THE UNIVERSITY OF MELBOURNE

Natural language generation from the knowledge graph

Jiachen Li 1068299

Supervised by Jianzhong Qi

Research Question

Can the pre-trained language models be used to improve the current solution to natural sentence generation from the knowledge graph problem?

Pre-trained language models - BERT

- Bidirectional Encoder Representations from Transformers (BERT)
- Designed based on a multi-layer bidirectional Transformer encoder
- Pre-trained on BooksCorpus (800M words) and English Wikipedia (2,500M words)
- Learnt bidirectional text representation that captures syntactic and semantic information
- Easy to apply to the downstream task

NLG from the knowledge graph

- Convert unstructured data into a human-readable sentence.
- Knowledge graphs are built by many RDF triples in <subject, predicate, object> format.
- The current solution was the GTR-LSTM encoder-decoder-based model proposed by Trisedya et al.

RDF triples	⟨John Doe,birth place,London⟩ ⟨John Doe,birth date,1967-01-10⟩ ⟨London,capital of,England⟩
Target sentence	John Doe was born on 1967-01-10 in London, the capital of England.

Table 1: RDF based sentence generation.

Table 1 from Trisedya, Bayu Distiawan, Jianzhong Qi, Rui Zhang, and Wei Wang. "GTR-LSTM: A Triple Encoder for Sentence Generation from RDF Data." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1627–37. Melbourne, Australia: Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/P18-1151.

Why this research?

- NLG and BERT are trending
- NLG is used for Artificial Intelligence or question-answering systems
- NLG is also needed for Metaverse
- Metaverse has an exploding market
- The state-of-art GTR-LSTM did not use any pre-trained model at all

Scenario 1

- RDF triple pre-processor entity mapper triple pre-processor entity mapper
- Map entity to type
- Help handle unseen entities
- The baseline relied on API call
- Is a named entity recognition problem
- Proposed solve by NER-BERT
- Evaluate using BLEU, METEOR, and TER benchmarks
- Record running time, memory and internet usage
- Contribute an internet-free model with potentially higher accuracy

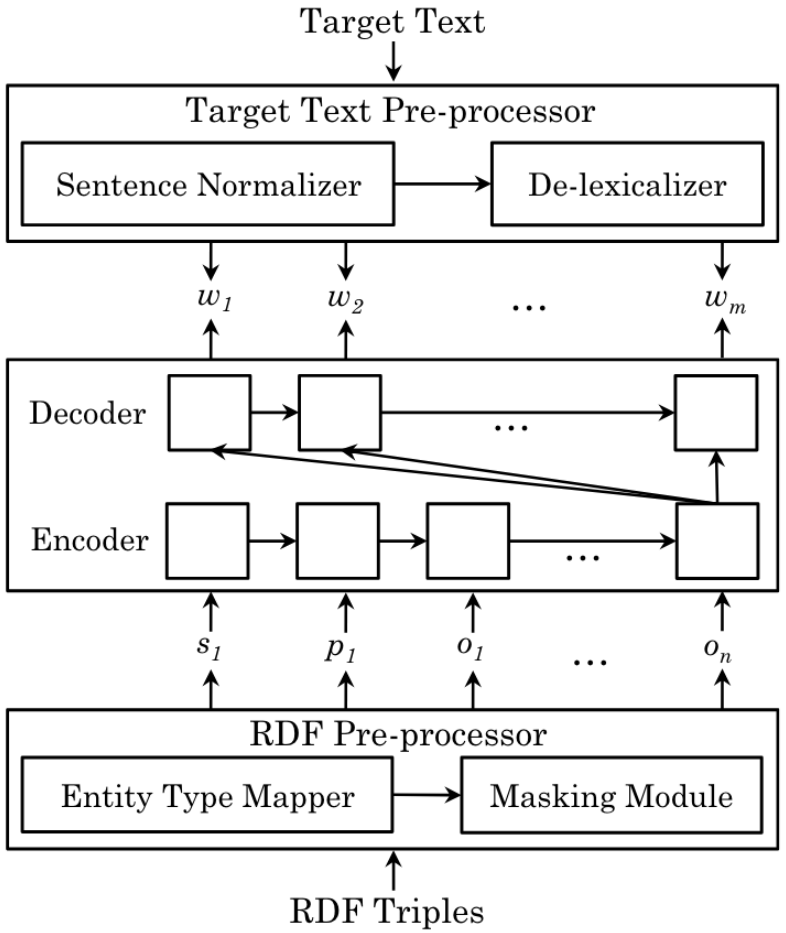

Figure 1: RDF sentence generation based on an encoder-decoder architecture.

Figure 1 from Trisedya, Bayu Distiawan, Jianzhong Qi, Rui Zhang, and Wei Wang. "GTR-LSTM: A Triple Encoder for Sentence Generation from RDF Data." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1627–37. Melbourne, Australia: Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/P18-1151.

Scenario 2

- Entity alignment
- Merge the same entities with different representations
- The baseline used joint Learning of Structure Embedding and Attribute Character Embedding
- Proposed to use the BERT embedding leant syntactic and semantic knowledge
- Evaluate using DWY-NB and other benchmarks in the original paper
- Record running time and memory usage
- Contribute a model with faster running speed, and likely higher accuracy

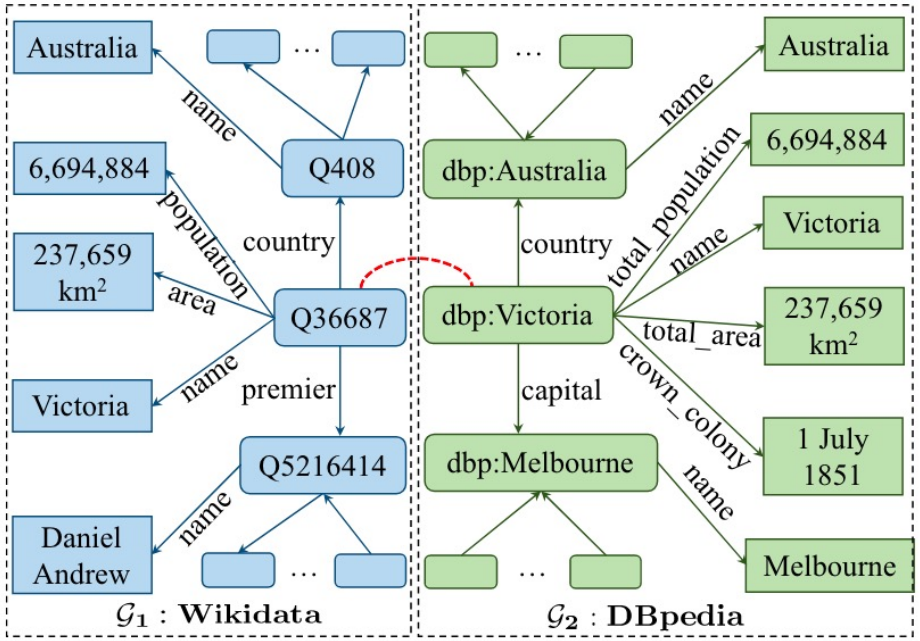

Fig. 1 An example of EA.

Fig 1 from Zhang, Rui, Bayu Distiawan Trisedy, Miao Li, Yong Jiang, and Jianzhong Qi. "A Benchmark and Comprehensive Survey on Knowledge Graph Entity Alignment via Representation Learning." arXiv, May 5, 2022. https://doi.org/10.48550/arXiv.2103.15059.

THE UNIVERSITY OF MELBOURNE

Thank you