Named Entity Extraction in ConnectionLens based on Large Language Models

Minh Hoang Duong[1] and Ioana Manolescu[2]

[1] Ecole Polytechnique, 91120 Palaiseau, France
[2] Inria, France

Abstract. The ConnectionLens system was developed in prior work to interconnect and analyze very heterogeneous datasets, leveraging in particular Named Entity Extraction (NEE-, in English and French. ConnectionLens comprised some NE modules based on small models whose quality is not very good, and a NE based on ChatGPT-4, of better quality but expensive. In this project, I investigated free, large language models (LLMs) and their exploitation for NE. A **NE benchmark** has been built, and several free models evaluated, through the Ollama framework. We concluded that Gemma2 9B parameters by Google is the most efficient and accurate model across all test cases, attaining a precisio of **92.86%** for English and **80.94%** for French.

Keywords: Entity Extraction · Machine Learning · Natural Language Processing.

1 Introduction

ConnectionLens [1] is a system that integrates data of various data models and formats, such as CSV, JSON, HTML, text, etc. into unified data graphs. An important feature of ConnectionLens is that it **extracts named entities from text fields** found in any of the data sources which are then categorized into Organization, Location or Person.

The following models were already available within ConnectionLens: StanfordNER, English and French Flair, and a model based on ChatGPT. In this research project, our goal has been to **investigate other Large Language Models (LLMs)** that could be used in ConnectionLens instead of ChatGPT. Our goal is a model **free of charge**, of **high quality**. Extraction speed is also desirable; however, the quality of the results was of paramount interest.

2 Evaluation method and prompt engineering

To determine how well a model performs, we need to compare its results with other results considered perfect, or **gold standard**. It can be quantified as a **confusion matrix** as shown in Figure 1.

In such a matrix, a number $N_{x,y}$, where $x, y \in \{\text{Person, Organization, Location, None}\}$ is the number of entities that is of type x by the **gold standard** and is y according to our model. The None category is that the entity was not recognized in the respective results. These can be aggregated into a single **accuracy score** that is computed by the following formula:

Gold std. vs Model	Organization	Location	Person	None
Organization	$N_{O,O}$	$N_{O,L}$	$N_{O,P}$	$N_{O,N}$
Location	$N_{L,O}$	$N_{L,L}$	$N_{L,P}$	$N_{L,N}$
Person	$N_{P,O}$	$N_{P,L}$	$N_{P,P}$	$N_{P,N}$
None	$N_{N,O}$	$N_{N,L}$	$N_{N,P}$	$N_{N,N}$

Fig. 1. Comparing a model's extraction results to the gold standard

$$\frac{N_{P,P}+N_{O,O}+N_{L,L}}{\sum_{x,y\in\{P,L,O,N\}} N_{x,y}}$$

In our study, based on recent encouraging results obtained by the team using ChatGPT for the NE task [2], we will take the results of ChatGPT, more specifically GPT-4o (https://openai.com/api/), as the **gold standard**, as it is the state-of-the-art LLM for a wide range of natural language processing tasks nowadays.

2.1 The gold standard and prompt engineering

Ideally, we would want the gold standard outputs, or the outputs of GPT-4o to be **deterministic** for testing purposes. Therefore, we must adjust the model's parameters to achieve it—mainly through setting the temperature to 0 and prompt engineering. The final prompt we devised contains instructions such as: do not split Person's names; split Locations into multiple elements; return a JSON dictionary with categories as keys, sort them in the order they appear in the text, etc.

We prepared two datasets for each language to experiment with the models on and evaluate. The **English corpus** is a collection of 300 sentences extracted from the PubMed library, while the **French corpus** consists of 10 French Press articles extracted from real news articles.

To measure the quality of results obtained with each model and prompt, for each dataset of each language, we ran the same task twice times, and compare the results with those of ChatGPT, using the same **confusion matrix** as above. In the end, we had **98.67%** for English and **96.94%** for French. Thus, we concluded that it is consistent enough for our purposes.

3 Results of Open-sourced models

We evaluated other open-sourced LLMs by comparing their results with the **gold standard** generated by GPT-4o. Here, we used Ollama, a platform used to run and configure multiple LLMs locally and effortlessly (https://ollama.com/). We will also use the same prompt engineering technique for other LLMs.

In the experiment, we ran the task on multiple LLMs, notably the Llama3 family (https://ai.meta.com/blog/meta-llama-3/), Mistral (https://mistral.ai/), and Gemma2 (https://blog.google/technology/developers/google-gemma-2/).

Results The results for running on the English medical dataset can be found in Figure 2 (left). This corresponds to a total of 462 entities and an accuracy

	Org	Loc	Per	None
Org	174	1	0	5
Loc	1	142	0	3
Per	1	3	113	0
None	13	6	0	0

	Org	Loc	Per	None
Org	139	0	0	31
Loc	0	58	0	8
Per	0	0	113	3
None	20	9	2	0

Fig. 2. Gemma2-9B results for English (left) and French (right).

of **92.86%**. Similarly, the results on the French dataset are shown in Figure 2 (right). This corresponds to a total of 383 entities and accuracy of **80.94%**.

Errors observed in the results of both models

- Difference in splitting long and confusing addresses and organizations, i.e. "Department of Neurology, Experimental Neurology and Neurosurgery" is seen as two entities: "Department of Neurology" and "Experimental Neurology and Neurosurgery"
- Common nouns such as "ville" or "ecologistes" are easier to get confused, i.e. in one of the text, the "ville" did something and it is seen as an organization, or "ecologistes" is inconsistently picked up by the model despite the context.
- Locations when they are mentioned as an organization, i.e. in "France worked with other countries", it is recognized an organization.

Although these limitations of Gemma2 do exist, it does not happen all the time and it is still optimal compared to the other models. It is also notable to mention that these mistakes appear more common in French tasks than in English.

Based on these results, we selected Gemma2 and included it in the core of ConnectionLens. The server side and API endpoints are handled with Ollama's hosting and serving features, while in the core Java code new sections have been added to handle the new model option.

References

1. Anadiotis, A.C., Balalau, O., Conceicao, C., Galhardas, H., Haddad, M.Y., Manolescu, I., Merabti, T., You, J.: Graph integration of structured, semistructured and unstructured data for data journalism. Information Systems **104**, 101846 (Oct 2022). https://doi.org/10.1016/j.is.2021.101846, https://inria.hal.science/hal-03150441
2. Barret, N., Gauquier, A., Law, J.J., Manolescu, I.: Exploring heterogeneous data graphs through their entity paths. Information Systems (2024), invited paper, revised version currently under evaluation