On the Automatic Assessment of Natural Language Expert Explanations in Medicine

Santiago Marro[1,*], Theo Alkibiades Collias[1], Elena Cabrio[1] and Serena Villata[1]

[1] *Université Côte d'Azur, Inria, CNRS, I3S, France 930 route des Colles - Bât. Les Templiers 06903 SOPHIA ANTIPOLIS cedex, France*

Abstract

The importance of explanations in decision-making, particularly in the medical domain, has been widely recognized. However, the evaluation of the quality of these explanations remains a challenging task. In this work, we propose a novel approach for assessing and evaluating the reasons provided in explanations about clinical cases. Our approach leverages an external knowledge base and a defined prevalence function to score each reason based on its pertinence in the domain. By applying a deterministic prevalence function, we ensure total transparency of the reasons' assessment, facilitating a precise explanation of the rationale behind the scoring hierarchy of each reason. We demonstrate the effectiveness of our approach in clinical cases, where medical experts explain the rationale behind a specific diagnosis and why other potential diagnoses are dismissed. Our methodology provides a nuanced and detailed evaluation of the explanation, contributing to a more comprehensive understanding of the decision-making process.

Keywords

NLP, Named Entity Recognition, Healthcare

1. Introduction

Explainable Artificial Intelligence (XAI) [1, 2] has emerged as a central topic in contemporary AI research, given the predominance of black box methods on the one hand, and their application to sensitive domains such as medicine and education on the other hand. AI systems support human decision-making like in medical diagnosis. Nonetheless, the efficacy of these systems is dependent on their capability to deliver explanations that are comprehensible and significant to the user [3, 4]. Recent work shows how the best-known XAI approaches fail to provide sound explanations, or that alternatively find explanations that can exhibit significant redundancy [5].

To address this challenging open issue, we propose a novel approach for an assessment and evaluation of the reasons employed in explanations, which satisfies transparency as well. More precisely, our goal is to automatically evaluate the relevance of all conceivable reasons that

22nd International Conference of the Italian Association for Artificial Intelligence, 6 - 9 Nov, 2023, Rome, Italy

[*] Corresponding author.

santiago.marro@inria.fr (S. Marro); theo-alkibiades.collias@inria.fr (T. A. Collias); elena.cabrio@inria.fr (E. Cabrio); serena.villata@inria.fr (S. Villata)

0000-0001-6220-0559 (S. Marro); 0000-0001-9374-7872 (E. Cabrio); 0000-0003-3495-493X (S. Villata)

could explain a particular event, and subsequently compare them with the reasons invoked by the explainer.

When applied to the medical field, our approach scrutinizes explanations provided in medical examinations, wherein medical residents elucidate a specific diagnosis of a patient, given the context (i.e., a clinical case detailing the patient's condition) and their medical expertise. Consequently, we generate an assessment that identifies the reasons employed in the explanation and evaluates them against the relevance scoring produced by our approach. Our approach leverages an external knowledge base, the Human Phenotype Ontology (HPO) [6], and a deterministic prevalence function to score each reason based on its pertinence in the domain. This function allows to elaborate the resulting reasons' scores, in a transparent way. We evaluate our approach on the Antidote Casimedicos dataset [7], a unique resource comprising 621 clinical case descriptions, each with a set of potential diagnoses, an indicator of the correct answer, and a detailed explanation of the decision-making process provided by medical professionals. The results obtained on this dataset show the effectiveness of the proposed approach.

While our methodology is assessed on a use case from the medical domain, it is abstract enough to be applied to any domain. We envision two potential scenarios where our methodology could be particularly beneficial: AI for education and online medical fora.

In the context of AI for education, our approach can assist medical resident students in learning how to solve medical cases and develop a logical and explainable reasoning process in order to explain a diagnosis. By providing a systematic and transparent way of evaluating the reasons given in explanations, we can help students understand the rationale behind a specific diagnosis and why other potential diagnoses are dismissed.

In online medical fora, our approach can help online users to distinguish good explanations from bad explanations present there. Users often discuss diagnoses and share their experiences, but the quality of these discussions can vary widely. With our approach, we can provide a systematic and transparent way of evaluating the reasons given in these discussions, helping users and moderators identify high-quality explanations and promote more informed discussions.

The research presented in this paper is driven by the necessity for a systematic and transparent methodology to assess the pertinence of reasons used in medical explanations. To the best of our knowledge, this is the first approach that leverages an external knowledge base, the Human Phenotype Ontology (HPO), and a deterministic prevalence function to evaluate the reasons for the potential diagnoses based on their relevance in the context of a specific clinical case and grounding on the HPO knowledge base.

The paper is organized as follows: after a comparison with the related work, we first describe our data preprocessing heuristics (Section 3.1), followed by the extraction and encoding of reasons from the clinical case into HPO terms (Section 3.2). Next, we describe the computation of the pertinence score for each reason using the prevalence function (Section 3.3) and discuss its deterministic nature. Then, we illustrate the sentence-matching approach employed to align the extracted reasons with those found in the explanation (Section 3.5). Finally, we demonstrate the generation of a pertinence assessment of the reasons following a template-based generation technique (Section 3.6).

