



HAL
open science

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

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

► To cite this version:

Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Dumenieu, Edwin Carlinet. [Poster] A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents. International Conference on Document Analysis and Recognition - ICDAR 2023, Aug 2023, San Jose, California, United States. 2023. hal-04191900

HAL Id: hal-04191900

<https://hal.science/hal-04191900>

Submitted on 30 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

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

S. Tual^{1,2}, N. Abadie¹, J. Chazalon², B. Duménieu³, E. Carlinet²
 (1) LASTIG, IGN-ENSG, Univ. Gustave Eiffel (France) (2) LRE, EPITA (France) (3) CRH, EHESS (France)



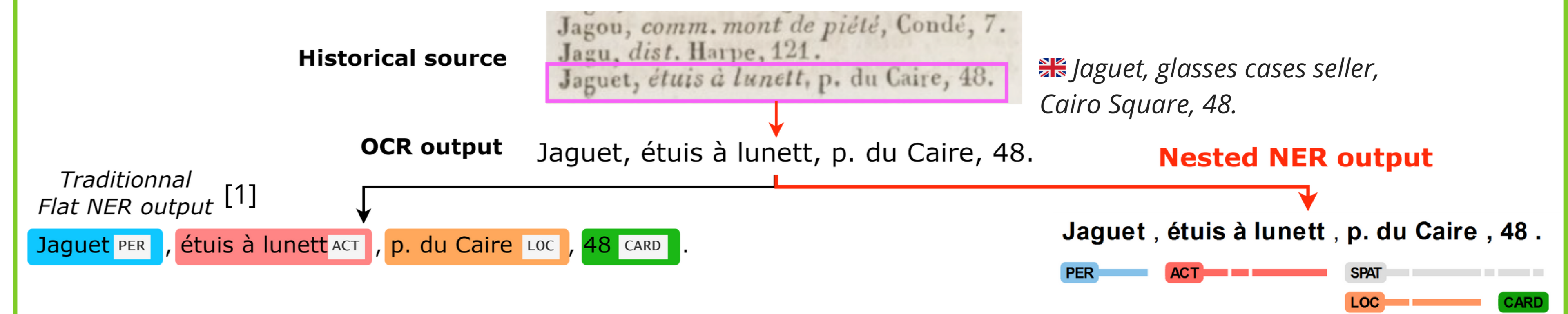
CONTEXT

The ANR SODUCO project: a study of the social dynamics of Paris from 1789 to 1950 on the basis of historical documents (maps and trade directories).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

