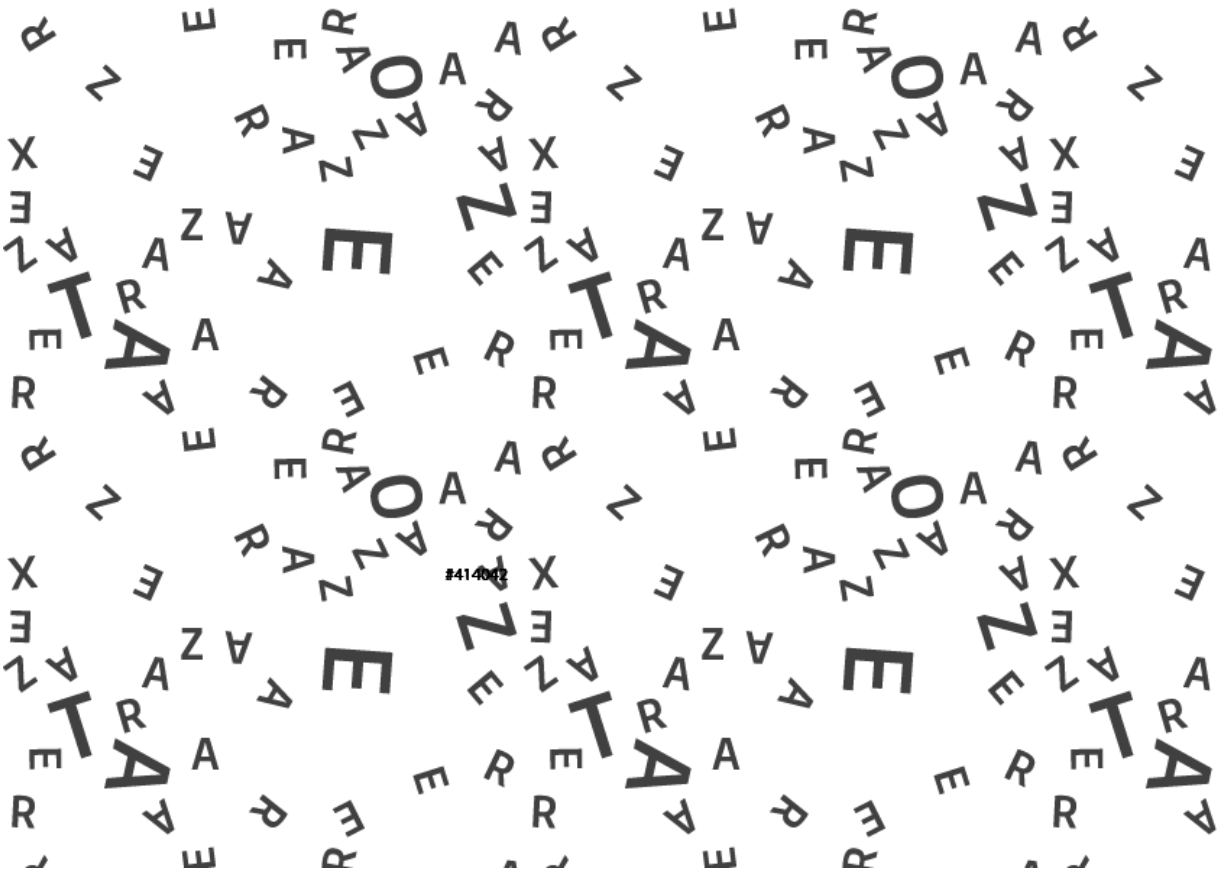
JARDET Quentin
MAYAKI Oummar

EIT
Evaluation d'outils de TAL

08/03/2019



Polytech Paris Sud
Maison de l'Ingénieur - bât 620 - Centre Scientifique d'Orsay - 91405 Orsay - France
tel : +33 (0)1 69 33 86 00 - fax : +33 (0)1 69 41 99 58 - www.polytech.u-psud.fr

Sommaire

Introduction

Ce document a pour objectif d'évaluer les différents outils de TAL (Transaction Application Language) à disposition. Ainsi, nous souhaitons analyser les outils open-source **CEA List LIMA** et **Stanford Core NLP**, afin d'obtenir une étude objective sur les points forts, les points faibles et les différences entre ces deux technologies.

