Rapport EIT

AMRAOUI Yassine - KIRILLOV Alexandre - MARTIN Brian

ET5 INFO

10/03/2019

Sommaire

1. Introduction

1.1 Problématique

L'objectif de ce projet est d'étudier deux frameworks d'analyse linguistique : CEA List LIMA et Stanford 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 Stanford 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

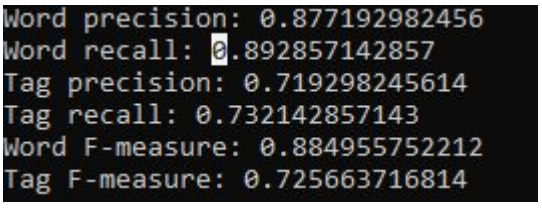
La plate-forme Stanford CoreNLP permet elle aussi l'analyse du langage naturel, mais contrairement à LIMA, elle est basée sur un modèle d'analyse statistique. Cette plate-forme supporte également de nombreux langages, tels que le français, l'anglais, l'allemand, l'arabe ou encore le chinois. Toutefois, 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.

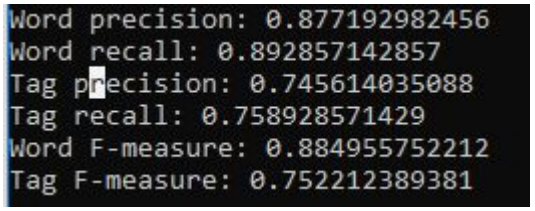
4. Description des expérimentations

4.1. Evaluation de l'analyse morpho-syntaxique

Dans cette partie, on transforme la sortie Lima brute sous la forme "Mot_Étiquette", puis on lui applique les étiquettes PTB pour ensuite passer aux étiquettes universelles. On réalise la même opération sur le fichier de référence, pour ensuite évaluer.



Les résultats avec les étiquettes PTB



Les résultats avec les étiquettes universelles

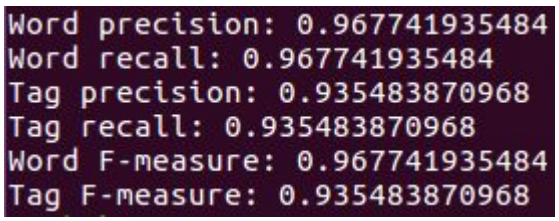
On remarque une légère amélioration des résultats sur les tags lorsqu'on utilise des tags universels. En effet, les résultats avec les étiquettes universelles sont supérieurs d'environ 2% à ceux avec les étiquettes PTB.

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.

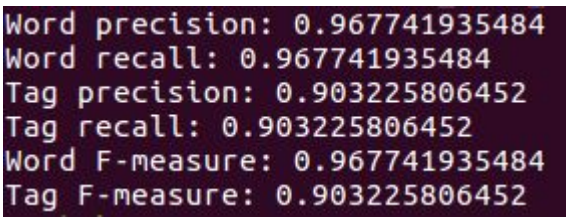
En ce qui concerne les mots, les performances ne changent pas puisque les mots ne sont pas modifiés lors du changement d'étiquettes.

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 précises pour la détection des tags dans un texte.

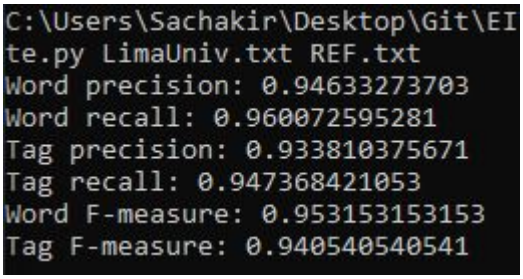
4.3. Evaluation du CEA List et de l'outil l'université de Stanford

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

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

L'objectif ici est de rapporter Lima en étiquettes universelles, en passant par un passage au format Stanford. Une fois les parties Lima et Stanford au format universel, on réalise une conversion du fichier contenant les tags de type "Ex" (Enamex, Timex, Numex etc.) au format universel, qui servira de point de comparaison pour les deux fichiers précédemment expliqués. Ainsi, on pourra comparer les performances des deux analyseurs sur les entités nommés dans la partie 4.4.

Evaluation Lima (Univ) avec Référence (Univ)



Evaluation Stanford (Univ) avec Référence (Univ)

```
C:\Users\Sachakir\Desktop\Git\EI
te.py StanUniv.txt REF.txt
Word precision: 0.84958677686
Word recall: 0.932849364791
Tag precision: 0.829752066116
Tag recall: 0.911070780399
Word F-measure: 0.889273356401
Tag F-measure: 0.868512110727
```

On remarque une différence de 10% entre Lima et Stanford (en faveur de Lima) en ce qui concerne la précision et de 3% (toujours en faveur de Lima) en ce qui concerne le rappel. Ceci s'explique par la différence de tokenisation entre ces deux plates-formes. En effet, Lima considère les entités composées de plusieurs mots, tel que "Consuela Washington", comme un seul token, comme le fait la référence humaine.

5. Conclusion

Avec les deux évaluations de Lima, une sur l'ensemble des étiquettes et une sur seulement trois étiquettes pour la partie V), nous pouvons remarquer que la précision est plus élevée avec seulement trois étiquettes.

Nous avons utilisé le fichier généré par Lima donné dans l'énoncé du projet. Il était impossible de générer les fichiers Lima par nous-mêmes sur les machines virtuelles car les étiquettes générées par Lima ne correspondaient pas aux étiquettes PTB et il était impossible d'utiliser ces fichiers générés pour l'évaluation.

