ParCzech 3.0 (ASR corpus)

Vladislav Stankov

ParCzech 3.0 (ASR part)

- Speech corpus of the Czech parliamentary speeches from The Czech Chamber of Deputies
- Stenographic protocols From 25th November 2013 to 1st April 2021

Basic information

- 20 257 mp3 files, each 14 minutes long
  - Originally about 4726 hours, more than 6 months …
- For each mp3 files there is list of its speakers
- Aligned sound with stenographic protocols
- Difficulties
  - Protocols are not exactly precise
  - Sound files have “safety” offset on the both ends
  - Sound files contain a lot of gaps (mainly created or just noise with no speech)

Solution

- GMM-based speech recognition to extract timing for individual words
    - != force alignment
    - Returns “recognized words” with timings
    - Some words may be badly recognized/misheard
    - Occasionally throughs out segments of length more than 1 minute …
        - Partially solvable by repeated run

- Align recognized words with stenographic transcripts to get timings for words
    - Global alignment with affine gap penalties (modified to work on words + parameter optimization)
    - Manually fix problems when word was misheard as group of words

Solution continued

- Split stenographic protocols into segments (mainly sentences)
    - Sometimes sentences are merged if the ending time of the sentence is ambiguous
- Create different statics on the segments (and corresponding sound parts) to filter out bad ones
    - Statics based on edit distance and missed words
    - Statistics based on the sound
- Clean the data based on statistics
    - Prefere data without noise for ASR training
- Create different test and dev sets with common train set
    - Based on speakers
    - Based on the original mp3 files
    - Random split

Results

- Data size after alignment (no filtering)
  - 3 071 hours (~ 65%, silence and offset reduced)
  - 1 391 785 segments
  - 22 153 778 words
- Clean data
  - About 43 % from the aligned corpus
  - 1 332 hours
  - 606 540 segments
  - 10 146 591 words
  - Min duration is 0.85 seconds and max is 54 seconds, avg duration is 8 seconds