Adapting a Pre-Neural Named Entity Recognizer and Linker to Historical Data

Damien Nouvel, Jean-Claude Zagabe Seruti

Adapting a Pre-Neural Named Entity Recognizer and Linker to Historical Data

Nouvel, Damien[1][0000−0001−8866−4028] and Zagabe Seruti, Jean-Claude[2]

[1] INALCO ERTIM damien.nouvel@inalco.fr http://damien.nouvels.net
[2] INALCO ERTIM claudezagabe@gmail.com

Abstract. This article describes the participation of ERTIM (INALCO) to the French NER task of CLEF HIPE 2020 lab with the mXS system, a combination of pattern mining and machine learning, implemented in 2010-2013. Due to multiple reasons, almost no upgrades or improvements were achieved since then, only a minimal linking module and some lexical entries were added. No training and almost no adaptation were implemented for this lab. Results on historical data show severe degradations, in particular concerning the recognition of organisations.

Keywords: Named Entity Recognition · French · Pattern Mining · Historical Texts

1 Introduction

Named Entity Recognition for French language was implemented by several resource building, evaluation campaigns or shared tasks. Many resources have already been described[2] and it undoubtly remains an important task in the field of NLP.

Our system, mXS[3], was developed during the QUAERO[3,5,8] evaluation campaign. This shared task, held in 2011, aimed at transcription and named entity recognition (only classification, no linking) of radio broadcast news in French. At this occasion, our system performed 3rd with competitive results.

Since its implementation, mXS was minimally updated. Recently, the TALAD[4] research project (NLP and Discourse Analysis) provided the opportunity to use mXS on political data (radio interviews). It is in the context of this project that we decided to participate in the CLEF HIPE 2020 lab [4].

[3] https://github.com/eldams/mXS
[4] https://web.u-cergy.fr/anr-talad/

2 The mXS System

Our system has already been exhaustively described [6], it combines data preprocessing, pattern mining and pre-neural machine learning. We only report the main characteristics of the system here.

2.1 Data preprocessing

The same preprocessing are used during training and inference (prediction) steps. For French, the tokenisation and POS tagging are done using TreeTagger [9], some postprocessing are implemented (e.g. proper nouns, numbers, verbs) to provide relevant features for named entity classification. The second step consists in looking up each token in a gazetteer which uses several large lexicons (in total more than 1M entries) containing names for French (mainly common and proper nouns, including multiword expressions).

Both preprocessing are considered as an enrichment process: raw textual data is transformed into sequences of disjunctive itemsets.

2.2 Pattern Mining

One of the main differences of mXS compared to other existing software is that named entity recognition is considered as a segmentation task rather than a classification task. In other words, the system tries to find positions where entities begin and end, while most systems consider this task as assigning a class (named entity category) to each element in the sequence.

Based on this assumption, during the training step the software runs a pattern mining algorithm [1] to discover rules that segment sequences and reach a minimum level of frequency and confidence in this regard. Those rules are stored in a separate file, and a single rule may introduce multiple segmentations.

2.3 Machine Learning

Once patterns are extracted, they are considered as features for a machine learning model implemented using the SciKit [7] software. The goal is to predict what is the probablity that a given named entity class would begin or end solely based on extracted patterns. Once this model is learned, it is therefore able to provide probabilities of named entity segmentation.

2.4 Named Entity Recognition

The recognition is a decoding step where individual segmentation probabilities are constrained (started entities must end within the considered sentence, nesting can be allowed and follow guidelines, etc.). The resulting annotation corresponds to the segmentation which respects the constraints and maximises the joint probabilities over the sequence.

2.5 Named Entity Linking

We did implement a simple entity linker, executed after recognition, and based on Wikipedia. As a first version, the system only links PERS entities. The main steps are:

1. The name is checked against a local database (very limited for now, it does only contain names and Wikipedia links of a small set of entities),
2. If no match, the wikiapi[5] python module, connected to the French Wikipedia, is searched for entries based on the entity label (as far as we understood, this API executes the search both in Wikipedia titles and pages),
3. Retrieved results are filtered with an approximate string matching based on the title of the Wikipedia page,
4. A similarity (cosine) between the BoW of processed document and the Wikipedia page content is computed,
5. The page which has highest similarity is selected.

3 Adaptation to the CLEF HIPE 2020 Lab

As previously mentioned, due to multiple reasons including the COVID19, we did not have much time to prepare the system for the CLEF HIPE 2020 lab. Actually we did not use the development and training data at all. Fortunately, this shared task was conducted using annotation guidelines mainly inspired by the QUAERO campaign, which facilitated our work.

