Seminar Topics: Information Extraction

English topics!

Viktor Hangya

hangyav@cis.lmu.de

Emerging and Complex Named Entity Recognition

- Irregular named entities:
  - complex noun phrases: `Eternal sunshine of the spotless mind` (movie)
  - ambiguous entities: `Among us` (can be a video game in some contexts)
  - emerging entities: `so... kktny in 30 mins?` (unseen entities)
- Project:
  - Discuss the difficulties of recognizing emerging and complex NE.
  - What techniques can solve these issues?
  - Is the problem solved or are there still difficulties?
- Resources:
  - Besnik et al., 2023, **Semeval-2023 task 2: fine-grained multilingual named entity recognition (multiconer 2)** *Proceedings of the 17th international workshop on semantic evaluation*
  - Derczynski et al., 2017, **Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition** *Proceedings of the 3rd Workshop on Noisy User-generated Text*

Information Extraction from Language Models

- Extract information from trained neural language models instead of texts.
- Language model training:
  - `I went to a concert Saturday [?]` → `night`
- Although they aren't directly trained to perform IE, they can answer questions or recognize entities:
  - `Who is the president of the United States? [?]` → `Joe Biden`
  - `Who is the president of the` **`[United States]`**^{LOC}`?`
- Project:
  - What kind of information is learned by pre-trained language models?
  - How can we extract them?
  - Is it superior to traditional techniques?
- Resources:
  - Liu et al., 2019, **Linguistic Knowledge and Transferability of Contextual Representations** *Proceedings of NAACL-HLT 2019*
  - Petroni et al., 2019, **Language Models as Knowledge Bases?** *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*
  - Cao et al., 2021, **Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases** *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

Political Framing and Persuasion Detection

- Manipulating people's view by presenting topics in certain ways:
  - Attack on reputation: name calling, casting doubt, etc.
  - Distraction: diverting the attention, accusing people without disproving their argument
  - Manipulation: emotionally loaded language, exaggeration, etc.
- Project:
  - How can the task be formulated (subtasks)?
  - Present a few models.
  - What issues should future work solve?
- Resources:
  - Piskorski et al., 2023, **SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup** *Proceedings of the 17th international workshop on semantic evaluation*
  - Cabot et al., 2017, **The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse** *Findings of the Association for Computational Linguistics: EMNLP 2020*

Similar Document Mining

- Multiple use cases:
  - Redundant content: same question multiple times on forums
  - Plagiarism, AI generated content without citation
  - Same content in different languages: valuable resource for machine translation
- Project:
  - Focus on document embedding based approaches
  - Discuss the monolingual case: paraphrase detection
  - Discuss the cross-lingual case: parallel sentence detection
- Resources:
  - Gao et al., 2021, **SimCSE: Simple Contrastive Learning of Sentence Embeddings** *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*
  - Duquenne et al., 2023, **SONAR: Sentence-Level Multimodal and Language-Agnostic Representations** *arXiv*
  - Wahle et al., 2022, **How Large Language Models are Transforming Machine-Paraphrased Plagiarism** *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*

Questions?

hangyav@cis.lmu.de