Learning Pattern-Based Extractors from Natural Language and Knowledge Graphs Applying Large Language Models to Wikipedia & the Linked Open Data (POSTER)

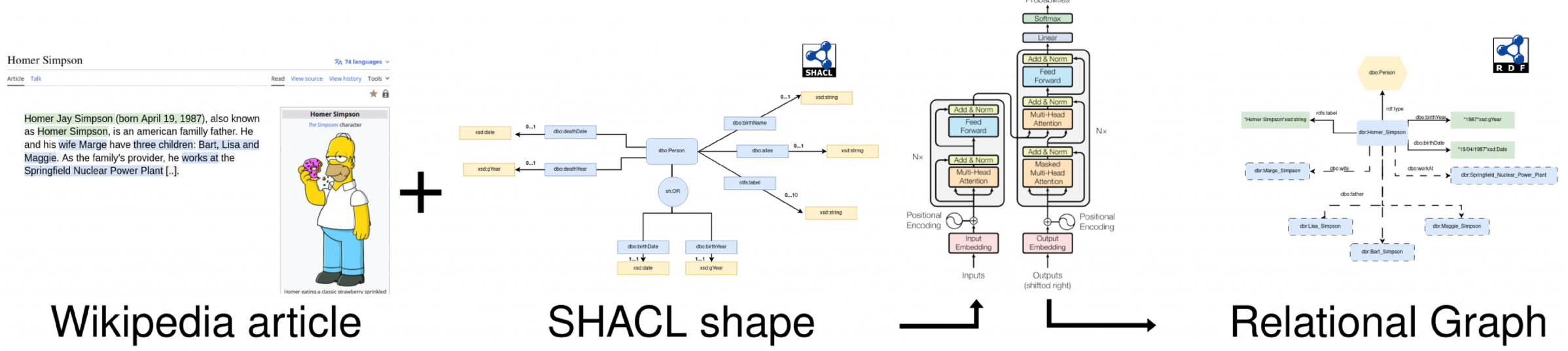
Celian Ringwald, Fabien Gandon, Catherine Faron, Franck Michel, Hanna Abi Akl

Learning Pattern-Based Extractors from Natural Language and Knowledge Graphs

Applying Large Language Models to Wikipedia & the Linked Open Data

Célian Ringwald[1], Fabien Gandon[1,2], Catherine Faron[1], Franck Michel[1], Hanna Abi Akl[1,2]

[1]Université Côte d'Azur, Inria, CNRS, I3S [2]Data ScienceTech Institute



Extractors for specific RDF patterns

Seq-to-seq transformer models have recently been successfully used for relation extraction, showing their flexibility, effectiveness and scalability on that task. In this context, knowledge graphs coupled with Wikipedia (e.g. DBpedia, Wikidata) allow us to leverage existing texts and corresponding RDF graphs to learn to extract such knowledge from text. The goal of this work is to learn efficient targeted extractors for specific RDF patterns by leveraging the latest language models and the dual base formed by Wikipedia on the one hand, and DBpedia & Wikidata on the other hand.

Research question:

Can we learn efficient customized extractors targeting specific RDF patterns from the dual base formed by Wikipedia on one hand, and DBpedia and Wikidata on the other hand?

Formalisation: Let Db be a dual base, $\subseteq W \times G$, where W is a set of Wikipedia articles and G a set of corresponding RDF graphs in DBpedia and Wikidata.
The goal is to learn: E_{Db}: $W \rightarrow L; (t, S) \mapsto g$,
where L is the LOD, t is an input text, S is a set of RDF patterns of interest represented as SHACL shapes, and g is an RDF graph implied by t and valid against S.

Related work

- Before LLM, RE task was solved by complex pipelines including multiple steps [1]. But this approach leads to error accumulations and propagation [2].
- Two main approaches proposed by the literature: Discriminative (based on encoder-only models) vs Generative RE (based on encoder-decoder or decoder-only models) [3].
- Generative RE demonstrate several successes: [4], [5], [6] but still face to limitations.

Incremental strategy

SRQ.1 – *Survey and follow latest trends in PLM-based KG extraction?*
- ☑ A pipeline for Scientific Literature exploration
- ☐ A systematic survey on the Relation Extraction task

SRQ.2 – *Which aspects of the task formulation impact the generation of triples with datatype properties?*
- ☑ "Well-written knowledge graphs": most effective RDF syntaxes for triple linearization in end-to-end extraction of relations from text
- ☐ Systematic evaluation of syntax and configuration impact on RE task with Encoder-Decoder models
- ☐ How to design the best possible prompt?
- ☐ How to integrate the SHACL shape into the prompt?

SRQ.3 – *How to jointly extract datatype properties and object properties for a KG?*
- ☐ HTML input + Copy-mechanism integration + Constrained decoding

SRQ.4 – *How to support fact extraction relying on different document granularity?*
- ☐ Encoder-decoder models with larger context and/or retrieval-based embeddings

SRQ.5 – *What is the best strategy to extract rare relations and under-represented instances of classes?*
- ☐ Data augmentation & Synthetic data

References

[1] J. L. Martinez-Rodriguez, A. Hogan, and I. Lopez-Arevalo. "Information extraction meets the Semantic Web: A survey". In: *Semantic Web* 11.2 (Feb. 5, 2020). Ed. by A. Hotho.

[2] F. Mesquita, M. Cannaviccio, J. Schmidek, P. Mirza, and D. Barbosa. "KnowledgeNet: A Benchmark Dataset for Knowledge Base Population". In: Proceedings of 2019 EMNLP-IJCNLP. Hong Kong, China: Association for Computational Linguistics, 2019.

[3] H. Ye, N. Zhang, H. Chen, and H. Chen. "Generative Knowledge Graph Construction: A Review". In: *EMNLP*. ACL, Dec. 1, 2022.

[4] P.-L. Huguet Cabot and R. Navigli. "REBEL: Relation Extraction By End-to-end Language generation". In: *Findings of the ACL: EMNLP 2021*. Findings 2021. ACL, Nov. 2021.

[5] M. Josifoski, N. De Cao, M. Peyrard, F. Petroni, and R. West. "GenIE: Generative Information Extraction". In: *Proceedings of the 2022 Conference of the NAACL*. ACL, 2022.

[6] C. Wang, X. Liu, Z. Chen, H. Hong, J. Tang, and D. Song. "DeepStruct: Pretraining of Language Models for Structure Prediction". In: *Findings of the ACL: ACL 2022*. ACL, May 2022.