2. Related Work

In this section, we discuss the related work on explanation selection and then we focus on the medical domain application scenario.

Explanation Selection. In the process of explanation selection, individuals choose what they perceive to be the most relevant causes from a larger set of causes for a particular event. This selection is not arbitrary and is guided by criteria such as temporality, abnormality, intention, and the differences between a fact and a foil [8]. Hilton [9] sustains this is due to the fact that causal chains are often too large to comprehend.Research shows that the primary way individuals select explanations is by contrasting a fact and a foil. The fact refers to the actual state of affairs, while the foil represents an alternative state that did not occur. The contrast between the fact and the foil forms the basis for explanation selection, with the explanation that highlights the greatest number of differences between the fact and the foil deemed to have the highest explanatory power [10]. Contrastive explanation is a concept that further elaborates on this idea. It posits that the differences between two events form the basis for explanation. This theory has garnered support from experimental research in cognitive science, which suggests that people perform causal inference, explanation, and generalization based on contrastive cases [11, 12]. Abnormality also plays a crucial role in explanation selection. Hilton and Slugoski [13] propose the abnormal conditions model, arguing that abnormal events are key in causal explanation. This model suggests that individuals use their perceived background knowledge to select conditions that are considered abnormal. This model has been supported by subsequent experimental studies [14, 15, 16]. In this paper, we introduce an approach that not only evaluates the relevance of each potential explanation for a given event but also incorporates the principles of abnormality and contrastive explanation into the calculation of the relevance score.

XAI for the medical domain. The importance of explanations in AI systems, particularly in the medical domain, has been extensively studied [17, 18, 19]. In the context of medical diagnosis, explanations often involve identifying the key reasons or symptoms that led to a specific diagnosis. The Human Phenotype Ontology (HPO) [6] provides a standardized vocabulary of phenotypic abnormalities encountered in human disease, which can be used to facilitate the assessment of explanations in this domain. Our work builds upon this ontology by developing an approach that assesses the selected reasons in explanations. The National Institutes of Health (NIH) Undiagnosed Diseases Program (UDP) [20] has also investigated the use of HPO in the context of diagnosing and evaluating patients with conditions that have eluded diagnosis. The clinical features of a patient are encoded into HPO terms, which are then used to retrieve a list of candidate diseases that might explain the patient's phenotype. This list is then examined by a clinician to identify the most likely diagnosis. Our methodology extends this approach by not only using HPO to facilitate diagnosis but also to evaluate the reasons given in explanations.

3. Assessing Reasons used in Explanations

Our approach to assessing the reasons used in explanations is visualized in Figure 1. We start with a clinical case of a patient, supplemented by an explanation provided by a medical expert, which elaborates on the specific diagnosis attributed to the patient. The objective is to evaluate the identified reasons in the clinical case and the external knowledge, and compare them with the reasons invoked by the expert to justify the medical diagnosis. To achieve this, we compute a *pertinence score* using a deterministic prevalence function (see Section 3.3), which ensures complete transparency, allowing us to explain why each reason is more or less pertinent than the others with respect to the given case.

The HPO Ontology [6] and, more precisely, its associated knowledge base (KB) will serve as our external knowledge source by providing a standardized vocabulary of phenotypic abnormalities, namely symptoms and findings encountered in human diseases. The HPO contains over 13,000 terms describing phenotypic abnormalities seen in human disease. It uses a directed acyclic graph structure to represent the relationships between terms, allowing for flexible descriptions. Most terms have textual definitions and synonyms. The ontology terms describe clinical abnormalities at different levels of specificity, from general (e.g. *Abnormal ear morphology*) to very precise (e.g. *Chorioretinal atrophy*). This KB facilitates the evaluation of the reasons in the explanation by providing standardized vocabulary and semantic relationships between phenotypic concepts relevant to human disease.

