MetaIE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks

Anonymous ACL submission

Abstract

Information extraction (IE) is a fundamental area in natural language processing where prompting large language models (LLMs), even with in-context examples, cannot defeat efficient small LMs tuned on very small IE datasets. We observe that while LLMs are not designed for user-specified information types, they have a decent sense of *important information*, i.e., the meta-understanding of IE. Therefore, we propose a novel framework MetaIE to build a small LM as a meta-model by learning to extract "important information", such that this meta-model can be effectively and efficiently adapted to all kinds of (few-shot) IE tasks. Specifically, we obtain the small LM via a symbolic distillation from an LLM. We construct the distillation dataset via sampling sentences from language model pre-training datasets and prompting an LLM to identify the typed spans of "important information". Extensive results on 13 datasets from 6 IE tasks confirm that MetaIE can offer a better starting point for few-shot adaptation and outperform other strong meta-models, including a multi-task model built upon multiple large IE benchmark training sets. Moreover, we provide comprehensive analyses of MetaIE, such as the size of the distillation dataset, the meta-model architecture, and the size of the meta-model.

1 Introduction

Large language models (LLMs), such as ChatGPT (OpenAI, 2023), benefit from the vast amount of training data and have demonstrated exceptional performance across various areas through in-context learning (ICL) (Dong et al., 2023). However, when it comes to information extraction (IE), LLMs, even with ICL examples, struggle to compete with smaller LMs, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), fine-tuned on very small training sets (Peng et al., 2023c; Wadhwa et al., 2023; Gao et al., 2024). This is usually regarded as a limitation of LLMs in following a specific extraction scheme (Xu et al., 2023).

We observe that while LLMs are not designed for user-specified information types, they have a decent sense of *important information*, i.e., the meta-understanding of IE. We formulate the IE tasks as a label-to-span matching and decompose the task as several label-specific instructions, such as "*Given an IE label* (l)*, extract a span from the input text*" as shown in Figure 1. Here l can be, for example, (1) *Person*, *Location*, *Organization* in named entity recognition (NER), or (2) *Tom births at* in relation extraction (RE) to verify if there is a certain relation between two entities by checking the other entity can be recognized or not. Following these label-to-span instructions, LLMs can handle all kinds of IE tasks and return imperfect yet semantically reasonable answers.

In this paper, we propose a novel framework MetaIE to build a small LM as a meta-model by learning to extract "important information", so this meta-model can be effective and efficient for all kinds of (few-shot) IE tasks. In this work, we aim at 6 diverse and popular types of IE tasks, namely, Named Entity Recognition, Relation Extraction, Event Extraction, Semantic Role Labeling, Aspect-based Sentiment Analysis, and Aspect Sentiment Triplet Extraction, which we believe represents the IE tasks well enough. Pioneer works have built meta-models for specific type of IE tasks, e.g., UniversalNER (Zhou et al., 2023) explores the potential of building a meta-model for different NER tasks. Our work is more ambitious at a larger scope of IE tasks.

MetaIE obtains the small LM via a symbolic distillation (West et al., 2022) from an LLM following the label-to-span scheme. We construct the distillation dataset via sampling sentences from language model pre-training datasets and prompting an LLM to identify the typed spans of "important information". In particular, we implement this idea

