Semester project
January 2025

Knowledge graph generation

Steve Taka
Benjamin Latrie
Abderahim Lagraoui

Supervised by :
Ms. Olha Nahorna

Table of contents

01
Introduction
What is a knowledge graph ? Why this project ?

02
Previous Work
Overview of last year's work

03
New Pipeline
Overview of the structured workflow

04
AI Methods
Model fine-tuning

05
Evaluation
Performance and quality measurement, comparison

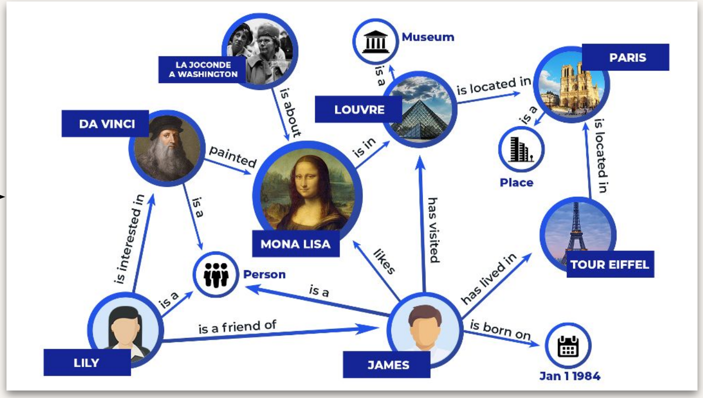
06
Environmental study
Analysis of ecological impact

07
Conclusion
Summary and perspectives

01
Introduction

Introduction

- Network of interconnected entities and relationships
- Represents and organizes knowledge
- Extracts, structures and organizes relevant information
- Easy to read and very visual



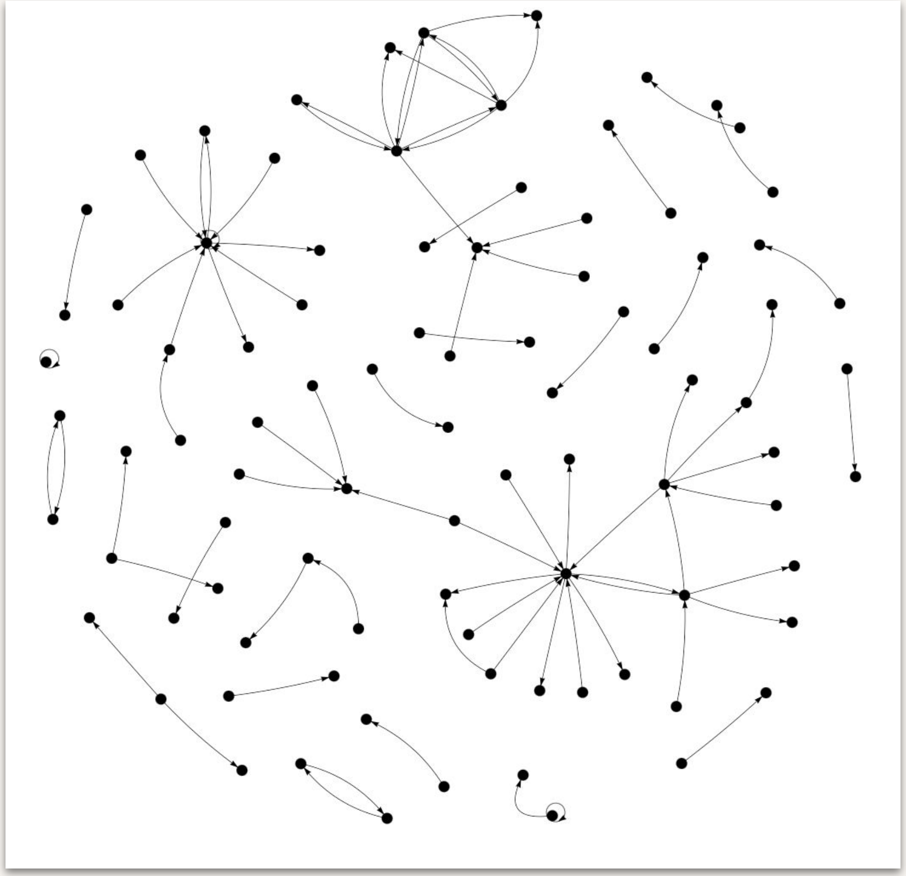
Triplet = (h, r, t)

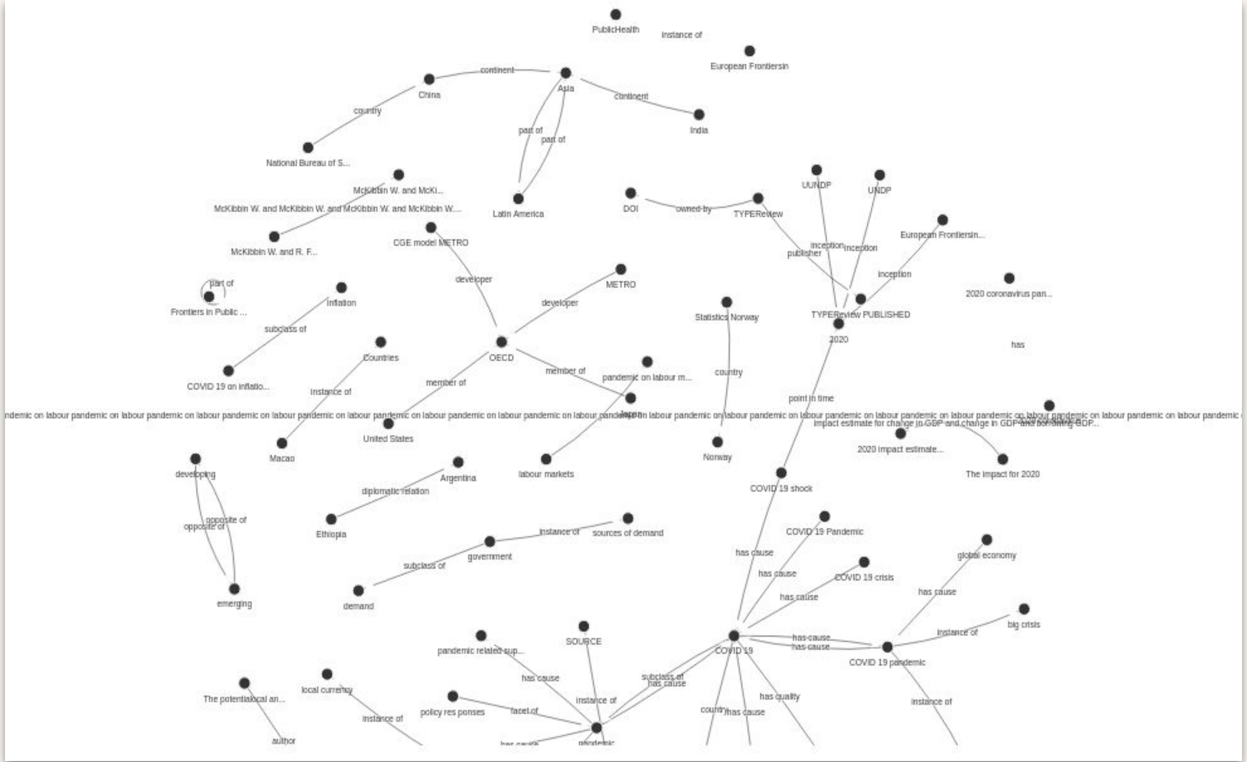
h, t are nodes
r is an edge

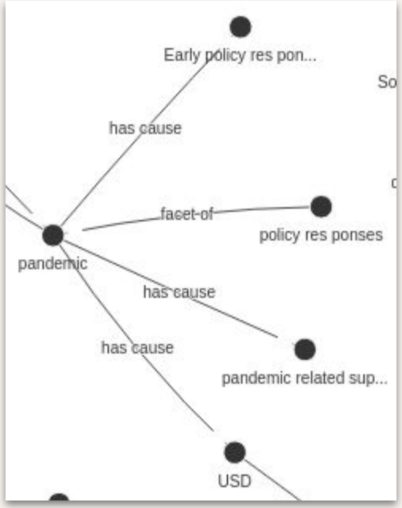
02
Previous Work

5 articles

- *“An Updated Assessment of the Economic Impact of COVID-19”* (**16 pages**)
- *“COVID-19: Fiscal Implications and Financial Stability in Developing Countries”* (**13 pages**)
- *“Understanding structural effects of COVID-19 ON THE GLOBAL ECONOMY”* (**36 pages**)
- *“COVID-19 outbreak: Impact on global economy”* (**13 pages**)
- *“The Economic Impact of COVID-19 around the World”* (**15 pages**)







"The COVID-19 pandemic and associated policy responses are likely to alter the global economy in a way that affects its ability to adjust to future shocks and changes."