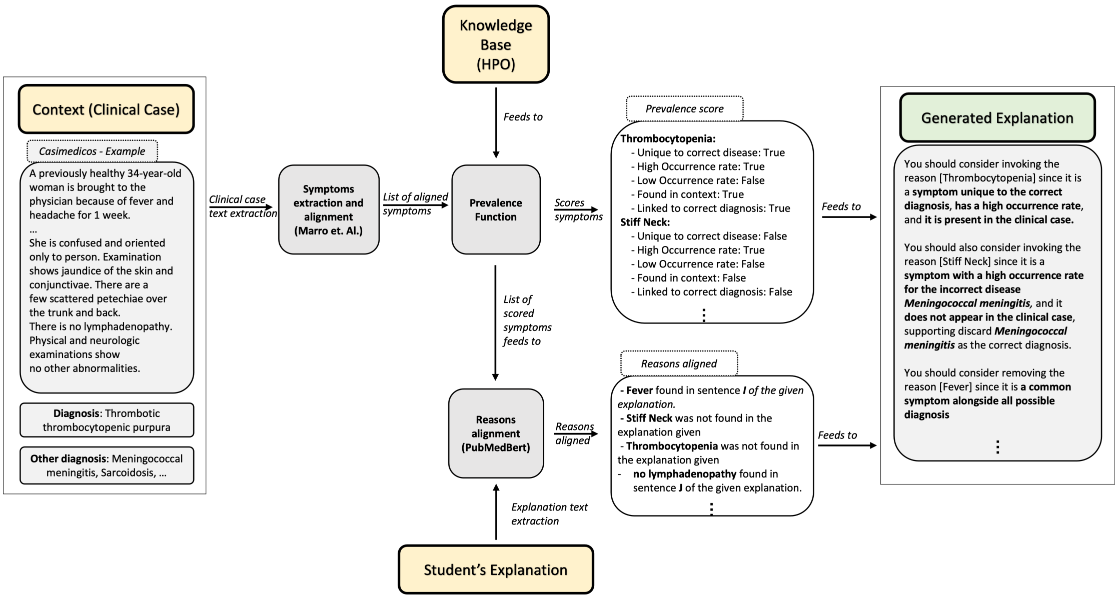
Our approach consists of two main steps: *(i)* the reasons given in the explanation are extracted from the clinical case and encoded into HPO terms (following the approach of Marro et al. [21]), in which a medical Named Entity Recognition (NER) step is performed, to then align them into HPO terms. This allows us to retrieve all the standardized information the ontology contains, such as the definition and the frequency of occurrence of that term in actual patient cases for each possible disease; *(ii)* the pertinence score for each reason is computed using the prevalence function, which takes into account the relevance of each reason in the context of the specific clinical case and the knowledge base.

3.1. Dataset and preprocessing

The foundation for our work is the Antidote Casimedicos dataset [7]. CasiMedicos [1] is a community and collaborative project where each member, developer or contributor adds something unique related to medical exams with the objective of enriching the information that will be available to students. In casiMedicos resource, the MIR exams were chosen and commented on by Spanish medical doctors in a voluntary effort to provide *answers* and *explanations* to the MIR exams. In the MIR exams source, there are 953 commented questions (153K words) that have been extracted from the MIR exams held between the years 2005, 2014, 2016, 2018, 2019, 2020, 2021 and 2022.

This dataset is unique in the medical domain as it consists of 621 theoretical, fictional clinical cases that provide real expert-crafted explanations, making it a valuable resource for students. Each clinical case is paired with a set of potential diagnoses, an indicator of the correct answer,

[1] https://www.casimedicos.com/



A previously healthy 34-year-old woman is brought to the physician because of fever and headache for 1 week.
...
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.

Diagnosis: Thrombotic thrombocytopenic purpura

Other diagnosis: Meningococcal meningitis, Sarcoidosis, ...

Thrombocytopenia:
- Unique to correct disease: True
- High Occurrence rate: True
- Low Occurrence rate: False
- Found in context: True
- Linked to correct diagnosis: True

Stiff Neck:
- Unique to correct disease: False
- High Occurrence rate: True
- Low Occurrence rate: False
- Found in context: False
- Linked to correct diagnosis: False

⋮

- **Fever** found in sentence ***I** of the given explanation.*
- **Stiff Neck** was not found in the explanation given
- **Thrombocytopenia** was not found in the explanation given
- **no lymphadenopathy** found in sentence **J** of the given explanation.

⋮

You should consider invoking the reason [Thrombocytopenia] since it is a **symptom unique to the correct diagnosis, has a high occurrence rate,** and **it is present in the clinical case.**

You should also consider invoking the reason [Stiff Neck] since it is a **symptom with a high occurrence rate for the incorrect disease *Meningococcal meningitis,*** and it **does not appear in the clinical case**, supporting discard ***Meningococcal meningitis*** as the correct diagnosis.

You should consider removing the reason [Fever] since it is **a common symptom alongside all possible diagnosis**

⋮

Figure 1: Overview of our approach for the automatic assessment of explanation's reasons.

and a detailed explanation of the decision-making process provided by medical professionals[2]. What sets this dataset apart from other Question Answering (QA) datasets in the medical domain is its inclusion of explanatory arguments for both the correct diagnosis or treatment and the reasons why other options are incorrect. These explanations, written by medical doctors, offer a rich source of information for research in XAI.

To prepare the data, we first enhance contextual information in the explanations by expanding abbreviated diagnosis references. It is common for the explainer to refer to diagnoses as "Answer 1", thus we implement a string replacement with the corresponding answer. Subsequently, as delineated in [7], the dataset encompasses various types of questions. For the purpose of evaluating the explanations provided by the experts, we manually filter out cases that solely discuss potential diagnoses of the patients, yielding a total of 206 clinical cases.

