Comparaisons et combinaisons d'extracteurs d'entités spatiales sur un corpus multilingue.

Zijian Wang[1], Yoann Dupont[2], Tian Tian[2], Gaël Lejeune[2]
Sorbonne Université, STIH,
[1] wangzijian1994@hotmail.com
[2] prenom.nom@sorbonne-universite.fr

24 Juin 2020

Plan

Contexte : Humanités Numériques et TAL

EN en humanité numérique : données

Extracteurs d'entités nommées existants

Résultats et discussion

Contexte : Humanités Numériques et TAL

Projet Antonomaz

- ANalyse auTOmatique et NumérisatiOn des MAZarinades [1]
  DIM STCN

1. http://lejeunegael.fr/antonomaz.html

Projet Antonomaz

- ANalyse auTOmatique et NumérisatiOn des MAZarinades [1]
  DIM STCN
- Le Traitement Automatique des Langues
- Les Humanités Numériques
- $\rightarrow$ leur interaction

Deux axes de réflexion :

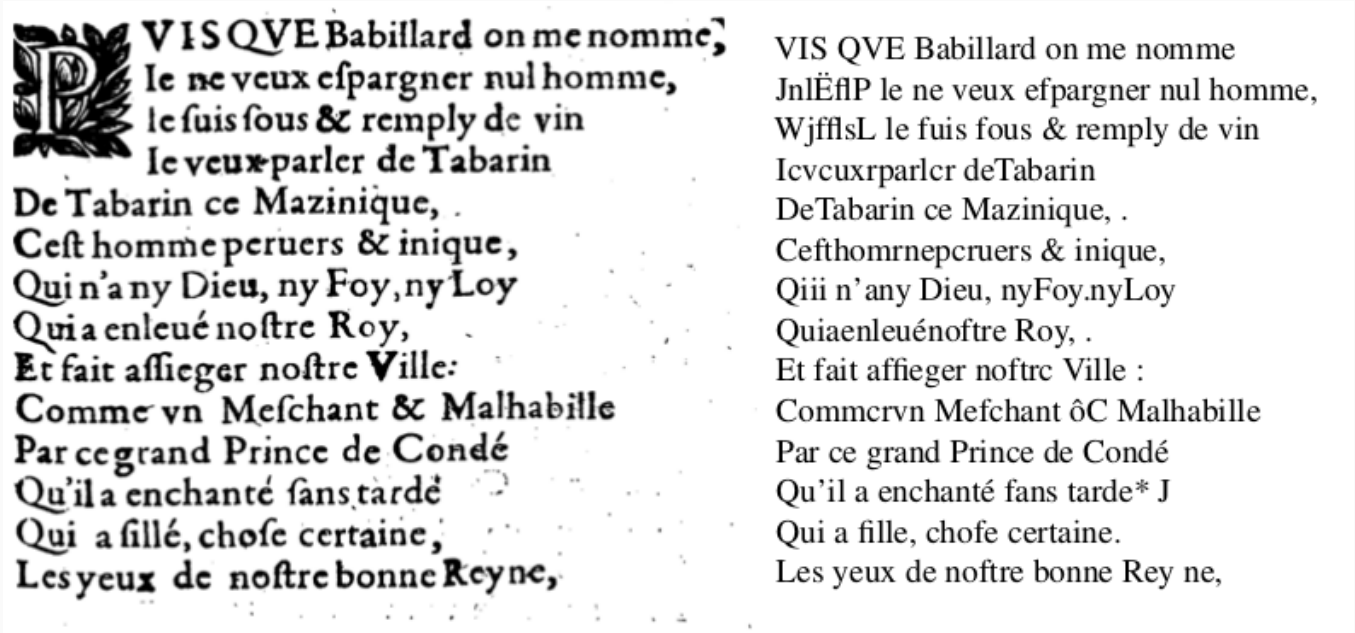
- *Input* : Documents numérisés, quelles contraintes ? (thèse de JB Tanguy 2019-...)
- *Output* : Adaptation/combinaison de systèmes de détection d'Entités Nommées (thèse de Caroline Parfait 2020-...)