Pipeline

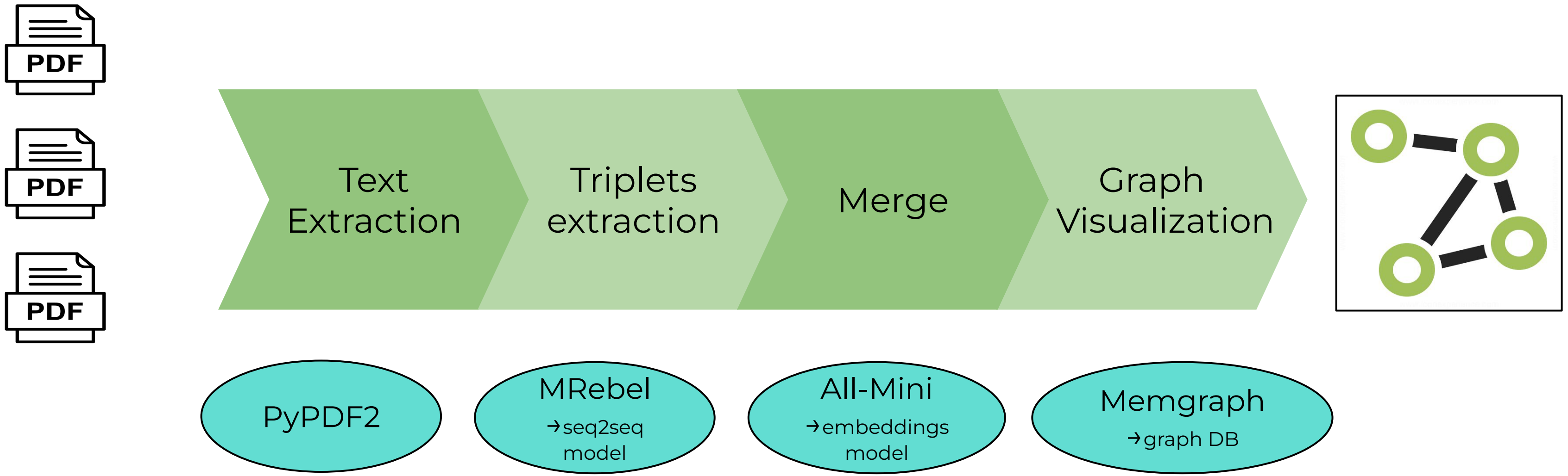


Fig 1: Pipeline from previous work

Good Points

- Quite **fast**
- Good basis for improvements
- Good **web interface**
- Use of models that combine speed and efficiency (MRebel, All-Mini)
- Use of **Memgraph**

Weaknesses

- Problem if articles are in **different languages**
- No way of assessing the **quality** of a generated graph
- Some of the triplets generated don’t seem **relevant**
- **Lack of important information** in the generated graph
- Merge module a bit **laborious**
- Presence of a **major error** in the code

03
New Pipeline

New Pipeline

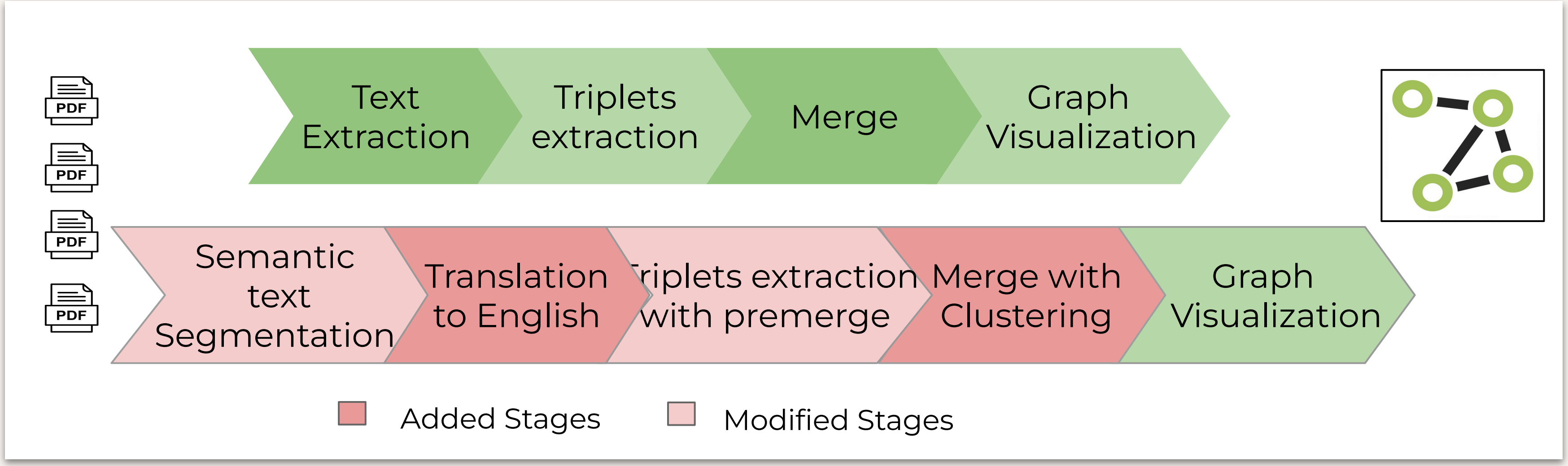


Fig 2: Proposed approach

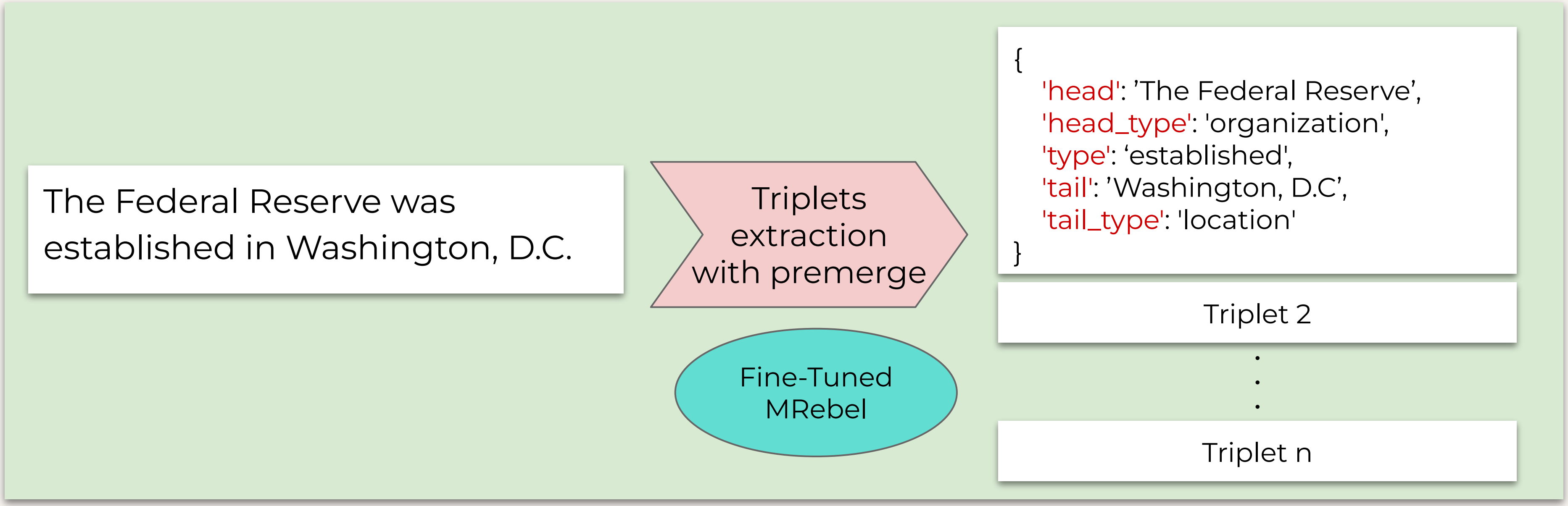
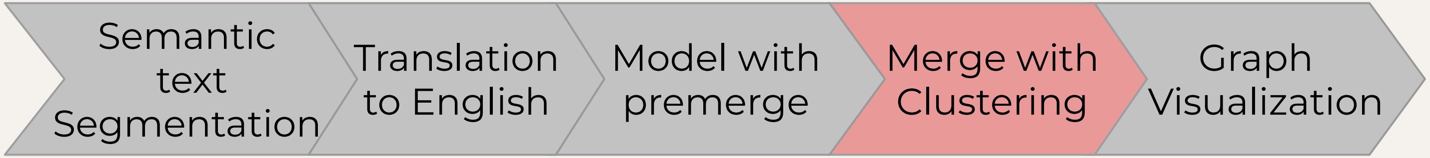