3.2. Identification and Alignment of Potential Causes

The initial phase of our approach (Figure 1) consists in identifying all potential causes within the given context. In our medical scenario, the context is represented by clinical cases, and we regard all symptoms as potential causes that could explain the patient's diagnosis. To address this, we utilize the clinical information extraction pipeline proposed by Marro et al. [21], which performs two key steps:

(i) *Medical named entity recognition*: Marro et al. first identify medical concepts and abnormalities described in the clinical case text using a named entity recognition (NER) system. Their NER module detects mentions of symptoms, findings, and other phenotypic concepts, labelling

[2] https://github.com/ixa-ehu/antidote-casimedicos

them with semantic tags like "Sign or Symptom" or "Finding." Marro et al. trained this module on a dataset of 314 annotated clinical cases, achieving promising results for symptom detection with an F1-Score of 0.86 using a transformer-based model. The NER model can accurately label symptoms described in layperson terminology, an important capability since clinical cases are often written for student exams, where terms may not appear standardised.

(ii) *Alignment to ontology terms*: Next, Marro et al. map the detected semantic concepts to the standardized terminology of the Human Phenotype Ontology (HPO) [6]. This mapping enables linking colloquial descriptions like "shortness of breath" to the formal HPO concept of "Dyspnea." They compute contextual embeddings for each identified symptom and find the closest HPO match via cosine similarity. Their top-performing alignment approach attained 0.53 accuracy in aligning symptoms to equivalent ontology terms in the top 5 matches.

The output we obtain from their pipeline is a list of symptoms extracted from the case text and aligned to ontology concepts. We select the *Sign or Symptom* and *No Sign or Symptom* entities, as these denote the presence or absence of abnormalities relevant for diagnosis. This aligned set of reasons extracted from the specific clinical case serves as our starting point. We later expand this list by retrieving additional related symptoms from the HPO knowledge base.

Ultimately, Marro et al.'s techniques provide us with an initial set of ontology-grounded reasons consisting of the aligned symptoms and no symptoms identified within the context of the particular clinical case description. We subsequently feed this encoded evidence as input to the prevalence function, which will evaluate and score the relevance of these reasons from the case alongside other reasons derived from the external HPO knowledge.

3.3. Prevalence Function

The Prevalence Function[3] is a central component of our approach, designed to systematically assess the pertinence of each possible reason that could explain a given event. This function is inspired by cognitive processes involved in explanation selection, aiming to replicate these processes in a transparent and replicable manner. The function takes into account the possible reasons found in the clinical case (see Section 3.2) and a KB that serves to find other possible reasons outside of the context but still relevant to the case. It then evaluates all possible reasons based on a set of predefined conditions, each of which contributes to the final prevalence score of the key reason. These conditions include whether the key reason is linked to the correct or incorrect diagnosis, its occurrence rate, and whether it is unique to the correct diagnosis, or shared among all possible diagnoses. In line with the abnormal conditions model, our function assigns a higher score to key reasons that are unique to the correct diagnosis and have a low occurrence rate. This reflects the idea that abnormal conditions, i.e., conditions that do not usually occur, are more likely to be the cause of an event. Moreover, our approach integrates the concept of contrastive explanation. For instance, if a symptom associated with an incorrect diagnosis has a high occurrence rate and does not appear in the clinical case, it can be invoked to discard the incorrect diagnosis. This aligns with the idea that the differences between two events form the basis for the explanation.

The computation of the prevalence function starts with the acquisition of the set of potential reasons to be evaluated. In the context of medical diagnosis, these reasons correspond to

[3] The full system implementation will be available upon acceptance

symptoms, which can be identified either within the clinical case or within the Human Phenotype Ontology (HPO) as symptoms associated with each potential diagnosis. The ontology provides not only a list of symptoms for each diagnosis but also pertinent information about each symptom, such as its occurrence rate, definition, and synonyms. This information facilitates the definition of three distinct sets of reasons, which serve as the basis for the computation of the Prevalence function:

- $SymptomsOfCorrectDiagnosis$: symptoms that belong to the correct disease;
- $SymptomsOfIncorrectDiagnosis$: symptoms that belong to all the incorrect diseases;
- $PresentSymptoms$: symptoms found in the case description.

The Prevalence Function is then used in conjunction with the additional disease information and symptom sets obtained from the HPO to produce a list of key reasons and their calculated prevalence scores. This allows us to provide a robust and transparent framework for assessing the quality of the reasons on which the explanations are grounded.

3.4. Algorithm explanation

The Prevalence Score Function, as outlined in Algorithm 0, is designed to assess the relevance or pertinence of a given key reason in the context of a clinical case (CC) and a knowledge base (KB). The function operates by assigning a score to the key reason based on its presence in the correct and incorrect diagnoses, its occurrence rate, and its presence in the clinical case.

The function begins by initializing the score to zero and setting several boolean variables to false (lines 2-6). It then retrieves the symptoms associated with the correct and incorrect diagnoses from the knowledge base (lines 7-8) and identifies the symptoms present in the clinical case using Named Entity Recognition (NER) (line 9).

