2018-00725 - Leveraging Noisy, Incomplete, or Implicit Labels for Automatic Speech Recognition

Level of qualifications required: Graduate degree or equivalent
Function: PhD Position

Context
This PhD position is in collaboration between:
- Emmanuel Vincent (emmanuel.vincent@inria.fr), Multispeech team, Inria (http://team.inria.fr/multispeech), Nancy
- Armelle Brun (armelle.brun@loria.fr), Kiwi team, Loria (http://kiwi.loria.fr/), Nancy
- A third advisor working for a large company

The selected candidate will be recruited by the company and he/she will be expected to spend about half of his/her time at Inria/Loria and the remaining half in the company (partially abroad).

Assignment
Great progress has been made in automatic speech recognition over the last decade [3, 7]. This is due to the maturity of machine learning techniques (e.g., end-to-end deep learning), the increase in computational power and, crucially, the availability of huge labeled datasets in the order of tens of thousand hours and speakers for supervised learning. Labeling consists of transcribing each utterance as a sequence of words. This is a time-consuming process, that incurs a huge cost and limits the availability of commercial speech recognition interfaces to about 100 languages, compared to the 7,000 languages currently spoken in the world.

Several attempts have been made to move beyond the current supervised learning paradigm. Semi-supervised learning leverages both labeled and unlabeled data, and can improve performance at no extra labeling cost. Active learning aims to automatically identify among a larger pool of unlabeled data those to be labeled in priority by a human. These two approaches have been exploited in combination, resulting in a transcription cost reduction of about 70% [9, 2]. Transfer learning (a.k.a. model adaptation) has also been explored [6], particularly for cross-language learning, but it reaches a ceiling due to the differences between languages. Overall, these
approaches are useful but there is still a long way toward achieving the desired cost-utility tradeoff and expanding the practical applicability of current speech recognition technology to a thousand or more languages.

The goal of this PhD is to explore weakly supervised learning approaches, that would exploit more information than semi-supervised learning, at little or no extra labeling cost. Examples of weak labels that could be exploited include, by decreasing order of human involvement:

1. erroneous labels — labelers are asked to work quickly and hence can make mistakes on certain words,
2. utterance-level labels — labelers are given an automatic transcript and asked to check, e.g., whether it is globally correct or whether the meaning is globally correct,
3. incomplete labels — labelers can refuse to label certain utterances, which indirectly informs about these utterances, e.g., they may be too complex or unintelligible,
4. implicit labels — the behavior of users interacting with the voice command system indirectly informs about the success of speech recognition, e.g., if a user utters the same command twice in a row, then it was probably misrecognized in the first place,
5. transformed labels — third-party close captioned data are used, which are summaries of the exact transcripts resulting from a more or less systematic summarization method.

The amount and distribution of noise in the labels, the reason for not labeling certain utterances, and the behavior in response to certain errors are specific to each labeler or user. Hence, training a weakly supervised speech recognition system on this data amounts to learning user behavior and specificities at the same time as learning the speech recognition task.

To address these issues, the PhD student will develop new methods at the intersection of speech recognition, weakly supervised learning and labeler/user modeling in crowdsourcing applications [10, 5, 8]. Results will be evaluated on a combination of existing close captioned corpora [1], simulated labeling errors from a supervised corpus (e.g., [4]), and real crowdsourcing experiments.

Main activities

The PhD student will develop new methods at the intersection of speech recognition, weakly supervised learning and labeler/user modeling in crowdsourcing applications [10, 5, 8]. Results will be evaluated on a combination of existing close captioned corpora [1], simulated labeling errors from a supervised corpus (e.g., [4]), and real crowdsourcing experiments.

Skills

MSc in machine learning or computer science

Experience with Python programming language