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

|  |
| --- |
| **ET5 INFO**    10/03/2019 |
|  |

**Sommaire**

[**1. Introduction**](#_uytg7d3ogrjm) **3**

[**2. Etat de l’art**](#_wj7syoo3ve9l) **4**

[**3. Description des deux plateformes**](#_ay6oudpluq9z) **5**

[**4. Description des expérimentations**](#_qgxvo7j8tttz) **6**

# 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 basée 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 boîte 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, 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 Évaluer et Comparer**

Après avoir récupéré 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) associé.

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 tagués universellement comme décrit juste au-dessus.

# 2. Etat de l’art

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

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

En ce qui concerne l’état de l’art, il existe différentes approches d’analyse linguistique, dont Lima et Core NLP que l’on décrira dans la partie 3, et les suivantes :

- **NLTK** (Natural Language ToolKit) a été pendant longtemps la librairie python standard pour le NLP (Natural Language Processing). Elle contient différents algorithmes de classification comme LIMA et Stanford Core NLP. Néanmoins, la librairie est trop ancienne et présente des limites de performances.

- **TextBlob** est une librairie python pour le NLP caractérisé par son interface NLTK intuitif.

- **Genism** est une librairie codée en python pour l’analyse de similarités des documents, basé sur le traitement de lots en mémoire.

- **SpaCy** est une librairie performante créée par l’entreprise Explosion AI

- **UIMA** est un standard de OASIS (Organisation pour l’avancement des normes d’information structurée) développé de base par IBM, et destiné à fournir des composants pour les analyses d’informations non structurées. Il peut notamment décrire des modèles de conception et suggérer des représentations de données.

- **Gate** est une boîte à outils Anglaise pour le NLP codé en Java et très largement utilisée dans le monde aussi bien par les chercheurs que par les étudiants. Il se caractérise par la présence d’un système d’extraction d’information de type ANNIE (A Nearly-New Information Extraction System)

- **ALEP** est une librairie pour coder des applications NLP, qui est reconnue pour son efficacité dans le domaine de la grammaire bilinguale et des vérifications stylistiques.

- **UIUC** **Curator** est une librairie de Chicago visant à simplifier des tâches comme les taggers vocaux. Codé sous Linux, il ne peut pas être utilisé sous Windows sans machine virtuelle.

- **Clear TK**, **DKPro Core** et **JCoRe** sont des packages dont les solutions sont encore trop complexes et lourds.

**2.2. Choix des plates-formes**

Nos expériences porteront sur les plates-formes LIMA et Stanford CoreNLP car se sont des plates-formes libres et très utilisées. De plus, la différence de méthodes d’analyse, l’une étant basée sur des règles et l’autre sur un modèle statistique, rend la comparaison intéressante.

# 3. Description des deux plateformes

**3.1. CEA List LIMA**

LIMA est une plate-forme d’analyse multi-langage développée au CEA LIST par le laboratoire LVIC. Créée à l’origine pour développer des applications industrielles basées sur le NLP, cette plateforme permet le traitement de 10 langues différentes. D’autre part, LIMA permet également de tester et évaluer divers modules linguistiques et la production de nouvelles ressources linguistiques.

Basée sur des règles formelles et des ressources validées par des experts linguistes, LIMA est donc une plate-forme qui regroupe une architecture, un lot d’outils et de ressources ainsi qu’un environnement de développement adapté au NLP. Sa stratégie de développement est organisée autour de trois grands objectifs : le multilinguisme, la modularité (adaptation à un grand nombre d’applications et possibilité d’ajouter de nouvelles fonctionnalités), et l’efficacité (capacité à traiter en un temps raisonnable un quantité importante de données).

La force de LIMA réside dans sa modularité. Le lot d’outils fournis par la plate-forme, allant de la tokenization à l’analyse syntaxique tout en prenant en compte de nombreux phénomènes linguistiques tels que l’absence de délimiteurs ou le manque de voyelles, permet un travail sur des textes rédigés en différentes langues, et même sur des documents multimédias et se basant sur de la reconnaissance d’images.

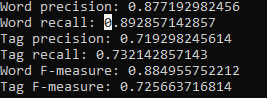
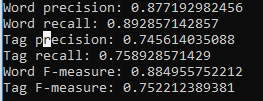
**3.2. Stanford Core NLP**

La plate-forme Stanford CoreNLP permet elle aussi l’analyse du langage naturel, mais contrairement à LIMA, elle est basée sur un modèle d’analyse statistique. Cette plate-forme supporte également de nombreux langages, tels que le français, l’anglais, l’allemand, l’arabe ou encore le chinois. Toutefois, l’anglais est la seule langue sur laquelle tous les composants d’analyse peuvent être utilisés.

