Sentence-Level Rationale Extraction Methods

For Closed-domain Question Answering Explainability

Lize Pirenne, Samy Mokeddem, Damien Ernst

University of Liège

Rationale extraction

Layer **8**, Head **6**

Question: What is Sauvignon blanc?
Context: Sauvignon blanc is a green-skinned grape variety that originates from the city of Bordeaux in France. The grape early origins as an indigenous grape in South West France. It is possibly a descendant of Savagnin. Sauvignon blanc is planted The grape is also a component of the famous dessert wines from Sauternes and Barsac. Sauvignon blanc is widely cultivated in Oregon, Washington, and California in the US. Some New World Sauvignon blancs, particularly from California, may also be Fumé.
Depending on the climate, the flavor can range from aggressively grassy to sweetly tropical. In cooler climates, the grape has a peppers and nettles with some tropical fruit (such as passion fruit) and floral (such as elderflower) notes. In warmer climates, i only slight grapefruit and tree fruit (such as peach) notes.
Answer: strong**Sauvignon blanc is a green-skinned grape variety that originates from the city of Bordeaux in France. The to its early origins as an indigenous grape in South West France.**strong

The attention of Gemma-2b heads can innately find some rationales.

Benefits of Rationales

- **Explainability** [7]
  Context size can now reach the million tokens in Large Language Models, and their size decrease QA performances [6]. There is a need for efficient fast-checking!
- **Efficient prompt engineering**
  One can target only the most promising sentences as input [5].
- **Explanation Regularization** [4]
  We can use the ability to find rationales as an additional objective to optimize [2].

Problem Statement

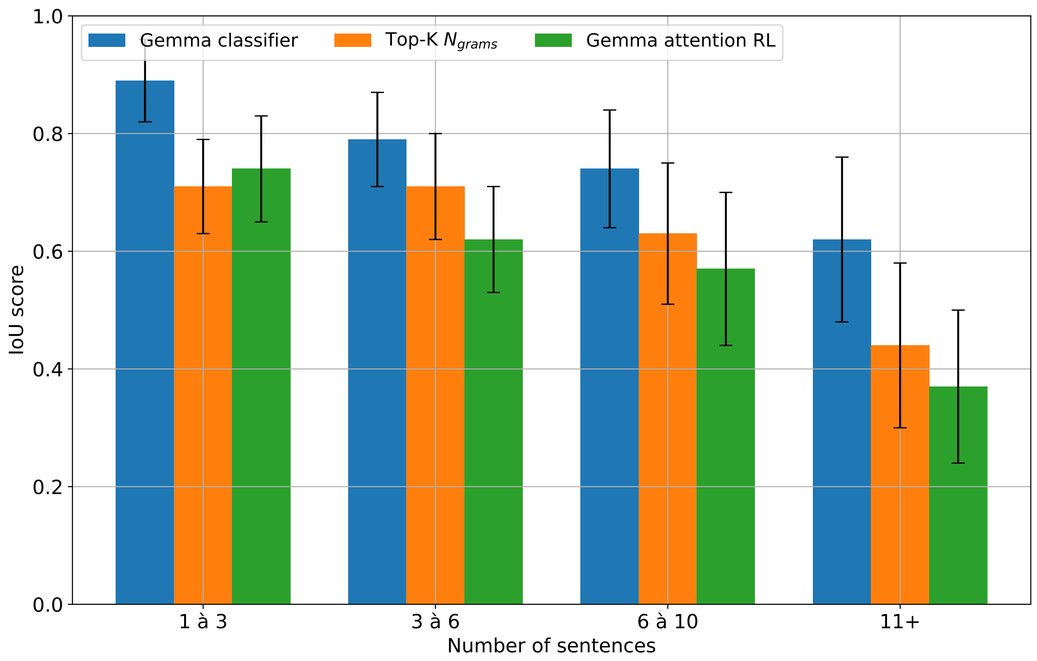
The task is to produce the rationales P given a question Q, a context C and possibly the answer A. The model is characterized by $M(Q, C, A) = P$. Our attention model also produces A, so the model becomes $M(Q, C) = A, P$. To measure the performances of the models for the task of **rationale extraction**, we have introduce the IoU score extended for sentences.

