Information Extraction – Seminar Topics

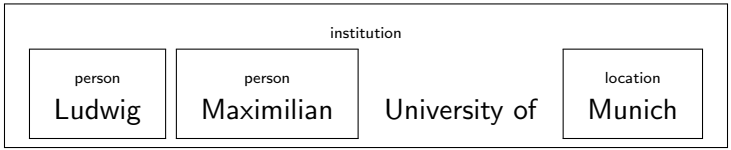
WS 2019–2020, English topics

Jindřich Libovický
libovicky@cis.lmu.de

October 17, 2019

Recognitions of nested entities

- Named Entites can overlap or be nested in each other

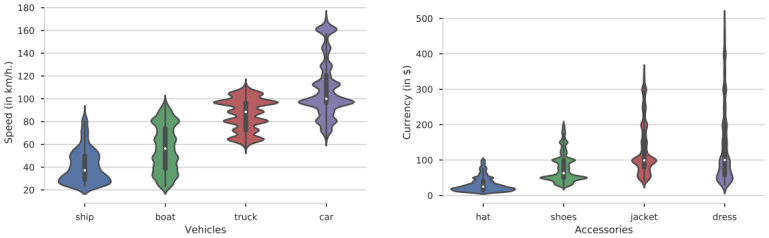


- How to adjust standard (LSTM-CRF) neural architectures to handle it?

Recommended Paper:
Straková, Jana, Milan Straka, and Jan Hajic. "Neural Architectures for Nested NER through Linearization." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
`https://www.aclweb.org/anthology/P19-1527.pdf`

Extracting word attributes from large amount of text

- Texts contains a lot of statements about entities–usually true or at least plausible
- With a use of huge amounts of text, NLP tools and simple statistics, we can get distributions over qualitative properties can be extracted



Recommended Paper:
Elazar, Yanai, et al. "How Large Are Lions? Inducing Distributions over Quantitative Attributes." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019
`https://www.aclweb.org/anthology/P19-1388/.pdf`

Fake Reviews Detection

- Internet is full of fake reviews of products, restaurants, ...
- Humans are not particularly good in detecting what reviews are fake
- Metadata + textual features (pre-trained embeddings, BERT) can come to rescue

Recommended Paper:
Kennedy, Stefan, et al. "Fact or Factitious? Contextualized Opinion Spam Detection." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2019.
`https://www.aclweb.org/anthology/P19-2048.pdf`

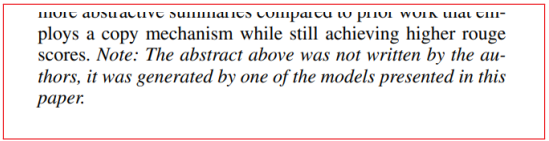
Depression and Self-Harm Risk Assessment in Online Forums

- People write various things on Reddit—e.g., that they depressed or have suicidal thoughts
- Can we from their previous posts predict they will have suicidal thought in the near future?
- There is a manually currated dataset collected from Reddit and NLP methods reaching 50% F_1 points

Recommended Paper:
Yates, Andrew, Arman Cohan, and Nazli Goharian. "Depression and Self-Harm Risk Assessment in Online Forums." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.
`https://www.aclweb.org/anthology/D17-1322.pdf`

Unsupervised Text Summarization

- Following recent success of very large language models (BERT, GPT-2, CTRL)
- Transofrmer as a language model can learn long-range dependencies, there is no need for task-specific architectures, just let the LM generate ...


more abstractive summaries compared to prior work that employs a copy mechanism while still achieving higher rouge scores. *Note: The abstract above was not written by the authors, it was generated by one of the models presented in this paper.*

Recommended Paper:
Subramanian, Sandeep, et al. "On extractive and abstractive neural document summarization with transformer language models." arXiv preprint arXiv:1909.03186 (2019).
`https://arxiv.org/abs/1909.03186`