6. Contributions individuelles

Nous avons tous contribués au rapport :

- **Yassine** : Introduction et état de l'art
- **Brian** : Description des deux plates-formes
- **Alexandre :** Conclusion
- **Tous** : Description des expérimentations

Concernant les scripts, **Alexandre** s'est spécialisé dans la partie V et le ReadMe. **Yassine** s'est spécialisé dans l'analyse des résultats, les algorithmes et la mise en place des comparaisons, et la partie Lima (LimaToText.py pour transformet en Mot_Etiquette avec les etiquettes PTB, scripts pour Lima…) : étiquettes Lima, PTB et Universel. **Brian** s'est concentré sur les scripts de conversion Python : passage du fichier .ref en texte (RefToText.py), conversion des étiquettes PTB en étiquettes universelles (PTBToUniversal.py), conversion de la référence humaine en étiquettes universelles (TagsToLabels.py).

Annexe

4.3.1. Lima

```
1   Consuela Washington Consuela Washington NP  PROPN   Person.PERSON   _   14  SUJ_V   _   _
2   ,   ,   PONCTU  COMMA   _   _   3   Dummy   _   _
3   a   a   DET DET _   _   6   det _   _
4   longtime    longtime    NC  NOUN    _   _   5   ADJPRENSUB  _   _
5   House   house   NC  NOUN    _   _   6   ADJPRENSUB  _   _
6   staffer staffer NC  NOUN    _   _   14  SUJ_V   _   _
7   and and CONJ    CONJ    _   _   6   COORD1  _   _
8   an  a   DET DET _   _   9   det _   _
9   expert  expert  NC  NOUN    _   _   14  SUJ_V   _   _
10  in  in  PREP    ADP _   _   12  PREPSUB _   _
11  securities  security    NC  NOUN    _   _   12  ADJPRENSUB  _   _
12  laws    law NC  NOUN    _   _   9   COMPDUNOM   _   _
13  ,   ,   PONCTU  COMMA   _   _   14  Dummy   _   _
14  is  be  V   VERB    _   _   15  Dummy   _   _
15  a   a   DET DET _   _   16  Dummy   _   _
16  leading lead    V   VERB    _   _   17  ADJPRENSUB  _   _
17  candidate   candidate   NC  NOUN    _   _   16  COD_V   _   _
18  to  to  PREP    ADP _   _   19  PrepInf _   _
19  be  be  V   VERB    _   _   17  MOD_N   _   _
20  chairwoman  chairwoman  NC  NOUN    _   _   19  COD_V   _   _
21  of  of  PREP    ADP _   _   23  PREPSUB _   _
22  the the DET DET _   _   23  det _   _
23  Securities and Exchange Commission  Securities and Exchange Commission  NP  PROPN   Organization.ORGANIZATION   _   20
COMPDUNOM   _   _
24  in  in  PREP    ADP _   _   27  PREPSUB _   _
25  the the DET DET _   _   27  det _   _
26  Clinton Clinton NP  PROPN   Location.LOCATION   _   27  ADJPRENSUB  _   _
27  administration  administration  NC  NOUN    _   _   23  COMPDUNOM   _   _
28  .   .   PONCTU  SENT    _   _   _   _   _   _
```

Sortie Lima brut

```
ConsuelaESPACEWashington/PERSON ,/O a/O longtime/O House/O staffer/O and/O an/O expert/O in/O securities/O laws/O ,/O is/O a/O
leading/O candidate/O to/O be/O chairwoman/O of/O the/O SecuritiesESPACEandESPACEExchangeESPACECommission/ORGANIZATION in/O the/O
Clinton/LOCATION administration/O ./O Ms./O Washington/LOCATION ,/O 44/O years/O old/O ,/O would/O be/O the/O first/O woman/O and/O
```

Conversion Lima vers Stanford

```
ConsuelaESPACEWashington_NOUN , a longtime House staffer and an expert in securities laws , is a leading candidate to be chairwoman
of the SecuritiesESPACEandESPACEExchangeESPACECommission_NOUN in the Clinton_NOUN administration . Ms. Washington_NOUN , 44 years
```

Conversion en tag universel

4.3.2. Stanford

```
Consuela/PERSON Washington/PERSON ,/O a/O longtime/O House/ORGANIZATION staffer/O and/O an/O expert/O in/O securities/O laws/O ,/O is
/O a/O leading/O candidate/O to/O be/O chairwoman/O of/O the/O Securities/ORGANIZATION and/ORGANIZATION Exchange/ORGANIZATION
Commission/ORGANIZATION in/O the/O Clinton/PERSON administration/O ./O
```

Sortie Stanford brut

```
Consuela_NOUN Washington_NOUN , a longtime House_NOUN staffer and an expert in securities laws , is a leading candidate to be
chairwoman of the Securities_NOUN and_NOUN Exchange_NOUN Commission_NOUN in the Clinton_NOUN administration . Ms. Washington_NOUN ,
```

Conversion en tag universel

4.3.3. La référence

La référence à l'état brut

```
ConsuelaESPACEWashington_NOUN, a longtime House_NOUN staffer and an expert in securities laws, is a leading candidate to be
chairwoman of the SecuritiesESPACEandESPACEExchangeESPACECommission_NOUN in the Clinton_NOUN administration.
Ms. Washington_NOUN, 44 years old, would be the first woman and the first black to head the five member commission that oversees the
securities markets.
```

La référence au format Lima

```
ConsuelaESPACEWashington_NOUN, a longtime House_NOUN staffer and an expert in securities laws, is a leading candidate to be
chairwoman of the SecuritiesESPACEandESPACEExchangeESPACECommission_NOUN in the Clinton_NOUN administration. Ms. Washington_
```