Université de Lorraine - IDMC
Masters SC & TAL – M1
2019-2020

Fiche de projet tutoré / Project form

Detecting Fillers in Conversational Speech

Encadrement / Supervisors

1. équipe, laboratoire / team, lab
   Multispeech
2. encadrant · e principal · e (nom, email) / main supervisor (name, email)
   Imran Sheikh (imran.sheikh@inria.fr)
3. autres encadrant · es / other supervisors

Description / Description

Conversational dialog systems rely on Automatic Speech Recognition (ASR) systems. These ASR systems are trained on large amounts of conversational speech data comprising spoken utterances and their text transcriptions. Obtaining such human transcribed conversational speech corpora incurs a huge cost. Crowd sourced read speech corpora are becoming a popular alternative to build ASR systems [1], which can also automatically transcribe collections of conversational speech [2]. However, read speech corpora as well as the ASR models built on them lack some specific attributes of spontaneous conversational speech [3,4]. For instance, spontaneous utterances often contain fillers, like 'uhh' and 'umm', which will be mapped to other vocabulary words by an ASR trained on read speech. In order to address these shortcomings, a smaller amount of human transcribed conversational speech data can be utilized to bootstrap ASR for conversational speech.

This project will study methods for automatically identifying and annotating fillers in several hours of un-transcribed conversational speech data, given a small amount of human transcribed subset. Automatic detection of fillers and other disfluencies in spontaneous speech has been studied previously. The proposed approaches are typically based on sequence classifiers trained on features extracted from the speech signal or from the output of an ASR system [5,6]. However, different datasets and setup have been used to report their performance. Moreover, the effect of using the detection results for bootstrapping ASR for conversational speech has not been studied. Improvement in ASR for conversational speech is final objective of this project. This project will be a part of the bigger study on 'Weakly Supervised Learning for Conversational ASR' under the COMPRISE project [7].

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

Experiments will be conducted on the English subset of the Verbmobil corpus [8] of human-human dialogs. The corpus consists of about 30 hours of human transcribed speech. The corpus will be divided into subsets for training filler detector and ASR models, as well as the validation and test sets. Automatic transcriptions of Verbmobil corpus will be prepared

beforehand, using an English ASR pre-trained on several hundred hours of read speech [3].

The project will go through the following stages:
- A bibliography on detecting fillers and disfluencies in spoken conversations.
- Analysis of errors made by read speech ASR in filler regions of the Verbmobil corpus. (Scripts for obtaining ASR errors with respect to groundtruth will be made available.)
- Train a filler detector model based on features extracted from the read speech ASR output. Existing tool to extract features from ASR output and to detect errors in ASR output, developed as part of the comprise project, will be used in this task.
- Train a filler detector model based on features extracted from the speech signal. Open-source feature speech extractor openSMILE [5] and a sequence classifier built in Scikit-learn [9] or Pytorch library [10] will be used for this task.
- Use the filler detector models to introduce fillers in the automatic transcriptions of Verbmobil corpus.
- Analysis on the (minimum) amount of annotated conversational speech data required for effective learning of fillers.
- Report improvements in ASR trained on automatic transcriptions of Verbmobil corpus. (Students are not expected to conduct experiments on re-training ASR.)

Livrables et échéancier / Deliverable and schedule

October-December: Reading on related work, methods and tools. Prepare bibliography.
January: Introduction to corpus and tools.
February-March: Experiments on training and evaluating the filler detection model.
April-May: Compilation and analysis of the results, report writing and defense.

Bibliographie /References

[1] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 5206-5210.
[2] P. Zhang, Y. Liu and T. Hain, "Semi-supervised DNN training in meeting recognition," 2014 IEEE Spoken Language Technology Workshop (SLT), 2014, pp. 141-146.
[3] M. Benzeghiba et. al., "Automatic speech recognition and speech variability: A review", Speech Communication, Volume 49, Issues 10–11, 2007, Pages 763-786.
[4] K. Aono, K. Yasuda, T. Takezawa, S. Yamamoto and M. Yanagida, "Analysis and effect of speaking style for dialogue speech recognition," 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, St Thomas, VI, USA, 2003, pp. 339-344.
[5] R. Brueckner and B. Schulter, "Social signal classification using deep blstm recurrent neural networks," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 4823-4827.
[6] M. Lease, M. Johnson and E. Charniak, "Recognizing disfluencies in conversational speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1566-1573, Sept. 2006.
[7] COMPRISE project, https://www.compriseh2020.eu
[8] A. Kurematsu et. al., "VERBMOBIL dialogues: multifaced analysis", Sixth International Conference on Spoken Language Processing, Beijing, 2000, pp. 712-715.
[9] Scikit-learn: Machine Learning in Python, https://scikit-learn.org/stable/index.html
[10] PyTorch Machine Learning Framework , https://pytorch.org