A Benchmark of Nested Named Entity Recognition Approaches on Historical Documents

Solenn Tual

Reading Group - 27/01/2023

SODUCO ANR-18-CE38-0013

ANR SoDUCo

Social Dynamics in Urban Context

Historical sources :

- Maps
- Trade directories

Non-Commerçans. (Paris): 269

Chardin, R. Pavée, 16. — B. C.
Chardin, R. Michel Lepelletier, 21.
Chardon, (Ve.) R. S. Marc, 15.
Chevron, R. Chapon, 13.
Chimay, (Mme.) R. de Varennes, 31.
Choart-Duplessis, R. de Turenne, 31.

AMADOU ET ALLUMETTES. — Pour les ALLUMETTES OXIGÉNÉES. Voyez BRIQUETS PHYSIQUES.

DARRAS (Thomas), r. de la Vieille-Monnaie, 10. Briquets et veilleuses, mèches à quinquets, mèches soufrées, pierres, agaric de chêne, liége en planches, bouchons.

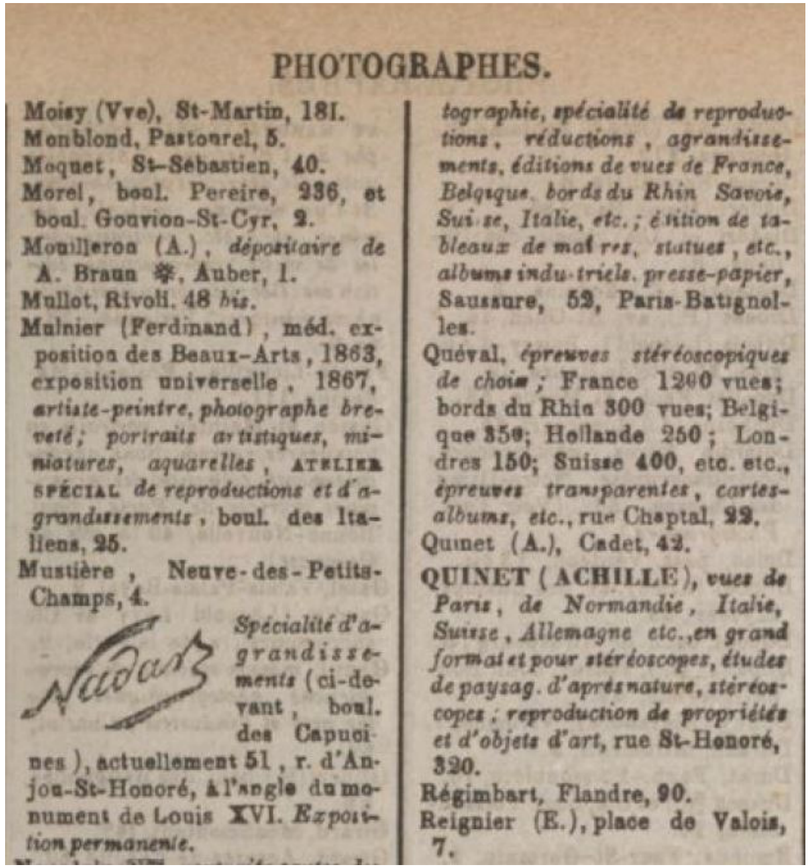
GALLIENNE jr., r. de la Beaumerie, 3. Brûle-tout, boîtes à briquet, mèches à vin et à quinquet, veilleuses mèches, soufrées; pierres, agaric, bouchons, liége.

LESOY, r. Aubry-le-Boucher, 43.

BAUDOYER (place).

IX Arr. *Hôtel-de-Ville.*) ← Rue Tixeranderie, pourtour St-Gervais, Saint-Antoine et Renaud-Lefèvre.

