Named Entity Recognition (NER)

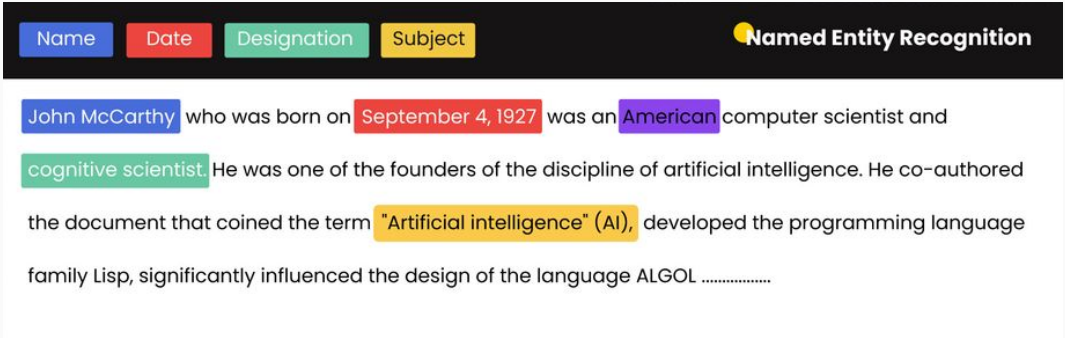
A focus on the GliNER Model Using Twitter Data During COVID-19

Olivier Caron

October 17, 2024

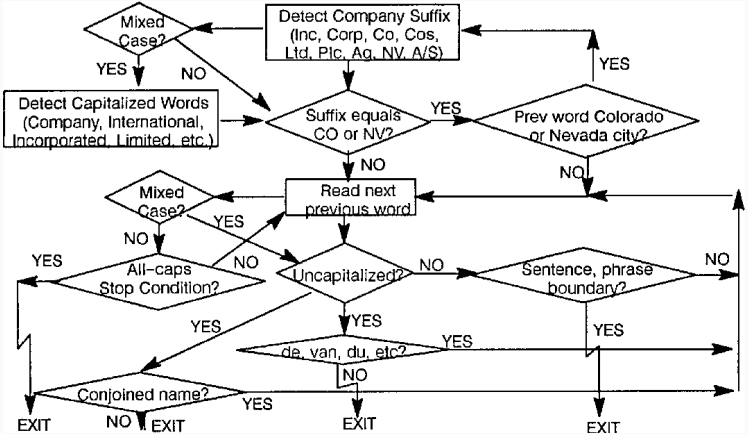
Introduction to Named Entity Recognition (NER)

- **Definition**: Named Entity Recognition (**NER**) is an information extraction task (Goyal, Gupta, and Kumar 2018) in Natural Language Processing (**NLP**) for identifying **entities** like names of **people**, **locations**, **organizations**, **dates**, and numerical expressions within text.



Historical overview of NER

- **Early Contributions**: Lisa Rau's (1991) work on extracting company names was pioneering, addressing the challenges posed by unknown words in NLP tasks.