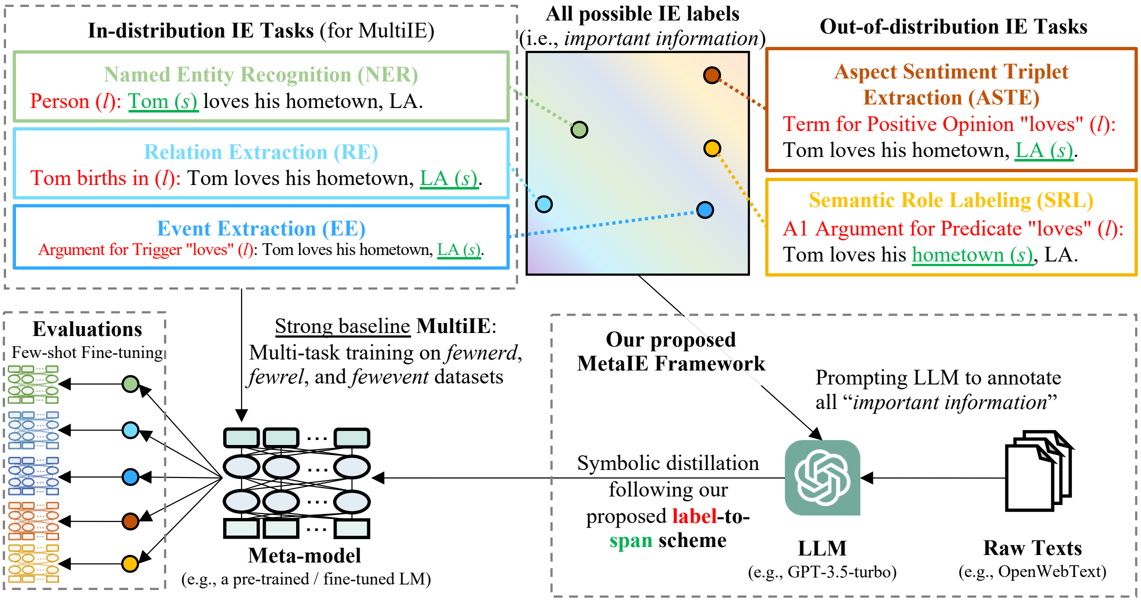

In-distribution IE Tasks (for MultiIE)
Named Entity Recognition (NER)
Person (l): Tom (s) loves his hometown, LA.
Relation Extraction (RE)
Tom births in (l): Tom loves his hometown, LA (s).
Event Extraction (EE)
Argument for Trigger "loves" (l): Tom loves his hometown, LA (s).
All possible IE labels
(i.e., important information)
Out-of-distribution IE Tasks
Aspect Sentiment Triplet Extraction (ASTE)
Term for Positive Opinion "loves" (l): Tom loves his hometown, LA (s).
Semantic Role Labeling (SRL)
A1 Argument for Predicate "loves" (l): Tom loves his hometown (s), LA.
Evaluations
Few-shot Fine-tuning
Strong baseline MultiIE: Multi-task training on fewnerd, fewrel, and fewevent datasets
Our proposed MetaIE Framework
Prompting LLM to annotate all "important information"
Symbolic distillation following our proposed label-to-span scheme
Meta-model
(e.g., a pre-trained / fine-tuned LM)
LLM
(e.g., GPT-3.5-turbo)
Raw Texts
(e.g., OpenWebText)

Figure 1: An overview of different transfer learning schemes involved in the experiments.

with 100,000 sentences from the OpenWebText corpus (Gokaslan and Cohen, 2019), which contains various webpage texts and is also a subset of the popular language model pre-training dataset. We feed these sentences to GPT-3.5-turbo for identifying "important information", which is then used to distill small LMs. It is worth mentioning that MetaIE is applicable to all types of small LMs and one only needs to convert the label-span pairs following the corresponding labeling scheme (e.g., BIO sequence labeling for encoders like RoBERTa, seq2seq labeling for encoder-decoders like BART).

Our evaluation focuses on the few-shot learning ability of the meta-model for different IE tasks. We mainly compare MetaIE with other meta-models, including (1) pre-trained language model, (2) a multi-task model built upon multiple large IE benchmark training sets, and (3) a task-specific model obtained via symbolic distillation from LLM for a single IE task. Large-scale training datasets for NER, RE, and event extraction (EE) tasks are used in single-IE-task and multi-IE-task pre-training, therefore, these datasets shall be considered as *in-task-distributional* for these two types of methods. For a more comprehensive evaluation, we further include *out-of-task-distributional datasets* from (1) semantic role labeling (SRL) (Carreras and Màrquez, 2005), (2) aspect-based sentiment analysis (ABSA) (Pontiki et al., 2014), and (3) aspect-sentiment triplet extraction (ASTE) (Xu et al., 2020), totaling 13 datasets across 6 IE tasks. In our experiments, MetaIE generally achieves the best performance, only *very occasionally* losing to task-specific distillation on some in-task-distributional datasets. This demonstrates that MetaIE is a strong and efficient method to distill the meta-understanding of IE from LLMs into small LMs. Remarkably, distilling from the LLM-produced dataset following the traditional human annotation schemes performs poorly. Therefore, the success of MetaIE, rather than from purely using LLMs, shall also come from our label-to-span scheme.

We have further conducted comprehensive analyses of MetaIE. We study the scaling-up rules to investigate the model and dataset size boundaries in obtaining the meta-understanding of IE. We showcase the diversity of the types of "important information" in the MetaIE distillation dataset. We show that the RoBERTa with sequence labeling framework is the best meta-model architecture compared with sequence-to-sequence and decoder-only models, at a similar scale.

Our contributions are three-fold:

- We are the first to build a small LM as a meta-model for all kinds of IE tasks.
- We propose a novel label-to-span scheme that unifies all IE tasks and applies symbolic distillation to distill the meta-understanding from an LLM to a small LM.
- We have a rigorous experiment design, which covers various IE tasks and meta-model methods. Comprehensive experiment results support the intuitive expectation and advantage of our MetaIE.

