TriplEx

Triple Extraction for Explanation

Mattia Setzu, Anna Monreale, Pasquale Minervini

Consiglio Nazionale delle Ricerche

TriplEx

- Explain Transformer-based models' predictions on three tasks:
  - Natural Language Inference (NLI)
  - Semantic Text Similarity (STS)
  - Text Classification (TC)
- Mainly aimed at users with no ML expertise
- Based on two staples of text explainability: token explanations and natural language explanations

Natural Language Inference

Premise
Wet brown dog swims towards camera.

Hypothesis
A dog is sleeping in his bed.

Does the premise entail, contradict, or is neutral to the hypothesis?

Example adapted from “Natural Language Inference with a Human Touch: Using Human Explanations to Guide Model Attention”, Stacey et al.

Text explanations: token importance

Premise
Wet brown dog swims towards camera.

Hypothesis
A dog is sleeping in his bed.

Token explanation:

Wet	brown	dog	swims	towards	camera.	A	dog	is	sleeping	in	his	bed.
0.1	0.02	0.2	0.7	0.04	0.01	0.01	0.001	0.04	0.8	0.01	0.01	0.02

Text explanations: token importance

Token explanation:

Wet	brown	dog	swims	towards	camera.	A	dog	is	sleeping	in	his	bed.
0.1	0.02	0.2	0.7	0.04	0.01	0.01	0.001	0.04	0.8	0.01	0.01	0.02

- By design[1, 2, 3, 4]
- Post-hoc[5, 6]

[1] Rationalizing neural predictions, Lei et al.
[2] Explaining Neural Network Predictions on Sentence Pairs via Learning Word-Group Masks, Chen et al.
[3] Explain Yourself! Leveraging Language Models for Commonsense Reasoning, Rajani et al.
[4] e-SNLI: Natural Language Inference with Natural Language Explanations, Camburu et al.
[5] A Unified Approach to Interpreting Model Predictions, Lundberg et al.
[6] Interpretation of NLP models through input marginalization, Kim et al.

Text explanations: token importance

Token explanation:

Wet	brown	dog	swims	towards	camera.	A	dog	is	sleeping	in	his	bed.
0.1	0.02	0.2	0.7	0.04	0.01	0.01	0.001	0.04	0.8	0.01	0.01	0.02

- By design[1, 2, 3, 4]
- Post-hoc[5, 6]

- Incomplete
- Brittle
- Ignore intra-text dependencies

[1] Rationalizing neural predictions, Lei et al.
[2] Explaining Neural Network Predictions on Sentence Pairs via Learning Word-Group Masks, Chen et al.
[3] Explain Yourself! Leveraging Language Models for Commonsense Reasoning, Rajani et al.
[4] e-SNLI: Natural Language Inference with Natural Language Explanations, Camburu et al.
[5] A Unified Approach to Interpreting Model Predictions, Lundberg et al.
[6] Interpretation of NLP models through input marginalization, Kim et al.

Text explanations: natural language

Premise
Wet brown dog swims towards camera.

Hypothesis
A dog is sleeping in his bed.

Example adapted from “Natural Language Inference with a Human Touch: Using Human Explanations to Guide Model Attention”, Stacey et al.

Text explanations: natural language

Premise
Wet brown dog swims towards camera.

Hypothesis
A dog is sleeping in his bed.

Natural Language explanation:
A dog cannot be sleeping while he swims.

Example adapted from “Natural Language Inference with a Human Touch: Using Human Explanations to Guide Model Attention”, Stacey et al.

Text explanations: natural language

Natural Language explanation:
A dog cannot be sleeping while he swims.

- By design: fine-tune a (conditioned[7]) language model[8,9,10]

- Brittle
- Require training data
- Require specific NLP knowledge from who explains

[7] NILE : Natural Language Inference with Faithful Natural Language Explanations, Kumar et al.
[8] Explanations for CommonsenseQA: New Dataset and Models, Shourya et al.
[9] Extractive and Abstractive Explanations for Fact-Checking and Evaluation of News, Kazemi et al.
[10] Learning to Explain: Answering Why-Questions via Rephrasing, Nie et al.