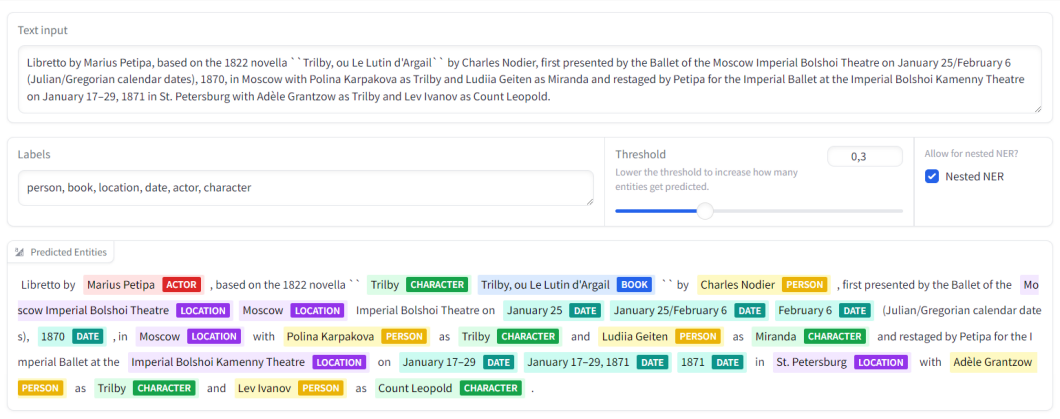


Historical overview of NER

- **Origins**: The term “Named Entitty” was formally introduced in the Sixth Message Understanding Conference (MUC-6), where it was recognized as essential for **extracting structured information**, such as names and temporal expressions (Grishman and Sundheim 1996).
- **Evolution**: Over time, methods evolved from rule-based approaches to **machine learning** and, eventually, to **deep learning**-based models (Nadeau and Sekine 2007; Liu, Chen, and Xia 2022).

GliNER: a Zero-shot NER Model

- **G**eneralist and **Li**ghtweight Model for **N**amed **E**ntity **R**ecognition (Zaratiana et al. 2024)



GliNER: Benchmark

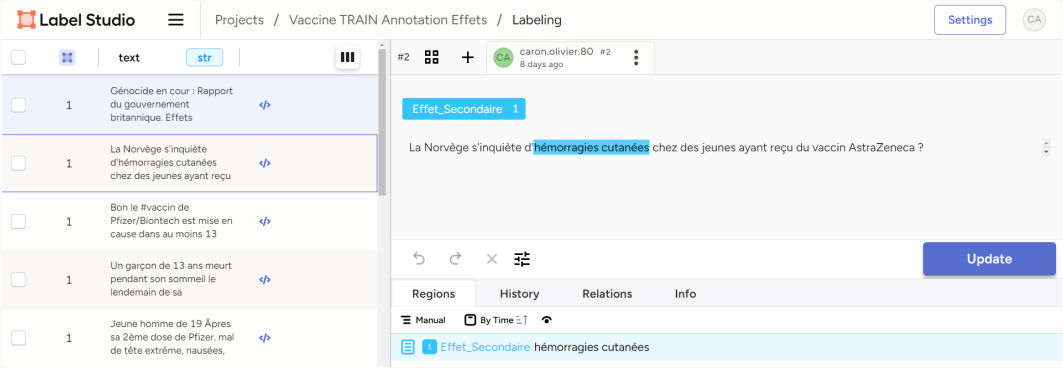
Model	Params	Movie	Restaurant	AI	Literature	Music	Politics	Science	Average
Vicuna-7B	7B	6.0	5.3	12.8	16.1	17.0	20.5	13.0	13.0
Vicuna-13B	13B	0.9	0.4	22.7	22.7	26.6	27.0	22.0	17.5
USM	0.3B	37.7	17.7	28.2	56.0	44.9	36.1	44.0	37.8
ChatGPT	–	5.3	32.8	52.4	39.8	66.6	68.5	**67.0**	47.5
InstructUIE	11B	**63.0**	21.0	49.0	47.2	53.2	48.1	49.2	47.2
UniNER-7B	7B	42.4	31.7	53.6	59.3	67.0	60.9	61.1	53.7
UniNER-13B	13B	48.7	36.2	54.2	60.9	64.5	61.4	63.5	55.6
GoLLIE	7B	**63.0**	**43.4**	**59.1**	62.7	67.8	57.2	55.5	58.0
GLiNER-S	50M	46.9	33.3	50.7	60.0	60.9	61.5	55.6	52.7
GLiNER-M	90M	42.9	37.3	51.8	59.7	69.4	68.6	58.1	55.4
GLiNER-L	0.3B	57.2	42.9	57.2	**64.4**	**69.6**	**72.6**	62.6	**60.9**

Table 1: **Zero-Shot Scores on Out-of-Domain NER Benchmark.** We report the performance of GLiNER with various DeBERTa-v3 (He et al., 2021) model sizes. Results for Vicuna, ChatGPT, and UniNER are from Zhou et al. (2023); USM and InstructUIE are from Wang et al. (2023); and GoLLIE is from Sainz et al. (2023).

