Knowledge-based detection of Automatically Generated Text

Ph.D. proposal

Topic

In the last five years, the Natural Language Processing field has been profoundly transformed by the research on neural models: starting from the word embeddings introduced in 2013 by T.Mikolov, the innovations based on deep neural networks have proved more and more successful. One of the most successful aspects of this research work is constituted by neural language models and in particular neural language generation. Models such as GPT-2 (Radford et al., 2018) are able to produce text that can easily be mistaken for a human produced text (see https://talktotransformer.com/ for examples). This ability can be used for malicious purposes, such as plagiarism, writing fake product reviews, and so on (Adelani et al., 2019). A recent work by (Gehrmann et al., 2019) indicates that the accuracy achieved by humans on the detection of neural automatically-generated texts was about 54%. This work also paves the way to further research from a strict NLP perspective since the authors found out that some clues can be given by the lack of synonyms and anaphoras, and also regularities in the syntactic structure of the generated sentences.

The objective of this PhD Thesis is to design and test models for the automated detection of automatically generated texts exploiting NLP techniques. In particular, we are interested in answering questions such as: is the lexicon produced by generators poorer than a human produced one? Are syntactic structures less ambiguous than human produced ones? And in particular how does the knowledge included in a sentence produced by an automated system compare to the knowledge that we can found in human redacted knowledge bases?
Answering these questions involves the use of various methods, from statistical analysis to classification methods to machine reading, neural language models and other techniques based on NLP and/or the Semantic Web, in particular for knowledge extraction and representation. This makes the RCLN team at LIPN a good environment where to carry out this research.

Context

The successful candidate will integrate the RCLN team at the Laboratoire d'Informatique de Paris Nord in Paris (Université Sorbonne Paris Nord, Villetaneuse Campus), a team working on Natural Language Processing and Knowledge representation and extraction. The referents for this Ph.D. subject will be Davide Buscaldi, MCF and Thierry Charnois, PU.

Profile

Candidates must have good knowledge in Machine Learning models and techniques, hold a Master in Computer Science and have good Python programming skills. Knowledge of libraries such as Keras, PyTorch and Huggingface and specific knowledges and experience in Natural Language Processing are a bonus.

Bibliographie

- Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. arXiv preprint arXiv:1907.09177.
- Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. arXiv preprint arXiv:1906.04043.

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.