Fig 3: Extracted triplets format



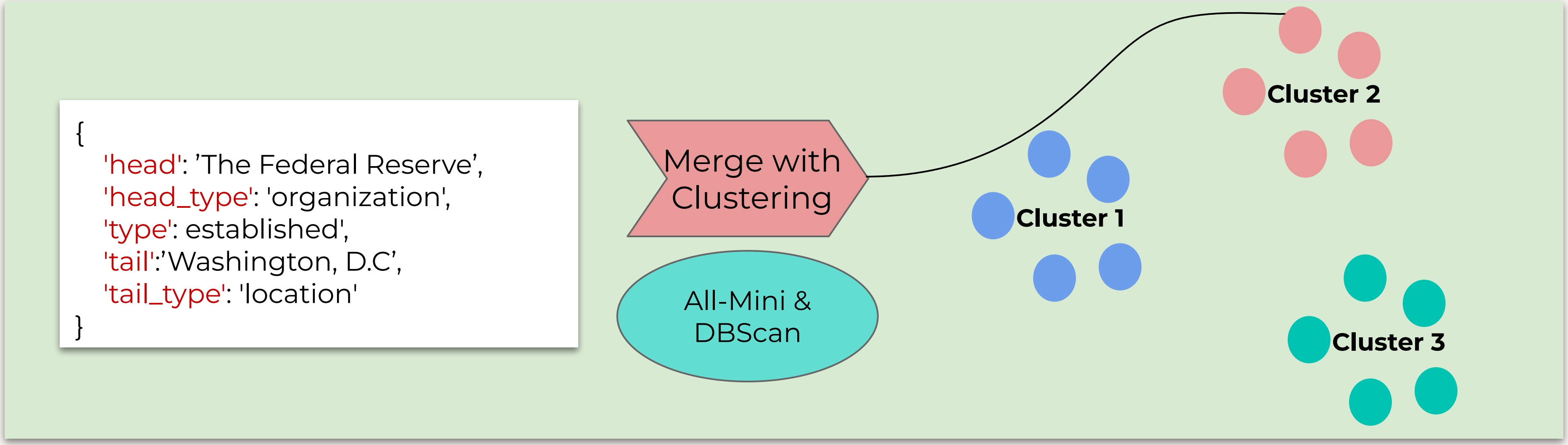


Fig 4: Merge with clustering process

Alternative for triplets generation

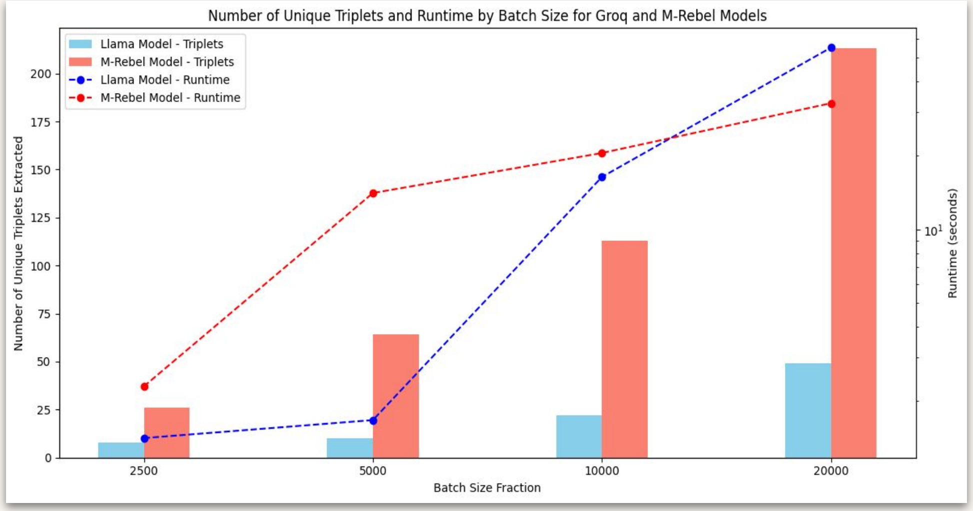


Fig 5 : Llama 3.1 and MRebel runtimes and graph sizes

- The output format of Llama is challenging to **standardize**.
- Llama **overheads** due to API connections. And it is limited with a **max_input** per minute parameter.
- Extracted triplets from Llama often feature lengthy **sentences** as **entities**.

04

AI Methods

Fine-Tuning of MRebel

To specialize it in economic data

→ 611 M parameters, F32

optimization technique : **LoRA**

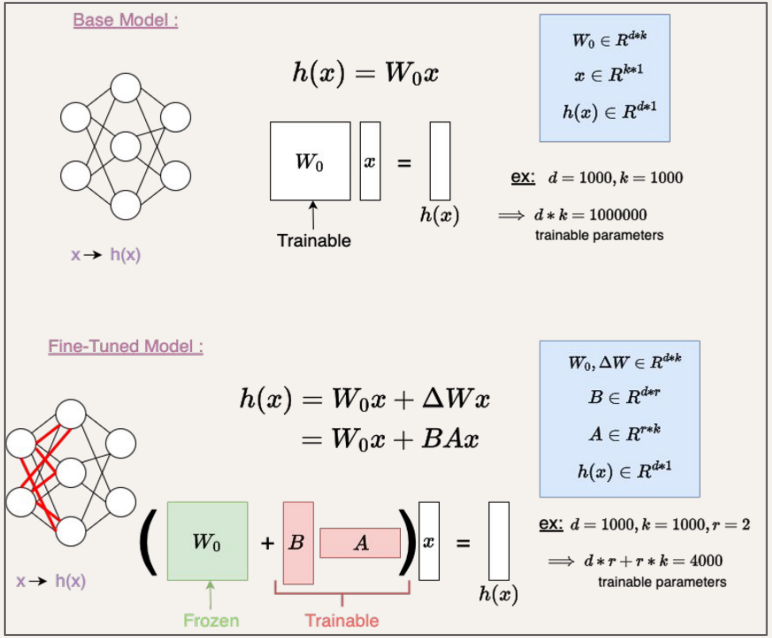


Fig 6: LoRA method

Fine-Tuning of MRebel

- ❑ Obtaining reference triplets for the training dataset using Llama3.1 (not optimal)
- ❑ Fast training thanks to LoRA
- ❑ Loss gradually decreases :

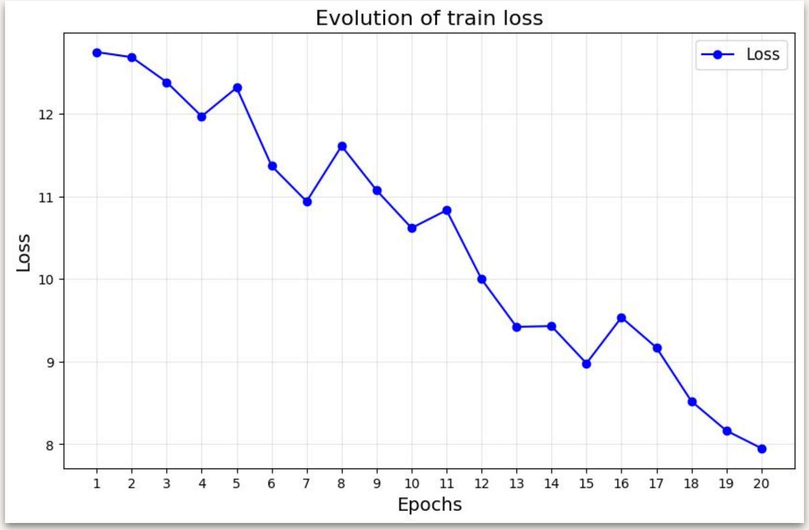


Fig 7: Training loss evolution

Fine-Tuning of All-Mini

To classify whether two triplets represent the same idea or not

- Use of a dataset of 3383 pairs of triplets, with 0 or 1 as the label (0 if merge of the 2 triplets required)
- Add of a classifier part (basic All-Mini used to extract semantic information from triplets)

