PROPOSITION SUJETS DE THESES

CONTRATS DOCTORAUX 2020-2023

Appel ciblé (merci de cocher la case correspondante):

x Contrat doctoral ministériel ED 536

□ **Contrat doctoral ministériel ED 537**

Directeur de thèse : Richard Dufour (richard.dufour@univ-avignon.fr)

Co-directeur : Jean-François Bonastre (jean-francois.bonastre@univ-avignon.fr)

Co-encadrant éventuel : Jane Wottawa (jane.wottawa@univ-lemans.fr)

Titre en français : Analyse automatique des erreurs des systèmes de transcription automatique de la parole par la réception des utilisateurs finaux

Titre en anglais : Automatic analysis of errors in automatic speech transcription systems by reception of end users

Mots-clés :

Co tutelle : ~~Oui~~ - Non **Pays :** -

Opportunités de mobilité à l'international du doctorant dans le cadre de sa thèse : Oui - ~~Non~~

Profil du candidat : informaticien avec attrait pour la linguistique et l'étude de comportements humains. Maîtrise d'au moins un langage de programmation courant, si possible expérience en traitement automatique du langage et de la parole, et/ou en apprentissage automatique, fouille de données.

Présentation détaillée du sujet :

Les systèmes de traitement automatique du langage (TAL) atteignent désormais un niveau de performance leur permettant une diffusion et utilisation par le grand public, comme la traduction automatique, la reconnaissance automatique de la parole (RAP), l'indexation de documents ... Malgré cette maturité technologique, les systèmes automatiques sont inévitablement sujets aux erreurs : ils sont ici censés fournir un niveau de service suffisant aux utilisateurs. La notion de performance a cependant toujours suscité un débat dans les communautés de TAL, car la plupart des questions et domaines scientifiques évoluent à travers les évaluations choisies. Très tôt, une très grande partie des domaines de recherche en TAL, avec l'apparition d'approches par apprentissage statistique, s'est réunie autour d'environnements expérimentaux [Galliers93], l'idée clé étant que les systèmes proposés pourraient être développés, évalués et comparés dans des environnements communs. Les métriques d'évaluation sont alors choisies et définies pour rendre les systèmes comparables : ceux qui obtiennent les performances les plus élevées sont considérés comme meilleurs. Bien qu'elles soient imparfaites et sujettes à discussion, les mesures faisant généralement consensus sont celles qui sont les plus faciles à appliquer, en particulier qui

ne nécessitent pas d'intervention humaine supplémentaire pour évaluer un nouveau système. En TAL, les mesures de référence ont la particularité de ne pas être conçues pour évaluer un système particulier mais pour s'appliquer de façon générique, par exemple le taux d'erreur-mot (WER) en RAP, les scores BLEU, NIST ou TER en traduction automatique... Au final, ces mesures reflètent principalement une partie de la performance des systèmes réels. Une possibilité pourrait être d'évaluer plus finement et indépendamment chaque système par des experts humains, comme dans la campagne d'évaluation MT IWSLT 2017 [Cettolo17], mais cette solution apparaît généralement trop coûteuse en temps et en argent, puisqu'elle nécessite une (voire plusieurs) évaluation humaine pour chaque sortie.

Un problème majeur soulevé dans les métriques d'évaluation en TAL concerne l'idée d'évaluation elle-même : avec l'objectif ultime d'obtenir un système sans erreur, ces métriques apparaissent souvent comme une sanction des systèmes eux-mêmes. Actuellement, nous construisons donc des machines qui minimisent les erreurs dans une tâche concernant une référence considérée, sans chercher à savoir, à comprendre et à évaluer l'impact de cette erreur d'un point de vue humain. L'impact de ces erreurs produites par les systèmes automatiques sur l'humain, et la manière dont les humains les reçoivent, les perçoivent, et les appréhendent au niveau cognitif, n'est alors jamais évalué, rendant ainsi l'analyse orientée système et non orientée vers l'humain, alors que ces mêmes systèmes sont conçus pour modéliser le langage humain.