2 Related Works

2.1 Information Extraction

Information extraction (IE) is one of the most popular and vital domains in natural language processing. Early IE systems are generally developed for a single IE dataset like NER (dos Santos and Guimarães, 2015), RE (Katiyar and Cardie, 2016), or EE (Chen et al., 2015). Due to the gap between the label sets and annotation styles of different IE datasets, few-shot IE frameworks (Ding et al., 2021; Han et al., 2018; Ma et al., 2023) are proposed to quickly learn models on new datasets. The IE models are pre-trained on a large scale of IE labels and then transferred to the target domain by fine-tuning on few examples. With the emergence of LLMs, researchers have started to train LMs on multiple IE tasks with unified formats (Lu et al., 2022; Paolini et al., 2021). LLMs fine-tuned for general purpose (OpenAI, 2023; Touvron et al., 2023) have also shown strong potential to understand new IE tasks with their instruction-following ability. However, these LLMs still lag behind supervised models (Xu et al., 2023), potentially due to the difficulty of specifying the required pattern for extraction in different datasets. Moreover, the cost of LLMs limits their application to IE on a large corpus. This paper aims to transfer the meta-understanding of IE from LLMs to lighter-weight models, which produce a flexible model with high adaptability to any target IE task.

2.2 Model Distillation

Model distillation (Hinton et al., 2015; Gou et al., 2021) is the process of transferring knowledge from large models (teacher models) to small ones (student models). Traditional distillation optimizes the similarity between logits produced by the teacher and student models (Hinton et al., 2015; Kim et al., 2019; Mirzadeh et al., 2020). Symbolic distillation (West et al., 2022; Li et al., 2023; West et al., 2023) for language models learns a student model on texts generated by the teacher model. In comparison with traditional distillation, symbolic distillation allows the student model to focus on one aspect of the teacher model (West et al., 2022), which can be some high-level ability, such as chain-of-thought reasoning (Li et al., 2023), with much smaller model size. For IE, symbolic model distillation has been successfully applied for an IE subtask, NER (Zhou et al., 2023), which distills an NER model that can extract entities in a broad domain. This paper aims to distill the cross-IE task ability of LLMs, i.e., meta-understanding of IE and proposes a meta-model that can effectively learn IE tasks with few examples.

2.3 Meta Learning

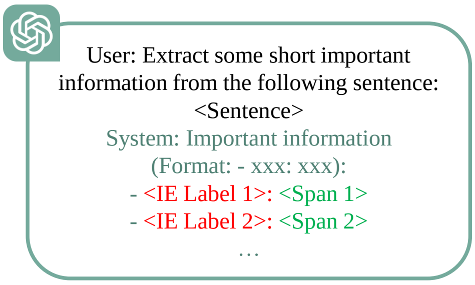
Meta-learning (Finn et al., 2017a) enables the models to learn new tasks better, i.e., stronger transfer learning ability. MAML (Finn et al., 2017b) proposes a framework to learn a better starting point for few-shot learning by utilizing multiple datasets for loss updating. Reptile (Nichol et al., 2018), similar to MAML, simplifies the meta-learning algorithm by performing stochastic gradient descent not only within each task but also across tasks, making it more efficient and easier to implement. The Prototypical Networks method (Snell et al., 2017) employs a distance-based classification approach, where it learns a metric space in which classification can be performed by computing distances to prototype representations of each class. While most meta-learning methods are experimented on classification tasks, pre-training on multiple datasets (Ding et al., 2021) and prototypical networks (Ji et al., 2022) have been applied for IE. While these methods focus on specific IE tasks like NER, we aim to optimize a starting point for general IE tasks by distilling from LLMs.

3 Our MetaIE Framework

3.1 Label-to-span Scheme

We formalize the IE task as given an IE label l (e.g., *Person* in NER), extracting a span s from a sentence $X = [x_1, \cdots, x_n]$. The span s can be represented as $x_{i:j}$ including the words from i-th to j-th. Denoting the IE process as a mapping $f_{IE}(\cdot)$, it can be represented as $s = f_{IE}(X|l)$. Machine learning-based methods aim to learn the mapping by optimizing a model M_θ with parameter θ. For a specific IE task (e.g., NER), the IE label set $\mathcal{L}^{(Task)}$ will contain l falling inside the task label, i.e., $(l \in \mathcal{L}^{(Task)})$. Based on the general definition of IE, the general IE label set $\mathcal{L}^{(IE)}$ can be any textual description, thus $\forall$Task, $\mathcal{L}^{(Task)} \subset \mathcal{L}^{(IE)}$.

In this paper, we aim to learn a meta-model that can be easily adapted to different IE tasks. In the current practice of IE, the "meta-model" is generally pre-trained in a single IE task with a large number of labels ($\mathcal{L}^{(pt)} \subset \mathcal{L}^{(Task)}$). Then, the meta-model can be fine-tuned on few-shot examples to quickly adapt to different downstream IE datasets



User: Extract some short important information from the following sentence: <Sentence>

System: Important information (Format: - xxx: xxx):

- <IE Label 1>: <Span 1>
- <IE Label 2>: <Span 2>

…

Figure 2: The prompt used in our experiments to build the dataset for symbolic distillation.

in the same task, such that $\mathcal{L}^{(ft)} \subset \mathcal{L}^{(Task)}$. We expand this learning scheme to a general meta-model that works for all existing and potentially new IE tasks. To achieve this goal, our intuition is to pretrain the model to learn the label-to-span mapping with the label set approximating the general IE label distribution $\mathcal{L}^{(pt)} \sim \mathcal{L}^{(IE)}$. As the label sets of all IE tasks are subsets of $\mathcal{L}^{(IE)}$, our meta-model will enjoy an efficient transfer to all IE tasks.