Dans un premier temps, nous allons faire le point sur l'état actuel des méthodes d'analyses linguistiques (leur fonctionnement, caractéristiques,...). Nous expliquerons aussi pourquoi nous avons choisi ces deux outils (Lima et Stanford Core) en tant que référence pour nos tests.

Dans un second temps, nous présenterons brièvement LIMA et Stanford Core pour offrir une idée générale de leur fonctionnement et donc des différences internes entre les deux.

Puis, nous effectuerons des séries de tests sur les deux plateformes. Les résultats seront analysés puis comparés. Nous pourrons ainsi déduire, les avantages et les désavantages de choisir Lima ou Stanford Core.

I. Etat de l'art

Les objectifs de Lima et Stanford Core sont de proposer des outils et des ressources pour le développement de logiciels et d'applications utilisant les technologies d'analyse du langage.

Il existe de nombreuses autres architectures que les deux choisis pour ce projet, mais elles n'ont pas toutes été construites afin de répondre aux mêmes problématiques. Les architectures utilisent généralement une combinaison de modules.

Certaines architectures se basent sur l'utilisation de base de données, mais ne sont pas très modulables de part l'obligation de normaliser les données pour la communication entre modules (MULTITEXT, LT XML library).

L'autre type d'architectures est le "TIPSTER-like architecture", celles-ci ont pour particularité d'avoir une interface commune pour l'ensemble de ses modules et une représentation des données particulière (le plus souvent par graphe d'annotations). Ce type d'architecture offre une bonne combinaison entre efficacité et modularité. Ainsi, LIMA, GATE, TEXTRACT sont de ce type.

Pourquoi choisir Lima et Stanford Core? Nous avons choisi ces deux plateformes pour deux raisons principales: Leur accessibilité au public et leur fonctionnement différent.

En effet, dans le cadre d'un cours pédagogique, nous ne disposons pas des ressources nécessaires pour obtenir une licence d'une plateforme payante comparable à Lima ou Stanford Core. C'est pourquoi, le choix d'un projet open-source est très avantageux. Les deux outils offrent une documentation détaillée sur leurs caractéristiques et sont faciles à expérimenter.

La seconde raison d'un tel choix est la différence entre Lima et Stanford Core. Dans le cas de Lima, un dictionnaire regroupant une grande quantité de mots est construit antérieurement. L'analyse des mots d'un texte se base donc sur des règles construites par des experts linguistiques.

Stanford Core n'utilise pas les mêmes mécaniques, en effet bien que les deux plateformes ont certaines étapes communes dans l'analyse d'un texte (tokenization), il y a de nombreuses différences. Il n'y a pas de dictionnaire de mots car Stanford Core utilise le machine learning pour l'annotation des textes (cela n'est pas le cas de tous les modules).

II. Présentation des plateformes

A. CEA list LIMA

Tel que présenté dans la première partie LIMA est une architecture composée de modules pour l'analyse linguistique. Sa caractéristique importante est qu'elle est flexible et supporte plusieurs langages tout en offrant une efficacité importante.

LIMA utilise un pipeline afin de faire circuler les données à travers différents traitements. L'ensemble des traitements suivent des règles établies par des experts linguistiques.

L'architecture a été construite pour avoir deux possibilités d'utilisations : directement avec LIMA et indirectement via une application intégrant LIMA. Bien sûr la seconde utilisation est très avantageuse car cela permet d'accéder en open-source à des outils de développement très intéressants.

Les modules sont très facilement remplaçables par les options de configurations. Les ressources, ainsi que les unités de processus sont accessibles durant le temps d'exécution. Toutes ces possibilités permettent donc de manager à notre bon vouloir les options de l'analyse avec les ressources souhaitées.

