Reconnaissance d'entités d'intérêt dans les signatures d'email Nextino@TextMine2023

MAËLLE BRASSIER

ASCELINE GOUDJO

Contexte

- Nextino : centre d'innovation dans la data protection
  - Projet en TAL autour de la détection des données personnelles
  - Similitude entre Données Personnelles et Entités d'intérêt

Entités d'intérêt	Données personnelles (Nextino)
Human	**NAME**
Organization	ORGANIZATION (à venir)
Function	FUNCTION (à venir)
Project	-
Location	**ADDRESS** (détails)
Reference_CEDEX	**ADDRESS** (Cedex)
Reference_CS	**ADDRESS** (Distribution)
Reference_Code_Postal	**ADDRESS** (ZipCode)
Phone_Number	**PHONE**
Email	**EMAIL**
Url	-
Social_Network	-
Reference_User	-

Nextino
AI Powered Innovation

Approche par hybridation

- Solution hybride déjà mise en place dans notre projet
  - Reproduction de cette approche pour le défi afin de tirer profit des caractéristiques de chaque type d'entité d'intérêt

Approche symbolique

Pour les données au format et à la structure homogènes

Conditional Random Field

Pour les données qui sont souvent dépendantes du context avoisinant

Approche neuronale

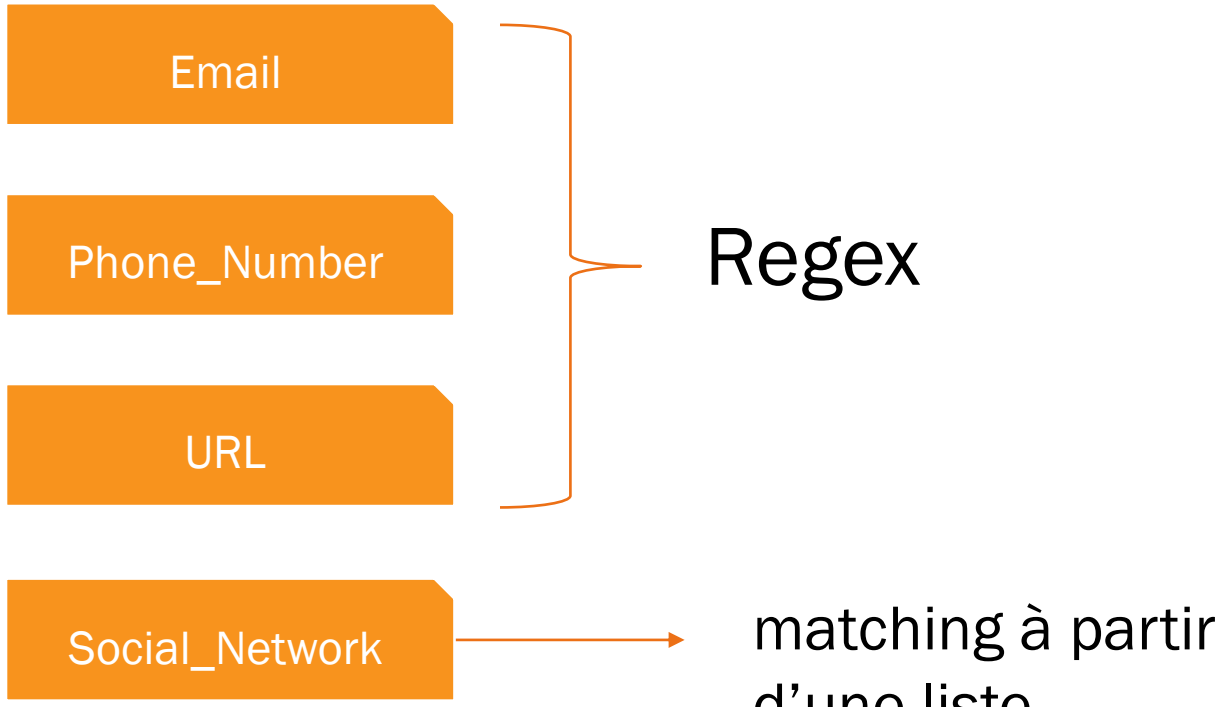
Pour les données dont on a besoin d'extraire une dimension sémantique

Nextino
AI Powered Innovation

Les approches
Approche symbolique

Approche symbolique

Se rapprocher des données et de leur représentation pour créer un ensemble de règles capable de les détecter de façon la + exhaustive possible



Les approches
Approche statistique

Approche statistique

Etude des similitudes et des rapprochements linguistiques pour déterminer la séquence de labels la plus probable étant donnée une séquence observée.

Méthode utilisée : CRF

Intérêt : détecter les données textuelles avec un caractère séquentiel et décrire un certain nombre de *Features* pour caractériser les éléments de contexte

Entités concernées :

- Location, Reference_CEDEX, Reference_CS, Reference_Code_Postal
  - Adaptation du modèle existant + train sur jeu TextMine
  - Exemples de features :
    - Is_street_name (from list)
    - Word_length
    - Is_digit
    - … description des features des tokens n+2 et n-2
- Reference User -> nouveau modèle
  - Exemples de features :
    - Contains_social_network
    - Is_url
    - Start_with_@
- Project -> nouveau modèle
  - Exemples de features
    - Trigger_words(département, licence, projet ,,,)
    - Is_digit

Nextino
AI Powered Innovation

Les approches
Approche neuronale

Approche neuronale

Utilisation d'un modèle pré-entrainé basé sur une architecture Transformer qui utilise le mécanisme d'attention pour modéliser les interactions entre les mots. S'appuie sur la probabilité d'une séquence de mots.