3.2 Distillation Dataset Construction

To apply a symbolic distillation of the meta-understanding of IE from LLMs, we prompt LLMs to create data for distillation by querying them to extract "important information" from texts as shown in Figure 2. Our expectation for the dataset is to cover as many l as possible to approximate the broad $\mathcal{L}^{(IE)}$ set to better distill the meta-model for all kinds of IE tasks. We query LLMs to annotate some raw corpora $\mathcal{X}$ to build the MetaIE dataset. Given each $X \in \mathcal{X}$, the LLM is instructed to generate a series of (l, s) pairs. We do not set any limitation to l to better approximate the broad $\mathcal{L}^{(IE)}$ set.

Implementation We select the paragraphs from OpenWebText (Gokaslan and Cohen, 2019), Since OpenWebText it is a popular dataset used in language model pre-training, we are not introducing new texts. We split the paragraphs by sentences and only use the first sentence of each paragraph for a higher diversity and to avoid the ambiguity caused by coreference. The LLM is instructed to formalize all (l, s) pairs in the prompting output as "- Place (l): New York (s)", which are extracted by regular expression matching. Considering there might be multiple spans returned for l, we split the span by conjunctions like comma.

Table 1 shows some statistics and example results of the labels returned by the LLM, illustrating a broad spectrum of IE domains, ranging from simple entities and events to complex relationships and contexts. The diversity in the n-gram categories showcases the model's ability to capture a wide array of query types. This variety underscores the comprehensive coverage and nuanced understanding that LLMs bring to the task of generating queries across different facets of the IE domain.

3.3 Distillation Framework

We illustrate the distillation with a sequence labeling model (dos Santos and Guimarães, 2015) that suits well for encoder-based language models (e.g., RoBERTa (Liu et al., 2019)). Given a sequence of words $X = [x_1, \cdots, x_n]$, the sequence labeling model will tag each word by outputting $Y = [y_1, \cdots, y_n]$. Following the traditional BIO labeling scheme, y_i will be B (begin), I (inner), and O (none). The model is trained on word tagging and the tags are decoded into spans by searching sequences that begin with B and continue by I. In traditional sequence labeling models, the B and I tags generally consist of label information such as B-place or I-person. In our case, we formalize the tagging in a query-dependent way since the model needs to handle arbitrary queries. We attach the label information as a prefix like "place: " to the beginning of the input text. The input text is then labeled by the BIO scheme, where the span label is indicated in the prefix. Finally, the BIO sequences are used to fine-tune the sequence labeling models. This distillation process can also be adapted to Seq2Seq encoder-decoder models and Causal LM-based decoder-only models. We use sequence labeling models for the main experiment based on their empirical advantage in IE tasks, which we also empirically find support in the analysis in Section 5.1.

4 Experiments

4.1 IE Tasks and Datasets

To deeply delve into the differences between different model distillation or meta-learning methods, we include a wide variety of tasks:

1. Named Entity Recognition (**NER**) extracts named entities with their labels from texts. We include 6 NER datasets that was studied in (Ushio and Camacho-Collados, 2021), i.e., (1) **CoNLL2003**, (2) **BioNLP2004**, (3) **WNUT2017**, (4) **MIT-Movie**, (5) **MIT-Restaurant**, (6) **BC5CDR**, which covers various domains: news, medical, social media, and

n-gram (Count)	Example IE Labels (Relative Frequency)
1-gram (270k)	Location (7.73%), Event (4.67%), Action (4.24%), Topic (3.57%), Subject (3.25%), Person (2.71%), Date (2.70%), Source (2.44%)
2-gram (44.5k)	Target audience (1.27%), Time period (0.998%), Individuals involved (0.992%), Action taken (0.877%), Political affiliation (0.762%), Parties involved (0.758%), Release date (0.697%), TV show (0.686%)
3-gram (16.9k)	Source of information (2.02%), Cause of death (1.17%), Call to action (1.02%), Date of birth (0.739%), Date and time (0.727%), Date of death (0.562%), Type of content (0.337%), Reason for arrest (0.325%)
4-gram (7.39k)	Purpose of the bill (0.325%), Location of the incident (0.271%), Name of the person (0.271%), Number of people killed (0.203%), Number of people affected (0.189%), Content of the bill (0.162%), Number of people arrested (0.149%), Source of the information (0.149%)
$\geq$ 5-gram (5.37k)	Dates of birth and death (0.13%), Age at the time of death (0.112%), Total number of votes cast (0.0931%), Feature: Auschwitz through the Lens of the SS (0.0931%), Number of people on board (0.0745%), Name of the person involved (0.0745%), Date and time of publication (0.0745%), Action taken by President Obama (0.0745%)

Table 1: Example IE Labels, Counts, and Relative Frequency in our constructed symbolic distillation dataset, grouped by the number of tokens.

reviews.