B. Stanford Core NLP

Standford Core NLP est une infrastructure logicielle basée sur une machine virtuelle Java contenant les principaux outils pour faire de l'analyse linguistique. C'est l'un des outils d'analyse virtuelle les plus utilisés aujourd'hui. Son succès est dû à sa facilité d'utilisation et d'implantation dans un éventuel autre système.

L'analyse linguistique du Standford Core NLP est basé sur un système d'annotations. Le texte en entrée est stocké dans un *Annotator,* puis chaque traitement est appliqué à cet objet, ce qui aura pour effet d'annoter le texte. Les traitements sont réalisés dans l'ordre, et on obtient en sortie le texte annoté.

Le système est fourni avec les outils d'analyse linguistiques dédiés à la langue anglaise, mais des modules tout aussi fournis permettent de traiter la langue chinoise. Des modules légèrement moins fournis permettent aussi de traiter le français, l'allemand et l'arabe, bien qu'il soit possible de modifier ou de créer des modules permettant le traitement d'autres langues.

III. Expérimentations

A. CEA list Lima

Dans cette partie, nous allons effectuer différents tests, afin d'étudier les performances de Lima que ce soit en précision ou en rappel.

- **1er test**

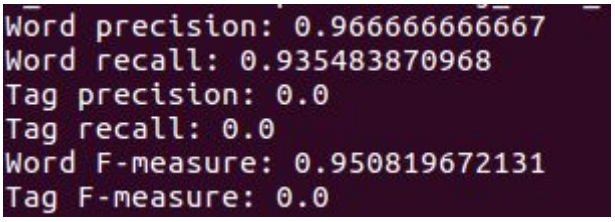
Nous allons tout d'abord effectuer une première expérimentation simple sur une seule phrase:

```
'When it's time for their biannual powwow, the nation's manufacturing titans typically jet off to the sunny confines of resort towns like Boca Raton and Hot Springs.'
```

Grâce à Lima, nous avons généré un fichier .conll qui représente l'ensemble des mots de la phrase associés à un *TAG* et un *TYPE*, ainsi que d'autres informations mais que nous ne traiterons pas ici (i.e. liaison sujet-verbe). Voici un extrait de ce fichier.

```
1       When    when    ADV     ADV     __      _       _       _       _
2       it      it      PRON    PRON    _       _       3       SUJ_V   _       _
3       's      be      V       VERB    _       _       _       _       _       _
```

Ces données seront traitées afin qu'elles puissent être plus facilement comparables. Nous allons donc adopter le format de MOT_ETIQUETTE.



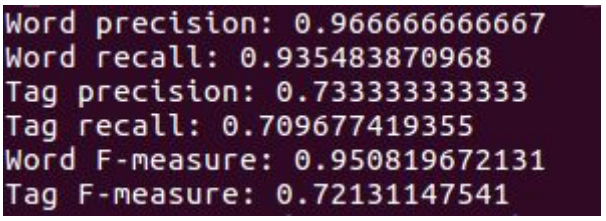
Résultats 1er test

Comme nous pouvons le voir ici, le rappel de mot et la précision sont corrects mais pas parfaits, mais surtout nous avons des résultats sur les tags très faibles. Le premier test nous a été demandé avec le fichier de référence fourni, or celui-ci utilise les tags universels (ou PTB). Comme LIMA possède ses propres tags (qui ne correspondent ni à PTB ni aux tags universels), alors le résultat est donc biaisé par la différence des tags.

Les résultats de ce test n'auront donc que très peu de valeur.

- **2nd test**

Nous allons effectuer un second test avec les mêmes conditions que le premier sauf que nous avons traduit toutes les étiquettes de LIMA dans le format universel. Nous n'aurons donc plus le problème du test précédent.



