Valentin-Gabriel Soumah
soumahvg@gmail.com

Coreferee

Une Bibliothèque Python pour la Résolution de Coréférence :

Ajout du support pour le français

Plan

1) La Résolution de Coréférence
2) Les Outils Existants
3) Le fonctionnement de l'Outil
    a) La détection des Mentions
    b) La résolution d'Anaphore
    c) Appariement des Noms
    d) Construction des chaines de coréférence
4) Démonstration

1) La Résolution de Coréférence

- La **Coréférence** : Plusieurs syntagmes (**les mentions**) désignent le même référent. Une personne, un lieu, une date…
- En Français ces mentions sont généralement des syntagmes nominaux ou pronominaux.
- Une mention est liée à une mention précédente : celle-ci est son **antécédent**.
- Toutes les mentions qui coréfèrent mises bout à bout forment **une chaine de coréférence**.
- L'identification automatique de toutes les mentions qui peuvent potentiellement coréférer est appelée la **détection de mention.**

1) La Résolution de Coréférence

Jean 1 vient juste d'être engagé par Renault 2 .

L'entreprise 2 lui 1 a offert un salaire très confortable.

Le nouvel employé 1 et sa fiancée Marie 3 attendaient cette opportunité pour se 3 marier.

Ils 3 prévoient d'organiser la cérémonie en mai.

2) Les Outils Existants

Pour les langues autres que le français

- NeuralCoref (par Huggingface) : Anglais
- Stanford CoreNLP:
  - Anglais
  - Chinois
- AllenNLP : Anglais
- BERT and SpanBERT for Coreference Resolution : Anglais

2) Les Outils Existants

Pour le français

- DecoFRE (Grobol) : Entraîné sur ANCOR
- French-CRS (Mirzapour) : Entraîné sur ANCOR
- COFR (Rodrigo, Landragin, Oberle, Amalia…): Entraîné sur DEMOCRAT

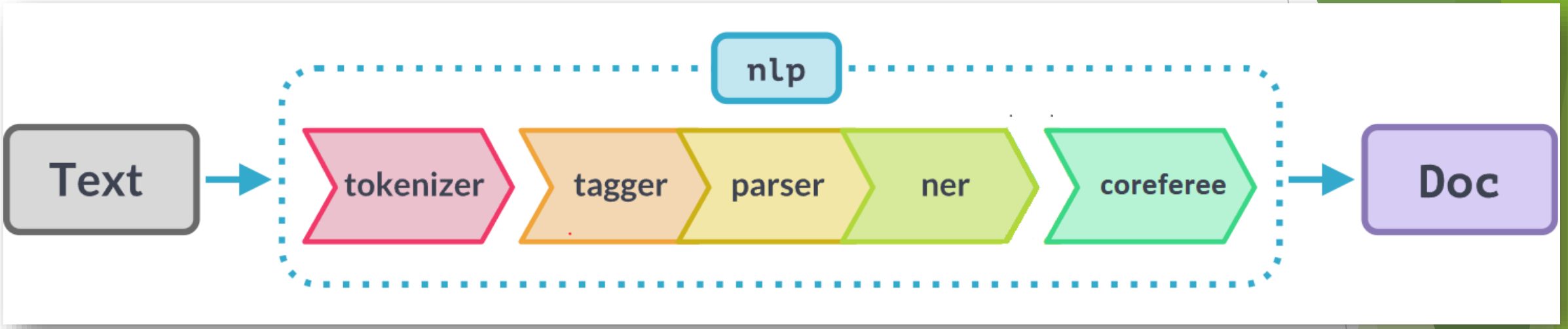
ANCOR (LI, LLL, Lattice): Premier large corpus en français annoté en coréférence et distribué librement. Corpus Oral

DEMOCRAT (Lattice, LiLPa, ICAR, IHRIM): Plus récent, corpus diachronique écrit comportant divers genres discursifs sur plusieurs siècles

Pourquoi coreferee?

- Bibliothèque basée sur spaCy
  - spaCy est utilisé par des universitaires et entreprises
  - Incorporable dans d'autres applications
- Mise en place et utilisation simple
- Bibliothèque extensible à de nouvelles langues
  - Modèles entraînables avec un corpus de taille limité
- Modèles pré-entraînés

3) Le Fonctionnement de l'Outil



- S'ajoute à la fin de la pipeline NLP de spaCy
- Utilise les composants précédents
- Trois modèles
  - fr_core_news_sm
  - fr_core_news_md
  - fr_core_news_lg
- Procède en deux étapes
  - Détection des mentions
  - Appariement des mentions
    - Résolution d'anaphore
    - Appariement de Noms

3) Le Fonctionnement de l'Outil
a) La Détection de Mentions

- Identifier les segments susceptibles de coréférer
- Utilise les annotations de spaCy
  - Parties du discours
  - Analyse syntaxique (en dépendance)
  - Morphologie

- Les mentions sont les têtes du syntagmes :
  - ~~Le maître de la classe de CE1~~
  - **maître**

