Constrained Linked Entity ANnotation using RAG (CLEANR)

Benedikt Kantz[1]*, Stefan Lengauer[1], Peter Waldert[1], Tobias Schreck[1]

[1]Institute of Visual Computing, Graz University of Technology
* Corresponding author, `benedikt.kantz@tugraz.at`

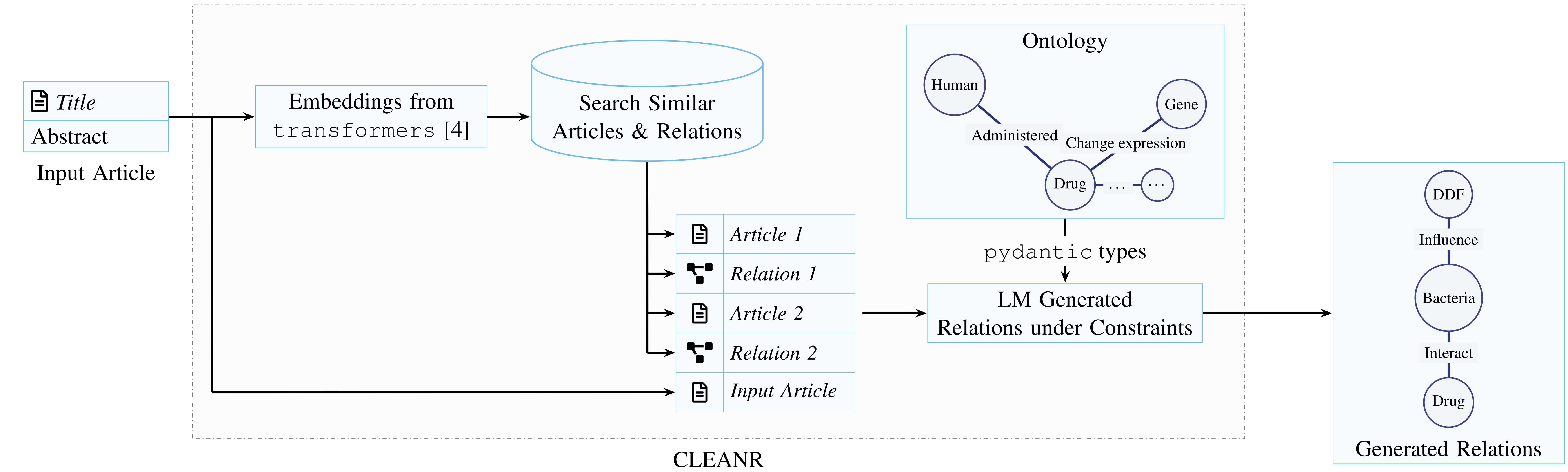
Idea

- Use general Language Models (LMs) like Hermes [1] . . . but constrain them!
- Few-shot tuning based on similar texts, constraints based on ontology!
- 👍 No fine-tuning, can be adjusted to arbitrary ontologies.
- 👎 Upper bounded by LM capabilities.

Related Relation Extraction (RE) Systems

RE system		
RAG4RE [2]	👍 Generalizable using LM	👎 no constraints, relies on LM adherence
REBEL [3], . . .	👍 Good performance	👎 long training time, specialized

System Overview



CLEANR

Evaluation Results

Subtask 6.2.1 Test Results

Model	RAG	LoRA	Reorder	P	R	F_1	P_{micro}	R_{micro}	$F_{1,micro}$
Hermes LLama 3.2 3B	×	✓	×	0.14	0.05	0.07	**0.74**	0.12	0.21
	✓	×	✓	0.18	0.09	0.11	0.57	0.17	0.26
		✓	✓	0.16	0.09	0.11	0.73	0.16	0.26
Hermes LLama 3.1 8B	✓	×	✓	**0.22**	0.13	0.15	0.57	0.26	**0.36**
OpenAI GPT 4-mini	✓	×	✓	0.20	**0.14**	**0.15**	0.39	**0.30**	0.34

Subtask 6.2.2 Test Results

Model	RAG	LoRA	Reorder	P	R	F_1	P_{micro}	R_{micro}	$F_{1,micro}$
Hermes LLama 3.2 3B	×	✓	×	0.13	0.04	0.06	0.68	0.09	0.17
	✓	×	✓	0.17	0.08	0.10	0.55	0.16	0.24
		✓	✓	0.15	0.08	0.10	**0.71**	0.15	0.24
Hermes LLama 3.1 8B	✓	×	✓	**0.23**	0.13	0.14	0.56	0.25	**0.34**
OpenAI GPT 4-mini	✓	×	✓	0.21	**0.15**	**0.15**	0.39	**0.28**	0.33

Subtask 6.2.3 Test Results

Model	RAG	LoRA	Reorder	P	R	F_1	P_{micro}	R_{micro}	$F_{1,micro}$
Hermes LLama 3.2 3B	×	✓	×	0.03	0.02	0.02	**0.18**	0.04	0.07
	✓	×	✓	**0.04**	**0.03**	0.02	0.04	0.02	0.03
		✓	✓	0.02	0.02	0.02	0.17	**0.05**	**0.08**
Hermes LLama 3.1 8B	✓	×	✓	0.03	0.03	**0.02**	0.06	0.04	0.05
OpenAI GPT 4-mini	✓	×	✓	0.00	0.00	0.00	0.00	0.00	0.00

