Automatiser l'écriture inclusive

Alice Hammel et Marjolaine Ray

Université de Paris – M1 Linguistique Informatique

24/06/2021

Outline

Table of Contents

Description de la tâche

- L'écriture inclusive
  1. Le problème du masculin neutre
     Les garçons et les filles sont studieux.
  2. Usage actuel : une représentation inclusive dans les communications d'organisations
     Les garçons et les filles sont studieux · ses.
     - Le cas des épicènes : artiste, personne, PDG
     - La création de mots qui contiennent plusieurs exposants. Les déterminants : le / la, les noms : déput é - e, écri vain · vaine
  3. Références génériques ou spécifiques à des groupes mixtes.
- L'objectif de la tâche et son intérêt
  - Automatiser la transformation de textes "traditionnels" en textes avec écriture inclusive

Approche

1. Création d'un corpus annoté "désinclusifié" à partir de textes écrit en écriture inclusive
2. Apprentissage SVM et forêt aléatoire
3. Evaluation $\rightarrow$ 95% d'*accuracy*
4. Création d'un outil de conversion de phrases

Table of Contents

Création du corpus

1. Récupération des textes
2. Détection des formes en écriture inclusive
   - Tokenization
   - Expressions régulières
   - Elimination mots-composés
   - Relecture et correction
3. "Désinclusification" de ces formes
   - Relecture et correction
4. Extraction des features

Extrait du corpus

```
# sent_id = 248
# text_no_ei =  Combien de fois montre-t-on l ' ertzaina qui charge contre des manifestants ?
 id          form        lemma upostag                                                xpostag feats  head     deprel deps misc
  0       Combien      combien     ADV                                        ADV__PronType=Int     _     4      advmod    _    _
  1            de           de     ADP                                                  ADP     _     4        case    _    _
  2          fois         fois    NOUN                          NOUN__Gender=Fem|Number=Sing     _     4      nummod    _    _
  3   montre-t-on  montre-t-on     ADV                                                  ADV     _     6        amod    _    _
  4             l            l    NOUN                                                 NOUN     _     6        amod    _    _
  5             '            '   PROPN                                                PROPN     _     0        ROOT    _    _
  6      ertzaina     ertzaina   PROPN                                                PROPN     _     6   flat:name    _    _
  7           qui          qui    PRON                                       PRON__PronType=Rel     _     9       nsubj    _    _
  8        charge      charger    VERB  VERB__Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin     _     6   acl:relcl    _    _
  9        contre       contre     ADP                                                  ADP     _    12        case    _    _
 10           des           un     DET                  DET__Definite=Ind|Number=Plur|PronType=Art     _    12         det    _    _
 11  manifestants  manifestant    NOUN                          NOUN__Gender=Masc|Number=Plur     _     9     obl:arg    _    ei=manifestant-e-s
 12             ?            ?   PUNCT                                                PUNCT     _     6       punct    _    _
```

Création et vectorisation des exemples

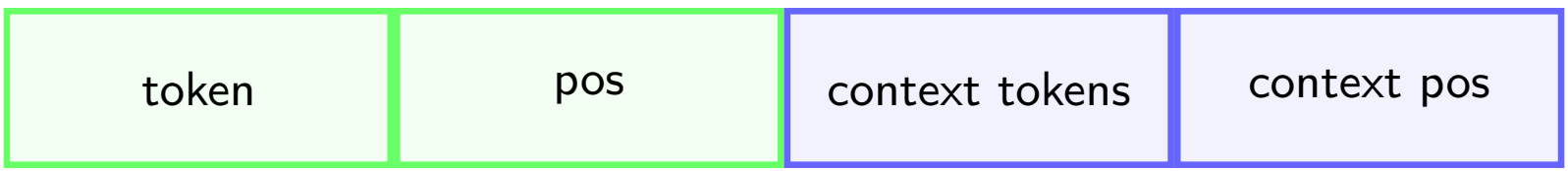
14 197 de classe positive, 4 134 506 tokens de classe négative

Création et vectorisation des exemples

14 197 de classe positive, 4 134 506 tokens de classe négative

Enfin , certains passagers peuvent intervenir ou être révoltés .

1 , PUNCT
2 certains DET__Gender=Masc - Number=Plur
3 passagers NOUN__Gender=Masc - Number=Plur
4 peuvent
VERB__Mood=Ind - Number=Plur - Person=3 - Tense=Pres - VerbForm=Fin
5 intervenir VERB__VerbForm=Inf



Baseline

Modèle (très) naïf qui classe dans la classe positive tous les noms, pronoms, adjectifs et participes passés.

