deductions [12], or argument schemes [178], are embedded within AA frameworks.

Argumentation Frameworks based Explanations refer to explanations built by Argumentation framework and can be classified into two categories [55] :

Intrinsic Explanations are built together with the model, making it explainable by design.

Post-hoc Explanations are generated after the prediction and aim to explain the reasoning of it.

Abstract argumentation frameworks are limited in real case scenarios because they require to be grounded on a formal representation of data (i.e., arguments) whereas real cases often are presented through arguments expressed in unstructured form such as natural language.

2.3.3 Explanatory Argumentation in Natural Language

Within the field of Natural Language Processing, the study of Explanatory Argumentation is gaining traction as researchers focus on automating generation and analysis of explanations. The challenge for Argumentative Explanations in natural language resides in the adaptation of existing methods to real cases scenario (i.e., moving from textual data to argumentation frameworks). Some contributions in Explanatory Argumentation combine natural language and AF, mostly in a post-hoc configuration [223, 54] whereas only a few proposed a intrasinc approach [51]. Closer to the NLP community, some contributions started to interest in XAI with approaches focusing on classification [176], Natural Language Inference (NLI) [33] or Natural Language Generation (NLG) [145, 29]. These NLG approaches may generate errors on the veracity of the data, which is problematic for sensitive domains.

2.3.4 Explanatory Argumentation and Medicine

The importance of explanatory argumentation is particularly pronounced in the medical domain, where AI systems (i.e., blackboxes) are increasingly leveraged to support decision-making processes. Whether it is advising treatments, diagnosing conditions, or predicting patient outcomes, AI is required to provide clear and understandable explanations to ensure that medical professionals can trust and validate system's suggestions. In this context, explanations should align with clinical reasoning processes, integrating evidence from medical literature, patient history, and expert guidelines. More specifically, autonomous systems should ideally be based on validated and accepted medical knowledge

sources such as database and ontologies from the medical community. Previous approaches propose verbal explanations for very specific cases such as the diagnosis of pneumonia [40], limited to this pathology which cannot be made domain agnostic. Letham et al. [118] present rule-based medical natural language explanations, which require a lot of upstream expert work and system maintenance, and are very difficult to scale. Although the need of XAI in medicine is already well identified [10], a gap in the literature is noticeable with almost no contributions producing robust natural language explanations (i.e., well-defined structure, integrated medical knowledge, expert validation). Therefore, trustworthy XAI in medicine remains a major challenge that has not yet been addressed.

The intersection of Explanatory Argumentation, NLP, and medicine represents some challenges. By leveraging NLP techniques, it is possible to build systems capable of producing explanations that are both technically sound and accessible to non-expert users. Current research is focused on addressing several challenges in this area, including :

Capturing expert knowledge. How can we integrate clinical expertise into AI models to ensure that generated explanations are aligned with medical best practices ?

Improving interpretability. How can we make explanations clear and understandable to diverse audiences, including clinicians, medical students, and patients ?

Evaluating explanations. Wich metrics and methods should be used to assess the quality and effectiveness of explanations, particularly in life-critical domains like healthcare ?

The goal is to create AI systems that not only perform well on the predition task, but also that are capable of engaging in explanatory dialogues, providing explanations that are coherent, logical, and evidence-based.

2.4 Medical Resources

Medical knowledge forms the backbone of effective decision-making in healthcare. In the context of AI systems, especially those focusing on Argument Mining and Explanatory Argumentation, incorporating accurate, comprehensive, and up-to-date medical knowledge is essential. Due to the hudge variety of applications related to the medical domain and the difficulty to harvest the expert knowledge, data is represented in over different formats.

Structured medical knowledge provides standardized, well-organized information that can be easily integrated into AI systems for reasoning, decision support, and generation tasks. These sources include databases, medical ontologies, classification systems, and terminologies that are foundational to computational healthcare.

2.4.1 Standard Vocabularies

Huge efforts are made in the standardization of medical terminology to numerically identify medical concepts. A standardised, widely adopted system is SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) [60] [2]. SNOMED CT is a compre-

2. https://www.snomed.org/value-of-snomedct

hensive clinical terminology used across the globe. SNOMED CT's hierarchical structure allows precise encoding of medical concepts. The International Classification of Diseases (ICD) [163], managed by the World Health Organization, is another significant classification system. It categorizes diseases and health conditions, enabling consistent documentation, research, and decision-making across healthcare systems. ICD codes are often employed in AI systems to annotate and categorize medical arguments, ensuring alignment with globally recognized diagnostic standards. Additionally, such resources as LOINC (Logical Observation Identifiers Names and Codes) [131] are crucial for standardizing clinical measurements and laboratory tests. By providing consistent identifiers, LOINC facilitates integration of diverse data sources, enhancing interpretability and consistency of AI-driven models. The Unified Medical Language System (UMLS) [25] integrates multiple biomedical terminologies into a unified framework. UMLS is widely used for mapping medical concepts across different vocabularies, making it a crucial tool in NLP applications that require standardization of clinical terms across different systems.

2.4.2 Medical Knowledge Bases and Ontologies

Ontologies and knowledge bases provide rich, structured information about medical concepts, their relationships, and clinical guidelines. These resources are often used to enhance reasoning capabilities of AI models, allowing them to incorporate expert knowledge into decision-making. The Human Phenotype Ontology (HPO) [110] provides a standardized vocabulary of phenotypic abnormalities encountered in human disease organized in a knowledge base. The main difference with other vocabularies is that HPO provide diseases and associated medical terms to enhance precision of phenotype-driven diagnostic tools. It is commonly used in rare disease diagnostics and gene-disease association studies. Another controlled and hierarchically-organized vocabulary and knowledge base produced by the National Library of Medicine is Medical Subject Headings (MeSH) [88] thesaurus. MeSH is used for indexing, cataloging, and searching of biomedical and health-related information. Focusing on clinical drugs, DrugBank [109] and RxNorm [121] propose comprehensive databases containing informations on drugs, their standardizes names, their mechanisms, interactions, and pharmacological properties.

2.4.3 Natural Language Resources for Medicine

While structured data provide standardization and consistency, a significant amount of medical knowledge is still embedded in unstructured formats, such as clinical notes, research articles or textbooks. As Natural Language Processing techniques enable extraction of relevant information from unstructured sources, the community interest about unstructured data is growing. Biomedical literature databases initially made for medical experts and researchers became one of the most popular sources of trustworthy unstructured data to make NLP algorithms learn about medicine and related domains. BioBERT [117] model pretrained on biomedical text or ClinicalBERT [92], fine-tuned on clinical narratives, are good example of the usage of unstructured medical knowledge. More recently, BioGPT [123], a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature such as databases like PubMed [3], which indexes thousands of research articles,

3. https ://pubmed.ncbi.nlm.nih.gov/

clinical trials, and case studies. Other data focusing on patients are Electronic Health Records (EHRs) that contains vast amounts of patients data, including notes, lab results, and medical histories. NLP methods are often employed to extract actionable insights from these records, converting unstructured texts into structured data that can be used for decision support and explainability. The main concern about such documents is related to privacy, thus, it is often required to obtain clinician and/or patient approval to make this data available to use or to anonymize. Similar to EHRs, Clinical Notes and Radiology reports are rich in detailed patient information, including symptom descriptions, diagnostic reasoning, and treatment plans. They are often written in narrative form, containing critical observations that are key to diagnostic decision-making. AI systems, particularly those focused on explainability, utilize NLP to process these unstructured texts, generating human-readable explanations.

2.4.4 Explanatory Medical Ressources

When focusing specifically on medical data with expert explanations, very few resources may be retrieved. These resources are needed to develop automatic explanation generation systems based on language models. They are also useful for empirical analysis of the structure of explanations. Most of the resources introduced above focus on clinical evidence or diagnoses, while explanatory data, particularly those which justifies a medical decision in detail, are very rare. Most resources focus on explaining answers to student exams. This is the case of MIR Asturias [4], a resource containing questions taken from previous years at MIR (Medical Intern Resident) exams. The MIR exam is the test required for doctors to gain access to a position as a specialist in training in the Spanish National Health System. As the documents are not structured, they have to be processed to extract the different sections of the document (e.g., questions, answers, explanations, etc.). In the same area, Casimedicos [5] also presents MIR questions with explanations written by expert volunteers. Finally, the SAEI (Andalusian Society of Infectious Diseases) resource presents clinical cases on infectious diseases, with comprehensive differential diagnosis explanations. SAEI's detailed differential diagnoses make it particularly valuable for models that require fine-grained expert knowledge. These resources represent some of the few available data that contain expert-generated medical explanations, higly valuable to generate and assess explanations automatically.

4. https://www.curso-mir.com/
5. https://www.casimedicos.com/

CHAPTER 3

Natural Language Explanation Generation

This chapter describes how expert knowledge can be retrieved from natural language text and injected into the explanation generation process. More precisely, I present how to automatically extract information relevant to a medical diagnosis in order to align it with expert knowledge bases. By focusing on features that are decisive for diagnosing a disease, two trends have been identified according to their complexity of interpretation. First, a symptomatic approach based on patient signs and symptoms already provides a good basis to understand and explain a given diagnosis. Then, medical findings, such as vital signs and medical measurement, are converted using a newly introduced dataset that defines the boundaries and terminology of these findings. Once these features have been detected, converted and aligned with a recognised knowledge base, I use them to generate reliable and grounded explanations. This chapter includes the work published in the Journal of Biomedical Semantics (2024)[140], which is built on the initial publication presented at the International Conference on Agents and Artificial Intelligence (ICAART-2023)[126].

In the medical field, the ability to explain diagnostic decisions or recommendations can enhance medical training, improve trustworthiness of AI systems, and ultimately support better clinical outcomes [10]. Therefore, automatically generating natural language explanations has recently gained attention, and applying this to the medical domain could be highly beneficial for many stakeholders, particularly medical students in educational settings. As a challenging starting point, being able to generate explanations for medical exams could be valuable in helping students understand clinical cases and could improve their critical thinking. In this chapter, I explore the potential of generating explanations for MedQA [96], which presents clinical case exams (comprising a case description, a question, and a set of possible answers with only one correct option) from a symptom-based perspective. Specifically, my goal is to justify why a given diagnosis is correct and why the alternative diagnoses are incorrect.

Given the critical nature of the medical domain, a primary criterion for automatically generating explanations is that they are grounded in verified knowledge, ensuring the provision of accurate and accepted evidence. This requirement presents challenges for generative models and large language models, as discussed in Section 2.1 due to their difficulties in controlling the generation process, including issues like hallucinations and biases. As a result, retrieving information from reliable sources and aligning it with widely accepted knowledge bases becomes a necessary step before generating explanations. This process involves detecting medical information from natural language text and aligning it with structured knowledge. Therefore, I propose a new transformer-based pipeline named SYMEXP to automatically extract medical entities (i.e., layperson symptoms) from clinical cases and align them with a medical ontology. Clinical cases often include basic patient information such as age, gender, symptoms that justify the visit, and possibly vital signs, test results, or medical measurement. In this work, I only focus on diagnosis type questions (i.e., "Which of the following is the most likely diagnosis ?"), where the objective is not to predict the diagnosis but to explain it. The correct and incorrect answers are already known as they are provided by medical expert within the dataset. To achieve this, it is necessary to identify the relevant information from the clinical case and align it with trusted medical ontologies. For instance, the HPO [110] contains valuable information about diseases and their associated symptoms. More specifically, I retrieve the answers of the medical exam withing the ontology and use the aligned detected entities from the clinical case to support or discard diagnoses.

While medical ontologies are generally reliable and contain a large amount of information, they are often specialized and limited to specific types of knowledge, making it difficult to convert certain clinical case information without prior interpretation. During my analysis of clinical cases, I discovered that medical findings could be interpreted as symptoms when they exceed normal values, but most databases lack information about abnormal values and conversion vocabularies. This is why I propose, in this contribution, a finding-to-medical term conversion system using a newly introduced database, verified by a medical expert.

Finally, as discussed in Section 2.3, explanation mechanisms are complex. Delivering explanations in educational contexts requires a different approach than in other scenarios, such as persuasive argumentation in conflict resolution. In this work, I explore how to craft explanations suitable for educational purposes by applying argumentation to enhance template-based explanations [98].

This Chapter is organised as follows : in Section 3.1 I introduce existing medical knowledge and vocabularies together with two new resources used to ground explanations in knowledge. In Section 3.2 I present the proposed architecture, while in Section 3.3 I describe the system implementation, results and error analysis. In Section 3.4 I discuss the argumentation patterns selected to generate natural language explanations, and I propose an overview of the related contributions in Section 3.5. Finally, I raise some conclusions in Section 3.6.

3.1 Ressources

The availability of high-quality clinical data, both in the form of annotated clinical cases and medical knowledge bases, is crucial for the success of this study. In this section, I present the foundational data used to generate grounded natural language explanations. The first part introduces the newly created dataset of 314 clinical cases, annotated medical entities and layperson symptoms, developed through a detailed annotation process using labels extracted from the UMLS [25] medical metathesaurus. This dataset is employed for training a Named Entity Recognition (NER) model, which is an essential step in automatically identifying relevant medical information in clinical cases. The second part details the creation and expert validation of a medical findings database, which is designed to interpret 100 findings (i.e., health measurements, observations and test results). These findings, together with detected symptoms are subsequently integrated into the generated explanations by converting them into relevant medical concepts from the Human Phenotype Ontology. Both resources play complementary roles in the pipeline. The dataset is crucial for training the NER models, while the findings database is used to convert medical findings to standardized medical terms to ensure consistency and accuracy in the generated explanations.

3.1.1 Medical entities dataset

To train and evaluate our language model based approach to identify and extract medical entities from natural language clinical cases, we rely on the existing MedQA dataset [96]. MedQA is a free-form multiple-choice open question answering dataset for solving medical problems collected from the professional medical board exams. More specifically, the questions and their associated answers were collected from the National Medical Board Examination in the USA (United States Medical Licensing Examination or USMLE), Mainland China (MCMLE), and Taiwan (TWMLE). Each question is preceded by a clinical case that introduces the patient and is followed by a set of options, only one of which is correct. Additionally, MedQA provides the correct answer, key words, and pieces of text relevant to the answer, as well as supporting evidence from the textbook "Harrison's Principles of Internal Medicine". Two examples extracted from the original contribution are available in Figure 3.1. In this work, we only focus on the clinical cases and the questions in English (i.e., *USMLE*). In total, the MedQA-USMLE dataset consists of 12,723 unique questions on different topics, ranging from questions like "Which of the following symptoms belongs to schizophrenia ?" to questions about the most probable diagnosis, treatment or outcomes for a certain clinical case. This latter group, which aims to test medical residents on making accurate diagnoses, is particularly suitable for generating explanations from a symptom-

Question	A 27-year-old male presents to urgent care complaining of pain with urination. He reports that the pain started 3 days ago. He has never experienced these symptoms before. He *denies gross hematuria or pelvic pain*. He is sexually active with his girlfriend, and they consistently use condoms. When asked about recent travel, he admits to recently returning from a boys' trip" in Cancun where he had *unprotected sex* 1 night with a girl he met at a bar. The patients medical history includes type I diabetes that is controlled with an insulin pump. His mother has rheumatoid arthritis. The patients temperature is 99 F (37.2 C), blood pressure is 112/74 mmHg, and pulse is 81/min. On physical examination, there are no lesions of the penis or other body rashes. No costovertebral tenderness is appreciated. A urinalysis reveals no blood, glucose, ketones, or proteins but is *positive for leukocyte esterase*. A urine microscopic evaluation shows a *moderate number of white blood cells* but no casts or crystals. A urine culture is negative. Which of the following is the most likely cause for the patient's symptoms?
Options	**A: Chlamydia trachomatis**, B: Systemic lupus erythematosus, C: Mycobacterium tuberculosis, D: Treponema pallidum
Evidence	At least one-third of male patients with *C. trachomatis* urethral infection have *no evident signs or symptoms of urethritis*. ... Such patients generally have *pyuria* ..., a *positive leukocyte esterase test*, ...
Question	A 57-year-old man presents to his primary care physician with a 2-month history of *right upper and lower extremity weakness*. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had *increasing difficulty* with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his *family have had musculoskeletal problems*. His right upper extremity shows *forearm atrophy* and *depressed reflexes* while his right lower extremity is *hypertonic with a positive Babinski sign*. Which of the following is most likely associated with the cause of this patients symptoms?
Options	A: HLA-B8 haplotype, B: HLA-DR2 haplotype, **C: Mutation in SOD1**, D: Mutation in SMN1, E: Viral infection
Evidence	1. The manifestations of ALS ... *insidiously* developing *asymmetric weakness*, usually first evident distally in one of the limbs. 2. ... *hyperactivity of the muscle-stretch reflexes (tendon jerks)* and, often, *spastic resistance to passive movements* ... 3. *Familial ALS (FALS)*... clinically indistinguishable from sporadic ALS... Genetic studies have identified mutations in multiple genes, including cytosolic enzyme *SOD1*...

Figure 3.1 – Two examples of the MedQA original contribution (Jin et al., 2021). The correct answer among options is marked in bold font. Key words in the question and evidence text to help answer the questions are highlighted in italic font. Evidence for both examples are from the textbook "Harrison's Principles of Internal Medicine".

based perspective. After filtering the MedQA-USMLE, 314 unique clinical cases associated with the list of possible diagnoses where extracted to constitute the MedQA-USMLE-Symp dataset. These clinical cases are student training exams formulated as question answering documents always composed by the patient description, a question i.e. about the patient diagnosis in this scenario, and, a set of options with only one (known) correct answer. An example of a clinical case extracted from the filtered dataset is available in Example 3.1.1.

Exemple 3.1.1 – **Clinical Case :** A 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days. The pain is worse with inspiration, and she feels she cannot take deep breaths. She has not had shortness of breath, palpitations, or nausea. She had an upper respiratory tract infection 10 days ago and took an over-the-counter cough suppressant and decongestant and acetaminophen. Her temperature is 37.2°C (98.9°F), pulse is 90/min, and blood pressure is 122/70 mm Hg. The lungs are clear to auscultation. S1 and S2 are normal. A rub is heard during systole. There is no peripheral edema. An ECG shows normal sinus rhythm and diffuse, upwardly concave ST-segment elevation and PR-segment depression in leads II, III, and aVF.
Question : Which of the following is the most likely diagnosis ?
Answers : [Acute pericarditis, Aortic dissection, Gastroesophageal reflux disease, Myocardial infarction, Peptic ulcer disease, Pulmonary embolism, Unstable angina pectoris]
Correct Answer : Acute pericarditis

The clinical case contains much information about the patient state and generally contextualize the situation, introducing the patient sex and age, potential symptoms or events triggering the clinical visit and, sometimes, exposing the patient vital signs, direct observations or medical measurements.

Annotation of the MedQA-USMLE-Symp Clinical Cases. The aim of this dataset is to help language models to identify medical informations, therefore it is needed to annotate the MedQA-USMLE with an extra layer of medical entities. To keep the annotation consistent

with standard textual annotations in the medical domain [34, 5, 139], the proposed annotations are based on an existing vocabulary from UMLS [25], more specifically UMLS Semantic Types. Among the extensive variety of labels offered by the UMLS Semantic Types, we chose these specific ones to suit our data and diagnosis based questions : *Sign or Symptom*, *Finding*, *No Symptom Occurrence*, *Population Group*, *Age Group*, *Location* and *Temporal Concept*. Our selection of these seven labels was informed by consultations with medical experts and determined by their explanatory power in our specific context. We mainly focus on the *Sign or Symptom* and *Findings* labels as they offer critical insights for the diagnosing task. However, in anticipation to the development of a more fine-grained approach as future research, we have conducted a comprehensive annotation across all seven labels, allowing for future reuse and exploration of these additional dimensions. It is important to note that annotating layperson symptoms and medical findings is a subjective task and annotators could interpret differently the same patient statement, therefore, in order to obtain high quality annotations we created a set of guidelines [1] to help them. Among the annotations conflicts, a well identified difficulty concerned the boundaries of entities, making us specify for most of the labels the annotation details with rules with examples as describe in the following paragraphs :

Overall instructions. As we aim to have a granular annotation, entities (e.g., *Sign or Symptom*) have to be as small as possible and need to be separated when multiple occurrences appear in a single sentence. As shown in Example 3.1.2, symptoms (i.e., entities in bold) are self-contained, separated in multiple components and do not include the punctuation in the components boundaries. Example 3.1.3 shows a case of intricated components where separating them will waste important informations therefore, components are annotated as one bigger symptom.

Exemple 3.1.2 – A 45-year-old woman has a 2-week history of increased **anxiety**, **abdominal discomfort**, **irritability**, and **difficulty concentrating**; she was robbed at knifepoint in a parking lot 3 weeks ago.

Exemple 3.1.3 – He also has a 1-year history of **joint and muscle pain** in his calves and a 1-month history of intermittent, diffuse **abdominal pain**.

Sign or Symptom / No Occurence of Sign or Symptom. Following the definition of UMLS we adapted it for the Sign or Symptom component described as "*an observable disease or condition in the clinical case including the symptoms from the past, the symptoms not related to the patient (family antecedents) and considering all kinds of diseases (physical, mental, ...)*". As shown in Examples 3.1.4, 3.1.5 and 3.1.6, symptoms can appear as expert medical vocabulary (e.g., "insomnia"), disease names (e.g., "Crohn disease") or layperson descriptions of symptoms (e.g., "difficulty concentrating").

Exemple 3.1.4 – A 45-year-old woman has a 2-week history of increased **anxiety**, **abdominal discomfort**, **irritability**, and **difficulty concentrating**; she was robbed at knifepoint in a parking lot 3 weeks ago.

Exemple 3.1.5 – She had an extensive abdominal operation 5 years ago for **Crohn disease**.

1. https://github.com/Wimmics/MEDQA-USMLE-Symp

Exemple 3.1.6 – She says that despite the test results, she has had **anxiety**, **insomnia**, and a **preoccupation** with **cancer** since noticing the **lump**.

Quantifiers are often found at the boundary of an entity and when defining a symptom should not been annotated (e.g., when we find "diffuse abdominal pain", we only annotate "abdominal pain" as in exemple 3.1.3). The label *No Occurence of Sign or Symptom* respect the same rules than *Sign or Symtom* except that this symptom is not observed in the clinical case or is resolved as in the examples 3.1.7, 3.1.8 and 3.1.9.

Exemple 3.1.7 – He has not had **chest pain** or **shortness of breath**.

Exemple 3.1.8 – On cardiac examination, no **murmurs** or **gallops** are heard.

Exemple 3.1.9 – Her **dysuria** has resolved.

The labels *Sign or Symptom* and *No Symptom Occurrence* are associated only to the text snippet defining the symptom in a sentence. The boundaries of these components may include the body part affected by the symptoms when they are strongly related as shown in Exemple 3.1.3. This decision is driven by the fact that annotating "pain" alone will loose too much information about the diagnosis. When a symptom is related to multiple body parts, we annotate the *locations* separately as described further.

Findings. Findings are the aggregation of vital signs, clinical measurements and laboratory test. They consist of informations discovered by direct observation or measurement of an organism's attribute or condition. Findings often (but not always) contain a numerical value or a quantitative indicator that need to be interpreted by a medical expert. They appears within many patterns but contains in most of the case three parts (i.e., Finding Name, Unit, Value). For instance, in "Her *temperature is 39.3°C*" the Findings Name is "Temperature", Unit is "°C" and Value is "39.3". An example of findings annotation from the guidelines is shown in Example 3.1.10. It is noticeable that some findings, mostly tests, does not have numerical values but rather boolean values "test of the stool for occult blood is **positive**" and still need to be further interpreted by a medical expert to be useful in the diagnostic prediction or explanation.

Exemple 3.1.10 – Her **temperature is 37.2°C (98.9°F)**, **pulse is 90/min**, and **blood pressure is 122/70 mm Hg**, **test of the stool for occult blood is positive** Her **temperature is 39.3°C (102.8°F)**, **pulse is 104/min**, **respirations are 24/min**, and **blood pressure is 135/88 mm Hg**.

Location. To provide a complete annotation and grasp all the useful informations for further works we annotated the locations in the human body relying on the three following UMLS semantic network labels **Body Location or Region**, **Body Part, Organ, or Organ Component** and **Body Space or Junction**. Example 3.1.11 differentiate the symptom from the body part where it occurs, while in previous example 3.1.3 the symptom pain is strongly related to the associated body part, i.e., abdominal, therefore we annotate both of them as a symptom. As a criterion to annotate coherently the symptoms, is assumed that if the symptom semantics changes when removing the location, then we include the location in the component.

Exemple 3.1.11 – A grade 3/6 harsh systolic ejection murmur is heard at the **left upper sternal border**.

Temporal Concept. Clinical cases concerning diseases often refer to the past of the patient or mention disease's duration. Therefore, we decide to annotate the *Temporal Concept* from UMLS semantic network, including the duration, references to patient's history or changes in symptoms is used to tag time-related information, including duration and time intervals. The temporal concept component includes the time descriptors as "**for** 3 days" or "10 days **ago**" and the scales (day, weeks, ...) as in Example 3.1.12 and 3.1.13. We also consider the adverbs like "sudden" in Example 3.1.14 as temporal concept, because it can play a relevant role in the diagnosis.

Exemple 3.1.12 – **For 8 weeks**, a 52-year-old man with a **5-year history** of type 2 diabetes mellitus has had deep burning pain in the ball of his right foot and big toe when the foot is raised above chest concentration.

Exemple 3.1.13 – A 56-year-old man has had the painful weeping rash shown **for 2 days**.

Exemple 3.1.14 – A 4-year-old girl has the **sudden** onset of abdominal pain and vomiting.

Population and Age Groups. Finally, given their main role in the diagnosis, we also annotated the gender and age components in our clinical cases with the *Age Group* and *Population Group* labels. These labels are the easiest to identify due to the document architecture, describing the clinical case by introducing the patient population group and age as shown respectively in examples 3.1.15 and 3.1.16. The component also includes the "year-old" specification because of the possibility to encounter age measurement in months or weeks, especially for newborns.

Exemple 3.1.15 – A **37-year-old** woman comes to the physician because of shortness of breath for 3 months.

Exemple 3.1.16 – A 16-year-old **boy** is admitted to the emergency department because of a knife wound to the left side of his chest.

Guidelines were provided together with two different and representative examples, available with colored entities in Appendice A.1.

Edge cases. As for the symptoms, in some clinical cases we can encounter some enmeshed components as illustrated in Example 3.1.17. Here both "red" and "ulcerated" refer to the "lesion", but they cannot be split in two self-contained symptoms' component. The solution we choose is to include both components in one as in the example.

Exemple 3.1.17 – Physical examination shows a 6-mm, **red, ulcerated lesion** with heaped borders.

Example 3.1.18 shows a tricky example of a symptom triggered by a specific event. This action is potentially relevant, so it should be included it in the annotated component.

Exemple 3.1.18 – The **symptoms are moderately exacerbated by exertion**.

Example 3.1.19 shows a clinical case where symptoms are put forward through common words as "stops working" or "lot of energy for work". These layperson symptoms are considered as hard to be detected automatically.

Exemple 3.1.19 – He has had a **lot of energy for work** but often is **distracted** to the point that he does not complete assigned tasks. He frequently **stops working** on his own tasks to attempt to develop greater efficiency in his shop. He states that he is **delighted with his newfound energy** and reports that he now **needs only 4 hours of sleep nightly**.

Example 3.1.20 represents a comparison between two body parts which are enmeshed. Since they cannot be divided as "left pupil is larger than the right" and "left pupil reacts sluggishly to light", we consider it as a single long symptom. This example respects the overall instruction asking fos the smallest possible and self-contained components.

