Rationale Trees: Towards a Formalization of Human Knowledge for Explainable Natural Language Processing

Andrea Tocchetti[1,*], Jie Yang[2] and Marco Brambilla[1]

[1]Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Via Giuseppe Ponzio 34, Milano, 20133, Italy

[2]Delft University of Technology, Mekelweg 5, Delft, 2628 CD, Netherlands

Abstract

As powerful and complex language models are being released to the public, understanding their behaviour is more important than ever. Although Explainable Artificial Intelligence (XAI) approaches have been widely applied to NLP models, the explanations they provide may still be complex to understand for human interpreters as these may not be aligned with the reasoning process they apply in language-based tasks. Furthermore, such a misalignment is also present in most XAI datasets as they are not structured to reflect such a fundamental property. Striving to bridge the gap between model and human reasoning, we propose ad hoc formalizations to structure and detail the thought process applied by human interpreters when performing a set of NLP tasks of interest. Hence, we define *rationale mappings*, i.e., representations that organize humans' analytical reasoning steps when identifying and associating the essential parts of the texts involved in a language-based task leading to its output. These are organized in tree structures referred to as *rationale trees* and characterized for each task to enhance their expressiveness. Furthermore, we describe their data collection and storage process. We argue these structures would result in a better alignment between model and human reasoning, hence improving models' explanations, while still being suited for standard explainability processes.

Keywords

Human Knowledge, Knowledge Formalization, Argumentation Mining, Argumentation Theory, Natural Language, Natural Language Processing, Explainable AI, Explainability,

1. Introduction

The recent spread of Large Language Models (LLM) with the release of ChatGPT raised the research community's interest in questioning the capabilities, understandability, and explainability of AI models like never before. Hence, researchers began exploring the potentiality of such systems with a particular focus on Natural Language Processing (NLP) tasks [1, 2, 3, 4, 5]. Likewise, understanding and explaining models' behaviour has always been of fundamental interest to the AI community [6, 7]. Such a multi-faceted scenario spreads across various technical and human-centred research fields, like computer science, philosophy, and many more.

XAI.it 2023 - Italian Workshop on Explainable Artificial Intelligence, 6–9 November, 2023, Rome

*Corresponding author.

andrea.tocchetti@polimi.it (A. Tocchetti); j.yang-3@tudelft.nl (J. Yang); marco.brambilla@polimi.it (M. Brambilla)

0000-0002-1681-5859 (A. Tocchetti); 0000-0002-0350-0313 (J. Yang); 0000-0002-8753-2434 (M. Brambilla)

Consequently, two fundamental objectives must be considered when explaining models: the faithful representation of their behaviour and the design of humanly understandable explanations. Acknowledged the plethora of explainability methods available in the literature [8, 9], recent trends in Explainable Artificial Intelligence (XAI) revealed the increasing importance and research interest in human-centred AI [10, 11]. As human actors are progressively more involved in different aspects of the explainability process, researchers developed various approaches to collect and employ human knowledge to assess and improve explanations [12]. Hence, an effort to bridge the gap between humans and models in XAI is of fundamental interest to the AI community.

In the context of NLP, researchers defined explanations by identifying the most important words in the input(s) [13], providing the most influential training examples affecting an outcome [14], or generating textual explanations [15]. One might think such explanations are always interpretable for humans as they rely on their ability to understand and reason on natural language. However, they may sometimes be too complex [16, 17], or their structure may not be intuitive enough for a human interpreter to promptly understand the model's behaviour. For example, saliency map scores assigned to the words of a sentence in a sentiment analysis task may not be fully understood, as these scores represent which parts of the input were deemed important and do not represent the actual model's reasoning [18]. Similar representations (i.e., highlights and textual explanations) are employed when collecting human knowledge to train or improve models and evaluate their explanations [19] as these are pretty simple for humans to describe. Despite the practicality of collecting human rationale abiding by these representations, they may not fully explain the actual human reasoning applied to perform the NLP task at hand or the intrinsic logic humans use when reasoning on the texts involved.

Striving to provide an even more complete representation of human rationale for a set of NLP tasks of interest (i.e., Sentiment Analysis, Text Summarization, Natural Language Inference, Claim Verification, and Question Answering), we propose ad hoc formalizations to structure human knowledge defined by drawing inspiration from Argumentation Mining [20, 21] and the recent literature in Data Structuring in XAI. These are referred to as *rationale mappings*. Since we identified them as fundamental discerning factors, we analyzed and organized the tasks based on their type (i.e., text classification or generation) and the number of inputs (i.e., single or multiple inputs). We inspected the processes and the nature of the considered NLP tasks from both the human and model perspective and designed the formalizations. A standard structure is described and then characterized for each of the considered tasks to reduce complexity and enhance expressiveness when possible. These are further hierarchically organized in tree structures, referred to as *rationale trees*. Such representations organize the reasoning steps humans apply when identifying and reasoning on the essential parts of the texts they are provided with. The process applied to collect such structures and an example are described for each of the considered tasks. In the end, a characterization of how these structures will be stored is also provided. To the best of the authors' knowledge, we are the first to provide ad hoc human knowledge formalizations applied to various NLP tasks. In summary, we answer the following research questions

- **RQ1**: How can we structure human knowledge to represent the analytical reasoning steps humans apply to NLP tasks?

- **RQ2**: Can *rationale mappings* be further organized to provide an even more comprehensive and detailed representation of the rationale they describe?

The remainder of the paper is structured as follows. Chapter 2 details the literature on collecting, structuring, and employing human knowledge in XAI and argumentation mining. Chapter 3 classifies the NLP tasks of interest based on their features and describe the structure of *rationale mappings*, the way they are organized into *rationale trees*, and their characterization for each task. Finally, Chapter 4 summarizes the article's content and provide insights about future works and developments.

