UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition

ELYAZIJI BOUTAINA

Dataset	BERT-base	InstructUIE-11B	UniNER-7B
ACE05	**87.30**	79.94	86.69
AnatEM	85.82	88.52	**88.65**
bc2gm	80.90	80.69	**82.42**
bc4chemd	86.72	87.62	**89.21**
bc5cdr	85.28	89.02	**89.34**
Broad Twitter	58.61	80.27	**81.25**
CoNLL03	92.40	91.53	**93.30**
FabNER	64.20	78.38	**81.87**
FindVehicle	87.13	87.56	**98.30**
GENIA	73.3	75.71	**77.54**
HarveyNER	**82.26**	74.69	74.21
MIT Movie	88.78	89.58	**90.17**
MIT Restaurant	81.02	**82.59**	82.35
MultiNERD	91.25	90.26	**93.73**
ncbi	80.20	86.21	**86.96**
OntoNotes	**91.11**	88.64	89.91
PolyglotNER	**75.65**	53.31	65.67
TweetNER7	56.49	**65.95**	65.77
WikiANN	70.60	64.47	**84.91**
wikiNeural	82.78	88.27	**93.28**
Avg	80.09	81.16	**84.78**

Table 3: F_1 on 20 datasets used in Wang et al. (2023a). BERT-base results are from Wang et al. (2023a). InstructUIE results are from our reevaluation.

Data Construction Prompt

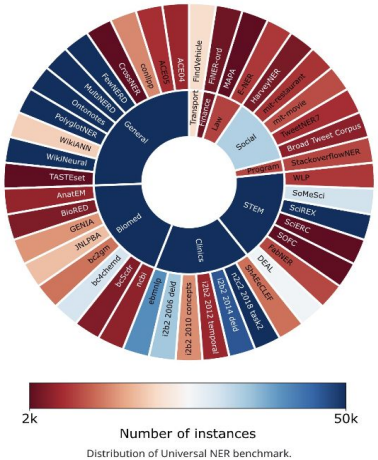
System Message: You are a helpful information extraction system.
Prompt: Given a passage, your task is to extract all entities and identify their entity types. The output should be in a list of tuples of the following format: [("entity 1", "type of entity 1"), ...].

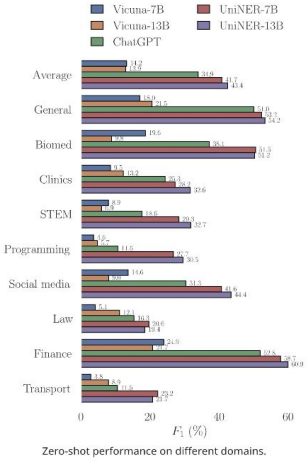
Passage: {input_passage}

Figure 1: Data construction prompt we use to generate entity mentions and their types for a given passage.

Frequency	Entity types
Top 1% (74%)	person, organization, location, date, concept, product, event, technology, group, medical condition, ...
1%-10% (19%)	characteristic, research, county, module, unit, feature, cell, package, anatomical structure, equipment, ...
10%-100% (7%)	attribute value, pokemon, immune response, physiology, animals, cell feature, FAC, input device, ward, broadcast, ...

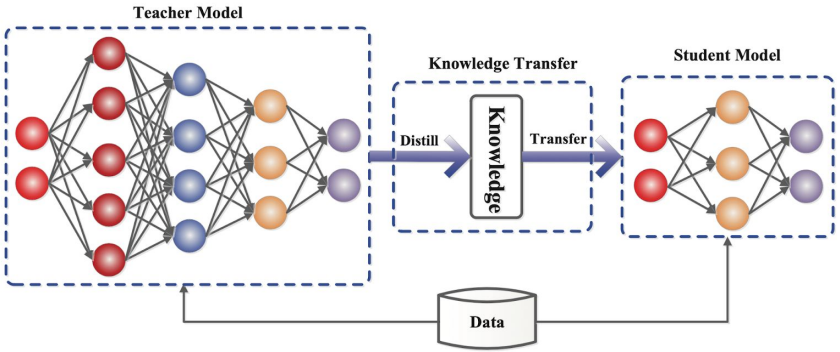
Table 1: Examples of entities across different frequency ranges - top 1%, 1-10%, and 10-100%, along with the percentage of total frequencies for each range.





