|  |
| --- |
|  |
| Rapport EIT |
|  |
| AMRAOUI Yassine – KIRILLOV Alexandre – MARTIN Brian |

**ET5 INFO**

8/03/2019

Rapport EIT

AMRAOUI Yassine – KIRILLOV Alexandre – MARTIN Brian

**Sommaire**

**1. Introduction**

**1.1 Problématique**

L’objectif de ce projet est d’étudier deux frameworks d’analyse linguistique : CEA List LIMA et Standford Core NLP.

CEA List LIMA est une plateforme d’analyse linguistique basé sur des règles formelles et des ressources validées par des experts linguistes, tandis que Standford Core NLP est basé sur des statistiques dans la mesure où la méthode d’apprentissage de cette dernière boite à outils linguistiques utilise des corpus annotés.

**1.2 Expérimentation des deux plateformes d’analyse linguistique**

Ces deux plateformes d’analyse linguistique ont des finalités identiques. En effet, les modules suivants sont présents dans les deux cas :

* Une étape de tokenisation pour découper les chaînes de caractères en mots
* Une étape d’analyse morphologique pour voir si chaque mot découpé fait partie de la langue ou non, et lui associer une propriété syntaxique
* Une étape d’analyse morpho-syntaxique pour désambiguïser les mots
* Une étape d’analyse syntaxique pour voir la relation entre les mots
* Une étape de reconnaissance des entités nommées pour identifier les dates, lieux, etc.

Néanmoins, leurs moyens d’analyse de ces étapes ne se basent pas sur les mêmes méthodes : règles pour l’un, statistiques pour l’autre.

Pour chacune des deux plateformes, il faudra d’abord commencer par lancer l’analyse linguistique sur les phrases afin d’en sortir l’analyse que nous aura fourni chacun de ces frameworks sous forme de matrice.

**1.3 Evaluer et Comparer**

Après avoir récupérer l’output brut de Stanford et Lima, le but est de coder un programme permettant de ressortir pour chaque phrase son POS tag (ou entité nommée) associée.

Cependant, il est nécessaire d’avoir un point de comparaison universel pour donner un sens à la comparaison : d’où l’utilité des tags universels, que l’on va remplacer à la place des tags propres à chaque plateforme linguistique à côté des mots. Ainsi, on se retrouvera avec des tags universels à côté de chaque mot, pour Lima et Stanford respectivement.

Enfin, on évaluera les performances des deux analyseurs grâce à des fichiers contenant des mots tagés universellement comme décrit juste au-dessus.

**2. Etat de l’art**

**1.1 Les différentes approches d’analyse linguistique existantes**

L’objectif de ce projet est d’étudier deux frameworks d’analyse linguistique : CEA List LIMA et Standford Core NLP.