The 13th Conference „Data Analysis Methods for Software Systems”
December 1 – 3, 2022, in Druskininkai, Lithuania.

COST Action CA18231

Enrich Knowledge Graphs and Test Pre-trained Language Models in Graph2seq Tasks

Grazina Korvel [1], Alkiviadis Katsalis [2], Konstantinos Diamantaras [2], Elena Lloret [3]

1 Institute of Data Science and Digital Technologies, Vilnius University

2 Department of Information and Electronic Engineering, International Hellenic University, Thermi, Greece

3 Department of Software and Computing Systems, University of Alicante, Alicante, Spain

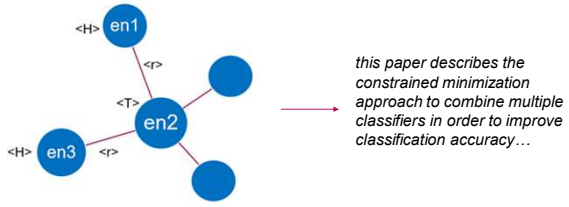
grazina.korvel@mif.vu.lt

Motivation

Modern natural language generation methods usually take the form of an encoder-decoder, which encodes the input sequences into latent space and predicts a collection of words based on the latent representation. Sequence-to-sequence (Seq2Seq) learning is one of the most widely used encoder-decoder based paradigms in this field. Recently, researchers find that **structural knowledge is beneficial to addressing some troublesome challenges**, e.g., long-dependency problems, and thus propose the **graph neural network (GNN) techniques**. Since structural knowledge is important for text generation and existing transformer-based models perform brilliant in text generation tasks, the combination of both seems promising.

Representation Learning on Graph with linearized input

$G = (H, r, T)$ – a graph
where H is the head node, r is the relationship and T is the tail node.



<H> en1 <r> rel1 <T> en2 <H> en3 <r> rel2 <T> en2

Fig. 1: Prepare input for language model (BART, T5, Pegasus) in a graph format for abstract generation task

Dataset

AGENDA Dataset: KGs are paired with scientific abstracts extracted from proceedings of AI conferences. Each sample contains the **paper title, a KG, entities, type of entities and the corresponding abstract.**

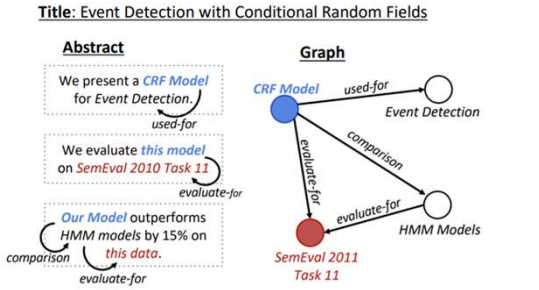


Fig. 2: A scientific text showing the annotations of an information extraction system and the corresponding graphical representation [1].

The challenges

Disregarding success, Graph2Seq models also inherit challenges, for example encoding relations between distant nodes. The outstanding issue is **how to incorporate information from the text and transform it properly into a graph.**

Our hypothesis

Enriching the graphs with **external knowledge** and testing pretrained language models can benefit text generation tasks. Using sentence transformers, the KGs can be enriched with **extra relationships** between available entities in the dataset. **Transformation of the title** from text to graph can also be beneficial.

Experimental set-up and result

- The similarities between entities that were not present in the existing KG of AGENDA dataset are found.
- BART-large pre-trained language model is fine-tuned in abstract generation problem.

Our approach METEOR: 0.35
SoT [2] METEOR: 0.257

References

[1] Koncel-Kedziorski, Rik, et al. "Text generation from knowledge graphs with graph transformers." arXiv preprint arXiv:1904.02342 (2019).
[2] Ribeiro, Leonardo FR, et al. "Investigating pretrained language models for graph-to-text generation." arXiv preprint arXiv:2007.08426 (2020).

Acknowledgements: This research work is part of the European Commission ICT COST Action “Multi-task, Multilingual, Multi-modal Language Generation" (CA18231). Moreover, it has been also part of R&D project “CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by “ERDF A way of making Europe” and by the Generalitat Valenciana through the project “NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)“.