1. `http://lejeunegael.fr/antonomaz.html`

Contraintes sur l'*Input*

Gérer la variation :

- Multilinguisme
- Hétérogénéité
- Bruitage



PVIS QVE Babillard on me nomme,
Ie ne veux eſpargner nul homme,
Ie ſuis ſous & remply de vin
Ie veux parler de Tabarin
De Tabarin ce Mazinique,
Ceſt homme peruers & inique,
Qui n'a ny Dieu, ny Foy, ny Loy
Qui a enleué noſtre Roy,
Et fait aſſieger noſtre Ville:
Comme vn Meſchant & Malhabille
Par ce grand Prince de Condé
Qu'il a enchanté ſans tarde
Qui a fillé, choſe certaine,
Les yeux de noſtre bonne Reyne,

VIS QVE Babillard on me nomme
JnlËflP le ne veux efpargner nul homme,
WjfflsL le fuis fous & remply de vin
Icvcuxrparlcr deTabarin
DeTabarin ce Mazinique, .
Cefthomrnepcruers & inique,
Qiii n'any Dieu, nyFoy.nyLoy
Quiaenleuénoftre Roy, .
Et fait affieger noftrc Ville :
Commcrvn Mefchant ôC Malhabille
Par ce grand Prince de Condé
Qu'il a enchanté fans tarde* J
Qui a fille, chofe certaine.
Les yeux de noftre bonne Rey ne,

Figure 1 – Source : Abiven et Lejeune 2019

Contraintes sur l'*Output*

Question des "conditions de laboratoire"

- Applicabilité des Systèmes état de l'art
- Évaluation/Interprétation
- Valeur ajoutée pour l'utilisateur

Focus aujourd'hui : Entités Nommées de lieu en Allemand (CLEF-HIPE 2020 P.Ortiz, Y.Dupont, G.Lejeune, T.Tian [2])

2. https://impresso.github.io/CLEF-HIPE-2020/

Les entités nommées : définitions

Enrichir les données textuelles avec une couche "sémantique" pour permettre de "raisonner" automatiquement.

- Noms d'organisation, noms de personne
- Noms de fêtes, noms de films/chansons,
- Noms de protéines ...

Les entités nommées : définitions

Enrichir les données textuelles avec une couche "sémantique" pour permettre de "raisonner" automatiquement.

- Noms d'organisation, noms de personne
- Noms de fêtes, noms de films/chansons,
- Noms de protéines ...
- Noms de lieux

Les entités nommées : définitions

À (Boncourt)$_{lieu}$ (en 1886)$_{date}$:
La boulangerie de (M. Staempfli)$_{personne}$, (rue du parc, 68)$_{lieu}$.
la (confédération du (Rhin)$_{lieu}$)$_{lieu??}$

Adaptation au domaine

Adapter les systèmes d'extraction d'entités nommées standard sur :

- des corpus anciens
- non nativement numériques
- multilingues

EN en humanité numérique : données

Textes anciens : problèmes

transcription à la main -> données bruités

segmentation des textes initiaux -> tiret séparant les mots tâches,

espace en trop -> caractères en trop, OOV (out-of-vocabulary)

Corpus

Corpus d'apprentissage et d'évaluation de la tâche CLEF-HIPE
En français et allemand, annotés en entités de type *Personne, Organisation, Location, Product et Time*

	tokens	*nb doc*	*nb seg*	*nb d'entités nommées annotées*				
				Pers	*Loc*	*Org*	*Time*	*Prod*
train Fr	*166217*	*158*	*19183*	*3067*	*2513*	*833*	*273*	*198*
dev Fr	*37592*	*43*	*4423*	*771*	*677*	*158*	*69*	*48*
train De	*86960*	*104*	*10353*	*1747*	*1170*	*358*	*118*	*112*
dev De	*36175*	*40*	*4186*	*664*	*428*	*172*	*73*	*53*