2. Related Works & Background

2.1. Human Knowledge and Reasoning in NLP and XAI

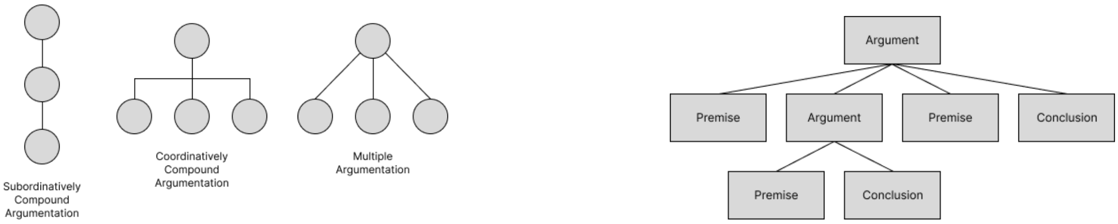
Natural Language Processing (NLP) is a research field aimed at interpreting, analyzing, and manipulating natural language data to learn, understand and produce human language content [22, 23]. NLP include a broad variety of tasks, some aimed at human language understanding (e.g., Coreference Resolution, Natural Language Parsing, etc.), while more complex tasks focus on classifying (e.g., Sentiment Analysis, Natural Language Inference, etc.) or generating (e.g., Text Summarization, Question Answering, etc.) text. In language-based tasks, humans can reason on and understand the provided text(s) through their inherent linguistic knowledge to generate a desired output. Such data are usually collected as couples of input-output texts and employed to train models capable of achieving specific tasks [24]. However, such a data collection approach does not include how humans performed the task and reasoned on the provided text(s). In particular, whenever model explainability is desired, human actors are also requested to provide a description or evidence of the thought process they applied. These are usually collected as free text or highlights of the input's words and sentences [12]. In particular, while the first is more expressive and readable, the latter provides a compact, sufficient, and comprehensive representation [19]. Such information can be used to train so-called self-explainable models [25], i.e., models capable of providing explanations for their outputs [19], or to assess the explanations extracted through other XAI techniques [26, 19]. Even though explanations are mainly collected through crowdsourcing approaches using the aforementioned representations, a wider variety of formats is available whenever an explanation is provided to a human interpreter. In the context of Explainable AI, NLP tasks' explanations are represented as saliency maps [27], declarative representations (i.e., trees and rules) [28], examples [29], or machine-generated natural language [30]. While lay users can easily understand the latter [16], the others may not be directly understandable to human interpreters since a deep understanding of XAI may be required [16, 17]. Although the similarities in the shape explanations are provided by and provided to humans, there's a significant gap when it comes to their interpretability. Furthermore, although some explanations may be humanly understandable, they are not structured to match human reasoning. Hence, a misalignment between how humans think when performing a task involving natural language and how explanations provide evidence for the model's reasoning can be identified.

2.2. Data Structuring in NLP and XAI

Over the last few years, various datasets organized human knowledge applied to the explainability of NLP tasks in the form of free-text [31], highlights of the most important words or sentences [32, 33], or a combination of both [34]. Although such simple structures were proven effective, researchers demonstrated that an enhanced level of detail also contributes to improving models' performances [35, 36] and understandability [37]. Most structures have been designed in the context of Question Answering as it is one of the most complex NLP tasks. Lamm et al. [37] defined annotation triples for Question Answering tasks by identifying relationships between the question and the provided passage. The annotator selects the passage entailing the answer, then chooses a short text span with the answer within the entailed text and marks the equivalent noun phrases in the question and the answer. Finally, entailment patterns are extracted. In the context of machine reading comprehension, Ye et al. [36] defined quadruples of question, paragraph, answer, and a textual explanation that motivates the human reasoning applied to build the annotation. WorldTree [38] and WorldTree V2 [39] are explanation graphs that motivate answers to science questions. They are built by defining and labelling relationships between words in the question, answer, and explanations generated through domain and world knowledge. Although the described processes and structures significantly advance the state-of-the-art in question answering in the corresponding contexts, these are task-specific and their aligned with human reasoning has yet to be proven.

2.3. Argumentation Mining

Argumentation Mining is the process of detecting arguments in a textual document, their relationships, and their internal structure [21, 40]. The basic argumentation unit is an *argument* whose structure involves implicit or explicit premises and a conclusion or, more generally, a set of at least two propositions [20]. For each *argument*, a schema defining relations between prepositions following human reasoning patterns is defined. In particular, Pragma-Dialects theory [41] describes argumentation structures that represent the relation between arguments through coordination, subordination, or forming multiple arguments, as depicted in Figure 1(a).



(a) Pragma-Dialect Theory

(b) Argumentation Tree Example

Figure 1: (a) The argumentation structures described by the Pragma-Dialect theory. (b) An example of the structure of an argumentation tree proposed by Mochales and Moens [20]. Each argument is supported by one or more premises and a conclusion. Furthermore, arguments can be premises for other arguments.

While a more complex graph structure is usually employed [40], Mochales and Moens [20] applied the Pragma-Dialects theory to define a tree-structure representation in which every tree and sub-tree represents a single argumentation structure. In such a setting, all arguments are uniquely related to another argument of a tree for which they represent a premise. An example of such a structure is represented in Figure 1(b).

3. Formalization

3.1. NLP Task Classification

This article considers five different Natural Language Processing tasks: Sentiment Analysis, Text Summarization, Natural Language Inference, Claim Verification, and Question Answering. We identified which features make these tasks substantially different (e.g., objective, process, number of inputs, type of output, etc.). Considering such differences, our research acknowledged the similarity in the nature of the inputs (i.e., all these tasks accept free-text inputs) and the number of outputs (i.e., all these tasks accept a single output) while identifying a significant difference in the process, the type of task (i.e., whether the task generates or classifies text), and the number of inputs (i.e., whether the task handles one or multiple inputs). While the process is unique for each considered NLP task, the type and number of inputs can be used to categorize them. Table 1 reports the outcome of this classification.

Task	Task Type	N Inputs	Input(s) Type	Output Type
Sentiment Analysis	Classification	Single	Free-text	Discrete
Text Summarization	Generation	Single	Free-text	Free-Text
Natural Language Inference	Classification	Multiple	Free-text	Discrete
Claim Verification	Classification	Multiple	Free-text	Discrete
Question Answering	Generation	Multiple	Free-text	Free-text

Table 1
A tabular representation classifying each NLP task of interest based on the features.