Résultats 2ème test

Cette fois-ci le résultat est tout autre, nous pouvons d'abord déduire que LIMA affiche un meilleur score sur la précision que sur le rappel. Nous pouvons noter que Lima possède plusieurs étiquettes qui ne seront représentées que par une seule étiquette universelle. Nous perdons donc de la précision lors de cette traduction de tag.

B. Stanford Core NLP

Cette partie est dédiée au tests de précision et de rappel pour évaluer les performances de l'outil Standford Core NLP.

Dans toute cette partie, nous reprendrons le même texte d'exemple que dans la partie dédiée à LIMA pour faire nos tests.

```
'When it's time for their biannual powwow, the nation's manufacturing titans typically jet off to the sunny confines of resort towns like Boca Raton and Hot Springs.'
```

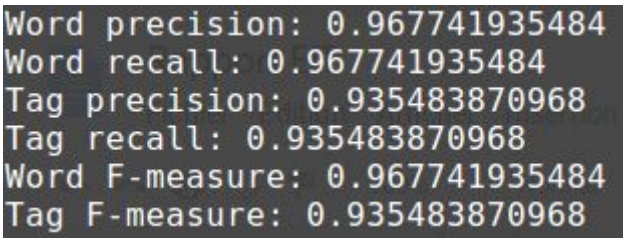
- **1er test** : **Étiquettes Penn Treebank (PTB)**

En appliquant l'outil, on obtient le texte sortie constitué de chaque mot du texte de base suivi de son étiquette, séparée par un underscore "_".

```
When_ADV it_PRON 's_VERB time_NOUN for_ADP their_PRON biannual_ADJ powwow_NOUN ,_. the_DET nation_NOUN s_PRT manufacturing_VERB titans_NOUN typically_ADV jet_VERB off_PRT to_PRT the_DET sunny_ADJ confines_NOUN of_ADP resort_NOUN towns_NOUN like_ADP Boca_NOUN Raton_NOUN and_CONJ Hot_NOUN Springs_NOUN ._.
```

Les étiquettes fournies par l'outil sont les étiquettes du format PTB.

En appliquant le script d'évaluation *evaluate.py* on obtient les résultats de précision de l'outil sur le texte d'exemple.

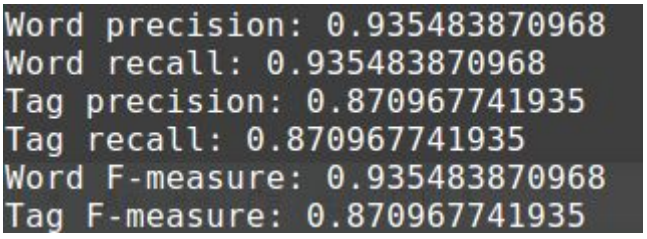


Résultats 1er test

On obtient une précision et un rappel de 96.77%, ce qui est un résultat qui nous semble correct étant donné qu'il n'est pas faible, mais pas parfait non plus.

- **2ème test** : **Étiquettes Universelles**

Dans ce second test, il s'agit de réitérer l'expérience en appliquant cette fois-ci un pré-traitement au texte de sortie de l'outil Standford afin de remplacer les étiquettes PTB par les étiquettes universelles.



Résultats 2e test

On constate que les résultats, bien que légèrement inférieurs, sont toujours corrects.

C. Comparaisons

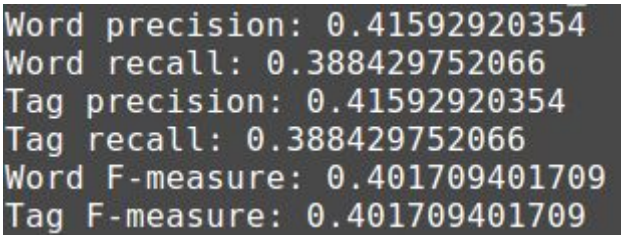
Cette partie est dédiée à la comparaison des deux outils d'analyse linguistique sur un corpus de référence commun. Elle servira plus précisément à comparer l'efficacité de leur reconnaissance d'entités nommées.