Input data is simply converted into raw text (with some minor French tokenisation adaptations) and passed to mXS. Output of our system as tagged texts are converted to the expected column format. A simple script checks and filters out unwanted named entities. For linking, another process runs our linker as previously described, and we used the wikimapper[6] python module to obtain the corresponding WikiData key.

4 Classification Results and Discussion

We report our system official results [4] (off-the-shelf without any upgrade) for NER in Table 1. In this paper, we do not report the linker results, since they are very low because this part is very simple and only links PERS entities. Performance is clearly very distant from other systems (best system F1 was 0.84 on COARSE-LIT-micro-strict). We still think it is interesting to have those results as an evaluation of how much an pre-neural system is degraded on unseen historical data when it has been minimally maintained and was neither updated nor trained on the domain and available data. As a side-product of our participation, we provide a valuable estimate of the expected performance before adaptation.

[5] https://pypi.org/project/wikiapi/
[6] https://pypi.org/project/wikimapper/

Table 1. NER tasks results.

Task	Rank	F1	Precision	Recall
COARSE-LIT-micro-strict	25th	0.316	0.435	0.248
COARSE-LIT-micro-fuzzy	23rd	0.439	0.604	0.344
FINE-LIT-micro-strict	7th	0.303	0.418	0.238
FINE-LIT-micro-fuzzy	7th	0.412	0.568	0.324

Clearly, our main problem is recall, as expected. Table 2 reports detailed results per type. We note that PERS, TIME and LOC types have a limited degradation, while ORG suffer from a very severe loss. Given the fact that the lab focused on historical data, this reveals that locations and persons are quite robust over time periods, as depicted below in Figure 1, what is not the case for organisations.

Table 2. Litteral micro precision, recall and F1 scores for French coarse NER per type.

Type	F1	Precision	Recall
ALL	0.435	0.583	0.346
PERS	**0.55**	0.628	0.49
TIME	0.462	0.336	**0.736**
LOC	0.421	**0.718**	0.297
ORG	0.117	0.158	0.092
PROD	0.082	0.25	0.049

We conducted some additional experiments reported in Table 3, with the coarse strict metric (in Table 1, COARSE-LIT-micro-strict) computed using the provided scorer[7]. First, we corrected some scripts which led to minimal improvements (`up`). In a second step, we added to mXS's lexicon dev and train data entries with their types (`up,lex`), but this did almost not change scores. Finally, we re-trained the machine learning component of the system (`up,lex,train`) with only CLEF HIPE 2020 lab data or merged with the data the system was initially trained on (QUAERO). Those last experiments more significantly improved the system performance, especially for precision, but recall still remains an issue.

