Learning to Rank Context for Named Entity Recognition Using a Synthetic Dataset

Arthur Amalvy, Vincent Labatut and Richard Dufour

November 10, 2023

Long Range Context in NER

- Transformers can't process large documents at once
  - Some information is lost at prediction time!
- Long range context retrieval is beneficial for NER performance in novels [ALD23]
- How to train a NER context retriever when no data is available?

Generating a Synthetic Dataset

Using an LLM, we generate a dataset where each example has the form:

(input sentence, context sentence, relevance)

Generating Positive Examples

- We empirically determined which types of context sentences were useful
- We designed prompts for each of these

Entity type	Useful contexts
PER	description, action
LOC	description, movement towards
ORG	description

Generating Positive Examples

Input sentence

*"**One-Eye**'s handicap in no way impairs his marvelous insight"*

Positive generated context sentence

*"**One-Eye** is a wise and mysterious character with a penchant for coming up with invaluable insights after the fact"*

Relevance: 1

Generating Negative Examples: Negative Sampling

Input Sentence

"I am afraid that I have been tempted into too great length about the ***Italian Catherine****; but in truth she has been my favourite."*

Negative sampled context sentence

"said Alice, as she swarm about, trying to find her way out."

Relevance: 0

Generating Negative Examples: Positive Example Swapping

Input sentence

"We left in pretty good time and came after nightfall to ***Klausenburgh****."*

Swapped positive example

"Forley was an adventurous and daring individual who was never afraid to take risks."

Relevance: 0

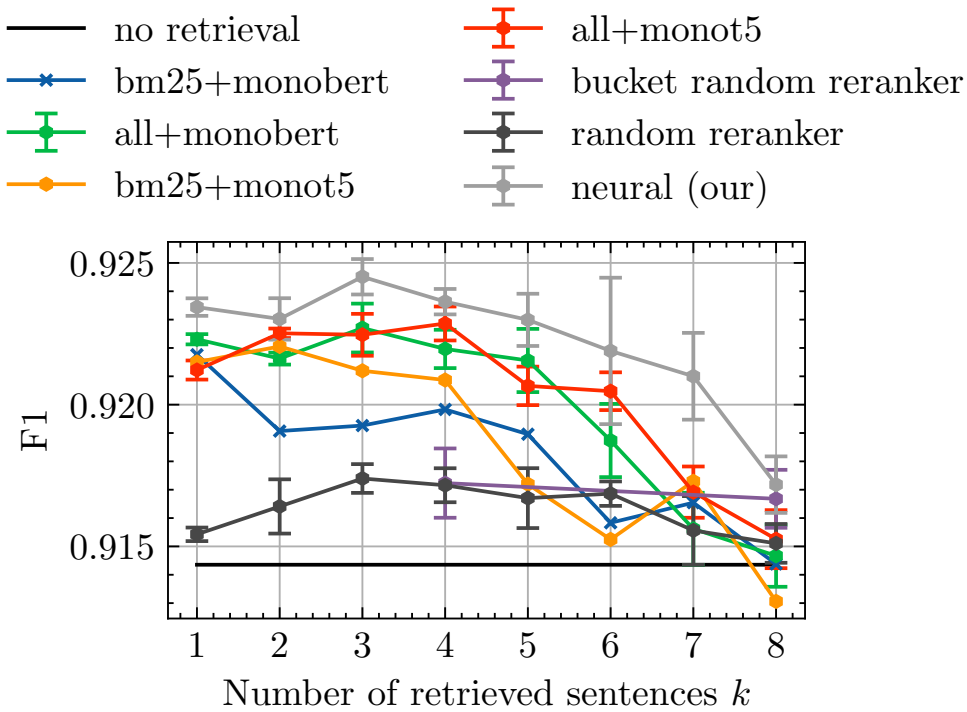
Experiments

Research questions:

- Can our re-ranker trained on synthetic data improve NER performance?
- Does the size of the LLM matters?

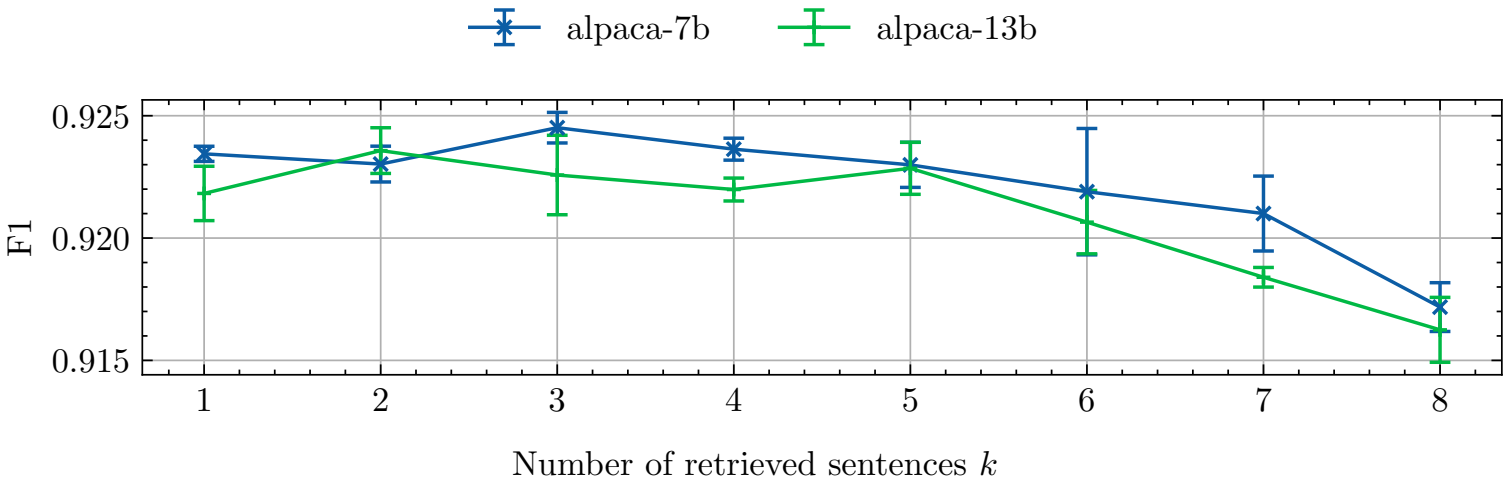
1. We train a BERT-based neural re-ranker on our synthetic context retrieval dataset (~2700 examples)
2. We train a BERT-based NER model on our NER dataset
3. At NER inference time, we retrieve context with our re-ranker

Comparison with existing supervised Re-rankers



- Our model is better than or equivalent to retrieval models trained on MSMarco

Does the size of the LLM matters?



- Quantitavely, Increasing the size of the LLM is not beneficial
- Qualitatively, no observable difference between both datasets

Conclusion

- We trained a neural retriever for NER using only generated data
- For NER, the performance of this retriever is higher than or equivalent to supervised models trained on manually annotated data
- Increasing the size of the LLM does not improve results
- More results and experiments in the article!

Références

[ALD23] A. Amalvy, V. Labatut **and** R. Dufour. **?**The Role of Global and Local Context in Named Entity Recognition**?** **in***61st Annual Meeting of the Association for Computational Linguistics*: 2023. DOI: 10.18653/v1/2023.acl-short.62.