Ph.D. Position in Computer Science

Automatic analysis of errors in automatic speech recognition systems from end-users reception

LIA - Avignon University (France)

Main laboratory: Laboratoire Informatique d'Avignon (LIA – EA 4128), France

Duration: 3 years - **Start time:** September 2021 - **Closing date:** 31 March 2021

Salary: 1500 euros/month (approximately)

Title in French: Analyse automatique des erreurs des systèmes de reconnaissance automatique de la parole par la réception des utilisateurs finaux.

Project context: This Ph.D. position is part of the French research project DIETS (Automatic diagnosis of errors of end-to-end speech transcription systems from users perspective) funded by the ANR (French National Research Agency) which aims at analyzing finely recognition errors by taking into account their human reception, and understanding and visualizing how these errors manifest themselves in an end-to-end ASR framework. The main objectives are to propose original automatic approaches and tools to visualize, detect and measure transcription errors from the end-users perspective.

Candidate profile: The applicant must hold a Master degree in Computer Science. Mastery of at least one common object programming language (Java, C++...) and one scripting language (Python, Perl…) are mandatory, furthermore experience in automatic language and speech processing, or machine learning, data mining are appreciated. He or she should also show interest in linguistics and the study of human behavior.

Interests for the candidate:
- Very favorable and collaborative work environment in an internationally recognized research laboratory in language processing and machine learning.
- Implementation, analysis and proposals for innovative approaches to different ASR systems (classical and end-to-end frameworks).
- Development of complementary metrics to WER that are user-oriented.
- Transdisciplinary scientific work allowing openness to other disciplines (e.g. linguistics and cognitive sciences).

Applications should be sent to:
- Richard Dufour (richard.dufour@univ-avignon.fr) - LIA, Avignon University
- Jane Wottawa (jane.wottawa@univ-lemans.fr) - LIUM, Le Mans University
and should include:
- a detailed CV (education and research experiences),
- a cover letter specifying the candidate's research interests on this proposed Ph.D. thesis,
- Bachelor (Licence) and Master grades in detail,
- at least one reference that could be contacted for recommandation.

Keywords: Automatic transcription, error analysis, end-to-end speech recognition, perceptive tests, machine learning.

General context

Natural language processing (NLP) systems are now reaching a level of performance that many of them are now accessible to the general public, such as machine translation (MT), automatic speech recognition (ASR), document indexing... Despite this technological maturity, automatic systems are inevitably prone to errors. Although imperfect and subject to discussion, the metrics usually making consensus are those being the easiest to apply widely and not requiring additional human intervention to evaluate a new system, e.g. the word-error rate (WER) in speech recognition. We now build NLP systems that minimize errors in a task regarding a considered reference, without trying to know, to understand, and to evaluate the impact of this error from a final user perspective. The impact of these errors made by automatic systems on humans, and the way in which they are perceived, is almost never evaluated. The analysis, and the evaluation metrics are only system-oriented instead of human-oriented, although these systems are designed to model human language for human users.

The classical ASR pipeline integrates modules each with a well-defined role (acoustic model, language model...). In recent years, deep learning allowed computational models with many processing layers in order to learn data representations corresponding to different levels of abstraction [LeCun15]. Recently, a great interest in deep neural end-to-end architectures developed for various NLP tasks, including ASR [Amodei16, Tomashenko19...] emerged. In end-to-end approaches, multiple levels of abstraction (acoustic, phonetic, lexical, syntactic...) can be incorporated into a single neural network model. In ASR, end-to-end models attempt to map an acoustic signal to a word (or phone/character) sequence directly using neural network models. Error analysis of ASR systems is not a new challenge. The automatic detection of errors in transcription systems has been studied for many years, such as in [Ghannay15, Errattahi18]. Although the approaches differ, a majority starts from the observation that transcription errors must be treated in their globability (binary problem) in order to correct them, aiming to improve the tasks in which the transcripts are then used, ignoring the nature of the error and its impact on end-users. Also, while classical ASR pipelines have been widely studied, analysis of end-to-end ASR systems' errors and a qualitative analysis of their errors is still in its early stages, information (acoustic, linguistic…) contained and conveyed in end-to-end ASR systems being difficult to identify (*deep* architectures).

It is a question of putting the human back at the heart of the system, on the one hand to better understand the ASR systems and their errors, and on the other hand to be aware of the flaws and limits of the systems proposed from the final user point-of-view.

Objectives

The main objective of the thesis is to finely analyze transcription errors from the point of view of their reception by the user. The thesis will have three complementary parts:

1. Approaches for error detection in transcripts of end-to-end ASR systems. This should lead to original confidence measures.
2. Detailed analysis of transcription errors in French, whether human or automatic, with a traditional or end-to-end system, in order to understand how errors are viewed from a

human perspective. This will shed light on new classes of errors, guided by their difficulty, or ease, to be understood by end users.

3. Realization of a new body of automatic transcriptions where errors are annotated using precise linguistic information, and information collected during perceptual tests to reflect how users perceive (and possibly correct) these errors. Carrying out different perceptual tests, by confronting humans with these transcription errors.

It will be a question of laying the first bases of a new and transversal research, at the crossroads between linguistics, computer science and cognitive sciences, for the evaluation of automatic systems and the understanding of NLP systems based on deep architectures. The Ph.D. student will then have the opportunity to learn and propose innovative approaches in automatic speech processing for the understanding of architectures with deep neural networks, but also to have an openness and skills in linguistics and on the implementation of perceptual tests.

Références bibliographiques :

[Amodei16] Amodei D., Ananthanarayanan S., Anubhai R., Bai J., Battenberg E., Case C., ... Chen J. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In Inter. conf. on machine learning (pp. 173-182).

[Errattahi18] Errattahi, R., Deena, S., El Hannani, A., Ouahmane, H., & Hain, T. (2018). Improving ASR Error Detection with RNNLM Adaptation. In IEEE SLT (pp. 190-196).

[Ghannay15] Ghannay, S., Esteve, Y., & Camelin, N. (2015). Word embeddings combination and neural networks for robustness in asr error detection. In EUSIPCO (pp. 1671-1675).

[LeCun15] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[Mdhaffar19] Mdhaffar, S., Estève, Y., Hernandez, N., Laurent, A., Dufour, R., & Quiniou, S. (2019). Qualitative evaluation of ASR adaptation in a lecture context: Application to the PASTEL corpus. In Interspeech, 569-573.

[Tomashenko19] Tomashenko N., Caubrière A., Estève Y. (2019). Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech. In Interspeech.