3.2. RQ1 - Rationale Mappings

When reasoning on text, humans are so used to finding logical, syntactical, and semantical connections between words that they are unaware of such behaviour. A simple example is the capability of humans to find all the expressions that refer to the same entity in a text (so-called Coreference Resolution in NLP). Such a task rarely requires complex human reasoning as it can be promptly achieved thanks to the linguistic flexibility and knowledge we have developed. On the other hand, extensive human reasoning may be necessary for complex language-based activities, like question answering, in which a human interpreter must understand the paragraph, the question, and the relations between their content, to answer it. We consider such reasoning fundamental building blocks to define and structure human rationale in Natural Language Processing tasks. We refer to them as *rationale mappings*, i.e., representations that organize humans' analytical reasoning steps when identifying and associating the essential parts of the texts involved in a language-based task leading to its output. In particular, we characterise three types of mappings common to the considered NLP tasks:

- *External mappings* represent the reasoning a human interpreter applies between two terms and/or parts of text belonging to **different** texts.
- *Internal mappings* represent the reasoning a human interpreter applies between two different terms and/or parts of text in the **same** text.
- *Resolution mappings* are *internal mappings* representing anaphora or coreference resolution reasoning between two terms and/or parts of text in the **same** text.

We define the structure of *rationale mappings* by combining the literature about data structuring in XAI and the argument structure described by Mochales and Moens [20]. In our definition, we constrained the number of propositions and extended it with their relationship, finally merging them into a single representation. Hence, we define *rationale mappings* as triples

$$\langle \quad text,\ text,\ label \quad \rangle$$

where *text* is a word or a set of consecutive words from any text involved in the human reasoning applied to the language-based task and *label* is a term that defines the relationship between the *texts*. The latter is defined based on the type of mapping. In *external mappings*, they are specific to the NLP task to which the mappings are applied, i.e., when the task involves a discrete output (i.e., a finite and well-defined set of outputs is possible) or specific terms that describe the applied approach, these are employed as labels as they represent both human- and model-understandable concepts. Otherwise, more generic linguistic labels are considered, i.e., *semantic* or *syntactic*, respectfully representing the semantic or syntactic similarity between texts. Whenever a *semantic* label is applied, mappings can be extended to include a textual description of the rationale a human interpreter applies, enhancing the level of detail. Such generic labels are also applied to *internal mappings* as they define a syntactic or semantic relationship between the texts.

External mappings may be simplified in specific cases, hence defining these *mappings* as couples

$$\langle \quad text,\ label \quad \rangle$$

where *text* is a word or a set of consecutive words from any text involved in the human reasoning applied to the language-based task and *label* is a term that specifies the *text*. In particular, we consider simplifications only when the nature of the task and the labels allow for them. The fact that the two texts in a mapping coincide is not considered a simplification, even though it might be helpful for what concerns data storage. *Internal mappings* are not subject to any simplifications as there is no meaningful overlap between *texts* and *label*. However, they may be subject to slight changes to improve their expressiveness when applied to specific tasks. *Resolution mappings* can't be simplified as it is necessary to specify the type of resolution used and the parts of text involved. Further clarifications will be made for each considered NLP task in their dedicated sections.

3.3. RQ2 - Rationale Trees

While defining such mappings is useful to understand the human reasoning applied to a task, these can be further hierarchically structured to describe better the rationale involved in a specific instance of a task. Hence, mappings are organized in a tree structure in which each *rationale mapping* is a tree node with different meanings and constraints based on its type. We refer to these structures as *rationale trees*. The root node represents the (generic) relationship between the input(s) and the output, i.e., a standard input-output representation of the task. Each other node (i.e., internal nodes and leaves) details the mapping between the texts in its parent node. In particular, considering a parent node *p* and its child node *c* defined as

p ⟨ *p_text_I, p_text_II, p_label* ⟩
c ⟨ *c_text_I, c_text_II, c_label* ⟩

and assuming that their corresponding *texts* (i.e., *p_text_I* and *c_text_I*, and *p_text_II* and *c_text_II*) are extracted from the same text, either one of the following constraints is enforced.

- *c_text_I* ⊂ ***p_text_I***, i.e., *c_text_I* is a word or a set of consecutive words that are a subset of *p_text_I*, **and** *c_text_II* ⊂ ***p_text_II***, i.e., *c_text_II* is a word or a set of consecutive words that are a subset of *p_text_II*.
- *c_text_I* ⊂ ***p_text_I***, i.e., *c_text_I* is a word or a set of consecutive words that are a subset of *p_text_I*, **and** *c_text_II* ⊂ ***p_text_I***, i.e., *c_text_II* is a word or a set of consecutive words that are a subset of *p_text_I*. The same can be applied considering *p_text_II*.

Such conditions define a structure in which the deeper the node, the more specific the rationale it describes. Furthermore, while *external* and *internal mappings* can either be internal nodes or leaves that detail the parent node's rationale, *resolution mappings* can only be leaf nodes and define rationale to be applied to their parent and sibling nodes whenever meaningful. Child nodes are considered to be in a coordinative relationship towards their parent node, simultaneously contributing to specifying the parent's node mapping. Moreover, while *external mappings* can have both *internal* and *external mappings* as child nodes, *internal mappings* can only have other *internal mappings* as child nodes since they are mainly employed to detail the rationale applied in *external mappings* and not vice-versa. Additionally, *resolution mappings* can be child nodes for both *internal* and *external mappings*. A generic example of a *rationale tree* is depicted in Figure 2.

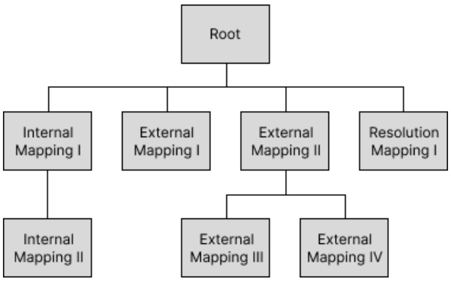


Figure 2: An example of a generic *rationale tree* organizing the mappings abiding by the described rules.

The only condition enforced between sibling nodes is that their *text_I* and *text_II* should not completely overlap simultaneously, i.e., considering any pair of sibling nodes *s1* and *s2* defined as

s1 ⟨ *s1_text_I, s1_text_II, s1_label* ⟩
s2 ⟨ *s2_text_I, s2_text_II, s2_label* ⟩

Task	Labels	Simplification
Sentiment Analysis	Positive, Negative	Yes
Text Summarization	Extractive, Abstractive	Yes
Natural Language Inference	Neutral, Contradiction, Entailment	No
Claim Verification	Support, Refute	No
Question Answering	Syntactic, Semantic	No

Table 2
A tabular representation summarizing some of the features of each NLP task of interest.

and assuming that their corresponding *texts* (e.g., *s1_text_I* and *s2_text_I*, and *s1_text_II* and *s2_text_II*) are extracted from the same text, the following constraints are enforced.