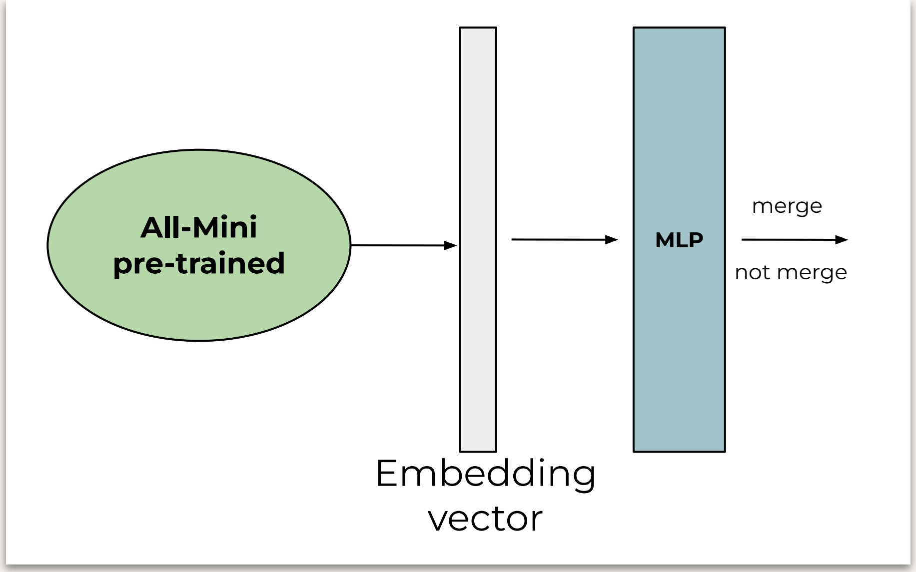
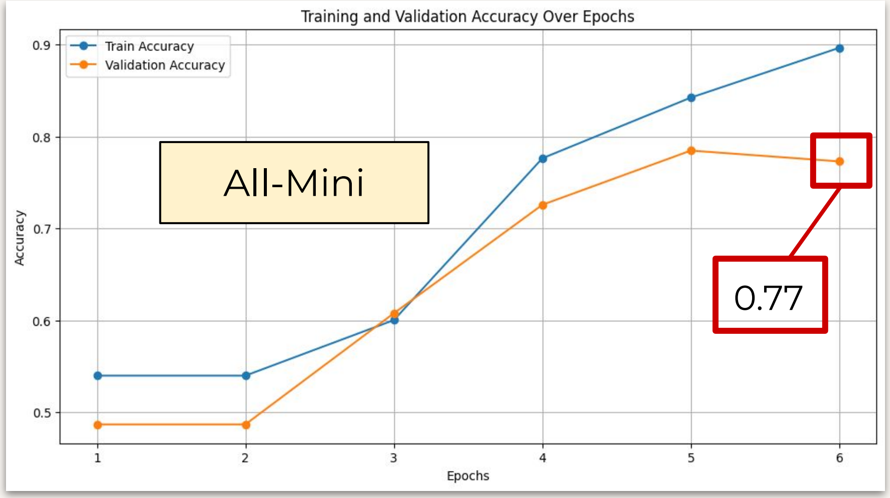


Fig 8: Fine-Tuned All-Mini architecture

Performance



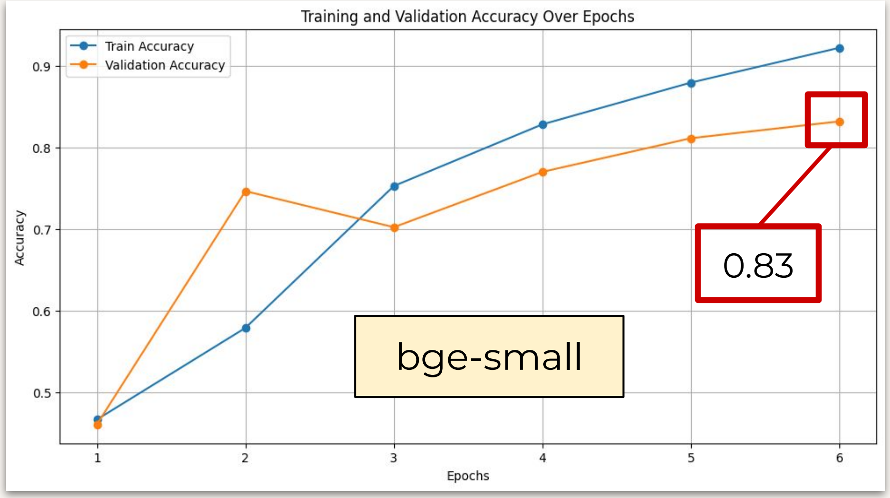


Fig 9: Accuracy for **All-Mini** and **bge-small** Fine-tuning

→ The fine-tuning of **bge-small** seems more effective, but a longer training period is observed.

05
Evaluation

Evaluation metrics

PyKEEN
Python library for
KG evaluation

$\boxed{h + r \approx t}$

Mean Rank :

$$\mathrm{MR} = \frac{1}{|T|} \sum_{(h,r,t)\in T} \mathrm{rank}(h,r,t)$$

Mean
Reciprocal Rank :

$$\mathrm{MRR} = \frac{1}{|T|} \sum_{(h,r,t)\in T} \frac{1}{\mathrm{rank}(h,r,t)}$$

Hits@10 :

$$\mathrm{Hits@10} = \frac{1}{|T|} \sum_{(h,r,t)\in T} 1(\mathrm{rank}(h,r,t) \leq 10)$$

No labeled data

→ Proposition of adapted TF - IDF (Term Frequency - Inverse Document Frequency)

$$w_{x,y} = tf_{x,y} \times \overbrace{\log\left(\frac{N}{df_x}\right)}$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

- Get n keywords
- Check the percentage of keywords in the KG

Quantitative Results

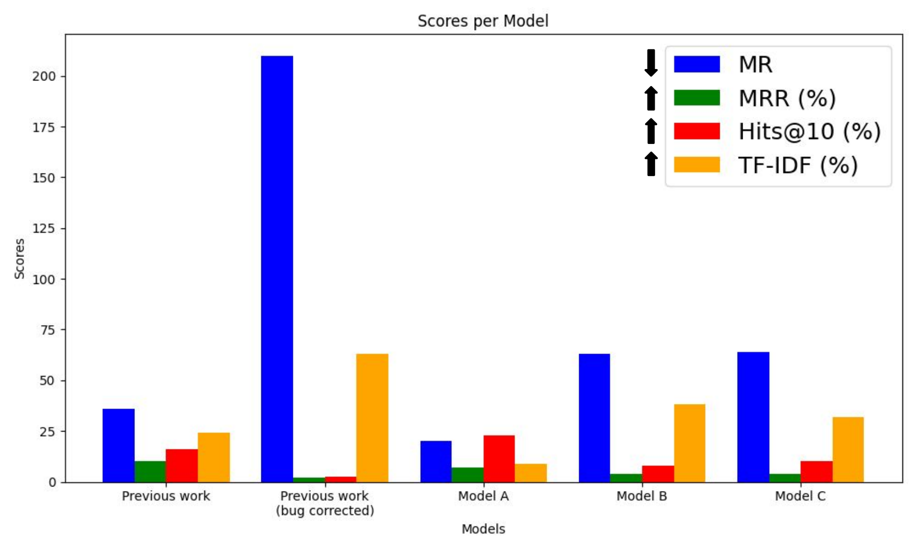


Fig 10: Metrics comparison

<u>Model A</u> :
MRebel & All-Mini
*(**Both** fine-tuned)*

<u>Model B</u> :
MRebel & All-Mini
*(**Only** All-Mini fine-tuned)*

<u>Model C</u> :
MRebel & bge-small
*(**Only** bge-small fine-tuned)*

➢ MRebel fine-tuning **not efficient** (data quality)

