[Poster] A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents

Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Dumenieu, Edwin Carlinet

A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents

S. Tual [1,2], N. Abadie [1], J. Chazalon [2], B. Duménieu [3], E. Carlinet [2]

(1) LASTIG, IGN-ENSG, Univ. Gustave Eiffel (France) (2) LRE, EPITA (France) (3) CRH, EHESS (France)

SAN JOSE, CALIFORNIA, USA 2023

ANR SODUCO: ANR-18-CE38-0013

CODE, DATA, PAPER
soduco/paper-nestedner-icdar23-code

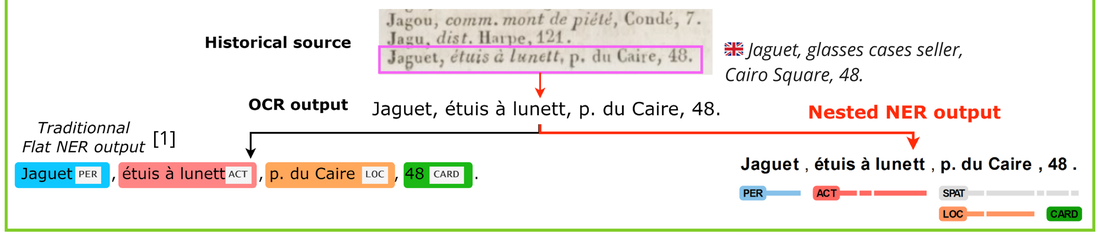
CONTEXT

The ANR SODUCO project: **a study of the social dynamics of Paris from 1789 to 1950** on the basis of historical documents (maps and trade directories).

MOTIVATIONS & RESEARCH QUESTIONS

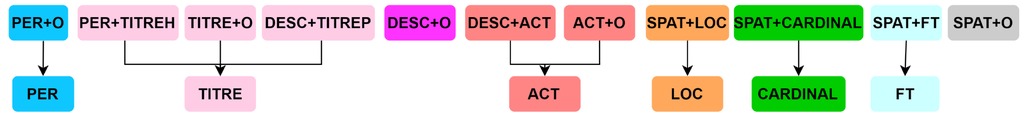
- Can we extract tree-structured entities from directory entries with sufficient quality?
- Which nested NER approaches are the most powerful and suitable for this task?
- What is the impact of the additional knowledge provided by nested NER methods on traditional flat NER methods?

PIPELINE



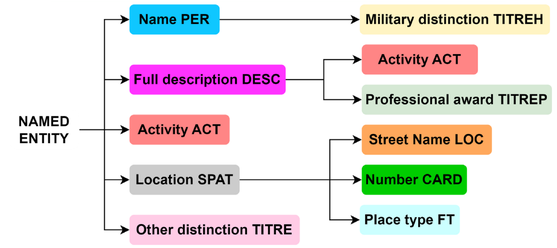
BASELINE

Nested entities are mapped with corresponding flat NER entities proposed by *Abadie et al.* [1]



[1] Abadie, N., Carlinet, E., Chazalon, J., Duménieu, B. *A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories*. DAS 2022.

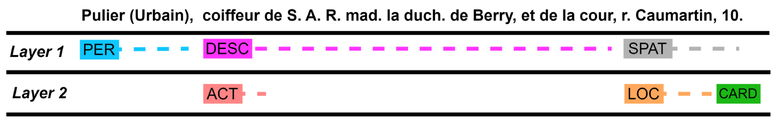
LABELS & THEIR HIERARCHY



STATE-OF-THE-ART APPROACHES

Approach 1 (M1) : Independant NER layers [2]

- One flat NER model (layer) recognises one entity level.
- No constraint on an entity tag or borders regarding other levels of entities. ✗

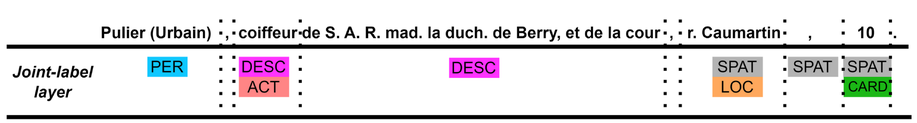


Pulier (Urbain), hairdresser of H.R.H Duchess of Berry and to the court, Caumartin Street, 10

[2] Jia, L., Liu, S., Wei, F., Kong, B., Wang, G. *Nested Named Entity Recognition via an Independent-Layered Pretrained Model*. IEEE Access 9, 109693–109703 (2021).

Approach 2 (M2) : Joint-labelling approach [3]

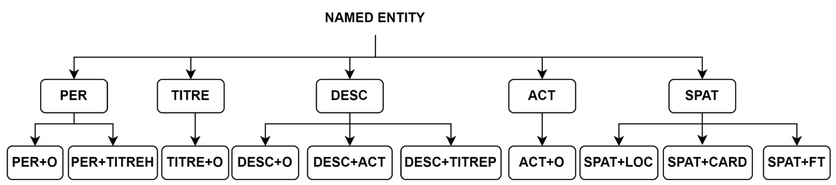
- One model recognises all entity levels using *joint-labels*.
- **Respect the hierarchy of labels**. ✓
- All classification errors have the same cost (*Categorical Cross Entropy Loss*).



[3] Agrawal, A., Tripathi, S., Vardhan, M., Sihag, V., Choudhary, G., Dragoni, N. *BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling*. Applied Sciences 12(3), 976 (2022).

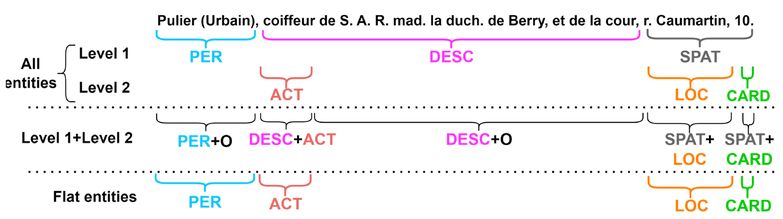
A NEW HIERARCHICAL APPROACH USING JOINT-LABELLING (M3)

We implement the *Hierarchical Cross Entropy Loss* [4] in the BERT-based model to **consider the semantic distance between *joint-labels*,** computed on the tree shown bellow.



[4] Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A. *Making better mistakes: Leveraging class hierarchies with deep networks*. CVPR (2020)

EVALUATION



Metrics

- Precision
- Recall
- F1-Score

DATASET

For each entry, two transcriptions variants:

- REF - OCR noise corrected manually
- NOISY - with OCR noise

Train: 6004 entries
Dev: 668 entries
Test: 1669 entries

RESULTS

F1-SCORE = 94,1 % **on noisy entries with M3**

- **M2 / M3**: ensures that **only valid nested entities are produced** regarding our class hierarchy.
- **M3 : increases performance on the less represented classes.**
- Additional knowledge doesn't improve performance on flat NER.

EXTRA MATERIAL

Comparison of IO and IOB2 labels to perform same class entities segmentation: results are not conclusive on this dataset.
Data, code, paper and materials are available on Git-Hub.