3) Le Fonctionnement de l'Outil
a) La Détection de Mentions

- Les Noms Indépendants (Peuvent introduire la référence dans le discours):
  - Noms Propres
  - Noms Communs
  - Adjectifs Substantivés (Le premier, le grand…)
  - Pronom Possessif de la 3ème personne (Le sien..)
  - Une des pommes, certains des hommes..
- Les Anaphores (dont cataphores) :
  - Pronom Personnel troisième personne (sauf « on »)
  - Pro-adverbes/ Pronoms adverbiaux (« y », « en »)
  - Déictiques avec usage potentiellement anaphorique (« ici »)
  - Pronoms démonstratifs
  - Déterminants possessifs
  - Pronom réfléchi

3) Le Fonctionnement de l'Outil
a) La Détection de Mentions

Les Eléments non pris en charge

- Les premières et seconde personnes
- Le pronom « on »
- Pronoms relatifs (« que », « qui »…)
- Pronoms interrogatifs (« comment », « quoi »..)
- Pronoms démonstratif neutre (« ça », « cela »…)
- Syntagmes adverbiaux temporels (« en 1990 »…)

3) Le Fonctionnement de l'Outil

- Deux étapes successives pour résoudre la coréférence
  - Résolution d'Anaphore
  - Résolution de coréférence pour les paires de noms indépendants

3) Le Fonctionnement de l'Outil
b) La résolution d'anaphore

- Deux types de paires:
  - Nom Indépendant – Anaphore
  - Anaphore – Anaphore
- Deux étapes
  - Règles basées sur les annotations de spacy (pos-tagging, parser, entités nommées…) pour ne garder que les paires grammaticalement et sémantiquement possibles
  - Classement des paires restantes par un ensemble de 5 réseaux de neurones .

3) Le Fonctionnement de l'Outil
b) La résolution d'anaphore -

En entrée du réseau de neurones

- Tableau de traits (one-hot)
  - traits syntaxique, morphologiques, types d'entités nommées …
  - Pareil pour les dépendants de la tête
- Tableau de position (position dans la phrase, profondeur dans l'arbre syntaxique…)
- Vecteurs sémantique spaCy des têtes et de leurs dépendants
- Tableau de compatibilité
  - Position relative des deux têtes comparées
  - Nombre de traits partagés entre les deux têtes
  - Similarité cosinus des vecteurs sémantiques des deux têtes

*** 5 réseaux initialisés avec des poids différents :**

- **Entraînés sur DEMOCRAT (textes du XIXème au XXIème)**

En sortie

- 5 scores entre 0 et 1 : Probabilité que la paire coréfère
- Moyenne des 5 scores

3) Le Fonctionnement de l'Outil
c) Appariement des Noms

4 manières de lier les paires de nom

- Référence absolue : Noms propres
  - Match exact ou partiel du nom
  - Emmanuel Macron (…) Monsieur Macron
- Noms Propres – Noms communs
  - Restrictions grammaticales et sémantiques (entités nommées)
  - Emmanuel Macron (…) Le président de la République
- Mêmes substantifs
  - Restrictions selon le déterminant (un != le)
  - Le président de la République (…) Le président français
- Structure grammaticale lie les noms
  - Appositions
    - Emmanuel Macron, le président français (…)
  - Copule
    - Macron est le président français.

3) Le Fonctionnement de l'Outil
d) Construction des chaines de coréférence

- Les Paires sont mises bout à bout pour former une chaine
- Règles d'exclusion au niveau de la chaine entière
  - Pour anaphores
  - Pour paires de noms
- Anaphores choisies parmi celles avec le plus haut rang attribué par le modèle et compatibles avec la chaine entière

Emile Zola est l'auteur d'*Illusions perdues*. Le célèbre écrivain a écrit son premier roman à 30 ans.

- Emile – auteur
- Emile – écrivain
- Écrivain – son

⇨ Emile, auteur, écrivain, son

4) Démonstration

- Github de la bibliothèque : https://github.com/msg-systems/coreferee
- Github du projet d'ajout du français : https://github.com/Pantalaymon/coreferee_french
- Télécharger dans un dossier les fichiers :
  - demonstration_coreferee.ipynb
  - build_mentions.py
- Installer Python 3.9 (apt-get pour Linux sinon sur le site Python)

Ubuntu

```
apt-get update
apt-get install python3.9
```

Windows / MacOS

Télécharger sur le site officiel :
https://www.python.org/downloads/release/python-395/

4) Démonstration

- Créer un environnement virtuel

Ubuntu / MacOS

```
python3.9 -m venv coreferee-env
source coreferee-env/bin/activate

python3 -m pip install coreferee
python3 -m spacy download fr_core_news_lg
python3 -m coreferee install fr
```

Anaconda Prompt

```
conda create –n coreferee-env python=3.9
conda activate coreferee-env

python -m pip install coreferee
python -m spacy download fr_core_news_lg
python -m coreferee install fr
```

- Lancer le notebook avec cet environnement

Ubuntu / MacOS

```
pip install jupyter
```

Anaconda Prompt

```
conda install jupyter notebook
```

```
cd path/to/dir_with_notebook
jupyter notebook demonstration_coreferee.ipynb
```

Merci pour votre participation !

Des Questions ?