Challenges for NLP shared tasks in the chatGPT era

Julio Gonzalo[1]

[1]Universidad Nacional de Educación a Distancia, Madrid, Spain

1. Abstract

The generative AI tsunami has turned Natural Language Processing upside down, and evaluation challenges (such as the ones conducted in SemEval, Evalita and IberLEF) are no exception. Since the introduction of chatGPT less than ten months ago, we have seen technical papers claiming that it annotates better than crowdworkers for many types of tasks, and at the same time that crowdworkers may be using chatGPT to complete their assignments in a substantial proportion. The best generative models are routinely passing exams from university degrees, but we have also learned that contamination is a key issue when evaluating large language models: models have probably seen the answers of the exams. Everything is fascinating and confusing.

In the talk we will discuss this situation, and we will review some evaluation challenges that are equally relevant in the post chaGPT era. We will pay special attention to evaluation metrics, learning with disagreement scenarios and the problem of measuring the relative performance of large language models between languages.

2. Short Biography

Julio Gonzalo is director of the Research Center in Natural Language Processing and Information Retrieval and deputy Vicerrector of Research at UNED (Madrid, Spain). Along his career he has worked on topics such as online reputation monitoring, Information Access technologies for Social Media, interactive cross-language search, toxicity and misinformation in Social Media, computational creativity and semantic similarity. He has also worked extensively in the design and assessment of evaluation metrics for a wide range of Artificial Intelligence problems, which led to a Google Faculty Research Award (together with Enrique Amigó and Stefano Mizzaro) for his work in this area. He has recently been co-chair of ACM SIGIR 2022 and co-funder and co-chair of IberLEF (2019-2022), the annual evaluation campaign for NLP systems in Spanish and other Iberian languages. He is currently leading an initiative of the Spanish government to measure the gap in Artificial Intelligence between Spanish and English.

Personal Website. https://sites.google.com/view/nlp-uned/people/julio-gonzalo

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

julio@lsi.uned.es (J. Gonzalo)