Objectif of the paper

- This paper explores a method known as **targeted distillation with mission-focused instruction tuning** to create smaller student models with minimal parameters that learn from a larger teacher model to excel in specific applications like open information extraction.
- Case Study : The broad application was in NER (**Named entity recognition) task**, which is one of the common Natural Language Processing (NLP) jobs capable of extracting information from a text given specific entities..



Creation of dataset

While earlier work " **Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing(Jung et al., 2023)** "employs language models to generate inputs,to achieve that researchers often **assume that they know the specific domains of the test data.** This approach involves **prompting the language model to generate data for each known domain.** However, this method becomes inadequate when dealing with distillation tasks across diverse domains w**here the distribution of test data is not well-defined.**

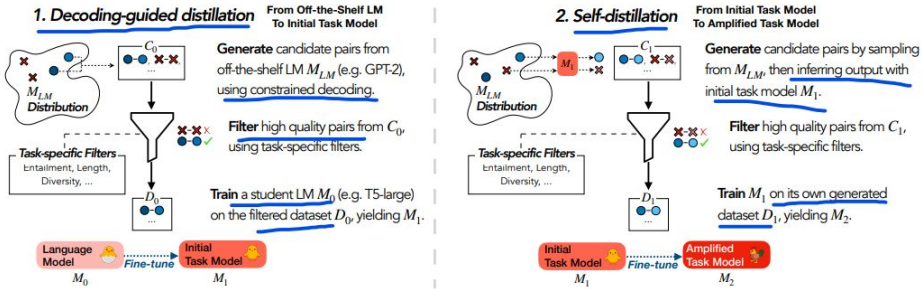


Figure 1: Overview of IMPOSSIBLE DISTILLATION. Starting from a small, off-the-shelf LM, we gradually produce higher-quality dataset and task model, **outperforming even the 200 times larger GPT-3 in both summarization and paraphrasing.**

Paraphrase (News Domain)	Paraphrase (Biomedical Domain)
Sentence *x*: At issue is a change in work rules that the company says will help reduce a massive surplus of processed steel.	**Sentence *x***: It is likely that the evidence from other settings, such as those in which birth size was decided by fetal ultrasound, will yield similar estimates.
Paraphrase *y*: The dispute is over a proposed change to the company's working conditions that the company says will help it reduce the amount of surplus steel.	**Paraphrase *y***: The findings should be expected to be generalizable to other settings, including those in which birth size is determined by fetal ultrasound.
Summary (Reddit Domain)	**Summary (Biomedical Domain)**
Sentence *x*: I've mentioned this to a few other people, and it seems that everyone else thinks this is completely weird, I don't know why.	**Sentence *x***: Additionally, the in vivo assays using P. berghei infected mice can be used as an alternative to screen more potent compounds for treating malaria.
Summary *y*: I've been telling people about it and they all think it's a weird thing to do.	**Summary *y***: The in vivo studies can be used as a platform to screen novel antimalarial compounds for use in malaria therapy.

Table 1: Samples in DIMSUM+. **All input-output pairs are generated by ∼1.6B LMs, without human supervision.** IMPOSSIBLE DISTILLATION distills a task-specific dataset and model from off-the-shelf LMs across domains, without scale or supervision. More examples in Appendix E.