Exemple 3.1.20 – The **left pupil is larger than the right and reacts sluggishly to light**.

Mental diseases, exemplified in Example 3.1.21, were identified as harder to annotate because of the peculiarity of their symptoms, so we follow and match as much as possible the symptoms available from the external knowledge database (i.e., the HPO) to annotate these examples.

Exemple 3.1.21 – He has been your patient since early adolescence, and he has a **history of truancy**, **shoplifting**, and two **attempts to run away from home**. He **dropped out of high school** in his senior year. He was **fired from his most recent job** because he **threatened a coworker with a hammer**.

Finally, in some patient introduction, the gender is revealed later and using the gender pronouns as in example 3.1.22

Exemple 3.1.22 – After the seizure, she was confused and had difficulty thinking of some words.

Annotation Settings To address the annotation process of the MedQA-USMLE-Symp dataset, we conducted a semi-automatic annotation using the UMLS database. Since we decided to annotate with labels inspired by UMLS, each clinical case was processed through the UMLS system, which provided all detected entities along with their Concept Unique Identifiers (CUI) and semantic types. The semantic type was then used to disambiguate the entities and generate pre-annotated files. After the definition of the detailed annotation guidelines [2] in collaboration with clinical doctors, three annotators with a background in computational linguistics carried out the annotation of the pre-annotated 314 clinical cases.

To assist the annotators, we initiated the process with a pre-automatic annotation using the Brat visualization tool [188] together with QuickUMLS [3], a tool that leverages UMLS data and pre-annotate them into the Brat visualization tool. This automatic annotation was then manually corrected and completed by annotators using the HPO for diseases and relevant symptoms. To ensure the reliability of the annotation task, the inter-annotator agreement (IAA) has been calculated on an unseen shared subset of 10 clinical cases annotated by three annotators, obtaining a Fleiss' kappa [69] of 0.70 for all of the annotated labels, 0.61 for *Sign or Symptom*, 0.94 for *Location*, 0.71 for *Population Group*, 0.66 for *Finding*, 0.96 for *Age Group* and 0.96 for *No Symptoms Occurrence*. We can see

2. https://github.com/Wimmics/MEDQA-USMLE-Symp
3. https://ir.cs.georgetown.edu/resources/quick-umls.html

a substantial agreement for *Sign or Symptom*, *Finding* and *Population Group*, and an almost perfect agreement for *Location*, *Age Group* and *No Symptoms Occurrence*. Table 3.1 reports on the statistics of the final dataset, named MedQA-USMLE-Symp. [4] The accuracy of the annotations provided by the three annotators has been validated by a clinical doctor. Of the seven entity labels, only three contain medical vocabulary (*Sign or Symptom*, *Finding*, and *No Symptom Occurrence*) and they have been evaluated by this expert. More specifically, we randomly sampled 10% of the data (i.e., 30 cases) and we asked the clinician to verify whether the entity was correctly labeled and whether there were any missing or extra words. The results of the validation showed that 98% of the data was labeled correctly. Errors were distributed randomly, being the majority of them annotation errors with missing/extra letters from the labels (e.g., "itchy rash" annotated as "tchy rash" or "generalized joint pain" annotated as "eneralized joint pain"). Less than 2% of the instances were evaluated as incorrectly labeled (e.g., a *Finding* that was labeled as a *Sign or Symptom* or vice versa).

We also manually annotated our test set with HPO terms equivalent to the annotated symptoms in the clinical cases. Out of the 162 symptoms identified, 88 were aligned with the concepts in the ontology in order to evaluate the alignment task define further in Section 3.2. These annotations are available within the project repository.

TABLE 3.1 – Statistics of the MedQA-USMLE-Symp dataset.

Label	# Entities
Sign or Symptom	1579
Finding	1169
Temporal Concept	567
Location	498
Population Group	364
Age Group	304
No Symptom Occurrence	264

Knowledge Base of Diseases and Relevant Symptoms. In the previous step, we focused on identifying and annotating medical entities present within the clinical documents, expressed in natural language. However, this alone is not sufficient to explain the correctness or incorrectness of diagnoses. To generate automatic explanations of diagnoses from a symptom-based perspective, it is also necessary to identify diseases and their associated symptoms using existing knowledge bases. To achieve this, we utilize the HPO knowledge base to determine whether a symptom detected in the previous step is relevant to a given disease (i.e., one of the options in the clinical case question-answering). Specifically, we employ the HPO to retrieve : *(i)* each clinical case diagnosis proposed as an option for the question "Which of the following is the most likely diagnosis ?" and *(ii)* the symptoms (referred to as *terms* in the HPO) associated with each diagnosis. As previously described in Section 2.4, the HPO is a comprehensive, structured vocabulary for describing phenotypic abnormalities encountered in human diseases, frequently used in genetic and rare disease research. We chose to use the HPO over other resources such as SNOMED CT because

4. https://github.com/Wimmics/MEDQA-USMLE-Symp

the HPO includes additional valuable information, such as the frequency [5] of symptom occurrence. This information is defined in collaboration with ORPHA [6] and provides a occurrence rate of symptoms such as :

— Excluded (0%) ;
— Very rare (1-4%) ;
— Occasional (5-29%) ;
— Frequent (30-79%) ;
— Very frequent (99-80%).
— Obligate (100%) ;

The occurrence rates are particularly valuable in our application scenario for generating fine-grained explanations that encourage critical thinking of medical students. The HPO integrates multiple sources of symptoms, including ORPHA and OMIM [7]. The HPO ontology is extensive and contains also structural informations such as hierarchical links between symptoms (e.g., symptom S2 subclass of symptom S1), genes or definitions. Since the HPO knowledge base is structured, we developed an automatic script to retrieve automatically the ontology terms associated to each clinical cases options (i.e., diseases). Concerning the options, a first batch were identified using the HPO search engine through exact term matching whereas the rest of options were retrieved manually for the test set of the MedQA-USMLE-Symp dataset. It's worth noting that some diseases could not be found in the HPO, usually because they were either too abstract or too specific (e.g., "seizure disorder," "primary hypersomnia"). The missing diagnosis alignment impact on explanation generation is discussed further in Section 3.4.

3.1.2 Medical Findings database

While annotating the MedQA-USMLE-Symp dataset, the "Findings" entities appeared to be a valuable source of information about the patient condition and often the key to distinct two diagnoses. As presented in the previous section, *finding* is a group that includes patient measurements, vital signs, test and analysis results or observation. If findings are already mentioned in some biomedical vocabularies such as LOINC, the majority of them does not contain findings but the interpretation of them (e.g., *fever* wich is the result of a high "Temperature", specifically more than 37.3°C). Moreover, a consequent number of HPO terms associated to clinical cases options are also the consequences of a finding and can be inferred from detected findings in the MedQA-USMLE-Symp dataset to enhance the explanatory power of this approach. As observed and discussed with medical experts, findings interpretation involves two steps : i) identifying the normal boundaries, and ii) linking them to the appropriate medical term. This complexity drives the need to build a specific database that encompasses the most frequent occurring medical findings within our clinical environment. These data will be used to automatically convert detected findings to medical terms (e.g., temperature is 39°C → fever).

Database description. The medical findings database is designed to facilitate the interpretation of medical test results, converting raw findings, such as "Platelet count is 100,000

5. https://hpo.jax.org/app/browse/term/HP:0040279
6. https://www.orpha.net/consor/cgi-bin/index.php?lng=FR
7. https://www.omim.org/

TABLE 3.2 – Low findings values and their corresponding LOINC codes, medical terms, HPO codes, ICD-10 codes and SNOMED CT codes for glucose and platelet count.

Finding	LOINC	Value	Medical term	HPO	ICD-10	SNOMED
Glucose (Glu)	97510-2	70 mg/dL	Hypoglycemia	HP :0001943	E16.2	271327008
Platelet count	74775-8	150000 mcL	Thrombocytopenia	HP :0001873	D69.6	74576004

TABLE 3.3 – High findings values and their corresponding LOINC codes, medical terms, HPO codes, ICD-10 codes and SNOMED CT codes for glucose and platelet count.

Finding	LOINC	Value	Medical term	HPO	ICD-10	SNOMED
Glucose (Glu)	97510-2	99 mg/dL	Hyperglycemia	HP :0003074	R73.9	-
Platelet count	74775-8	450000 mcL	Thrombocytosis	HP :0001894	D75.83	6631009

platelets per microliter of blood", into equivalent medical terms, in this case, "Thrombocytopenia". To this goal, it is necessary to determine whether a given value is classified as "high" or "low" with respect to its normal values, and subsequently associate it to a relevant medical term. In this study, we define "normal values" as those provided by known medical sources [8] and from the MED-USMLE tests themselves [96], bearing in mind that these values are simplifications and do not account for different ethnicity, gender-specific findings, or age-related variations. In order to ensure the comprehensiveness of the database, to foster future reuse of the resource and to maintain compatibility with existing systems, we have also associated each medical term with corresponding medical codes from the International Classification of Diseases version 10 (ICD-10), the HPO, the international health terminology standard SNOMED CT (July 2023 release) and, the findings names are associated with their LOINC codes. As introduced in Section 2.4, the ICD is a globally recognized system for categorizing diseases and other health conditions, maintained by the World Health Organization (WHO). The SNOMED CT is a comprehensive, multilingual clinical terminology system that covers a wide range of medical concepts, including diseases, symptoms, and procedures. Finally, since LOINC proposes a standard vocabulary for findings, we recently added LOINC codes, obtained from the webapp [9] to the database. To obtain LOINC codes we used the SearchLOINC version 2.78 and retrieved automatically the first result if exist. A representative example of the final database can be found in Table 3.2 for low values, and Table 3.3 for high values. As we convert findings into medical terms for each boundary, appear that some of the terms could not be found in some of the vocabularies due to their symptoms-centered architecture. Missing codes does not remain a problem due to the existence of unified medical encoding systems such as UMLS.

Semi-automatic database creation. Developing a new knowledge resource specifically tailored for our requirements in the medical domain presents a considerable challenge, particularly given the significant manual effort and human involvement necessary to conceive, collect, align, and verify the data. Moreover, obtaining expert assistance in the medical field can be difficult due to the demanding nature of the work and the workload of medical professionals. Consequently, we propose a semi-automatic method for generating a data-

8. https://emedicine.medscape.com/article/2172316-overview
9. https://loinc.org/search

base by harnessing the capabilities of state-of-the-art generative large language models, such as ChatGPT [30, 154], which are pre-trained on huge amounts of text, including medical documents. As these LLMs proved their performances without being specially trained for specific tasks or domain it appear to be a perfect tool to draft a new database, using the accumulated knowledge from the training literature. To automatize the creation, we constrain the language model to generate knowledge in a tabular format expressed as free text (i.e., Markdown format) as the following example :

```
| Finding    | Low reference value    | High reference value  |
| [FINDING] | ?                       | ?                      |
```

This free text is subsequently parsed using regular expressions, allowing for the extraction of structured data to be incorporated into the database directly. This database is then refined and verified by domain experts as explained below. It is important to notice that applying a semi-automatic approach to fill in the first and basic version of the database already significantly reduces the manual effort. To assess this semi-automatic approach we addressed both : i) an automated evaluation through comparison with an existing database, and ii) a human evaluation involving a medical expert for correction and validation of the database.

Automated evaluation : If no existing database or knowledge base provide an alignment of findings boundaries and their interpretation in medical terms (i.e., temperature is considered high when above 37.3°C and can be converted to the medical term fever), some of them provide non-structured studies over laboratory reference ranges. Therefore, a first automatic evaluation of our semi-automatic approach is based on this manually restructured database of laboratory reference ranges for healthy adults [10], in order to see if such Large Language Model would be useful for assisting the database creation. This database provides reference ranges for various categories such as Electrolytes, Hematology, Lipids, Acid base, Gastrointestinal function, Cardiac enzymes, Hormones, Vitamins, Tumor markers, and Miscellaneous, but does not specify the medical terms associated with these values. To evaluate our method, we used in the same condition the generative language model several times on the Electrolytes category as the gold standard of our test set of 22 findings and 44 boundaries. Since all the values are numerical we compute the accuracy on the mean value of the multiples runs for each finding and compare with the gold. This proxy score gave us an understanding of the efficacy of LLMs in generating factual data in the medical domain, particularly medical findings related values. This evaluation also assesses whether this method could be used "on the fly" to predict detected elements not present in the final database. It is important to highlight that the values are not strict and that they will strongly depend on the patient but to make a first simplified approach we established a prediction threshold at 20% around the gold value. For example the high reference value for Zinc is 100 µmol/L, so accuracy will be 1 for a prediction of 100 µmol/L, 0.8 for a prediction of 80 or 120 µmol/L and 0 if the prediction is above 120 or under 80 µmol/L. Using the ChatGPT-3 language model, our method achieves a model accuracy of 78% and 80% for low and high values, respectively, with a mean based on five predictions for Electrolytes. To obtain the prediction, 5 instances of the LLM were run and a vote for the majority

10. https://emedicine.medscape.com/article/2172316-overview

TABLE 3.4 – Comparison of version 3.5 and 4 of ChatGPT language models with varying runs and threshold settings, illustrating the impact on low and high accuracy metrics.

LM	Run	Threshold	Low acc.	High acc.
v-3.5	3	20%	0.64	0.73
	3	50%	0.79	0.82
	5	20%	**0.64**	**0.79**
	5	50%	0.78	**0.84**
v-4	3	20%	0.63	0.72
	3	50%	0.8	0.82
	5	20%	0.63	0.73
	5	50%	**0.8**	0.81

selected the answer (i.e., 3 instances proposed 100 and 2 instances proposed 80 make the system select 100). We processed an error analysis and looked at the LLM predictions and it showed that some findings are not known by the LLM but most of the incorrect predictions occurs due to the gray zones around gold values. This gray zone is the fact that the same finding will be interpreted differently according to the patient (gender, ethnicity, age) and combine with the potential units where the finding could be expressed with, lead to most of the LLM errors. However, the final database was meticulously validated by an expert (medical doctor).

We experimented with both version 3.5 and 4 of the ChatGPT LLM for the semi-automatic database creation. Observing the results in Table 3.4, we found that both ChatGPT-3 and ChatGPT-4 showed good performance in generating boundaries values for medical findings, suggesting that these models can be reliably employed later in the proposed pipeline. It is worth noticing that the number of run predictions used to calculate the average has no impact on the result. To get some insights on the generated data, as visualized in Figure 3.2, a few findings account for the majority of appearances in clinical cases.

Human evaluation : In order to meet the medical requirements, we employed the expertise of a clinician to evaluate a subset of our medical findings database, more precisely the 25 most frequent findings (as they appeared in our data) and 25 random findings (not among the 25 most frequent). The goal of this evaluation was to validate the normal range boundaries and associated medical terms for both “high” and “low” values across these 50 findings. This process yielded a total of 100 unique medical terms for validation, following the structure depicted in Table 3.2. This subset represents a third of our entire findings database, which has been extracted from our clinical cases.

Additionally, we provided the medical expert with a form to assess the relevance of our approach to translate medical findings into medical terms. The form contained three key questions : i) Is it medically sound and feasible to translate a finding such as “Temperature is 39°C” into a medical term like “Fever,” and what are the limitations of such an approach from a medical standpoint ?, ii) Are there always corresponding “high” and “low” values ?, and iii) How precise should we be when defining boundary values and units ? To summarize the results of the expert analysis, we concluded that i) this approach is considered to be medically accurate and relevant for the majority of the cases, except for some findings that

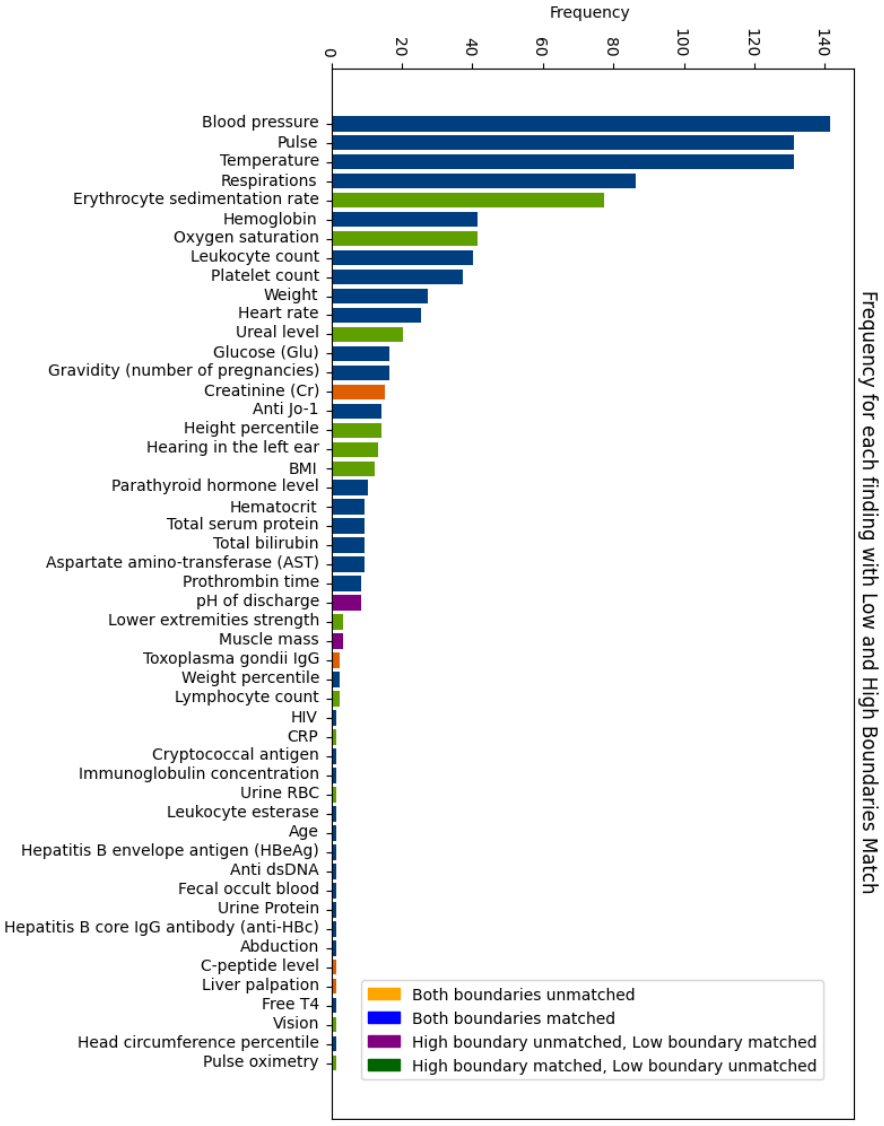

Figure 3.2 – Overview of our matched findings, ordered by occurrence in MedQA-USMLE-Symp clinical cases.

are context dependent such as “Acid-fast stain” that need to be associated to a bacteria to make sense. The second point also highlight that ii) some findings could not have both boundaries, such as the finding “vision”, that could only be lower than normal and therefore have only one boundary and associated term. Finally the medical expert emphasized that iii) biological values are not strict and often associated to a shallow acceptable value echoing with our gray zones thresholds used to calculate the accuracy of the LLM predictions. However, this later point is not the case in MedQA clinical cases, as they present clear-cut cases which support the students' training.

Following the expert's corrections, we discovered that our database creation using the ChatGPT-4 algorithm showed a good performance, achieving an accuracy rate of 78%. An analysis of the matched findings, ordered by finding occurrence, is visualized in Figure 3.2. This figure shows that errors were predominantly associated with the less represented findings, thereby highlighting a limitation of LLMs, i.e., their ineffectiveness to return knowledge from underrepresented data even in a contextual setup.

3.2 Proposed Architecture

The primary goal of this contribution is to generate natural language explanations that justify why a given diagnosis is correct and why other options are incorrect, based on the details of their clinical case. In this section I present the methods used to build the SYMEXP explanatory pipeline [11] which is based on the following components : First, clinical cases are analysed by i) a Medical Named Entity Recognition model, trained over the newly annotated MedQA-USMLE-Symp dataset presented in Section 3.1. This model mainly extracts two important informations namely symptoms and findings. Findings are then provided to the ii) Medical Finding Translation module in order to convert them into medical terms (e.g., temperature is 39.3°C $\rightarrow$ fever). After that, both detected symptoms and converted findings are provided to a iii) Medical Term Alignment module with the aim to align them with the HPO vocabulary and assign them to correct or incorrect diagnoses (i.e., options of the clinical case). Finally, we generate iv) template-based Natural Language Explanation, relying on the aligned symptoms and findings with the ontology and the occurrences rates from OMIM and ORPHA. I propose an overview of the architecture, called SYMEXP, in Figure 3.3.

3.2.1 Entities Identification

In order to accurately diagnose a patient's condition, it is important to identify the symptoms that are most relevant to the possible diagnoses. This means searching through the symptoms that have been detected and reported in the clinical case, and determining which ones are most likely to be related to the patient's condition. This can be done by considering the individual symptoms and their potential relevance to the possible diagnoses. It is also important to consider any additional information that may be available, such as the patient medical measurements, observation or test (i.e., findings), to be able to fully

11. We do not report any experiment carried out on live vertebrate (or higher invertebrates), humans or human samples. As we rely on standard benchmarks in the field of AI in medicine, it is not possible to identify patient/participant information as the clinical cases are not real cases but they are explicitly conceived for training medical residents. Our research does not concern either human transplantation research, nor it reports results of a clinical trial.

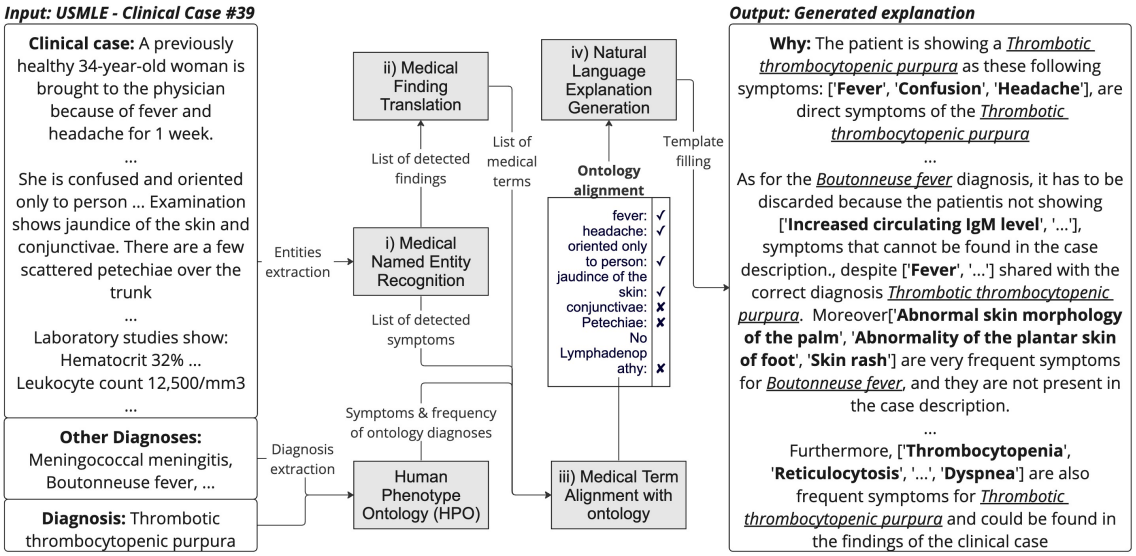

Figure 3.3 – Overview of our full pipeline for terms detection, conversion, alignment, and NL explanation generation module. The steps are i) Medical Named Entity Recognition, ii) Medical Finding Translation, iii) Medical Term Alignment with ontology and iv) Natural Language Explanation Generation.

explain the diagnosis. This first part of the pipeline identify a list of medical terms detected from the clinical case and converted from detected findings in order to prepare them to be further align with the HPO.

Entities detection. As introduced before, we rely on the USMLE dataset described in Section 3.1. In USMLE clinical cases, patient often express the symptoms in its own words and this layperson vocabulary is not well detected by standard medical NER systems [171]. In order to extract a maximum of information from clinical cases, we propose a neural approach based on pre-trained Transformer Language Models, fine-tuned on manually annotated entities from our proposed MEDQA-USMLE-Symp dataset (Section 3.1), so as to incorporate layperson terms and findings into our training set. More specifically, we cast the detection problem as a sequence tagging task. Following the BIO-tagging [170] scheme, each token is labeled as either being at the **B**eginning, **I**nside or **O**utside of a component. This translates into a sequence tagging problem with five labels, i.e., *B-Sign-or-Symptom*, *I-Sign-or-Symptom*, *B-Finding*, *I-Finding* and *Outside*.

Findings Interpretation. While symptoms entities can be aligned with the concepts in the ontology without any extra processing (e.g., pain in the head → headache), findings show more complexity to be relevant for diagnoses. Findings vary in the way they are presented but often include the value—unit pair, allowing them to be compared with "normal values" to make sense and represent the patient's condition. Our NER system does identify findings in clinical cases (results are showed in Section 3.3) but no existing vocabulary or database provides automatic interpretation of them. This is why we have created this new database of findings, presented in Section 3.1, presenting the most common

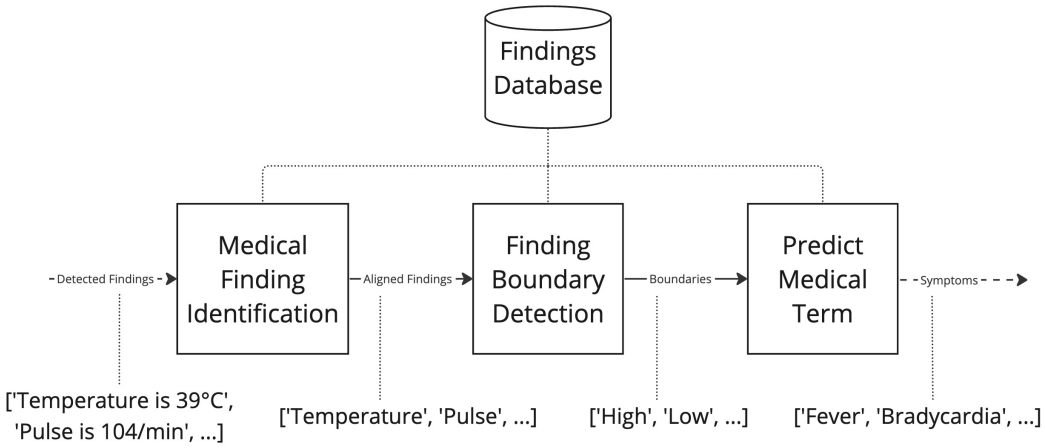

Figure 3.4 – Findings to medical terms converter module.

findings, their normal values and the terms associated with abnormal values. As discussed below, the findings, after conversion, strongly resemble symptoms that we could detect with our NER (e.g., temperature is 39°C → fever). This is why we need an automatic conversion system to increase the number of terms to be aligned, and therefore, the number of potential arguments (i.e., symptoms) for our explanations. Figure 3.4 shows in detail the conversion module (ii) taking the detected findings and returning a list of medical terms to be added to the list of detected symptoms. This new list of medical terms and layperson symptoms will be sent to the HPO alignment module (iii). More specifically, this module convert medical findings expressed in natural language, such as "Platelets count is 50000 mcL" into medical terms commonly found in physicians' vocabulary or medical ontologies, for instance, "Thrombocytopenia". To do so, we first rely on the previously detected findings and perform three key steps : a) medical findings identification, b) findings boundary detection, and c) prediction of the associated medical term.

