RIBEIRO
Nathan
ET5 IIM

Rapport de Traitement Automatique des Langues

Objet du projet

L'objectif de ce projet est d'effectuer une analyse comparative des systèmes de POS tagging et de détection d'entités nommées du package python nltk et du CoreNLP Java de Stanford.

L'analyse de ces systèmes va être effectuée comparaison avec deux corpus annotés par des humains, un pour le POS tagging et un pour la détection d'entités nommées.
Le premier est annoté avec des étiquettes au format CoNLL-U (https://universaldependencies.org/format.html) et le second avec des étiquettes au format CoNLL-2003.
Pour la partie POS tagging, les étiquettes seront toutes converties au format universel depuis les format CoNLL-U et PTB pour comparaison. Pour la partie détection d'entités nommées les étiquettes générées par NLTK et CoreNLP seront converties au format CoNLL-2003.

Lors de la comparaison, on s'attachera à déterminer la précision et le rappel des deux systèmes par rapport à la référence humaine.
Précision : nb documents correctement attribués à la classe i / nb documents attribués à la classe i
Rappel : nb documents correctement attribués à la classe i / nb documents appartenant à la classe i
Note : A cause de la structure de la comparaison, la précision et le rappel seront proches, et même indissociables dans le cas de la détection d'entités nommées.

Résultats

<u>POS Tagging :</u>

- NLTK :
  - Precision : 0.6820395039044557
  - Rappel : 0.736946595195553
- CoreNLP :
  - Precision : 0.6897565457050988
  - Rappel : 0.745284891800675

<u>Détection d'entités nommées :</u>

- NLTK :
  - Precision : 0.8355184743742551
  - Rappel : 0.8355184743742551
- CoreNLP :
  - Precision : 0.8443583631307111
  - Rappel : 0.8443583631307111

On peut déjà noter à l'aide de ces résultats que le système CoreNLP semble avoir un léger ascendant sur le système NLTK, mais il convient de chercher à déterminer les causes de cette supériorité statistique et d'éventuelles lacunes non visibles ici.

Points forts, faiblesses, pistes

POS Tagging :

On peut remarquer assez rapidement que le corpus annoté utilisé pour l'analyse du POS tagging effectue des groupements de mots contrairement à NLTK ou au CoreNLP qui gardent tous les mots séparés.

Ainsi on a par exemple :

New England Journal of Medicine NOUN

qui devient :

New NOUN
England NOUN
JournalNOUN
of ADP
Medicine NOUN

ou :

55 years old ADJ

qui devient :

55 NUM
years NOUN
old ADJ

On peut supposer qu'il serait possible d'améliorer ce résultat en effectuant une passe de détection d'entités nommées avant le POS tagging ce qui rendrait peut-être possible de détecter au moins les entités et d'éviter de les éclater lors du POS tagging.

Une erreur qui semble exclusive et systématique au système NLTK est celle d'étiqueter une phrase ad positionnelle comme prétérit.

Par exemple :

open ADJ
only ADV
to ADP
institutions NOUN

devient :

open ADJ
only ADV
to PRT
institutions NOUN

Plus exactement d'étiqueter "to" comme prétérit systématiquement. L'intégralité des "to" traité par NLTK ont été étiquetés PRT là où près de la moitié ne l'étaient pas.

Les deux systèmes semblent régulièrement mal étiqueter des conjonctions en phrases ad positionnelles ou déterminants, des verbes en adjectifs, des noms en adjectifs ou en nombres. Même si le système NLTK semble plus propice aux erreurs d'étiquetage de verbes et de noms en adjectifs

Il semble donc dans l'ensemble que le système CoreNLP est globalement plus juste que le système NLTK en POS tagging et n'a pas d'erreurs qui lui sont spécifiques

Détection d'entités nommées :

On peut commencer par noter l'utilisation par NLTK du tag PTB GPE (Geo-Political Entity) qui n'est pas utilisé de façon consistante dans la littérature du Penn Treebank qui est un premier point de divergence avec la référence mais aussi le CoreNLP et n'était pas inclus dans les dictionnaires utilisés lors des tests.

On peut continuer en remarquant que les deux systèmes semblent ne pas inclure de "the" avant leurs entités nommées dans leurs entités nommées, par exemple :

```
the     B-LOC
United  I-LOC
States  I-LOC
```

devient :

```
the     O
United  B-LOC
States  I-LOC
```

On peut finalement remettre en question la validité de l'évaluation des entités nommées en observant des cas très étranges d'étiquetage dans la référence, tels que :

```
General         O
Motors  O
recovered       O
from    O
their   O
bankruptcy      O

the     O
United  O
States  O

the     O
US      O
government      O

a       O
year    O
after   O
the     O
appointment     O
of      O
Akerson         O
,       O

Republican      O
Senator         O
Bob     O
Corker  O

Japan   O
's      O
Nikkei  O
```

Tokenisation :

Un certain nombre d'erreurs d'évaluations peuvent être déplorées par un problème de tokenisation qui est commun aux deux évaluations et n'est pas directement lié aux fonctionnalités de POS tagging ou de détection d'entités nommées. Les deux processus pour les deux systèmes utilisent le tokeniser du package NLTK en pré traitement et celui ci sépare parfois un unique élément en deux, décalant les éléments de la phrase. Cette séparation semble être uniquement effectuée pour la contraction du verbe "is" avec un nom, par exemple :

Donoghue /// 's
government /// 's
funds /// '

Performance :

Un dernier point de divergence entre les deux systèmes est la performance, en effet (tout du moins en utilisant python avec nltk comme interface vers CoreNLP) les traitements de CoreNLP ont une durée entre 10 et 100 fois plus grande, ce qui pourrait un facteur déterminant de choix entre les deux systèmes lorsque de grandes quantités de données doivent être traitées à cause du surcoût en temps, en machines et en énergie pour un gain de précision et de rappel de l'ordre du pourcent.

Contributeurs

Nathan RIBEIRO