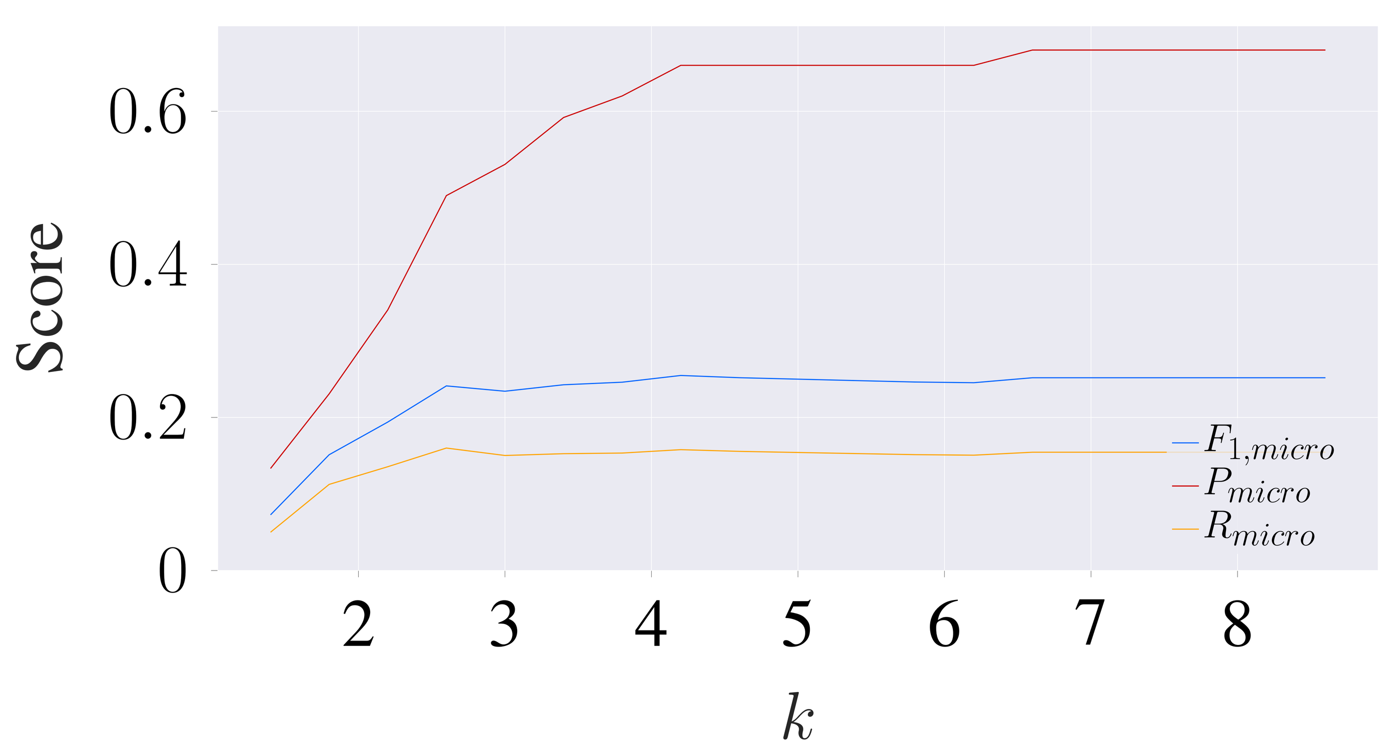
Discussion

- 👍 High precision for tasks without locality (6.2.1, 6.2.2), among best approaches (for this metric)!
- 👍 Precise with general models, even better when fine-tuned (optional).
- 👎 Low Recall $\Longrightarrow$ Low F_1, well below baseline.

Error Analysis

- LM power: Could be limited, used only small and low-budget models – seems to scale well!
- Scores @ k reveal that model output does not really improve after 3-4 relations.

Scores @ k for Task 6.2.1



Improvements

- Compare larger models (Hermes LLama 3.1B 70B, reasoning models?): employ domain-specific retrieval approaches and models.
- Generation hyperparameters: Modify temperature of generation to reduce repeated outputs, use beam-search for improved constrained generation.
- Improve system prompt/prompt fine-tuning by adding schema / . . .

Links

Paper

Code

References

[1] Ryan Teknium, Jeffrey Quesnelle and Chen Guang. *Hermes 3 Technical Report*. 2024. arXiv: 2408.11857 [cs.CL]. URL: https://arxiv.org/abs/2408.11857.

[2] Sefika Efeoglu and Adrian Paschke. *Retrieval-Augmented Generation-based Relation Extraction*. 2024. arXiv: 2404.13397 [cs.CL]. URL: https://arxiv.org/abs/2404.13397.

[3] Pere-Lluís Huguet Cabot and Roberto Navigli. 'REBEL: Relation Extraction By End-to-end Language generation'. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204.

[4] Nils Reimers and Iryna Gurevych. 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084.

LaTeX TikZposter