Internship call – Studying text representations of spontaneous speech transcripts in different speaking contexts

Place of work **Paid internship** at Telecom-Paris, Palaiseau (Paris outskirt).

Duration 5-6 months

Starting date ASAP, from 1st February 2021.

Context This work is part of the MSCA ITN - Marie Sklodowska-Curie Actions project ANIMATAS (`animatas.eu`) part of European Union's Horizon 2020 research and innovation programme under grant agreement No 765955. The intern will work with doctoral student Tanvi Dinkar and Prof. Chloé Clavel of the S2a [SSA] team at Telecom-Paris at the social computing [SocComp] lab.

Candidate profile

- → **Technical skills:** Python, solid theoretical and practical (eg. Pytorch, TensorFlow) background in machine learning. Basic knowledge of NLP methodology is appreciated, but not required.
- → Proficiency in English is a must.
- → The candidate should be passionate to work on an interdisciplinary field, i.e both in the implementation of state-of-the-art (SOTA) NLP as well as the in the analysis of language data.
- → **Note** This is a research internship and candidates should be willing to read papers and collaborate with a doctoral student. This work may lead to further scientific publications.

How to apply Please submit pdf files of your CV to:
tanvi.dinkar@telecom-paris.fr and
chloe.clavel@telecom-paris.fr.
If your application is selected, you will then be contacted for further information and interview details.

About the position People rarely write in the same manner with which they speak. When speaking, people tend to repeat themselves, interrupt each other (eg."Uh-huh"), speak in ungrammatical sentences, and are in general, disfluent. *Disfluencies* are defined as interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something said previously [1]. *Fillers* (eg. "uh" and "um" in English) are a kind of disfluency that fill a pause in an utterance or conversation.

Within the Social computing team, you will implement SOTA language models to see how they perform on spontaneous speech transcripts in various speaking contexts, specifically looking at disfluencies and fillers. The intern will learn about NLP tasks, specifically language modelling, and get practical exposure

to the training of NLP systems (using for example, the Hugging Face library and Pytorch). The intern also will learn about working with well known NLP datasets and the nuances of working with speech transcripts.

Concretely, the objectives are:

→ To design and implement a system that trains SOTA pre-trained language models on various speaking contexts.

→ To evaluate the representations learnt of disfluencies and fillers, by using for example, systems like VisualBERT.

→ To work with a doctoral student to design an evaluation protocol for more in depth analysis of representations of spontaneous speech phenomena.

→ To produce open source and reproducible code that will be associated with the project and possible paper.

Suggested reading

→ BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [2].

→ The importance of fillers for text representations of speech transcripts [3].

→ What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models [**?**].

→ Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data [4].

→ Hierarchical pre-training for sequence labelling in spoken dialog [5].

→ `https://ruder.io/research-highlights-2020/` section 5: Evaluation beyond accuracy.

Bibliography

[1] Scott H. Fraundorf, Jennifer Arnold, and Valerie J. Langlois. Disfluency, 2018. URL https://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0189.xml.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[3] Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. The importance of fillers for text representations of speech transcripts, 2020.

[4] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proc. of ACL*, 2020.

[5] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*, 2020.