- if $s1\_text\_I \equiv s2\_text\_I \Rightarrow s1\_text\_II \neq s2\_text\_II$.
- if $s1\_text\_II \equiv s2\_text\_II \Rightarrow s1\_text\_I \neq s2\_text\_I$.

Such conditions define a structure where the same mapping can't be duplicated, although they still allow a fine granularity in the differences between the mappings associated with a parent node.
The following sections describe the formalizations, detailing a set of features of interest. In particular, the simplest ones are summarized in Table 1, while the most complex ones are detailed in the corresponding sections, summarized in Table 2, and explained below.

- **Labels**, i.e., the concepts applied as labels when defining the mappings. These are mainly employed in *external mappings*, although *internal mappings* may sometimes benefit from such labels.
- **Mappings Interpretation**, i.e., a task-specific description for *internal* and *external mappings*, if needed. Whenever no specific interpretation is provided, we consider them aligned with their general description.
- **Simplifications**, i.e., whether any simplification can be applied to the mappings, their description and structure.
- **Mapping Guidelines**, i.e., the process a human interpreter applies to define *mappings* and a *rationale tree* for a task of interest. We consider human interpreters to be performing the task themselves, although we do not include details of such a process in the guidelines. The same approach can be applied even when the interpreter is provided with all the texts involved in the task.
- **Example**, i.e., an example of a *rationale tree* collected by applying the process to a task entry from a specified NLP dataset. Each mapping type is identified by its starting letters (e.g., EM stands for *external mapping*). An example is provided for Sentiment Analysis, while all the others can be found in Appendix A as they follow the same principles.

3.4. Sentiment Analysis

Sentiment Analysis is an NLP task in which a human interpreter defines the output by assigning a sentiment (either *positive*, *negative*, or sometimes *neutral*) to an input sentence or text. For such a task, *internal mappings* are defined as

$$\langle \;\; \textit{input\_text}, \textit{input\_text}, \textit{label} \;\; \rangle$$

where *input_text* is a word or a set of consecutive words from the input text and *label* is the sentiment between the *texts* (i.e., *positive* or *negative*, in our case). On the other hand, *external mappings* are defined as

$$\langle \;\; \textit{input\_text}, \textit{output\_text}, \textit{label} \;\; \rangle$$

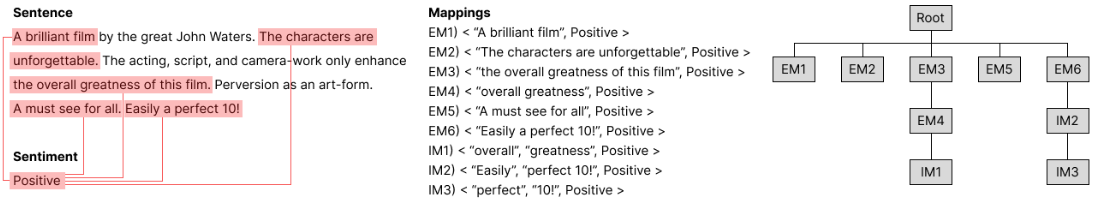
where *input_text* is a word or a set of consecutive words from the input text, and *output_text* and *label* represent the sentiment associated with the *input_text*. Acknowledged the overlapping between *output_text* and *label*, *external mappings* are applied a simplification. Hence, they are defined as

$$\langle \;\; \textit{input\_text}, \textit{label} \;\; \rangle$$

where *input_text* is a word or a set of consecutive words from the input text and *label* is the sentiment associated with *input_text*.
A human interpreter providing *rationale mappings* for a Sentiment Analysis task performs the following assignments.

- They define *external mappings* between the input and the output texts.
- For each of the previously defined *external mapping*, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same process is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define any *resolution mapping* that was applied, as child nodes or sibling nodes based on where they are applied.

We picked a data point from the Large Movie Review Dataset [42] and built its *rationale tree* as an example, represented in Figure 3.


Sentence

A brilliant film by the great John Waters. The characters are unforgettable. The acting, script, and camera-work only enhance the overall greatness of this film. Perversion as an art-form. A must see for all. Easily a perfect 10!

Sentiment

Positive

Mappings

EM1) < "A brilliant film", Positive >
EM2) < "The characters are unforgettable", Positive >
EM3) < "the overall greatness of this film", Positive >
EM4) < "overall greatness", Positive >
EM5) < "A must see for all", Positive >
EM6) < "Easily a perfect 10!", Positive >
IM1) < "overall", "greatness", Positive >
IM2) < "Easily", "perfect 10!", Positive >
IM3) < "perfect", "10!", Positive >

Figure 3: An example of a *rationale tree* organizing the mappings of a data point from the Large Movie Review Dataset. Internal mappings were omitted from the initial representation for clarity purposes.

3.5. Text Summarization

Text Summarization is an NLP task in which a human interpreter is provided with an input text and they provide a summarized output text. Two different approaches can be applied and combined together. An *extractive* approach reports parts of the input text into the output text, maintaining the same syntax. An *abstractive* approach formulates the output text to have the same semantics as parts of the input text while using a different syntax. For such a task, *external mappings* are detailed as

$$\langle \quad \textit{input\_text}, \textit{output\_text}, \textit{label} \quad \rangle$$

where *input_text* is a word or a set of consecutive words from the input text, *output_text* is a word or a set of consecutive words from the output text, and *label* is the summarization approach (i.e., *abstractive* or *extractive*) applied to *input_text* to generate *output_text*.
Whenever an extractive approach is applied, *external mappings* can be simplified as such an approach involves reporting the same text from the input in the output text. Hence, they are defined as couples

$$\langle \quad \textit{input\_text}, \text{“}\textit{extractive}\text{”} \quad \rangle$$

where *input_text* is a word or a set of consecutive words from the input text.
A human interpreter providing *rationale mappings* for a Text Summarization task performs the following assignments.