1 Lhuoty (Vve), *vins.*
2* Privé, *distillateur.*
Lemoine-Chanel et Leroy, *nouveautés.*
Chantrier, *court.-gourm.*
24* Longpré aîné, *bijoutier en or et argent.*
Saint-Ouser, *émailleur.*
Cellier (A.), *graveur-ciseleur.**
Rousseau (J.), *bijoutier en or.**
Benoît, *orfèvre-fabr.*
Lérésy, *doreur.*
30 Bouton, *fab. de cuir verni.**
31 Pardon, *vins.*
Bourguille, *fabr. de presses.*
Vaudain, *passementier.*
Fininojne, *bronze doré.*
Babé aîné, *fabr. de boutons.**
Gaulin, *chapelier.*
Moisy, *tabletier.*
49* Cendrier aîné, *prop.*
Desmarests, *fab. boîtes d'emballage.*
Ferrand, *lapidaire.*
7 École communale de jeunes filles.
Berthelot, *vins.*
6 Versiaen, *serrurier-mécanicien.*
8 Michel, *brossier.*
9 Labottiere, *serrurier.*
10 Sacrez, *vins.*
12 Baudoin, *épic.*
13 Lejard, *clouteries et crépins.*
14 Isaduel (Vve), *fab. de et tapisseries.*
10 Lainé jeune, *vins.*
Jumelles omnibus et entreprise générale des Omnibus.
11 Melouzay, *vins en gros, et à Bercy, Port, 31.*
12 Combaud, *coiffeur.*
Mommain (F.), *vins en gros.*
13 Dufailly, *sculpt. fabr. de carton-pierre.*



PHOTOGRAPHES.

Moisy (Vve), St-Martin, 181.
Monblond, Pastourel, 5.
Moquet, St-Sébastien, 40.
Morel, boul. Pereire, 236, et boul. Gouvion-St-Cyr, 2.
Mouilleron (A.), *dépositaire de* A. Braun ✻, Auber, 1.
Mullot, Rivoli, 48 *bis.*
Mulnier (Ferdinand), méd. exposition des Beaux-Arts, 1863, exposition universelle, 1867, *artiste-peintre, photographe breveté; portraits artistiques, miniatures, aquarelles,* ATELIER SPÉCIAL *de reproductions et d'agrandissements*, boul. des Italiens, 25.
Mustière, Neuve-des-Petits-Champs, 4.

Nadar

Spécialité d'agrandissements (ci-devant, boul. des Capucines), actuellement 51, r. d'Anjou-St-Honoré, à l'angle du monument de Louis XVI. *Exposition permanente.*

tographie, spécialité de reproductions, réductions, agrandissements, éditions de vues de France, Belgique, bords du Rhin Savoie, Suisse, Italie, etc.; édition de tableaux de maîtres, statues, etc., albums industriels, presse-papier, Saussure, 52, Paris-Batignolles.
Quéval, *épreuves stéréoscopiques de choix;* France 1200 vues; bords du Rhin 300 vues; Belgique 350; Hollande 250; Londres 150; Suisse 400, etc. etc., *épreuves transparentes, cartes-albums, etc.*, rue Chaptal, 22.
Quinet (A.), Cadet, 42.
QUINET (ACHILLE), *vues de Paris, de Normandie, Italie, Suisse, Allemagne etc., en grand format et pour stéréoscopes, études de paysag. d'après nature, stéréoscopes; reproduction de propriétés et d'objets d'art*, rue St-Honoré, 320.
Régimbart, Flandre, 90.
Reignier (E.), place de Valois, 7.

Paris trade directories

CORPUS

213 directories printed between 1789 and 1950

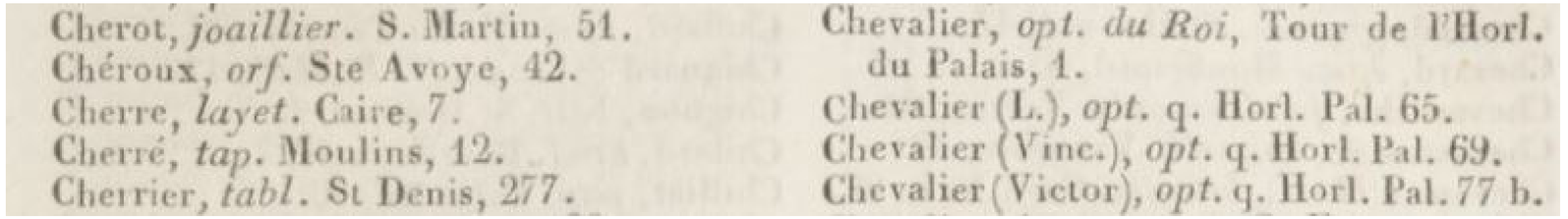
STRUCTURE

1 directory = thousands of pages
1 page = lists of hundred of entries with several entities enumeration patterns