➢ Models with All-Mini and bge-small :
→ **nearly identical** performance

➢ Previous work better with the error fixed is bad ? →
- Too many nodes

Quantitative Results

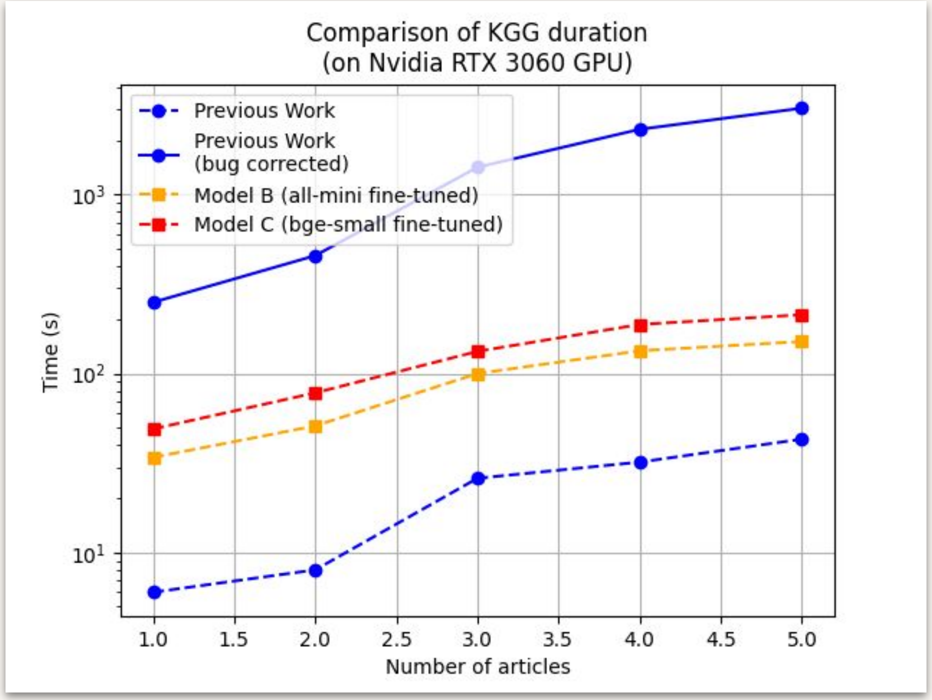
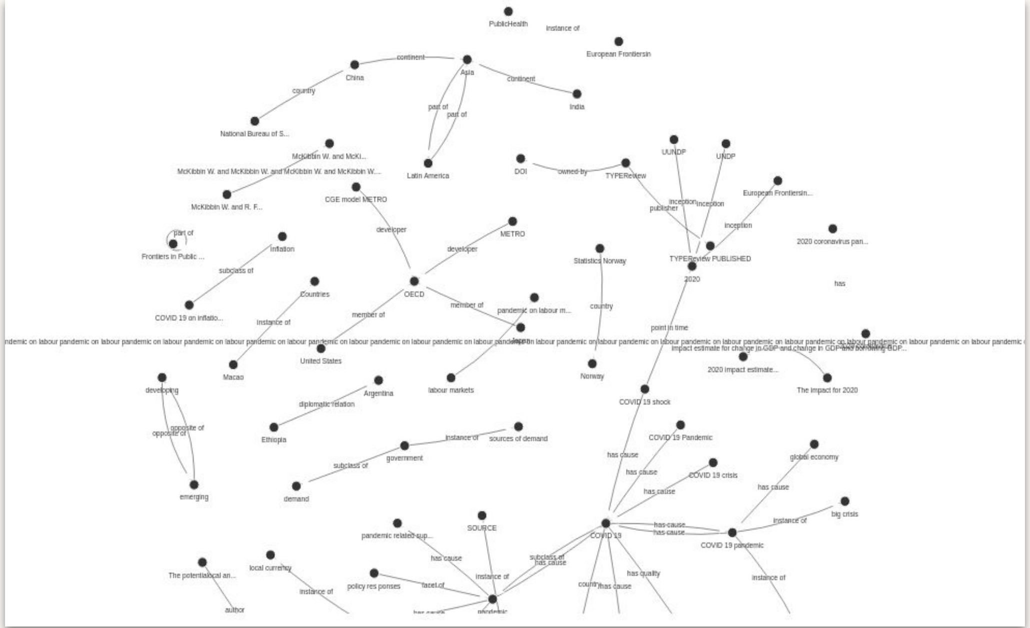


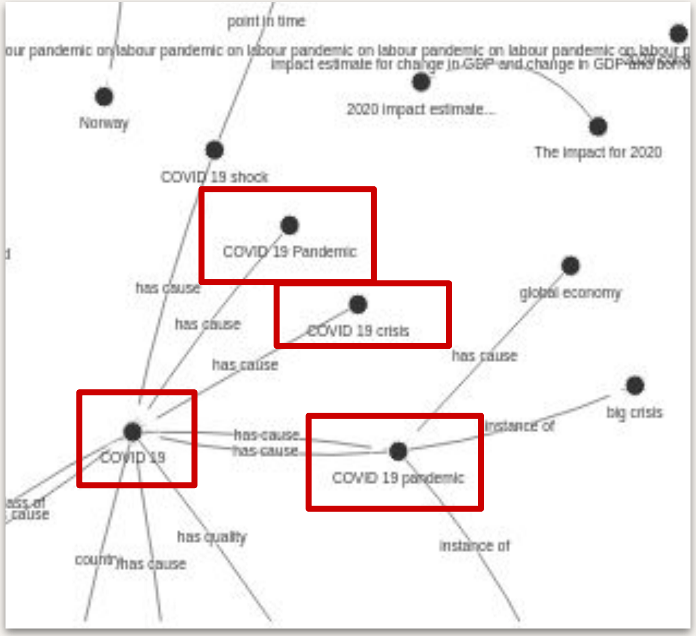
Fig 11: Generation time comparison

- ➢ Previous work with error is obviously **the fastest**
- ➢ Model B **slightly faster** than model C
- ➢ By correcting the bug, the previous work takes much more time