TriplEx: a semantically-principled approach

- Natural language-like explanations, enriched with importance
- No need for additional data or expertise
- Semantically meaningful explanations through domain knowledge, i.e. knowledge bases[11]

[11] WORDNET: a lexical database for english, Miller

Our setting

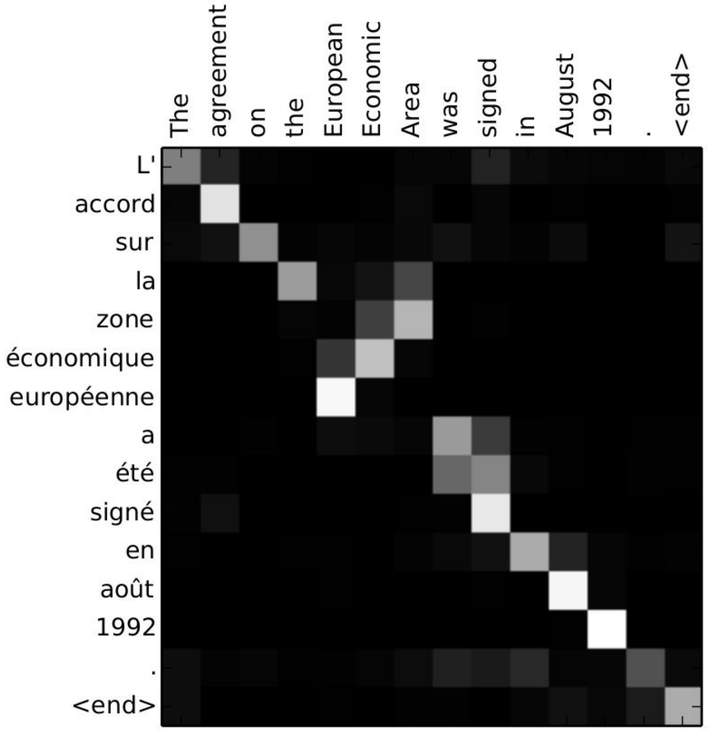
- Explain Transformer-based models
- Applications in NLI, Text Classification (TC), and Semantic Text Similarity (STS)
- Explanations as abstraction

TriplEx

Transformer-based models

Transformers (BERT,[12] RoBERTa[13], ALBERT[14]) implement a form of internal relevance, *attention* between input tokens[15]:

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



[12] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al.
[13] RoBERTa: A Robustly Optimized BERT Pretraining Approach, Liu et al.
[14] ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, Lan et al.

[10] Attention Is All You Need, Vaswani et al.

TriplEx

Explanations through

- information extraction
- semantic perturbation

Key idea #1: rely on information, not tokens!