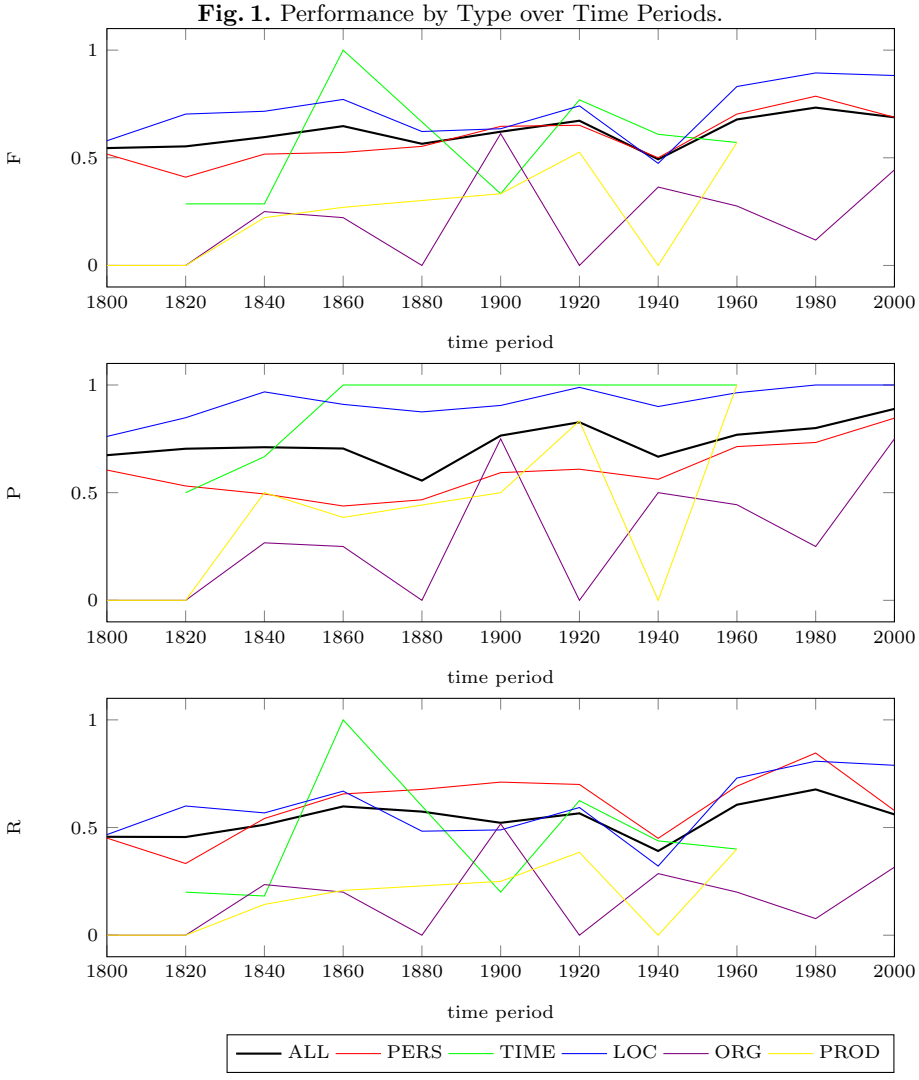
From those results, we first notice to what extent a system's performance is degraded when changing the dataset, both for the domain (historical) and annotation scheme (there are notable differences between QUAERO and HIPE). Despite our efforts, the system does not reach the organizers baseline (CRF). Secondly, even with a modular system where lexicon is an external resource, it is not sufficient to update only this component, the machine learning model parameters have to be recomputed. Finally, from the fact that scores are better

[7] https://github.com/impresso/CLEF-HIPE-2020-scorer

Table 3. Litteral coarse strict results after system upgrades.

Type	F1	Precision	Recall
up	0.288	0.36	0.24
up,lex	0.293	0.367	0.244
up,lex,train(HIPE)	**0.548**	0.733	0.438
up,lex,train(QUAERO,HIPE)	0.449	0.533	0.388

Fig. 1. Performance by Type over Time Periods.



when training with CLEF HIPE 2020 lab data only we make the assumption that QUAERO and HIPE datasets are somehow heterogeneous.

Figure 1 depicts F1, precision and recall of entity types over time periods (20 years, each point is plotted at the beginning of the period). For person and location types, results are quite stable. This is not the case for other types, especially for the organisation type which fluctuates a lot. The global score (ALL) is quite stable and increases smoothly, at the exception of the 1940-1960 period where there is an important loss.

5 Conclusion

This paper describes the results obtained by our old-fashioned mXS system for the CLEF HIPE 2020 lab (named entity classification and linking). Diverse reasons prevented us from updating and training the system for this shared task. Performance is severely degraded compared to what was obtained during the QUAERO evaluation campaign (2012). Our analysis showed that organisation entities have higher performance losses performance decrease compared to person and location entities. Our participation provides and estimation of how much a system trained on recent radio transcripts decreases its quality when faced with historical data.

References

1. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: Proceedings of the eleventh international conference on data engineering. pp. 3–14. IEEE (1995)
2. Ehrmann, M., Nouvel, D., Rosset, S.: Named Entity Resources-Overview and Outlook. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3349–3356 (2016)
3. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Impresso Named Entity Annotation Guidelines (Jan 2020). https://doi.org/10.5281/zenodo.3604227
4. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020). Lecture Notes in Computer Science (LNCS), vol. 12260. Springer (2020)
5. Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., Quintard, L.: Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. In: Proceedings of 5th International Joint Conference on Natural Language Processing. pp. 518–526 (2011)
6. Nouvel, D., Antoine, J.Y., Friburger, N.: Pattern Mining for Named Entity Recognition. In: Language and Technology Conference. pp. 226–237. Springer (2011)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

8. Rosset, Sophie, Grouin, Cyril, Zweigenbaum, Pierre: Entités nommées structurées : guide d'annotation Quaero. Tech. Rep. 2011-04, LIMSI-CNRS (2011)
9. Schmid, H.: Probabilistic Part-Of-Speech Tagging using Decision Trees. In: New methods in language processing. p. 154 (2013)