Le but d'un système de reconnaissance automatique de la parole (RAP) est d'associer une séquence de mots à une séquence acoustique. Le pipeline classique de ces systèmes intègre des modules ayant chacun un rôle bien défini (modèle acoustique, modèle de langage...). Au cours des dernières années, l'apprentissage profond a permis à des modèles informatiques comportant de nombreuses couches de traitement d'apprendre des représentations de données correspondant à différents niveaux d'abstraction [LeCun15]. Dans les approches de RAP dites de bout en bout, plusieurs niveaux d'abstraction (acoustique, phonétique, lexical, syntaxique...) peuvent ainsi être intégrés dans un modèle par réseau de neurones unique. Dans la reconnaissance vocale, les modèles de bout en bout tentent de mapper un signal acoustique à une séquence de mots directement à l'aide de modèles neuronaux profonds.

Dans le contexte de la transcription automatique, la métrique d'évaluation classique est le taux d'erreur-mot (WER), calculé à partir d'une comparaison entre les transcriptions automatiques et leurs transcriptions manuelles (références). Le WER prend en compte, de manière uniforme, la somme des erreurs de substitution (mot transcrit à la place d'un mot de la transcription manuelle), d'insertion (mot transcrit ajouté par rapport à la transcription de référence) et de suppression (mot de la référence oublié dans l'hypothèse fournie par le système de RAP), divisé par le nombre total de mots dans la transcription de référence. Des campagnes d'évaluation ont été proposées à la communauté scientifique pour comparer les performances des systèmes sur un référentiel commun (actualités diffusées [Gravier12], données Youtube multi-genres [Bell15]..). S'il est alors possible d'évaluer n'importe quel système de RAP, ces campagnes restent un cadre d'évaluation et d'analyse des erreurs peu performant, les informations essentielles extraites de l'analyse des erreurs étant leur fréquence, les mots éventuellement supprimés, substitués ou insérés, ou encore les mots hors-vocabulaire (OOV)... Certains travaux cherchent à rendre la métrique du WER plus précise, ou tout du moins plus adaptée à une tâche visée, comme dans [Jannet15, Mdhaffar19], ou encore d'analyser les erreurs selon des catégories linguistiques génériques (erreurs d'accord, erreurs grammaticales...), comme dans le projet ANR VERA. Au final, il s'agit toujours d'une analyse d'erreur orientée système, les utilisateurs étant finalement ignorés, alors que chaque erreur peut avoir un impact cognitif différent. Très récemment, les auteurs de [Li19] ont publié un

ensemble de données annoté sur l'intelligibilité (c'est-à-dire la capacité d'un énoncé à comprendre) sur les sorties de transcription de différents corpus au moyen de deux systèmes de RAP. Néanmoins, ce corpus est assez limité en termes de perception humaine : seules les erreurs de transcription (substitution, suppression et insertion) sont fournies, sur lesquelles ils estiment que l'intelligibilité peut être estimée.

De nombreux travaux se sont concentrés sur la proposition d'approches automatiques pour la détection des erreurs, plus que sur une analyse détaillée de ces erreurs, faite par des linguistes experts par exemple. La détection automatique des erreurs dans les systèmes de transcription est étudiée depuis de nombreuses années, comme dans [Tam14, Ghannay15, Errattahi18a]. Bien que les approches diffèrent, la plupart des travaux considèrent les erreurs dans leur globalité (problème binaire mot correct/mot erroné), souvent pour les corriger ou les ignorer, dans l'objectif d'améliorer les tâches pour lesquelles les transcriptions sont utilisées (sous-titrage, traduction automatique, indexation...), ignorant la nature de l'erreur et son impact sur les utilisateurs finaux. Certaines approches proposent d'utiliser, pour détecter les erreurs, différentes caractéristiques acoustiques et linguistiques extraites des systèmes de RAP. Dans les pipelines de RAP classiques, cela est rendu possible par le fait que différents modules se succèdent pour effectuer une transcription : il est donc possible d'extraire des informations dans chacun d'eux, ce qui n'est pas applicable aux systèmes de RAP de bout en bout. Dans [Guo19], les auteurs ont proposé d'utiliser un modèle de langage de rescoring pour améliorer les transcriptions de bout en bout, mais cette première tentative n'exploite pas d'informations internes au système. À notre connaissance, aucun travail ne propose de détection automatique d'erreurs exploitant les informations à l'intérieur de ce type d'architecture par apprentissage profond. Notons cependant que très récemment, les chercheurs ont proposé une approche pour détecter automatiquement les erreurs de transcriptions manuelles [Wang19b], alors que de tels systèmes n'ont jusqu'à présent été proposés que sur les transcriptions automatiques. Bien que cet outil de détection d'erreur soit toujours conçu pour les systèmes ASR conventionnels, il montre qu'il existe un intérêt, scientifique et industriel dans le cas de cette étude, à prendre en compte l'humain.