Then, it checks if the key reason is present in the symptoms of the correct diagnosis (lines 11-15). If it is, the function increments the score and sets the variable *linkedToCorrectDiagnosis* to true. If not, *linkedToCorrectDiagnosis* is set to false.

Next, the function checks if the key reason is present in the symptoms of the incorrect diagnoses (lines 17-20). If it is, the variable *linkedToIncorrectDiagnosis* is set to true. If not, it is set to false.

The function then checks the occurrence rate of the key reason (lines 22-29). If the key reason has a high occurrence rate (more than 70%), the variable *hasHighOccurrenceRate* is set to true, and if it is linked to the correct diagnosis, the score is incremented. If the key reason has a low occurrence rate (less than 30%), the variable *hasLowOccurrenceRate* is set to true.

It then checks if the key reason is unique to the correct diagnosis or shared with other diagnoses (lines 31-40). If the key reason is unique to the correct diagnosis and has a low occurrence rate, the score is incremented twice. If the key reason is shared with other diagnoses, the score is decremented.

Finally, the function checks if the key reason is present in the symptoms of the incorrect diagnoses but not in the present symptoms (lines 42-48). If the key reason has a high occurrence rate, the score is incremented. Otherwise, the score is decremented.

The final score represents the prevalence of the key reason in the context of the specific clinical case, providing a measure of its relevance or pertinence.

Algorithm 1 Prevalence Function

```
procedure PrevalenceFunction($Symptom$, $CC$, $CorrectDisease$, $IncorrectDiseases$, $KB$)
    $score = 0$
    $uniqueToCorrectDiagnosis = False$
    $sharedToOtherDiagnosis = False$
    $hasLowOccurrenceRate = False$
    $hasHighOccurrenceRate = False$
    $SymptomsOfCorrectDiagnosis$ = $KB(CorrectDisease)$
    $SymptomsOfIncorrectDiagnosis$ = $KB(IncorrectDiseases)$
    $PresentSymptoms$ = $NER(CC)$
    if $Symptom$ is in $SymptomsOfCorrectDiagnosis$ then
        $linkedToCorrectDiagnosis = True$
        score = score + 1
    else
        $linkedToCorrectDiagnosis = False$
    end if
    if $Symptom$ is in $SymptomsOfIncorrectDiagnosis$ then
        $linkedToIncorrectDiagnosis = True$
    else
        $linkedToIncorrectDiagnosis = False$
    end if
    if $Symptom$ has a high occurrence rate (more than 70%) then
        $hasHighOccurrenceRate = True$
        if $linkedToCorrectDiagnosis == True$ then
            score = score + 1
        end if
    else if $Symptom$ has a low occurrence rate (less than 30%) then
        $hasLowOccurrenceRate = True$
    end if
    if $Symptom$ is in $SymptomsOfCorrectDiagnosis$ then
        if $Symptom$ is not in $SymptomsOfIncorrectDiagnosis$ then
            $uniqueToCorrectDiagnosis = True$
            $score = score + 1$
            if $hasLowOccurrenceRate == True$ then
                $score = score + 1$
            end if
        else if $Symptom$ is in $SymptomsOfIncorrectDiagnosis$ then
            $sharedToOtherDiagnosis = True$
            $score = score - 1$
        end if
    end if
    if $Symptom$ is in $SymptomsOfIncorrectDiagnosis$ then
        if $Symptom$ is not in $PresentSymptoms$ then
            if $Symptom$ has a high occurrence rate then
                $score = score + 1$
            else
                $score = score - 1$
            end if
        end if
    end if
    return $score$
end procedure
```

3.5. Reason Alignment via Sentence Matching

A crucial step in our approach is the alignment of potential causes (i.e., reasons) identified in the clinical case with those actually invoked in the expert's explanation. This alignment (visualized as the "Reasons alignment module" in Figure 1) is achieved through a sentence matching technique. The objective of this step is to discern which of the potential reasons identified were actually utilized by the experts in their explanation, thereby enabling subsequent suggestions of modifications to enhance the explanation's pertinence.

Our approach to sentence matching is inspired by the work of Lu et al. [22], particularly their creation of an intermediate dataset using a distance metric for fine-tuning their sentence-matching model. In their work, Lu et al. [22] employ the Jaccard distance to identify sentences with high similarity between complex and simplified texts. We adapt this strategy to our context, aiming to locate similar reasons between the clinical case and the explanations provided by the explainer. In our adaptation of Lu et al.'s approach, we aim to identify similar reasons between the clinical case and the explanations provided by the experts. However, our methodology diverges in two key aspects: the choice of distance metric, and the preprocessing of the texts for comparison. Instead of employing the Jaccard distance, we opt for a process that begins with the detection of medical-named entities within both texts. Following this, the texts are segmented into individual sentences.