The first step a) involve accurately identifying the relevant finding within the input sentence. In our example, the finding "Platelets count" is explicitly stated, whereas in some cases, like "respirations are 22/min," the finding may be incomplete or represented by a synonym, such as "Respiration rate" or "Breathing rate," rather than simply "Respiration". This step enables the alignment of the finding sentence with an entry in our database while filtering out potential errors arising from sentences that either lack findings or contain multiple findings. The second step b) determine whether the detected finding value falls within normal ranges or should be classified as not applicable. If the value falls outside of the normal boundaries, we investigate whether there is an associated medical term. If the finding is not applicable, we do not proceed further. Finally, in the third step c), we predict which medical term, if any, is associated with the detected finding and boundary classification.

On-the-fly generation of knowledge. As we created an expert verified database of findings boundaries and associated medical terms, most of the findings detected in the MEDQA-USMLE-Symp dataset already exist and can be retrieved from the database. However, as the semi-automatic process of generating the draft database (detaild in Section 3.1) showed that it eased the work of the expert by being accurate (i.e., Table 3.4) we decided to integrate this process as an option if the knowledge about detected finding is not found in the current version of the database. In other words, if the user allows it, we generate on-the-fly the missing findings normal values and medical terms for exessive values, highlighting the part generated by the LLM. We proposed three methods to generate the findings knowledge using LLM by I) Input-Output (IO) Zero-Shot Prompting [212] that will serve us as a baseline, II) mimicking the doctors reasoning with a Chain of Thought (CoT) [213], and III) using Self Consistency (SC) with IO and CoT [211].

The first method, Input-Output (IO) Zero-Shot Prompting, provides a basic approach where the medical term is directly predicted from the detected finding, without any intermediate steps. This method solely relies on the capabilities of the LLM for its predictions. The second method, Chain of Thought (CoT), seeks to emulate the process of medical professionals. It divides the prediction task into two phases : firstly, determining the boundaries (both low and high) associated with the detected finding, and secondly, correlating this value range with the appropriate medical term. Lastly, the Self Consistency (SC) method enhances the decision-making process by repetitively applying the previous methods and employing a voting mechanism to select the most reliable outcome. For medical terms, this involves a count-based voting system where the most frequent occurring term is chosen. For determining value boundaries, we experimented with two approaches, i.e., an average on all predictions, and a count-based voting system.

The converted findings are then injected together with the detected symptoms into the medical term alignment algorithm. This will ensure the inclusion of findings interpretation within the generated explanations in the last step of the pipeline (Figure 3.3).

3.2.2 Medical term alignment

The medical term alignment module (Fig. 3.3, component (iii)) associates, whenever possible, the pertinent symptoms or translated findings mentioned in the clinical case description with a term of a diagnosis found in the HPO knowledge base. To the best of our knowledge, only Manzini et al. [125] proposed a solution, named DASH, to align automatically layperson symptoms with an ontology. Our approach differs from Manzini et al. because we align not only the symptoms but also the findings with the concepts of the ontology. In addition, we explore a contextual approach where DASH compares only the symptoms with HPO terms. We compare our approach regarding to Manzini et al. further in the result paragraph and describe their methodology in detail in Section 3.5. Our proposed framework consists of two different steps, where : *(a)* we retrieve from the HPO the required diagnosis information (i.e., the terms and how common they are), then the symptoms in the case are detected and extracted using the modules introduced in the previous section ; *(b)* the relevancy of each symptom is assessed by matching the detected medical term with the ones retrieved from the HPO, e.g., “Platelets count is 50000 mcL” converted into Thrombocytopenia to the HPO concept HP :0001873 [12]. The matched terms

12. https://hpo.jax.org/app/browse/term/HP:0001873

are then used to generate natural language argument-based explanations for correct and incorrect diagnoses.

Regarding the matching module *(b)*, we experimented with two different methods to align our detected entities with terms in the HPO by *(i)* similarly to DASH, directly comparing the computed embeddings of the detected entities with the embeddings of the terms in the HPO, and *(ii)* by taking into account the context in which the entities are detected and applying the same context to every term in the HPO. The reasoning behind the latter is that the corresponding entities in the HPO should not change the semantics of the sentence with respect to the other symptoms. To align our detected symptoms and converted findings with the equivalent HPO terms, we calculate the cosine distance of each embedding of the HPO terms with respect to the embedding of the detected symptom.

It is worth noticing that, for task *(ii)*, it is necessary to calculate the context embeddings "on-the-fly" because each context is unique and depends on the clinical case in which it has been detected. However, to avoid recomputing all HPO term embeddings on the fly for each new context (i.e., the ontology contains 10,319 unique terms), we propose to generate all the HPO terms embedding at once and store them. Therefore, this module takes both symptoms and converted findings, detected by the previous module and looks for the context [13] of these symptoms in the clinical case.

The context C is embedded using sentence embedding methods and saved separately from the symptoms S, and the two embeddings are merged together $(C + S)$ to form the reference R. This same context embedding C is added in the same way to each HPO term embedding $T_1, T_2, \ldots, T_i$ to form the candidates $C_1, C_2, \ldots, C_i$. We compute and retrieve the five best cosine distances between C and R to address a fair comparison with the other systems.

3.2.3 Explanation generation

We propose a template-based explanation generation module based solely on the symptoms and findings that are relevant to explain the diagnosis. To do this we propose several templates that tackle different aspects of explanations, going from explaining why a patient was given a certain diagnosis, to explaining why the alternatives cannot be considered as viable options. We support our explanations with statistical information obtained from the HPO such as the frequency of each symptom incidence, and we propose to look for possible symptoms that were not detected by the system but are frequent for a certain disease. These explanations are built from the HPO ontology, where the answers (i.e., diseases) are retrieved, taking care to separate the correct answers from the incorrect ones. Each disease has a list of associated terms (with codes) among which we look for codes in common with the output of the alignment module. These aligned terms identified as relevant for our diseases are then saved and, if possible, we associate them with the frequencies provided by ORPHA and OMIM. They will serve as arguments for the explanations why the correct answer is correct and why the others are not. Finally, the terms that are missing in our system but have a high frequency of occurrence according to the ontology are also saved to populate our supplementary template. The detailed templates and examples are described in Section 3.4.

13. The context consists of the sentence(s) containing the symptom and the entire clinical case.

3.3 System Implementation

In this section, we report on the experimental setting, the obtained results and the error analysis for the named entities detection, the finding translation and the symptom alignment methods. It is worth noticing that the presented pipeline can be applied also to different clinical cases datasets, ensuring the generalisability of the proposed approach. This pipeline named SYMEXP is a part of the wider ANTIDOTE project, introduced in Section 1.1 and is currently deployed online [14] within the ANTIDOTE project showcase. The ANTIDOTE software suite implementation is detail further in Chapter 5.

3.3.1 Experimental setting

In order to make all our work reusable and reproducible, we present the implementation details of SYMEXP. Section 2.1 presents an overview of natural language representations and completes the implementation details of the proposed pipeline.

Named Entities Recognition. For the entity detection task, we experiment with different transformer-based language models such as BERT [58], SciBERT [20], BioBERT [117], PubMedBERT [77] and UmlsBERT [132] initialized with their respective pre-trained weights. All the models we employ are specialized in the scientific or biomedical domain, with the exception of BERT which will serve us as a baseline. To fine-tune the LMs, we use the PyTorch implementation of Huggingface [216] (v4.18). For BERT, we use the uncased base model with 12 transformer blocks, a hidden size of 768, 12 attention heads, and a learning rate of 2.5e-5 with Adam optimizer for 3 epochs. The same configuration was used to fine-tune SciBERT, BioBERT, PubMedBERT and UmlsBERT. For SciBERT, we use both the cased and uncased versions, and for BioBERT we use version 1.2. Batch size was 8 with a maximum sequence length of 128 subword tokens per input example. Both the dataset and the guidelines used to train our NER model are available in this project repository [15].

Finding converter. In our experiments, we adopted the large language model from OpenAI, ChatGPT gpt-3.5-turbo-0301 and gpt-4 [30, 154]. We employed the snapshot of gpt-3.5-turbo from March 1st, 2023. This model was used for both joint and combined baselines, employing classic handcrafted prompts described in Appendice A.2. These two models were the most recent at the time of the experiments and are currently replaced by gpt-4o-mini in the available online version.

For the CoT steps that mimics doctors' reasoning, we used the FuzzyWuzzy [16] Python package version 0.18.0 for the task of medical finding identification. This package leverages the Levenshtein distance to calculate the differences between sequences in a user-friendly package. Concerning the finding values detection using string matching, we employed a Python regular expression with the regex package version 2022.10.31 :

```
\b\d+(?:\.\d+)?\b
```

14. http://antidote.i3s.unice.fr/symexp/
15. https://github.com/Wimmics/MEDQA-USMLE-Symp/tree/main/MEDQA-USMLE-Symp_corpus
16. https://pypi.org/project/fuzzywuzzy/

We experiment as an alternative a NER approach, using med7 [111] with the "en_core_med7_lg" model, trained on MIMIC-III free-text electronic health records, and Spacy version 3.5.2. All experiments were conducted using Python 3.10.11 directly in a Google Collaboratory Pro notebook. The medical finding database, validated by medical expert is available in our project repository [17].

Ontology alignment. Regarding the matching module, we experimented with two different methods to align the detected entities with the terms in the HPO by *(i)* directly comparing the computed embeddings of the detected entities with the embeddings of the terms in the HPO, and *(ii)* by taking into account the context in which the entities are detected and applying the same context to every term in the HPO. In the experimental setting of both tasks *(i)* and *(ii)*, we use the static pre-trained embeddings GloVe 6B as well as BERT, SciBERT, BioBERT and UmlsBERT with the same configurations as in the medical NER task. Each embedding is calculated with Sentence Transformer Document Embeddings using the flair framework [4], with the same Python environment as the previous modules.

We defined a test set of 23 cases from the MEDQA-USMLE-Symp dataset where *(i)* we retrieved from the HPO the symptoms related to the diagnoses for each case, and *(ii)* we manually aligned the annotated symptoms in the case to the concepts from the HPO. This resulted in 162 symptoms aligned to a specific term in the HPO that serve us as a testing set for our matching module.

As detailed further in Section 3.5, the system proposed by [125] offers a similar approach to translating layperson terms to medical terms in the HPO. However, their work does not take into account the context in which a symptom is mentioned in the text and does not provide any solution for findings interpretation. To compare with this approach and due to the unavailability of their model, we rely on their online demo, which outputs only the top 5 ranking of the HPO terms that are closest to the input symptom. To perform a comparison with our pipeline, we first compute the accuracy of the aligned symptoms using our symptom alignment module and then replaced it with Manzini et al. [125] proposed system (DASH). Results are shown later, in Table 3.9.

Since a symptom can be composed of several words (e.g., "shortness of breath"), we separated the symptom into words that we encode by either using each word as an input on GloVe [157], or extracting directly from the contextualized models the representation of the symptom by summarizing the hidden states of the last four layers in the model. We then sum the vectors of each word to get an n-gram representation of the symptom. We also explore sentence embeddings, by making use of Sentence-BERT [172], a model that derives semantically meaningful sentence embeddings (i.e., semantically similar sentences are close in vector space) that can be compared using cosine similarity. Sentence-BERT can be used with different pre-trained models, in this work we focus on the models BERT [58], SciBERT [20], UMLSBERT [132] and S-PubMedBert by [57]. The first represents a competitive baseline in our experiments since it is the state-of-the-art model for comparing sentences cross-domain, while the three latter models are pre-trained on scientific or medical data or both.

To tackle both tasks, we make use of the MedQA-USMLE-Symp dataset introduced in Section 3.1. The annotations are converted into two datasets, one for each part of

17. https://github.com/Wimmics/MEDQA-USMLE-Symp/tree/main/Findings-database

TABLE 3.5 – Results for entity recognition in macro multi-class precision, recall, and macro f1-score.

Model	P	R	F1
BERT	0.85	0.84	0.84
BioBERT v1.2	0.84	0.85	0.84
UmlsBERT	0.85	0.85	0.85
PubMedBERTbase	0.83	0.84	0.83
SciBERT cased	0.85	0.85	0.85
SciBERT uncased	**0.85**	**0.86**	**0.86**

the pipeline. The first dataset is used for the symptom detection task, and it is in the CoNLL format for token-wise labels. The second dataset, for the symptom alignment task, is converted into a csv format, where each symptom in the clinical case description and available related knowledge (i.e., the list of symptoms and their frequencies for each possible diagnosis associated with the case) extracted from the HPO are paired. Finally, we rely on the HPO ontology, utilizing the requests package version 2.27.1 and the public HPO endpoint [18]. It is important to note that this service is no longer available on the HPO and is replaced by a local knowledge base to download.

3.3.2 Results

In the following, we report on the results obtained for our pipeline presented in Figure 3.3, focusing on medical entities recognition, finding conversion and alignment with ontology. We compare our methodology with the DASH system proposed by Manzini et al. [125].

Medical NER. As introduced before, the first task addressed in our pipeline is to detect the medical named entities. The results for the symptom detection task are shown in Table 3.5 in macro multi-class precision, recall, and F1 score. We can observe that all models perform similarly, with the best results from the specialized SciBERT [20] model. The biggest difference in performance is given by comparing SciBERT uncased with PubMedBERT, with the SciBERT model performing better. Interestingly, BERT performs closely to the specialized models, and, in some cases, it outperforms them. This may be due to the fact that the clinical cases from our dataset are written for medical exams at the med school. They contain some technical specialized words, but overall the symptoms are described in layperson terms. It is also worth noticing that the majority of our labels do not belong to medical terminology (i.e., *Age* and *Population Group*, *Location* and *Temporal Concept*). *Sign or Symptom* and *Finding* are the only labels that require specialized vocabulary.

Overall, SciBERT uncased is the best-performing model (in bold) with a macro f1-score of 0.86, outperforming the other approaches for each of the categories. In Table 3.6 and Table 3.7 we report on the performances for each entity with the best-performing models SciBERT and BERT. The *Sign or Symptom* and **Finding** detection task obtains a 0.82 and 0.86 of macro f1-score. In the work of [147], the authors also detect symptoms obtaining

18. https://hpo.jax.org/api/hpo/search

TABLE 3.6 – Results for entity recognition using our best performing model (SciBERT uncased) in P, R, and f1-score.

Entity	P	R	F1
Other	0.93	0.91	0.92
Age Group	1.00	0.97	0.98
Finding	0.85	0.88	0.86
Location	0.74	0.80	0.77
No Symptom Occurrence	0.79	0.72	0.75
Population Group	0.88	0.95	0.91
Sign or Symptom	0.83	0.82	0.82
Temporal Concept	0.78	0.87	0.82
Weighted avg	0.89	0.89	0.89
Macro avg	0.85	0.86	0.86

an f1 score of 0.61. However, these results can not be directly compared since the datasets on which both models were fine-tuned are different : we train on clinical cases, while they use dialogues between doctors and patients. Moreover, given that the dataset they use is not released, we can not evaluate our approach on their data to compare the results.

TABLE 3.7 – Results for entity recognition using BERT uncased in P, R, and f1-score.

Entity	P	R	F1
Other	0.92	0.91	0.92
Age Group	1.00	0.97	0.98
Finding	0.87	0.87	0.87
Location	0.74	0.80	0.77
No Symptom Occurrence	0.79	0.72	0.75
Population Group	0.88	0.95	0.91
Sign or Symptom	0.83	0.82	0.82
Temporal Concept	0.78	0.87	0.82
Weighted avg	0.89	0.89	0.89
Macro avg	0.85	0.86	0.86

Raza et al. [171] proposed a transformer-based NER system employing DistillBERT [177] that is able to recognize a wide range of clinical entity types, encompassing medical risk factors, vital signs, drugs, and biological functions. Their approach, which primarily relies on the Case Report dataset MACROBAT [41], focuses on doctors vocabulary. To make a fair comparison, we evaluated the output of their model, BioEN, at a token level using our own test set, specifically comparing the accuracy of the *Signs or Symptoms* labels. The results highlight a significant gap between the two approaches in terms of performance : out of 285 gold tokens, BioEN detected only 79, whereas our model identified 260. This disparity is primarily due to our specific focus on the detection of data encoded in layperson vocabulary.

Finding converter. Here we describe the results of the prediction task of the medical terms associated to the detected findings. The efficacy of our medical finding to medical term on-the-fly conversion module is detailed in Table 3.8. This table presents the accuracy of our Finding Converter module in identifying boundaries, both in terms of values and associated terms. The mentioned 78% accuracy refers to the accuracy achieved in determining values, considering the previously discussed 20% threshold. The accuracy figures are computed based on the final version of the generated database, which achieved an accuracy of 78% for "low" boundaries (88%) and "high" boundaries (68%), after the doctor validation. The proficiency of the model in predicting "low" boundaries could be attributed to their higher frequency and often singular appearance as the defining boundary for a medical finding. For instance, the "vision" finding exemplifies this trend, as it only has a "low" boundary, represented by "blindness", with no corresponding "high" boundary. The context added by the intermediate steps seems to fine-tune the language model's knowledge and helps in generating more suitable responses. The Self-Consistency method does not improve the results.

TABLE 3.8 – Results for "on-the-fly" findings to medical terms prediction using the generative LLM ChatGPT.

Prompting Method	**Accuracy**
IO ChatGPT 4	0.64
IO ChatGPT 3	0.52
CoT ChatGPT 4	**0.66**
CoT ChatGPT 3	0.52
SC IO ChatGPT 4	0.64
SC IO ChatGPT 3	0.54
SC CoT ChatGPT 4	**0.66**
SC CoT ChatGPT 3	0.54

The results of the symptom alignment module, that aim to associate the detected entities in the clinical case with the HPO ontology, are summarised in Table 3.9. As baseline models, we propose to use the same methods but without the context of the symptoms, similarly to *DASH* [125]. In Table 3.9, we show only the best-performing baseline *PubMedBERT no context* obtaining similar results to *DASH* (0.41 and 0.37, respectively). Adding contextual representation to the embeddings results in a significant improvement (up to 0.70 in accuracy) supporting the hypothesis that context plays an important role when translating layperson terms to formal medical terms.

3.3.3 Error Analysis

The main limitation of adopting the HPO as medical knowledge base concerns the number of symptoms associated with each diagnosis. For some diagnoses, we have multiple symptoms, while for others we can have only one or none. We notice that in those cases where the diagnosis is a mental disease, the model tends to make more mistakes. Inspecting the HPO for this kind of diagnoses, we find that either the diagnosis does not appear in the HPO ontology or the symptoms tend to be more general, including a lot of common symptoms like "changes in appetite" or "low energy", that alone may not be relevant but

TABLE 3.9 – Results for DASH and our symptom alignment method using different embeddings with and without context (accuracy score).

Model	Accuracy
DASH	0.37
BERT + no context	0.39
SciBERT + no context	0.39
UMLSBERT + no context	0.44
S-PubMedBERT no context	0.53
BERT + context	0.53
SciBERT + context	0.57
UMLSBERT + context	0.59
S-PubMedBERT + context	**0.70**

all together indicate a precise diagnosis. Moreover, some relevant symptoms may not be described explicitly but encoded in the clinical cases as *Findings*. These findings, even translated into a medical term, do not appear in the symptoms list extracted from the HPO, since this ontology focuses on pathological terms. The finding "Gravidity" (i.e., number of pregnancies) exemplifies this insight because being "Multiparous" is not pathological but is the medical term associated to the "high" boundary of the finding. Therefore it would be useful for the explanation but it does not match in our system because of the HPO limitations. Moreover, a diagnostic can be supported by a less specific interpretation of a finding, e.g., the thrombotic thrombocytopenic purpura can be explained by a patient "Arrhythmia" defined as "A irregular heartbeat / A problem with the rate or rhythm of your heartbeat" but our system will detect either a "Bradypnea" or a "Tachypnea" that are both kinds of Arrhythmia. A possible extension of this work consists in a deeper investigation of the ontologies to find a way to align with different granularity the detected finding.

Given that we rely on the HPO only, some diseases or diagnoses are not present in the knowledge base, preventing us to generate the associated explanations. Combining the HPO with more specialized medical knowledge bases is a future direction for this work, both to complete the information we have, and also to integrate new diagnoses.

3.4 Argumentation patterns for Explanations Generation

As I discussed in Chapter 2, argumentation explanations are an interesting direction in an educational context and are perfectly suited to explain examinations focused on diagnosis. Therefore, I present in this section the final brick in our explanatory pipeline : the generation of natural language explanations. While the previous sections introduced how to identify complex medical concepts through clinical cases, the generation of explanations is a different challenge. Given the critical aspect of medical data, it is necessary to control the generation of explanations to make them clear and true. While many approaches focus on natural language generation [154, 30] and some on explanation generation [33, 145, 29, 113, 115], all of them are based on neural models. Although these LMs are very powerful and extremely convincing, they face certain uncontrollable yet limitations that make their use incompatible with the medical field (e.g., hallucinations, bias) [13]. In order to provide

high quality explanations, we decided to create argumentation patterns to generate our explanations [100, 35, 61]. We therefore created 3 patterns to i) explain the correct answer based on the detected, converted and aligned entities, ii) explain the set of incorrect answers with the discriminative symptoms and finally iii) draw attention to important symptoms missing in the clinical case (or not detected). It's worthy to note that these templates have been developed with the help of a doctor to make them as useful as possible, while keeping them flexible enough not to make the templates obvious. Let us consider the following clinical case, where in bold we highlight the **symptoms**, in italic the *findings* and we underline the relevant symptoms and findings supporting the correct answer.

Clinical case. A previously healthy 34-year-old woman is brought to the physician because of **fever** and **headache** for 1 week. She has not been exposed to any disease. She takes no medications. Her *temperature is 39.3°C (102.8°F)*, *pulse is 104/min*, *respirations are 24/min*, and *blood pressure is 135/88 mm Hg*. She is **confused** and **oriented only to person**. Examination shows **jaundice of the skin** and **conjunctivae**. There are a few scattered **petechiae** over the trunk and back. There is **no lymphadenopathy**. Physical and neurologic examinations show **no other abnormalities**. *Test of the stool for occult blood is positive*. Laboratory studies show : *Hematocrit 32% with fragmented and nucleated erythrocytes Leukocyte count 12,500/mm3 Platelet count 20,000/mm3 Prothrombin time 10 sec Partial thromboplastin time 30 sec Fibrin split products negative Serum Urea nitrogen 35 mg/dL Creatinine 3.0 mg/dL Bilirubin Total 3.0 mg/dL Direct 0.5 mg/dL Lactate dehydrogenase 1000 U/L Blood and urine cultures are negative*. A CT scan of the head shows **no abnormalities**. Which of the following is the most likely diagnosis ?

This example is extracted from the MEDQA-USMLE-Symp dataset and the (already known) correct diagnosis is **Thrombotic thrombocytopenic purpura**, whilst the other (incorrect) options are Disseminated intravascular coagulation, Immune thrombocytopenic purpura, Meningococcal meningitis, Sarcoidosis and Systemic lupus erythematosus.

Why Pattern. After empirical observation of clinical cases, we have identified that one of the easiest ways to justify the correct diagnosis is to support it through the symptoms identified in the case. Often, one or more of these symptoms are unique, or occur in 100% of cases. This intuition to justify the correct response to explain the diagnosis will be confirmed later in Chapter 4. Indeed, these later investigations identifies in Section 4.3 argumentation patterns in doctors explanations, and analyse of the distribution of explanatory labels (i.e., Table 4.6) highlighting that only less than 5% of clinical cases question answering exams do not justify the correct answer (and explain differently). Given these informations, we prepared the following "Why" template :

Définition 3.4.1. (Why the correct diagnosis is correct) The patient is showing a [CORRECT DIAGNOSIS] as these following symptoms [**PERFECT MATCHED SYMPTOMS, MATCHED SYMPTOMS, MATCHED FINDINGS**] are direct symptoms of [CORRECT DIAGNOSIS].

Moreover, [**OBLIGATORY SYMPTOMS**] are obligatory symptoms (always present, i.e., in 100% of the cases) and [**VERY FREQUENT SYMPTOMS**] are very frequent symptoms (holding on 80% to 99% of the cases) for [CORRECT DIAGNOSIS] and are present in the case description. [19]

19. Sources from the HPO : https://hpo.jax.org/app/browse/term/HP:0040279

In Template 3.4.1, the [CORRECT DIAGNOSIS] represents the correct answer to the question "Which of the following is the most likely diagnosis ?" and therefore the correct diagnosis of the described disease. The [**SYMPTOMS**] / [**FINDINGS**] in bold represent the medical terms automatically detected through the first module of our pipeline, and they are also underlined when they are considered as relevant by our matching module (i.e., they are listed among the symptoms for the disease in the HPO knowledge base). Both [**PERFECT MATCHED SYMPTOMS**] and [**MATCHED SYMPTOMS**] in Template 3.4.1 are considered relevant but they differ in the confidence level the system assigns to the matched symptoms. This allows us to integrate a notion of granularity in our explanations and to rely on the symptoms or raw findings detected in the clinical case that strongly match with a symptom in the HPO. If the system does not detect any relevant symptom, no explanation is generated for the correct answer. This isn't necessarily a problem, as in some cases it's difficult to identify the correct answer from the symptoms in common, and so sometimes an approach such as eliminating incorrect answers leads to the deduction of the correct answer (i.e., by elimination). Furthermore, we employ the information about the symptom frequencies (retrieved through the HPO) in the [**OBLIGATORY SYMPTOMS**] and [**VERY FREQUENT SYMPTOMS**] to generate stronger evidence to support our natural language argumentative explanations. Sometimes the frequencies are not available in the HPO, in which case we do not display them in our final explanation.

The following example show some explanatory arguments automatically generated by our system.

Exemple 3.4.1 – The patient is showing a [Thrombotic thrombocytopenic purpura] as these following symptoms [**Headache**, **Fever**, **Confusion (Oriented to persons)**, **Thrombocytopenia (Platelet count 20,000/mm3)**, **Reticulocytosis (Jaundice of the skin)** and **Decreased serum creatinine (Creatinine 3.0 mg/dL)**] are direct symptoms of [Thrombotic thrombocytopenic purpura].

Moreover [**Reticulocytosis (Jaundice of the skin)** and **Thrombocytopenia (Platelet count 20,000/mm3)**] are very frequent symptoms (holding on 80% to 99% of the cases) for [Thrombotic thrombocytopenic purpura] and are present in the case description.

When filling the [**SYMPTOMS** and **FINDINGS**] span in Template 3.4.1, we inject only the terms matched in the HPO for the [**PERFECT MATCHED SYMPTOMS**], and we combine the HPO with the detected symptoms and findings in the case description for the [**MATCHED SYMPTOMS** and **MATCHED FINDINGS**] in this form : [**matched term in HPO (detected term in the clinical case)**] (e.g., in Example 3.4.1 : **Confusion (Oriented to persons)** and **Thrombocytopenia (Platelet count 20,000/mm3)**.).

Why not Template. Explaining why one diagnosis is the correct one is important, but it's also in some cases necessary to use another approach more based on deduction [135]. This intuition is also confirmed by the next chapter (i.e., Table 4.6), where I observe that 29% of the explanations given by doctors explain each incorrect options. We, therefore, propose to provide explanations based on the relevant symptoms for the incorrect options by contrasting them with the clinical case at hand.

Définition 3.4.2. (Why this incorrect diagnosis is incorrect) Concerning the [INCORRECT DIAGNOSIS] diagnosis, it has to be discarded because the patient in the case description is not showing [**INCORRECT DIAGNOSIS SYMPTOMS / FINDINGS FROM THE HPO (MINUS DETECTED SYMPTOMS IN CASE)**] symptoms.

Despite [**SHARED CORRECT SYMPTOMS / FINDINGS**] symptoms shared with the [CORRECT DIAGNOSIS] correct diagnosis, the [INCORRECT DIAGNOSIS] diagnosis is based on [**INCORRECT DIAGNOSIS SYMPTOMS**].

