EIT

Rapport

Comparaison d'outils d'analyse

Auteurs :
Mathias Bazin et Nathan Bonnard

Contents

1 Introduction

1.1 Objectifs

Nous allons dans ce rapport comparer deux outils d'analyse de textes : LIMA du CEA List et l'outil de l'université de Stanford, notemment sur l'attribution des POS-tags et la reconnaissance d'entités nommées.

Nous tenterons de mettre en évidence les points forts et limites de chaque outil et de comprendre les différences entre les résultats.

1.2 Etat de l'Art

Il existe aujourd'hui différents outils d'analyse de textes. Les plus utilisés sont CEA List LIMA, Stanford CoreNLP, NLTK et SpaCy. Tous utilisent l'approche statistique sauf Lima, d'où l'intérêt de les comparer à Lima.

Les résultats qu'ils obtiennet different pour chaque corpus de test de réference mais on obtient en général au maximum autour de 95 pourcents de reussite pour le pos tagging pour les meilleurs avec prétraitement et pré-entrainement (cf.$https : //aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)$ et 90 pour la reconnaissance d'entités nomméees ($http : //www.aclweb.org/anthology/N16 - 1030$).

La volonté d'étudier Lima et Stanford coreNLP est surtout due au fait que ces outils sont open source et utilisent des approches différentes.

1.3 Les outils

LIMA

LIMA est un outil développé par le CEA List qui fait l'analyse à partir de règles de grammaire préalablement écrites par des experts linguistes.

Lima commence par prendre le texte donné et le segmente en éléments unitaires appelés tokens. Ensuite vient l'etape d'analyse morphologique après laquelle le programme associe à chaque token du texte une étiquette (ou POS-tag) qui indique la fonction grammaticale du mot dans la phrase. Enfin Lima reconnait les entitées nommées et leur associe une catégorie.

Le point fort de Lima est que le fonctionnement par règle assure une détection stricte et donc ceci devrait en théorie donner une forte précision dans les résultats. Par contre si les règles sont trop strictes on pourrait s'attendre à voir un plus faible score de rappel. De plus, la rédaction de ces règles est une tâche très compliquée qui nécessite le travail de linguistes experts pour être faite.

Outil Stanford

L'outil développé par Stanford passe par une approche statistique en utilisant des algorithmes d'apprentissage qui sont d'abord entrainés sur des corpus annotés.

Le point fort de cette approche est qu'elle est plus souple, par l'apprentissage on s'attend à obtenir des résultats sur n'importe quel type de corpus, même des textes remplis de fautes de frappe/fautes d'orthographe, du moment que l'algorithme a été entrainé sur des corpus adaptés. On s'attend donc à trouver des résultats avec un fort rappel mais peut-être une précision moindre.

Un autre point à souligner est le fait que pour entrainer l'algortithme, il faut lui fournir un corpus de textes annotés. Bien que la tâche d'annnotation puisse être longue dt fastidieuse, elle est moins compliquée que la rédaction de règles et ne nécessite pas forcement d'être faite par des experts linguistes.

2 Résultats et commentaires

2.1 LIMA étiquettes morpho-syntaxiques

Pour l'évaluation sur les fichiers wsj0010.sentence on obtient les résultats suivants

Catégorie	Score (en pourcentage)
Word precision	96.6
Word recall	93.5
Tag precision	76.6
Tag recall	74.1
Word F-measure	95.0
Tag F-measure	75.4

On voit que la précision et le rappel au niveau des tags sont tous les deux aux alentours des 75 pourcents, ce qui est plutôt bon.

Malheuresement l'évaluation en utilisant les étiquettes universelles s'est montrée très compliquée. En effet, lorsque l'on analyse le texte avec lima, il ne segmente pas certains mots qu'il garde dans un seul token en tant que mot composé pour assigner un seul tag à ce dernier, certains mots sont aussi perdus, notemment les "'s" parfois. Bien que cela soit logique puisqu'il est naturel de considérer par exemple "Rust Belt" comme une seule entité, je crains que cela ait affecté en mal les résultats donnés par le script d'évaluation car nous allons le voir, les résultats sont moins bons. Certains fichiers contenaient aussi des caractères erronés qui n'étaient clairement pas à leur place comme un] apres un nom d'étiquette, ce qui a forcé le nettoyage parfois manuel de certains fichiers. Il est possible que cela ait affecté le résultat mais ces manipulations étaient nécessaires au déroulement du test.
Ces resultats ont étés donnés par l'évaluation sur les fichiers type wsj.sample.pos.lima.toText.