Subsequently, we compute sentence embeddings using only the identified named entities. This computation leverages the Sentence Transformers method [23], using various pre-trained models specialized in scientific text. To align sentences from the clinical case with those in the explanations, we employ cosine similarity. A match is considered valid only when the cosine distance is sufficiently close, ensuring that only highly similar sentences are matched, thereby enhancing the precision of our reason alignment process. Our embedding calculation method is also sensible to negation agents by representing the absence of a symptom in a clinical case with the label "No Sign or Symptom" in the NER stage. As an example of this, the sentence "A 62-year-old man with *no* history of alcohol abuse…" will have a different numerical representation than "A 62-year-old man with a history of alcohol abuse…".

3.6. Template-Based Explanation Generation

In the final step of our pipeline, we employ a template-based generation approach to articulate the pertinence of each reason. This approach allows us to generate natural language explanations that are understandable by human users. Each template is designed to address a specific combination of features associated with a reason, and the appropriate template is selected based on the values of these features for each reason. The features considered in our approach are:

- *uniqueToCorrectDiagnosis* indicates whether the reason is unique to the correct diagnosis.
- *sharedToOtherDiagnosis* indicates whether the reason is shared with other diagnoses.
- *hasLowOccurrenceRate* indicates whether the reason has a low occurrence rate.
- *hasHighOccurrenceRate* indicates whether the reason has a high occurrence rate.
- *linkedToCorrectDiagnosis* indicates whether the reason is directly linked to the correct diagnosis.

- *linkedToIncorrectDiagnosis* indicates whether the reason is linked to an incorrect diagnosis.
- *presentInClinicalCase* indicates whether the reason is present in the clinical case.

Based on the values of these features, a template, like the following, is selected to generate the explanation:

- *Template 1*: "You should consider invoking the reason `[reason]` since it is unique to the correct diagnosis, has a high occurrence rate, and is present in the clinical case."
- *Template 2*: "You should also consider invoking the reason `[reason]` since it is a symptom with a high occurrence rate for the incorrect disease `[disease]`, and it does not appear in the clinical case, supporting discard `[disease]` as the correct diagnosis."
- *Template 3*: "You should consider removing the reason `[reason]` since it is a common symptom alongside all possible diagnoses."

This template-based generation approach allows us to generate explanations that are informative and specific to the context of each reason, thereby ensuring the interpretability of the proposed approach.

4. Evaluation

In this section, we first present the experimental setting we propose to assess our approach, and then we discuss the obtained results. Finally, we apply our approach to a clinical case to discuss the final outcome of the pipeline.

4.1. Experimental Setting.

The main experimental component of our task is the named entity-based sentence matching. This task can be decomposed into two subtasks: first, the generation of tuples of similar sentences, and second, the fine-tuning of Language Models (LMs) using the generated dataset. For tuple generation, we implemented two approaches:

Baseline string distance method : As a baseline, we employed a simpler method based on string distance between full sentences, similar to the work of Lu et al. [22]. We first separated each sentence from the clinical case and associated explanation. We then computed the Levenshtein distance between all combinations of case and explanation sentences. If the distance between a pair surpassed a defined threshold (0.5), we kept that tuple as a match. Otherwise, we discarded it. This baseline relies only on the full sentence text, without considering named entities.

Named entity-based method : Subsequently, we generated tuples using the medical-named entities identified by Marro et al.'s pipeline [21]. Their system detects not only symptom mentions but also their absence, allowing us to model the presence or lack of Signs/Symptoms differently. The identified entities are joined into sentences and embedded using Sentence-BERT [23]. We compute cosine distances between all case and explanation sentence pairs, keeping those above a threshold (0.975) as matches.

For both cases, a dataset composed of

$$(case\_id, answer\_sentence_i, case\_sentence_j, \{True|False\})$$

tuples is generated. For the fine-tuning of the LMs, we employ the PyTorch implementation provided by Hugging Face [24]. The experiments were conducted with a batch size of 8, a maximum sequence length of 256, and a learning rate of 2.5e-5 over 4 epochs. We selected all-mpnet-base-v2 [23] as our baseline, and fine-tuned models such as BioBERT v1.2 [25], S-PubMedBert-MS-MARCO [26], and BioBERT-mnli-snli-scinli-scitail-mednli-stsb [27] as more domain-specific LMs under the same experimental setting. Despite each transformer model achieving its best results with a different cosine similarity threshold for performing the named entity-based matching, we kept a threshold value of 0.975 to ensure the matching of sentences with the highest possible semantic coherence.

4.2. Results.

In this section, we present the obtained results on the Casimedicos dataset. We evaluate the quality of explanations written by experts, highlighting both successful and unsuccessful examples. The performance of the sentence matching task is quantified in terms of macro-average precision, recall, and F1 score, as shown in Table 1.

We adopt the all-mpnet-base-v2 model as our baseline, being it the state-of-the-art for sentence embedding computation across various domains. However, our results indicate that domain-specific models outperform this baseline across all metrics. In particular, the model based on PubMedBert [28, 26] demonstrates superior performance, achieving the highest scores in precision, recall, and F1 score (highlighted in bold in Table 1). These results underline the importance of domain-specific models in achieving high-quality sentence matching.