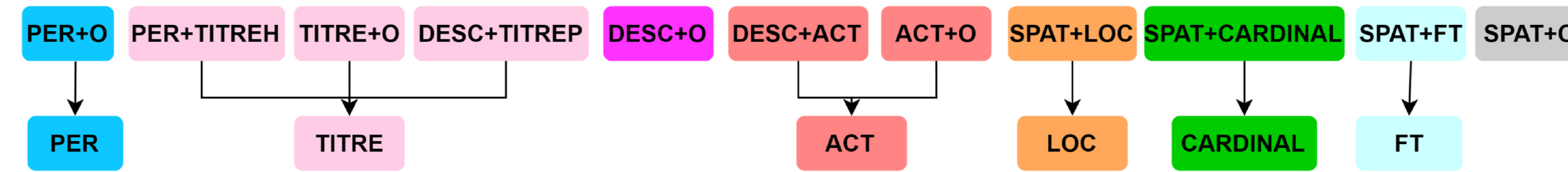
PIPELINE



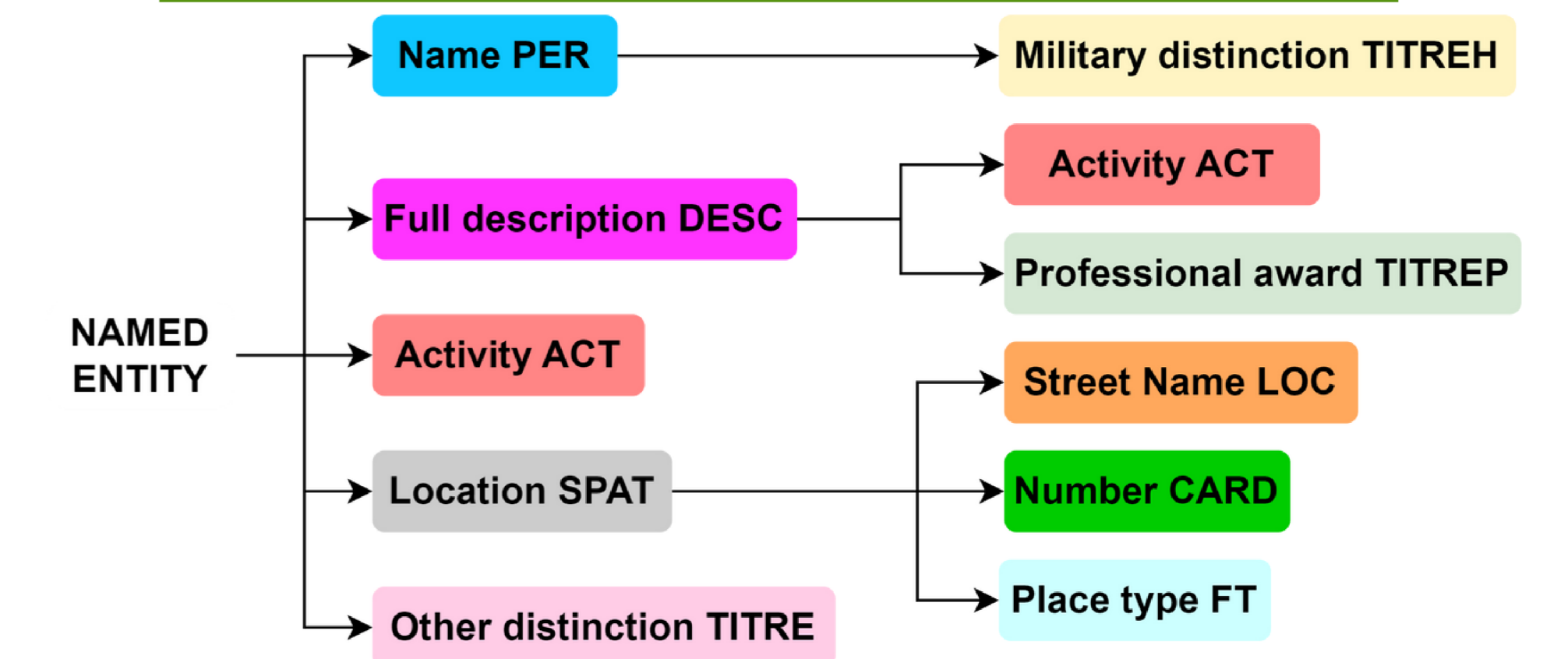
Nested entities are mapped with corresponding flat NER entities proposed by Abadie et al. [1]

[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.

BASELINE



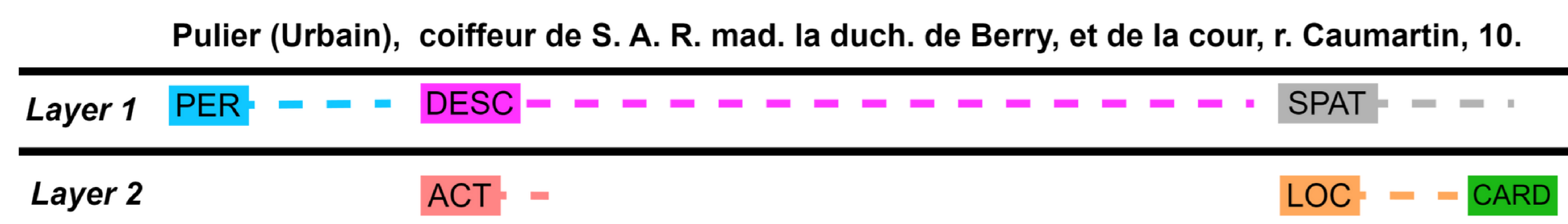
LABELS & THEIR HIERARCHY



STATE-OF-THE-ART APPROACHES

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

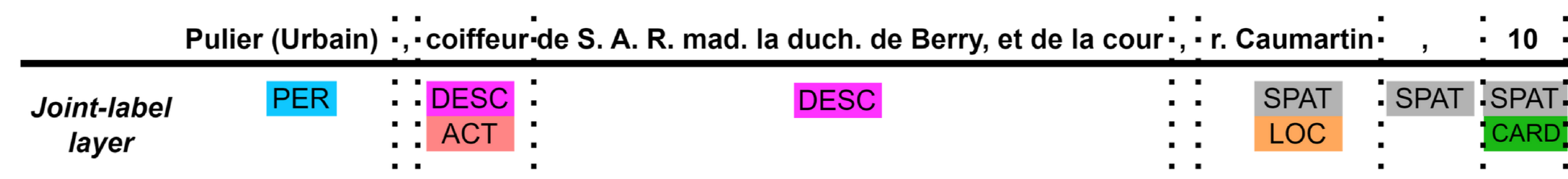
- One flat NER model (layer) recognises one entity level.
- No constraint on an entity tag or borders regarding other levels of entities. ✗



[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).