Moreover, [**OBLIGATORY SYMPTOMS**] are obligatory symptoms (always present, i.e., in 100% of the cases) and [**VERY FREQUENT SYMPTOMS**] are very frequent symptoms (holding on 80% to 99% of the cases) for [INCORRECT DIAGNOSIS], and they are not present in the case description.

Template 3.4.2 can be applied to each incorrect possible answer of the case, individually. The incorrect answer corresponds to the [INCORRECT DIAGNOSIS] and [**INCORRECT DIAGNOSIS SYMPTOMS / FINDINGS**] are all relevant terms associated with this disease in the HPO knowledge base, without the terms in common with the correct answer. Again, in the template, we use the frequencies provided by the HPO to provide further evidence to make our explanatory arguments more effective. The template includes therefore with [**OBLIGATORY SYMPTOMS**] and [**VERY FREQUENT SYMPTOMS**] the mandatory and very frequent symptoms of the incorrect diagnosis, which are missing in the clinical case description. The following explanations are automatically generated for (one of) the incorrect diagnoses of the clinical case we introduced at the beginning of this section.

Exemple 3.4.2 – Concerning the [Meningococcal meningitis] diagnostic, it has to be discarded because the patient in the case description is not showing [**Stiff neck, Nuchal rigidity or CSF pleocytosis, Increased CSF protein, Hypoglycorrhachia**] symptoms.

Despite [**Petechiae, Fever, Headache**] symptoms shared with the [Thrombotic thrombocytopenic purpura] correct diagnosis, the [Meningococcal meningitis] diagnosis is based on [**Stiff neck, Nuchal rigidity or CSF pleocytosis, Increased CSF protein and Hypoglycorrhachia**].

Moreover, [**Stiff neck, Nuchal rigidity, CSF pleocytosis, Increased CSF protein or Hypoglycorrhachia**] are very frequent symptoms (holding on 80% to 99% of the cases) for [Meningococcal meningitis] and are not present in the case description.

Example 3.4.2 shows the natural language explanation of why the possible answer [Meningococcal meningitis] is not the correct diagnosis given the symptoms discussed in the clinical case description. In case the disease is not found in the HPO, we do not generate the associated explanation.

Additional Explanatory Arguments. In order to enrich our explanations with additional explanatory arguments to improve critical thinking in the medical residents, we also generate another template. Indeed, in some clinical cases, it is possible that the detected terms are not sufficient to explain the diagnosis or sometimes the informations are missed by the proposed system.

In some situations, SYMEXP is not able to abstract some findings that are important for the diagnosis as for the **Thrombotic thrombocytopenic purpura**, a Very frequent symptom is “Arrhythmia”, defined as “Any cardiac rhythm other than the normal sinus rhythm”. Our system will detect a “Tachycardia” that, by definition is a kind of “Arrhythmia” (i.e., high boundary). Template 3.4.3 aims at drawing the medical residents’ attention to (statistically) important symptoms that are missing or not explicitly mentioned in the clinical case description :

Définition 3.4.3. Furthermore, [**CORRECT DIAGNOSIS VERY FREQUENT TERMS (MINUS MATCHED TERMS)**] are also frequent symptoms for [CORRECT DIAGNOSIS] and could be found in the findings of the clinical case.

Example 3.4.3 is generated by our system and brings attention to "Arrhythmia". This additional explanatory argument complements the explanation we generate for the correct and incorrect diagnoses in the case presented at the beginning of this section.

Exemple 3.4.3 – Furthermore, [**Arrhythmia, Generalized muscle weakness**, and **Microangiopathic hemolytic anemia**] are also frequent symptoms for [Thrombotic thrombocytopenic purpura] and could be found in the findings of the clinical case.

All examples are extracted from our dataset and can be tested online on the latest version of SYMEXP [20]. Chapter 5 reuses the previous examples and shows the implementation of these templates in the website interface.

3.5 Related Work

This section presents and discusses existing work on the generation of natural language explanations and the role of verified sources of medical knowledge such as ontologies and thesauri in generating such explanations. First, I present available medical resources, describing medical vocabularies, structured knowledge bases and NLP datasets about medicine, and explanation. Then I review the Information Extraction systems working on medical data and, more specifically, on symptoms. Also, I describe existing approaches to align medical entities with standardized vocabulary and knowledge. Finally, I focus on the existing methods to generate natural language explanations.

3.5.1 Medical data and linguistic resources

As introduces in the Background chapter 2.4, a considerable amount of research effort focused on the construction of robust and trustworthy sources of knowledge like ontologies and vocabularies. We differentiate vocabularies, that aims at identifying concepts and disambiguate them from databases or knowledge bases who rather store data about a specific concept or area. Finally, within the natural language, some manually annotated datasets are proposed to train and evaluate language model approaches for the clinical scenario (i.e., entities detection, diagnosis classification, etc...).

Vocabularies. Several of these vocabularies are centered around clinical terms, such as SNOMED CT [60] [21], ICD (i.e., ICD-10) codes [163] and the HPO [110], making them useful for diagnosis prediction. The later one, the HPO, propose a knowledge base gathering many diseases and their associated phenotypes (or terms). On the other hand more specific in-domain tasks can be solved using specialized vocabularies. RxNorm [121], for instance, is devoted to clinical drugs, CPT [88] is built around procedural terminology, and MeSH [88] is designed for cataloging and searching biomedical information. Furthermore, Bodenreider et al. [25] proposed a Metathesaurus based on the aforementioned vocabularies and many others, into a unified structure. This integrated resource includes names, relationships, attributes, and other details related to biomedical and health-related concepts. Focusing on vital signs, health measurements, and observations, LOINC [131] proposed an international

20. http://antidote.i3s.unice.fr/symexp/
21. https://www.snomed.org/value-of-snomedct

standard to identify them but do not provide the associated normal values (i.e., normal hematocrit for men is 40 to 54% ; for women it is 36 to 48% [206]).

Datasets. In parallel to the efforts made in creating reliable medical vocabularies, significant advancements have been made in the compilation of medical datasets, particularly in natural language. Notably, this has been accomplished through shared tasks such as i2c2 (renamed as n2c2) for Information Extraction (IE) [191, 86], MEDIQA used in Natural Language Inference, Recognizing Question Entailment (RQE), and Question Answering (QA) [21], and SemEval 22 with IE and NLI tasks [104, 101, 204].

Finding Datasets. While Johnson et al. [97] proposed the MIMIC-III dataset consisting of textual data about vital signs, medications, laboratory measurements, observations and more, other efforts have focused both on structured and unstructured data. For instance, eICU and PhysioNet [160, 143] are two contributions that have been key in enhancing the body of available medical datasets by collecting respectively anonymized structured data from patients (including vital sign measurements, care plan documentation, diagnosis information, treatment information) and signals archive. Simultaneously, resources like the UK Biobank and the Cancer Imaging Archive [189, 47] include both medical images and textual data.

Medical NER Datasets. Numerous contributions focused on identifying medical named entities from article abstracts, primarily from PubMed. These approaches to Named Entity Recognition target various biomedical aspects, ranging from Part-of-Speech (PoS) tagging with the Extended GENIA dataset [152], to more detailed entity annotations on full articles, as in the CRAFT corpus [15]. The AnatEM corpus [161] and some of the BioNLP Shared Tasks [106, 108] concentrate on entities and relations, while other approaches [183, 119, 107, 162, 112, 75] specifically address gene, protein, and species entities.

Medical Findings NER Datasets. However, only a limited number of studies have focused on disease and medical findings annotation, e.g., the NCBI disease corpus [59] and our MedQA-USMLE-Symp dataset [126], which is annotated with UMLS symptoms and findings tags and described in the Resources Section 3.1. Despite these two resources, the issue of matching medical findings to symptoms is still an open research question. This highlights the need for further research in this area to improve the understanding and adoption of medical findings for more accurate and comprehensive diagnostic and explanatory tasks.

3.5.2 Information Extraction on medical text

Many robust off-the-shelf pipeline toolkits like Spacy [90], MedSpacy [67], and CLAMP [185], have been recently proposed for text processing, and in particular, to process medical text. Notably, MedSpacy is a specialized extension of Spacy, custom-built for clinical language processing. CLAMP stands out due to its capability for NER and its interactive interface for annotating clinical text. However, their rule-based approach for NER in the medical domain makes it complex to apply it to named entities not originally considered in the tool, and new rules need to be defined.

Recent approaches cast NER as a sequence labeling task, where transformer-based models have shown remarkable performance, especially when fine-tuned on specific domains. Naseem et al. [146] showed that pre-training the ALBERT model on a large-scale biomedical corpus enhances the model's ability to capture the context found in biomedical NER tasks. This specialized approach has resulted in these models outperforming non-specialized counterparts and achieving top-tier results on several datasets. BioELECTRA [102] exemplifies this trend by pre-training a biomedical language model using biomedical text and vocabulary with the ELECTRA architecture [48]. Other BERT-based models, such as SciBERT [20], BioBERT [117], PubMedBERT [77], and BioMed-RoBERTa [81], which is based on RoBERTa, have also been designed for the biomedical domain.

Other approaches like UmlsBERT [132] integrate domain knowledge from the Unified Medical Language System into a contextual embedding model. The model's strength lies in its ability to associate different clinical terms with similar meanings in the UMLS knowledge base, creating meaningful input embeddings by leveraging information from the semantic type of each word. I compare in Section 3.3 the representations of symptoms found in clinical cases with different contextual embeddings, seeking to identify a representation that aligns with the one provided in the Human Phenotype Ontology [110].

Raza et al. [171] propose the Bio-Epidemiology-Ner (BioEN) pipeline, an approach inspired by [126], where they fine-tune a DistilBERT [177] model, a simplified and more computationally efficient version of BERT, for the task of biomedical NER. They adapt the last layer of the pre-trained DistilBERT model to their specific biomedical task and adjust the input and output dimensions accordingly. However, the labels they use are not derived from any certified ontology or medical source, making this approach ad hoc to their NER labels and limiting its reusability. Furthermore, their approach does not account for the broader scope of medical findings, which include vital signs and analysis results, essential elements to analyse clinical cases.

Finally, with the goal to predict the correct diagnosis and explain these predictions using feature attribution methods, Ngai et al. [147] identify clinical information, where among the entities, symptoms are detected. They use the detected symptoms to know their intent (or pertinence) for the diagnostic across five labels *confirm/deny/unsure* of symptom, *closing* the discussion or *other*. Although they offer diagnostic explanations based on natural language dialogues, the explanations remain mathematical structures (as discussed in Section 2.3) and cannot be interpreted in the same way as natural language explanations.

3.5.3 Medical term alignment

As discussed in Section 2.3, generating explanations automatically grounded on medical knowledge is a long term challenge require interaction with expert sources of informations. As many vocabularies already exists and are widely used for prediction and inference tasks, trying to align non-structured data with structured and standardized knowledge for generation purpose can be viewed as a starting point. Focusing on explaining (already known) correct and incorrect diagnosis predictions is a good first step on explanatory argumentation and reduce the amount of knowledge to align with the natural language document. A symptomatic approach, based on the patient symptoms within the document can already give an explanation of diagnoses, mostly when we know which one is correct.

If detecting symptoms is partially tackled by the literature, an alignment is required to elucidate the connections between symptoms, that can be expressed in layperson terms, and diseases. The alignment task is not straightforward when using real case scenario in natural language involving patients interaction such as dialogue, Electronic Health Records, clinical notes or online resources. Layperson symptoms are symptoms expressed by the patient with his words and therefore, often does not match easily the medical vocabulary due to the difference of knowledge between patients and doctors about medicine. Manzini et al. [125], for instance, propose an automated system for translating layperson terminology to HPO terms. This system leverages a neural network and a vector space to generate and compare vector representations of medical and layperson terms. The main limitation of this approach is that it translates layperson terms without considering the context, potentially missing relevant information that may change the semantics of the term.

In addition to symptoms, medical explanations often rely on the results of health measurements, observations or vital signs (i.e., medical findings). Consequently, it is crucial to take into account and interpret these data. Looking at medical vocabularies and knowledge bases, it is noticeable that many symptoms are the result of abnormal values of medical findings such as abnormal temperature will be resumed as fever when too high of hypothermia when too low. Several recent studies focus on the automatic detection of medical findings in digitized patient records, such as Electronic Medical Records (EMR)[156] or EHRs[73, 124, 74]. However, none of these studies, to the best of our knowledge, concentrate on training exams. These exams can be clinical cases that utilize a different structure from EMR or EHR and often show a more narrative text, presenting symptoms, patient history, and lab results as part of a broader storyline.

Earlier contributions focusing on the extraction of medical findings and vital signs proposed rule-based approaches [156, 124, 74]. Although they obtained good results, they still require specialists to create and refine the rules, limiting their generalisability to other medical tasks. In contrast, the approach proposed by Gavrilov et al. [73] employs a deep learning strategy, training a model on Russian data using Bloom's embedding methods implemented in SpaCy. While these works showed good performance in detecting vital signs, their applicability range remains limited. First, they primarily focus on detecting the six fundamental vital signs : blood pressure, heart rate, respiratory rate, body temperature, height, and weight. Even if these vital signs are the most used in the literature [156, 124, 74, 73], some complementary analysis such as laboratory analysis are needed to confirm or discard a diagnosis. Since these findings are numerous and evolve with time, a rule-based system would require a large number of experts to create and maintain the rules. Genes et al. [74] offers a NER assigning also a quality score to each entity, computed according to a set of rules for each vital sign.

3.5.4 Medical explanations generation

Natural language explanation generation has received a lot of attention in recent years, grounding on the progress of generative models to train specific explanatory systems. Camburu et al. [33] generates explanations by justifying a relation (i.e., *entailment*, *contradiction* or *neutral*) for a premise-hypothesis pair by training a Bi-LSTM on their e-SNLI dataset (i.e., the Stanford Natural Language Inference [28] dataset augmented with an explanation layer which explains the SLNI relations). Kumar et al. [113] propose to generate short

explanations with GPT-2 [165], learned together with the input by a classifier to improve the final label prediction, using e-SNLI [33]. If these solutions propose interesting results, they are not applicable to the medical domain given that explaining a medical diagnosis is a sensible task which can hardly be restrained to the above-mentioned three basic relations (considered in [33] and [113]). On the other hand, Narang et al. [145] propose an approach based on the T5 model [166] to generate an explanation after a prediction. The problem with these approaches based on neural models is that we do not master the internal knowledge of these models, which can generate errors on the veracity of the data. Again, this solution is not applicable to the medical scenario, since explanations are required to be structured following precise argumentative structures [100, 35, 61] and to be grounded on verified medical knowledge.

Other approaches keep more control on explanations, using templates based explanations [173, 37]. For instance, Abujabal et al. [1] use templates and inject the reasoning steps and query of their Q&A system. To the best of our knowledge, no related work generates natural language post-hoc explanations under the form of arguments for the medical domain.

3.6 Conclusion

Through the presented SYMEXP pipeline, I explored the creation of natural language explanations to justify a correct diagnosis and discard others, using template-based symptomatic and argumentative explanations. To archive this result, we introduced two new resources, i) a new expert validated database of medical findings (i.e., vital signs, health measurements, observations, test results), interpreted terms and normal values and ii) a new dataset of clinical case question-answering documents, annotated with UMLS vocabulary entities related to diagnoses (i.e., symptoms, findings). More precisely, proposed pipeline *(i)* automatically identifies relevant medical entities in a clinical case description so as to explain a diagnosis using symptoms and medical findings. Medical findings are *(ii)* automatically interpreted and converted (if possible) to medical terms, relying on the expert database. If needed, we generated missing findings boundaries and associated terms on the fly with a LLM. Symptoms and converted findings are then *(iii)* aligned with the HPO medical knowledge base to be associated to the correct and incorrect diagnoses proposed as potential answers to the test. Finally, *(iv)* explanations are automatically generated based on natural language explanatory arguments patterns highlighting *why* a certain answer is the correct diagnoses and *why* the others are *not*.

Extensive experiments on a dataset of 314 clinical cases in English on various diseases show good results (0.86 on symptom detection, 0.78 on findings conversion and 0.70 on relevant symptom matching), outperforming competitive baselines and state-of-the-art approaches. Given the sensitivity of the medical domain and the fact that this system is intended as an example of AI in education and training, our explanations have a didactic goal which is exemplified through the enrichment of the data available in the clinical case description with further verified information from the HPO knowledge base. In our work, we have decided to adopt a method based on templates to generate explanations in order to avoid any hallucination problems associated with LLMs. I explore explanations assessment criteria and how they appear in medical question answering document, supporting

the choice of template-based argumentation according to the user requirement. Although this approach has its own limitations, such as being design-dependent, it provides a robust and verified strategy which is more suitable to the medical domain.

However, this first step into generating grounded explanations for diagnoses could be derived and completed from different aspects. First, as introduced in Chapter 2, explanations expectations will strongly depend on an explainee and, therefore, being able to focus on user-specific explanations is a great challenge. With the help of recent improvement in the natural language generation domain, making explanation process interactive could be an interesting path to explore. These interactions might lead to a closer human inference, to the best explanation and to better students understanding of clinical cases. Then, as discovered during the error analysis, investigating the structure and relations of ontologies might be helpful to make the explanation (and even predictions) systems more flexible and usable in real case scenario. For instance, one of the frequent errors in our experiments concerns cardiac arrhythmia, often a symptom linked to a diagnosis in ontologies but difficult to detect because it is often expressed "too" precisely via the symptoms "tachycardia" or "bradycardia", which are both forms of cardiac arrhythmia (low or high rhythm). Taking into account their relations, we will be able to improve the matching module performances by allowing more or less abstraction on matching concepts. This could be also completed by ameliorating medical knowledge. As we rely only on the HPO, we probably lack some knowledge and using a bridge with other vocabularies and knowledge bases (i.e., UMLS unified vocabulary) will enhance the arguments of our explanations. Moreover, given the difference in our performances to predict low and high boundaries (88% and 68%, respectively), it may be interesting to address further experiments, comparing our proposed automated methods and re-submit our results to undergo another expert evaluation. Finally, even though clinical doctors have been involved in the definition of the annotation guidelines, a user evaluation with med residents is required to get their feedback on our explanatory arguments.

CHAPTER 4

Assessing Argument based Natural Language Explanations

This chapter shows how, after the step addressing the generation of argumentative explanations in natural language, to automatically assess such explanations, relying on argumentation patterns. Specifically, I propose a fully automatic pipeline to extract argumentation structures from natural language explanations, and return a set of argumentation characteristics about the structure of the explanations. I investigate the structure of the argumentation within natural language explanations to highlight some patterns (e.g., the excessive usage of implicit claims) which can be used to assess explanations, depending on the target use case. To focus on the structure rather than the content, Argument Mining methods provide a suitable solution to extract the argumentation structure of a text (i.e., the retrieved arguments and their relations). In this work, I propose a set of patterns based on both empirical analysis of the gold argumentation graphs of medical explanations and the literature on argumentation-based explanations in philosophy. This chapter presents the contribution (currently under review) submitted to the ACM SIGAPP Symposium on Applied Computing (SAC-2025).

XAI approaches are detailed in Section 2.3 and can be broadly categorized into model-agnostic methods [174, 122, 181], which can be applied to any machine learning model, model-specific methods [192, 2, 94, 214] tailored to particular types of models, and by design methods [142], where the architecture of the model explains the predictions. This issue is particularly relevant in sensitive scenarios like medicine, law and defence. Natural language explanations may be constructed in different ways. One of the research directions in this domain consists in constructing argument-based natural language explanations, given the justification ability of argumentation in decision making [55, 201]. This was the object of the contribution I presented in Chapter 3 to generate natural language argument-based explanations.

In medicine, argument-based explanations may be employed to aid students to elucidate their reasoning behind a diagnosis [140]. In this chapter, we focus on argumentation structure of natural language explanations written by humans to justify the outcome of a diagnosis. More precisely, we propose a number of criteria to characterise the argumentation structure of these explanations. The proposed framework automatically tags the explanations with respect to these criteria, based on argumentation literature [198, 207, 22] and the empirical analysis of medical explanations in clinical case question-answering (QA). The proposed architecture first extracts the argumentation structure of the natural language explanation in the form of an argumentation graph, second, it automatically detects argumentation patterns in the extracted argumentation graphs. More specifically, we propose an approach based on pre-trained transformers to extract argumentative structures, finetuned on annotations of clinical cases from the Casimedicos dataset, which we assess through rules on argumentation graphs.

This Chapter is organised as follows : in Section 4.1 I introduce our proposed pipeline, named ABEXA, to characterise explanations from the argumentation viewpoint. In Section 4.2, I describe the first part of the system to extract the argumentation structure from medical documents. Then, in Section 4.3, I define the set of patterns I decide to use for assessing explanations, while in Section 4.4 I share the implementation details of both steps of the pipeline and I discuss the obtained results. Section 4.5 presents an overview of the related contributions. Finally, I conclude the Chapter in Section 4.6.

4.1 Natural Language argument based explanations assessment

In this section, we introduce our pipeline for assessing natural language explanations of medical question answering. Our approach aims to automatically identify patterns in explanations by analyzing their argumentation structure, relying on a manually defined set of criteria. The aim is to identify specific argumentation patterns in student explanation so that they can improve them and better structure their explanations.

More specifically, we have developed a pipeline that takes as input a QA consisting of a clinical case, its question (e.g., on diagnosis, treatment), a set of answers and the explanation. The argument structure is extracted from this document in the form of a graph using argument mining techniques based on transformers. Once the argumentation graph is obtained, we apply a set of rules to the graph to detect specific argumentation patterns and return them to the user. The full pipeline is illustrated in Figure 4.1.

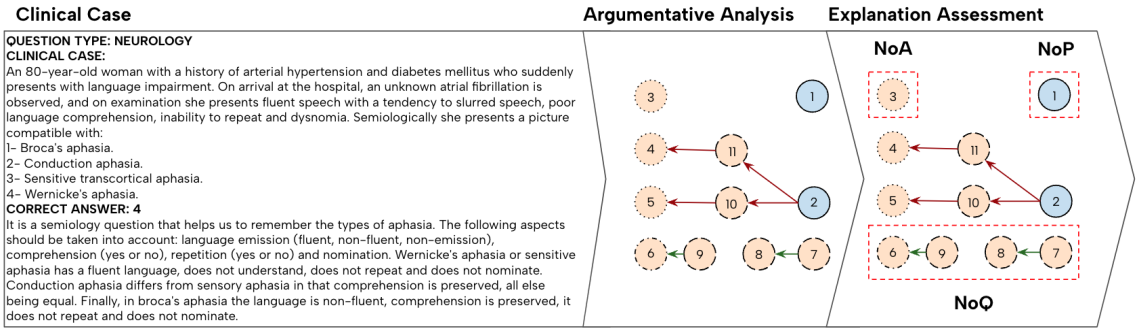

Figure 4.1 – Overview of our proposed ABEXA pipeline. Argumentation patterns are detected in the Explanation Assessment module, highlighted with red-dotted squares. The tags NoA, NoP and NoQ stands for No Answer, No Premise, No Question used in the explanation. We detailed all patterns in Section 4.3.

4.2 Argument Mining

The generation of the argumentation graph of a natural language explanation is necessary to be able to later assess the explanation itself and the detected patterns. In this section, we first discuss the dataset used in our experimental setting, and then we describe the results obtained with our pipeline in an end-to-end configuration.

4.2.1 Datasets

In order to assess natural language explanations in the medical domain, we decided to focus on two datasets to evaluate our approaches. We decided to select these datasets because both describe medical data and are annotated with argumentative structures.

Casimedicos Dataset : The Casimedicos dataset [3] has been annotated by Sviridova et al. [195] adding the argument components (i.e., premise, claim) and relations (i.e., support, attack). This dataset is made over Spanish medical exams with a medical context and a specific question over 64 different topics such as Psychiatry, Infectious diseases, Pediatry, ... Each document ends with a question which may vary from guessing the diagnosis to proposing a treatment or an intervention. Then four to five options are provided, according to the question type, where one is the correct answer and the other are not. Finally, Casimedicos is provided with an expert written explanation for each QA. As described in Section 2.4, the explanations are written by volunteers, who are different for each clinical case, and this is mirrored in the explanations as well, which result to be of different granularity. In comparison to the USMLE dataset used to generate explanations in Chapter 3, Casimedicos questions are broader than diagnostic questions, including all types of medical questions. An example of a Casimedicos QA is shown in example 4.2.1 where the discussed parts are indicated in bolds. Casimedicos annotations are presented as a conll file exemplified in Apendice A.3.

Exemple 4.2.1 – **QUESTION TYPE** : NEUROLOGY
CLINICAL CASE : In a woman with an epileptic seizure presenting with the following clinical features : epigastric aura, unpleasant odor, disconnection from the environment, motor automatisms (sucking, swallowing, opening and closing of one hand) and postcritical amnesia, what is

your diagnostic suspicion ?
OPTIONS :
1- Generalized non-convulsive seizure or typical absence.
2- Continuous partial epilepsy.
3- Amyotonic crisis.
4- Complex partial temporal lobe seizure.
CORRECT ANSWER : 4
EXPLANATION : Clearly the answer is 4, with a very characteristic clinic of temporary seizures.

AbstrRCT Dataset : The AbstRCT dataset [129] propose an argumentative annotation of “Claim” and “Premise” components, “Support” and “Attack" relations over abstracts of Randomized Clinical Trial (RCT) available on PubMed [1]. PubMed is a free search engine accessing primarily the MEDLINE database [2] of references and abstracts on life sciences and biomedical topics. We showed in Example 4.2.2 an example of the AbstRCT dataset, the annotations have the same format than Casimedicos.

Exemple 4.2.2 – To investigate the effects of medroxyprogesterone acetate (MPA) on appetite, weight, and quality of life (QL) in patients with advanced-stage, incurable, non-hormone-sensitive cancer. Two hundred six eligible patients were randomized between double-blind MPA 500 mg twice daily or placebo. Appetite (0 to 10 numerical rating scale), weight, and QL (European Organization for Research and Treatment of Cancer Quality of Life Questionnaire [EORTC-QLQ-C30]) were assessed before the start of treatment (t = 0), and 6 weeks (t = 6) and 12 weeks (t = 12) thereafter. One hundred thirty-four patients (68 MPA and 66 placebo) were assessable at t = 6 and 99 patients (53 MPA and 46 placebo) at t = 12. A beneficial effect of MPA on appetite was observed after both 6 weeks (P = .008) and 12 weeks (P = .01) of treatment. After 12 weeks, a mean weight gain of 0.6 +/- 4.4 kg was seen in the MPA, versus an ongoing mean weight loss of 1.4 +/- 4.6 kg in the placebo group. This difference of 2.0 kg was statistically significant (P = .04). During the study, several areas of QL deteriorated in the total group of patients. With the exception of an improvement in appetite and possible also a reduction in nausea and vomiting, no measurable beneficial effects of MPA on QL could be demonstrated. The side effects profile of MPA was favorable : only a trend toward an increase in (usually mild) peripheral edema was observed. In weight-losing, advanced-stage non-hormone-sensitive cancer patients, MPA exhibits a mild side effects profile, has a beneficial effect on appetite, and may prevent further weight loss. However, general QL in the present study was not measurably influenced by MPA treatment.

Overall, the Casimedicos Dataset includes 4125 Claims and 1932 Premises, with 2431 Support relations and 1106 Attack relations annotated across its 553 clinical cases. Similarly, the AbstrRCT Dataset consists of 700 annotated abstracts from PubMed, containing a total of 1488 Claims and 2985 Premises, along with their corresponding 2402 Support and 355 Attack relations.