Extrait du corpus français

```
# language = fr
# newspaper = EXP
# date = 1918-04-22
# document_id = EXP-1918-04-22-a-i0077
# segment_iiif_link = https://iiif.dhlab.epfl.ch/.../1186,1881,474,79/full/0/default.jpg
Lettre          O       _
de              O       _
la              O       _
Su              B-loc   NoSpaceAfter
.               I-loc   _
_               I-loc   NoSpaceAfter
sss             I-loc   _
allemands       I-loc   EndOfLine
# segment_iiif_link = https://iiif.dhlab.epfl.ch/.../1190,1967,493,52/full/0/default.jpg
(               O       NoSpaceAfter
Nous            O       _
serons          O       _
heureux         O       _
de              O       _
publier         O       _
de              O       _
temps           O       _
à               O       EndOfLine
# segment_iiif_link = https://iiif.dhlab.epfl.ch/.../1165,1995,517,53/full/0/default.jpg
autre           O       NoSpaceAfter
,               O       _
sous            O       _
cette           O       _
rubrique        O       NoSpaceAfter
,               O       _
des             O       _
```

Extrait du corpus français - Exemple de bruit

```
Lettre          O          _
de              O          _
la              O          _
Su              B-loc      NoSpaceAfter
.               I-loc      _
_               I-loc      NoSpaceAfter
sss             I-loc      _
allemands       I-loc      EndOfLine
```

Forme non-standard :

"Su. _sss allemands" -> "Suisse allemande"

Extracteurs d'entités nommées existants

Extracteurs multilingues d'entités nommées

- Modèles appris et configurés sur une langue "standard" (la langue journalistique).
- Adaptation avec apprentissage automatique à partir de données annotées dans la langue/le domaine cible

nom	langue initiale	algorithme	corpus d'apprentissage
SpaCy	anglais	CNN	Conll2003 (journal)
Standford NLP	anglais	CRFs	Conll2003 (journal)
SEM	français	CRFs	French Treebank

SpaCy

Présentation :

- réseau de neurones avec CNN.
- multilingue, mais qualité variable.

SpaCy

Présentation :

- réseau de neurones avec CNN.
- multilingue, mais qualité variable.

Adaptation :

- entraînement depuis zéro (blank).
- adaptation depuis modèle existant (de_core_web_sm).

SEM

Présentation :

- SEM : outil utilisant CRF, Wapiti.
- Enrichissements textuels, gestion de format.
- Uniquement pour le français.

SEM

Présentation :

- SEM : outil utilisant CRF, Wapiti.
- Enrichissements textuels, gestion de format.
- Uniquement pour le français.

Adaptation :

- Enrichissement des lexiques existants.
- Entraînement sur le corpus.
- Beaucoup de travail manuel.

Bi-LSTM-CRF

Présentation :

- architecture état-de-l'art pour l'étiquetage de séquences.
- LSTM bi-directionnel (contextes gauche et droit) + CRF (cohérence des étiquettes).
- "plug and play" : utiliser des embeddings pré-entraînés et adapter reste du réseau sur données.

Bi-LSTM-CRF

Présentation :

- architecture état-de-l'art pour l'étiquetage de séquences.
- LSTM bi-directionnel (contextes gauche et droit) + CRF (cohérence des étiquettes).
- "plug and play" : utiliser des embeddings pré-entraînés et adapter reste du réseau sur données.

Adaptation :

- Concaténation de 3 représentations : FastText original, Flair + FrELMo (ELMo entraîné sur OSCAR-fr).
- Entraînement sur le corpus (représentations figées).

Résultats et discussion

Résultats

système	P	R	F
Spacy (blank)	64,8%	53%	58,3
Spacy (pré-entraîné)	70,9%	59,1%	64,42
SEM	64,4%	43,8%	52,1
Bi-LSTM-CRF	**63.1%**	**66,6%**	**64,8**

Quelques observations

- Réseaux de neurones meilleurs que CRF (métriques)
- RN plus facilement adaptables que CRF (pré-entrainement vs ré-entraînement)
- adapter un modèle pré-entrainé améliore les résultats (divers papiers)

Plafond de verre ?