L2 MIASHS, Université de Lorraine

Ingénierie linguistique : TD 9

Chuyuan Li, Marie Cousin

Mars 17, 2023

1 Reprise du TD8

1. Combien y a-t-il d'EN dans les exemples suivants :
    1. Et si Éliane Houlette, la procureur du parquet national financier (PNF), tenait entre ses mains le sort de la prochaine élection présidentielle ?
    2. À l'Apple store, ce ne sont pas des jobs ordinaires
    3. Pour ses dix ans, le musée du Quai Branly pourrait bien changer de nom et porter celui de Jacques Chirac, le président de la République.
    4. Emmanuel et Brigitte Macron vont au Japon la semaine prochaine
    5. Le conseil de la vie universitaire de l'Université de Lorraine
2. Trouver deux exemples d'EN avec la polysémie.
3. Proposer des éléments caractéristiques des EN pour les personnes, les lieux et les organisations. Au niveau des EN elles-mêmes.
4. En fonction de leur contexte d'apparition.

2 Pratique avec NLTK et Spacy

2.1 Jeu de données

Ces deux phrases sont votre jeu de données. Pour chaque question, vous pouvez tester avec ces deux phrases.
test1 = "Marie says that it is raining in Paris but not in Nancy."
test2 = "European authorities fined Google a record $5.1 billion on Wednesday for abusing its power in the mobile phone market and ordered the company to alter its practices."

2.2 NLTK

NLTK (Natural language toolkit) est une librairie python qui permet de manipuler la langue naturelle. (https://www.nltk.org/index.html). En plus d'implémenter un très grand nombre d'outils pour le TAL, elle permet l'interface avec plusieurs resources linguistiques (corpus, dictionnaires...).

1. Prérequis : installez Python3, nltk (*pip3 install nltk*). La commande *nltk.download()* permet de télécharger une ressource.
2. La fonction *nltk.ne_chunk* tague les entitées nommées d'un texte. Elle prend un entrée un texte segmenté (tokens) et tagué en catégories morphosyntaxiques (POS-tag). Implémenter la fonction *preprocess()* qui prend en entréee un texte, et retourne une liste de tuples (token, POS-tag).
3. Reprendre le texte de l'exercice 1. Après l'avoir segmenté et tagué, appliquez *ne_chunk* et affichez le résultat. Qu'observez vous ?
4. L'arbre obtenu en sortie de *ne_chunk* est de type *nltk_tree*. Pour récupérer la liste des entitées nommées obtenues, vous pouvez procéder de cette manière:

- générer tous les sous-arbres de l'arbre obtenu (fonction *subtrees()*)
- parcourir les sous-arbres, et récupérer ceux qui ont pour label ['PERSON', 'GPE', 'DATE', 'ORGANIZATION'] (fonction *label()*)
- afficher la liste des entitées nommées

2.3 Spacy

La librairie spacy permet également d'interagir avec des données en langue naturelle. Son implémentation de la reconnaisssance d'entités nommées repose sur un modèle statistique entraîné sur le corpus Ontonotes5.

Installer Spacy et le package en suivant ces commandes :

```
>>> pip install -U pip setuptools wheel
>>> pip install -U spacy
>>> python -m spacy download en_core_web_sm
```

1. Spacy fonctionne avec des objets *nlp* pour lesquels un ensemble de méthodes sont implémentées. Chargez le texte précédent ('test1' or 'test2') en objet *nlp* et stocker dans une variable ('doc1' ou 'doc2') .

2. Les entitées nommées de l'objet créé sont récupérables en appelant la propriété *.ents* de l'objet *nlp*. Vous pouvez obtenir le texte et le label de chaque EN en utilisant *.text* et *.label_*.
   Récupérer les entités nommée en parcurant le résultat de *.ents*. Comparez avec celles obtenues avec NLTK. Qu'observez vous ?