4.2.2 Methodology

Our goal is to provide an end-to-end pipeline, taking a natural language text as input and returning the argumentation graph to the Explanation Assessment module. As

1. https://pubmed.ncbi.nlm.nih.gov/
2. https://www.nlm.nih.gov/medline/medlineoverview.html

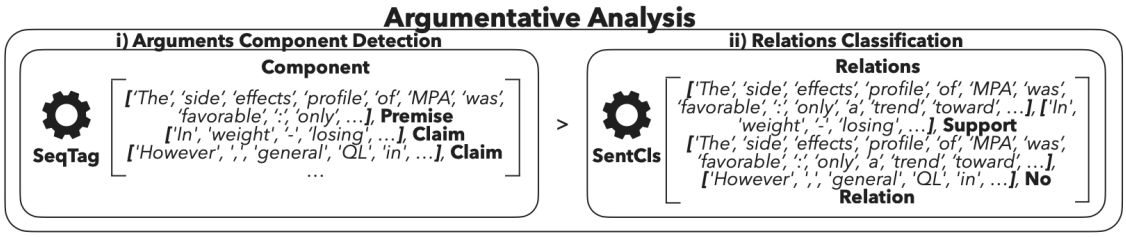

Figure 4.2 – Argumentative Analysis Pipeline with i) Argument Components Detection forwarded to the ii) Relations Classification

a baseline, we selected the two state-of-the-art approaches for end-to-end argument mining : i) the ACTA system [36] trained over the AbstRCT dataset, taking advantage of transformer-based Language Models such as BERT [58] and SciBERT [20] ; ii) the Multi-Task Argument Mining (MT-AM) approach of Morio et al. [144], a state-of-the-art model for AM over scarce resources and various corpora. Both of the approaches tackled the tasks of Argument Component Detection and Relation Classification, but Morio et al. opted for a subdivision in Span Identification and Component Classification, following Stab et al. [187].

Morio et al. [144] compared their approach to the original ACTA implementation but their model did not perform better, so we decided to follow the ACTA approach [129] for both subtasks due to the higher results on the overall pipeline. The Argumentative Analysis pipeline is shown in Figure 4.2. Therefore, we experimented first on the AbstRCT dataset and tackled the Argument Component Detection by casting it into a sequence tagging problem, using the 5 classes **BIO** scheme tag [170], detecting the **B**eginning and **I**nside of "Claim" and "Premises" as well as the **O**utside of a component. We slightly modified the original neural architecture employing the BERT-based model followed by a RNN, here we relied on a GRU [43] and a Conditional Random Field (CRF [114]) by removing the CRF, achieving then better results on the AbstRCT dataset. For the Relation Classification, we tackle this task as a sequence classification problem, predicting for each pair of detected components if the relation is "Attack", "Support" or "No Relation".

In addition to SciBERT, we also addressed experiments using more recent BERT-based transformers such as ClinicalBERT [208], Bio_ClinicalBERT [7] and PubMedBERT [77].

All these pre-trained models have been finetuned independently for each task namely i) Argument Component Detection as a Sequence-Tagging problem, and ii) Relation Classification casted into a Sequence Classification problem.

4.2.3 Evaluation and results

For the end-to-end pipeline evaluation, we proposed to evaluate each documents one by one, aligning the predicted components with the gold components to retrieve the expected relations. As we are tackling the span identification and the classification of the component label in a single sequence labeling step, we need to align the detected component with the component expected in the gold standard. This step, illustrated in Figure 4.3, is necessary because the relations gold standards of the Casimedicos dataset are built over the gold of argument components and if the prediction of the Argument Component task is not

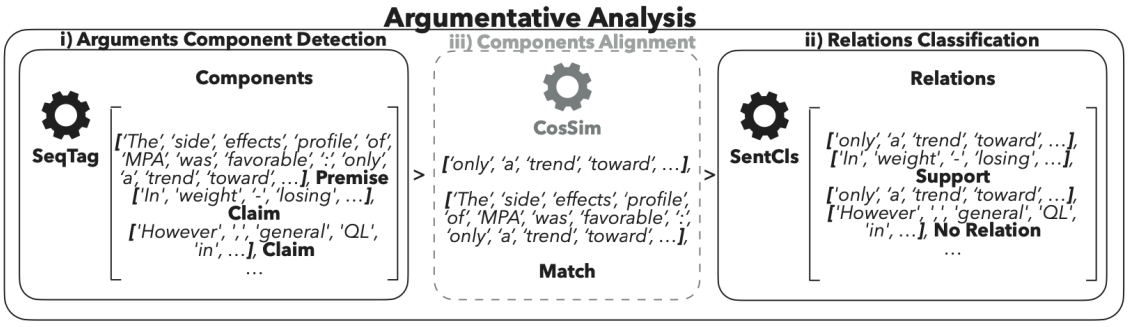

Figure 4.3 – Argument Mining end-to-end pipeline with alignment step.

precisely the gold one it still could be considered as correct. We exemplify this scenario in example 4.2.3 where the Detected Component (DC) missed some tokens compared to the Gold Component (GC).

Exemple 4.2.3 – **GC** : The side effects profile of MPA was favorable : only a trend toward an increase in (usually mild) peripheral edema was observed.
DC : only a trend toward an increase in (usually mild) peripheral edema was observed.

To automatically address this alignment, we proposed some heuristics. More precisely, we evaluated two different methods starting to count similar tokens with different thresholds (50, 80 and 100% of similar tokens), and to compare the similarity of the detected and gold components (through cosine and Levenshtein distance metrics). Once the components have been aligned with our goldstandard, we could recreate our new goldstandard dataset for relation prediction task, as shown in Example 4.2.4.

Exemple 4.2.4 – **Initial Gold** : **T6**[The side effects profile of MPA was favorable : only a trend toward an increase in (usually mild) peripheral edema was observed.] → Support → **T7**[In weight-losing, advanced-stage non-hormone-sensitive cancer patients, MPA exhibits a mild side effects profile, has a beneficial effect on appetite, and may prevent further weight loss.]
New Gold : **T1**[only a trend toward an increase in (usually mild) peripheral edema was observed.] → Support → **T2**[In weight-losing, advanced-stage non-hormone-sensitive cancer patients, MPA exhibits a mild side effects profile, has a beneficial effect on appetite, and may prevent further weight loss.]

This example of goldstandard reconstruction aims at evaluating our performances and does not change the detected argument components but only retrieve the expected argumentative relations. In Table 4.1, we report the results of our end-to-end argument mining pipeline, using two evaluation methods, i.e., Flexible and Strict. The Flexible evaluation considers that each *NoRelation* relationship will be valid even if the component is not detected during the argument component detection step, whereas Strict will consider these relationships to be false because the component has not been detected. For a fair comparison with Morio et al., all the experiments are tested over the neoplasm subset of abstRCT dataset.

This new architecture allows to enhance the result of Mayer et al. [128] from 0.55 to 0.57 for our equivalent SciBERT T50, having all models reaching similar good performances

	ArgComp	Alignment	RelClass	Flexible	Strict
Morio ST*	0.8937	-	-	-	0.3191
Morio MT*	0.8923	-	-	-	0.3394
Mayer	0.85	-	0.68	-	0.55
Us SciBERT T50	0.83	**0.9315**	0.73	0.6528	**0.5730**
Us SciBERT T80	0.83	0.8805	0.73	0.6130	0.4778
Us SciBERT T100	0.83	0.3017	0.73	0.2403	0.0374
Us SciBERT COS	0.83	**0.9315**	0.73	**0.6562**	0.5701
Us SciBERT LEV	0.83	**0.9315**	0.73	0.6516	0.5637
Us PubMedBERT T50	0.82	0.9198	0.71	0.6272	0.5224
Us PubMedBERT T80	0.82	0.8717	0.71	0.5930	0.4492
Us PubMedBERT T100	0.82	0.2945	0.71	0.2448	0.0424
Us PubMedBERT COS	0.82	0.9227	0.71	0.6430	0.5469
Us PubMedBERT LEV	0.82	**0.9271**	0.71	**0.6343**	**0.5523**
Us BioClinicalBERT T50	0.81	0.9213	0.71	0.6301	0.5392
Us BioClinicalBERT T80	0.81	0.8790	0.71	0.6051	0.4742
Us BioClinicalBERT T100	0.81	0.3324	0.71	0.2870	0.0498
Us BioClinicalBERT COS	0.81	0.9257	0.71	**0.6428**	0.5525
Us BioClinicalBERT LEV	0.81	**0.9300**	0.71	0.6296	**0.5407**
Us ClinicalBERT T50	0.73	0.8222	0.50	0.5298	0.3629
Us ClinicalBERT T80	0.73	0.7857	0.50	0.5168	0.3227
Us ClinicalBERT T100	0.73	0.3147	0.50	0.2686	0.0467
Us ClinicalBERT COS	0.73	0.8338	0.50	0.5468	0.3889
Us ClinicalBERT LEV	0.73	**0.8528**	0.50	**0.5524**	**0.4163**

TABLE 4.1 – Argument mining results for component detection, component alignment, and end-to-end evaluation (Flexible and Strict) using Macro F1 over the subset of the AbstRCT dataset on neoplasm.

over the end-to-end pipeline. According to experiments, SciBERT and PubMedBERT perform the best, mostly with the T50, COS and LEV configurations for alignment methods. SciBERT is still the best performing model. Finally, we fine-tuned both SciBERT and PubMedBERT over the Casimedicos dataset with T50, COS and LEV alignment methods. Obtained results are reported in Table 4.2.

	ArgComp	Alignment	RelClass	Flexible	Strict
Us SciBERT T50	0.91	0.8721	0.46	0.4639	0.327
Us SciBERT COS	0.91	0.9202	0.46	0.4826	0.3862
Us SciBERT LEV	0.91	0.9506	0.46	**0.4904**	**0.4312**
Us PubMedBERT T50	0.90	0.8609	0.49	0.4746	0.3219
Us PubMedBERT COS	0.90	0.9183	0.49	0.5002	0.3995
Us PubMedBERT LEV	0.90	0.8609	0.49	**0.5084**	**0.4406**

TABLE 4.2 – Argument Mining results expressed with Macro F1 over the Casimedicos dataset.

The results over Casimedicos are way lower than on AbstRCT from 0.63 with PubMedBERT combined with LEV alignment versus 0.51 in Casimedicos. We assume it is mostly due to the fact that AbstRCT is using Randomized Clinical Trials, that have an imposed structure whereas medical QA vary more in the structure of the text. Also the medical QA explanations and questions are written by different authors and on differents topics, making the argumentative structure more complex to learn than in RCT where the entire document is written by the same author. Finally, we observed that Casimedicos contains a lot of coreferences, linking the components by formulations like "Option 1 can be discarded...". Therefore we retrained and reevaluate our pipeline on a new version of Casimedicos where coreferences are detected using a set of regex and replacing these occurrences by the content of the answer like in Example 4.2.5. Results of coreference experiments are shown in Table 4.3.

Exemple 4.2.5 – Among the other 2, it is important to know that 5 is correct
Among the other 2, it is important to know that **5- Microsatellite instability and DNA error repair genes should be studied.** is correct

	ArgComp	Alignment	RelClass	Flexible	Strict
Us SciBERT T50	0.73	0.8002	0.52	0.4355	0.2648
Us SciBERT COS	0.73	0.8741	0.52	0.4600	0.3407
Us SciBERT LEV	0.73	0.8978	0.52	**0.4656**	**0.3672**
Us PubMedBERT T50	0.72	0.7864	0.52	0.4494	0.2708
Us PubMedBERT COS	0.72	0.8655	0.52	0.4731	0.3509
Us PubMedBERT LEV	0.72	0.8827	0.52	**0.4849**	**0.3835**

TABLE 4.3 – Argument Mining results without coreferences expressed with Macro F1 over the Casimedicos dataset.

Coreferences resolution on the best performing model for Casimedicos is showing a minor improvement on relation classification (from 0.49 to 0.52) but a degradation of the results from 0.5084 to 0.4849 of macro f1 through the Flexible evaluation.

4.3 Explanation Assessment

We propose here a new modular system to assess explanations from the argumentation structure viewpoint. Our goal is not to classify explanations as being good and bad, but proposing different criteria to aid users in semi-automatic assessment of argumentative explanations, according to their needs and use cases.

In this section, we first describe the dataset focusing on explanations, before introducing the set of argument-based heuristics. Finally, we provide the results over the entire pipeline.

4.3.1 Argument components in Casimedicos' explanations

The Casimedicos dataset [3] contains QA documents with both the Question, the potential Answers and the Explanation written by an external doctor. The explanations are not written by the same author as the questions. Therefore, explanations differ a lot the ones from the others, in terms of writing style and argumentation process used to justify the correct answer. We report in Table 4.4 the argumentation distribution of components and relations in Casimedicos with respect to their presence in the explanation.

Label	Total	Mean	In Explanations	Mean per Explanation
Claim	4125	8.93	3003	5.948
Premise	1932	4.18	470	0.935
Support	2431	4.36	-	-
Attack	1106	1.98	-	-

TABLE 4.4 – Distribution of argumentation components and relations in the Casimedicos dataset, by Cases and by Explanations.

It might be noticed that these documents (Question, Answers and Explanation) contain twice more claims than premises. They introduced more than five times more claims than premises in the Explanation showing that explanations are mainly focused on making assertions (claims) rather than providing supporting reasons or evidence (premises). When we look at the interactions between the different components, we can notice that not all components are used in the document nor in the explanation, as shown in Table 4.5.

Label	Total	Mean per Case
Claim	1527	2.74
Premise	1446	2.60

TABLE 4.5 – Total and average per Case of unused Argumentative Components.

4.3.2 Argumentation-based patterns for explanations

In this section, we propose a set of graph-based criterias to assess the argumentation structure of natural language explanations. These criteria may then be used by users to select those explanations which satisfy the criteria which are relevant for their precise use cases. The proposed argumentative explanation tags and their corresponding descriptions are used to assess the structure of explanations. The criteria are as follows :

— **Inconsistent (Inc) :** indicates a conflict or incoherence in the argumentation graph.
— **No Explanation (NoE) :** is applied when the explanation is composed of a simple argument component, it does not contain any component or is an empty text.
— **No Question (NoQ) :** is triggered when less than 80% of the components in the document are linked to the question.
— **No Premise (NoP) :** appears when less than 80% of the components from the explanation are linked to a premise.
— **No Answer Positive (NAP) :** is activated if the correct answer is not addressed in the explanation.
— **No Answer Negative (NAN) :** is triggered when less than 80% of incorrect answers are tackled.
— **No Length (NoL) :** applies when the explanation is shorter than 238 characters.
— **Too Long (ToL) :** is used for explanations exceeding 616 characters.
— **Implicit (Imp) :** is triggered when more than 50% of the components are introduced in the explanation.

In the following argumentation graphs, as visualized in Figure 4.4, orange nodes represent claims and blue nodes represent premises. The border of a node expresses where the component appears in the text of Casimedicos : the full line is for components in **Questions**, the dashed line for **Explanation**, and the dotted line for the **Answers** (where the half dashed half dotted line is for the **correct answer**). Green and red edges between two nodes represent the support and attack argument relations, respectively. The following paragraphs details each criterion with examples.

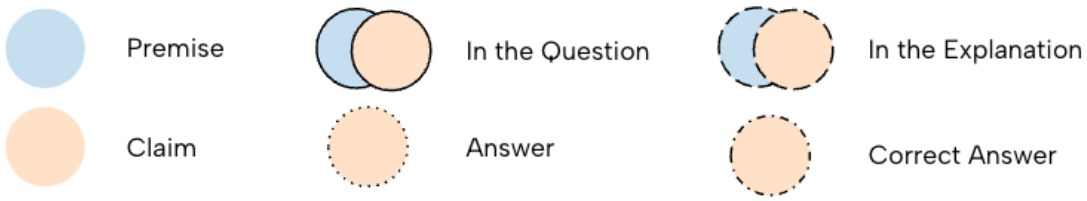

Figure 4.4 – Components types and positions in documents.

Incosistent (Inc) By analysing the explanations, we founded out that some of the argumentation graphs show inconsistent argumentation patterns, as shown in Figure 4.5. In this example, claim 12 both indirectly attacks (by attacking the premise 13 wich supports claim 14) and support claim 14. To detect this kind of patterns, we created a set of graph-based rules that trigger the Inconsistent tag if in a QA document, one component is both (indirectly) attacked and (indirectly) supported by another argument component. Only 3 documents where identified in Casimedicos showing such inconsistency.

No Explanation (NoE) The No Explanation (NoE) tag aims to detect wether an explanation is empty (this situation does not appear in the Casimedicos dataset) or if the explanation contains one single component. In this case, the argumentation graph results to be poor and probably many (incorrect) answers are not tackled in the explanation.

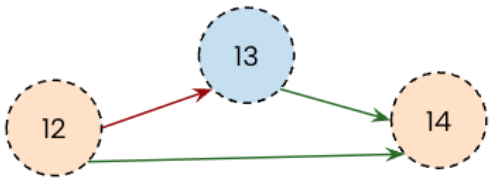

Figure 4.5 – Inconsistent graph where the Claim 12 both Attack (indirectly) and Support Claim 14.

Exemple 4.3.1 – Clearly the answer is 4, with a very characteristic clinic of temporary seizures

Exemple 4.3.2 – They are undoubtedly describing the typical lesions (both on skin and oral mucosa) of a lichen planus (5)

Exemple 4.3.3 – It never ceases to amaze us the strange ways of biology and that the axiom "the longest is the right one" is fulfilled in this case

Examples 4.3.1, 4.3.2 and 4.3.3 shows the kind of explanations with only 1 claim in the explanation. Over the goldstandard argumentation graphs of Casimedicos, 39 cases appears to be tagged NoE as only one component was present in the explanations. Over these claims, 25 of them are backed by the Question. Figure 4.6 illustrates this scenario of example 4.3.1.

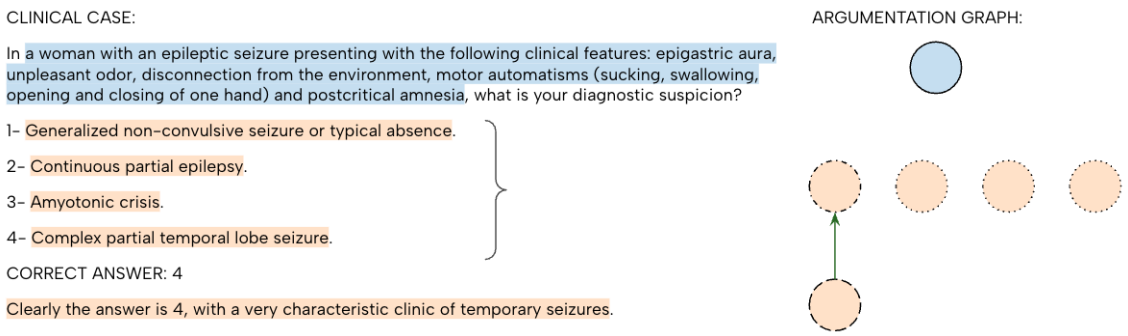

Figure 4.6 – Example of the Question 452_149 with only one component (claim) in the explanation, not backed by any components from the question.

No Question (NoQ) As a second criterion, we decided to focus on the cases where the elements provided in the Question are not used (or at least linked) in the formulation of the Explanation. Example 4.3.4, and the related argumentation graph visualized in Figure 4.7, exemplify this criterion by showing the claims from the Explanation are mainly related to the answers rather than the Question.

Exemple 4.3.4 – **QUESTION TYPE** : ENDOCRINOLOGY
CLINICAL CASE : 14-year-old girl who consults for decreased growth for 2-3 years previously normal (provides data) and that other girls her age have greater physical and sexual development. Lately she has had headaches and visual problems that she notices in class and when studying. She

has not had menarche or polydipsia or polyuria. Parents with normal height. Examination : short stature at -2.1 standard deviations, normal body proportions, little pubic hair and breast development. Campimetry shows left temporal partial hemianopsia. Bone age : delay of 2 years. General laboratory tests were normal. Gonadotrophins (FSH and LH) and estradiol are low. What do you think is the most appropriate response ?
1- Decreased growth and sexual development, delayed bone age, headache and visual alteration suggest hormonal deficit and involvement of the optic chiasm.
2- As she is a girl of pubertal age, it is most likely that her decreased growth and sexual retardation are due to Turner syndrome.
3- She must not have a hypothalamic tumor because of the absence of polyuria and polydipsia. She probably has constitutional delay and her visual problem is refractive.
4- A growth hormone deficiency may explain the developmental delay and low estradiol. To evaluate if she needs glasses, due to her headaches and visual disturbances.
5- She could have a craniopharyngioma, but it would be rare if she had not shown symptoms before. Also, it would not justify low gonadotrophins and estradiol.
CORRECT ANSWER : 1
Answer 2 is false (Turner syndrome : low estradiol and elevated gonadotrophins), brain tumors affecting the hypothalamus-pituitary gland do not give low gonadotrophins (5 false), and it seems obvious that refractive defects do not give hemianopsia (3 and 4 false).

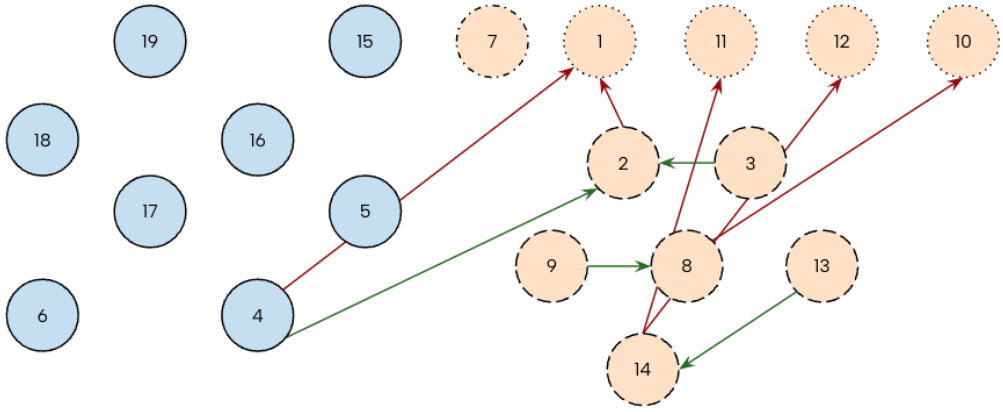

Figure 4.7 – Example of the Question 56_76 with the correct answer (claim 7).

We can observe that numerous components are not used in this explanation, and most of them are from the question. We observed 498 occurrences of this tag over 553 using a threshold of 20% of unlinked components.

No Premise (NoP) In the explanatory literature [85, 38], the Explanans (i.e., the statements that provide the explanation, based on relevant facts) plays a central role in the explanation. Therefore we decided to highlight when an explanation does not use the Premises, mainly present in the Question. We noticed that most of the Premises are not linked to any component (1446/1932 according to Table 4.4 and 4.5). For instance, in

Example 4.3.4 visualized in Figure 4.7, we can notice that the majority of unemployed components are Premises, and only 1/8 is related to the explanation. Figure 4.8 shows a correlation between the lack of use of premises, strongly correlated with the lack of use in the explanation of the Question element.

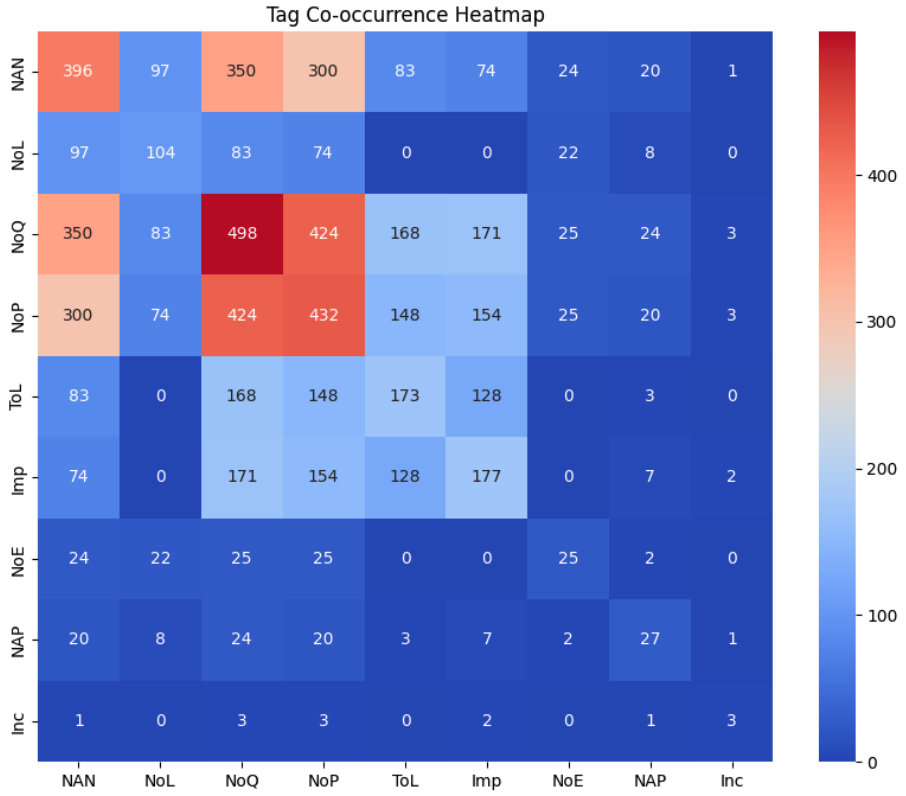

Figure 4.8 – Co-occurence matrix of tags over gold labels.

In this example, the Claims introduced by the doctor in the Explanation are implicit, and the understanding of the explanation could improve by making a more heavy use of the Premises available in the Question. This tag is the second most observed tag in Casimedicos and it is strongly correlated to the NoQ pattern, according to the co-occurrence matrix visualized in Figure 4.8.

No Answer Positive (NAP) This tag is assigned when the claim contained in the correct answer (or at least one of the claims of the correct answer) is not linked to any other components in the document. We focused on this tag because it represents the most common way to explain the answer, observed in 526 case out of 553. For instance, in example 4.3.4 and its visualisation in Figure 4.7, we can see that the argumentation of the explanation does not tackle the correct answer. In this case, the doctor explains the outcome of the case by discarding the incorrect answers and does not tackle the correct one. Another reason relies on the fact that we spot some human errors during the manual annotations, as in the graph visualized in Figure 4.9, which shows one annotation error of the example 4.3.5 where the annotator missed a relation (symbolised with green dotted arrow).

Exemple 4.3.5 – **QUESTION TYPE** : DIGESTIVE SYSTEM
CLINICAL CASE :
18-year-old young man with a history of asthma, allergy to pollens, mites and cat hair, comes to the emergency room referring sensation of food detention at retrosternal level with practical inability to swallow his own saliva. He refers similar episodes on other occasions that have subsided spontaneously within a few minutes. Which of the following is the most likely diagnosis ?
1- Barrett's esophagus.
2- Distal esophageal ring (Schatzki).
3- Infectious esophagitis.
4- Eosinophilic esophagitis.
CORRECT ANSWER : 4
This is intermittent dysphagia. Barrett's esophagus does not necessarily imply peptic stricture, but assuming it does, it is progressive. Infectious esophagitis is more typical of immunocompromised patients. That leaves distal esophageal ring and eosinophilic esophagitis ; both are possible, but the insistence on the patient's atopic burden indicates the likelihood of the latter.

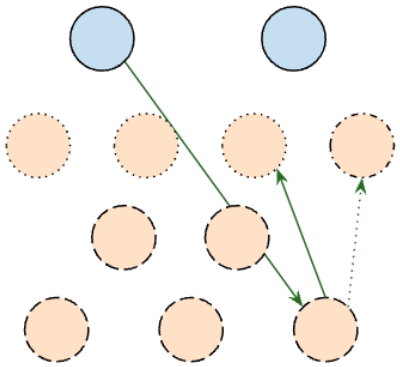

Figure 4.9 – Example of the Question 273_71 with the green-dotted link missed by the annotator in the annotation.

