UBIQUITOUS KNOWLEDGE PROCESSING

Master Thesis

TECHNISCHE UNIVERSITÄT DARMSTADT

Building a Corpus of Creative Paraphrases

Motivation

Creative language is ubiquitous in everyday life, and while deep learning models are adept at handling many types of speech, many aspects of creativity remain challenging. This is largely due to the lack of sufficient data: creative components such as metaphor, irony, humor, and sarcasm are notoriously difficult to annotate, and thus current resources are insufficient. This work aims to explore data collection for creative data, the necessity of good data sources, and methods for overcoming difficulties in creative data collection.

Task Description

- Build a fundamental theoretical understanding of different types of creative data: metaphors, idioms, irony, sarcasm, and humor.
- Develop annotation methodology for annotation creative data [2,3], as well as explore methods for generating new, novel creative paraphrases.
- Use crowd-based services such as Amazon Mechanical Turk to build large, high-quality datasets.
- Apply state-of-the-art methods: can we advance the field by providing better data? sources?

References

- [1] G. Lakoff, M. Johnson. Metaphors We Live By. 1987
- [2] G. Steen et al. A method for linguistic metaphor identification: From MIP to MIPVU. 2010
- [3] T. Chakrabarty, D. Ghosh, A. Poliak, S. Muresan. Figurative Language in Recognizing Textual Entailment. ACL Findings, 2021

Analysis ■■■■■

Programming ■■■□□

Literature ■■■■□

Contact

Prof. Dr. Iryna Gurevych

Kevin Stowe, PhD.

thesis@ukp.informatik.tu-darmstadt.de