- They define *external mappings* between the input and the output texts.
- For each of the previously defined *external mapping* that is assigned the *“extractive”* label, they recursively define *internal mappings* detailing the texts involved until a desired level of detail is achieved. Instead, for each of the previously described *external mapping* that is assigned the *“abstractive”* label, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same approach is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define any *resolution mapping* that was applied, as child nodes or sibling nodes based on where they are applied.

We picked a data point from the CNN/Daily Mail Dataset [43] and built its *rationale tree* as an example, represented in Figure 6 in Appendix A.

3.6. Natural Language Inference

Natural Language Inference is an NLP task in which a human interpreter is provided with two texts, an *hypothesis* and a *premise*, and they define whether they are in an *entailment*, *contradiction*, or *neutral* relationship. For such a task, *external mappings* are defined as

$$\langle \quad \textit{premise\_text}, \textit{hypothesis\_text}, \textit{label} \quad \rangle$$

where *premise_text* is a word or a set of consecutive words from the premise, *hypothesis_text* is a word or a set of consecutive words from the hypothesis, and *label* is the relationship (i.e., *entailment*, *contradiction*, or *neutral*) between *premise_text* and *hypothesis_text*.
A human interpreter providing *rationale mappings* for a Natural Language Inference task performs the following assignments.

- They identify *external mappings* between the premise and the hypothesis texts.
- For each of the previously defined *external mappings*, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same approach is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define whether any *resolution mapping* that was applied, as child nodes or sibling nodes based on where they are applied.

We picked a data point from the e-SNLI Dataset [34] and built its *rationale tree* as an example, represented in Figure 4 in Appendix A.

3.7. Claim Verification

Claim Verification is an NLP task in which a human interpreter is provided with two texts, i.e., a *claim* and an *evidence*, and they define whether the *evidence supports* or *refutes* the *claim*. For such a task, *external mappings* are defined as

$$\langle \quad claim\_text, \; evidence\_text, \; label \quad \rangle$$

where *claim_text* is a word or a set of consecutive words from the claim, *evidence_text* is a word or a set of consecutive words from the evidence, and *label* is the relationship (i.e., *support* or *refute*) between *claim_text* and *evidence_text*.
A human interpreter providing *rationale mappings* for a Claim Verification task performs the following assignments.

- They identify *external mappings* between the claim and the evidence.
- For each of the previously defined *external mappings*, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same approach is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define any *resolution mapping* that was applied, as child nodes or sibling nodes based on where they are applied.

We picked a data point from the FEVER Dataset [44] and built its *rationale tree* as an example, represented in Figure 5 in Appendix A.

3.8. Question Answering

Question Answering is an NLP task in which a human interpreter is provided with a *question* and a *paragraph*, and they provide an *answer* to the *question* through the *paragraph*. For such a task, *mappings* are defined as

$$\langle \quad text, \; text, \; label \quad \rangle$$

where *text* is a word or a set of consecutive words from the same (in *internal mappings*) or different (in *external mappings*) texts, i.e., the *question*, the *paragraph*, or the *answer*, and *label* describes whether there's a *semantic* or *syntactic* relationship between the *texts*. Similarly to *internal mappings*, whenever a *semantic* label is applied, the mapping is further detailed by collecting comments detailing the relationship between the *texts*.
Rationale trees increase in complexity in Question Answering tasks as the process is more convoluted than the other considered NLP tasks. First of all, a new type of mapping has to be defined. *Abstractive mappings* define which word or set of consecutive words of the question contributed to defining its class among the following question types.

- Yes/No Question, i.e., questions looking for confirmation in the paragraph.
- Wh-Question, i.e., questions looking for the answer based on the type of wh-question (e.g., Who, What, etc.).

- Choice Question, i.e., questions picking the answer among the ones proposed in the question based on the paragraph.
- Disjunctive Questions, i.e., questions looking for confirmation in the paragraph.

Such mappings are introduced to be aligned with the question-answering process in which a human interpreter identifies which information they should look for to answer the question before reading the paragraph [45, 46, 47]. *Abstractive mappings* are defined as couples

$$\langle \quad question\_text, question\_class \quad \rangle$$

where *question_text* is a word or a set of consecutive words from the question and the *question_class* describes the question class chosen from a list of values defined from the question types described, i.e., *yes/no question*, *disjunctive question*, *choice question*, and *wh-question.* The latter is further detailed based on the type of wh-question, defining a *specialization* (described in Table 3 in Appendix A). Moreover, each *rationale tree* can only have one *abstractive mapping* and must be a child node of the root node.

A human interpreter providing *rationale mappings* for a Question Answering task performs the following assignments.

- They define an *abstractive mapping* associated with the question.
- They define *external mappings* between the question and the paragraph. The same is done for *internal* and *resolution mappings* in these texts. These are recursively refined until the desired level of detail is achieved.
- They define *external mappings* between the paragraph and the answer. The same is done for *internal* and *resolution mappings* in these texts. These are recursively refined until the desired level of detail is achieved.
- They detail the previously defined *abstractive mapping* by defining *external mappings* between the question and the answer. These are recursively refined until the desired level of detail is achieved.

We picked a data point from the SQuAD 2.0 Dataset [48] and built its *rationale tree* as an example, represented in Figure 7 in Appendix A.

4. Conclusions

This article described a novel approach to structuring human knowledge for a set of Natural Language Processing tasks of interest. We reported on the literature about human knowledge and data structuring in NLP and XAI and Argumentation Mining that extensively inspired our work. We explained the concept of *rationale mapping*, its specializations, and how these can be structured into *rationale trees* to describe the reasoning process a human interpreter applies in language-based tasks. Task-specific mappings, potential simplifications, and extensions were detailed for each one. We argue these representations contribute towards representing human knowledge to be applied to XAI tasks while also being a suitable way of shaping explanations provided by XAI methods or self-explaining models (e.g., LLMs). Future work will involve collecting and assessing the understandability of *rationale trees* for datasets of interest, improving and detailing the labels applied to some of the proposed mappings, and exploring the applicability of *rationale trees* to other NLP tasks.

References

[1] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, D. Yang, Is chatgpt a general-purpose natural language processing task solver?, 2023. URL: https://arxiv.org/abs/2302.06476. doi:10.48550/ARXIV.2302.06476.

