"Espaces de représentation sémantique distribues pour les tâches de traduction automatique (compréhension et génération de la parole) dans les systèmes d'interaction vocale"

Keywords: word embeddings, deep neural networks, statistical machine translation, spoken language understanding, natural language generation

L'émergence de l'accès universel à la société numérique implique le développement de systèmes d'interaction multilingues : par exemple pour la recherche d'information dans des corpus multimédia multilingues ou pour le développement de systèmes de dialogue multilingues (dont les tâches peuvent aller des systèmes de question/réponse jusqu'à la résolution de problèmes complexes, comme l'aide aux usagers d'une compagnie). Dans ce contexte la traduction automatique n'est pas limitée au passage entre deux langues humaines, ainsi la compréhension et la génération de la parole peuvent être vues comme des exemples de tâches de traduction de la parole et seront étudiées dans le cadre de cette thèse. La recherche d'une solution optimale pour l'ensemble des tâches visées sera bien sur un élément clef de cette étude.
Depuis une dizaine d'années, les approches les plus performantes pour la traduction automatique sont basées sur l'utilisation de modèles probabilistes. Pour être efficaces, de telles approches nécessitent de disposer de larges bases de données d'exemples (dans ce cas, des corpus de phrases parallèles entre les langues source et cible), ce qui n'est pas toujours possible, en particulier dans les domaines spécialisés. Par ailleurs, dans le cas de la traduction de la parole, les systèmes doivent baser leurs hypothèses sur les sorties imparfaites des systèmes de reconnaissance de parole. Il est donc important de baser la décision sur un maximum d'informations (et pas uniquement sur l'identité des mots présents).
Des approches récentes ont montre l'intérêt d'intégrer l'information sémantique pour réaliser la traduction automatique de la parole par des méthodes statistiques [1]. Les gains en performance restent toutefois limités et une grande marge d'amélioration est encore possible. De plus, la nécessité d'analyser au préalable le texte réduit les possibilités d'application de ces approches dans le contexte des systèmes d'interaction vocale, où l'étape de reconnaissance de la parole en diminue la faisabilité. En effet, l'extraction fine de caractéristiques est fortement perturbée par le niveau élevé de bruit dans les textes à traiter issus d'une étape de décodage automatique.
Aussi l'émergence de nouvelles approches fortement automatiques pour la représentation des données textuelles, par exemple à l'aide de réseaux neuronaux profonds [2], présente une nouvelle opportunité pour développer des approches permettant d'envisager l'utilisation de nouveaux paramètres sur une grande échelle pour guider et améliorer la traduction par la prise en compte renforcée d'information syntactico-sémantiques [3]. Dans le contexte des systèmes de dialogue homme-machine il sera possible d'évaluer la pertinence des approches envisagées sur des tâches et des corpus de tailles contrôlables.

Références :
[1] Dekai Wu et Pascale Fung, Can Semantic Role Labeling Improve SMT?, EAMT, 2009
[2] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, 2013.
[3] David Chiang, Kevin Knight et Wei Wang, 11,001 New Features for Statistical Machine Translation, NAACL-HLT, 2009

Pré-requis : Master en informatique avec une composante sur les méthodes d'apprentissage automatique et/ou sur l'ingénierie de la langue
Encadrant : Prof. Fabrice Lefèvre (co-encadrants : Stéphane Huet et Bassam Jabaian)
Lieu : LIA-CERI-Univ. Avignon