Creation of dataset

- **To address this limitation ,they propose an alternative:** Directly sampling inputs from a large corpus across diverse domains, and then using an LLM to generate outputs.
- The authors sample inputs from the **Pile corpus**, which is a compilation of 22 distinct English subdatasets. They process the articles in the Pile corpus by chunking them into passages, each with a **maximum length of 256 tokens**. From these passages, **they randomly sample 50,000 passages to use as inputs** for their experiments.
- Following the sampling of inputs, the authors utilize **ChatGPT** (specifically, the **gpt-3.5-turbo-0301** variant of the model) to generate entity mentions and their associated types as a **list of tuples** based on the Data Construction Prompt. they set the generation temperature to 0 , to generate deterministic outputs.

Frequency	Entity types
Top 1% (74%)	person, organization, location, date, concept, product, event, technology, group, medical condition, ...
1%-10% (19%)	characteristic, research, county, module, unit, feature, cell, package, anatomical structure, equipment, ...
10%-100% (7%)	attribute value, pokemon, immune response, physiology, animals, cell feature, FAC, input device, ward, broadcast, ...

Table 1: Examples of entities across different frequency ranges - top 1%, 1-10%, and 10-100%, along with the percentage of total frequencies for each range.

Data Construction Prompt

System Message: You are a helpful information extraction system.
Prompt: Given a passage, your task is to extract all entities and identify their entity types. The output should be in a list of tuples of the following format: [("entity 1", "type of entity 1"), ...].

Passage: {input_passage}

Improvements employed

- The authors adopted a **conversation-style tuning format :** Taking an input (text), for example, we could ask, “What describes the PERSON in this text?” **Then it generates a JSON list representation of all the entities in the text. This method proved more effective than the traditional NER-style tuning.**
- **Supervised fine tunning :** As they are training with multiple datasets, significant challenge arises as there might be discrepancies in label definitions among these datasets, **resulting in label conflicts**. **To address this issue,** we propose to use **datasets specific instruction tuning** templates to provides the model with context that helps it understand and interpret labels within the context of each dataset.

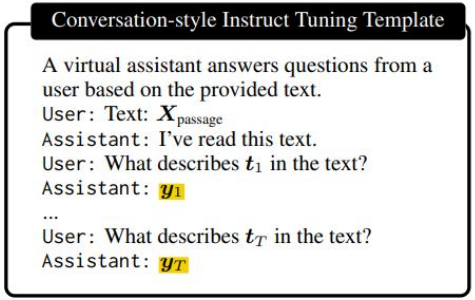


Figure 2: The conversation-style template that converts a passage with NER annotations into a conversation, where $X_{passage}$ is the input passage, $[t_1, ..., t_T]$ are entity types to consider, and y_i is a list of entity mentions that are t_i. The conversation is used to tune language models. Only the highlighted parts are used to compute the loss.

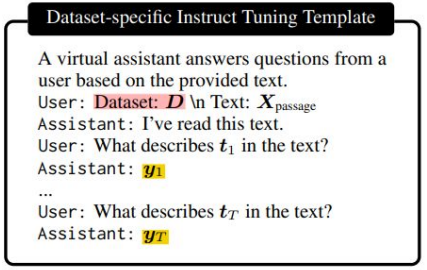
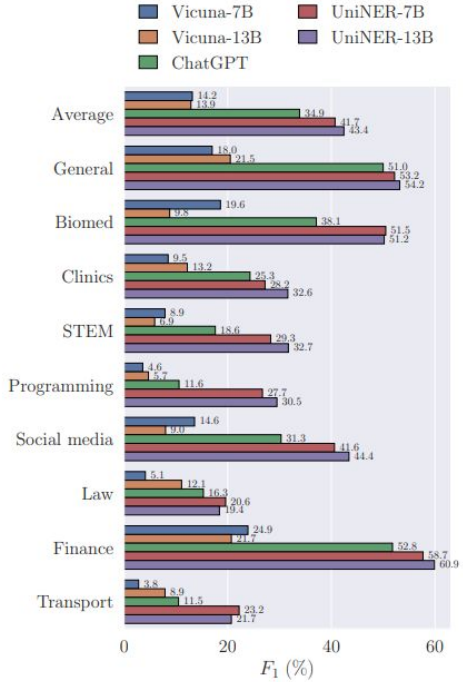


Figure 3: The dataset-specific instruction tuning template. We add the dataset name D (colored in red) as part of the input to resolve conflicts in label definitions.