We noticed this tag with 27 occurrences over 553 cases. It means that only 27 cases does not tackle the correct answer in the explanation text.

No Answer Negative (NAN) Similar to the NAP pattern, the No Answer Negative (NAN) criterium aims to show the cases where at least one incorrect answer is not linked to any other components in the explanation. This is the most common scenario because the Casimedicos dataset contains four to five answers for each question, and only one is correct, making many components to tackle in the explanation to justify why the correct answer is correct, but also to explain why the wrong answer is not correct. Example 4.3.6 shows that the explainer only tackles the correct option.

Exemple 4.3.6 – **QUESTION TYPE** : NEUROLOGY
CLINICAL CASE :
In a woman with an epileptic seizure presenting with the following clinical features : epigastric aura, unpleasant odor, disconnection from the environment, motor automatisms (sucking, swallowing, opening and closing of one hand) and postcritical amnesia, what is your diagnostic suspi-

cion ?
1- Generalized non-convulsive seizure or typical absence.
2- Continuous partial epilepsy.
3- Amyotonic crisis.
4- Complex partial temporal lobe seizure.
CORRECT ANSWER : 4
Clearly the answer is 4, with a very characteristic clinic of temporary seizures.

Not/Too Long (NoL-ToL) We now introduce two tags computing the length of the Explanation in characters. We addressed an empirical analysis to define the default criteria in terms of length of the Explanation (i.e., Figure 4.10). This analysis shows that, in the Casimedicos dataset, Explanations have a 25th percentile (Q1) of 238 characters, a Median (50th percentile) of 398 characters, and a 75th percentile (Q3) of 616 characters. This analysis justifies our defaults setting of 238 and 616 characters to detect long and short explanations. These tags occur rarely with only 104 and 173 for NoL and ToL, respectively.

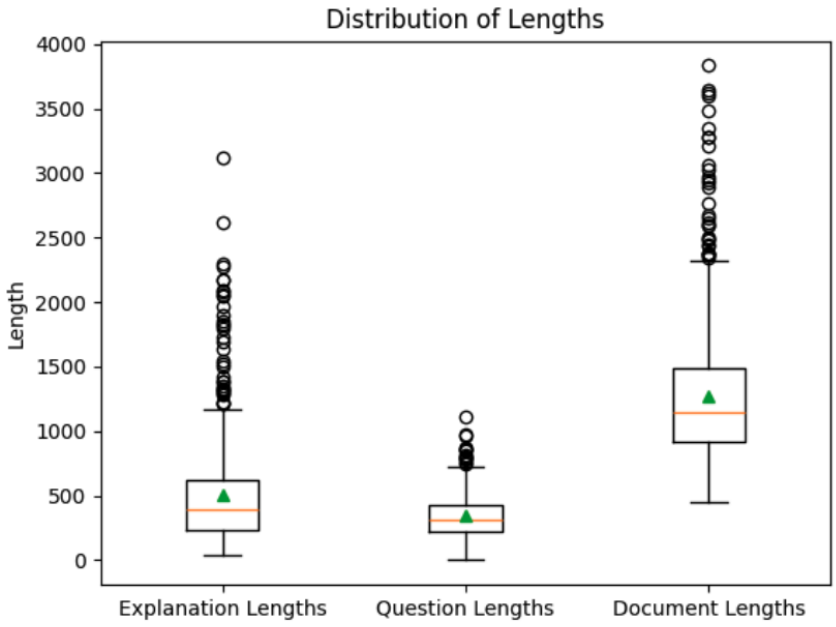

Figure 4.10 – Casimedicos Explanations, Questions and Document (entire QA) lengths distribution.

Implicit (Imp) Finally, we decided to add a final criterion to detect an overuse of components introduced in the explanation. This case is exemplified in Figure 4.7 and Example 4.3.4 where almost all the linked components are from or to a component introduced in the Explanation. This criterion relies on the fact that the argumentation scheme is based on the explainer knowledges rather than the Premises or the Question of the QA

document, like in the following example : “and it seems obvious that refractive defects do not give hemianopsia (3 and 4 false)”. We detected 177 documents with the Implicit tag.

4.4 Experimental settings and results

In this section, we report on the results obtained by our experimental setting of the proposed ABEXA pipeline which starts with the argument mining module to the explanatory assessment one on the extracted argumentation graphs.

We replicated the experimental settings used by Mayer *et al.*[128] for the argument component module, based on the results from the updated version of ACTA [36], presented in details in Chapter 5. We also removed the CRF[114] layer from the original approach. This task is still casted as a sequence tagging problem using the BIO-tagging scheme for modeling. Thus, for token-level representation of contextualized sentences, we use the pre-trained bidirectional transformer language model SciBERT [20] rather than BERT base [58], which we fine-tuned for three epochs using an Adam optimizer with a learning rate of 2e-5. The sentence representation is then passed into a RNN, specifically a GRU [43].

Regarding relation prediction, we followed the approach of Mayer et al.[129] and casted it as a sequence classification problem, utilizing a bi-directional transformer. The problem is tackled by generating all possible combinations between components and passing them through a softmax linear layer to classify each combination into one of three target classes : *Support*, *Attack*, and *NoRelation*. As Casimedicos has more components and therefore way more pairs to train the model than AbstRCT, we have therefore sub-sampled 30% of label *NoRelation* to make training faster and less expensive, without influencing the results. We also use the SciBERT uncased base model with pre-trained weights for sentence representation, fine-tuned with a learning rate of 2e-5, a batch size of 8, and a maximum sentence length of 256 sub-word tokens per input example for three epochs. The weight factor for each of the three classes in the weighted cross-entropy loss is the normalized number of training samples for each class.

To tag the explanations along with the argumentative criteria we defined, we created a script in Python 3.9.19 that retrieve all the documents' ann file and create the associated argumentation graph using networkx [3] version 3.2.1. Each graph is then provided to a set of functions, one per tag, that return True or False if the heuristic threshold is reached. Each node of the argumentation graphs is associated to their types (i.e., Claim or Premise) and their labels (i.e., their position in the clinical case like Question, Answer, Correct Answer, Explanation). Each edge is headed with the type Attack or Support.

The occurrence of each tags in the Casimedicos dataset are showed in Table 4.6, and the tags co-occurrence matrix is available in Figure 4.8 and in Figure 4.11 for the pipeline outputs.

3. https://pypi.org/project/networkx/

Label	Gold	Gold ratio (%)	Pred	Pred ratio (%)
NAN	396	71.61	86	73.50
NoL	104	18.81	25	21.37
NoQ	498	90.05	100	85.47
NoP	432	78.12	92	78.63
ToL	173	31.28	37	31.62
Imp	177	32.00	41	35.04
NoE	25	4.52	3	2.56
NAP	27	4.88	5	4.27
Inc	3	0.54	1	0.85

TABLE 4.6 – Label Occurrences from Heatmaps with Ratios over 553 documents for Gold and 117 for Pred. Ratio represent the percentage of documents over the dataset with this tag.

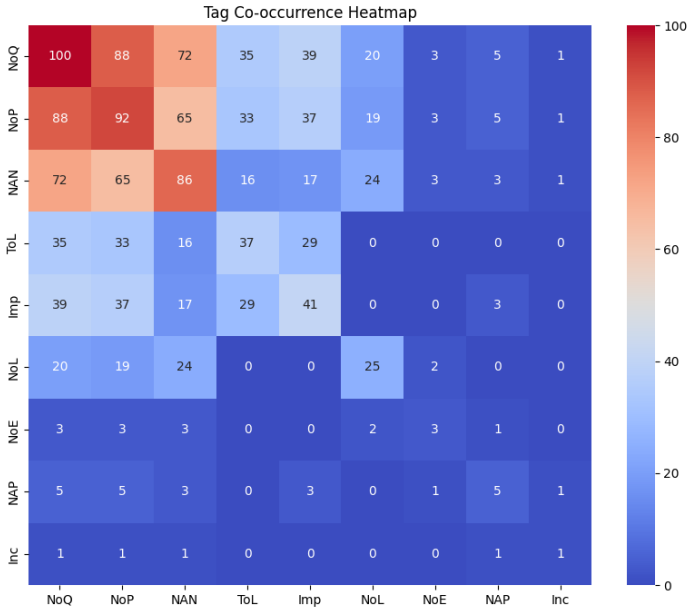

Figure 4.11 – Co-occurence matrix of taggs over prediction labels.

According to Table 4.6, our pipeline detects about the same proportion of tags. As presented in Section 4.2, the results on the Casimedicos dataset are lower than AbstRCT, with a best performing model that reach 0.5084 of macro F1 score. However, we can observe that the detected tags respect the same proportions and do not differ from the goldstandard documents.

4.5 Related Work

In the medical domain, some contributions already proposed Argument Mining pipelines [129, 36, 144]. A recent contribution of Kawarada et al. [103] employs LLMs to tackle argument mining in the medical domain as a single text-to-text task. Concerning the assessment of argument structures, many approaches have been proposed in the literature [203]. However, to the best of our knowledge, no work focuses on assessing argument-based natural language explanations. This is indeed the goal of this contribution, showing an application in the medical domain.

The assessment of explanations is still an open research question [134]. Recent approaches in the literature identified dimensions that allows to get more insights over an explanation [179]. According to the literature, a key element to evaluate an explanation is the user itself [134, 87, 42]. Another element is multidimentionality [179, 31] of explanations, i.e., being true and convincing. In this line, Zhou et al. [224] provides a high-level overview of explanatory methods in Machine Learning (ML) without focusing on natural language data. Focusing on textual data, Schuff et al. [179] identify two ways to evaluate explanations in natural language. A first method is to use a proxy score, computed automatically such as Accuracy, Recall and F1 metrics. Recent works [220, 136] focus on explanation generation in natural language across the multi-hop question answering task which identifies relevant paragraphs, determines supporting facts, and then predicts the correct answer. Due to the nature of the HotpotQA dataset, they evaluate the explanation using the F1 score [136] or both F1 score and Exact Match (EM) metrics [68, 120]. Other approaches [149, 200] predict supporting facts together with the answer to tackle interpretable reading comprehension, reporting Accuracy, Precision and Recall. Focusing on student peer discussions, Chou et al. [45] proposed an explanation assessment system trained on an expert annotated data to classify good and bad explanations using the Accuracy and Standard deviation metrics. The second kind of metrics used to evaluate natural language explanations are word embedding based metrics, inspired from natural language generation such as BLEU [155] or BERT-score [222]. Some approaches [33, 145] propose to use the BLEU metric on top of Perplexity and Accuracy evaluations. Clinciu et al. [49] investigate the correlations between automatic metrics and human ratings were computed using the Spearman correlation coefficient observing that embeddings based metrics performed better than word-overlap ones but still relatively far behind human rating.

Human evaluation is central to evaluate explanations but still suffer from the subjectivity and dimensionality of the task. Only few approaches proposed a human evaluation on top of the automatic ones, using crowdsourced solutions [145, 179]. Human evaluation criteria are mainly explanation utility, consistency, correctness and usability, and mental effort [179].

4.6 Conclusion and Discussion

In this chapter we presented a full pipeline named ABEXA to assess and characterize natural language explanations within the context of medical question answering. This approach tackles the assessment from the argumentation prism, providing an Argument Mining module to retrieve the argumentative structure from the raw text. Our system

returns a list of tags associated to the QA document, highlighting certain argumentation-based characteristics of the explanations. To the best of our knowledge we achieved state-of-the-art results for argument mining on clinical cases accounting for 0.65 of macro f1 in our end-to-end configuration. We have also analysed the extracted graphs and obtained similar proportion of detected patterns, showing that our approach can automatically detect argumentation patterns. The provided framework allows the user to configure the system according to the required level of granularity and the targeted use case.

There are multiple aspects that could be explored in the continuation of this work at the crossroad between explanable AI, argumentation, and medicine. As ABEXA evaluates medical explanations in natural language, a potential future line research consists in coupling it with the generation of diagnosis explanations discussed in chapter 4 to characterise the generated explanations. It would be interesting to explore its adaptation to other specific domains such as law or politics to understand both the argumentation patterns and see how they differ in between domains. Then, according to the trend around large language models, another future work line will be to explore how to generate explanations which enhance the existing method depending on one of the selected criteria. This framework would be a contribution in AI and education to enhance the medical residents' critical thinking and argumentative skills.

CHAPTER 5

Implementation of Argumentation-Driven Explainable AI for Medicine

This chapter introduces the tools that I developed or participated in the development to achieve argumentation-driven XAI in medicine. More specifically, it focuses mainly on argumentation tools such as the extension of the Argumentative Clinical Trial Analysis (ACTA) features, introduced in Chapter 4, to ease the clinician review of the literature. Then it presents MedMT5, an investigation on how multilingual text-to-text large language models specialized on medical data can enhance the performance of the argument mining task over non-english data. Finally, I present the ANTIDOTE software suite that showcase the tools developed for the project, described in Chapter 1. This chapter brings together the contributions published at the International Joint Conference on Artificial Intelligence (IJCAI 2022) [141] and European Conference on Artificial Intelligence (ECAI 2024) [36] demo tracks as well as the paper we published in the Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) [72].

This thesis is a contribution in the context of the ANTIDOTE CHIST-ERA EU project, which aimed to develop an explicable AI where low-level features of the deep learning process are combined with higher-level patterns of human argumentation. With the goal of providing an autonomous system for generating high-quality explanations for AI predictions in natural language and applied to the medical field, this project led us to create and deploy together the ANTIDOTE software suite. Thanks to the participation of a large number of multidisciplinary experts, the project successfully covered explanatory argumentation from theory to application. As part of this project, I was directly involved in the development of argument-based XAI technological solutions, which I will presented in this chapter. More precisely, in section 5.1, I first present our argumentation structure extraction tool, initially designed for the EBM context. I will then discuss, in Section 5.2, our contribution to the multitask and multilingual medical LLM MedMT5, before presenting the results of the ANTIDOTE project in Section 5.3. Finally, I will draw some conclusions and future challenges in Section 5.4.

5.1 Argumentation Mining for medical documents

Argument mining has shown many potential applications in helping clinicians to make informed decisions through evidence-based medicine [127]. While AM helps decision-making in healthcare, I have also shown in the previous Chapters that it has a major role to play in explanatory argumentation. As discussed in Chapter 3 Section 3.6, generating argumentative explanations in natural language represents an open challenge, expecially for the medical domain. To meet this challenge, I have collaborated to the development of Version 2.0 and 3.0 of the ACTA tool to automatically extract argumentation structures from natural language text. It is worth noticing that we have used ACTA to assess explanations (Chapter 4).

5.1.1 ACTA

Identifying argumentation structures within natural language appear to be a great support in many domains such as evidence-based medicine [127], wich aims at making decisions about the care of individual patients based on the explicit use of the best available evidence in the patient clinical history and the medical literature results. Argumentation represents a natural way of addressing this task by *(i)* identifying evidence and claims in text, and *(ii)* reasoning upon the extracted arguments and their relations to make a decision. ACTA is a modular architecture introduced by Mayer et al. [128] relying on fine-tuned transformer-based architecture to extract argumentation structures from natural language document. The tool is designed for the medical domain, to ease the work of clinicians in analyzing Randomized Clinical Trials (RCT). It is design for assisting clinicians keeping up to date with the latest discoveries and literature by providing a quick view on the main argumentation of RCTs. Alternatively to keyword-based search in clinical trial abstracts, it empowers the clinician with the ability to retrieve the main claim(s) stated in the trial, as well as the premises (or evidence) linked to this claim. As a result, the clinician does not need to read the whole abstract, but is provided with a structured "summary" of the abstract under the form of a graph. The overall architecture of ACTA is visualized in Figure 5.1.

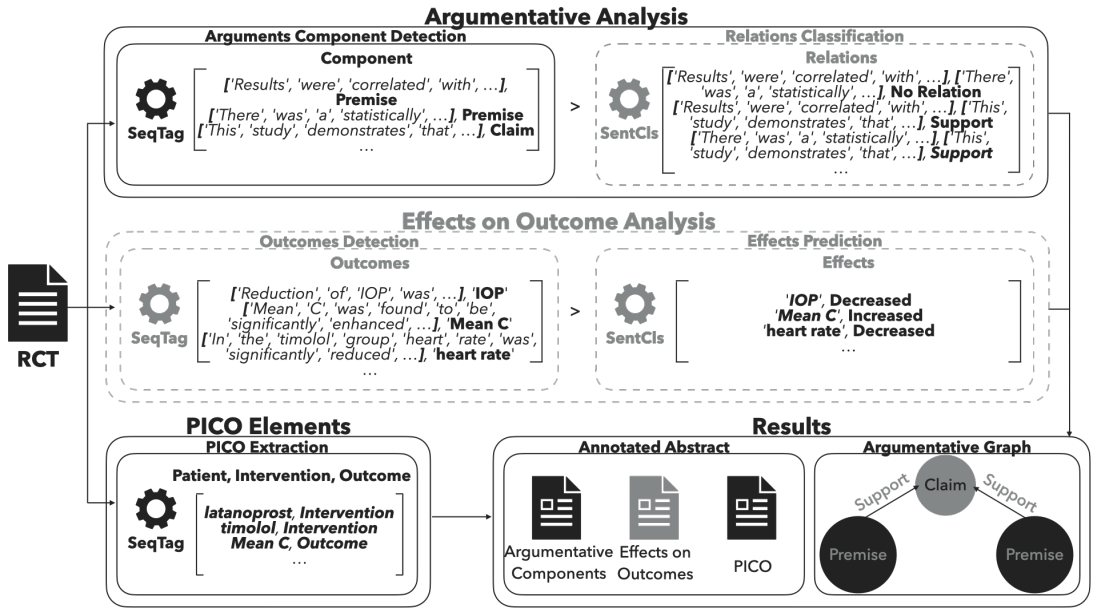

Figure 5.1 – ACTA 3.0 pipeline (the newly introduced modules are in grey). Tasks SeqTag and SentCls means sequence tagging and sentence classification respectively.

ACTA Features. ACTA has been improved since the first version, each time with the addition of new features to make it more useful, particularly in the field of evidence-based medicine. The main use of ACTA lies in its ability to extract argumentation structures (i.e., argumentation components and relations between components) directly from RCTs expressed in natural language. Initially, ACTA detects the simplest components (i.e., claim, premises) and predicts whether a relationship exists between them, without specifying the nature of this relationship. In addition, to improve the decision-making process, ACTA is proposing to detect PICO [1] elements directly in the RCTs. Then, given that the system is designed for medical experts, it allows the use of the PubMed [2] search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Finally, ACTA provides a graph representation of the argumentation structure, with the nodes being the components and the arcs the relationships. This graph is accompanied by the initial RCT text, with the option of highlighting the argument components or PICO elements directly in the text. These analyses can be run on a single RCT, on a selection of documents directly from the PubMed search or by copying and pasting the raw text into the platform. The Figure 5.2 shows some visualisations of the latest version of ACTA, sharing the same features.

5.1.2 Towards ACTA 3.0

We have improved ACTA by adding new features, including performance by reworking the architecture and adding more customizable parameters to the pipeline, features for

1. PICO is a framework to answer health-care related questions in evidence-based practice. Elements comprise patients/population (P), intervention (I), control/comparison (C) and outcome (O) information.
2. https://pubmed.ncbi.nlm.nih.gov/

Figure 5.2 – ACTA 3.0 demonstration system. Argument analysis results are shown within the graph and in a textual representation (i.e., in the initial text on top right part) where blue components indicate premises and yellow components claims. Relations between nodes are headed, represented by attack or support labeled arrows. The lower section reveals the PICO elements together with Effect on Outcomes, also highlightable within the top right part of the visualisation.

argument mining and evidence-based medicine such as relations labeling and effect on outcome prediction, distribution and access with the deployment of a public API and the distribution of ACTA as a python package.

Effect on Outcomes. We introduced a new module to analyse the reported effects an intervention has on the outcomes (O of PICO framework) in the clinical trial abstract. As introduced by Mayer et al. [130], the automatic identification of effects on outcomes could be beneficial to enrich the arguments with valuable medical information and to provide structured and machine-processable data, which can serve as input to a computational model of argument system [14]. The effect on outcome analysis is divided into i) outcome detection tackled as a sequence tagging, and ii) the effect prediction casted into a sentence classification problem. We then are able to identify 6 classes of effects on outcomes namely *Increased*, *Decreased*, *Improved*, *NoOccurrences* or *NoDifferences*. In the interface of the tool, these effects can also be highlighted within the inputted RCT text.

Relations labeling. As a main enhancement of ACTA's argumentative analysis, we also implemented a new relation classification methodology, described in [130], to predict the relation labels (i.e., attack, support and no relation) between components. The implemented method is based on a SciBERT sequence classification of pairs of components previously detected, trained over the relations of the AbstRCT dataset. Each combination of pairs is predicted with the labels Attack, Support or No Relations making the possibility to extract and visualize the graph, with only the Attack and Support relations which are visible.

Public distribution. As ACTA is a set of modules, each performing a particular subtask, some modules could be useful for the development of different tools and/or in different domains. To do so, the entire architecture had to be redesigned in order to make it available for the community with the possibility to use the pipeline or a module as a tool for other argument mining tasks. Therefore, in ACTA 2.0 [141], we proposed to separate each component of the full pipeline and make all of them available independently to allow anyone to use only the bricks of interest. To do so, on top of reworking the code architecture [3], the need of a REST API was identified to interface the modules with the web. The API documentation is available online [4] and an overview is available in Figure 5.3. Following the ambition to make argument mining easier to access, the ACTA modules and pipeline were released as an open-source library [5] within the last contribution about ACTA 3.0 [36].

5.1.3 Implementation details and results

We report here on the implementation details of the ACTA modules and the results obtained for each of the AM tasks on the tool.

3. We later decided to follow Black (https://github.com/psf/black) and Flake (https://flake8.pycqa.org/en/latest/) formatting standard.
4. http://antidote.i3s.unice.fr/acta/api/docs/
5. https://gitlab.com/wimmics-antidote/antidote-acta

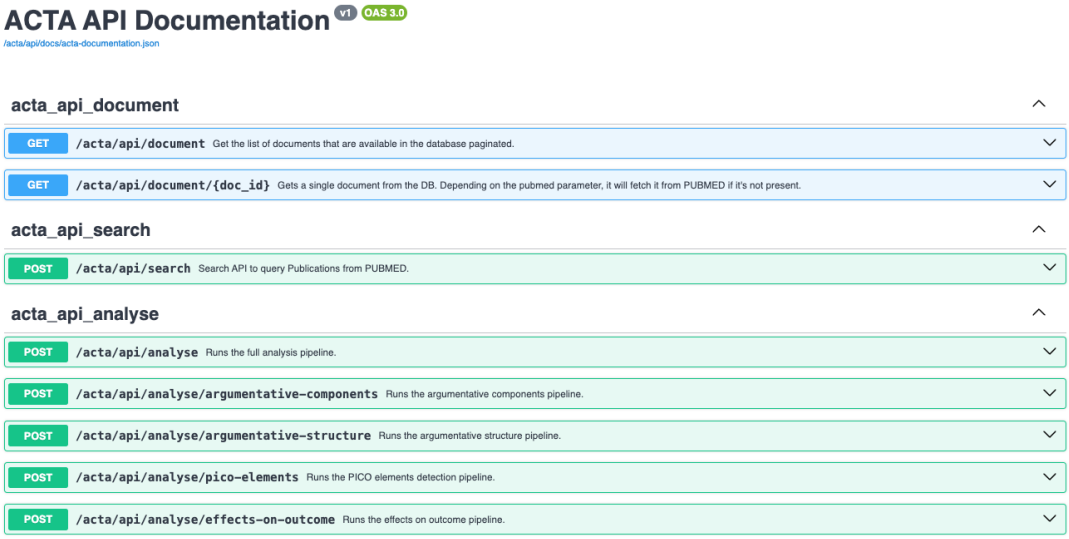

Figure 5.3 – ACTA 3.0 API documentation. External users can investigate each endpoint and try requests from the API page. Each ACTA component is described (i.e., argument components detection, relations classification, PICO element detection, outcomes detection and effect on outcome prediction) with the expected parameters to run the modules.

Argumentative Analysis. For the argumentative analysis, we distinguish two complementary stages : the extraction of argumentation components and the classification of relationships between components. We tackled the argument component detection as a sequence tagging problem based on a pre-trained bi-directional transformer language model. The sentence representation is then passed into a Recurrent Neural Network, here a GRU [43] and then to a CRF [114]). The best performing model after refactoring the architecture with the Huggingface is DEBERTa-v3 [84], achieving a macro f1-score of 0.81, 0.82, and 0.82 for the AbstRCT-Neoplasm, AbstRCT-Glaucoma, and AbstRCT-Mixed tests sets respectively [130]. This model is fine-tune over the AbstRCT dataset [6] (i.e., 500 abstracts of randomized controlled trials on neoplasm treatment annotated with claim, premise and their relations) during three epochs with an Adam optimizer and a learning rate of 2e-5.

Concerning the relation classification step, we also rely on a bi-directional transformer, but we changed the initial representation of the sequence classification task to jointly model the relations by classifying all the argumentation component combinations. This new representation is passed to a linear layer with a softmax which classifies it into the three target classes (*Support*, *Attack* and *NoRelation*). The best performing model for this task is SciBERT [20] uncased base model fine-tuned with a learning rate of 2e-5, batch size of 8, maximum sentence length of 256 sub-words tokens per input example during 3 epoch. The weight factor for each of the 3 classes in the weight cross entropy loss is the normalized number of training samples of this class. This configuration achieves a macro f1-score of 0.68, 0.70 and 0.70 for the AbstRCT-Neoplasm, AbstRCT-Glaucoma, and AbstRCT-Mixed tests sets.

6. https://gitlab.com/tomaye/abstrct/

Evidence-based Medicine features. PICO elements are detected in the same way as argumentation components with a token-level representation inputted into a bidirectional GRU followed by a CRF. As a transformer architecture, we use the BERT base model [58] with pre-trained weights. We fine-tune the entire model with an Adam optimizer and a learning rate of 2e-5 for three epochs above the EBM-NLP corpus [150] with coarse labels. The dataset splits are the same than in [130] without sentences containing less than 10 WordPiece [218]. The obtained f1-score on the test set is 73.4.

More specifically about the outcome analysis (i.e., the reported effects an intervention has on the outcomes), we tackled both outcome detection, and effect prediction with pre-trained bidirectional transformer language models. The outcomes detection is also presented sequence tagging problem relying on BIO-tagging scheme [170] to detect the outcome boundaries. It follows the same configuration than argument components detection with the SciBERT uncased base model fine-tuned with a learning rate of 2e-5, batch size of 8, maximum sentence length of 256 sub-words tokens per input example during 3 epoch. We then address the effect prediction as sentence classification, where each outcome together with the component it occurred in is provided as input into the effect classifier. The same pre-trained transformer model types as for relation classification (based on SciBERT combined to a bidirectional GRU and a final CRF) are used to predict one among the fives effect classes (*Improved*, *Increased*, *Decreased*, *NoDifference*, *NoOccurrence*). The outcome detection and effect classification tasks together reach a macro f1-score of 0.80.

Architecture and API. It is important to note that ACTA was initially developed before international initiatives to make deep learning, specifically NLP tasks more accessible (e.g., Huggingface [7] [215]). Therefore, we refactored and revisited the ACTA pipeline [141], improving and updating technical aspects of the original code to increase its overall stability, documentation, and compatibility, especially with newer models available on Huggingface. Finally, as implemented in the explanatory assessment pipeline in Chapter 4, we discovered that removing the CRF layer of the argument components detection also perform slightly better on some datasets leading us to the development of an optional parameter to remove it.