2. Relation Extraction (**RE**) extracts named entities, and in addition, identifies the relationships between them. We include 2 popular datasets, (1) **ADE** (Gurulingappa et al., 2012) and (2) **CoNLL2004** (Carreras and Màrquez, 2004) representing RE on medical and news domain. We evaluate the performance of RE models on both relation detection and the detection of entities involved in the relations.
3. Event Extraction (**EE**) extracts event triggers and their arguments. We use the standard **ACE2005** dataset (Walker et al., 2006) preprocessed by OMNIEVENT (Peng et al., 2023a) for EE evaluation. We compare the model performance on both event detection (ED) and event argument extraction (EAE) tasks, following the consistent evaluation framework proposed by Peng et al. (2023b).
4. Semantic Role Labeling (**SRL**) extracts predicates (verbs) and their arguments. We select the **CoNLL2005** (Carreras and Màrquez, 2005) dataset for SRL. We follow previous works to learn backbone LMs on samples from the Brown training dataset and then test them on Brown and WSJ test datasets.
5. Aspect-based Sentiment Analysis (**ABSA**) extracts aspect terms and the sentiment polarity towards them. We select **SemEval2014** (Pontiki et al., 2014) as the dataset for ABSA, with its two subsets: **14res** and **14lap** including reviews about restaurants and laptops.
6. Aspect Sentiment Triplet Extraction (**ASTE**) extracts aspect terms and the corresponding opinion terms that contain the sentiment polarity towards them. We use the same **SemEval2014** dataset as for ABSA, on which aspect-sentiment triplets are further annotated by Xu et al. (2020).

For a fair comparison, we formalize all those tasks as $s = f_{IE}(X|l)$, which can be found in the Appendix A. For each task, we query each possible label to extract (l, s) pairs. For spans conflicting with each other, as we run label-wise extractions, we only keep the one with a higher BI sequence probability. For tasks that extractions are dependent on each other (e.g., RE, EE, SRL, ASTE), we follow (Paolini et al., 2021) to run multi-stage extractions for these tasks. As ACE2005 involves too many labels, we report the unlabeled performance on detecting the triggers and arguments for all methods for comparison.

4.2 Evaluation Metric: Few-shot Fine-tuning Performance

We use the few-shot fine-tuning performance on all IE tasks to evaluate the meta-model's quality. Specifically, all methods in our evaluation will provide us a backbone LM. We then conduct few-shot fine-tuning from the training dataset for fine-tuning with sample details in Appendix B. Finally, we evaluate them on the test dataset using the micro F1 score as the evaluation metric. For multi-task pretraining baselines, tasks without large-scale annotations (SRL, ABSA, ASTE) are **out-of-distribution**

Method Category	Method	NER					
		ConLL2003	BioNLP2004	WNUT2017	MIT-Movie	MIT-Restaurant	BC5CDR
LLM Prompting	ICL	59.68	48.08	36.51	46.08	60.62	59.82
FT	Vanilla	32.58	36.06	33.87	57.65	63.40	18.15
Task-level ML+FT	Transfer						
	Human	71.61	54.58	43.15	**64.80**	69.17	72.02
	LLM	67.74	45.62	45.36	59.59	69.19	73.14
	Task Distillation	**74.86**	**56.18**	**50.09**	**65.70**	**71.48**	71.01
IE-level ML+FT	MultiIE	63.94	52.47	44.29	58.43	69.38	71.20
	MAML	66.97	53.09	46.14	60.57	68.86	72.58
	MetaIE	71.49	**55.76**	44.33	**65.64**	**71.33**	**75.21**

Method Category	Method	RE (NER)		RE		ED	EAE
		ADE	CoNLL2004	ADE	CoNLL2004	ACE2005	ACE2005
LLM Prompting	ICL	63.55	58.47	39.02	31.34	60.47	28.79
FT	Vanilla	25.97	62.13	15.67	33.52	67.46	32.86
Task-level ML+FT	Transfer						
	Human	41.56	**69.27**	20.53	37.51	**72.79**	35.77
	LLM	35.43	66.93	14.35	35.07	65.17	34.86
	Task Distillation	66.99	68.66	**41.92**	41.58	67.34	34.56
	NER Distillation	67.35	**69.88**	32.73	35.68	66.17	32.86
IE-level ML+FT	MultiIE	53.26	**69.14**	18.23	39.65	71.16	35.23
	MAML	56.95	**69.28**	38.65	42.07	68.22	35.84
	MetaIE	**69.29**	**69.47**	**40.43**	**43.50**	69.85	**36.83**

Table 2: Few-shot transferring performance (F1 score) of different meta-learning sources **(IE-level+FT)** on IE tasks. Other methods are included for reference: **1) LLM Prompting**: Performance of the *large* teacher model; **2) Task-Level ML+FL:** Performance of meta-learning that only focuses on the target IE task. **Bold:** Performance of the *small LM* that is not significantly different from the best one. ($p < 0.05$).

tasks.