[2] W. Jiao, W. Wang, J.-t. Huang, X. Wang, Z. Tu, Is chatgpt a good translator? a preliminary study, 2023. URL: https://arxiv.org/abs/2301.08745. doi:10.48550/ARXIV.2301.08745.

[3] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. URL: https://arxiv.org/abs/2302.04023. doi:10.48550/ARXIV.2302.04023.

[4] X. Yang, Y. Li, X. Zhang, H. Chen, W. Cheng, Exploring the limits of chatgpt for query or aspect-based text summarization, 2023. URL: https://arxiv.org/abs/2302.08081. doi:10.48550/ARXIV.2302.08081.

[5] B. Zhang, D. Ding, L. Jing, How would stance detection techniques evolve after the launch of chatgpt?, 2022. URL: https://arxiv.org/abs/2212.14548. doi:10.48550/ARXIV.2212.14548.

[6] W. Samek, K.-R. Müller, Towards Explainable Artificial Intelligence, Springer International Publishing, Cham, 2019, pp. 5–22. URL: https://doi.org/10.1007/978-3-030-28954-6_1. doi:10.1007/978-3-030-28954-6_1.

[7] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable ai: The new 42?, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2018, pp. 295–303.

[8] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, 2020. URL: https://arxiv.org/abs/2006.00093. doi:10.48550/ARXIV.2006.00093.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models, 2018. URL: https://arxiv.org/abs/1802.01933. doi:10.48550/ARXIV.1802.01933.

[10] Q. V. Liao, K. R. Varshney, Human-centered explainable ai (xai): From algorithms to user experiences, 2021. URL: https://arxiv.org/abs/2110.10790. doi:10.48550/ARXIV.2110.10790.

[11] U. Ehsan, P. Wintersberger, Q. V. Liao, M. Mara, M. Streit, S. Wachter, A. Riener, M. O. Riedl, Operationalizing human-centered perspectives in explainable ai, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411763.3441342. doi:10.1145/3411763.3441342.

[12] A. Tocchetti, M. Brambilla, The role of human knowledge in explainable ai, Data 7 (2022). URL: https://www.mdpi.com/2306-5729/7/7/93. doi:10.3390/data7070093.

[13] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, 2016. URL: https://arxiv.org/abs/1606.04155. doi:10.48550/ARXIV.1606.04155.

[14] X. Han, B. C. Wallace, Y. Tsvetkov, Explaining black box predictions and unveiling data artifacts through influence functions, 2020. URL: https://arxiv.org/abs/2005.06676. doi:10.48550/ARXIV.2005.06676.

[15] I. Ampomah, J. Burton, A. Enshaei, N. Al Moubayed, Generating textual explanations

for machine learning models performance: A table-to-text task, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3542–3551. URL: https://aclanthology.org/2022.lrec-1.379.

[16] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.

[17] N. Feldhus, L. Hennig, M. D. Nasert, C. Ebert, R. Schwarzenberg, S. Moller, Constructing natural language explanations via saliency map verbalization, ArXiv abs/2210.07222 (2022).

[18] H. Schuff, A. Jacovi, H. Adel, Y. Goldberg, N. T. Vu, Human interpretation of saliency-based explanation over text, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22, Association for Computing Machinery, New York, NY, USA, 2022, p. 611–636. URL: https://doi.org/10.1145/3531146.3533127. doi:10.1145/3531146.3533127.

[19] S. Wiegreffe, A. Marasović, Teach me to explain: A review of datasets for explainable natural language processing, 2021. URL: https://arxiv.org/abs/2102.12060. doi:10.48550/ARXIV.2102.12060.

[20] R. Mochales, M.-F. Moens, Argumentation mining, Artificial Intelligence and Law 19 (2011) 1–22. doi:10.1007/s10506-010-9104-x.

[21] R. M. Palau, M.-F. Moens, Argumentation mining: The detection, classification and structure of arguments in text, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL ’09, Association for Computing Machinery, New York, NY, USA, 2009, p. 98–107. URL: https://doi.org/10.1145/1568234.1568246. doi:10.1145/1568234.1568246.

[22] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimedia Tools and Applications 82 (2023) 3713–3744. URL: https://doi.org/10.1007/s11042-022-13428-4. doi:10.1007/s11042-022-13428-4.

[23] J. Hirschberg, C. D. Manning, Advances in natural language processing, Science 349 (2015) 261–266. URL: https://www.science.org/doi/abs/10.1126/science.aaa8685. doi:10.1126/science.aaa8685. arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa8685.

[24] M. Sabou, K. Bontcheva, A. Scharl, Crowdsourcing research opportunities: Lessons from natural language processing, in: Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, i-KNOW ’12, Association for Computing Machinery, New York, NY, USA, 2012. URL: https://doi.org/10.1145/2362456.2362479. doi:10.1145/2362456.2362479.

[25] D. Alvarez-Melis, T. S. Jaakkola, Towards robust interpretability with self-explaining neural networks, 2018. arXiv:1806.07538.

[26] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Lin-

guistics, Online, 2020, pp. 3256–3274. URL: https://aclanthology.org/2020.emnlp-main.263. doi:10.18653/v1/2020.emnlp-main.263.

[27] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv:1312.6034.

[28] N. Pröllochs, S. Feuerriegel, D. Neumann, Learning interpretable negation rules via weak supervision at document level: A reinforcement learning approach, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 407–413. URL: https://aclanthology.org/N19-1038. doi:10.18653/v1/N19-1038.

[29] H. Guo, N. Rajani, P. Hase, M. Bansal, C. Xiong, FastIF: Scalable influence functions for efficient model interpretation and debugging, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10333–10350. URL: https://aclanthology.org/2021.emnlp-main.808. doi:10.18653/v1/2021.emnlp-main.808.

[30] W. Ling, D. Yogatama, C. Dyer, P. Blunsom, Program induction by rationale generation: Learning to solve and explain algebraic word problems, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 158–167. URL: https://aclanthology.org/P17-1015. doi:10.18653/v1/P17-1015.

[31] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7740–7754. URL: https://aclanthology.org/2020.emnlp-main.623. doi:10.18653/v1/2020.emnlp-main.623.

