PROJET D'EIT - ET5 INFO

Evaluation d'outils d'analyse linguistiques

Anthony WOZNICA

Correspondence:
anthony.woznica@u-psud.fr
Full list of author information is available at the end of the article

Abstract

Les premiers travaux scientifiques en matière de Traitement Automatique des Langues remontent aux années 1950 avec la volonté d'élaborer une méthodologie de traduction automatique. Ces premiers travaux, initiés par Alan Turing dans son article "Intelligence" ont donné naissance aux critères d'évaluation de la qualité d'un agent conversationnel.

Initialement fondés sur un ensemble de règles (approche heuristique du langage), les outils de reconnaissance ont évolué dans les années 1980 en des méthodologies reposant sur des approches statistiques avec, entre autres, des approches par apprentissage statistique (autrement appelées, Machine Learning).

A l'heure actuelle, la tendance se dessine en faveur d'approches statistiques plus élaborées, basées sur des architectures en réseaux de neurones, parmi lesquelles il est possible de citer les RNN, rendus obsolètes par les LTSM puis enfin par les architectures dites Transformer.

L'objectif de ce document est d'élaborer une méthodologie de comparaison entre une approche statistique et une approche heuristique pour le TAL. Ces approches seront réalisées respectivements avec les outils open-source STANFORD CORE NLP et CEA LIST LIMA.

CEA List Lima: Développé CEA au sein du Department of Ambient Intelligence and Interactive Systems, l'outil Lima est un analyseur linguistique basé sur des règles et dictionnaires, supportant 5 langues. L'analyse se repose sur le principe de traitement en cascade configurable, correspondant au différents niveaux d'analyse linguistique. Lima permet également, grâche à son approche par unités de traitement (pipes), de tenir compte de manques de ponctuation, d'absence de voyelles ou de la concaténation de mots.

Stanford Core NLP: Développé par l'Université de Stanford, le Stanford Core NLP Toolkit se présente comme un framework JAVA d'analyse linguistique, reposant sur l'apprentissage statistique à partir de corpus annotés.

Keywords: Traitement Automatique des Langages; CEA List Lima; Stanford Core NLP

Documents

L'ensemble des résultats et du code ayant permis de fournir ces résultats est disponible sur https://github.com/thewozn/EIT-PPS-2019

Introduction

L'évaluation de plateformes d'analyse linguistique exige d'identifier et d'analyser les résultats de chaque module de l'analyse linguistique standard avant de procéder à une comparaison avec un corpus de référence.
Ainsi, lorsque cela est possible, nous observerons et interpréterons les résultats de chacun des modules suivants:

- **Tokenisation,** ou découpage des chaînes de caractères du texte en mots.
- **Analyse morphologique,** vérification de l'appartenance d'un mot à la langue et association de propriétés syntaxiques à chaque mot (catégories grammaticales).
- **Analyse morpho-syntaxique,** désambiguisation des mots non reconnus par l'analyse morphologique.
- **Analyse syntaxique,** identification des principaux constituants de la phrase et des relations qu'ils entretiennent entre eux.
- **Reconnaissance d'entités nommées,** expression linguistique référentielle associée aux noms propres et aux descriptions définies. Dans le cadre du rapport, seules sont considérées les références à des Lieux, des Personnes ou des Organisations. La reconnaissance associe un token à ces entités.

Ainsi, la comparaison s'effectuera tant sur les entités nommées que sur les entités grammaticales usuelles. Une analyse sera également effectuée sur un corpus conséquent pour la labélisation d'entités nommées. Les expérimentations seront menées et imagées à l'aide de l'extrait de texte suivant:

"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. Not this year. The National Association of Manufacturers settled on the Hoosier capital of Indianapolis for its fall board meeting. And the city decided to treat its guests more like royalty or rock stars than factory owners. The idea, of course: to prove to 125 corporate decision makers that the buckle on the Rust Belt isn't so rusty after all, that it's a good place for a company to expand."

On y dénombre en tout 7 entités nommées, à savoir:

Table 1 Entités nommées de l'échantillon

Entite Nommée	Type	Occurrences	Proportion
Boca Raton	LOCATION	1	14.28%(1/7)
Hot Springs	LOCATION	1	14.28%(1/7)
National Association of Manufacturers	ORGANIZATION	1	14.28%(1/7)
Hoosier	PERSON	1	14.28%(1/7)
Indianapolis	LOCATION	1	14.28%(1/7)
125	NUMBER	1	14.28%(1/7)
Rust Belt	LOCATION	1	14.28%(1/7)

On note également qu'il n'y a pas de répétitions, ou d'ambiguité sur les entités nommées (Par exemple, un cas hypothétique où Hoosier serait de type PERSON, mais Hoosier pourrait également être de type LOCATION).