The default backbone LM we used for fine-tuning is `RoBERTa-Large` (Liu et al., 2019), which is a traditional bidirectional encoder used for learning IE tasks formalized as sequence tagging. The learning rate is set to 2×10^{-5} with AdamW (Loshchilov and Hutter, 2019) as the optimizer and a cosine annealing learning rate scheduler (Loshchilov and Hutter, 2017). We fine-tune the backbone LM with batch size 64 for a single epoch to avoid overfitting.

4.3 Compared Methods

We first include a comparison with the teacher model **GPT-3.5-turbo** via **LLM Prompting** with in-context learning (**ICL**). For ICL, we provide 5 examples in the prompt of our query. Based on previous discoveries on LLM-based IE (Peng et al., 2023c; Wadhwa et al., 2023; Gao et al., 2024), we shall expect that fine-tuned small LMs work better than the LLM.

We compare our **MetaIE** with a variety of methods from the following three categories

1. **Vanilla** LM fine-tuning (**FT**), i.e., directly using the vanilla pre-trained LM as the backbone LM in fine-tuning.
2. **Task-level** Meta-learning (**ML**)+**FT**. It is expected to have a strong performance to other datasets in the same IE task but poor generalization to other IE tasks.
   - **Transfer (Human)** is a baseline that trains the backbone LM on large-scale human annotations of a specific IE task. Specifically, we use *FewNerd* (Ding et al., 2021) for NER, *FewRels* (Han et al., 2018) for RE, and *FewEvents* (Ma et al., 2023) for EE.
   - **Transfer (LLM)** uses the same datasets in **Transfer (Human)** but queries the LLM to annotate them following the human workflow. This baseline aims to compare the quality of annotation from humans and LLMs following the conventional annotation schema.
   - **Task Distillation** distills from LLMs by querying answers for specific IE tasks. We implement this by providing in-context task-specific examples to control the LLM-produced data similar to the label IE task. The input texts are set to be the same as MetaIE to avoid bias.
   - **NER Distillation** applies the model distilled following **Task Distillation** but tests them on non-NER tasks to evaluate its cross-task transferability.
3. **IE-level** Meta-learning (**ML**)+**FT** aims to learn

Method Category	Method	SRL		ABSA		ASTE	
		Brown	WSJ	14RES	14LAP	14RES	14LAP
LLM Prompting	ICL	28.79	31.56	53.04	35.62	58.94	44.87
FT	Vanilla	52.59	56.47	24.46	10.32	39.17	41.50
Task-level ML+FT	NER Distillation	43.65	51.29	10.77	11.21	40.06	38.40
IE-level ML+FT	MultiIE	52.26	56.63	38.22	35.28	24.91	40.49
	MAML	52.69	56.23	40.22	34.45	30.83	40.95
	MetaIE	**54.50**	**58.49**	**50.96**	**39.71**	**43.30**	**43.10**

Table 3: Few-shot transferring performance (F1 score) of different meta-learning sources **(IE-level+FT)** on IE tasks *without* the large scale of human annotations for task-level meta-learning.

an IE model with strong transferability to all IE tasks. Our **MetaIE** also falls into this category.

- **MultiIE** merges the multiple human-annotated IE datasets (*FewNerd*, *FewRels*, *FewEvents*) to train a backbone LM, which represents a multi-task baseline with human annotations.
- **MAML** (Finn et al., 2017b) is a traditional meta-learning baseline that merges gradients on different datasets to build a model that can be quickly transferred to these datasets. We use the datasets in **MultiIE** for **MAML** in the experiment.

For all baselines, the data number for meta-learning is controlled to the same as MetaIE by sampling towards a fair comparison. The main method category for comparison is **IE-level ML+FT**, which focuses on meta-learning for all IE tasks. **LLM-Prompting** is experimented with to show the performance of the large teacher model as a reference. **Task-level ML+FT** is also used as a reference to showcase the ability of task-specific ML, as it utilizes different meta-learning sources for different IE tasks.

4.4 Result Discussion

The result from our experiments is presented in Tables 2 and 3. The vanilla model is poorly transferred by fine-tuning to all kinds of IE tasks. The model with meta-learning on a single IE task, NER, is only well-transferred to other NER datasets but poorly-transferred to other IE tasks. Among IE-level meta-learning methods, the MultiIE model can be transferred to in-domain IE tasks with outstanding performance but still fails to be transferred to out-of-domain IE tasks, either with regular pre-training or meta-learning frameworks like MAML. In contrast to all these baselines, our MetaIE shows a strong transferability to all IE tasks, especially on out-of-domain tasks for MultiIE. Thus, the experiment results are highly consistent with our claim in IE task transferability that wider pre-training label set $\mathcal{L}^{(IE)}$ will enable macro transferability of the model to all IE tasks.

Besides the main discovery, we can also observe that LLM-based meta-learning outperforms the pre-training on human annotation. Take NER as an instance, while both label sets satisfy $\mathcal{L} \subset \mathcal{L}^{(NER)}$, the $\mathcal{L}$ proposed by LLMs is much more diverse than the fixed set in human annotated datasets, which again verifies the importance of the label distribution, even in task-specific distillation.