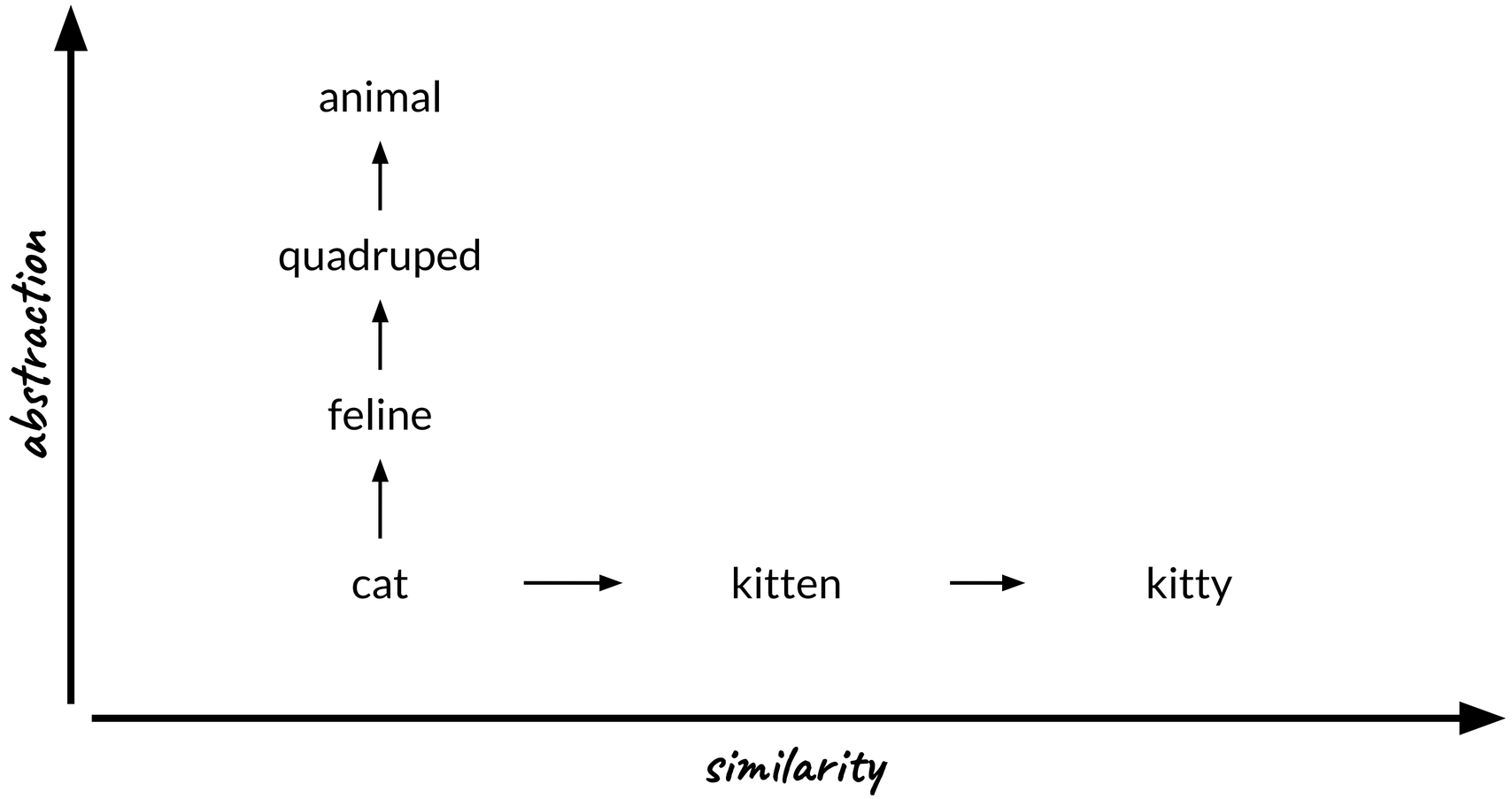
Information extraction is a well-developed field, we have several algorithms at our disposal:

- OpenIE[16]
- MinIE[17]
- OllIE[18]

they extract (subject, predicate, object) triples

[16] Open information extraction from the web, Etzioni et al.
[17] Minie: minimizing factsin open information extraction, Gashteovski et al.
[18] Open language learning for information extraction, Mausam et al.

Key idea #2: semantically meaningful perturbations



[11] WORDNET: a lexical database for english, Miller

Explanation algorithm

1 | Triples extraction from premise/text

Premise for NLI, whole text for STS, TC

2 | Triples semantic perturbation

3 | Model query

Explanation algorithm

1 | Triples extraction from premise/text

2 | Triples semantic perturbation

Generate all perturbations with the given knowledge base

3 | Model query

[11] WORDNET: a lexical database for english, Miller

Explanation algorithm

1 | Triples extraction from premise/text

2 | Triples semantic perturbation

3 | Model query

Return the perturbation with highest distance, i.e. the highest abstraction for NLI, and the most dissimilar for STS, preserving the model's classification

Explanation algorithm

1 | Triples extraction from premise/text

2 | Triples semantic perturbation

3 | Model query

4 | [Optional] Enriching with Triple alignment

Assign an average attention score to each triple, either w.r.t. the premise (NLI) or the whole text (STS)

Premise Cairo is now home to some 15 million people - a burgeoning population that produces approximately 10,000 tonnes of rubbish per day, putting an enormous strain on public services. In the past 10 years, the government has tried hard to encourage private investment in the refuse sector, but some estimate 4,000 tonnes of waste is left behind every day, festering in the heat as it waits for someone to clear it up. It is often the people in the poorest neighbourhoods that are worst affected. But in some areas they are fighting back. In Shubra, one of the northern districts of the city, the residents have taken to the streets armed with dustpans and brushes to clean up public areas which have been used as public dumps.
Hypothesis 15 million tonnes of rubbish are produced daily in Cairo.
Label Contradiction

Subject	Predicate	Object
Cairo	is	home to some 15 million people
Government	encourage	finance in waste sector
Finance	is	in waste sector
4000 tonnes	are	left
People	are	in poor neighborhood

