Traitement morpho-syntaxique du Latin

Alexis Nasr, Tristan Vigliano

Ce stage s'inscrit dans le cadre d'une collaboration avec le laboratoire CIELAM. Il vise à proposer des outils de TAL pour le Latin, plus particulièrement, un étiqueteur en parties de discours, un analyseur morphologique, un lemmatiseur et un analyseur syntaxique. On s'intéressera en particulier à une caractéristique du Latin qui est son riche système de déclinaisons et la souplesse de l'ordre des mots dans la phrase.

Dans un sytème classique de TAL, les deux opérations d'analyse syntaxique et d'analyse morphologique sont réalisés par deux modules distincts. Dans le cas du Latin, cette distinction est problématique du fait que ces deux opérations sont intimement liées. On étudiera différentes manières de réaliser cette tâche de manière conjointe en recourant par exemple à l'apprentissage multi-tâche Caruana (1997).

On étudiera aussi la possibilité de recourir à l'enrichissement de données Feng et al. (2021) pour mieux traiter l'ordre libre des mots.

En ce qui concerne la nature des modèles utilisés, différentes options seront étudiées, parmi lesquelles la possibilité de traiter les différents niveaux d'analyse à l'aide de l'analyse par transition, comme proposé par Dary and Nasr (2021).

Références

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1) :41–75.

Dary, F. and Nasr, A. (2021). The reading machine : a versatile framework for studying incremental parsing strategies. In *The 17th International Conference on Parsing Technologies*.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for nlp. *arXiv preprint arXiv :2105.03075*.