[32] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 1631–1642. URL: https://aclanthology.org/D13-1170.

[33] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, Eraser: A benchmark to evaluate rationalized nlp models, 2019. URL: https://arxiv.org/abs/1911.03429. doi:10.48550/ARXIV.1911.03429.

[34] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, 2018. URL: https://arxiv.org/abs/1812.01193. doi:10.48550/ARXIV.1812.01193.

[35] T. Khot, P. Clark, M. Guerquin, P. Jansen, A. Sabharwal, Qasc: A dataset for question answering via sentence composition, 2020. arXiv:1910.11473.

[36] Q. Ye, X. Huang, E. Boschee, X. Ren, Teaching machine comprehension with compositional explanations, 2020. arXiv:2005.00806.

[37] M. Lamm, J. Palomaki, C. Alberti, D. Andor, E. Choi, L. B. Soares, M. Collins, Qed: A framework and dataset for explanations in question answering, 2020. arXiv:2009.06354.

[38] P. A. Jansen, E. Wainwright, S. Marmorstein, C. T. Morrison, Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference, 2018.

arXiv:1802.03052.

[39] Z. Xie, S. Thiem, J. Martin, E. Wainwright, S. Marmorstein, P. Jansen, WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5456–5473. URL: https://aclanthology.org/2020.lrec-1.671.

[40] L. Carstens, F. Toni, Towards relation based argumentation mining, in: Proceedings of the 2nd Workshop on Argumentation Mining, Association for Computational Linguistics, Denver, CO, 2015, pp. 29–34. URL: https://aclanthology.org/W15-0504. doi:10.3115/v1/W15-0504.

[41] F. van Eemeren, R. Grootendorst, A systematic theory of argumentation: The pragma-dialectical approach (2003). doi:10.1017/CBO9780511616389.

[42] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015.

[43] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, B. Xiang, Abstractive text summarization using sequence-to-sequence RNNs and beyond, in: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 280–290. URL: https://aclanthology.org/K16-1028. doi:10.18653/v1/K16-1028.

[44] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[45] E. Riloff, M. Thelen, A rule-based question answering system for reading comprehension tests, in: Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems - Volume 6, ANLP/NAACL-ReadingComp '00, Association for Computational Linguistics, USA, 2000, p. 13–19. URL: https://doi.org/10.3115/1117595.1117598. doi:10.3115/1117595.1117598.

[46] M. A. Calijorne Soares, F. S. Parreiras, A literature review on question answering techniques, paradigms and systems, Journal of King Saud University - Computer and Information Sciences 32 (2020) 635–646. URL: https://www.sciencedirect.com/science/article/pii/S131915781830082X. doi:https://doi.org/10.1016/j.jksuci.2018.08.005.

[47] P. Katyayan, N. Joshi, Design and development of rule-based open-domain question-answering system on squad v2.0 dataset, 2022. arXiv:2204.09659.

[48] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for squad, 2018. arXiv:1806.03822.

A. Appendix

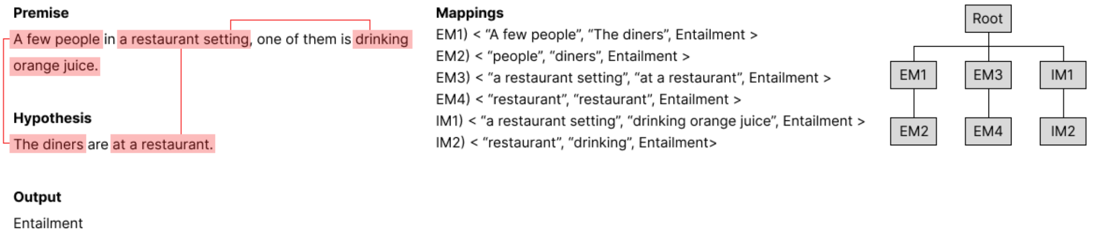

Premise
A few people in a restaurant setting, one of them is drinking orange juice.
Hypothesis
The diners are at a restaurant.
Output
Entailment
Mappings
EM1) < "A few people", "The diners", Entailment >
EM2) < "people", "diners", Entailment >
EM3) < "a restaurant setting", "at a restaurant", Entailment >
EM4) < "restaurant", "restaurant", Entailment >
IM1) < "a restaurant setting", "drinking orange juice", Entailment >
IM2) < "restaurant", "drinking", Entailment>
Root
EM1
EM3
IM1
EM2
EM4
IM2

Figure 4: An example of a *rationale tree* organizing the mappings of a data point from the e-SNLI Dataset.

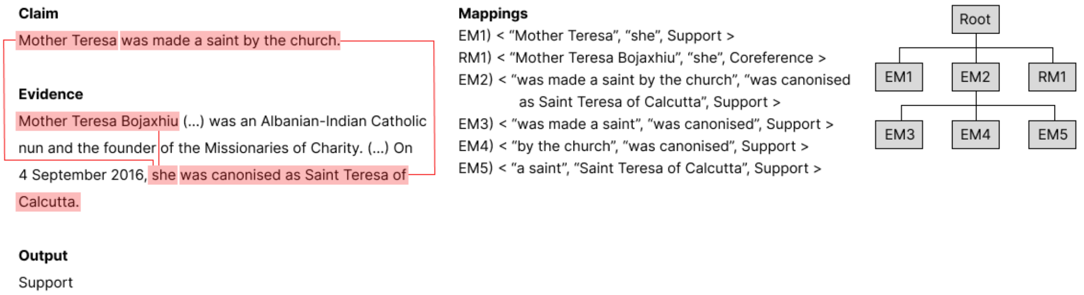

Claim
Mother Teresa was made a saint by the church.
Evidence
Mother Teresa Bojaxhiu (...) was an Albanian-Indian Catholic nun and the founder of the Missionaries of Charity. (...) On 4 September 2016, she was canonised as Saint Teresa of Calcutta.
Output
Support
Mappings
EM1) < "Mother Teresa", "she", Support >
RM1) < "Mother Teresa Bojaxhiu", "she", Coreference >
EM2) < "was made a saint by the church", "was canonised as Saint Teresa of Calcutta", Support >
EM3) < "was made a saint", "was canonised", Support >
EM4) < "by the church", "was canonised", Support >
EM5) < "a saint", "Saint Teresa of Calcutta", Support >
Root
EM1
EM2
RM1
EM3
EM4
EM5

Figure 5: An example of a *rationale tree* organizing the mappings of a data point from the FEVER Dataset. We included the evidence defined in the dataset and collected from Wikipedia, removing the text that wasn't deemed useful for clarity purposes.