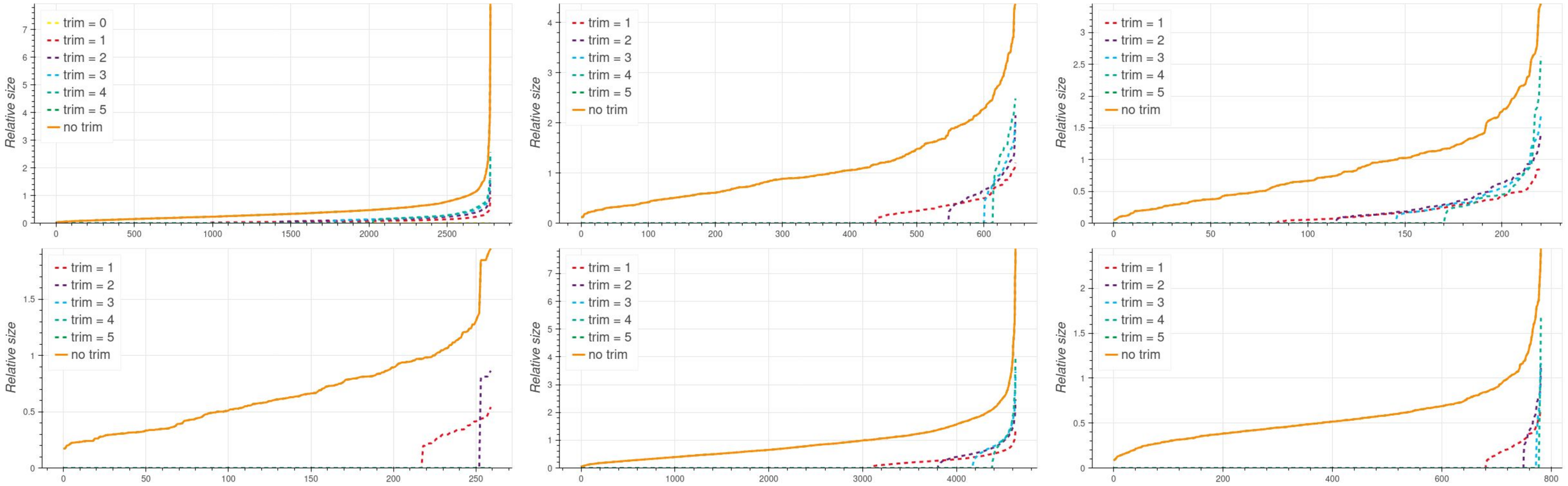
Premise Cairo is now home to some 15 million people - a burgeoning population that produces approximately 10,000 tonnes of rubbish per day, putting an enormous strain on public services. In the past 10 years, the government has tried hard to encourage private investment in the refuse sector, but some estimate 4,000 tonnes of waste is left behind every day, festering in the heat as it waits for someone to clear it up. It is often the people in the poorest neighbourhoods that are worst affected. But in some areas they are fighting back. In Shubra, one of the northern districts of the city, the residents have taken to the streets armed with dustpans and brushes to clean up public areas which have been used as public dumps.
Hypothesis 15 million tonnes of rubbish are produced daily in Cairo.
Label Contradiction

Attention score	Subject	Predicate	Object
0.0567	4000 tonnes	are	left
0.0500	Cairo	is	home to some 15 million people
0.0458	People	are	in poor neighborhood
0.0433	Finance	is	in waste sector
0.0411	Government	encourage	finance in waste sector