Evaluation baseline sur corpus de **test**

ACCURACY	65.5%
RECALL	88.1%
PRECISION	19.5%
F1 SCORE	0.32

	pred 0	pred 1
gold 0	17916	10436
gold 1	341	2541

Modèle 1 : Forêt aléatoire

Ensemble d'arbres de décisions entraînés sur des sous-ensembles des données et des features.

Pourquoi ?

1. Efficacité pour données avec beaucoup de dimensions et une frontière de décision complexe
2. Rapidité de l'apprentissage
3. Interprétabilité
4. Facilité de paramétrage
5. Capacité de généralisation

Modèle 1 : Forêt aléatoire – Choix des hyperparamètres et évaluation

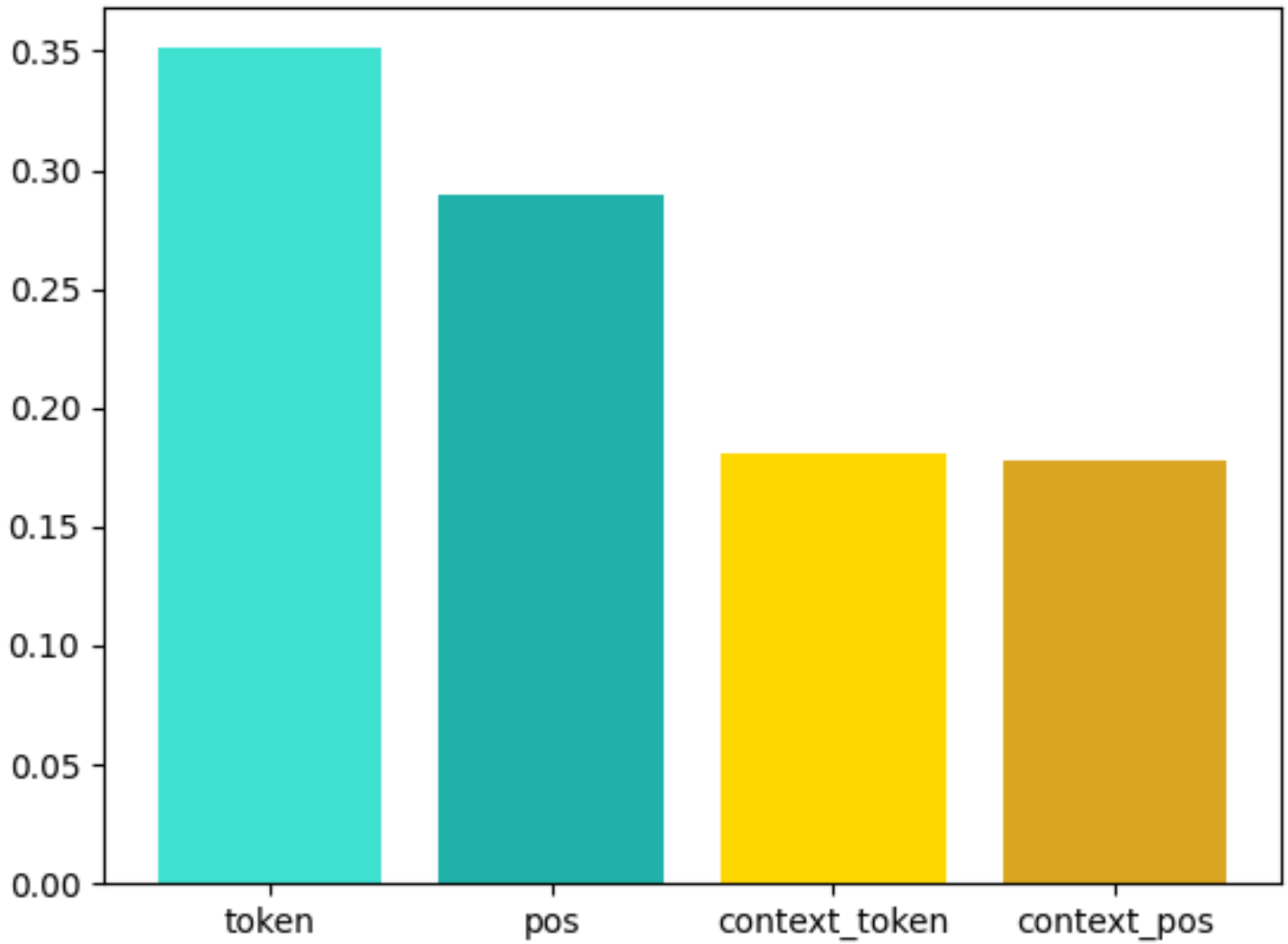
Choix des hyperparamètres `n_estimators` et `max_depth` par recherche de grille avec validation croisée.