Le texte de test sera donc *formal-tst.NE.key.04oct95_small.txt* fournit par mail et le texte servant de référence sera *formal-tst.NE.key.04oct95_small.ne* fournit lui aussi par mail. Ces deux fichiers ne seront cependant pas utilisés en l'état.

En effet, il est nécessaire d'effectuer un prétraitement à la fois sur le texte de référence de base et sur le texte une fois traité par l'outil que l'on veut tester. En effet, les sorties des deux outils sont dans un format différent, et le script d'évaluation ne fonctionne que lorsque les entrées sont dans le format Mot_Etiquette. De plus, dans le cas de Standford, tous les mots du texte entrée sont contenus dans le texte sortie, ce qui n'est pas le cas de LIMA.

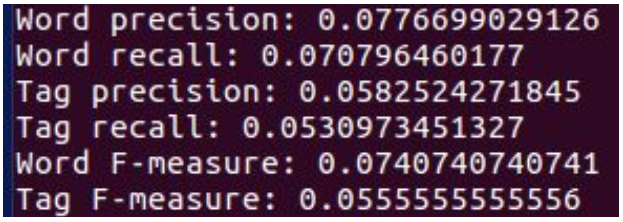
Nous obtenons les résultats suivants :

- **Standford**



Résultat de Standford

- **Lima**



Résultat de LIMA

On observe dans les deux cas que les résultats sont relativement faibles. Bien que le doute soit encore permis dans le cas de Standford, il est très difficile de croire que les résultats obtenus pour LIMA sont corrects. Il est difficile d'établir une réelle comparaison dans ce cas.

Cela est probablement dû à la manière dont sont créés les fichiers utilisés pour l'évaluation, ou à la manière dont l'évaluation elle même est faite. En effet, en comparant "manuellement" les fichiers utilisés pour l'évaluation dans le cas de LIMA, on observe que les différences ne sont pas si nombreuses au point d'avoir une précision de 7%. Si cette hypothèse s'avère, alors cela n'est malheureusement pas de notre ressort.

Certaines erreurs sont à noter, comme l'erreur d'associer "Clinton" ou "Washington" à un lieu et non une personne (LIMA).

Nous pouvons cependant déjà observer des différences entre les deux outils. En effet, Lima regroupe les entités formées de plusieurs mots (par exemple, *Consuela Washington*). Or ce n'est pas le cas de Stanford, ainsi nous perdons cette information.

Dans le cas où Lima et Stanford auraient des résultats (précision et rappel) proches, Lima seraient plus intéressants de part l'avantage de regrouper les mots composés.

Conclusion

Nous n'avons malheureusement pas réussi à obtenir de résultats sur le test qui nous permet de comparer les performances des outils de CEA list Lima et Stanford Core NLP. Cependant, cela ne signifie pas qu'il nous est impossible de les comparer en tant que tel.

En effet, on peut dire que le résultat de l'analyse linguistique de LIMA est plus riche en informations et fournit un résultat structuré sous forme de tableau. Cependant, cela donne un résultat avec parfois trop d'informations, là où le résultat de l'analyse de Standford est plus simple et possiblement moins riche. Ce dernier présente aussi l'avantage de ne pas "perdre d'information" lorsque l'on effectue une analyse des entités nommées.

Nous pouvons aussi soulever la différence d'utilisation des ***postags***, en effet, Lima possède ses propres étiquettes qui sont différentes de celles de Stanford. D'un côté nous avons **Stanford utilisant les étiquettes universelles** donc directement comparables avec de nombreuses références mais de l'autre nous avons **Lima qui utilisent un panel plus fourni** d'étiquettes. Dans ce cas-ci, une nouvelle fois Lima gagne en précision de l'information par rapport à Stanford.

Un point sur lequel il serait intéressant de comparer les deux outils serait le rapport entre leur degré d'efficacité et le sujet du corpus de référence utilisé (article de journal, article scientifique, article sportif, …).