Concerning the API, we used the Flask framework [8] version 3.0 to create the webserver using Python 3.8. Each module takes as input a JSON file, where for the argument components, the PICO elements and Outcome Detection modules, the field "text" must be filled in with the medical text to be analyzed. For the relation classification module, the input JSON file must have the field "candidates" filled with the list of all of the argumentation components text and type (claim or premise) for which the user wants to predict the relation (support or attack). For the effect prediction module, both the original text and the selected outcomes have to be provided in the "text" and "outcomes" fields respectively. For every module, a JSON file is produced as output with the corresponding results, either being the detected component spans or the predicted labels. All the results, including the argumentative analysis together with PICO elements and effects on outcomes, can be downloaded as a JSON file for each of the processed abstracts.

7. https://huggingface.co/
8. https://flask.palletsprojects.com/en/3.0.x/

5.2 MedMT5

With the advancement of NLP and mainly in generative tasks [145, 29, 154], many language models and large language models are proposed to generate text, but the development of medical applications remains a hot topic. Thus, a number of LLMs have recently been adapted to the medical domain, so that they can be used as a tool for mediating in human-AI interaction. In this section, I present my contribution to the development of the LLM MedMT5, a language model trained on multilingual and multitask medical data. I first introduce MedMT5 in and then focus in detail on my contributions. More specifically, I explain how we retrieved the French training data and how we evaluated the performance of the model.

5.2.1 The MedMT5 LLM.

A number of specialised models have appeared, with on the one hand encoder models such as SciBERT [20], BioBERT [117] or PubmedBERT [77] and text-to-text models on the other, with SciFive [158], BioGPT [123], Med-PaLM [182], PMC-LLaMA [217] or ClinicalGPT [209]. However, the development of all the aforementioned text-to-text LLMs has been focused on a single language, usually English. As a consequence, there is a lack of high-quality multilingual evaluation benchmarks for the medical domain. Thus, although there have been efforts to generate evaluation data in languages other than English [209, 39], they have consisted largely in monolingual approaches.

In order to address these issues, we have compiled, to the best of our knowledge, the largest multilingual corpus for training LLMs adapted to the medical domain. Our corpus includes 3B words in four languages, namely, English, Spanish, French, and Italian. While relatively small when compared to English existing datasets [217], it allowed us to build MedMT5, the first open-source text-to-text multilingual model for the medical domain. Medical mT5 is an encoder-decoder model developed by continuing the training of publicly available mT5 [219] checkpoints on medical domain data for English, Spanish, French, and Italian. Additionally, we have also created two new multilingual sequence labeling (argument component detection) and generative question answering datasets for the evaluation of multilingual LLMs in the medical domain.

As a part of the ANTIDOTE project, MedMT5 open-source models and data are released as (i) the collection of the largest publicly available in-domain medical multilingual corpus for Spanish, French, and Italian languages [9]. It also provides (ii) two new datasets for Spanish, French, and Italian on Argument mining [10] and generative Question Answering tasks, generated taking their original English versions as a starting point [11]. Finally, (iii) the public release of two Medical mT5 versions : a 770M [12] and 3B [13] parameters.

9. https://hf.co/datasets/HiTZ/Multilingual-Medical-Corpus
10. https://hf.co/datasets/HiTZ/multilingual-abstrct
11. https://hf.co/datasets/HiTZ/Multilingual-BioASQ-6B
12. https://hf.co/HiTZ/Medical-mT5-large
13. https://hf.co/HiTZ/Medical-mT5-xl

5.2.2 French data collection.

Other benefits of our Medical mT5 models include the comparatively low hardware requirements needed for both fine-tuning on downstream tasks (the large 770M version easily fits in a 24GB V100 GPU) and for inference (a 12GB GPU should be enough). As an example, a LLaMA 7B model [217] requires at least a 80GB A100 GPU using LoRA [91] or a more demanding 4 80GB A100 GPUs without it. If ANTIDOTE project partners have enough computing power to train large models such as T5 in our case, we still need to collect data. As MedMT5 intend to be multilingual the training data must also be multilingual and collected from verified sources in each language. In the end, we collected around 3 billion tokens (a word can be split into several tokens depending on the model, detailed in Section 2.1). For the French data, we managed to collect a total of 7,192,779 sentences and 670,972,717 words were compiled using the data sources listed in bold in Table 5.1. PubMed data was extracted using the `Bio.Entrez` package [14]. We already relied on PubMed to browse RCTs for ACTA in Section 5.1 and decided to retrieve the french literature as it comprises more than 37 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites. [15] Other collaborators also retrieved from PubMed in English, Spanish and Italian, using similar methodologies. Science Direct offers a collection of scientific and medical publications which can be extracted via their official API [16]. We filtered relevant articles with the keyword "Médecine", and the obtained XML documents were parsed to extract the `<dc:description>` tag. As for Spanish data, we took advantage of Wikipedia as a source of medical knowledge to obtain HTML formatted data from the category "Category :Médecine". The EDP French/English Parallel Medical Corpus [95] provides bilingual content from journals that address domains such as dentistry and life sciences. From this source, we downloaded the dataset labeled "EDP French corpus, text format". Finally, Google Patents is a comprehensive repository of patent data from around the world. Google Patents data were retrieved by filtering using the IPC code and abstract language. The data request was queried on Google Big Query with the following request :

```
SELECT publication_number, abstract.text
FROM `patents-public-data.patents.publications`,
 UNNEST(abstract_localized) as abstract,
 UNNEST(ipc) as ipc
WHERE abstract.language = "fr"
 AND (ipc.code LIKE 'A61B%'
   OR ipc.code LIKE 'A61C%'
   OR ipc.code LIKE 'A61F%'
   OR ipc.code LIKE 'A61H%'
   OR ipc.code LIKE 'A61K%'
   OR ipc.code LIKE 'A61L%'
   OR ipc.code LIKE 'A61M%'
   OR ipc.code LIKE 'A61P%')
```

14. https://biopython.org/docs/1.75/api/Bio.Entrez.html
15. Accessed in September 2024.
16. https://dev.elsevier.com/

Every collected data was cleaned and converted into JSONL format. These transformations involved the use of regex patterns to eliminate HTML tags and rectify encoding anomalies. A final French language verification step was undertaken by applying the `langdetect` package (version 1.0.9). Concerning the other languages, we collected 1,107,800,000 words for English, 956,050,000 for Spanish and 143,278,000 for Italian as detailed by sources in Table 5.1.

Source	English	Spanish	French	Italian
ClinicalTrials	127.4M	-	-	-
EMEA	12M	13.6M	-	-
PubMed	968.4M	8.4M	**1.4M**	2.3M
Medical Crawler	-	918M	-	67M
SPACC	-	350K	-	-
UFAL	-	10.5M	-	-
WikiMed	-	5.2M	-	-
Wikipedia	-	-	**5M**	13.3M
Science Direct	-	-	**15.2M**	-
EDP	-	-	**48K**	-
Google Patents	-	-	**654M**	-
Drug instructions	-	-	-	30.5M
E3C Corpus - IT	-	-	-	11.6M
Medicine descriptions	-	-	-	6.3M
Medical theses	-	-	-	5.8M
Medical websites	-	-	-	4M
Supplement description	-	-	-	1.3M
Medical notes	-	-	-	975K
Pathologies	-	-	-	157K
Medical test simulations	-	-	-	26K
Clinical cases	-	-	-	20K

TABLE 5.1 – Combined data sources and word counts for English, Spanish, French, and Italian.

5.2.3 Evaluation of the French capacities.

Since the MedMT5 contribution also provides a multilingual and multitasking benchmark, we evaluated the model on the Sequence Labelling and Abstractive Question Answering tasks. The two challenges of this model in the context of explanatory argumentation are to see whether it can generate better explanations in natural language and whether it allows better extraction of argumentation structures.

Abstractive Question Answering. In order to be able to use MedMT5 for generating medical explanations (e.g., potentially replacing our Chapter 3 templates-based approach), we explored the text generation capabilities of Medical mT5 on the BioASQ question answering dataset. We use the BioASQ-6B English Question Answering dataset [199] to generate parallel French, Italian and Spanish versions. Given a biomedical question and a set of snippets of text with relevant information about the question, the model must

generate the *ideal* answer. A set of ideal gold answers are provided to assess the performance of the models. We machine translated the questions and ideal answers into French, Italian and Spanish using the NLLB200 3B parameter model [52]. Previous work typically evaluates the performance on this task using the ROUGE score [199] to compare the gold standard answer with the answer generated by the model. However, we find this metric not appropriate for medical domain tasks as it does not address crucial aspects of the generation such as factuality, potential harm, and bias [182]. Consequently, we involved medical professionals to analyze the answers produced by the models.

During annotation, medical doctors were displayed the question, the ideal gold answers and the answers generated by each model. If required, they could also inspect the snippets that provide context to answer each of the questions. We narrowed the evaluation to Medical-mT5-large, mT5-large, FlanT5-large and SciFive. The evaluation was conducted by medical doctors proficient/native speakers of English, French and Spanish. For each question, doctors were asked to rank the answers generated by the models as the best, second-best, third-best, and worst answer.

For the French language, three French clinicians analyzed 186 answers, of which 47 were done by 2 doctors to calculate IAA (Cohen's Kappa Score : 0.28 and Average Spearman's Rank Correlation : 0.48), which indicates a low level of agreement. This exercise provided interesting insights with respect to the performance of the models in text generation tasks in the medical domain. First, medical doctors could not in general establish significant differences between the answers generated by each of the models ; predictions were far too similar, and all tended to fail on the same questions. Two Spanish medical doctors proficient or natives in English and Spanish also analyzed 50 English examples and 252 Spanish. As an example, Table 5.2 shows the answers to two different questions. As it can be observed, the answers generated by each model are very similar, and the doctors ended up ranking them primarily based on style.

Sequence Labeling. The MedMT5 model is based on the T5 architecture and allows tasks other than the generation presented above. In the context of the ANTIDOTE project and to evaluate the applications of this model to the extraction of argumentation structures, we have evaluated the model on the Sequence Labelling task. We therefore evaluated the performance of MedMT5 in 3 different configurations : Single Task Monolingual Supervised Results, Multi-Task Supervised Results and Zero-shot Cross-Lingual Transfer Results. This evaluation compares our model with the same alternatives as for the generation, namely SviFive and FlanT5. The implementation details are detailed in the contribution by Garcia et al. [72].

The result of this Sequence Labeling evaluation shows that Medical mT5 stands out for its superior performance in the evaluation for Spanish, French, and Italian languages, especially for the multitask and the zero-shot transfer settings. FlanT5 and SciFive, are systematically superior on English data which can be explained since both have been pre-trained with a much larger amount of English-only data specific to the medical domain. However, these capabilities can help mitigate the scarcity of manually annotated medical data for other target languages. Despite Medical-mT5-xl being larger than Medical-mT5-large (3B vs 738M), its performance is worse in the single-task evaluation setting. This behaviour is not observed in the multi-task and zero-shot experiments, leading us to hy-

Ideal Answer	Which type of urinary incontinence is diagnosed with the Q tip test ? Stress urinary incontinence is diagnosed with the Q tip test. The test evaluates urethral mobility.
mT5-large	Q-tip test is used for urodynamic assessment of patients with severe urinary incontinence.
FlanT5-large	The Q tip test is used for assessing urethral mobility.
SciFive	Q tip test is used for evaluation of stress urinary incontinence.
MedMT5-large	Q-tip test is used for evaluation of stress urinary incontinence.
Ideal Answer	Which are the main manifestations of Ohdo syndrome ? Severe ID, absent or deficient language, skeletal manifestations including bilateral patella dislocations.
mT5-large	Skeletal manifestations in Ohdo syndrome are a case with bilateral patella dislocations where surgical intervention has been indicated.
FlanT5-large	The main manifestations of Ohdo syndrome are : 1) severe ID, 2) absent or deficient language and 3) milder, clinical manifestation in heterozygotes.
SciFive	Ohdo syndrome is characterized by severe ID, absent or deficient language and, milder, clinical manifestation in heterozygotes.
MedMT5-large	The main manifestations of Ohdo syndrome are : 1) absent or deficient language and 2) mildder clinical manifestation in heterozygotes.

TABLE 5.2 – Examples of answers generated by each model for two different BioASQ questions.

pothesize that the larger Medical-mT5-xl model is more prone to overfit in the single-task supervised setting.

Argument Mining. While annotated medical data is already rare, manually annotated medical data in languages other than English are even rarer. As we discussed in Chapter 4, argumentative annotations are a good example because they are hard to produce, especially in languages other than English. It is therefore essential to develop models capable of generating predictions in languages other than those used for fine-tuning. We evaluate this ability to perform cross-linguistic transfer from scratch by refining MedMT5 and the reference models on the AbsRCT [128] Neoplasm dataset in English, and then evaluating them on the Neoplasm, Glaucoma and Mixed datasets for Spanish, French and Italian. The results are presented in table 5.3. The results show that Medical mT5 outperforms all other models. In addition, Medical-mT5-xl performed significantly better than Medical-mT5-large. The state-of-the-art result for this task in English is reported in my Chapter 4 for the Argument Component Detection task, reaching 0.83 by finetuning the SciBERT model.

5.2.4 Discussion

In this contribution, we presented MedMT5, the first open source multilingual text-to-text LLM for the medical domain. Its development required the compilation of a new

Lang	Dataset	$mT5_{XL}$	SciFive	$FlanT5_{XL}$	$mDeBERTa_{V3\ base}$	$MedMT5_{large}$	$MedMT5_{XL}$
ES	Neoplasm	71.4	69.8	67.9	65.1	**72.4**	71.7
ES	Glaucoma	**74.1**	71.5	70.6	68.3	72.4	73.2
ES	Mixed	**69.4**	67.0	66.7	60.9	68.1	68.8
FR	Neoplasm	71.6	68.6	69.9	60.5	72.4	**72.8**
FR	Glaucoma	75.8	74.5	71.0	68.7	72.3	**76.7**
FR	Mixed	**73.0**	68.5	68.2	59.3	70.4	72.4
IT	Neoplasm	70.6	63.1	67.3	62.4	72.9	**73.2**
IT	Glaucoma	76.7	71.6	72.0	70.2	75.4	**79.0**
IT	Mixed	69.9	62.5	66.9	62.1	71.7	**71.9**
AVERAGE		72.5	68.6	69.0	64.2	72.0	**73.3**

TABLE 5.3 – Zero-shot F1 scores for Argument Mining. Models have been trained in English and evaluated in Spanish, French and Italian.

corpus of 3B words in English, French, Italian and Spanish specific to the medical domain. In addition, motivated by the lack of multilingual references, we generated evaluation references for French, Italian and Spanish for argument extraction and abstract question answering.

Regarding the languages we selected, we would like to point out that data acquisition in the medical field is extremely difficult. Furthermore, the choice of languages was also influenced by the availability of native language doctors to carry out the manual evaluation of the response to abstract questions.

Extensive experimentation on sequence labelling tasks shows that Medical mT5 outperforms reference text-to-text models of similar size in both multitasking and zero-shot multilingual evaluation contexts. This is particularly interesting as these parameters fully exploit the multilingual nature of a text-to-text model such as Medical mT5.

Furthermore, our experiments on abstract question answering show the inherent difficulty of generative task evaluation for this specific domain, where complex issues such as veracity and truthfulness are difficult to capture by automatic measures. Manual evaluation is also not ideal, as the doctors were unable to clearly distinguish the quality of the responses generated by the different models.

5.3 The ANTIDOTE software suite

The goal of the ANTIDOTE project was of providing a unified computational framework for jointly learning clinical predictions and the associated argumentative justifications, fostering a natural interaction with clinicians through explanatory dialogues.

In this contribution, we present the ANTIDOTE demonstration [36], a software suite proposing different tools for argumentation-driven explainable Artificial Intelligence for digital medicine. Our system offers the following functionalities. First, we tackled argumentative analysis for the medical domain from different perspectives such as training specialized language models (i.e., the ACTA tool [128, 141], presented in Section 5.1), proposing multilingual [17] language models based on data-transfer [221] methods and with Multi-scale Convolution Neural Network [190] for Argument Structure Learning (ASL).

17. https://huggingface.co/datasets/HiTZ/multilingual-abstrct

ANTIDOTE also tackles explanation generation of clinical diagnoses relying on medical knowledge across the SYMEXP pipeline [140], details in Chapter 3. Finally, a multilingual large language model for the medical domain and the first multilingual benchmark for medical question-answering is provided within MedMT5 (i.e., described in Section 5.2 and [72]) and the MedExpQA dataset [6]. Experimental results demonstrate the efficacy of ANTIDOTE across different tasks, highlighting its potential as an asset in medical research and practice and fostering transparency, which is crucial for informed decision-making in healthcare.

Implementation details. ANTIDOTE is deployed in the form of a website [18], bringing together the tools available on the web or referring to links hosting datasets and models. Some of the models are hosted and distributed on Huggingface [19], along with their respective training datasets [20]. The rest of the datasets are available on versioning platforms such as gitlab and github on each of the project pages in question. The suite also includes SYMEXP [21] and ACTA [22], whose implementation details are given in Chapter 3. To implement ANTIDOTE, we used basic web technologies such as HTML, CSS, Javascript and the Bootstrap version 5.3.3 [23] framework for the front-end. For the back-end, API, we used the Flask framework [24] version 3.0 to create the webserver using Python 3.8.

5.4 Challenges

In this Chapter, I presented the tools I have developed through my research and explain how they have been used in the context of the ANTIDOTE research project. However, despite this set of tools, the automatic generation of natural language argumentative explanations remains an open challenge. While argument mining is effective for structuring explanations, it must be combined with generative models to produce coherent explanations in natural language. The limitations of existing tools, notably large language models, lie in their inability to retrieve knowledge from their internal state. Combining this kind of solution with medical knowledge bases, vocabularies or databases therefore seems to be the most interesting research line to explore in order to develop more efficient and secure tools, taking the advantage of LLM to generate human-like sentences. Finally, the concept of explanation is mostly introduced in a dialogue context, and chatbot-type assistants could be the kind of tools needed to enhance the impact of XAI in the field of education.

18. http://antidote.i3s.unice.fr/
19. https://huggingface.co/HiTZ
20. https://huggingface.co/datasets/HiTZ
21. http://antidote.i3s.unice.fr/symexp/
22. http://antidote.i3s.unice.fr/acta/
23. Accessed in February 2024
24. https://flask.palletsprojects.com/en/3.0.x/

Chapter 6

Conclusion et Perspectives

Generating explanations in an argumentative way is essential to enhance understanding from human users on the object of the explanation, but it is also a hard task in all application domains, particularly in sensitive areas such as medicine. This thesis tackles the generation of qualitative argumentative explanations in natural language applied to the medical domain.

The research questions I answered focus on the advancements in automatic generation of natural language explanations, on grounding these explanations on reliable medical sources and on enabling assessment of argumentation strength of medical explanations. In particular, the research questions introduced in Chapter 1 were addressed, resulting in the following contributions :

1. Generating grounded explanations. To enable argumentation-based generation of medical explanations, this thesis introduces in Chapter 3 the SYMEXP pipeline for extracting and aligning knowledge with trustable sources to generate natural language explanations. Starting from a clinical case, I automatically detect named entities such as layperson symptoms and health measurement using natural language processing methods. Particular attention is paid to medical measurements, test results and observations (so-called medical findings) in order to detect whether these findings are in normal range or not. Therefore, a set of 100 most common findings, their normal values and their conversion into medical terms is introduced, exploiting the knowledge of a medical expert. If a finding is identified as abnormal, I retrieve the medical term equivalent to that abnormal value (e.g., temperature is 30°C $\rightarrow$ fever). These converted findings and detected symptoms are then mapped to standardized medical ontologies such as the HPO to align them with validated knowledge. By linking case details to external sources, additional relevant information is obtained, such as symptoms that frequently coincide with certain diagnoses. Finally, I propose a first solution to generate argumentative explanations based on templates, designed with the help of medical experts, to explain symptomatically why the correct diagnosis is correct and why other options are not, while highlighting the important elements missing from the clinical case.

To automatically detect the different medical entities described in clinical cases, I have experimented with different transformer-based language models, such as SciBERT, BioBERT, PubMedBERT and UmlsBERT, initialized with their respective pre-trained weights

specialized in the biomedical domain. I considered the symptom detection problem as a sequence labeling task, using a BIO scheme tag. The experiments show that the SciBERT model performs best, with an f1 macro-score of 0.86 for symptoms named entity recognition. To convert findings correctly, I use regexes to find them and interpret their values according to my database. If it is not possible to idenfiy in the the database the corresponding information for a finding, I proposed a pipeline based on the GPT LLM to interpret such finding on the fly. Conversion with the database performs in the best configuration, with an accuracy of 0.78 on our clinical cases and 0.66 using the generative model. To accurately match detected symptoms and converted findings to the HPO medical ontology, I computed embeddings for each symptom and calculate the cosine distance with each HPO term to find the closest match. Our context-aware integration approach, which sums the embeddings for symptoms and context sentence, performs significantly better than a basic context-free method such as the existing tool DASH, improving accuracy from 0.37 to 0.70 on our data. By taking contextual information into account, I achieved a more reliable alignment between layperson symptom descriptions and official HPO terminology. Overall, specialized neural models and context-aware integration technique effectively extract and align salient clinical entities, providing a solid basis for assessing the relevance of symptoms to potential diagnoses and generating explanatory arguments based on the validated medical knowledge.

2. Explanatory argumentation assessment. To be able to investigate explanations suitable for an explanee, I have developed in Chapter 4, an automatic system named ABEXA to assess argumentation and characterize medical explanations. For a given clinical case, I extract the argumentation structure of the case and its explanation to detect argumentation patterns prepared beforehand. More specifically, I use argument mining techniques to adapt the ACTA tool, originally designed for RCTs, for clinical cases. This tool (detailed in Chapter 5) allows me to recover argument components such as claims and premises, as well as the attack and support relations between such components. I then used a set of argumentation patterns parameterizable according to the level of granularity on this argumentation graph to retrieve information about the type of argumentation which is employed. The pipeline explores the structure of the graph to detect the use of argumentation patterns such as the use of answers, question or facts presented in the question or the size of the arguments.

To automatically detect argumentation structures, I adapted the ACTA's architecture and experiment with different transformer-based language models for the argument extraction and relation prediction tasks, training the models on clinical cases rather than RCTs (i.e., using the Casimedicos dataset instead of AbstRCT). I evaluated the performance of each task individually, as well as the entire pipeline, using two different approaches to address the issue of misdetected components. The end-to-end argument mining task results in a 0.65 macro f1 scores on the AbstRCT dataset using the SciBERT model, and 0.51 on Casimedicos using the PubMedBERT model. For component detection and relation prediction tasks, I obtained respectively 0.83 and 0.73 on AbstRCT and 0.86 and 0.51 on Casimedicos using the same models. I also highlighted that coreference disambiguation does not ameliorate the results. On the contrary, with coreference disambiguation the end-to-end results drop from 0.51 to 0.49. In detecting patterns of argumentation in the explanation, I found some interesting points about medical explanations. First of all, the

most frequent argumentation patterns aim to justify the correct answer rather than to discriminate between all the incorrect answers in 95% of cases. Then, the most frequent and strongly correlated patterns are the lack of use of an argumentation component from the question and the lack of use of premises (mostly introduced in the question), representing 90% and 78% of Casimedicos clinical cases, respectively. Finally, when evaluating the proposed end-to-end pipeline, fewer argumentative components and relations in clinical cases are detected, but the results on patterns show similar results to the gold ones in terms of proportion.

3. Explanatory argumentation applications. Finally, I have participated in the implementation and dissemination of various tools for explanatory argumentation, as presented in Chapter 5. In particular, I have introduced two pipelines for generating (SYMEXP) and assessing (ABEX) explanations in the medical domain, as well as more fundamental tools such as ACTA for the argumentation structure extraction, and MedMT5 which is a new multilingual and multitasking LLM. More specifically, I have contributed to the improvement of ACTA by developing a model based on transformers to predict effect-on-outcome in RCT, performing 0.80 macro f1. The analysis is divided into two parts : the detection of outcomes (considered as a sequence labeling task based on the BIO scheme tag), and the classification of the entities through 5 classes : Increased, Decreased, Improved, NoOccurrences and NoDifferences. The classification of relations has also been updated by implementing a system for classifying relationships between components, so that in addition to detecting a relation, it is possible to determine whether it is a support or an attack relation. In addition, I extended the tool by developing a REST API with an endpoint for each of the tasks performed by ACTA, making the pipeline reusable in other underlying pipelines. I also took part in a major overhaul of ACTA's code, adapting it from older versions of transformers to new technologies based on the Huggingface initiative, offering new options to facilitate the development of underlying pipelines. These technology updates have enabled us to update ATCA's results on argument mining tasks for RCT, showing a macro f1 performance of 0.81, 0.82, 0.82 for argument component detection and 0.68, 0.70, 0.70 for relations classifications for the respective AbstRCT-Neoplasm, AbstRCT-Glaucoma, and AbstRCT-Mixed tests sets. I also took part in the development of MedMT5 by collecting and evaluating French aspect of the multilingual model on generation and sequence labelling tasks. The model showed modest performance on the complex generation task, but the result ameliorates when applied to other languages on the sequence modelling task to detect argumentation components. Finally, I took part in the development of the ANTIDOTE software suite, by implementing central platform grouping together all the tools proposed during the project.

6.1 Perspectives

While this thesis tackled significant aspects of explanatory argumentation in natural language, it also opens to several lines for future research and potential improvements.

Leveraging Generative and Dialogue Models. Given that this work explores the generation of explanations, the investigation of LLMs is a promising next step, particularly if they

are supported by argumentation theory. Indeed, although I have tackled the task of generating argumentative explanations, my approach is based on natural language templates. This solutions might not provide enough flexibility, which is required in other application scenarios. To overcome this limitation, a promising direction consists in the integration of generative models or dialogue models. This would enable the explanee to interact with the system, asking follow-up questions or seeking clarification, thereby facilitating a more dynamic and personalized form of explanatory argumentation.

As we discussed earlier, it is crucial to adapt explanations to the explanee, and the explanation will be very different if we are explaining a diagnosis to a doctor, a student or a patient. This gap in explanee's knowledge is an important factor that should be identified in order to generate an explanation that is adapted to the interlocutor, enabling greater comprehension of the explanation.

Expanding Knowledge. Focusing on external knowledge, this thesis demonstrated how ontologies such as the HPO can align expert knowledge with the knowledge found in medical documents. A promising future direction consists in expanding the use of external knowledge sources, for instance by exploring vocabularies shared between ontologies and concept codes. This approach would not only allow us to rely on the content of the knowledge bases but also on its structure. These structures would allow to grasp greater granularity allowing to understand the hierarchy of concepts, for instance, being able to find that "tachiccardia" is a form of "arrhythmia" and consider the correct alignment.

Additionally, I introduced a database to translate medical findings into symptoms or medical terms, but a significant limitation still holds : the applicability of the findings across diverse populations. The medical field evolves differently across geographic regions, and standard values are often derived from data representative of a typical population based on the entity that collected it. Therefore, normal values do not reflect all populations and will be a way less accurate for under-represented populations (i.e., different ethnic and age groups). Addressing this issue will require the collection of more data in collaboration with medical centers that cater to various population groups, ensuring that explanations and medical findings can be adapted to a wider range of individuals.