Lima CEA List

Cette section à pour objet de couvrir les expérimentations effectuées sous Lima sur l'extrait de texte mentionné précédemment.

Avant de débuter, nous activerons les loggers *specificEntitiesXmlLogger* et *disambiguatedGraphXmlLogger*, qui permettent respectivement de fournir des fichiers contenant la reconnaissance d'entités nommées et les résultats de l'analyse morpho-syntaxique.

Extraction d'entités nommées

La lecture de la sortie de l'analyseur LIMA dans le fichier *sample.text.se.xml* fournit les résultats suivants:

Table 2 Reconnaissance d'entités nommées

Entite Nommée	Type	Nombre d'occurrences	Proportion
Indianapolis	LOCATION	1	14.28%(1/7)
Boca Raton	LOCATION	1	14.28%(1/7)
Rust Belt, Rust	LOCATION	2	28.56%(2/7)
National Association of Manufacturers	ORGANIZATION	1	14.28%(1/7)
resort	LOCATION	1	14.28%(1/7)
125	NUMBER	1	14.28%(1/7)

On note que seules 5 des 7 entités nommées du texte ont été reconnues et correctement typées. Observons les deux entités non/mal reconnues:

- **resort**, qui est un faux-positif.
- **Hot Springs**, qui n'est pas reconnu.

Dans le premier cas, celui de l'entité **resort**, on observe l'expression *the sunny confines of*, dont la succession d'adjectifs pourrait laisser penser que le n-gramme de l'outil CEA List a pu atteindre une limite, et ainsi considérer le terme suivant comme une entité nommée (dans la sémantique, si Resort était effectivement une entité nommée, le n-gramme *the sunny confines of resort* aurait toujours eu un sens).

Le cas de **Hot Springs** semble plus délicat, puisqu'il n'est nulle mention de ce nom dans le fichier **.se.xml** contenant les entités nommées. Il en convient donc de se pencher sur le fichier **.conll** généré par Lima:

```
26  Hot Springs  Hot Springs  NP  PROPN  _  _  27  Dummy  _  _
```

On y retrouve Hot Springs sous l'étiquette NP (nom propre), portant à croire que l'entité aurait du être présente dans le fichier .se.xml généré. Néanmoins, on note également que contrairement aux entités nommées du document, l'entité Hot Springs ne présente aucune étiquette spécifique à son type, portant à croire que Lima n'a pas pu l'interpréter. En revanche, l'analyse sémantique effectuée pour l'extraction d'entités nommées (lien article) aurait du détecter que Hot Springs fait référence à un lieu. En revanche, on peut supposer qu'il s'agisse d'un manque de densité du corpus, ou d'un manque de mentions de Hot Springs dans le corpus.

On ajoute la phrase *"The city of Hot Springs is the place to be."* en début de corpus afin de valider ou d'invalider cette hypothèse, retournant le résultat suivant, dans lequel l'entité Hot Springs est effectivement reconnue:

Table 3 Reconnaissance EN après ajout d'éléments au corpus

Entite Nommée	Type	Nombre d'occurrences	Proportion
National Association	ORGANIZATION	1	9.09%(1/11)
Boca Raton, Boca	LOCATION	2	18.18%(2/11)
Hoosier	PERSON	1	9.09%(1/11)
Rust Belt, Rust	LOCATION	3	27.27%(3/11)
125	NUMBER	1	9.09%(1/11)
Indianapolis	LOCATION	1	9.09%(1/11)
Hot Springs	LOCATION	2	18.18%(2/11)

Analyse morpho-syntaxique

On s'intéresse à présent au POS-tagging des entités nommées. A noter, Lima nous fournit originellement le texte sous forme d'étiquettes PennTreeBank, tenant compte du pluriel des noms communs, rendant l'exercice de reconnaissance d'entités plus difficile mais plus précis qu'avec des étiquettes universelles. On obtient, pour l'échantillon de test, le texte étiquetté normalisé est disponible dans le dossier **Results/** du projet.

