

Towards a Framework for Evaluating Explanations in Automated Fact Verification

Neema Kotonya and Francesca Toni

LREC-COLING 2024

Introduction

- Opacity of deep learning models necessitates explainability.
- Explanation evaluation has not been widely explored.
- How we decide to explain should inform how we evaluate explanations.
- The focus of this paper is explainability for automated fact checking, but what will be discussed is relevant to knowledge-intensive NLP tasks.

Contributions

We conceptualize evaluation of explanations through the prism of explanation style.

We offer the following contributions:

- We define three classes of explanation and offer formal definitions for properties commonly attributed to explanations.
- We argue the case for concrete metrics for evaluating explanations informed by the properties that we define.

Rationalization as Explanation

- Offers explanations at the level of individual predictions.
- Rationales are the input features that are the most salient for the model's prediction (as in Table 1).

claim: "The typical life span of a daffodil exceeds two years."

evidence: "Daffodil is the common name for plants of the narcissus genus, which describes perennial plants [...] A perennial plant has a minimum life span of two years."

label: VERIFIABLE

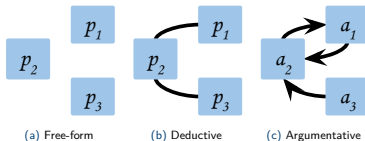
explanation: A daffodil is a member of the narcissus genus, such plants have life spans of at least two years.

Table 1: Example of a rationalizing explanation.

Explanations by Degree of Structure

We consider three classes of explanation:

- Free-form explanations
 - “highlighted” tokens
- Deductive explanations
 - chains of facts
- Argumentative explanations
 - explanations as contrastive evidence with “winning arguments”



Properties of Free-Form Explanations

We define a single property for free-form explanations:

- Coherence

Claim: Glasgow is the capital of Scotland.

Explanation:

p_1 : Glasgow is the capital and also the most populous city in Scotland.

p_2 : Glasgow was founded before Edinburgh, therefore it is Scotland's first and capital city.

p_3 : Glasgow was European Capital of Culture in 1990.

p_4 : Edinburgh is the capital of Scotland and the location of the Scottish Parliament.

Properties of Deductive Explanations

We define the following properties for deductive explanations:

- Non-circularity (only for directed \mathcal{R})
- Non-redundancy
- Relevance
- Full connectivity

p_3 : A daffodil plant can live for more than two years.

Verdict: Verified

Explanation $\langle \mathcal{P}, \mathcal{R} \rangle$, where:

$\mathcal{P} = \{p_1, p_2, p_3\}$, for:

p_1 : Daffodil is the common name for plants of the narcissus genus, which are perennial.

p_2 : A perennial plant has a minimum life span of two years.

$\mathcal{R} = \{(p_1, p_2), (p_2, p_3)\}$.

Properties for Argumentative Explanations

Properties for argumentative explanations:

- 1 Dialectical non-circularity
- 2 Dialectical faithfulness. Inspired by the notion of *dialectical strength* for arguments in an argumentative explanation, in the spirit of (Baroni et al., 2019).
- 3 Acceptability.

Argumentative Explanations - Dialectical Non-circularity

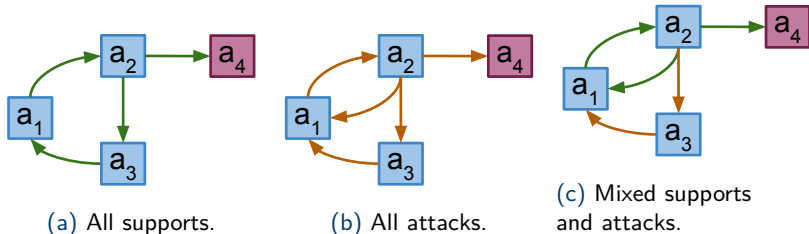


Figure 2: Examples of dialectical circularity for three argumentative explanations. Attacks are shown in orange and supports are shown in green. Argument a_4 with conclusion \hat{y} is purple.

Argumentative Explanations - Dialectical Faithfulness and Acceptability



Figure 3: An illustration of argumentative explanations for top, high, and low confidence (binary) predictions. Attacks are shown in orange and supports are shown in green. Argument a_4 with conclusion \hat{y} is purple. Left-most and middle explanations are acceptable, no argument attacks a_4 , the only argument with conclusion \hat{y} . The right-most explanation is not acceptable, as there is no argument attacking a_4 in the explanation.

Metrics

We define metrics for evaluating each explanation class:

Metric	Explanation	Range
Coherence	Free-form	min: 0, max: 1
Weak relevance	Deductive	min: 0, max: 1
Strong relevance	Deductive	min: 0, max: 1
Non-redundancy	Deductive	min: 1, max: 0
Circularity	Argumentative	min: 1, max: 0
Acceptability	Argumentative	min: 0, max: 1

Conclusion

- Explanations of increasing structure (free-form, deductive, argumentative).
- Properties of explanations
 - Varying the structure of explanations introduces new properties.
- Formalizing properties allows us to derive more intuitive metrics for evaluation.