The comparison with the teacher model also shows the student model generally outperforming the teacher model under few-shot supervision. Thus, we conclude fine-tuning a distilled student IE model to perform better than inference by the teacher LLMs with few-shot in-context examples. This further verifies the advantage of model distillation in meta-learning which enables more efficient and effective transfer.

5 Further Analysis

5.1 Distillation Framework Comparison

We compare student models following different distillation frameworks (because of their architectures) to investigate how this factor affects the distillation effectiveness.

Seq2Seq implements the distillation by learning to extract a group of spans based on the IE label as in the distillation dataset. We include two Seq2Seq models: `BART-Large` (Lewis et al., 2020) and `T5-Base` (Raffel et al., 2020), which contain the same scale of parameters as in the `RoBERTa-Large` in our previous experiments.

CausalLM is similar to **Seq2Seq** but only uses the decoder model instead of the encoder-decoder as in **Seq2Seq**. We also include two CausalLM-based models with similar parameter scales: `GPT2-Medium` (Brown et al., 2020) and `OPT-350M` (Zhang et al., 2022).

Framework	Model	ConLL2003	BioNLP2004	WNUT2017	MIT-Movie	MIT-Restaurant	BC5CDR
Seq-Labeling	BERT	63.01	52.39	32.71	61.75	62.50	66.24
	RoBERTa	**71.49**	**54.88**	44.33	**65.64**	**71.33**	**75.21**
Seq2Seq	BART	**71.39**	47.18	**46.74**	62.76	67.98	65.90
	T5	64.01	42.35	40.74	55.05	53.60	38.67
CausalLM	GPT	57.20	37.29	36.89	52.14	60.46	61.03
	OPT	52.39	37.64	34.48	53.07	53.59	52.86

Table 4: Comparison between different frameworks on MetaIE distillation.

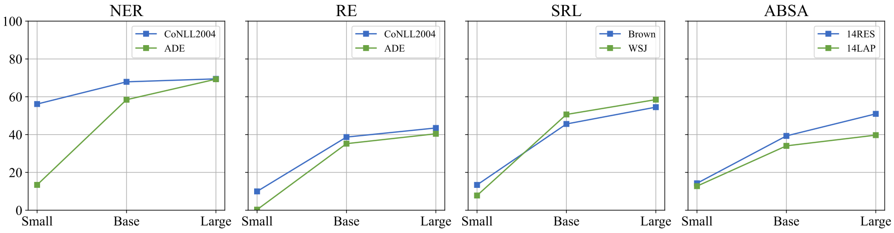


Figure 3: The size analysis of the student model scale on different IE tasks and domains.

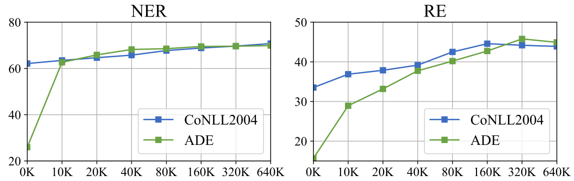


Figure 4: The size analysis of the distillation data scale on different IE tasks and domains.

We also include another sequence labeling model `BERT-Large-Cased` (Devlin et al., 2019) as a baseline to explore the influence of the backbone model quality on the learning performance. For all models, we pre-train them using our MetaIE dataset with the same hyperparameters.

We compare the performance of different distillation frameworks on NER as an example and the result is demonstrated in Table 4. Sequence labeling models perform the best in few-shot transfer learning, which indicates their advantage in the distillation of meta-understanding of IE. This can be attributed to the consistency of sequence labeling with the extraction nature. We thus conclude distilling IE knowledge to a traditional sequence labeling model is better than those popular generative models. Between sequence labeling models, RoBERTa outperforms BERT, showing a better student model also benefits the distillation procedure.

5.2 Size Analysis

We explore how size is a factor affecting the distillation quality. For the model scale, we compare among `RoBERTa-Small`, `RoBERTa-Base`, and `RoBERTa-Large`. For the data scale, we increase the sampling size to $640K$ and pre-train the student model with different amounts of data.

The analysis of **model size** is presented in Figure 3, we can observe the performance of a student model can be scaled up by more parameters. Also, for simple tasks (like NER) with a general domain (like CoNLL2004), a tiny student model is competent for the distillation. However, for specific domains or complex tasks, the student model needs more parameters for generalization.

The analysis of **data size** is presented in Figure 4, we observe the existence of a threshold between $80K \sim 160K$ to endow the student model with the meta-understanding of IE. Also, a small amount of metadata $(10K)$ significantly benefits the transfer.