Lima sur wsj complet

Catégorie	Score (en pourcentage)
Word precision	56.4
Word recall	51.8
Tag precision	52.4
Tag recall	48.1
Word F-measure	54.0
Tag F-measure	50.2

Lima sur wsj complet et tags universels

Catégorie	Score (en pourcentage)
Word precision	56.4
Word recall	51.8
Tag precision	55.4
Tag recall	50.9
Word F-measure	54.0
Tag F-measure	50.2

Comme dit précédemment, on voit que les resultats sont bien moins bons mais cela est surement du à la modification des fichiers. Des changements sur la ponctuation pouvaient avoir de l'influence sur les résultats (ajouter des espaces avant et après augmentait les résultats par exemple) mais il a été choisi de modifier le moins possible les fichiers pour rester fidèle aux textes de réference.
On peut néanmoins comparer les resultats entre les tags PTB et Universels : on voit que les résultats en tags universels sont légèrement meilleurs poiur les tags, une hypothèse est que comme les tags universels sont plus généraux et moins précis que les tags Lima, la possibilité de se tromper est moindre. Par exemple Lima pourrait se tromper entre deux types d'adjectifs mais en universel il n'en existe qu'un type donc pas d'erreur.

2.2 Désambiguïsation morphosyntaxique de l'université de Stanford

Les résultats sont les suivants sur le fichier wsj0010.sample

Stanford sur wsj

Catégorie	Score (en pourcentage)
Word precision	96.7
Word recall	96.7
Tag precision	93.5
Tag recall	93.5
Word F-measure	96.7
Tag F-measure	93.5

Stanford sur wsj en tags universels

Catégorie	Score (en pourcentage)
Word precision	96.7
Word recall	96.7
Tag precision	93.5
Tag recall	93.5
Word F-measure	96.7
Tag F-measure	93.5

Les résultats sont exactement les mêmes, ce qui est plutot étrange. On pourrait penser qu'il existe une correspondances assez bonne entre les tags PTB utilisés par Stanford et les tags universels.

2.3 Reconnaissance d'entités nommées

Encore une fois il a fallu de nombreuses transformations pour faire correspondre les fichiers obtenus en sortie d'analyse avec les fichiers de réference pour pouvoir utiliser le script d'évaluation. En effet, Lima produisait des erreurs ConllDumper dont la source n'a pas été trouvée. Pour permettre l'évaluation il a été une nouvelle fois nécessaire de modifier les fichiers obtenus. Malgré cela, il a été possible de soutirer ces résultats à Lima et Stanford sur le fichier formal.small.

Stanford - Entités nommées

Catégorie	Score (en pourcentage)
Word precision	74.4
Word recall	84.1
Tag precision	61.2
Tag recall	69.2
Word F-measure	78.9
Tag F-measure	65.0

LIMA - Entités nommées

Catégorie	Score (en pourcentage)
Word precision	24.5
Word recall	22.4
Tag precision	11.3
Tag recall	10.3
Word F-measure	23.4
Tag F-measure	10.8

Ici on remarque que les scores de précision et de rappel sont très bas pour Lima, en effet le fichier contient beaucoup de termes ambigus qui peuvent facilement être confondus entre organisation et location comme le terme House qui peut à la fois désigner les deux. La méthode des règles semble être en diffuiculté dans ce cas.

Pour ce qui est de Stanford les scores sont raisonnablement élevés, à peu près au niveau de l'état de l'art dans ce domaine. On peut en déduire que l'approche statistique fonctionne mieux pour ce qui est de la reconnaissance d'entités nommées, ce qui semble intuitif car il n'y a pas vraiment de règle qui définit la forme que peut prendre un nom propre.

3 Répartition des tâches

Mathias : Partie technique et développement des scripts

Nathan : Partie analyse et rédaction du rapport

4 Conclusion

A travers ce projet nous avons pu apprendre beaucoup sur le fonctionnement de ces deux outils d'analyse et nous avons réflechi sur les points forts et points faibles des différentes approches qu'ils utilisent.

Nous sommes néanmoins perplexes face aux résultats obtenus, en particulier pour Lima. Nous esperons que les modifications et les erreurs trouvées dans les fichiers de réferences n'ont pas trop déformé les scores que Lima aurait du obtenir.