Authôt.com
Vous parlez. Nous écrivons.

Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

Proposition de Stage de Recherche
Master 2 – 2018-2019

Détection automatique de la thématique et adaptation des modèles de langage

Domaines : Traitement du Langage Naturel, Modèle de langage, Reconnaissance Automatique de la Parole.

Mots clés : Word embedding, TALN, Modèle de langage, Parole

Lieu du stage : IRIT-UPS, 118 route de Narbonne, 31062 Toulouse

Contact : sferreira@authot.com sebastien.ferreira@irit.fr

Description :

Authôt est une start-up française spécialisée dans la retranscription de la parole et le sous-titrage depuis 6 ans. Travaillant avec de nombreux secteurs : Éducation, Grand comptes, Médias, Productions audiovisuelles, etc. Nous permettons, entre autres, de répondre aux normes d'accessibilité exigées par la loi de février 2005.

Aujourd'hui, grâce à notre plateforme en ligne et sécurisée de retranscription automatique et de sous-titrage, nous divisons par deux le temps de travail lié à la retranscription d'un enregistrement audio en texte et le sous-titrage d'une vidéo. **L'enjeu est de progresser dans cette productivité.**

Les systèmes de transcription automatique de la parole fonctionnent grâce à deux modèles : le modèle acoustique et le modèle de langage (ML). Le ML aide à déterminer le mot le plus probablement prononcé en modélisant une logique dans la succession de mots. Par exemple, il est plus probable que la phrase « le chien aboie » soit prononcé plutôt que « le chien maison ».

De plus, les systèmes de transcription automatique de la parole utilisent un vocabulaire fini. Le vocabulaire recense tous les mots que le système sera en mesure de reconnaître : un mot non recensé ne peut pas être détecté. Le problème principal est qu'un vocabulaire efficace nécessite un compromis entre le nombre de mots qu'il contient et sa couverture lexicale. En pratique, les systèmes de transcription automatique de la parole sont conçus pour maximiser la couverture lexicale sur un nombre de thème varié tout en limitant le nombre de mots de mots possible (généralement entre 50K à 100K mots). Afin d'obtenir une transcription de qualité, la couverture lexicale doit dépendre de la thématique du fichier transcris : un vocabulaire ayant une bonne couverture pour un match de foot aura une couverture beaucoup plus faible pour un cours de médecine.

Comme **les fichiers soumis à une tâche de transcription sont très variés** chez Authôt, nous cherchons à intégrer une solution qui détecte automatiquement la thématique d'un fichier afin d'adapter le modèle de langage et le vocabulaire.

Objectif :

L'objectif de ce stage sera de **détecter automatiquement une/ou plusieurs thématiques** abordées. Puis de compléter le lexique avec les mots manquants et extraire des corpus de texte pour adapter le modèle de langage afin d'améliorer la qualité de la transcription automatique.

Profil du candidat :

Nous recherchons un étudiant pour un stage de fin d'études (M2 ou 3e année d'école d'ingénieur). Un(e) candidat(e) ayant des connaissances en **reconnaissance de formes** et **apprentissage automatique** et intéressé par la recherche. Des connaissances en Traitement Automatique du Langage Naturel et en Reconnaissance Automatique de la Parole sont un plus pour bien comprendre les transcriptions en sortie des systèmes, et pour pouvoir les adapter.

Bibliographie en lien avec le stage :

Lecorvé, Gwénolé. "*Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole*". Diss. INSA de Rennes, 2010.

Bougouin, Adrien. "État de l'art des méthodes d'extraction automatique de termes-clés." *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*. 2013.

Thomas, Justine Raju, Santosh Kumar Bharti, and Korra Sathya Babu. "Automatic keyword extraction for text summarization in e-newspapers." *Proceedings of the International Conference on Informatics and Analytics*. ACM, 2016.

Bharti, Santosh Kumar, and Korra Sathya Babu. "Automatic Keyword Extraction for Text Summarization: A Survey." *arXiv preprint arXiv:1704.03242* (2017).

Fallery, Bernard, and Florence Rodhain. "Quatre approches pour l'analyse de données textuelles: lexicale, linguistique, cognitive, thématique." *XVI ème Conférence de l'Association Internationale de Management Stratégique AIMS*. AIMS, 2007.

Vosse, Theo. "Detecting and correcting morpho-syntactic errors in real texts." *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992.

Federico, Marcello, and Nicola Bertoldi. "Broadcast news LM adaptation using contemporary texts." *Seventh European Conference on Speech Communication and Technology*. 2001.

Bellegarda, Jerome R. "Statistical language model adaptation: review and perspectives." *Speech communication* 42.1 (2004): 93-108.

Arisoy, Ebru, and Murat Saraçlar. "Lattice extension and vocabulary adaptation for Turkish LVCSR." *IEEE transactions on audio, speech, and language processing* 17.1 (2009): 163-173.

Aronowitz, Hagai. "Online vocabulary adaptation using contextual information and information retrieval." *Ninth Annual Conference of the International Speech Communication Association*. 2008.

Bertoldi, Nicola, and Marcello Federico. "Lexicon adaptation for broadcast news transcription." *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*. 2001.