WIP*: an Evaluation Metric for Logic-to-Text Generation

Eduardo Calò, Guanyi Chen, Elias Stengel-Eskin,
Albert Gatt, Kees van Deemter

UU NLP Group Meeting

September 28, 2023

* LoTE: https://en.wikipedia.org/wiki/Celtis_australis
GLEM: https://starwars.fandom.com/wiki/Glem | https://github.com/GreekPerspective/glem
ELoT: https://en.wiktionary.org/wiki/ELOT

Elias

https://esteng.github.io/

Met at ACL earlier this year
Extremely interested in this idea, started collaboration

PhD at Johns Hopkins University in semantic parsing, human-robot interaction, ambiguity
Postdoc at University of North Carolina, Chapel Hill

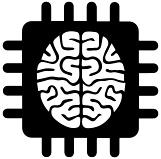
Overview

1. Introduction
2. Framework
3. Methodology
4. Plan

Logic-to-Text Generation

$$\exists x(Problem(x) \land \forall y(Researcher(y) \rightarrow Interested(y, x)))$$

$$\Downarrow$$



$$\Downarrow$$

- There is an x, s.t. x is a problem and, for all y, if y is a researcher, then y is interested in x.

- There is a problem that every researcher finds interesting.

- Every researcher finds a problem interesting.

...

Research Questions

Can we **automatically evaluate** the *quality* of text generated from logical formulae?

- Why yet another metric?
- How do we define quality?

Motivations

- **Lack** of appropriate evaluation **metrics** for **logically rich texts**
- **Existing metrics** focus on **different** linguistic **aspects** (lexical similarity, etc.)

	Reference	Realization	Metric Score	Metric
(1)	*There are no cubes.*	*Nothing is a cube.*	0.000	ROUGE-L
(2)	*Everybody eats apples.*	*Nobody eats apples.*	0.999	BERTSCORE

- **Human evaluation** is **expensive** and time-consuming
- Rise of **LLMs requires** better **metrics** to evaluate their outputs
- **Neural models** still **struggle with** processing **logically rich text**

Definitions

1. **Faithfulness**: *A generated text T is faithful to the original formula F if (at least) one of the interpretations of T reflects (i.e., is logically equivalent to) the meaning expressed by F.*
2. **Viciousness**: *A generated text T read by reader R in a domain D is viciously ambiguous in D if (i) R's most prominent (i.e., likely by a large margin w.r.t. the others) interpretation for T in D does not reflect (i.e., is not logically equivalent to) the meaning expressed by the original formula F, or (ii) R perceives distinct interpretations for T in D as equally prominent.*
3. **Naturalness**: *A generated text T is natural if T is morphosyntactically well-formed, fluent, and natural sounding.*

Why These Dimensions?

- Evaluation of **Content** and **Form** separate
- **Overlapping** and/or conflating of **different dimensions** in previous literature (we give precise definitions)
- Empirical **evidence** that **ambiguity** needs to be **addressed** in our task [Calò et al., 2022, Calò et al., 2023]

Sentence	*If b is a tetrahedron, then b is a tetrahedron and it is not the case that c is a tetrahedron.*
Interpretation 1	*(If b is a tetrahedron, then b is a tetrahedron) and (it is not the case that c is a tetrahedron).*
Interpretation 2	*If (b is a tetrahedron), then (b is a tetrahedron and it is not the case that c is a tetrahedron).*
Original Formula	$\mathit{Tet}(b) \rightarrow (\mathit{Tet}(b) \wedge \neg \mathit{Tet}(c))$

Tasks Framing

1. **Faithfulness**: Natural language inference: how faithful is T to F?
2. **Viciousness**: Text classification: how vicious is T w.r.t. F?
3. **Naturalness**: Reference-less structural evaluation: how natural is T?

Faithfulness

Premise: *Every researcher finds a problem interesting.*

Hypothesis: $\exists x(\mathit{Problem}(x) \wedge \forall y(\mathit{Researcher}(y) \rightarrow \mathit{Interested}(y, x)))$

Faithfulness: 0.99

Data:

- Retrieve various parallel datasets

- Homogenization of the datasets

- Augmentation (e.g., swapping formula operators) to create negative samples

Viciousness

Text: *Every researcher finds a problem interesting.*
Original interpretation: $\exists x(Problem(x) \land \forall y(Researcher(y) \rightarrow Interested(y, x)))$
Viciousness: 0.75

Data gathering (experimental setup):
Text: *Every researcher finds a problem interesting.*
Interpretation 1: *There is a problem that every researcher finds interesting.*
Interpretation 2: *Each researcher finds a (possible different) problem interesting.*
User selection: Interpretation 2
Confidence: 0.85
(Selection not matching original interpretation and user quite confident → T viciousness high)

Naturalness

We are **yet to decide** what to use to measure naturalness of T. Options include:

- Perplexity
- Grammatical error detection
- Readability metrics
- ...

Plan & Problems (I)

Faithfulness:

✓ Gather existing parallel datasets

✓ Homogenize datasets

1. Downsample datapoints (e.g., those with exact same structure, etc.)
2. Data perturbation to create unfaithful samples
3. Dataset quality inspection
4. Train model

Plan & Problems (II)

Viciousness:

✓ Selection of ambiguities to focus on

1. Setup human annotation experiment:
   - Expand existing framework [Stengel-Eskin et al., 2023] to accommodate more ambiguities and interpretations
   - Recruit participants (Prolific)
2. Train model

Subsequent:

- Experiments on **reference-less metrics** for naturalness
- **Correlation**: Gather human judgments on an independent set of texts generated from formulae on the three dimensions and compute correlations between these judgments and the scores of our metric to show effectiveness of our metric

References I

- Calò, E., Levy, J., Gatt, A., and Van Deemter, K. (2023).
  Is shortest always best? the role of brevity in logic-to-text generation.
  In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 180–192, Toronto, Canada. Association for Computational Linguistics.

- Calò, E., van der Werf, E., Gatt, A., and van Deemter, K. (2022).
  Enhancing and evaluating the grammatical framework approach to logic-to-text generation.
  In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 148–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

- Stengel-Eskin, E., Rawlins, K., and Durme, B. V. (2023).
  Zero and few-shot semantic parsing with ambiguous inputs.

Questions
&
Feedback Time