CONTENT

1 entry =

- Name
- Description including activity
- Address(es) : street name, street number, kind of geographical place
- Military or professional title



Cherot, *joaillier*. S. Martin, 51.
Chéroux, *orf*. Ste Avoye, 42.
Cherre, *layet*. Caire, 7.
Cherré, *tap*. Moulins, 12.
Cherrier, *tabl*. St Denis, 277.

Chevalier, *opt. du Roi*, Tour de l'Horl. du Palais, 1.
Chevalier (L.), *opt*. q. Horl. Pal. 65.
Chevalier (Vinc.), *opt*. q. Horl. Pal. 69.
Chevalier (Victor), *opt*. q. Horl. Pal. 77 b.

Pipeline

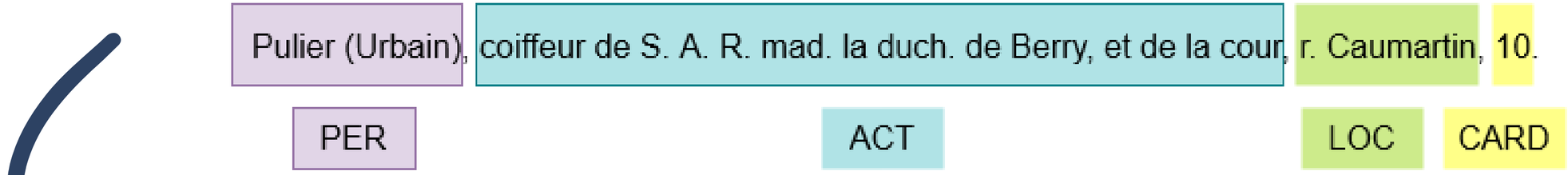
Already existing pipeline

- Entry detection and segmentation
- Text transcription using Optical Character Recognition (OCR) → **Noisy text**
- Named Entity Recognition → **Only one entity type by word**

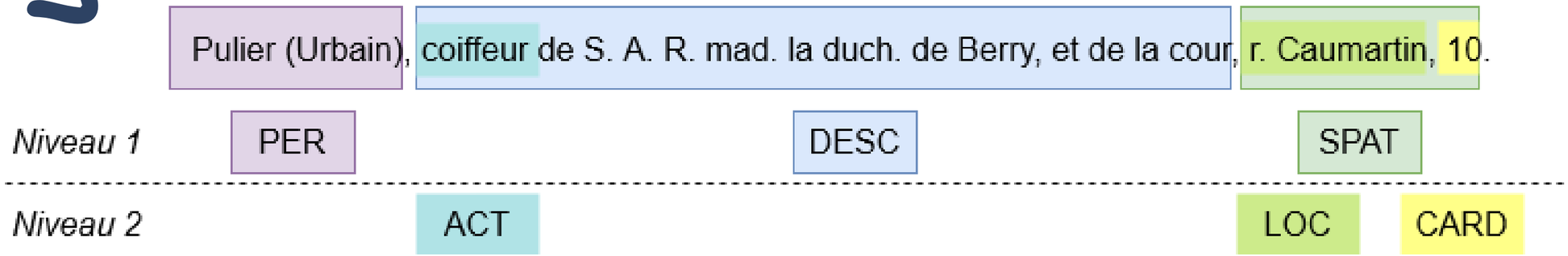
Abadie & al. (2022)