Validation

Complexity

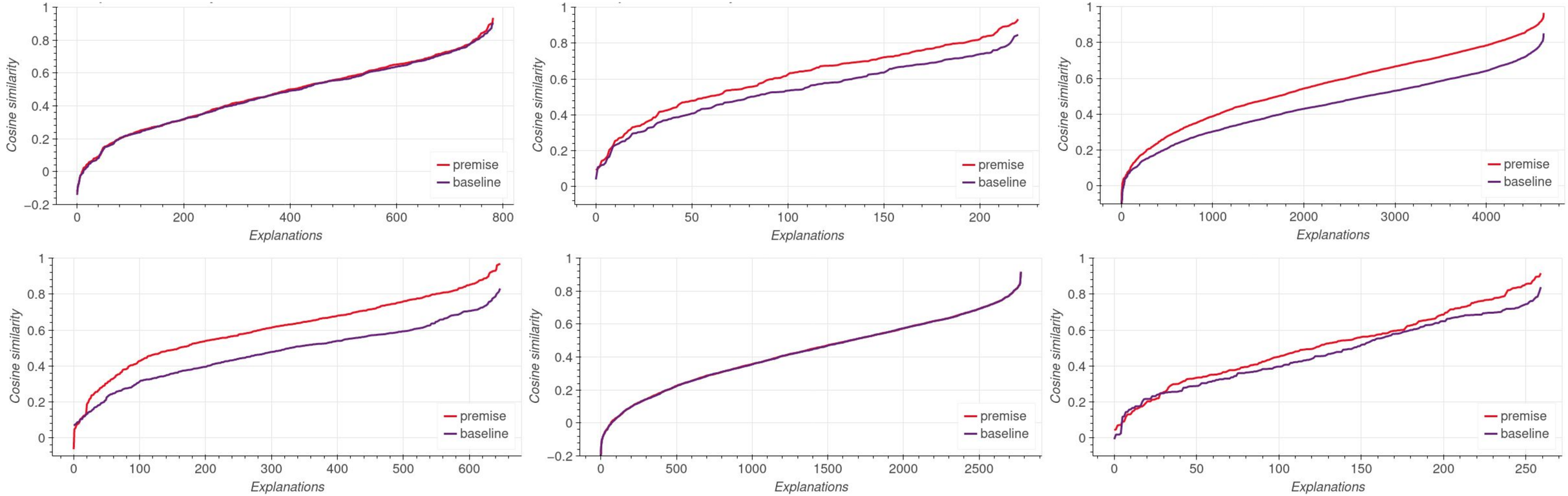
Relative explanation size, i.e., the ratio between the explanation and premise/input text length.



[19] GLUE: A multi-task benchmark and analysis platform for natural language understanding, Wang et al.
[20] Superglue: A stickier benchmark for general-purpose language understanding systems, Wang et al.
[21] Hidden factors and hidden topics: understanding rating dimensions with review text, McAuley & Leskovec

Information retention

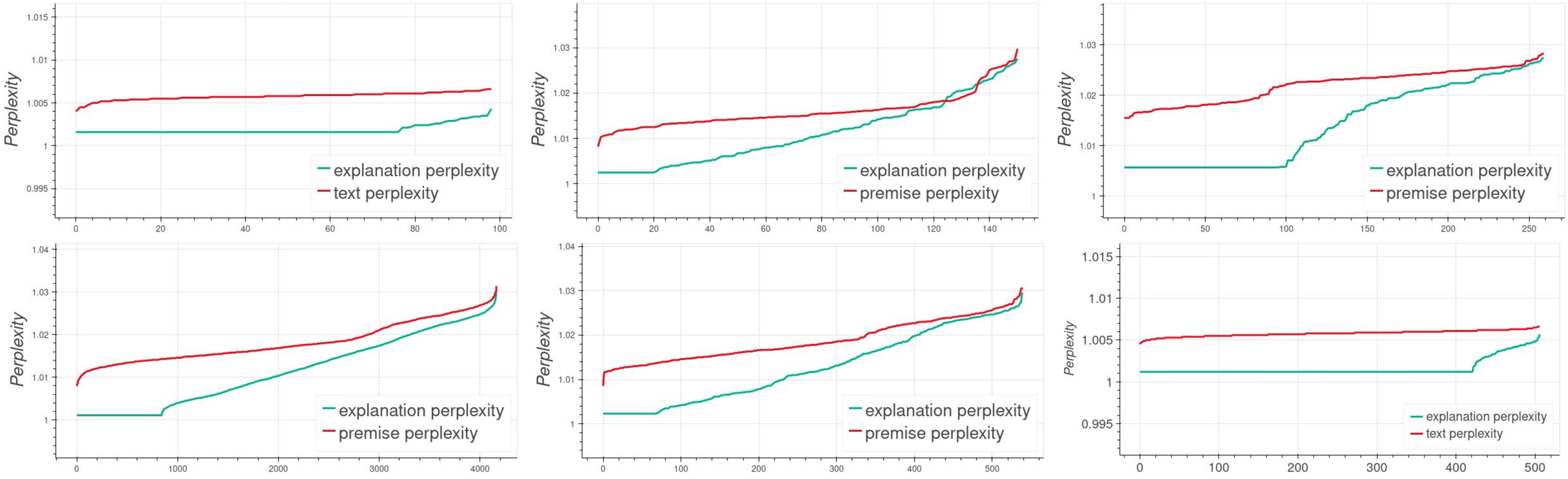
How similar are the explanations to the original text?[14]



[22] Sentence-bert: Sentence embeddings using siamese bert-networks., Reimers & Gurevych

Perplexity

How realistic are the explanations?[15]



[23] How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, Liu et al.