Table 1
Results for named entity-based matching in macro multi-class precision, recall, and F1-score.

Tuples method	Model	P	R	F1
String Distance	all-mpnet-base-v2	0.80	0.84	0.82
String Distance	BioBERT cased v1.2	0.81	0.82	0.82
String Distance	BioBERT-mnli-snli-scinli-scitail-mednli-stsb	0.82	0.84	0.83
String Distance	S-PubMedBert-MS-MARCO	0.82	0.83	0.83
Named Entity	all-mpnet-base-v2	0.84	0.70	0.74
Named Entity	BioBERT cased v1.2	0.84	0.76	0.80
Named Entity	BioBERT-mnli-snli-scinli-scitail-mednli-stsb	0.85	0.79	0.82
Named Entity	S-PubMedBert-MS-MARCO	**0.89**	**0.85**	**0.87**

4.2.1. A Full Example.

To illustrate the outcome of our approach, we present a full clinical case, the expert's explanation, and the assessment of reasons from the CasiMedicos dataset. We consider a clinical case where the correct diagnosis is *Porphyria cutanea tarda*. The other potential diagnoses considered are *Epidermolysis bullosa acquisita*, *Acute intermittent porphyria*, and *Ulerythema ophryogenesis*.

Clinical Case: "A 62-year-old man with a history of significant alcohol abuse, carrier of hepatitis C virus, treated with Ibuprofen for tendinitis of the right shoulder, goes to his dermatologist because after spending two weeks on vacation at the beach he notices the appearance of tense blisters on the dorsum of his hands. On examination, in addition to localization and slight malar hypertrichosis."

Expert's Explanation: "Porphyria Cutanea Tarda: 60% of patients with PCT are male, many of them drink alcohol in excess, women who develop it are usually treated with drugs containing estrogens. Most are males with signs of iron overload, this overload reduces the activity of the enzyme uroporphyrinogen decarboxylase, which leads to the elevation of uroporphyrins. HCV and HIV infections have been implicated in the precipitation of acquired PCT. There is a hereditary form with AD pattern. Patients with PCT present with blistering of photoexposed skin, most frequently on the dorsum of the hands and scalp. In addition to fragility, they may develop hypertrichosis, hyperpigmentation, cicatricial alopecia, and sclerodermal induration."

Assessment of Reasons: The generated explanations for the top and bottom scoring reasons are as follows:

- You should consider invoking the reason *Elevated urinary delta-aminolevulinic acid* since it is a symptom with a high occurrence rate for the incorrect disease Acute intermittent porphyria, and it does not appear in the clinical case, supporting discard Acute intermittent porphyria as the correct diagnosis.
- You should also consider invoking the reason *Abnormal hair morphology* since it is a symptom with a high occurrence rate for the incorrect disease Epidermolysis bullosa acquisita, and it does not appear in the clinical case, supporting discard Epidermolysis bullosa acquisita as the correct diagnosis.
- You should consider invoking the reason *Alcoholism* since it is a symptom unique to the correct diagnosis and present in the clinical case.
- The symptom *Contact dermatitis* does not meet the criteria for a strong reason in this case.
- The symptom *Dry skin* does not meet the criteria for a strong reason in this case.
- The symptom *Dermal atrophy* does not meet the criteria for a strong reason in this case.

The Expert's Explanation for this case attributes the patient's condition to Porphyria Cutanea Tarda (PCT), citing factors such as the patient's gender, alcohol abuse, and the presence of blistering on photoexposed skin. These align with our top-scoring reasons in the Assessment of Reasons, demonstrating the agreement between the expert's explanation and our assessment. In the expert's explanation, the symptom "Abnormal hair morphology" is not mentioned. However, our methodology identifies it as a significant reason that could enhance the explanation. This symptom is common in the incorrect disease *Epidermolysis bullosa acquisita*, but it is not present in the clinical case. Therefore, its absence provides a strong reason to discard *Epidermolysis bullosa acquisita* as the correct diagnosis. This additional information could potentially enhance the expert's explanation by providing further evidence to support the correct diagnosis and rule out other alternatives. This demonstrates the capability of our approach not only to validate the reasons used by the expert but also to suggest new pieces of information that could enrich the explanation.

5. Concluding remarks

In this paper, we present a novel approach to recognizing the intricate cognitive processes that underpin the selection of explanations and strive to emulate these processes in a methodical and transparent manner. By incorporating the principles of abnormality and contrastive explanation, we ensure that our approach is attuned to the subtleties of explanation selection in real-world contexts, with a particular focus on the medical domain. Moreover, by leveraging the Human Phenotype Ontology and medical named entity recognition, we are able to identify and assess potential reasons in a systematic and data-driven manner. This not only allows us to assess the explanations that are consistent with the expert's perspective but also to suggest additional pieces of information that could enhance the explanation itself.