Goal

Flat Named Entity Recognition



Nested Named Entity Recognition



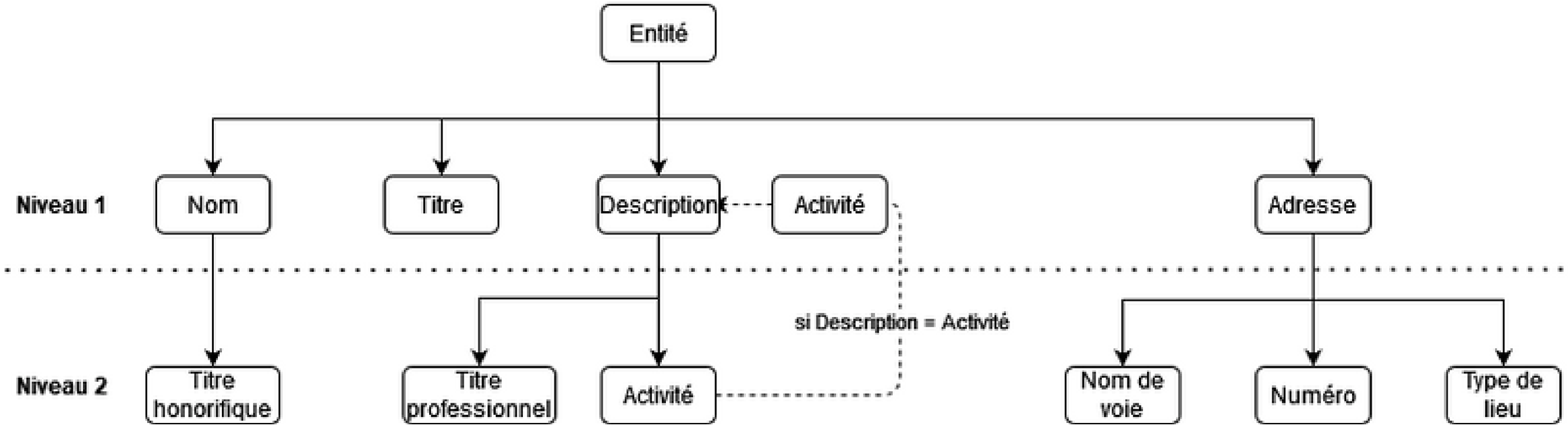
Create a high level semantic view of directories & searchable data

8572 annotated entries :

- reference = post-processed after OCR stage
- noisy

Noisy entry : Dufour, architecte, r. de Chartres-S.-Ho-2781 40anore, 12. (Elig.)

Reference entry : Dufour, architecte, r. de Chartres-St-Honoré, 12. (Elig.)



Entity type hierarchy

MODEL

CamemBERT → French BERT-based model

Martin et al. 2020

BERT : Bidirectional Encoder Representations from Transformers

- Transformer model for natural language processing
- Pre-trained on a vast amount of data

→ Can be fine-tuned for specific NLP tasks as NER

TOOL

HuggingFace API

A user-friendly Python library for NLP using Deep Learning pre-trained models

A1 : Independant flat NER layers

Jia et al. (2021)

Tags

Token	Niveau 1	Niveau 2
Dufour	I-PER	O
(	I-PER	O
Gabriel	I-PER	O
)	I-PER	O
,	O	O
libraire	I-ACT	O
,	O	O
r	I-SPAT	I-LOC
.	I-SPAT	I-LOC
de	I-SPAT	I-LOC
Vaugirard	I-SPAT	I-LOC

MODEL 1 MODEL 2

- 1 layer (= fine-tuned model) for each entity level
- Nested entities created in post-processing stacking predictions of each layer

A2 : NNER using joint-labelling

Agrawal et al. (2022)

Token	Joint-label
Dufour	I-PER+O
(	I-PER+O
Gabriel	I-PER+O
)	I-PER+O
,	O+O
libraire	I-ACT
,	O+O
r	I-SPAT+I-LOC
.	I-SPAT+I-LOC
de	I-SPAT+I-LOC
Vaugirard	I-SPAT+I-LOC