TriplEx

Triples Extraction for Explanation

TriplEx

- Explanation for Transformer-based text models
- Information extraction as basis for explanations
- Semantic perturbation

Future directions

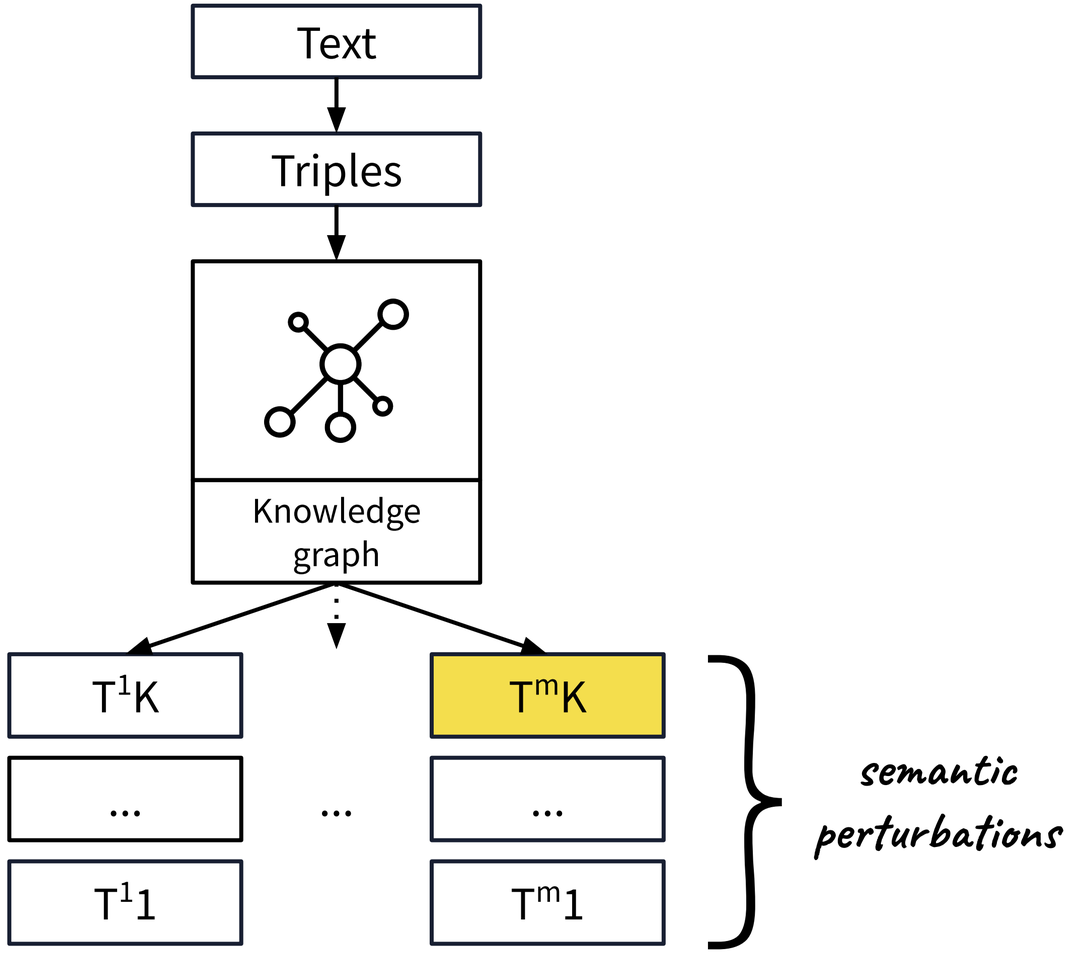
- Perturbation along more axes
- Text enrichment w/ commonsense KGs

github.com/msetzu/triplex

TriplEx

Explaining Transformer models
with triples:

- NLI and generic text classification tasks
- Semantic perturbation through knowledge graph
- Importance score for each triple



TriplEx: Triples for Explanation, Setzu et al.