Our work provides a significant step forward in the development of AI systems that can automatically assess explanations both in a cognitively plausible and contextually sensitive manner. It is driven by the necessity for a systematic and transparent approach to assessing the pertinence of reasons used in explanations, particularly in the medical domain. The deterministic nature of the prevalence function ensures the transparency of our approach.

A key advantage of our methodology is its modularity, with distinct components that can be modified or replaced to adapt the approach to new domains or tasks. The prevalence function encapsulates the reasoning model for scoring explanations, while the sentence matching technique, matching strategy, and knowledge base instantiate components tailored to assessing clinical rationales. Each of these modules could be swapped with alternatives better suited for a different domain or application. For example, the prevalence function could be re-designed to match commonsense rather than medical reasoning, while more general lexical resources like WordNet could replace domain-specific ontologies. This flexibility broadens the applicability of our techniques to assess explanations in a wide range of settings. Future work includes the application of this approach to education in medicine, by providing a systematic and transparent way of evaluating the reasons given in medical residents' explanations, and to online discussions in medical fora by helping users and moderators to identify high-quality and low-quality explanations. Additionally, our current implementation is limited to evaluating explanations based solely on symptoms. However, physicians may consider other types of clinical information, such as lab findings, family history, or patient demographics. To address this, we plan to expand the scope of our approach by extracting and encoding additional clinical entities, such as test results, into ontology concepts like HPO. This will enable assessing explanations across a broader range of relevant clinical data.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002. This work was supported by the CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002.

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[2] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.

[3] E. Reiter, Natural language generation challenges for explainable ai, arXiv preprint arXiv:1911.08794 (2019).

[4] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, V. Patkar, Argumentation-based inference and decision making–a medical perspective, IEEE intelligent systems 22 (2007) 34–41.

[5] J. Marques-Silva, A. Ignatiev, Delivering trustworthy ai through formal xai, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 12342–12350.

[6] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott, et al., The human phenotype ontology in 2017, Nucleic acids research 45 (2017) D865–D876.

[7] R. Agerri, I. Alonso, A. Atutxa, A. Berrondo, A. Estarrona, I. Garcia-Ferrero, I. Goenaga, K. Gojenola, M. Oronoz, I. Perez-Tejedor, G. Rigau, A. Yeginbergenova, Hitz@antidote: Argumentation-driven explainable artificial intelligence for digital medicine, 2023.

[8] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[9] D. Hilton, Social attribution and explanation, 2016.

[10] G. Hesslow, The problem of causal selection, Contemporary science and natural explanation: Commonsense conceptions of causality (1988) 11–32.

[11] B. Rehder, A causal-model theory of conceptual representation and categorization., Journal of Experimental Psychology: Learning, Memory, and Cognition 29 (2003) 1141.

[12] B. Rehder, When similarity and causality compete in category-based property generalization, Memory & Cognition 34 (2006) 3–16.

[13] D. J. Hilton, B. R. Slugoski, Knowledge-based causal attribution: The abnormal conditions focus model., Psychological review 93 (1986) 75.

[14] J. L. McClure, R. M. Sutton, D. J. Hilton, Implicit and explicit processes in social judgments: The role of goal-based explanations., Social judgments: Implicit and explicit processes 5 (2003) 306.

[15] J. Samland, M. R. Waldmann, Do social norms influence causal inferences?, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 36, 2014.

[16] D. J. Hilton, L. M. John, The course of events: counterfactuals, causal sequences, and explanation, in: The psychology of counterfactual thinking, Routledge, 2007, pp. 56–72.

[17] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations, KI-Künstliche Intelligenz 34 (2020) 193–198.

[18] C. Panigutti, A. Perotti, D. Pedreschi, Doctor xai: an ontology-based approach to black-box sequential data classification explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 629–639.

[19] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable ai: the new 42?, in: Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2, Springer, 2018, pp. 295–303.

[20] T. Gall, E. Valkanas, C. Bello, T. Markello, C. Adams, W. P. Bone, A. J. Brandt, J. M. Brazill, L. Carmichael, M. Davids, et al., Defining disease, diagnosis, and translational medicine within a homeostatic perturbation paradigm: The national institutes of health undiagnosed diseases program experience, Frontiers in medicine 4 (2017) 62.

[21] S. Marro, B. Molinet, E. Cabrio, S. 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, 2023, pp. 438–449.

[22] J. Lu, J. Li, B. Wallace, Y. He, G. Pergola, NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1079–1091. URL: https://aclanthology.org/2023.findings-eacl.80.

[23] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. 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, Association for Computational Linguistics, Online, 2020, pp. 38–45.

[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[26] P. Deka, A. Jurek-Loughrey, P. Deepak, Improved methods to aid unsupervised evidence-based fact checking for online health news, Journal of Data Intelligence 3 (2022) 474–504.

[27] P. Deka, A. Jurek-Loughrey, et al., Evidence extraction to validate medical claims in fake news detection, in: International Conference on Health Information Science, Springer, 2022, pp. 3–15.

[28] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing,

2020. arXiv:arXiv:2007.15779.