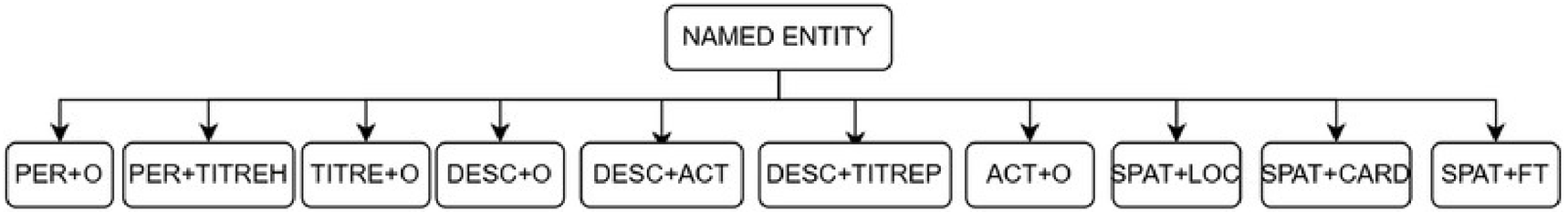
- One joint-label (= multiclasses) tag for each token
- One fine-tuned model to recognise all entities

A3 : NNER using joint-labelling and hierarchical loss

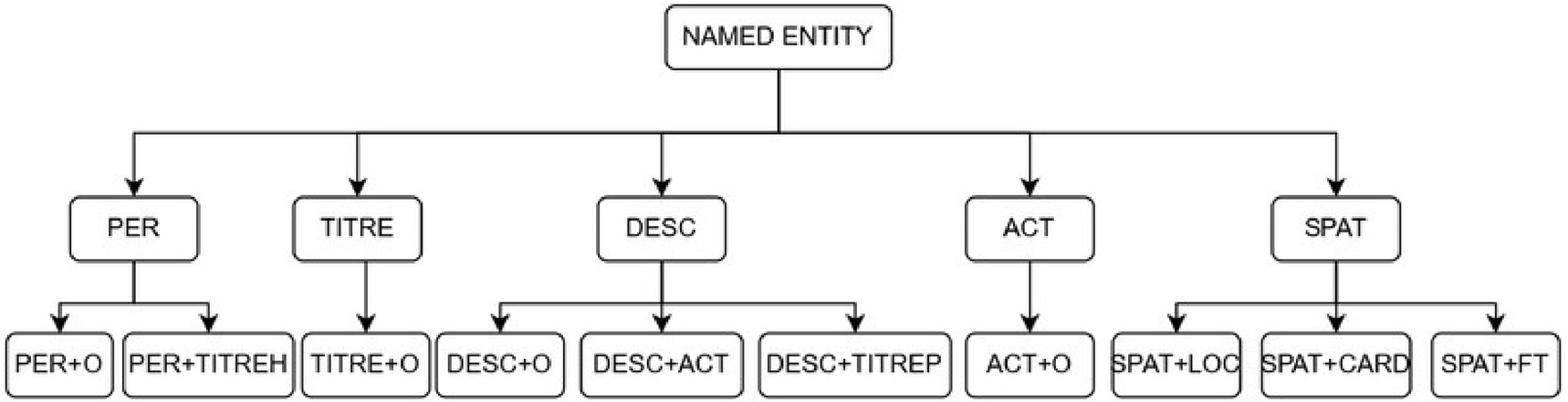
Agrawal et al. (2022) /// Bertinetto et al. (2020)

A3 considers semantic distance between reference and prediction (semantic distance is calculated in the hierarchical loss)

A2



A3



F1-Score (mean of 5-runs with fixed seeds)

		CamemBERT	*Pretrained CamemBERT*
Reference	A1	95.5	**95.6**
	A2	96.2	**96.3** ✓
	A3	**96.3** ✓	95.6
Noisy dataset	A1	93.8	**94.3** ✓
	A2	93.8	**93.9**
	A3	**94.1**	**94.1**

Results

A1

```
<PER>Dullaut</PER>,
<ACT>chandronnier</ACT>,
<SPAT>
    <LOC>r. de la Sourdière</LOC> I-
    <LOC>(</LOC>
    <TITREH>☞</TITREH>T
    <CARDINAL>T4 </CARDINAL>
</SPAT>
```

- Classe valide
- Classe de niveau 2 incompatible avec le niveau 1
- Ambiguité sur la nature de l'entité (erreur d'OCR ?)
- Entité non existante (O attendu)