Qualitative Results

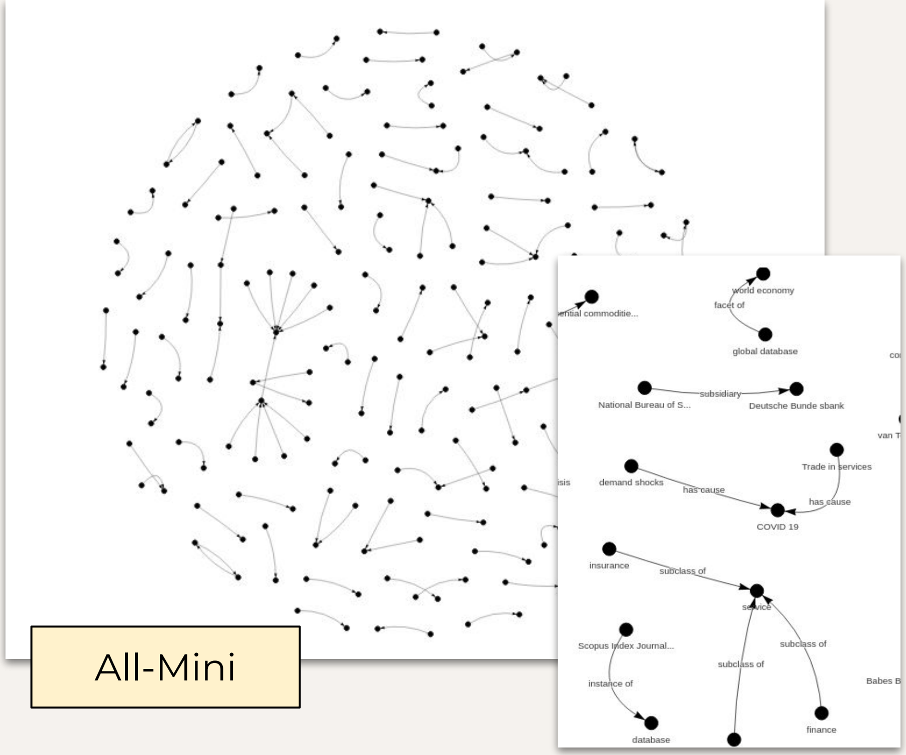


Previous work

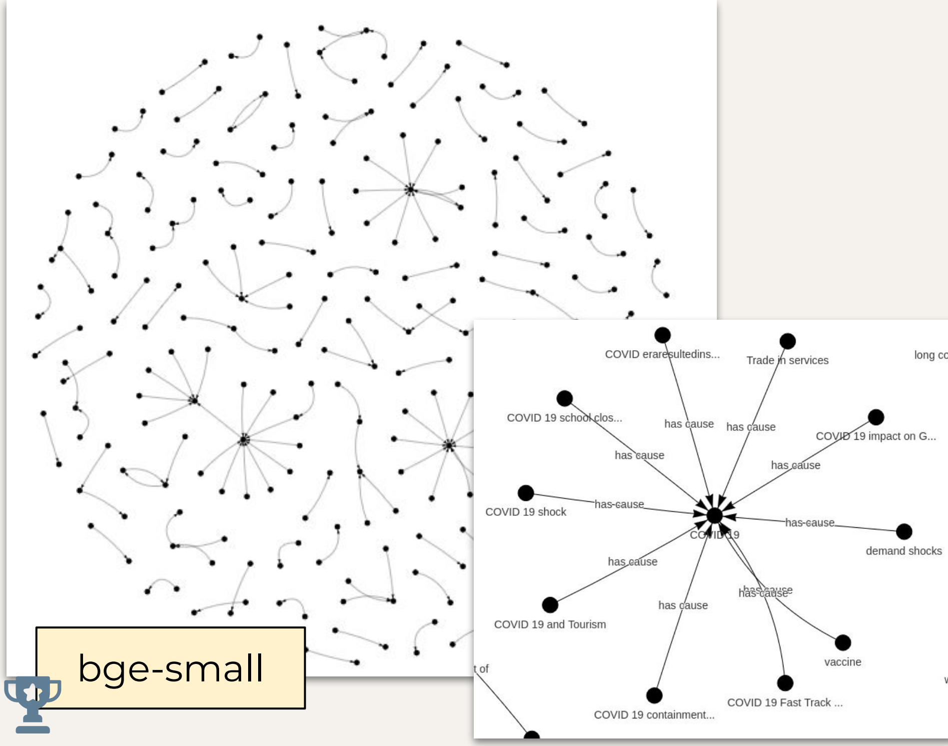


- ❏ Redundancies
- ❏ Too general
- ❏ Very long nodes

Qualitative Results



All-Mini



bge-small

06
Environmental study

Carbon footprint

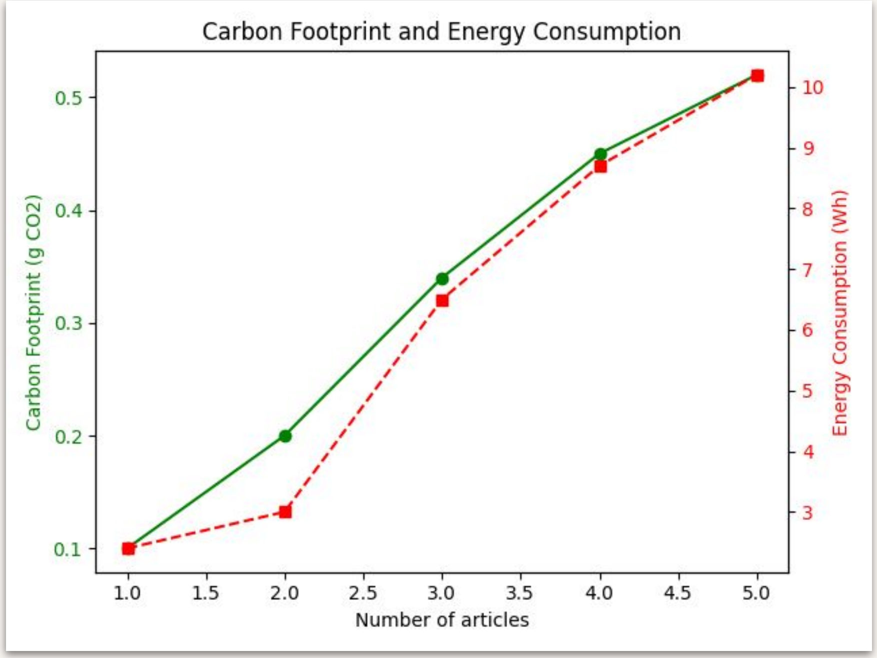
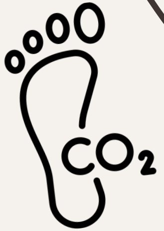


Fig 12: Carbon footprint and energy consumption



- ❏ Not too high
- ❏ More with all the model trainings
- ❏ Depends on the using rate

<u>E.g.</u> For 5000 uses after assumed deployment

Model	Carbon footprint (CO_2 eq)	Energy needed (Wh)	% of a flight Paris-London
Normal MRebel + fine-tuned bge-small	2,6 kg	51.10^3	5%

07
Conclusion

Conclusion

- The proposed approach outperforms the existing version.
- The training datasets can be better to obtain improved results
- Ontology schemas are a good possible approach to generate knowledge graphs.

Thank you for your attention !

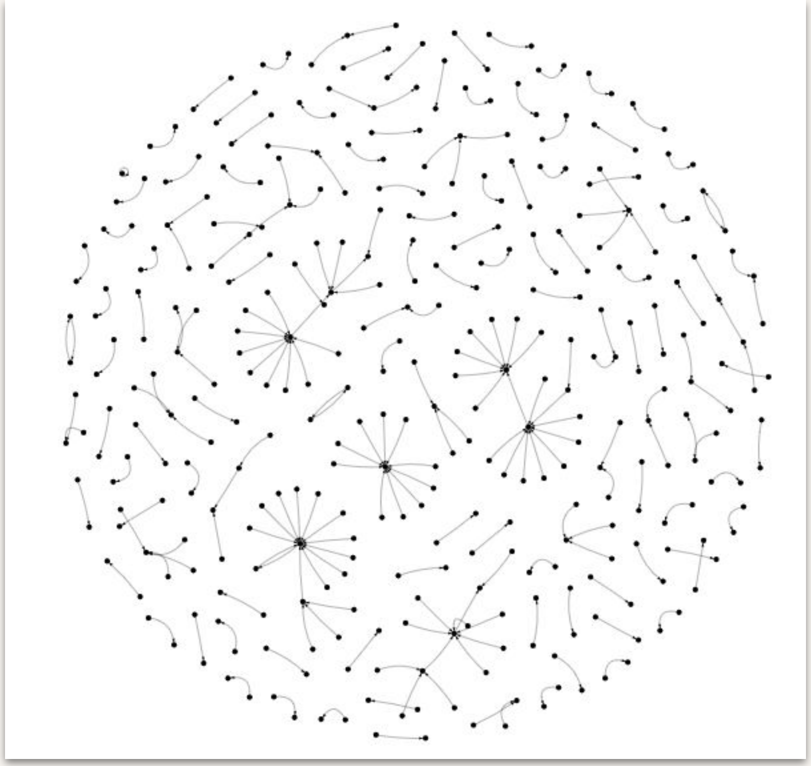
Our team

Steve

Benjamin

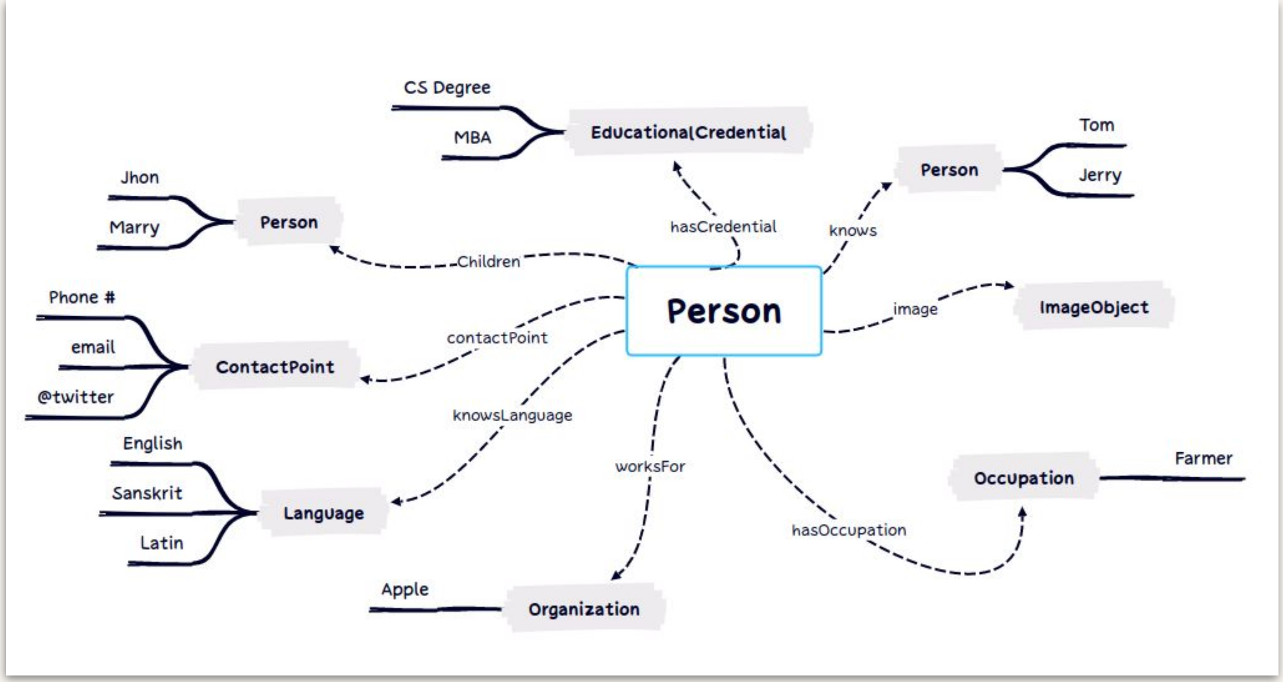
Abderahim

Scalability of graph generation



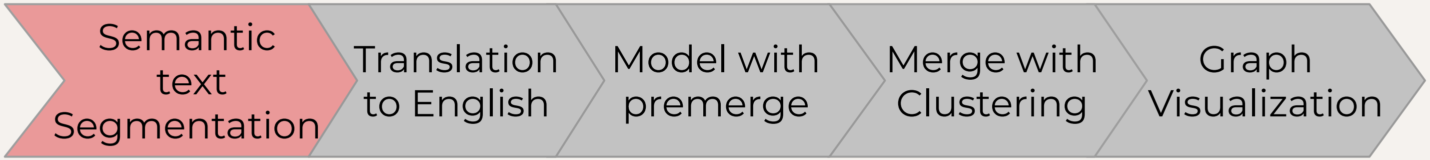
- ❖ **9 articles** (5 in english and 4 in french)
- ❖ **Nodes** : 349 - **Relations** : 229
- ❖ **Time** : 535 s
- ❖ **MR** : 110.68 - **MRR** : 0.026 - **Hits@10** : 0.054
- ❖ **TF-IDF** (30) : 0.43
- ❖ **TF-IDF** (60) : 0.45
- ❖ **TF-IDF** (100) : 0.42

