Do Language Models Learn about Legal Entity Types during Pretraining?

Claire Barale[1], Michael Rovatsos[1], Nehal Bhuta[2]
School of Informatics[1], School of Law[2], The University of Edinburgh

NLLP 2023
EMNLP 2023

Problem definition

- During pretraining, LMs learn signals from the corpus: **what kind of signals and cues do they capture?**
  Factual and abstract? Single token and multi-token entities?
  Generic and legal vocabulary? Semantic and syntactic signals?
- Legal entity typing is a foundation: (i) for incidental supervision and text labeling, (ii) for downstream tasks, (iii) for text comprehension
- **Task:** Entity Typing (predict the entity type) zero-shot
- **Dataset:** AsyLex, a dataset of refugee decisions
- **Models:** Encoder-only vs. Decoder-only

Objective: Evaluating the quality of entity knowledge learned during pretraining in LMs, a surrogate for legal knowledge

Key takeaways

- Llama2 performs well on certain generic entity types
- Optimized prompt templates are necessary
- Law-oriented LMs show inconsistent performance
- LMs can also type multi-token entities accurately
- Entities that belong to sub-domains of the law are difficult
- Llama2 frequently overlooks syntactic cues, contrary to BERT-based architectures

Vocabulary Overlap in the Pretraining Corpora

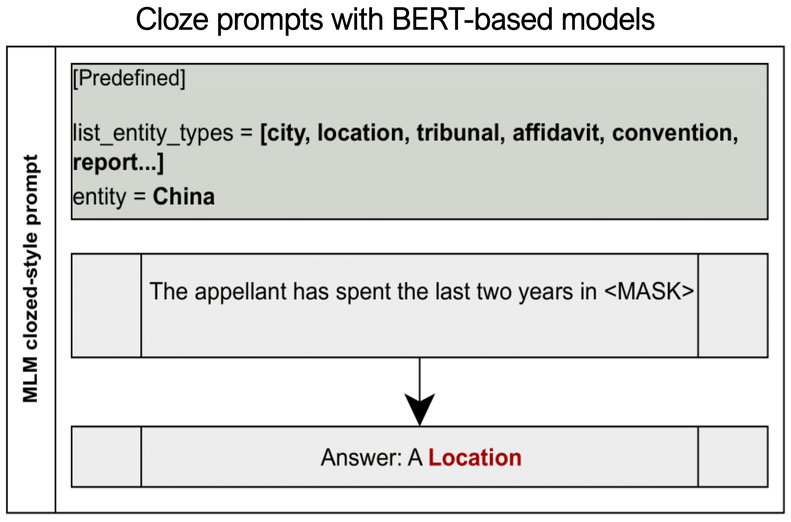
Vocabulary overlap (%) between the pretraining corpora. Gen stands for Generic and is sampled from sources similar to RoBERTa's pretraining corpus. Vocabularies are created with the top 10K most frequent tokens in a sample of 50K documents per model

Research Questions

1. How proficient are LMs at **acquiring knowledge** about **domain-specific** entities like legal entities during pretraining?
2. Can this acquired knowledge be considered sufficiently reliable for tasks such as **annotating new datasets** or serving as an indirect source of supervision?
3. How does the choice of **prompt type** impact the results obtained from knowledge queries?
4. To what extent does the variation in acquired knowledge differ across entity types? What categories of factual knowledge can LMs retrieve, and in what instances do they make errors? **Which signals are captured**, and which are missed?
5. Does **domain-specific pretraining** and jurisdiction-specific pretraining enhance the amount of factual knowledge compared to generic pretraining?

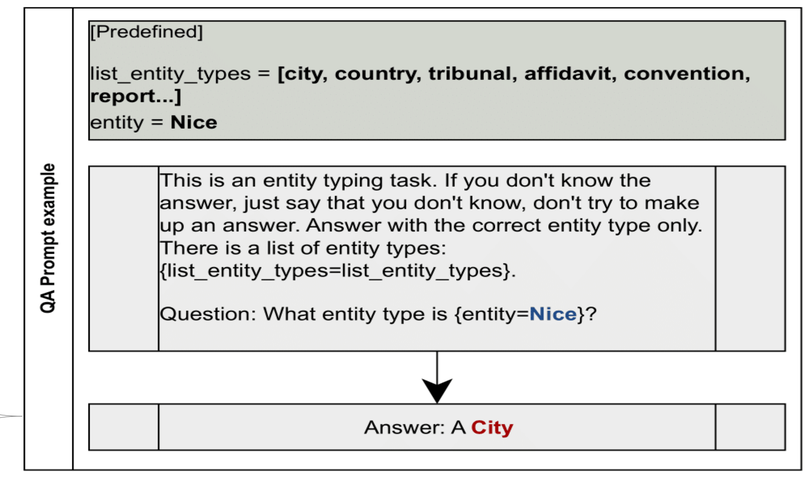
Prompt Templates and Models Used

Experiment 1: MLM
Cloze prompts with BERT-based models



Model	Corpus	# Param.
RoBERTa	Generic	125M
DeBERTa-v3	Generic	86M
CaseHOLD	Harvard Case Law	110M
Pile of Law	US, Canada, ECthR	340M
LexLM	US, Canada, UK, India	125M
Llama2-7B	Generic	7B

Experiment 2: Llama2
QA prompts



Results

Gen groups the results of RoBERTa and DeBERTa-v3, CH refers to CaseHOLD

Model	Gen	CH	PoL	LexLM	Llama2
Generic	11.59	8.51	20.42	11.97	63.26
Gen Legal	17.98	20.89	15.40	12.48	29.52
Refugee Law	10.01	5.93	4.68	4.92	13.03

Entity types prediction F1 scores averaged on 3 aggregated groups:

- Generic: location, date, norp (adjectives)
- Gen Legal - Legal entities applicable to most legal domains: org, law (citations), claimant_info, procedure, doc_evidence, law_case (precedents)
- Refugee Law - Legal entities specific to refugee law: credibility, determination, explanation, legal_ground, law_report (NGO reports)

Failure Cases Analysis

	Experiment 1	Experiment 2
Random prediction	70.71	22.22
Contextually accurate	12.43	-
Closely related	16.86	18.52
False Negative	-	33.33
Prompt error	-	25.93

Examples

under <mask> of the Republic of China, they cannot take on a second citizenship.
Predicted: lawsuit. Gold: law

What is *female claimant*? Predicted: female claimant. Gold: gender (claimant information)

Scan here for paper