Modèles utilisés : famille BERT

Intérêt : détecter les données qui ne disposent pas d'indice interne* et qui nécessitent de s'appuyer sur le contexte

*Les indices internes permettent d'identifier une Entité Nommée en se basant uniquement sur l'ensemble des caractères qui la composent et de sa forme graphique.

Entités concernées :

- Human
  - Changement du modèle existant (Bert_base_multilingual) + train sur jeu TextMine
  - Preprocessing :
    - Ajout marqueur retour à la ligne
    - Harmonisation des espaces etc.
- Organization -> nouveau modèle (Bert_base_multilingual)
  - Preprocessing :
    - ajout marqueur retour à la ligne
    - Harmonisation des espaces etc.
- Function -> nouveau modèle ; utilisation de CamemBERT

Nextino
AI Powered Innovation

Post Traitement
Les règles par Entité

Exemples de règles après application des différents modules de détection pour les tokens sans LABEL

HUMAN

Ajout du label *Human* si un **token similaire** a déjà été détecté *Human* dans la signature

Exemple:

- Ganelon*[Human]* Grenier*[Human]* (Remplaçante de Ganelon Grenier durant son congé maternité)[…]

⇨ Ganelon*[Human]* Grenier*[Human]* (Remplaçante de Ganelon***[Human]*** Grenier***[Human]*** durant son congé maternité)[…]

LOCATION

Ajout du label *Location* si le token suivant est *Location* et qu'il s'agit d'un **digit**

Exemple:

- 105 Avenue*[Location]* Morane*[Location]* Saulnier*[Location]*

⇨ 105***[Location]*** Avenue*[Location]* Morane*[Location]* Saulnier*[Location]*

FUNCTION

Ajout du label *Function* si le premier token a été détecté comme *Function* sur la même ligne

Exemple:

- Élisabeth*[Human]* Grimard *[Human]* **\n**Professeur*[Function]* des Universités**\n** SkiSurfers.fr*[Url]*

⇨ Élisabeth*[Human]* Grimard *[Human]* **\n**Professeur*[Function]* des***[Function]*** Universités***[Function]*****\n** SkiSurfers.fr *[Url]*

Résultats

	precision	recall	f1-score	support
email	1.0000	1.0000	1.0000	344
function	0.9638	0.8088	0.8795	1449
human	0.9238	0.9222	0.9230	1196
location	0.9382	0.9361	0.9372	2678
organization	0.7988	0.8966	0.8449	1537
phone_number	1.0000	0.9927	0.9964	688
project	0.2837	0.4758	0.3554	124
reference_cedex	0.9776	0.8973	0.9357	146
reference_code_postal	0.9140	0.9798	0.9458	347
reference_cs	1.0000	0.0541	0.1026	37
reference_user	1.0000	0.8182	0.9000	11
social_network	1.0000	0.9286	0.9630	28
url	1.0000	1.0000	1.0000	303
accuracy			0.9065	8888
macro avg	0.9077	0.8239	0.8295	8888
weighted avg	0.9167	0.9065	0.9078	8888

F1: 0.9078304117108689

- Les regex affichent les meilleures performances

Résultats

	precision	recall	f1-score	support
email	1.0000	1.0000	1.0000	344
function	0.9638	0.8088	0.8795	1449
human	0.9238	0.9222	0.9230	1196
location	0.9382	0.9361	0.9372	2678
organization	0.7988	0.8966	0.8449	1537
phone_number	1.0000	0.9927	0.9964	688
project	0.2837	0.4758	0.3554	124
reference_cedex	0.9776	0.8973	0.9357	146
reference_code_postal	0.9140	0.9798	0.9458	347
reference_cs	1.0000	0.0541	0.1026	37
reference_user	1.0000	0.8182	0.9000	11
social_network	1.0000	0.9286	0.9630	28
url	1.0000	1.0000	1.0000	303
accuracy			0.9065	8888
macro avg	0.9077	0.8239	0.8295	8888
weighted avg	0.9167	0.9065	0.9078	8888

F1: 0.9078304117108689

- Les regex affichent les meilleures performances

- Les autres catégories affichent des résultats satisfaisants même si + d'erreurs pour des catégories qui peuvent se rejoindre (organization, function, location)

Résultats

	precision	recall	f1-score	support
email	1.0000	1.0000	1.0000	344
function	0.9638	0.8088	0.8795	1449
human	0.9238	0.9222	0.9230	1196
location	0.9382	0.9361	0.9372	2678
organization	0.7988	0.8966	0.8449	1537
phone_number	1.0000	0.9927	0.9964	688
project	0.2837	0.4758	0.3554	124
reference_cedex	0.9776	0.8973	0.9357	146
reference_code_postal	0.9140	0.9798	0.9458	347
reference_cs	1.0000	0.0541	0.1026	37
reference_user	1.0000	0.8182	0.9000	11
social_network	1.0000	0.9286	0.9630	28
url	1.0000	1.0000	1.0000	303
accuracy			0.9065	8888
macro avg	0.9077	0.8239	0.8295	8888
weighted avg	0.9167	0.9065	0.9078	8888

F1: 0.9078304117108689

- Les regex affichent les meilleures performances

- Les autres catégories affichent des résultats satisfaisants même si + d'erreurs pour des catégories qui peuvent se rejoindre (organization, function, location)

- Grosses difficultés sur les project qui sont à la fois peu détectés et mal détectés et CS qui n'en détectent que 2 faute d'indices

Conclusion

- Approche hybride qui a permis de tirer profit des données
- Difficultés sur la catégorie Project qui demanderait une investigation plus poussée
- Analyse des annotations gold pour mieux comprendre les erreurs

Merci pour votre attention ! Avez-vous des questions ?