Evaluating causal extraction

Italo Luis da Silva

2023-12-01

King's College London

The Problem

Dataset: Fine Grained Causal Reasoning

Extract cause, effect and relation from text passage.

Example

The firm's gross margin is set to stabilize as Harley refocuses its efforts on more profitable markets, and our base case assumes that it stabilizes around 32% in 2029, helped by a more measured approach to entering new markets.

- Cause: Harley refocuses its efforts on more profitable markets
- Effect: The firm's gross margin is set to stabilize
- **Relation**: cause

Evaluation

The standard metric, Exact Match, is not a good metric for this task. This is correct but not an exact match:

Annotation

BB&T and SunTrust have completed their merger, forming Truist, which we believe will drive the next step up in profitability for the franchises.

Model prediction

BB&T and SunTrust have completed their merger, forming Truist, which we believe will drive the next step up in profitability for the franchises.

Cause Effect

Evaluation methods

- EM and F1 are not good
- ROUGE-L and BLEU
- Trained metrics like BertScore and BLEURT
- Goal: metric should be compatible with human evaluation
  - None of these are quite there yet.

Custom evaluation methods

- Entailment: fine tune a model[1] to predict whether the extracted text is entailed by the context
  - Uses synthetic data
- NLI: use a pre-trained NLI model to predict the entailment
  - Structured output is rewritten as a natural language sentence
- Valid: train a binary classifier to predict whether the extracted text is valid
  - Requires explicit human annotation
- Problem: perfect evaluation implicitly requires solving the problem!
  - Search problem $\Rightarrow$ decision problem

[1] Entailment and Valid are DeBERTa-v3 models. NLI is DeBERTa-MNLI.

Reinforcement learning

- How to improve the model?
- Supervised learning is limited as it tries to learn exact wording
- Use Reinforcement Learning to improve the fine-tuned extraction model
- Reward models: entailment, NLI and valid as before
    - Reward is the logit of the true class (entailment or valid)

Evaluation results

Models	Human	Token F1	EM
ChatGPT (10-shot)	35.13%	67.52%	31.95%
Supervised	64.38%	80.59%	54.31%
RL with entailment	59.23%	76.58%	47.06%
RL with valid	60.48%	78.65%	50.02%
RL with MLNI	-	75.65%	44.92%

Evaluation results (cont.)

Models	Human	ROUGE-L	BLEU
ChatGPT (10-shot)	35.13%	64.33%	61.76%
Supervised	64.38%	77.18%	75.83%
RL with entailment	59.23%	73.08%	73.42%
RL with valid	60.48%	75.47%	75.31%
RL with NLI	-	75.73%	75.49%

Evaluation results (cont.)

Models	Human	BLEURT[2]	BertScore F1
ChatGPT (10-shot)	35.13%	63.09%	89.84%
Supervised	64.38%	75.30%	95.52%
RL with entailment	59.23%	71.61%	94.84%
RL with valid	60.48%	73.71%	95.25%
RL with NLI	-	73.73%	95.35%

[2]BLEURT-20-D12

Next steps

- Automated evaluation is tricky, but human evaluation is time-consuming and expensive
- LLM-based evaluation: how can we use the LLM to evaluate the output?
  - How to prompt GPT-3.5 and GPT-4 for this?
  - Fine-tuned open source models (e.g. Llama 2)
- There's work on general metrics like readability, grammar, faithfulness and context relevance[3], but not so much for specialised evaluation.

[3] RAGAS: https://github.com/explodinggradients/ragas

Future work

- WhyQA: TellMeWhy dataset
  - Question Answering on the reasons behind actions in the text
- Prone to the same evaluation challenges, but worse
- Harder to evaluate, both automatically and manually
- Answers aren't necessarily substrings
- Wording can very a lot
- Some questions can't be answered

Thanks!

github.com/oyarsa/event_extraction