Internship : syntactic analysis of speech without transcription

Advisers: Alexis Nasr, Ricard Marxer & Benoit Favre
LIS/CNRS, Aix-Marseille University

Spring-Summer 2022

1 Description

Context Syntactic analysis, or syntactic parsing, consists in predicting a tree representation of the syntactic relationship between words of a sentence. A range of paradigms and methods have been proposed over the years for solving that task (see among others the methods presented by Zhang (2020)). Beyond text, parsing speech recordings is an important task for developing pervasive applications with spoken interactions (Tur and De Mori, 2011; Damonte et al., 2019; Tran and Ostendorf, 2021). It is also very difficult because of two main reasons: idealized models of language that were developed for text do not apply completely to speech, and automatically generated transcriptions are not devoid of errors. The later is problematic because syntax is deeply linked to the representation of linguistic content as a sequence of words.

Objectives The goal of this internship is to reconsider this axiom: what if we could perform syntactic parsing of speech recordings without relying on an explicit transcription. This study will explore an alternate representation of the speech signal as a sequence of automatically extracted symbols representing sub-lexical units. The extraction of these units will be performed using discrete representations learned from audio signal, such as the VQ-WAV2VEC model (Baevski et al., 2019).

The quantized speech segments will be fed to a transition-based parser, that typically considers attaching the current word to a partial syntax tree, with additional transitions that accumulate sub-lexical units to form tokens. Such parser can be trained with regular procedures for transition-based parsing (Dary and Nasr, 2021; Nivre, 2013).

Scientific program In order to learn a syntactic analyzer able to input this type of representations, we will use the ORFEO corpus[1] (Benzitoun et al., 2016). That corpus is composed of transcriptions of speech recordings annotated with syntactic analyses. The speech signal and alignments to the word transcripts are also available (when a word begins and ends in the speech signal). The idea is to map the speech signal to discrete units using a model such as VQ-WAV2VEC mentioned above and to project the syntactic annotations on sequences of such symbols. At the end of this step, it is possible to train a syntactic analyzer which inputs sequences of symbols originating from VQ-WAV2VEC and outputs a dependency tree derived from a sequence of transitions. This process can be divided into the following steps:

1. Learning discrete representations from a large set of speech recordings such as Mozilla common voice (Ardila et al., 2019), the EPAC corpus (Esteve et al., 2010) or the non-annotated part of the Orfeo corpus

2. Extract these representations on the part of Orfeo with syntax annotations

3. Transfer syntax annotations from words to the discrete representations

4. Create a new transition system to allow for sub-lexical units

5. Train and evaluate a dependency parser in those conditions

[1] https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html

In addition, it will be interesting to explore the possibility to self-train a parser on large quantities of unannotated speech transcripts, and to explore variations of the discretization strategies and transition systems.

2 Additional information

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

Meishan Zhang. A survey of syntactic-semantic parsing based on constituent and dependency structures. *Science China Technological Sciences*, pages 1–23, 2020.

Gokhan Tur and Renato De Mori. *Spoken language understanding: Systems for extracting semantic information from speech.* John Wiley & Sons, 2011.

Marco Damonte, Rahul Goel, and Tagyoung Chung. Practical semantic parsing for spoken language understanding. *arXiv preprint arXiv:1903.04521*, 2019.

Trang Tran and Mari Ostendorf. Assessing the use of prosody in constituency parsing of imperfect transcripts. *arXiv preprint arXiv:2106.07794*, 2021.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

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.

Joakim Nivre. Transition-based parsing. *Uppsala universitet*, 2013.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15), 2016.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Yannick Esteve, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. The epac corpus: Manual and automatic annotations of conversational speech in french broadcast news. In *LREC*. Citeseer, 2010.