6 Conclusions and Future Work

This paper presents a novel approach for distilling the meta-understanding of IE from LLMs into more efficient, smaller language models through a synthesized dataset, MetaIE. Our findings indicate that this method not only enhances the adaptability and efficiency of smaller models but also outperforms existing single-task and multi-task distillation methods in various IE tasks. The success of MetaIE underscores the potential of leveraging LLM's meta-understanding to improve the performance and versatility of smaller models in complex tasks, offering a promising direction for future research in model distillation and IE. Future work will explore a better way for meta-learning by distilling from LLMs and other meta-tasks can be trained based on distillation.

Limitation

Efficiency The efficiency of the unified label-to-span will be $O(|\mathcal{L}^{(Task)}|)$, which is lower than the traditional $O(1)$ (number of LM forwarding) BIO sequence labeler with label information in the labeling result. This will limit the application of our model to cases where $|\mathcal{L}^{(Task)}|$ is large. This efficiency is a trade-off for the ability to process any IE label, which enables the fast transfer of the BIO model to different IE tasks.

Bias in LLM-proposed labels As pointed out in previous works (Gallegos et al., 2023; Fang et al., 2023), LLMs have biases in their responses. This can also be observed in the statistics of our distillation dataset. Thus, the small meta-model might also inherit the bias and have better transferability to labels that LLMs prefer than others.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning, CoNLL 2004, Held in cooperation with HLT-NAACL 2004, Boston, Massachusetts, USA, May 6-7, 2004*, pages 89–97. ACL.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005, Ann Arbor, Michigan, USA, June 29-30, 2005*, pages 152–164. ACL.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 167–176. The Association for Computer Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3198–3213. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Cícero Nogueira dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop, NEWS@ACL 2015, Beijing, China, July 31, 2015*, pages 25–33. Association for Computational Linguistics.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: An examination of news produced by large language models. *CoRR*, abs/2309.09825.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017a. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017b. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *CoRR*, abs/2309.00770.

Jun Gao, Huan Zhao, Wei Wang, Changlong Yu, and Ruifeng Xu. 2024. Eventrl: Enhancing event extraction with outcome supervision for large language models. *CoRR*, abs/2402.11430.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Informatics*, 45(5):885–892.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4803–4809. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Bin Ji, Shasha Li, Shaoduo Gan, Jie Yu, Jun Ma, Huijun Liu, and Jing Yang. 2022. Few-shot named entity recognition with entity-level prototypical network enhanced by dispersedly distributed prototypes. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1842–1854. International Committee on Computational Linguistics.

Arzoo Katiyar and Claire Cardie. 2016. Investigating lstms for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Jangho Kim, Yash Bhalgat, Jinwon Lee, Chirag Patel, and Nojun Kwak. 2019. QKD: quantization-aware knowledge distillation. *CoRR*, abs/1911.12491.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2665–2679. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5755–5772. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Yixin Cao, and Aixin Sun. 2023. Few-shot event detection: An empirical study and a unified view. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11211–11236. Association for Computational Linguistics.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5191–5198. AAAI Press.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023a. OmniEvent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517, Singapore. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023b. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9206–9227.

Letian Peng, Zihan Wang, and Jingbo Shang. 2023c. Less than one-shot: Named entity recognition via extremely weak supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13603–13616. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Web Download. LDC Catalog No. LDC2006T06.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. 2023. Novacomet: Open commonsense foundation models with symbolic knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1127–1149. Association for Computational Linguistics.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *CoRR*, abs/2312.17617.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2339–2349. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *CoRR*, abs/2308.03279.

A Label-to-Span Formalization

NER

Person: John/B Smith/I loves/O his/O hometown/O ,/O Los/O Angeles/O

RE

Person: John/B Smith/I loves/O his/O hometown/O ,/O Los/O Angeles/O

John Smith births in: John/O Smith/O loves/O his/O hometown/O ,/O Los/B Angeles/I

EE

Trigger: John/O Smith/O loves/B his/O hometown/O ,/O Los/O Angeles/O

Argument for Trigger "loves": John/O Smith/O loves/O his/O hometown/O ,/O Los/B Angeles/I

SRL

Verb: John/O Smith/O loves/B his/O hometown/O ,/O Los/O Angeles/O

A1 Argument for Verb "loves": John/O Smith/O loves/O his/O hometown/B ,/O Los/O Angeles/O

ABSA

Positive Term: John/O Smith/O loves/O his/O hometown/O ,/O Los/B Angeles/I

ASTE

Positive Opinion: John/O Smith/O loves/B his/O hometown/O ,/O Los/O Angeles/O

Aspect for Opinion "loves": John/O Smith/O loves/O his/O hometown/O ,/O Los/B Angeles/I

B Few-shot Details

NER samples 5-shot examples that contain a certain type of entity for each entity type.

RE samples 5-shot examples that contain a certain type of relation for each relation type.

EE samples 5% examples from the original training dataset.

SRL samples 50-shot examples from the original training dataset.

ABSA samples 5-shot examples that contain terms with a certain sentiment polarity for each sentiment polarity type.

ASTE samples 5-shot examples that contain aspect-opinion triplet with a certain sentiment polarity for each sentiment polarity type.