`n_estimators` :150
`max_depth` : 30

	Baseline	**Random Forest**
Accuracy	65.4%	**88.7%**
Recall	88.1%	83.5%
Precision	19.5%	**44.2%**
F1	0.32	**0.57**

	pred 0	pred 1
gold 0	25312 (17916)	3040 (10436)
gold 1	473 (341)	2409 (2541)

Modèle 2 : Forêt aléatoire – Importance des features



Modèle 2 : Support Vector Machine

Caractérise les exemples d'entraînements pour maximiser la taille de l'espace entre les deux catégories. Il est couplé à un coeur Radial Basis Function

- Modèle linéaire, soft-margin avec un noyau non-linéaire car il y a peu de chance d'y avoir une frontière de décision linéaire dans notre cas.
- Problèmes de convergence du modèle et implication pour la recherche croisée (GridSearch)

Pourquoi?

-

Modèle 2 : Support Vector Machine

Défauts et avantages

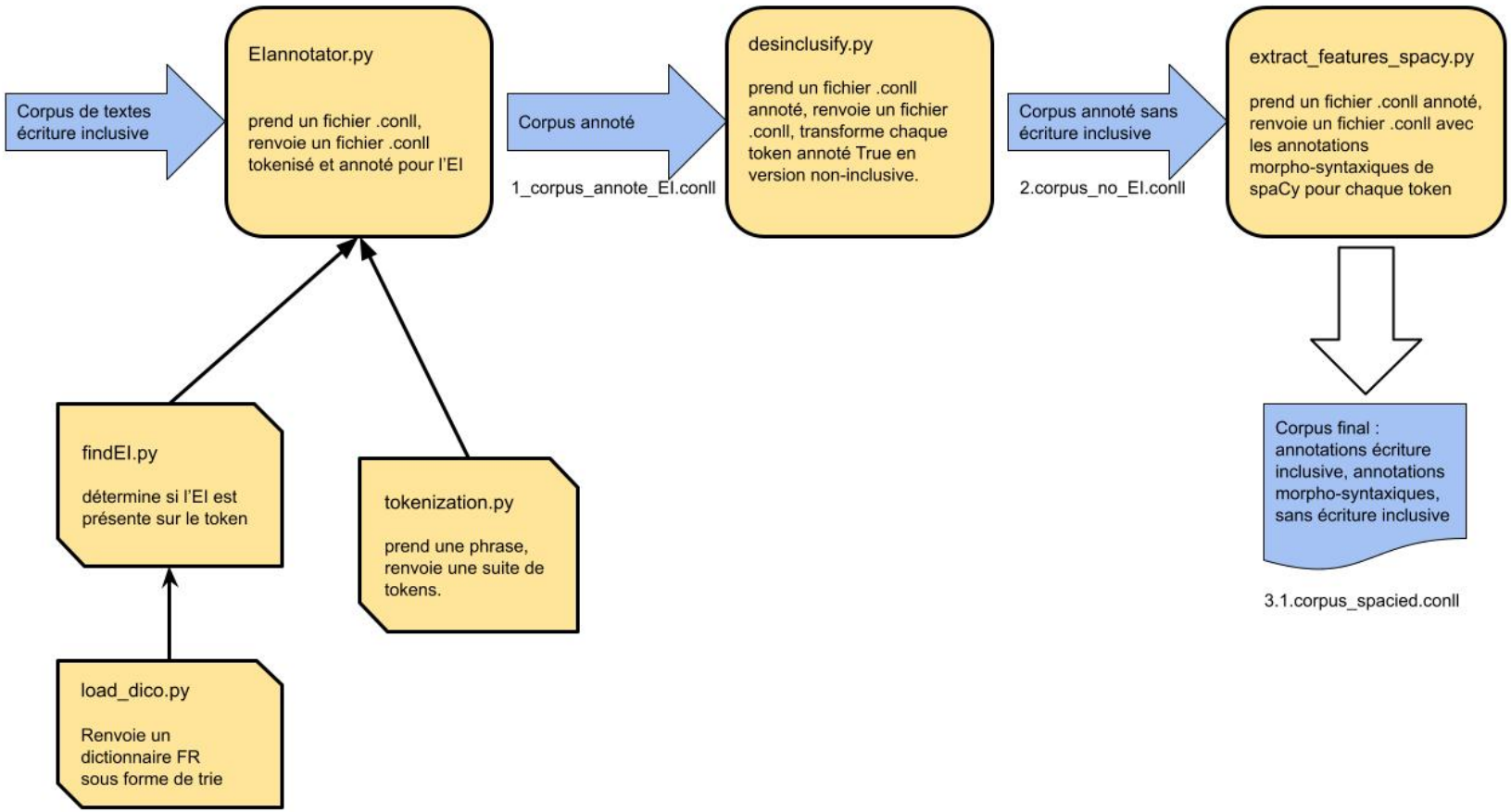
- Apprentissage lent
- Peu interprétable par rapport à RandomForest
- Des résultats excellents

