Tutoriel TALN 2022:
Quelques étapes souvent omises dans la préparation des corpus

Tian TIAN

Albeiro ESPINAL,

Yannis HARALAMBOUS

Plan

Outline

Chercheurs en TAL

- Tian TIAN :
  - Post-doctorante à IMT Atlantique, Brest;
  - Domaine de recherche : reconnaissance d'entités nommées, normalisation lexicale, méthode symbolique
- Albeiro ESPINAL :
  - Doctorant à IMT Atlantique. Consultant NLP à DSI Group;
  - Domaine de recherche : applications du traitement automatique de la langue aux processus d'embauche.
- Yannis HARALAMBOUS :
  - Enseignant-chercheur à IMT Atlantique, Brest;
  - Domaine de recherche : Fouille de texte, langages contrôlés et visuels, grapholinguistique

Outline

Objectifs du tutoriel

- Extraction de termes liés au domaine :
  - Définition du terme dans la terminologie
  - Définition de C-Values pour terme composé
- Extraction des entités nommées sous forme de mots composés
- Résolution d'anaphore

Le corpus à traiter :

- 60 articles du journal Le Monde sur l'ouragan Irma (2017).

Les outils à utiliser :

- Talismane,
- Grew.

Outline

Le Terme

- Un **terme**, définit en terminologie, est la ”représentation linguistique d’un concept”,
- Un **terme simple** contient un seul mot. Un **terme complexe** contient plus de deux mots.
- Frantzi et Ananiadou définissent la **termitude** comme une valeur (C-Value, Équation 1) définie en fonction de :
  - La fréquence des groupes de mots dans le corpus,
  - La fréquence des groupes de mots qu’ils contiennent,
  - La fréquence des groupes de mots qui les contiennent.

Groupe de mots: un motif spécifique de parties du discours (part-of-speech).

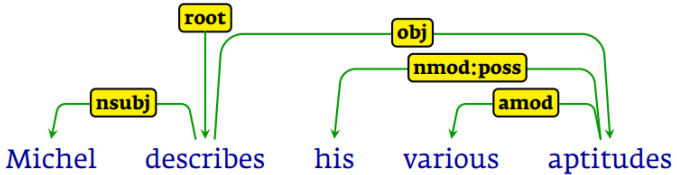
$$C(a) = \begin{cases} log_2|a| \cdot f(a) & si\ T_a = \emptyset \\ log_2|a| \cdot (f(a) - \frac{1}{\#T_a} \sum_{b \in T_a} f(b) & sinon. \end{cases} \quad (1)$$

Parties du discours traditionnelles

- **NC: Nom commun** (fille, maison, ouragan,...),
- **NP: Nom propre** (Paris, Emmanuel Macron, Thales,...),
- **PRON: Pronom** (je, moi, me, en, y, qui,...),
- **ADJ: Adjectif** (rouge, beau, premier,...),
- **V: Verbe** (parler, rire,...).
- **ADV: Adverbe** (aiment, souvent, alors, aussi,...),
- **ART: Article** (le, la, les),
- **PREP: Préposition** (à, de, en, entre, par,...),
- **CONJ: Conjonction** (et, ou, mais, comme, que,...),
- **INTERJ: Interjection** (ah!, euh!, hum!,...),

L'arbre de dépendances et la représentation CoNLL

- La manière standard de représenter un **arbre de dépendances** est la suivante :



- La représentation respective sous format **CoNLL** :

```
1 Michel Michel PROPN NNP Number=Sing 2 nsubj _ start_char=0|end_char=6
2 describes describe VERB VBZ Mood=Ind|Number=Sing|Person=3|Tense=Pres|
3 his he PRON PRP$ Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=P
4 various various ADJ JJ Degree=Pos 5 amod _ start_char=21|end_char=28
5 aptitudes aptitudes NOUN NNS Number=Plur 2 obj _ start_char=29|end_ch
6 . . PUNCT . _ 2 punct _ start_char=38|end_char=39
```

Outline

Présentation du corpus

Nous traiterons 60 articles du journal Le Monde sur l'ouragan Irma (2017).

