Repérage d'alternance codique

Projet pour le Master DAC - François Yvon

Sorbonne Université, CNRS, ISIR

L'alternance codique (ou code-switching) se définit par un changement de langue durant une interaction linguistique. Elle peut avoir lieu entre deux phrases successives, au sein d'une phrase (''That's comme ça that you thank me to have learned you english''), ou même au sein d'un mot, quand on greffe à une base lexicale d'une première langue une terminaison d'une autre langue (''ce projet, j'ai liké de ouf"). Ces phénomènes linguistiques sont fréquents dans les échanges entre locuteurs qui ont plusieurs langues en commun.

Dans ce projet on s'intéresse à une question en apparence simple : repérer les alternances codiques au sein de phrases, en attribuant à chaque mot la langue qui lui correspond. Ce problème est en fait difficile car (a) il n'existe pas pas beaucoup de données d'apprentissage (b) elles concernent peu de couples de langues (espagnol-anglais (spanglish), hindi-anglais (hinglish), français-anglais (franglish), turc-allemand, etc) ; par ailleurs l'immense majorité des phrases que l'on observe dans les textes est monolingue.

Le but de ce projet est d'améliorer MaskLID, un détecteur d'alternance codique assez simple qui exploite un classifieur pré-entrainé pour détecter des langues (plus de 2000 langues dans la version actuelle), mais n'intègre pas explicitement d'étape d'apprentissage. Parmi les améliorations possibles, on pourra s'intéresser à mieux exploiter les caractéristiques du classifieur préentrainer, à le calibrer, à ajouter des dépendances dans les décisions de classification de mots adjacents, de contraindre le nombre de langues dans un segment, etc. Ces améliorations seront systématiquement évaluées sur les jeux de données standard (par exemple provenant du benchmark LinCE).

Références :

— MaskLID: Code-Switching Language Identification through Iterative Masking (Kargaran et al., Proc. ACL 2024)
— LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation (Aguilar et al., Proc. LREC 2020)).