Finetuning - Labelling data

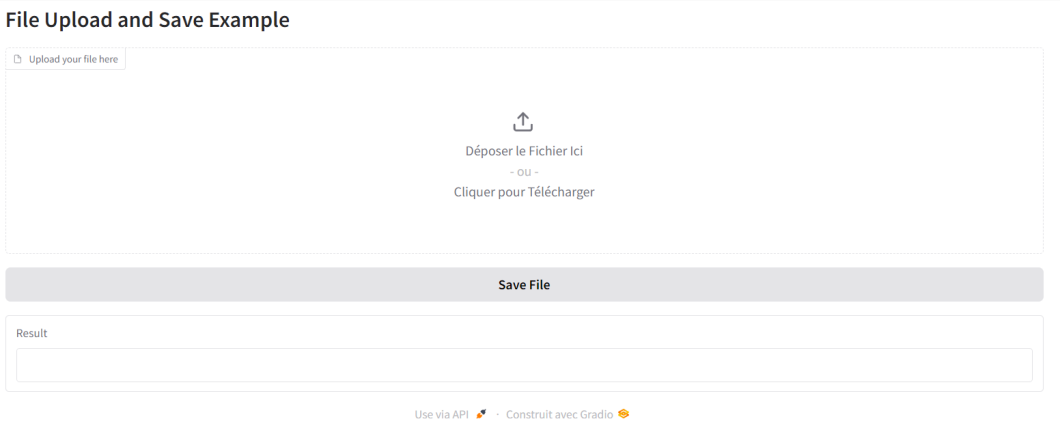
Start Label Studio with:

```
label-studio start
```

