You are invited to the 52nd edition of the PRAGUE COMPUTER SCIENCE SEMINAR

ONDŘEJ DUŠEK

Robust Data-to-Text Generation with Pretrained Language Models

The lecture will be followed by a discussion

February 9, 2023
4:15pm

Auditorium KN:E-107,
FEL CTU, Karlovo nám. 13,
Praha 2

ABSTRACT

The task of data-to-text generation amounts to describing structured data in fluent natural language sentences. The state-of-the-art approach in research systems today is finetuning pretrained neural language models (PLMs). This often leads to overfitting and hallucinations, i.e. situations where the PLM generates outputs that are not grounded in the input, replicating or amplifying training data noise. Rather than applying a PLM as black box for the whole data-to-text task, we aim at using PLMs for simple subtasks, aiming to achieve broad generalization and minimize hallucination.

First, we use a pipeline approach where the PLMs only work as text “editors”, rather than generators, taking advantage of their high output fluency. The data is converted into text in an initial preprocessing step, where we use simple handcrafted templates recounting the individual input facts (i.e. relations between entities). The PLMs then order the facts and fuse them into fluent sentences. This helps us generate without in-domain training data and achieve good fluency and accuracy. We further examine the capability of PLMs to produce accurate descriptions of individual facts from the data, in order to remove the last handcrafted step. Using a specially collected dataset, we show that PLMs finetuned to describe a variety of relations are very robust in verbalizing novel, unseen relations. The key to PLMs’ usability here is providing clear relation names on the input.

ABOUT THE PRAGUE COMPUTER SCIENCE SEMINAR

The seminar takes place once a month on Thursdays at 4:15pm (except June to September, and December) alternately in the buildings of Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo nám. 13, Praha 2 and Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, Praha 1. Its program typically consists of a one-hour lecture followed by a discussion. The lecture is based on an (internationally) exceptional or remarkable achievement of the lecturer, presented in a way which is comprehensible and interesting to a broad computer science community. The lectures are in English.

Ondřej Dušek is an Assistant Professor at Charles University in Prague, focusing on natural language generation (NLG) and dialogue systems. His recent research focuses on end-to-end neural NLG architectures, mostly applied to the data-to-text and dialogue response generation tasks. He is specifically interested in NLG semantic accuracy and semantic grounding, as well as ways of evaluating NLG accuracy. He co-authored more than 90 publications on NLG, dialogue, machine translation or speech synthesis. After obtaining his PhD in Prague, he spent 2 years as a postdoc at Heriot-Watt University in Edinburgh in 2016-2018, where he co-supervised a 2x Amazon Alexa Prize chatbot competition finalist team. He is currently the PI of the NG-NLG (Next-Generation Natural Language Generation) ERC Starting Grant, which aims to constrain neural models and combine them with knowledge graphs and semantic representations in order to produce fluent, accurate and explainable NLG systems.

Contact: info@praguecomputerscience.cz
Information: www.praguecomputerscience.cz