Domaine / Thématique : reconnaissance automatique de la parole, traitement du langage, apprentissage automatique, linguistique

Objectif : L'objectif principal de la thèse est d'analyser finement les erreurs de transcription du point de vue de leur réception par l'utilisateur. La thèse aura trois volets complémentaires :

1. Réalisation d'un nouveau corpus de transcriptions automatiques où les erreurs sont annotées par rapport à des informations linguistiques précises et aux informations collectées lors de tests perceptifs afin de rendre compte de la manière dont les utilisateurs perçoivent (et éventuellement corrigent) ces erreurs. Réalisation de différents tests perceptifs, en confrontant des humains à ces erreurs de transcription.
2. Analyse détaillée des erreurs de transcription, qu'elles soient humaines ou automatiques, avec un système classique ou de bout en bout, afin de comprendre comment les erreurs sont perçues d'un point de vue humain. Cela permettra de mettre en lumière de nouvelles classes d'erreurs, guidées par leur difficulté, ou leur facilité, à être appréhendées par les utilisateurs finaux.
3. Approches pour la détection des erreurs dans les transcriptions de systèmes de RAP de bout en bout.

Contexte et enjeux : À notre connaissance, aucun corpus intégrant une catégorisation fine des erreurs de transcription avec une indication sur le coût de ces erreurs à partir d'une analyse cognitive humaine n'a été mis à la disposition de la communauté scientifique, allant bien plus loin qu'une simple estimation d'intelligibilité. Le corpus sera alors fourni librement à la communauté scientifique, tout comme les outils développés dans le cadre de la thèse. Il s'agira de poser les premières bases d'une recherche nouvelle et transversale, à la frontière entre la linguistique, l'informatique et les sciences cognitives, pour l'évaluation de systèmes automatiques et la compréhension de systèmes de traitement automatique du langage par apprentissage profond. L'étudiant en thèse aura alors l'opportunité de se former et proposer des approches innovantes en traitement automatique de la parole pour la compréhension d'architecture avec réseaux de de neurones profonds, mais également d'avoir une ouverture, et des compétences en linguistique et sur la mise en place de tests perceptifs, cette dernière partie étant assurée dans l'encadrement par Jane Wottawa, maître de conférences en Linguistique.

Références bibliographiques :

[Bell15] Bell, P., Gales, M. J., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., ... & Woodland, P. C. (2015). The MGB challenge: Evaluating multi-genre broadcast media recognition. In IEEE ASRU (pp. 687-693).

[Cettolo17] Cettolo, M., Federico, M., Bentivogli, L., Jan, N., Sebastian, S., Katsuitho, S., ... & Christian, F. (2017). Overview of the iwslt 2017 evaluation campaign. In IWSLT (pp. 2-14).

[Errattahi18a] 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).

[Galliers93] Galliers, J.R., Jones, K.S. (1993). Evaluating natural language processing systems. Tech. report.

[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).

[Gravier12] Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., & Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In LREC.

[Guo19] Guo, J., Sainath, T. N., Weiss, R. J. (2019). A spelling correction model for end-to-end speech recognition. In IEEE ICASSP (pp. 5651-5655).

[Jannet15] Jannet, M. A. B., Galibert, O., Adda-Decker, M., & Rosset, S. (2015). How to evaluate ASR output for named entity recognition?. In Interspeech.

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

[Li18] Li, K., Xu, H., Wang, Y., Povey, D., & Khudanpur, S. (2018). Recurrent Neural Network Language Model Adaptation for Conversational Speech Recognition. In Interspeech (pp. 3373-3377).

[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.

[Wang19b] Wang, X., Yang, J., Li, R., Sadhu, S., Hermansky, H. (2019). Exploring Methods for the Automatic Detection of Errors in Manual Transcription. arXiv preprint arXiv:1904.04294.

Les sujets devront être adressés à

secretariat-ed@univ-avignon.fr

avant le 8 avril 2020