Evaluating the assessment. Looking at the assessment of explanations, I proposed a set of patterns to automatically characterize natural language explanations from the point of view of the argumentation employed. These criteria and default values were designed specifically for the medical field, particularly for Casimedicos clinical cases. However, it would be important to extend this analysis to broader datasets in medicine, and other domains, such as law, where the standards for a satisfactory explanation may differ significantly from those in medicine.

Additionally, since these criteria were developed to emphasize key argumentative points in student explanations, it would be beneficial to assess their actual impact on students through a user study. This could help to quantify how these criteria influence student learning and whether they lead to improved explanations writing in future exams. Such a study would also open up new research directions, such as generating better explanations based on the identified critical patterns. By integrating this pattern detection system, through LLMs or other generative models discussed before, we could potentially generate higher-quality natural language explanations.

Enhancing Tools for Explanatory Argumentation. Finally, I introduced a suite of tools and complete pipelines for explanatory argumentation in natural language. While these tools mark significant progress, there are still areas that require further exploration, especially to apply these tools in fields other than medicine. Although initial efforts have been made in this direction, the creation of domain-agnostic tools would be a major step towards a greater adoption of explanatory argumentation.

Finally, the need for a unified system is becoming increasingly urgent, as done in the context of the ANTIDOTE project. Future research should aim to integrate these tools into a unified system capable of predicting diagnoses, reasoning through complex cases and generating explanations adapted to the user's knowledge. The development of interactions between the explainer, the explanee and the explanation itself would be essential to achieve the ultimate goal of understanding and responding to the explanee's needs.

Bibliography

[1] Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. QUINT : Interpretable question answering over knowledge bases. In Lucia Specia, Matt Post, and Michael Paul, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 61–66, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[3] Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, Iker Garcia-Ferrero, Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, et al. Hitz@ antidote : Argumentation-driven explainable artificial intelligence for digital medicine. *arXiv preprint arXiv :2306.06029*, 2023.

[4] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[5] Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5) :922–930, 2013.

[6] Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. MedExpQA : Multilingual Benchmarking of Large Language Models for Medical Question Answering. *arXiv 2404.05590*, 2024.

[7] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv :1904.03323*, 2019.

[8] Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345. ACL, July 2020.

[9] Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, 2020.

[10] Julia Amann, Dennis Vetter, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Sara Gerke, Thomas K Gilbert, Thilo Hagendorff, Sune Holm, Michelle Livne, et al. To explain or not to explain ?—artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, 1(2) :e0000016, 2022.

[11] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[12] Abdallah Arioua, Nouredine Tamani, and Madalina Croitoru. Query answering explanation in inconsistent datalog knowledge bases. In *International Conference on Data Management in Cloud, Grid and P2P Systems*, pages 203–219. Springer, 2015.

[13] Sai Athaluri, Varma Manthena, Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Siri Duddumpudi. Exploring the boundaries of reality : Investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15, 04 2023.

[14] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Mag.*, 38(3) :25–36, 2017.

[15] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1) :1–20, 2012.

[16] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014.

[17] Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation ? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[18] Maria Becker, Katharina Korfhage, and Anette Frank. Implicit knowledge in argumentative texts : An annotated corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2316–2324, Marseille, France, May 2020. European Language Resources Association.

[19] Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. Enriching argumentative texts with implicit knowledge. In *Natural Language Processing and Information Systems : 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings 22*, pages 84–96. Springer, 2017.

[20] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*, 2019.

[21] Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy, August 2019. Association for Computational Linguistics.

[22] Trevor JM Bench-Capon and Paul E Dunne. Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15) :619–641, 2007.

[23] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[24] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2) :157–166, 1994.

[25] Olivier Bodenreider. The unified medical language system (UMLS) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1) :D267–D270, 2004.

[26] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5 :135–146, 2017.

[27] Andrei Bondarenko, Phan Minh Dung, Robert A Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial intelligence*, 93(1-2) :63–101, 1997.

[28] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. pages 632–642, 2015. Publisher Copyright : © 2015 Association for Computational Linguistics. ; Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 ; Conference date : 17-09-2015 Through 21-09-2015.

[29] Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. Learning to rationalize for nonmonotonic reasoning with distant supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14) :12592–12601, May 2021.

[30] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33 :1877–1901, 2020.

[31] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

[32] Elena Cabrio and Serena Villata. Five years of argument mining : A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433, 2018.

[33] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli : natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9560–9572, Red Hook, NY, USA, 2018. Curran Associates Inc.

[34] Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1) :1–19, 2021.

[35] Daniel G Campos. On the distinction between peirce's abduction and lipton's inference to the best explanation. *Synthese*, 180(3) :419–442, 2011.

[36] Cristian Cardellino, Theo Collias, Benjamin Molinet, Erwan Hain, Wei Sun, Rodrigo Agerri, Serena Villata, and Elena Cabrio. ANTIDOTE : ArgumeNtaTIon-Driven

explainable artificial intelligence fOr digiTal mEdicine. In *ECAI-24-Demos Proceedings - 27th European Conference on Artificial Intelligence*, Santiago de Compostela, Spain, October 2024.

[37] Giuseppe Carenini, Vibhu O Mittal, and Johanna D Moore. Generating patient-specific interactive natural language explanations. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 5, 6218 Georgia Avenue NW, Suite 1 PMB 3077 Washington, DC 20011, 1994. American Medical Informatics Association.

[38] Rudolf Carnap. The two concepts of probability : The problem of probability. *Philosophy and phenomenological research*, 5(4) :513–532, 1945.

[39] Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[40] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare : Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[41] J. Harry Caufield. MACCROBAT. 1 2020.

[42] Alison Cawsey. *Explanation and interaction : the computer generation of explanatory dialogues*. MIT press, 1992.

[43] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*, 2014.

[44] Noam Chomsky. On nature and language, 2002.

[45] Chih-Yueh Chou, Yu-Ting Chou, and Yu-Cheng Huang. Automated explanation quality assessment based on similarity calculations of representative explanations of different qualities. In *2024 IEEE 7th Eurasian Conference on Educational Innovation (ECEI)*, pages 97–100. IEEE, 2024.

[46] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv :1412.3555*, 2014.

[47] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (TCIA) : maintaining and operating a public information repository. *Journal of digital imaging*, 26 :1045–1057, 2013.

[48] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra : Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv :2003.10555*, 2020.

[49] Miruna Clinciu, Arash Eshghi, and Helen Hastie. A study of automatic metrics for the evaluation of natural language explanations. *arXiv preprint arXiv :2103.08545*, 2021.

[50] Oana Cocarascu, Antonio Rago, and Francesca Toni. Explanation via machine arguing. In *Reasoning Web International Summer School*, pages 53–84. Springer, 2020.

[51] Oana Cocarascu, Andria Stylianou, Kristijonas Čyras, and Francesca Toni. Data-empowered argumentation for dialectically explainable predictions. In *ECAI 2020*, pages 2449–2456. IOS Press, 2020.

[52] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind : Scaling human-centered machine translation. *CoRR*, abs/2207.04672, 2022.

[53] Kristijonas Čyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, 127 :141–156, 2019.

[54] Kristijonas Čyras, Dimitrios Letsios, Ruth Misener, and Francesca Toni. Argumentation for explainable scheduling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2752–2759, 2019.

[55] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI : A survey. *ArXiv*, abs/2105.11266, 2021.

[56] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9) :1342–1350, 2018.

[57] Pritam Deka, Anna Jurek-Loughrey, and P Deepak. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4) :474–504, 2022.

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[59] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus : a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47 :1–10, 2014.

[60] Kevin Donnelly et al. SNOMED CT : The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121 :279, 2006.

[61] Stefan Dragulinescu. Inference to the best explanation and mechanisms in medicine. *Theoretical medicine and bioethics*, 37 :211–232, 2016.

[62] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2) :321–357, 1995.

[63] Phan Minh Dung, Paolo Mancarella, and Francesca Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 171(10-15) :642–674, 2007.

[64] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv :1704.06104*, 2017.

[65] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2) :179–211, 1990.

[66] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639) :115–118, 2017.

[67] Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. Launching into clinical space with medspacy : a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438, 6218 Georgia Avenue NW, Suite 1 PMB 3077 Washington, DC 20011, 2021. American Medical Informatics Association.

[68] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. Hierarchical graph network for multi-hop question answering. *arXiv preprint arXiv :1911.03631*, 2019.

[69] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5) :378, 1971.

[70] J.B. Freeman. *Argument Structure : : Representation and Theory*. Argumentation Library. Springer Netherlands, 2011.

[71] Alejandro J García, Carlos I Chesñevar, Nicolás D Rotstein, and Guillermo R Simari. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Systems with Applications*, 40(8) :3233–3247, 2013.

[72] Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, et al. Medmt5 : An open-source multilingual text-to-text llm for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, 2024.

[73] Denis Gavrilov, Alexander Gusev, Igor Korsakov, Roman Novitsky, and Larisa Serova. Feature extraction method from electronic health records in russia. In *Conference of Open Innovations Association, FRUCT*, number 26, pages 497–500, Helsinki, Finland, e-ISSN 2343-0737, 2020. FRUCT Oy, FRUCT Oy.

[74] N Genes, D Chandra, S Ellis, and K Baumlin. Validating emergency department vital signs using a data quality engine for data warehouse. *The open medical informatics journal*, 7 :34, 2013.

[75] Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus : a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1) :1–17, 2010.

[76] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[77] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1) :1–23, 2021.

[78] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5) :1–42, 2018.

[79] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22) :2402–2410, 2016.

[80] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37) :eaay7120, 2019.

[81] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining : Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics.

[82] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task : Identification and reconstruction of implicit warrants. *arXiv preprint arXiv :1708.01425*, 2017.

[83] ZS Harris. Distributional structure, 1954.

[84] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3 : Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv 2111.09543*, 2021.

[85] Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2) :135–175, 1948.

[86] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1) :3–12, 2020.

[87] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1) :65, 1990.

[88] Joshua A Hirsch, Thabele M Leslie-Mazwi, Gregory N Nicola, Robert M Barr, Jacqueline A Bello, William D Donovan, Raymond Tu, Mark D Alson, and Laxmaiah Manchikanti. Current procedural terminology ; a primer. *Journal of neurointerventional surgery*, 7(4) :309–312, 2015.

[89] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[90] M Honnibal and I Montani. spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. neural machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 688–697, 2017.

[91] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA : Low-rank adaptation of large language models. *arXiv preprint*, 2106.09685, 2021.

[92] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert : Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv :1904.05342*, 2019.

[93] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing.* Chapman and Hall/CRC, 2010.

[94] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv :1902.10186*, 2019.

[95] Antonio Jimeno-Yepes, Aurélie Névéol, Mariana L. Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. Findings of the WMT 2017 biomedical translation shared task. In Ondrej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 234–247. Association for Computational Linguistics, 2017.

[96] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14) :6421, 2021.

[97] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1) :1–9, 2016.

[98] Ralph H Johnson. *Manifest rationality : A pragmatic theory of argument.* Lawrence Earlbaum Associates, New York, 2012.

[99] Karen Sparck Jones. Natural language processing : a historical review. *Current issues in computational linguistics : in honour of Don Walker*, pages 3–16, 1994.

[100] John R Josephson and Susan G Josephson. *Abductive inference : Computation, philosophy, technology.* Cambridge University Press, 1996.

[101] Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. Semeval-2023 task 7 : Multi-evidence natural language inference for clinical trial data. *arXiv preprint arXiv :2305.02993*, 2023.

[102] Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. BioELECTRA :pretrained biomedical text encoder using discriminators. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online, June 2021. Association for Computational Linguistics.

[103] Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. Argument mining as a text-to-text generation task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2002–2014, 2024.

[104] Vivek Khetan, Somin Wadhwa, Byron Wallace, and Silvio Amir. SemEval-2023 task 8 : Causal medical claim identification and related PIO frame extraction from social media posts. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2266–2274, Toronto, Canada, July 2023. Association for Computational Linguistics.

[105] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution : Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[106] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 workshop companion volume for shared task*, pages 1–9, Portland, Oregon, United States, 2009. Omnipress.

[107] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, page 70–75, USA, 2004. Association for Computational Linguistics.

[108] Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, 2013.

[109] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0 : the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1) :D1265–D1275, 2024.

[110] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, et al. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1) :D1207–D1217, 2021.

[111] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. Med7 : A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118 :102086, 2021.

[112] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The

chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1) :1–17, 2015.

[113] Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online, July 2020. Association for Computational Linguistics.

[114] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA, 2001.

[115] Veronica Latcinnik and Jonathan Berant. Explaining question answering models through text generation. *arXiv preprint arXiv :2004.05569*, 2020.

[116] John Lawrence and Chris Reed. Argument mining : A survey. *Computational Linguistics*, 45(4) :765–818, 2020.

[117] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240, 2020.

[118] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis : Building a better stroke prediction model. 2015.

[119] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus : a resource for chemical disease relation extraction. *Database*, 2016 :baw068, 2016.

[120] Xin-Yi Li, Wei-Jun Lei, and Yu-Bin Yang. From easy to hard : Two-stage selector and reader for multi-hop question answering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[121] Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. Rxnorm : prescription for electronic drug information exchange. *IT professional*, 7(5) :17–23, 2005.

[122] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[123] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT : generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 2022.

[124] Martín Díaz Maffini, Fernanda Aguirre Ojea, and Matías Manzotti. Automatic detection of vital signs in clinical notes of the outpatient settings. In *MIE*, pages 1211–1212, 2020.

[125] Enrico Manzini, Jon Garrido-Aguirre, Jordi Fonollosa, and Alexandre Perera-Lluna. Mapping layperson medical terminology into the human phenotype ontology using neural machine translation models. *Expert Systems with Applications*, 204 :117446, 2022.

[126] Santiago Marro, Benjamin Molinet, Elena Cabrio, and Serena Villata. Natural Language Explanatory Arguments for Correct and Incorrect Diagnoses of Clinical Cases. In *ICAART 2023 - 15th International Conference on Agents and Artificial Intelligence*, volume 1 of *Proceedings of the 15th International Conference on Agents and Artificial Intelligence - (Volume 1)*, pages 438–449, Lisbon (Portugal), Portugal, February 2023.

[127] Tobias Mayer. *Fouille d'arguments à partir des essais cliniques*. PhD thesis, Université Côte d'Azur, 2020.

[128] Tobias Mayer, Elena Cabrio, and Serena Villata. Acta : A tool for argumentative clinical trial analysis. In *IJCAI 2019-Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6551–6553, 2019.

[129] Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press, 2020.

[130] Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*, 118 :102098, 2021.

[131] Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. Loinc, a universal standard for identifying laboratory observations : a 5-year update. *Clinical chemistry*, 49(4) :624–633, 2003.

[132] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. UmlsBERT : Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1744–1753, Online, June 2021. Association for Computational Linguistics.

[133] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013.

[134] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence*, 267 :1–38, 2019.

[135] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artif. Intell.*, 267 :1–38, 2019.

[136] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. *arXiv preprint arXiv :1906.02916*, 2019.

[137] Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial intelligence and law*, 19 :1–22, 2011.

[138] Sanjay Modgil and Henry Prakken. The aspic+ framework for structured argumentation : a tutorial. *Argument & Computation*, 5(1) :31–62, 2014.

[139] Sunil Mohan and Donghui Li. Medmentions : A large biomedical corpus annotated with UMLS concepts. 2019.

[140] Benjamin Molinet, Santiago Marro, Elena Cabrio, and Serena Villata. Explanatory argumentation in natural language for correct and incorrect medical diagnoses. *Journal of Biomedical Semantics*, 15(1) :8, 2024.

[141] Benjamin Molinet, Santiago Marro, Elena Cabrio, Serena Villata, and Tobias Mayer. Acta 2.0 : A modular architecture for multi-layer argumentative analysis of clinical trials. In *IJCAI 2022-Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[142] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[143] George B Moody, Roger G Mark, and Ary L Goldberger. Physionet : a web-based resource for the study of physiologic signals. *IEEE Engineering in Medicine and Biology Magazine*, 20(3) :70–75, 2001.

[144] Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. End-to-end argument mining with cross-corpora multi-task learning. *Transactions of the Association for Computational Linguistics*, 10 :639–658, 2022.

[145] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5 ?! training text-to-text models to explain their predictions, 2020.

[146] Usman Naseem, Matloob Khushi, V. Balakista Reddy, S. Rajendran, Imran Razzak, and Jinman Kim. Bioalbert : A simple and effective pre-trained language model for biomedical named entity recognition. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2021.

[147] Hillary Ngai and Frank Rudzicz. Doctor XAvIer : Explainable diagnosis on physician-patient dialogues and XAI evaluation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 337–344, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[148] Huy Nguyen and Diane Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[149] Kosuke Nishida, Kyosuke Nishida, Itsumi Saito, and Sen Yoshida. Towards interpretable and reliable reading comprehension : A pipeline model with unanswerability prediction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[150] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[151] Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Playing the part of the sharp bully : Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics : ACL 2023*, pages 2758–2772, 2023.

[152] Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Hideki Mima, and Junichi Tsujii. The GENIA corpus : An annotated research abstract corpus in molecular biology do-

main. In *Proceedings of the human language technology conference*, pages 73–77, San Francisco, CA, USA, 2002. Citeseer, Morgan Kaufmann Publishers Inc.

[153] Cathy O'neil. *Weapons of math destruction : How big data increases inequality and threatens democracy*. Crown, 2017.

[154] OpenAI. Gpt-4 technical report, 2023.

[155] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[156] Olga V Patterson, Makoto Jones, Yiwen Yao, Benjamin Viernes, Patrick R Alba, Theodore J Iwashyna, and Scott L DuVall. Extraction of vital signs from clinical notes. *Studies in health technology and informatics*, 216 :1035–1035, 2015.

[157] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe : Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[158] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. SciFive : a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598, 2021.

[159] Steven T Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3) :280–291, 2012.

[160] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multicenter database for critical care research. *Scientific data*, 5(1) :1–13, 2018.

[161] Sampo Pyysalo and Sophia Ananiadou. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6) :868–875, 2014.

[162] Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, and Jun'ichi Tsujii. Towards exhaustive protein modification event extraction. In *Proceedings of BioNLP 2011 Workshop*, pages 114–123, United Kingdom, 2011. Oxford University Press.

[163] Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A Beck, Thomas E Feasby, and William A Ghali. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care*, 43(11) :1130–1139, 2005.

[164] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[165] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8) :9, 2019.

[166] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv :1910.10683*, 2019.

[167] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1) :5485–5551, 2020.

[168] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca : Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[169] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David A. Lagnado, and Francesca Toni. Argumentative explanations for interactive recommendations. *Artif. Intell.*, 296 :103506, 2021.

[170] Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999.

[171] Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12) :e0000152, 2022.

[172] Nils Reimers and Iryna Gurevych. Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[173] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1) :57–87, 1997.

[174] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you ?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[175] Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1) :1, 2020.

[176] Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. Thinking like a skeptic : Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 4661–4675, 2020.

[177] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. *arXiv preprint arXiv :1910.01108*, 2019.

[178] Isabel Sassoon, Nadin Kökciyan, Elizabeth Sklar, and Simon Parsons. Explainable argumentation for wellness consultation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems : First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1*, pages 186–202. Springer, 2019.

[179] Hendrik Schuff, Heike Adel, Peng Qi, and Ngoc Thang Vu. Challenges in explanation quality evaluation. *arXiv preprint arXiv :2210.07126*, 2022.

[180] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3) :379–423, 1948.

[181] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[182] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *arXiv preprint*, abs/2212.13138, 2022.

[183] Larry Smith, Lorraine K Tanabe, Cheng-Ju Kuo, I Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, et al. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2) :1–19, 2008.

[184] SP Somashekhar, M-J Sepúlveda, S Puglielli, AD Norden, Edward H Shortliffe, C Rohit Kumar, A Rauthan, N Arun Kumar, P Patil, Kyu Rhee, et al. Watson for oncology and breast cancer treatment recommendations : agreement with an expert multidisciplinary tumor board. *Annals of Oncology*, 29(2) :418–423, 2018.

[185] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3) :331–336, 2018.

[186] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1) :11–21, 1972.

[187] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3) :619–659, 2017.

[188] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, page 102–107, USA, 2012. Association for Computational Linguistics.

[189] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank : an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3) :e1001779, 2015.

[190] Wei Sun, Mingxiao Li, Jingyuan Sun, Jesse Davis, and Marie-Francine Moens. Dmon : A simple yet effective approach for argument structure learning. *arXiv preprint arXiv :2405.01216*, 2024.

[191] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5) :806–813, 2013.

[192] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[193] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Interspeech*, volume 2012, pages 194–197, 2012.

[194] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv :1409.3215*, 2014.

[195] Ekaterina Sviridova, Anar Yeginbergenova, Ainara Estarrona, et al. Casimedicos-arg : A medical question answering dataset annotated with explanatory argumentative structures. In *EMNLP Proceedings 2024*, 2024.

[196] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning : Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502, 2009.

[197] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai) : Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11) :4793–4813, 2020.

[198] S.E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

[199] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16 :138 :1–138 :28, 2015.

[200] Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. Select, answer and explain : Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9073–9080, 2020.

[201] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence : a survey. *The Knowledge Engineering Review*, 36 :e5, 2021.

[202] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[203] Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, et al. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 LREC/COLING*, pages 1519–1538. ELRA and ICCL, 2024.

[204] Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. Redhot : A corpus of annotated medical questions, experiences, and claims on social media. *arXiv preprint arXiv :2210.06331*, 2022.

[205] Toshiko Wakaki, Katsumi Nitta, and Hajime Sawamura. Computing abductive argumentation in answer set programming. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 195–215. Springer, 2009.

[206] H Kenneth Walker, W Dallas Hall, and J Willis Hurst. Clinical methods : the history, physical, and laboratory examinations. 1990.

[207] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[208] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning : a proof-of-concept trial. *Nature Medicine*, 29(10) :2633–2642, 2023.

[209] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. ClinicalGPT : Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *ArXiv preprint*, abs/2306.09968, 2023.

[210] Jianfang Wang. On freeman's argument structure approach. In *Chinese Conference on Logic and Argumentation*, 2016.

[211] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[212] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[213] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[214] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv :1908.04626*, 2019.

[215] T Wolf. Huggingface's transformers : State-of-the-art natural language processing. *arXiv preprint arXiv :1910.03771*, 2019.

[216] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online, October 2020, accessed 2024. Association for Computational Linguistics.

[217] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA : Towards building open-source language models for medicine. *arXiv preprint*, 2304.14454, 2023.

[218] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*, 2016.

[219] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5 : A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics, 2021.

[220] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa : A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv :1809.09600*, 2018.

[221] Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. Argument Mining in Data Scarce Settings : Cross-lingual Transfer and Few-shot Techniques. In *ACL*, 2024.

[222] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*, 2019.

[223] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications*, 117 :42–61, 2019.

[224] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations : A survey on methods and metrics. *Electronics*, 10(5) :593, 2021.

List of Figures

Appendices

APPENDICES A

A.1 Fully Annotated Clinical Case

This appendice presents two colored fully annotated clinical cases from the MEDQA-USMLE-Symp dataset in Examples A.1.1 and A.1.2. The entities are **Sign or Symptoms** in orange, **No Sign or Symptoms** in red, **Findings** in blue, **locations** in green, **temporal concepts** in pink, **Population group** in cyan, **Age group** in teal and finally, the **Sign or Symptoms** associated to the correct answer in **orange bold**. Dataset description and annotation details are available in Section 3.1.

Exemple A.1.1 –

Clinical Case. A 37-year-old woman is brought to the emergency department because of intermittent ***chest pain*** for 3 days. The pain is worse with inspiration, and she feels she cannot take deep breaths. She has not had shortness of breath, palpitations, or nausea. She had an upper respiratory tract infection 10 days ago and took an over-the-counter cough suppressant and decongestant and acetaminophen. Her temperature is 37.2°C (98.9°F), pulse is 90/min, and blood pressure is 122/70 mm Hg. The lungs are clear to auscultation. S1 and S2 are normal. A ***rub is heard during systole***. There is no peripheral edema. An ECG shows normal sinus rhythm and *diffuse, upwardly concave ST-segment elevation* and PR-segment depression in leads II, III, and aVF.

Question. Which of the following is the most likely diagnosis ?

Options. ['Acute pericarditis', 'Aortic dissection', 'Gastroesophageal reflux disease', 'Myocardial infarction', 'Peptic ulcer disease', 'Pulmonary embolism', 'Unstable angina pectoris']

Correct Answer. Acute pericarditis

Exemple A.1.2 –

Clinical Case. A previously healthy 34-year-old woman is brought to the physician because of ***fever*** and ***headache*** for 1 week. She has not been exposed to any disease. She takes no medications. Her temperature is 39.3°C (102.8°F), pulse is 104/min, respirations are 24/min, and blood pressure is 135/88 mm Hg. She is ***confused*** and ***oriented only to person***. Examination shows jaundice of the skin and conjunctivae. There are a few scattered ***petechiae*** over the trunk and back. There is no lymphadenopathy. Physical and neurologic examinations show no other abnormalities. Test of the stool for occult blood is positive. Laboratory studies show :

- Hematocrit 32% with fragmented and nucleated erythrocytes
- Leukocyte count 12,500/mm3
- Platelet count 20,000/mm3
- Prothrombin time 10 sec
- Partial thromboplastin time 30 sec

— Fibrin split products negative
— Serum
— Urea nitrogen 35 mg/dL
— Creatinine 3.0 mg/dL
— Bilirubin
— Total 3.0 mg/dL
— Direct 0.5 mg/dL
— Lactate dehydrogenase 1000 U/L

Blood and urine cultures are negative. A CT scan of the head shows no abnormalities.

Question. Which of the following is the most likely diagnosis ?

Options. ['Disseminated intravascular coagulation', 'Immune thrombocytopenic purpura', 'Meningococcal meningitis', 'Sarcoidosis', 'Systemic lupus erythematosus', 'Thrombotic thrombocytopenic purpura']

Correct Answer. Thrombotic thrombocytopenic purpura

A.2 Findings converter experiments prompts

This appendice shows the prompts used in our findings converter experiments. The label *[FINDING]* is replaced on the fly by the current finding name.

A.2.1 Prompt system

```
Ignore all instructions before this one.
You're a doctor assistant.
You have been doing medicine for 20 years.
Your task is now to return the medical terms associated to
findings.
```

A.2.2 IO configuration

```
Fill the following table by replacing "?":

| Finding | Medical term |
| [FINDING] | ? |

ONLY fill the table with ONE line, no extra sentences
Put "-" if the value is normal
```

A.2.3 CoT and SC configurations

Prompt 1 :

```
We focus on the finding '[FINDING]'

| Finding | Low reference value | High reference value |
| [FINDING]               | ?         | ?           |

ONLY fill the table with ONE line with LOW and HIGH, no extra
sentences.
```

Prompt 2 :

```
Now we want to associate a medical term to this finding
'[FINDING]'

| Finding | Medical term |
| [FINDING] | ? |

ONLY fill the table with ONE line, no extra sentences
Put "-" if the value is normal
```

A.3 Findings converter experiments prompts

Exemple A.3.1 – T1 Premise 44 294 a woman with an epileptic seizure presenting with the following clinical features : epigastric aura, unpleasant odor, disconnection from the environment, motor automatisms (sucking, swallowing, opening and closing of one hand) and postcritical amnesia
T2 Claim 334 387 Generalized non-convulsive seizure or typical absence
T3 Claim 392 419 Continuous partial epilepsy
T4 Claim 424 440 Amyotonic crisis
T5 Claim 445 482 Complex partial temporal lobe seizure
T6 Claim 509 589 Clearly the answer is 4, with a very characteristic clinic of temporary seizures
R1 Support Ent1 :T6 Ent2 :T5

These annotation shows components with lines starting with a “T” and relations with “R”. The second column show the label of T and R and the span boundaries for T and the entities linked for R. Finally, the last column for the T lines is the component content.