En comparaison au texte de référence, celui-ci obtient le score suivant:

Table 4 Scores Lima - Etiquettes PTB

Word precision:	0.972727272727
Word recall:	0.972727272727
Tag precision:	0.745454545455
Tag recall:	0.745454545455
Word F-measure:	0.972727272727
Tag F-measure:	0.745454545455

Il est possible de noter immédiatement un très fort taux de succès dans la segmentation, indiquant que le découpage du texte est très précis/que les étiquettes sont bien placées. Cela s'explique notamment par le principe fondamental de tokénisation d'une chaîne de caractères par Lima. En revanche, la désambiguation morpho-syntaxique semble avoir des taux de succès assez faibles en comparaison avec ceux de sa segmentation. Lorsque l'on compare les résultats, on remarque que l'essentiel des erreurs d'étiquettes provient de cas de conjugaison et/ou de déclinaisons de mots. Par conséquent, on peut supposer que lors du passage à des étiquettes universelles, le taux de succès devrait en toute logique augmenter.

Enfin, dernière remarque, certaines entités composées sont considérées comme des deux entités par le texte de référence, et comme une seule entité par Lima (Ex: Boca Raton_NP sous Lima, Boca_NNP Raton_NNP dans la référence), réduisant ainsi la précision. Ce point sera étudié plus tard, sur un corpus de plus grande taille.

Evaluation à l'aide d'étiquettes universelles

Nous nous appuierons sur une table de correspondance entre les étiquettes Penn Treebank et les étiquettes universelles.

La particularité de ces étiquettes universelles est qu'elles ne tiennent pas compte des déclinaisons dans les termes. En conséquence, un analyseur qui détecterait le verbe mais non sa conjugaison aurait tout de même un certain taux de succès.

En revanche, l'intérêt majeur des étiquettes universelles est de pouvoir réduire le nombre d'étiquettes et ainsi se focaliser sur les catégories. Cela permettra également de confirmer notre hypothèse précédente, comme quoi le taux de succès de l'analyseur morpho-syntaxique est biaisé par la présence de déclinaisons. Ainsi, on met à profit la **propriété de surjection des étiquettes PTB vers les étiquettes universelles**. Après conversion, les scores de l'analyseur CEA LIST LIMA sont disponibles sur le **Tableau 5**.

Conformément aux attentes, la précision de la segmentation n'a pas varié, mais la précision de l'analyse morpho syntaxique s'est améliorée (+6%). Néanmoins, on peut également constater que les textes analysés sous Lima ne sont pas adaptés à l'évaluation sans avoir au préalable traité le texte en profondeur. En effet, la précision aurait été bien meilleure si les mots composés étaient considérés comme

Table 5 Scores Lima - Etiquettes universelles

Word precision:	0.972727272727
Word recall:	0.972727272727
Tag precision:	0.818181818182
Tag recall:	0.818181818182
Word F-measure:	0.972727272727
Tag F-measure:	0.818181818182

deux tokens distincts. Cependant, cela serait au risque de poser des problèmes de sémantique dans le cas d'une application concrète de l'outil.

Stanford NLP Core

Cette section à pour objet de couvrir les expérimentations effectuées sous Stanford NLP Core sur l'extrait de texte mentionné en introduction.

Analyse morpho-syntaxique

Nous débuterons cette partie par une analyse morpho-syntaxique du corpus de texte donné en introduction. Les résultats sont les suivants:

Table 6 Scores Stanford Core - Etiquettes PTB

Word precision:	0.981818181818
Word recall:	0.990825688073
Tag precision:	0.954545454545
Tag recall:	0.963302752294
Word F-measure:	0.986301369863
Tag F-measure:	0.958904109589

On note une précision importante tant dans l'étiquettage que dans la segmentation. Lors du passage en étiquettes universelles, les résultats deviennent les suivants:

Table 7 Scores Stanford Core - Etiquettes universelles

Word precision:	0.990825688073
Word recall:	0.981818181818
Tag precision:	0.963302752294
Tag recall:	0.954545454545
Word F-measure:	0.986301369863
Tag F-measure:	0.958904109589

Avant de poursuivre l'analyse, il est essentiel d'éclaircir les notations:

- **Precision:** Proportion d'entités correctement prédites.
- **Recall:** Proportion d'entités correctement prédites par rapport au nombre total d'entités de la même étiquette (*sensibilité statistique*)
- **F-Measure:** Moyenne harmonique de la précision et du recall. Elle est un **estimateur de la qualité de la prédiction**.