$$IoU(P,\ S) = \frac{|P \cap S|}{|P \cup S|}$$

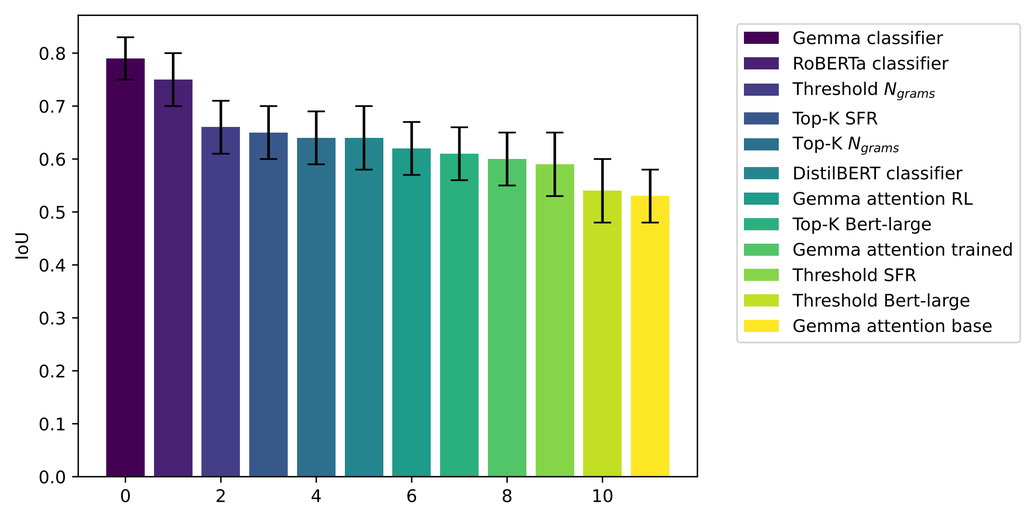
Method comparison

- **Classifiers based on Large Language Models perform better** than the rest for our task. Indicating that the ability to answer might be linked to rationale extraction.
- **Using Reinforcement Learning can lead to increased performances** when modifying attention weights with IoU plus Meteor [1] as reward score .
- The strong performances of a simple N-gram model in this custom Dolly Databricks [3] dataset show that either the answers are too straight-forward, or that humans prefer reusing the wording of their source.

Performances against context length



Study of best performing methods



Importance of context length

Upon inspection of our best performing models, we can see that **the number of sentences in the context play a major role** in the performances even at small scales.
It will become important to find solutions that scale with them in speed and in IoU. Indeed, the time for classifier

Future Works

Multi-rationale extraction The task becomes even more challenging when we have to identify which sentence referred to which sub-question.
More RL? Reward models are cumbersome with LLMs, is there a better way?

References

[1] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*. Ed. by Jade Goldstein et al. Association for Computational Linguistics, 2005, pp. 65–72.

[2] Aaron Chan et al. "UNIREX: A Unified Learning Framework for Language Model Rationale Extraction". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2867–2889.

[3] Mike Conover et al. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. 2023. (Visited on 02/05/2024).

[4] Brihi Joshi et al. "ER-test: Evaluating Explanation Regularization Methods for Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, 2022, pp. 3315–3336. DOI: 10.18653/V1/2022.FINDINGS-EMNLP.242.

[5] Satyapriya Krishna et al. "Post Hoc Explanations of Language Models Can Improve Language Models". In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh et al. 2023.

[6] Freda Shi et al. "Large Language Models Can Be Easily Distracted by Irrelevant Context". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 31210–31227. (Visited on 05/10/2024).

[7] Haiyan Zhao et al. "Explainability for Large Language Models: A Survey". In: *ACM Transactions on Intelligent Systems and Technology* 15.2 (Feb. 2024). ISSN: 2157-6904. DOI: 10.1145/3639372.