Figure: WordCloud créé à partir du corpus.

Le fichier **Irma.txt** contient le corpus que nous analyserons lors de ce tutoriel.

Première Étape : Calcul de la Termitude

Soit a un candidat de n mots ($|a| = n$) et T_a l'ensemble des candidats qui contiennent a et notons $\#X$ le cardinal d'un ensemble X et $f(x)$ la fréquence d'un terme dans le corpus. Alors on définit la C-valeur de a par

$$C(a) = \begin{cases} \log_2 |a| \cdot f(a) & \text{si } T_a = \emptyset, \\ \log_2 |a| \cdot \left(f(a) - \frac{1}{\#Ta} \sum_{b \in T_a} f(b) \right) & \text{sinon.} \end{cases}$$

Pour calculer efficacement la C-valeur de tous les candidats, on commence par les groupes les plus longs, ce qui nous donne les ensembles T_* pour le calcul des sous-groupes de ceux-ci.

Exercise

Trouver les termes du corpus Irma en implémentant la formule de calcul de la C-valeur et en l'appliquant au document **irma.surf.conll**. En utilisera :

$$\mathcal{M} = \begin{cases} \text{NC ADJ+} \\ \text{NC P DET? NC (ou P peut être ”de”).} \end{cases}$$

Exemples : ”innovation conceptuelle”, ”centrale de pharmacie” (pour N P N), ”droits de l'homme” (pour N P DET N).

Attention :

- Certains nombres sont taggés en tant que NC ou en tant que ADJ, les éviter
- ne garder que la préposition *de* et le déterminant défini dans les motifs
- éviter les valeurs _ en tant que NC.

Deuxième étape : Entités nommées

Quand les entités nommées occupent plusieurs tokens dans le CoNLL, remplacer :

```
17 l'le D DET n=s 18 det _ _
18 hôpital hôpital N NC g=m|n=s|s=c 16 obj.p _ _
19 de de P P _ 18 dep _ _
20 Saint _ N NPP s=p 19 obj.p _ _
21 - - PONCT PONCT _ 22 ponct _ _
22 Martin martin N NC g=m|n=s|s=c _ _ _ _
```

par

```
18 hôpital hôpital N NC g=m|n=s|s=c 16 obj.p _ _
19 de de P P _ 18 dep _ _
20 Saint-Martin _ N NPP s=p 19 obj.p _ _
```

Exercise

Réunir également les noms propres en plusieurs mots sans ponctuation :

```
16 New _ N NPP s=p 15 obj.p _ _
17 York York N NPP g=m|n=s|s=p 16 mod _ _
```

devient

```
16 New_York _ N NPP s=p 15 obj.p _ _
```

Écrire le code qui

1. Réunit les cas (NPP, “-”, NC),
2. réunit les cas (NPP, NPP).

Faites attention aux dépendances, *avant et après* l’entité nommée.

Troisième étape : Résolution d'Anaphores

Il s'agit de trouver le référent d'une expression référentielle. Pour simplifier on va s'intéresser juste aux pronoms, par exemple dans "Mon père est venu. Il est arrivé hier", on cherche à détecter que "Il" se réfère à "père" (on écrira "Mon père$_1$ est venu. Il$_1$ est arrivé hier").

Pour trouver des candidats, on appliquera la méthode suivante :

1. On cherchera les NC ou NPP qui se trouvent dans la phrase précédente, la même phrase ou la phrase suivante, en restant dans le même document ;
2. on ne gardera que ceux qui sont masculins et au singulier ;
3. on calculera la distance entre référence et référent, pour ceux qui se trouvent après la référence on multipliera par un facteur ρ (par exemple 1.5, mais qui peut varier) ;
4. on gardera celui de plus faible poids.

- Un tutoriel pour l'utilisation et l'installation d'outils se trouve sur le site Github : `https://github.com/albeiroep/tutorial_taln2022`,
- Vous pourrez comparer progressivement votre méthode "maison" artisanale à des outils existants (TermSuite, Spacy),
- Nous sommes ici pour vous accompagner si vous avez des questions,
- À vous de jouer!