La valeur qui nous intéresse en premier est celle nommée **F-measure**. Le fait qu'elle ne varie pas entre les deux tests montre que **la performance du Stanford POS Tagger ne varie pas lors du passage aux étiquettes universelles**.

D'autre part, on observe également que le les taux de précision et de recall ne font que s'inverser, ce qui montre qu'aucun changement n'est opéré lors du passage en étiquettes universelles.

Extraction d'entités nommées

La lecture de la sortie de l'analyseur LIMA dans le fichier sample.txt.ner.stanford fournit les résultats suivants:

Table 8 Reconnaissance d'entités nommées

Entite Nommée	Type	Occurrences	Proportion
Boca Raton	LOCATION	1	25%(1/4)
Hot Springs	LOCATION	1	25%(1/4)
National Association of Manufacturers	ORGANIZATION	1	25%(1/4)
Indianapolis	LOCATION	1	25%(1/4)

On note que seules 4 des 7 entités nommées ont été reconnues. A la différence de l'analyseur d'entités nommées Lima, celui-ci reconnaît bien la totalité de l'organisation *National Association of Manufacturers* ainsi que *Hot Springs*. Cependant, on observe un taux de reconnaissance bien inférieur à celui de l'outil du CEA. En effet, les entités **Hoosier, 125 et Rust Belt ne sont cette fois pas reconnues**. Deux de ces trois entités non reconnues sont de type **PERSON** et **NUMBER**, types d'entités peu présentes dans le corpus. En revanche, Rust Belt, qui est de type LOCATION n'est pas non plus reconnu. Une rapide analyse de la sémantique pourrait éventuellement expliquer la confusion: *"the buckle of the Rust Belt"* a tout autant de sens que *"the buckle of the rust belt"*.

En revanche, **Rust Belt est bien détecté comme une entité nommée par le POS Tagger,** ce qui en toute logique signifie qu'il y a bien détection du terme, mais que celui-ci n'est pas d'un type reconnu !

Cependant, on peut observer une forte dissention entre les scores NER et ceux du POS Tagger, montrant les limites de la reconnaissance par apprentissage statistique (qui, en toute logique, est moins bon pour la reconnaissance d'entités nommées, puisque celles-ci ne sont pas nécessairement présentes dans la langue d'apprentissage, donc il y a potentiellement peu voir aucune occurrence de ces dernières dans les textes d'entraînement).

Comme tout apprentissage statistique cependant, celui-ci s'améliore avec la taille des données en entrée. Nous aurons l'occasion d'expérimenter cela avec un corpus plus important en dernière partie de rapport.

Reconnaissance d'entités nommées du CEA List et de l'université de Stanford

On cherche à présent à comparer les performances de l'outil du CEA et de celui de l'Université de Stanford. Pour cela, on convertit les étiquettes de l'outil du CEA List en étiquettes de l'université de Stanford, on ne se contente plus d'ajouter l'étiquette à la fin de l'entité, mais on l'ajoute à chaque unité de l'entité comme suit:

Hot Springs_NP devient Hot_NP Springs_NP
Boca Raton_NP devient Boca_NP Raton_NP

On utilise un corpus de référence trouvable dans **Files/format-tst**.

Les scores sont les suivants:

Table 9 Scores CEA List NER

Word precision: 0.460784313725
Word recall: 0.485370051635
Tag precision: 0.428104575163
Tag recall: 0.450946643718
Word F-measure: 0.472757753562
Tag F-measure: 0.43922883487

Table 10 Scores Stanford Core NER

Word precision: 0.976149914821
Word recall: 0.986230636833
Tag precision: 0.948892674617
Tag recall: 0.958691910499
Word F-measure: 0.981164383562
Tag F-measure: 0.953767123288

NOTE: il y a un décalage de la ponctuation qui n'a pas été modifié et qui explique les résultats moyens de l'outil CEA et qui n'a pas été corrigé dans les temps.

En revanche, l'hypothèse selon laquelle la performance de l'apprentissage statistique s'améliore avec la quantité de données a ici été vérifiée. Ainsi, on note ici la force de la plateforme Stanford Core NER pour la reconnaissance d'entités nommées sur les grands corpus.