Internship: joint speech segmentation and syntactic analysis

Advisors: Alexis Nasr & Benoit Favre
LIS/CNRS, Aix-Marseille University

Spring-Summer 2022

1 Description

Context Segmenting texts into sentences is a standard task in natural language processing which does not pose great difficulty, in particular thanks to punctuation marks at the end of sentences (Read et al., 2012). The situation is more complex in the case of speech transcriptions where punctuation is generally absent, in particular if they are the result of automatic speech recognition. The segmentation process can be performed on the sole basis the word sequence, but the results usually are not very good[1] (Żelasko et al., 2018). In order to improve over lexical-only models, one can add prosody (in the form of F0, energy, pause duration...) and syntax (Favre et al., 2008). There is a chicken and egg problem in adding syntactic features to segmentation as syntactic parsers cannot handle unsegmented inputs and segmenting speech requires the result of parsing.

Objectives The goal of this internship is to develop a joint model of syntactic parsing and sentence segmentation for spoken recordings, based on lexical and prosodic features. The problem of the vicious dependency cycle between syntactic parsing and segmentation can be handled by using online transition-based parsing which does not assume a sentence boundary, as proposed for example in (Nasr et al., 2020). Compared to traditional transition-based parsing, this kind of parser adds a special transition for predicting sentence boundaries which flushes the current tree and starts a new one. In this context, speech-derived features, such as prosody, could be added to the classifier to inform its segmentation decisions. A potential benefit is that speech features might also help with predicting syntactic structures in addition to performing more accurate segmentation.

Scientific program The work will be carried out on a corpus of speech transcriptions annotated with syntactic trees, such as for example the data from the ORFEO project. First, speech-derived features, such as F0, energy and pause duration will be extracted using a standard toolkit. Then, the parser model will be adapted to handle this new source of information (Dary and Nasr, 2021). The resulting system will be trained jointly to perform both syntactic parsing and segmentation, and evaluated on both tasks.

Different ways of extracting speech features, such as simple features from kaldi[2], more advanced representations from OpenSmile Eyben et al. (2010) or unsupervised pre-trained representations such as huBert (Hsu et al., 2021), will be evaluated.

Different models for integrating speech features will also be compared.

2 Additional information

- Skills: Master-level computer science, an interest for linguistics, python programming, deep learning, Pytorch, rigor and tenacity.
- Location: the internship will take place at LIS/CNRS on the Luminy campus of Aix-Marseille University.

[1] See results reported at `https://github.com/benob/recasepunc`
[2] https://kaldi-asr.org/

- Dates: Spring-summer 2022, duration 5-6 months.
- Wages: regulatory internship salary (about 500 euros/month).
- Computation: the intern will have access to the Jean-Zay GPU cluster for running experiments.

References

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, 2012.

Piotr Żelasko, Piotr Szymański, Jan Mizgajski, Adrian Szymczak, Yishay Carmiel, and Najim Dehak. Punctuation prediction model for conversational speech. *arXiv preprint arXiv:1807.00543*, 2018.

Benoit Favre, Dilek Hakkani-Tur, Slav Petrov, and Dan Klein. Efficient sentence segmentation using syntactic features. In *2008 IEEE Spoken Language Technology Workshop*, pages 77–80. IEEE, 2008.

Alexis Nasr, Franck Dary, Frédéric Bechet, and Benoît Fabre. Annotation syntaxique automatique de la partie orale du orféo. *Langages*, (3):87–102, 2020.

Franck Dary and Alexis Nasr. The reading machine: a versatile framework for studying incremental parsing strategies. In *The 17th International Conference on Parsing Technologies*, 2021.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*, 2021.