Specialization	**Wh-Question Keywords**
Person	Who, Whose, Whom
Information	What, How
Location	Where
Time	When
Reason	Why, What for, How come, Why don't
Quantity	How many, How much, How far, How long, etc.
Choice	Which, Whom

Table 3
A table summarizing the specializations for the class of wh-questions. For each specialization, a list of keywords identifying the wh-question is provided.

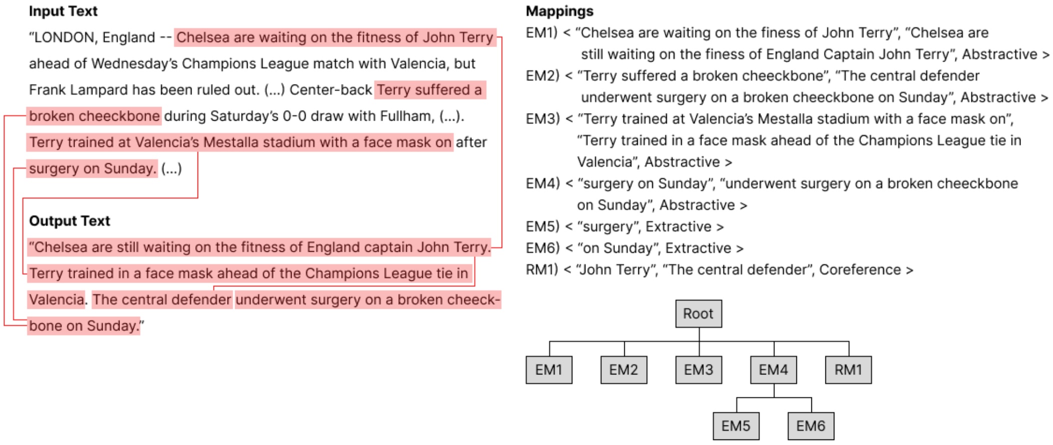

Input Text
"LONDON, England -- Chelsea are waiting on the fitness of John Terry ahead of Wednesday's Champions League match with Valencia, but Frank Lampard has been ruled out. (...) Center-back Terry suffered a broken cheeckbone during Saturday's 0-0 draw with Fullham, (...). Terry trained at Valencia's Mestalla stadium with a face mask on after surgery on Sunday. (...)
Output Text
"Chelsea are still waiting on the fitness of England captain John Terry. Terry trained in a face mask ahead of the Champions League tie in Valencia. The central defender underwent surgery on a broken cheeck-bone on Sunday."
Mappings
EM1) < "Chelsea are waiting on the finess of John Terry", "Chelsea are still waiting on the finess of England Captain John Terry", Abstractive >
EM2) < "Terry suffered a broken cheeckbone", "The central defender underwent surgery on a broken cheeckbone on Sunday", Abstractive >
EM3) < "Terry trained at Valencia's Mestalla stadium with a face mask on", "Terry trained in a face mask ahead of the Champions League tie in Valencia", Abstractive >
EM4) < "surgery on Sunday", "underwent surgery on a broken cheeckbone on Sunday", Abstractive >
EM5) < "surgery", Extractive >
EM6) < "on Sunday", Extractive >
RM1) < "John Terry", "The central defender", Coreference >
Root
EM1
EM2
EM3
EM4
RM1
EM5
EM6

Figure 6: An example of a *rationale tree* organizing the mappings of a data point from the CNN/Daily Mail Dataset Dataset. Although most *external mappings* could be further detailed, only one *external mapping* was refined for clarity purposes. For the same reason, part of the input text that wasn't deemed useful was omitted.

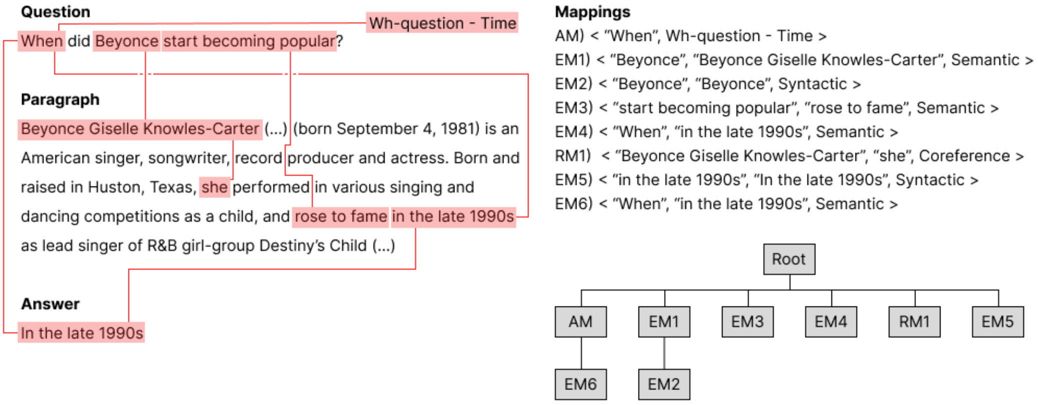

Question
Wh-question - Time
When did Beyonce start becoming popular?
Paragraph
Beyonce Giselle Knowles-Carter (...) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Huston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child (...)
Answer
In the late 1990s
Mappings
AM) < "When", Wh-question - Time >
EM1) < "Beyonce", "Beyonce Giselle Knowles-Carter", Semantic >
EM2) < "Beyonce", "Beyonce", Syntactic >
EM3) < "start becoming popular", "rose to fame", Semantic >
EM4) < "When", "in the late 1990s", Semantic >
RM1) < "Beyonce Giselle Knowles-Carter", "she", Coreference >
EM5) < "in the late 1990s", "In the late 1990s", Syntactic >
EM6) < "When", "in the late 1990s", Semantic >
Root
AM
EM1
EM3
EM4
RM1
EM5
EM6
EM2

Figure 7: An example of a *rationale tree* organizing the mappings of a data point from the SQuAD 2.0 Dataset.