Ce qui fait la force de cette plate-forme, c’est l’alliance entre simplicité d’utilisation (interfaces simples, peu de bagages externes requis) et la qualité de ses composants d’analyse. En effet, Stanford CoreNLP se concentre essentiellement sur les fonctionnalités les plus utiles afin de rester légère et simple d’utilisation, et ainsi pouvoir être utilisée en tant que composant dans un plus gros projet.

# 4. Description des expérimentations

**4.1. Evaluation de l’analyse morpho-syntaxique**

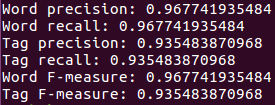
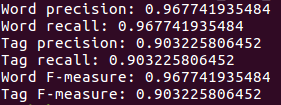
**Les résultats avec les étiquettes PTB Les résultats avec les étiquettes universelles**

On remarque une légère amélioration des résultats sur les tags lorsqu’on utilise des tags universels. En effet, les résultats avec les étiquettes universelles sont supérieurs d’environ 2% à ceux avec les étiquettes PTB.

Cette amélioration s’explique par le fait que dans le cas universel, on utilise plusieurs fois le même tag pour représenter des éléments différents. Par exemple, le tag NOUN remplace quatre tags PTB différents. Ainsi, le mot “resort” dans les fichiers testés avait le tag NNP dans le fichier lima, et NN dans le fichier référence, mais ces deux tags devenaient NOUN après le passage en étiquettes universelles.

**4.2. Evaluation du POS tagger**

Les résultats du fichier evaluate.py sur le fichier wsj\_0010\_sample.txt avec les étiquettes PTB et avec les étiquettes universelles respectivement :

**Les résultats avec les étiquettes PTB Les résultats avec les étiquettes universelles**

D’après ces résultats les paramètres qui diffèrent sont les Tag précision, Tag recall et Tag F-measure. La valeur des tags est plus basse avec les étiquettes universelles. Elle est environ égale à 0.93 pour les étiquettes PTB et elle est environ égale à 0.90 pour les étiquettes universelles. Les étiquettes PTB sont donc plus précisent pour la détection des tags dans un texte.

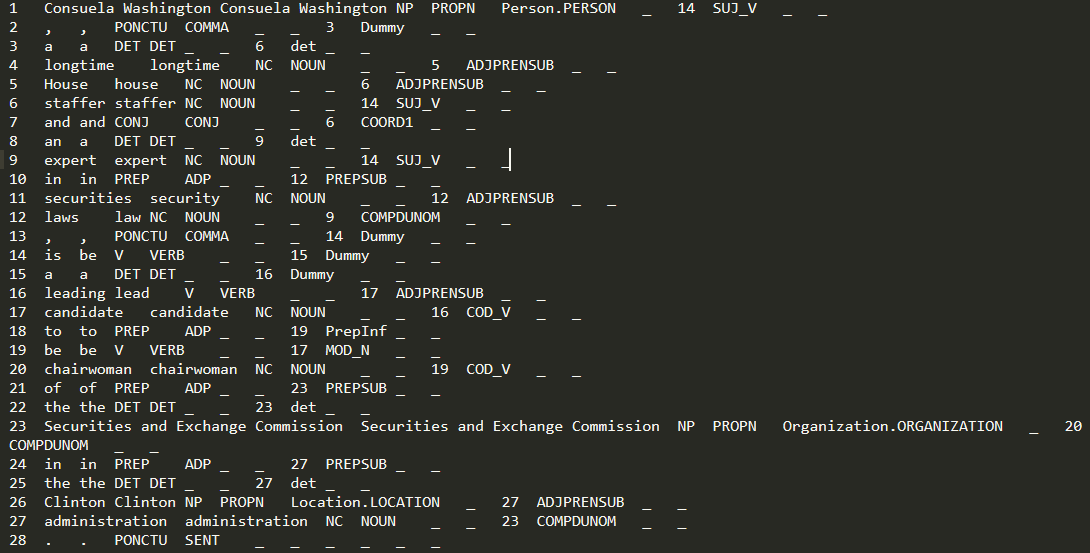
**4.3. Evaluation du CEA List et de l’outil l’université de Stanford**

La table de correspondance entre les étiquettes LIMA et les étiquettes Stanford est :

* Person.PERSON : PERSON
* Organization.ORGANIZATION : ORGANIZATION
* Location.LOCATION : LOCATION

**4.3.1. Lima**

Lima



Lima → Stanford



Stanford → Univ



**4.3.2. Stanford**

Stanford



Stanford → Univ



**4.3.3. La référence Enamex**

**4.3.4. Evaluation et comparaison**

**4.4. Analyse des résultats**

# 5. Conclusion

**5.1. Limitations**

**5.2. Perspectives d’améliorations pour les deux plateformes**