# NLP+LLMs

#### Time: Friday, September 22, 2023 - 10:30 to 12:00

#### Chair: TBA

## **Talks**

### Knowledge-Grounded Target Group Language Recognition in Hate Speech

Hate speech comes in different forms depending on the communities targeted, often based on factors like gender, sexuality, race, or religion. Detecting it online is challenging because existing systems are not accounting for the diversity of hate based on the identity of the target and may be biased towards certain groups, leading to inaccurate results. Current language models perform well in identifying target communities, but only provide a probability that a hate speech text contains references to a particular group. This lack of transparency is problematic because these models learn biases from data annotated by individuals who may not be familiar with the target group. To improve hate speech detection, particularly target group identification, we propose a new approach that incorporates explicit knowledge about the language used by specific identity groups. We leverage a Knowledge Graph (KG) and adapt it, considering an appropriate level of abstraction, to recognise hate speech-language related to gender and sexual orientation. A thorough quantitative and qualitative evaluation demonstrates that our approach is as effective as state-of-the-art language models while adjusting better to domain and data changes. By grounding the task in explicit knowledge, we can better contextualise the results generated by our proposed approach with the language of the groups most frequently impacted by these technologies. This helps us examine model outcomes and the training data used for hate speech detection systems, and handle ambiguous cases in human annotations more effectively. Overall, integrating semantic knowledge in hate speech detection is crucial for enhancing understanding of model behaviors and addressing biases derived from training data.

| Paula Reyero LoboKnowledge Media Institute (KMi) | PhD. Candidate at the Knowledge Media Institute (KMi) in the Open University, and Early Stage Researcher under the Marie Sklodowska-Curie Innovative Training Networks (ITN) programme, "NoBIAS- Artificial Intelligence without Bias".  Her work titled "Knowledge-grounded Target Group Language Recognition in Hate Speech" proposes a novel KG adaptation paradigm for applying semantic knowledge to specific downstream tasks. The use case on hate target identification demonstrates that a hybrid learning architecture is key to enhance model transparency, robustness, and handling of annotation errors. |
| --- | --- |

### Using Pre-trained Language Models for Abstractive DBpedia Summarization: A Comparative Study

Purpose:

This study addresses the limitations of current short abstracts of DBpedia entities, which often lack a comprehensive overview due to their creating method (i.e., selecting the first two-three sentences from the full DBpedia abstracts).

We propose an abstractive summarization approach using pre-trained language models, which can generate concise and informative summaries.

Methodology:

We leverage pre-trained language models to generate abstractive summaries of DBpedia abstracts in six languages (English, French, German, Italian, Spanish, and Dutch).

We performed several experiments to assess the quality of generated summaries by language models. In particular, we evaluated the generated summaries using human judgments and automated metrics (Self-ROUGE and BERTScore).

Additionally, we studied the correlation between human judgments and automated metrics in evaluating the generated summaries under different aspects: informativeness, coherence, conciseness, and fluency.

Findings:

Pre-trained language models generate summaries more concise and informative than existing short abstracts. Specifically, BART-based models effectively overcome the limitations of DBpedia short abstracts, especially for longer ones. Moreover, we show that BERTScore and ROUGE-1 are reliable metrics for assessing the informativeness and coherence of the generated summaries with respect to the full DBpedia abstracts. We also find a negative correlation between conciseness and human ratings. Furthermore, fluency evaluation remains challenging without human judgment.

Value: This study has significant implications for various applications in machine learning and natural language processing that rely on DBpedia resources.

By providing succinct and comprehensive summaries, our approach enhances the quality of DBpedia abstracts and contributes to the semantic web community.

| Hamada ZaheraUniversity Paderborn | Hamada M. Zahera is a postdoctoral researcher at the Data Science Group (DICE) at Paderborn University, Germany, where he obtained his Ph.D. in Computer Science in 2023. He completed his B.Sc. and M.Sc. degrees of Computer Science from Menoufia University, Egypt. His research interests focus on natural language processing, knowledge graphs and semantic Web. He has prior experience working with the Technical University of Munich (Germany), XEROX Research Centre Europe (France), and EPFL (Switzerland). He also serves as a program committee member for several prestigious conferences: NeurIPS, ICLR, ESWC, ACL and EMNLP. |
| --- | --- |