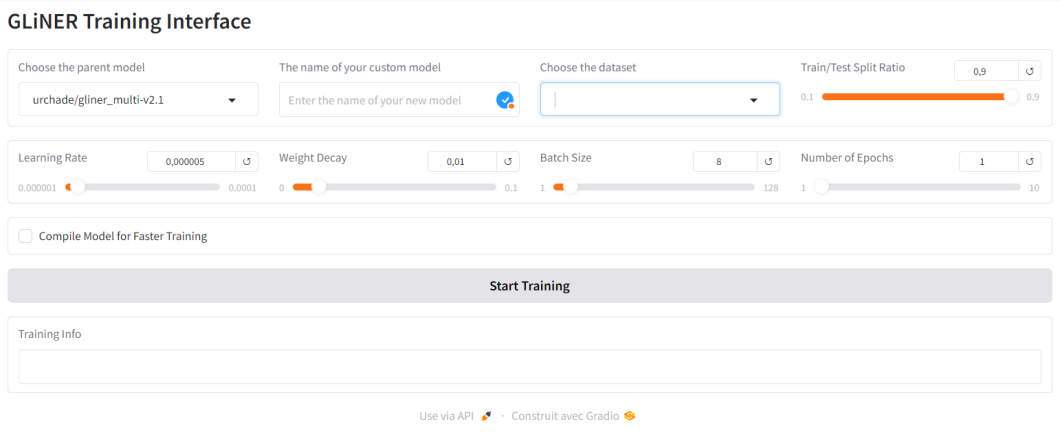


Finetuning - Upload annotaded data

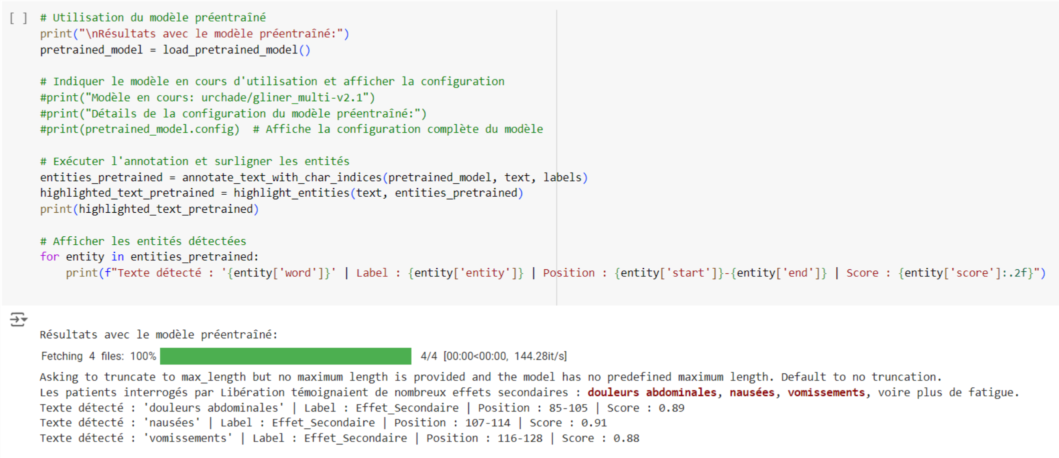
I used the GliNER Studio Colab notebook (Stepanov and Shtopko 2024) from Knowledgator.



Finetuning - Train the model



Finetuning - gliner_multi-v2.1



```
[ ] # Utilisation du modèle préentraîné
    print("\nRésultats avec le modèle préentraîné:")
    pretrained_model = load_pretrained_model()

    # Indiquer le modèle en cours d'utilisation et afficher la configuration
    #print("Modèle en cours: urchade/gliner_multi-v2.1")
    #print("Détails de la configuration du modèle préentraîné:")
    #print(pretrained_model.config)  # Affiche la configuration complète du modèle

    # Exécuter l'annotation et surligner les entités
    entities_pretrained = annotate_text_with_char_indices(pretrained_model, text, labels)
    highlighted_text_pretrained = highlight_entities(text, entities_pretrained)
    print(highlighted_text_pretrained)

    # Afficher les entités détectées
    for entity in entities_pretrained:
        print(f"Texte détecté : '{entity['word']}' | Label : {entity['entity']} | Position : {entity['start']}-{entity['end']} | Score : {entity['score']:.2f}")
```

```
Résultats avec le modèle préentraîné:
Fetching 4 files: 100% 4/4 [00:00<00:00, 144.28it/s]
Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.
Les patients interrogés par Libération témoignaient de nombreux effets secondaires : douleurs abdominales, nausées, vomissements, voire plus de fatigue.
Texte détecté : 'douleurs abdominales' | Label : Effet_Secondaire | Position : 85-105 | Score : 0.89
Texte détecté : 'nausées' | Label : Effet_Secondaire | Position : 107-114 | Score : 0.91
Texte détecté : 'vomissements' | Label : Effet_Secondaire | Position : 116-128 | Score : 0.88
```

