Semi-Automatic Annotation of Conversations in Audio-Visual Documents

Most human interactions occur through spoken conversations. If this interaction mode seems so natural and easy for humans, it remains a challenge for spoken language processing models as conversational speech raises critical issues. First, non-verbal information can be essential to understand a message. For example a smiling face and a joyful voice can help detecting irony or humor in a message. Second, visual grounding between participants is often needed during a conversation to integrate posture and body gesture as well as references to the surrounding world. For example, a speaker can talk about an object on a table and refer to it as this object by designing it with her hand. Finally, semantic grounding between participants of a conversation to establish mutual knowledge is essential for communicating with each other.

In this context, the MINERAL project aims to train a multimodal conversation representation model for communicative acts and to study communicative structures of audiovisual conversation.

As part of this project, we are offering a 5- to 6-month internship focused on semi-automatic annotation of conversations in audio-visual documents. The intern's first task will be to extend the existing annotation ontology for dialog acts, currently available for audio documents (through the Switchboard corpus for example), to incorporate the visual modality. In a second step, the intern will develop an automatic process for transferring annotations to new audiovisual datasets (such as meeting videos and TV series or movies) using transfer or few-shot learning approaches.

Practicalities:

The internship will be funded ~500 euros per month for a duration of 5 or 6 months and will take place at LISN within the LIPS team. This internship can potentially be followed by a funded PhD, based on performance and interest in continuing research in this area.

Required Qualifications:

- Master’s degree (M2) in Computer Science or related field.
- Experience with deep learning frameworks such as Keras or PyTorch.
- Knowledge of image processing would be an advantage.

To apply, please send your CV, a cover letter and your M1 and M2 transcripts (if available) by email to Camille Guinaudeau camille.guinaudeau@universite-paris-saclay.fr and Sahar Ghannay sahar.ghannay@universite-paris-saclay.fr

References:

[Albanie, 2018] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In Proceedings of the 26th ACM international conference on Multimedia. 2018

[Zhang, 2021] Sheng Zhang, Min Chen, Jincai Chen , Yuan-Fang Li, Yiling Wu, Minglei Li, Chuanbo Zhu. Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition. Knowledge-Based Systems. 2021.

[Fang, 2012] Alex C. Fang, Jing Cao, Harry Bunt and Xiaoyue Liu. The annotation of the Switchboard corpus with the new ISO standard for dialogue act analysis. Workshop on Interoperable Semantic Annotation. 2012.