A2 / A3

```
<PER>Dullaut</PER>,
<ACT> chandronnier</ACT>,
<SPAT>
    <LOC> r. de la Sourdière</LOC>
</SPAT>
<TITRE>·I</TITRE>
<SPAT>-</SPAT>
<TITRE> (·☞T4</TITRE>
```

- Classe valide
- Entité non existante (O attendu)
- Entité bruitée (limites incorrectes)

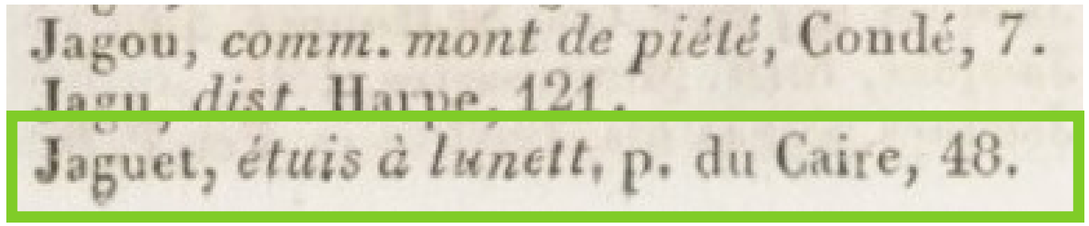
✗ **Entity types hierarchy** ✓

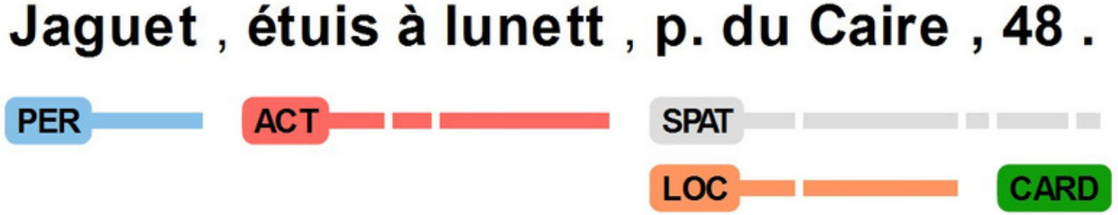
✗ **Fine-tuning time** ✓

Thank you for you attention

Nested named entities

High-level semantic view of directories


Jagou, *comm. mont de piété*, Condé, 7.
Jagu, *dist.* Harpe, 121.
Jaguet, *étuis à lunett*, p. du Caire, 48.



Jaguet, étuis à lunett, p. du Caire, 48.

OCR

NER

References

Nathalie Abadie, Edwin Carlinet, Joseph Chazalon et Bertrand Duménieu (mai 2022). « **A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories** ». In : t. Document Analysis Systems : 15th IAPR International Workshop, DAS 2022, La Rochelle, France.

Nested NER

Liruizhi Jia, Shengquan Liu, Fuyuan Wei, Bo Kong et Guangyao Wang (août 2021). « **Nested Named Entity Recognition via an Independent-Layered Pretrained Model** ». In : IEEE Access 9. Conference Name : IEEE Access

Ankit Agrawal, Sarsij Tripathi, Manu Vardhan, Vikas Sihag, Gaurav Choudhary et Nicola Dragoni (jan. 2022). « **BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling** ». In : Applied Sciences 12.3, p. 976

Luca Bertinetto, Romain Mueller, Sina Samangooei et Nicholas A. Lord (juin 2020). « **Making Better Mistakes : Leveraging Class Hierarchies with Deep Networks** ». In : The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Models

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah et Benoît Sagot (2020). « **CamemBERT : a Tasty French Language Model** ». In : Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. arXiv : 1911.03894, p. 7203-7219. doi : 10.18653/v1/2020.acl-main.645. url : http://arxiv.org/abs/1911.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser et Illia Polosukhin (déc. 2017). « **Attention Is All You Need** ». In : arXiv :1706.03762 [cs]. arXiv : 1706.03762. url : http://arxiv.org/abs/1706.03762.