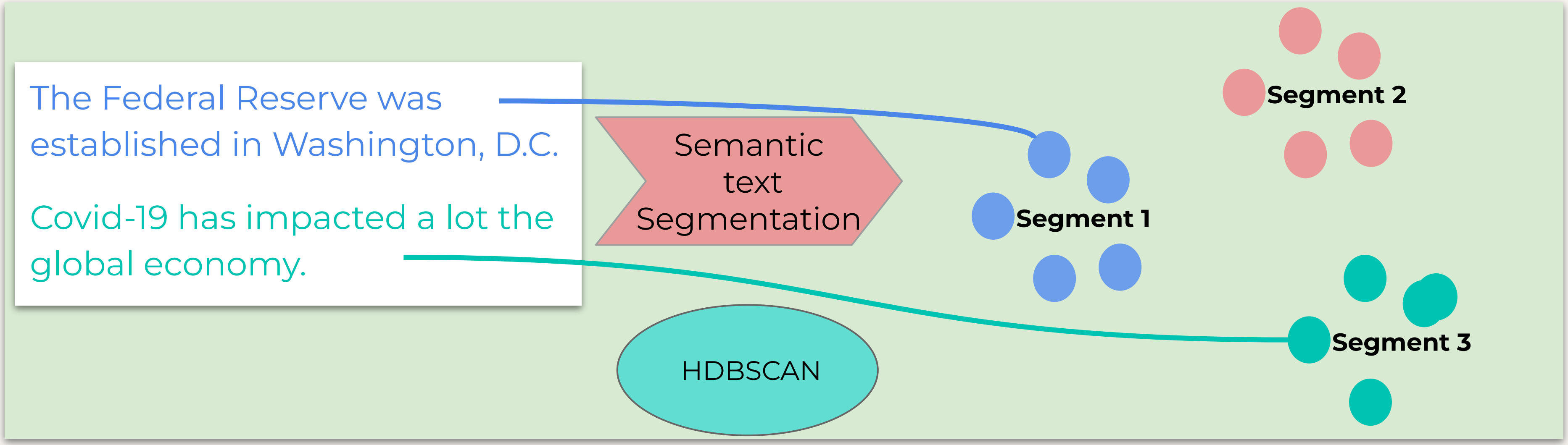
Ontology



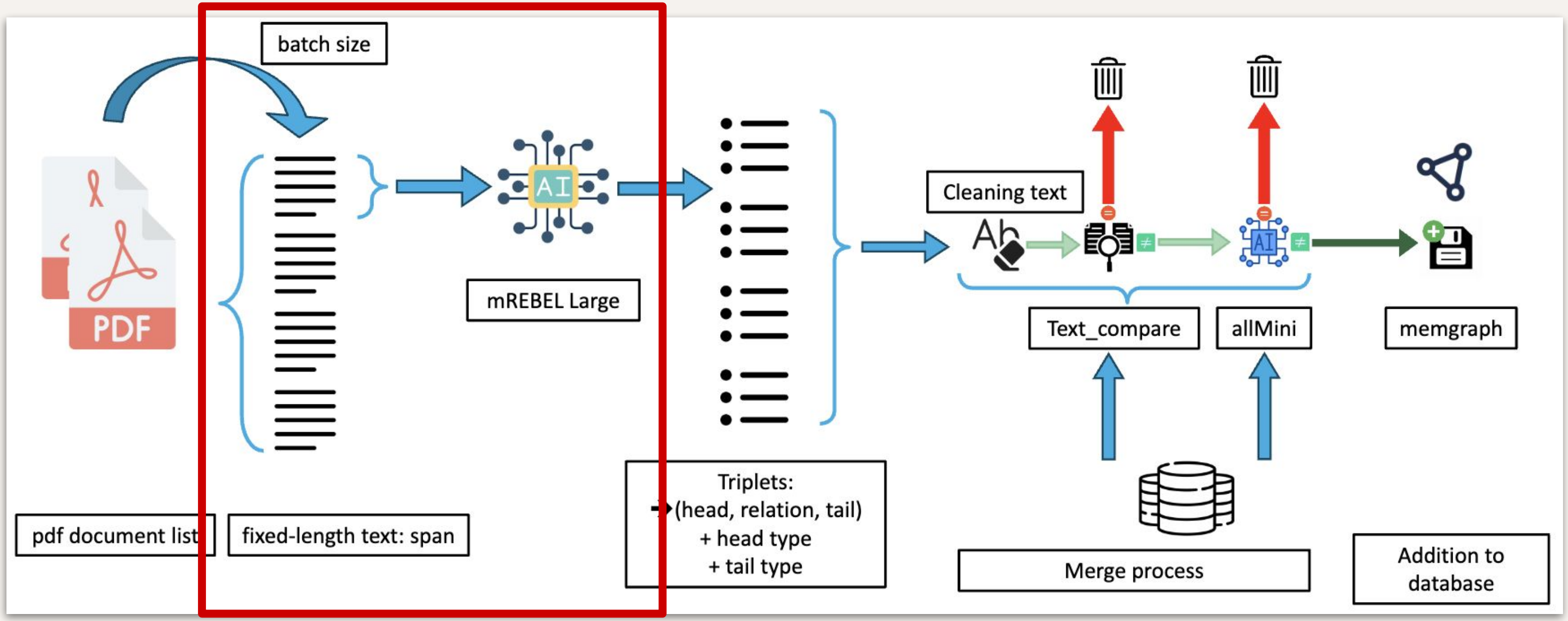
Ontology schema example

Semantic Segmentation Module





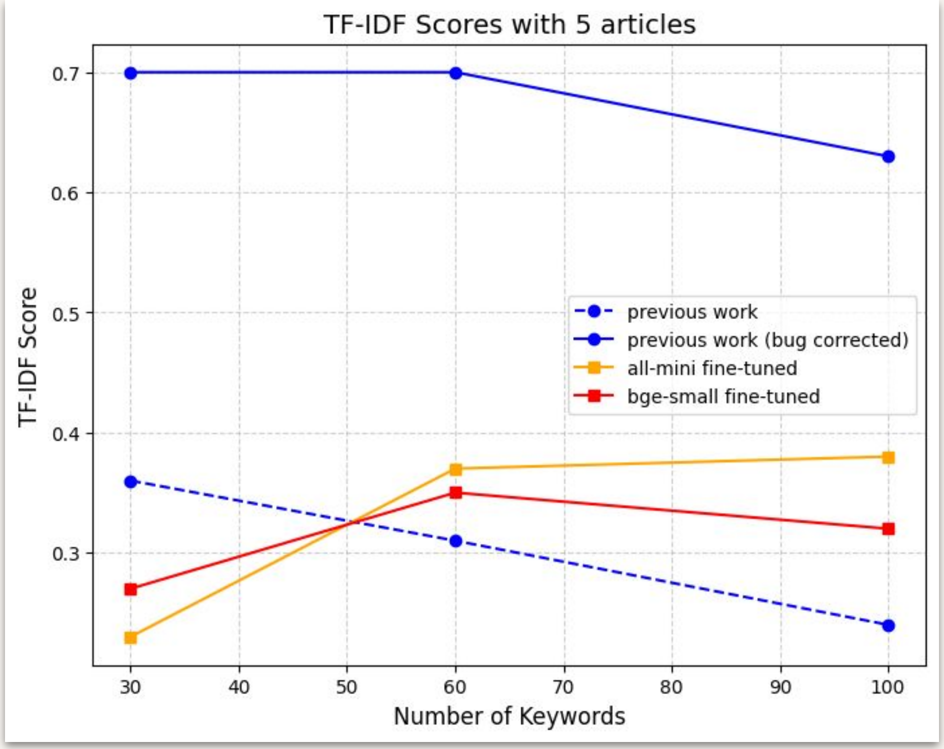
Error of previous work



- Batch size : 15000 chars
- Max length (for tokenizer) : 256 tokens ~ 1000 chars

Figure from last year's presentation

Variations in TF-IDF score



Example of keywords

Good : 2020, covid 19, pandemic, government, economic, china, policy, containment, ...

Bad : 19, country, total, world, table, tnum, wb, ...