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

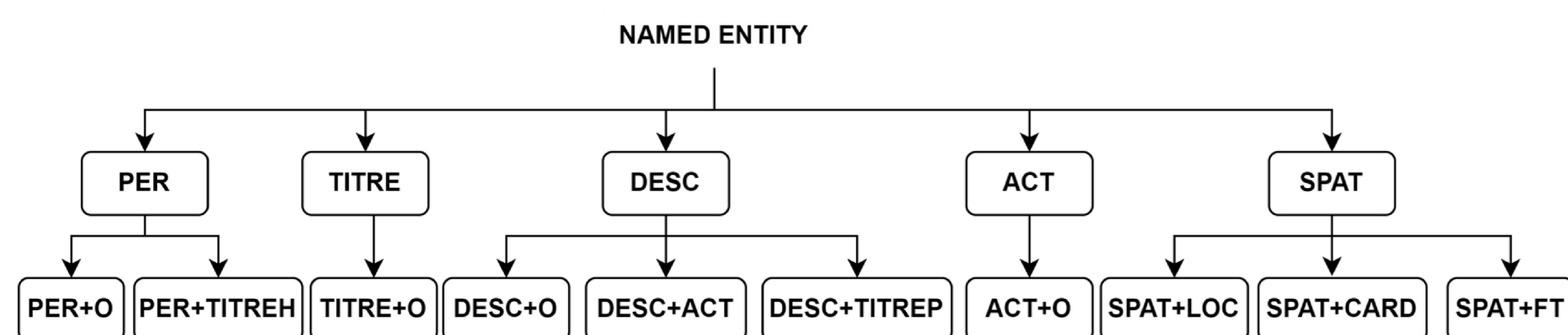
RESULTS

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.

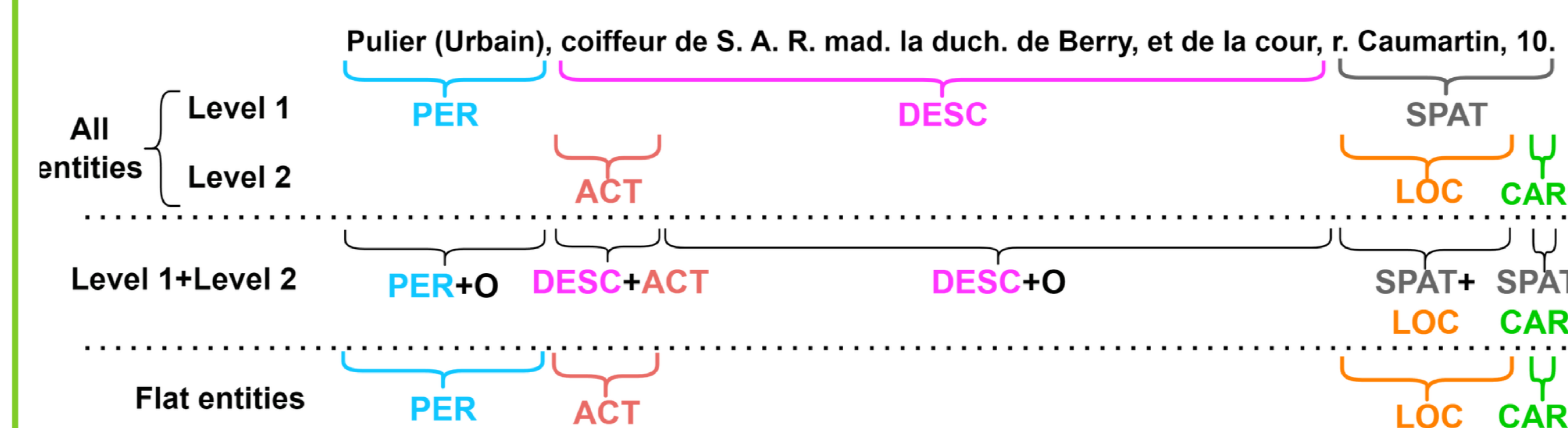
A NEW HIERARCHICAL APPROACH USING JOINT-LABELLING (M3)

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

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.