Improvements employed

- **UniNER's 7B and 13B versions outperform ChatGPT** in terms of average F1 scores, with scores of 41.7% and 43.4% compared to ChatGPT's 34.9%. **This indicates that the targeted distillation from diverse inputs yields better performance across various domains.**
- **UniNER-13B exhibits better performance compared to UniNER-7B**, **indicating that fine-tuning on larger models may lead to improved generalization.**
- **Negative sampling** was also used, which is a strategy that includes introducing data that is not associated with the goal of the model to avoid overfitting and improve the model's performance. When the UniNER does not recognize an entity, it provides an empty JSON list, similar to this []. **The frequency-based sampling emerging as the most effective method for enhancing model performance in the study.**



(a) Comparisons of zero-shot models on different domains. Our distilled models achieve better results than ChatGPT in all evaluated domains.

Identified weaknesses

- **UniNER-definition :** they employed ChatGPT not just to identify but also to define entities in brief sentences. **This method made the model more adaptable to entity type paraphrasing** but **it underperformed in traditional NER benchmarks.**
- **UniNER-all-in-one :** instead of focusing on one entity type per query, they also experimented with combining all entity types into a single query, prompting the model to provide a consolidated output.
- **The 2 Variants of UniNER result in lower performance**
- **UniNER-7B-type sometimes fails to recognize entities with similar semantic meanings.**

Sensitivity Analysis of Entity Types

Text: I'm visiting Los Angeles next week.

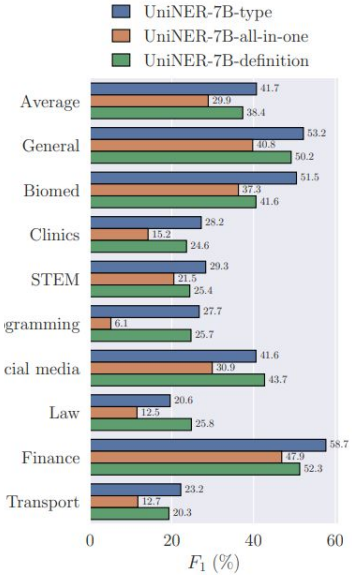
```
User: What describes city in the text?
UniNER-7B-type: ["Los Angeles"]
UniNER-7B-definition: ["Los Angeles"]

User: What describes place in the text?
UniNER-7B-type: []
UniNER-7B-definition: ["Los Angeles"]

User: What describes metropolis in the text?
UniNER-7B-type: []
UniNER-7B-definition: ["Los Angeles"]

User: What describes urban area in the text?
UniNER-7B-type: []
UniNER-7B-definition: ["Los Angeles"]

User: What describes human settlement in the text?
UniNER-7B-type: []
UniNER-7B-definition: ["Los Angeles"]
```



(b) Comparisons between UniNER-7B and two variants. UniNER-7B-definition is distilled on Pile data prompted with entity type definitions. UniNER-7B-all-in-one is tuned with the template where all entity types are asked in one query.

My Proposed solutions

1. Use advanced embeddings that capture deeper contextual meanings, enabling better differentiation between similar entities.

2. Hybrid Modeling Approach where we Combine the paraphrasing capability of UniNER-definition with with the precision of a more traditional NER approach through rule-based filtering.

3. Break down the all-in-one query into smaller, more manageable segments that are processed sequentially or in parallel. After initial entity recognition, use a consolidation step to merge the results into a single output.

4. Implement a feedback loop where the model's predictions are periodically reviewed and corrected, allowing continuous learning and adaptation to new entities and contexts.

Thank you !