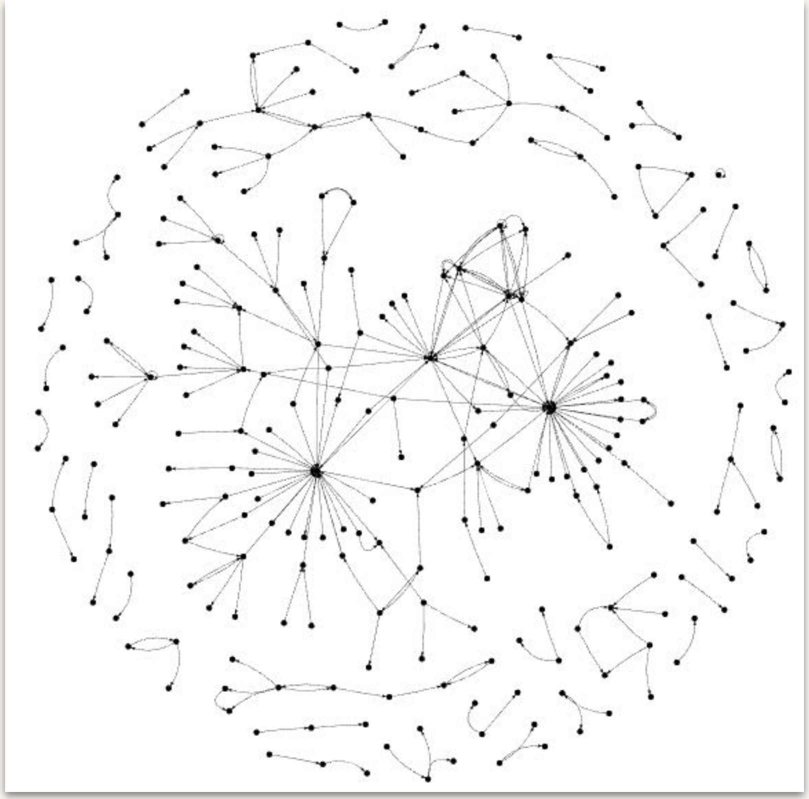
→ A majority of **good** keywords

Variation of graph dimensions

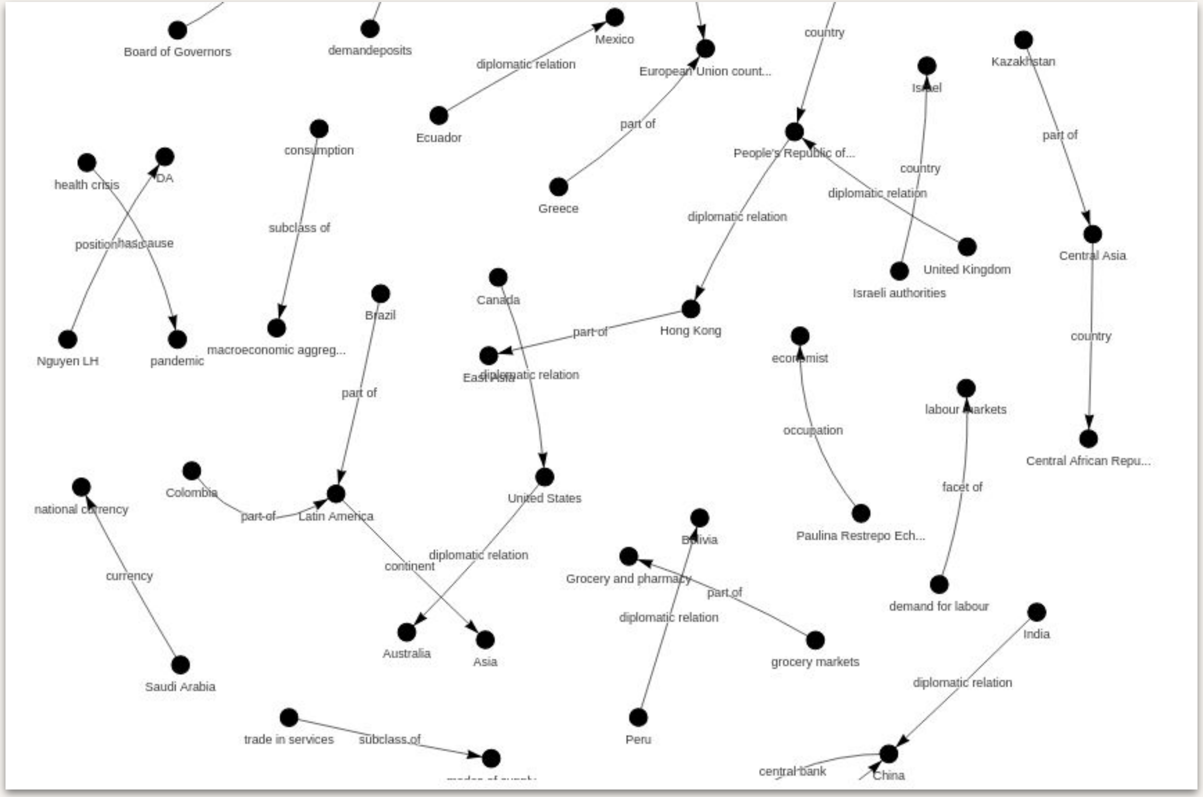
Nodes | Relations | Ratio

	1 article	2 articles	3 articles	4 articles	5 articles
Previous work	29 | 35 | 1.2	40 | 47 | 1.18	69 | 74 | 1.07	88 | 91 | 1.03	100 | 106 | 1.06
Previous work with corrected bug	276 | 360 | 1.3	351 | 507 | 1.44	698 | 940 | 1.35	923 | 1215 | 1.32	1067 | 1395 | 1.31
Fine-tuned MRebel + fine-tuned All-Mini	11 | 7 | 0.64	24 | 15 | 0.63	43 | 25 | 0.58	46 | 27 | 0.59	47 | 28 | 0.6
Basic MRebel + fine-tuned All-Mini	59 | 36 | 0.61	81 | 53 | 0.65	146 | 90 | 0.62	182 | 111 | 0.61	198 | 121 | 0.61
Basic MRebel + fine-tuned bge-small	77 | 50 | 0.65	108 | 74 | 0.69	180 | 115 | 0.64	227 | 142 | 0.63	247 | 156 | 0.63

Previous work graph (with corrected error)

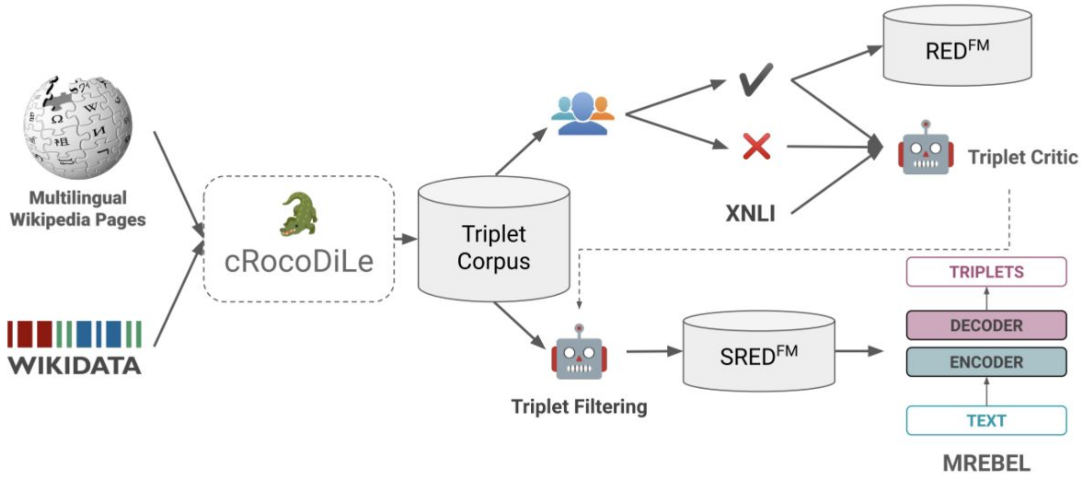


Fine-Tuned MRebel graph



MRebel

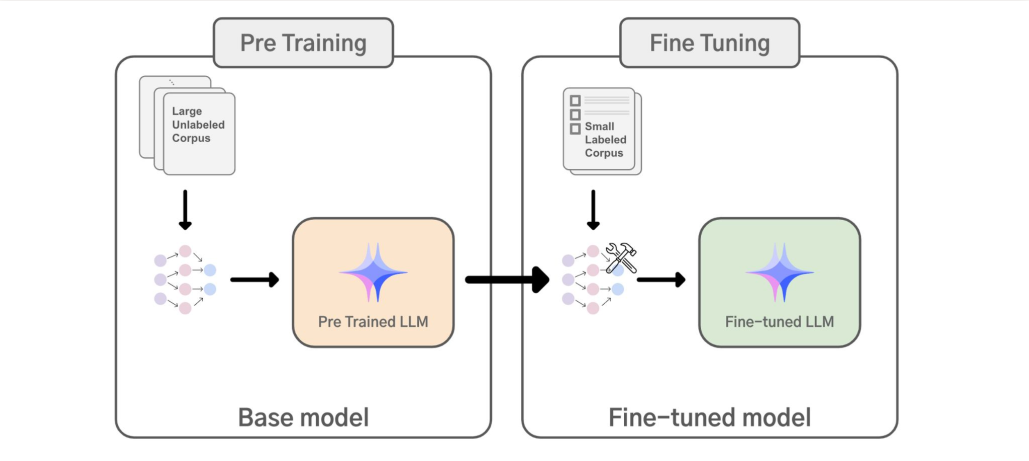
MRebel



Complete pipeline for the creation of REDFM, SREDFM and mREBEL.
Retrieved from *https://doi.org/10.48550/arXiv.2306.09802*

Figure from last year's presentation

What is Fine-Tuning ?



Objectives

- Improve the quality of the model for a specific task / domain
- Modify the role of a model (VGG + classifier for example)