Finetuning - Custom model for Covid-19 vaccines' side effects

[9]

```python
# Utilisation du modèle personnalisé
print("Résultats avec le modèle personnalisé:")
custom_model_path = "drive/MyDrive/Models/custom_model"
custom_model = load_custom_model(custom_model_path)

# Indiquer le modèle en cours d'utilisation et afficher la configuration
#print(f"Modèle en cours: {custom_model_path}")
#print("Détails de la configuration du modèle personnalisé:")
#print(custom_model.config)  # Affiche la configuration complète du modèle

# Exécuter l'annotation et surligner les entités
entities_custom = annotate_text_with_char_indices(custom_model, text, labels)
highlighted_text_custom = highlight_entities(text, entities_custom)
print(highlighted_text_custom)

# Afficher les entités détectées
for entity in entities_custom:
    print(f"Texte détecté : '{entity['word']}' | Label : {entity['entity']} | Position : {entity['start']}-{entity['end']} | Score : {entity['score']:.2f}")
```

```
Résultats avec le modèle personnalisé:
config.json not found in /content/drive/MyDrive/Models/custom_model
WARNING:huggingface_hub.hub_mixin:config.json not found in /content/drive/MyDrive/Models/custom_model
Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.
Les patients interrogés par Libération témoignaient de nombreux effets secondaires : douleurs abdominales, nausées, vomissements, voire plus de fatigue.
Texte détecté : 'douleurs abdominales' | Label : Effet_Secondaire | Position : 85-105 | Score : 0.97
Texte détecté : 'nausées' | Label : Effet_Secondaire | Position : 107-114 | Score : 0.97
Texte détecté : 'vomissements' | Label : Effet_Secondaire | Position : 116-128 | Score : 0.98
Texte détecté : 'fatigue' | Label : Effet_Secondaire | Position : 144-151 | Score : 0.83
```

Pharmaceutical Brands

Pfizer

AstraZeneca

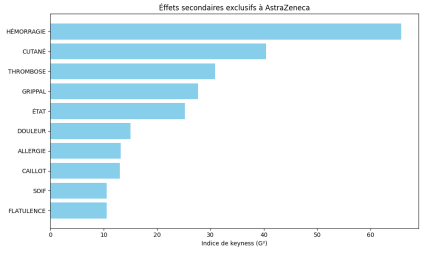
Moderna

Sanofi

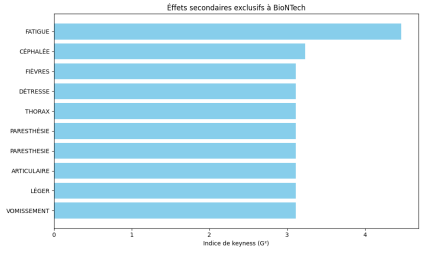
BioNTech

Pharmaceutical Companies

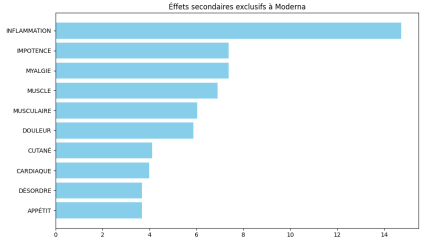
AstraZeneca



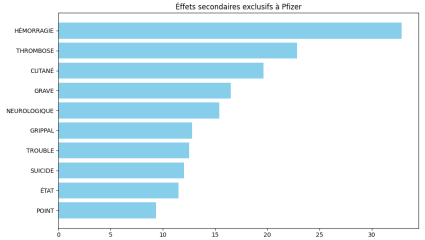
BioNTech



Moderna



Pfizer



StreamLit App for your data

- https://huggingface.co/spaces/oliviercaron/GLiNER_file
- Clone the repository on your own computer:
- https://github.com/oliviercaron/GliNER_streamlit

Bibliography

Goyal, Archana, Vishal Gupta, and Manish Kumar. 2018. “Recent Named Entity Recognition and Classification Techniques: A Systematic Review.” *Computer Science Review* 29 (August): 21–43. https://doi.org/10.1016/j.cosrev.2018.06.001.

Grishman, Ralph, and Beth M Sundheim. 1996. “Message Understanding Conference-6: A Brief History.” In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Liu, Xing, Huiqin Chen, and Wangui Xia. 2022. “Overview of Named Entity Recognition.” *Journal of Contemporary Educational Research* 6 (5): 65–68.

Nadeau, David, and Satoshi Sekine. 2007. “A Survey of Named Entity Recognition and Classification.” *Lingvisticae Investigationes* 30 (1): 3–26.

Rau, Lisa F. 1991. “Extracting Company Names from Text.” In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, 29–30. IEEE