Multi-Level Chained Transformer for Joint Automatic Speech Recognition and Subtitle Generation

Jakob Poncelet

KU Leuven, Department Electrical Engineering ESAT-PSI

Hugo Van hamme

KU Leuven, Department Electrical Engineering ESAT-PSI

Automatic Speech Recognition (ASR) is growing rapidly, powered by an increasing amount of data and computing resources and recent advancements in deep learning. Unsupervised pre-training on raw unlabeled audio, as in Wav2Vec 2.0, has drastically improved the speech recognition performance for many languages, although supervised finetuning with labeled speech remains necessary. On the other hand, the recent state-of-the-art Whisper model has shown great performance with supervised training on huge amounts of multilingual web-scraped audio and transcript pairs, although very tightly curated and filtered. The scale of the model and the required resources also has many drawbacks.

For Dutch, and many other languages, labeled ASR data remains rather scarce. On top of that, speech recognition datasets often only cover a small domain of read and prepared speech, and models trained on these datasets have difficulties to generalise to real-life conversational and spontaneous speech. One type of data that would solve these problems is subtitled data on TV. TV subtitles are abundantly present in Dutch, manually annotated, and span a very broad domain of speech, with news broadcasts, interviews, talkshows and soaps, hereby covering spontaneous speech and even dialectal speech. However, subtitles on screen should be regarded as weak supervision, as they are not true verbatim (exact) transcriptions, but contain many edits, corrections and rephrasings to improve readability and understanding, which render them ineffective for direct ASR training.

In this work, we expand on previous work [1] that used subtitled data to improve an end-to-end ASR model for Flemish (Belgian Dutch). By solving verbatim and subtitle transcription as two separate tasks with parallel decoders, jointly trained on ASR and subtitled data, it is shown that the inexact subtitle transcriptions can improve a shared speech encoder on spontaneous speech without deteriorating the verbatim ASR branch. On top of that, the model delivers both a verbatim transcription and a cleaned-up subtitle transcription.

We propose a cascaded multi-level Transformer model that consists of an ASR encoder-decoder chained to a subtitle encoder-decoder. The entire model is trained end-to-end on both data streams. The subtitle encoder can process either the hidden features generated by the ASR encoder or decoder. The subtitle decoder is a Multi-Transformer that utilises double cross-attentions to both the subtitle encoder and ASR encoder features. In contrast to the parallel model, the cascaded model can leverage the verbatim hypotheses and transform them into subtitle hypotheses, similar to machine translation, with internal features. Furthermore,

the subtitled data is backpropagated through the ASR branch, and the additional subtitle encoder allows separate CTC losses for both data streams without conflicting in domain.

We perform experiments with an improved Conformer-based ASR model, trained end-to-end with hybrid CTC/Attention. While the parallel model improves the ASR performance on the spontaneous track without worsening the performance on the clean track, the cascaded model improves on both tracks and strongly outperforms the baselines, while showing improved subtitling capabilities.

[1] Poncelet, J., & Van hamme, H., “Learning to jointly transcribe and subtitle for end-to-end spontaneous speech recognition”, SLT 2022.