Bilan des modèles

	Baseline	**SVM**	**Random Forest**
Accuracy	65.4%	**95.6%**	88.7%
Rappel	**88.1%**	70.8%	83.5%
Précision	19.5%	**79.6%**	44.2%
F1	0.32	**0.74**	0.57

Table of Contents

Organisation du code - Création du corpus



Organisation du code - Modèles

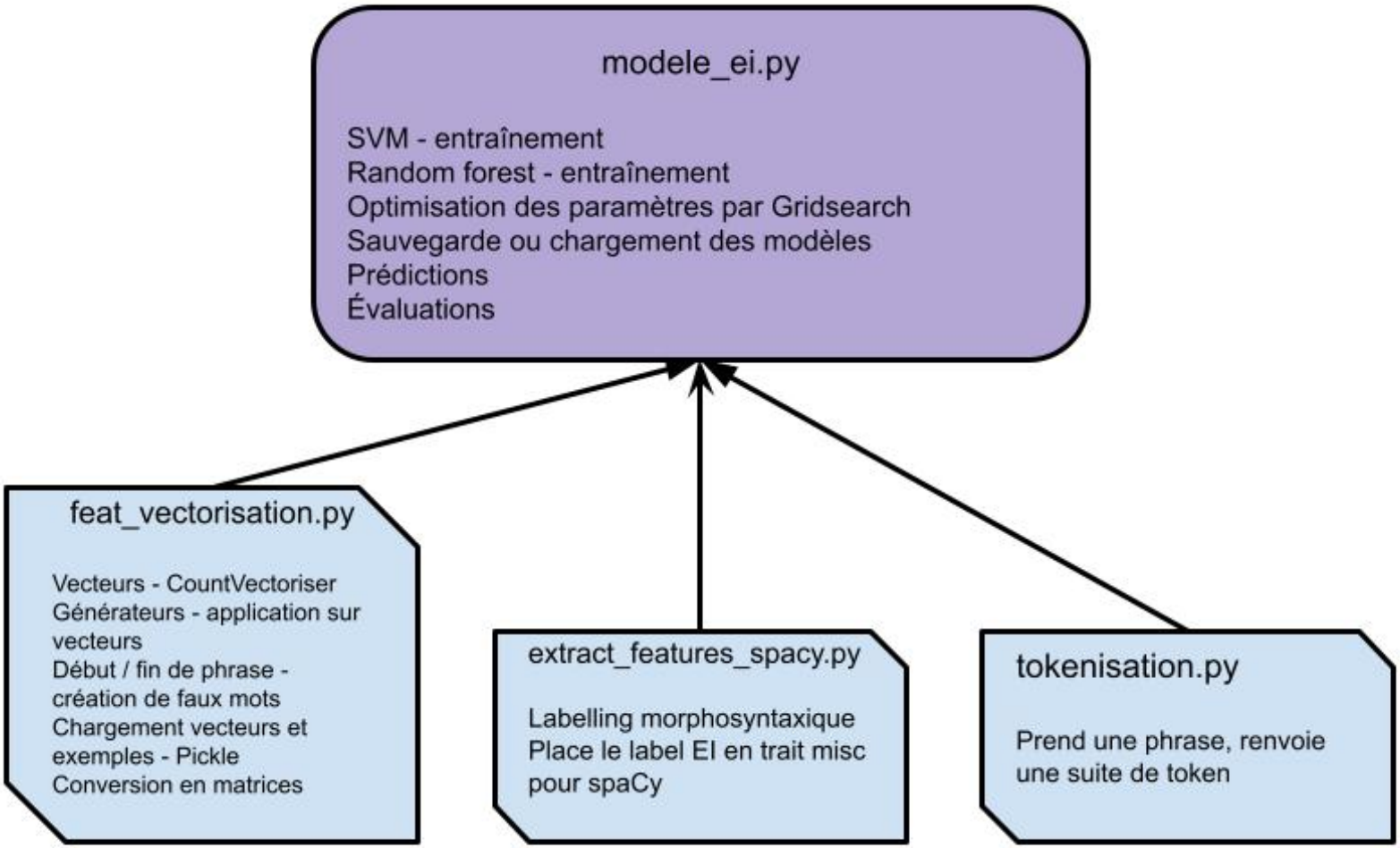


Table of Contents

Table of Contents

Pistes d'amélioration

1. Amélioration du corpus : diversifier, améliorer tokenisation, annotation plus exhaustive
2. Classifier des syntagmes plutôt que des tokens
3. Meilleur choix des features, plongements lexicaux, suffixes, infos sur le paradigme du lexème...
4. Implémenter un système de règles pour la transformation
5. Incorporer plus d'information sur le texte : plus grande fenêtre, RNN ?

Merci !