



INSTRUCTIE: A Bilingual Instruction-based Information Extraction Dataset

Honghao Gui^{1,2}, Shuofei Qiao^{1,2}, Jintian Zhang^{1,2}, Hongbin Ye¹, Mengshu Sun^{2,3},
Lei Liang^{2,3}, Jeff Z. Pan⁴, Huajun Chen^{1,2}, and Ningyu Zhang^{1,2}(✉)

¹ Zhejiang University, Hangzhou, China

² Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph, Hangzhou, China

³ Ant Group, Hangzhou, China

⁴ University of Edinburgh, Edinburgh, UK

{guihonghao, shuofei, huajunsir, zhangningyu}@zju.edu.cn

Abstract. Large language models can perform well on general natural language tasks, but their effectiveness is still suboptimal for information extraction (IE). Recent works indicate that the main reason lies in the lack of extensive data on IE instructions. Note that the existing datasets on IE instructions not only have limited coverage but also involve high construction costs. To address this issue, we introduce INSTRUCTIE, a bilingual instruction-based IE dataset, which covers 12 diverse domains. We propose KG2Instruction, a framework specifically for the automatic generation of such datasets. Additionally, we manually annotate the test set. Experimental results demonstrate that large language models trained with INSTRUCTIE can not only obtain better IE capabilities but also enhance zero-shot performance compared with baselines.

Resource Type: New Dataset

Source Repo: <https://huggingface.co/datasets/zjunlp/InstructIE>

DOI: <https://doi.org/10.5281/zenodo.10970777>

License: Attribution-NonCommercial-ShareAlike 4.0 International

Keywords: Knowledge Graph · Knowledge Graph Construction · Information Extraction · Dataset · Large Language Models

1 Introduction

Information Extraction (IE) aims to extract structured data from text sources which can boast extensive applications across various fields such as Knowledge Graph (KG) construction, and question-answering systems [28]. IE tasks are highly diverse due to their varying objects (entities, relations, events, etc.), heterogeneous structures, and demand-specific patterns. Traditional approaches [15, 20, 35, 57, 58] design specific architectures for different IE tasks. Generative IE [22, 23, 25, 32] unifies various IE tasks into a sequence-to-sequence text generation framework. Although these methods have held promising capabilities in the past, a notable inherent limitation is their constraint to pre-defined labels along with the once-and-for-all training pattern as illustrated in Fig. 1 (a). Such inflexibility significantly hampers their adaptability, especially in the ever-evolving real world where more scalable solutions are demanded.

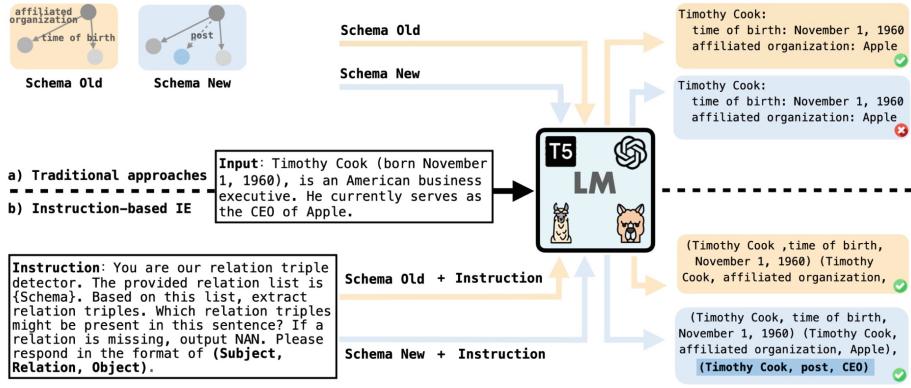


Fig. 1. Comparison of traditional information extraction (IE) approaches with Instruction-based IE in handling emergent classes (unseen during training). Dashed lines represent the addition of a new class (e.g., `post`). Traditional approaches often struggle to accommodate the evolving demands of user extraction requirements. In contrast, Instruction-based IE demonstrates the capability to comprehend instructions, discern changes in requirements, and effectively extract newly added classes.

With the emerging development of Large Language Models (LLMs) [2, 31, 40], it is possible to achieve generalized Instruction-based information extraction (IE) capabilities. For example, as shown in Fig. 1 (b), the IE system should be capable of interpreting natural language instructions and producing the expected responses accordingly [13]. Recently, some studies [47, 49, 52] make performance gains in low-resource settings by designing prompt-based frameworks, e.g. leveraging models like ChatGPT for in-context learning. Other works like GoLLIE [34], InstructUIE [48], and [6, 51] are proposed which are trained with IE-based instruction data. Despite previous advancements, recent researches [5, 11, 13, 17, 18, 27, 43, 44, 46, 53] indicate that the effectiveness of LLMs for IE is still suboptimal, mainly due to the limited availability of datasets with comprehensive IE instructions. These existing datasets not only have restricted coverage but also entail high construction costs.

To address this issue, we develop a framework called KG2Instruction for automatically generating information extraction instruction datasets across different domains. KG2Instruction first generates relationship triples by aligning the knowledge graph (KG) with existing corpora. Subsequently, it addresses the incompleteness of the KG by supplementing missing triples using an existing information extraction (IE) model incrementally trained on a small amount of manually annotated data. Finally, a natural language inference model is used to filter out unreal triples. Furthermore, using KG2Instruction, we construct a bilingual IE instruction dataset named INSTRUCTIE, which covers 12 domains and 123 types of relationships, containing 364,074 instances. Additionally, we manually annotated 2,000 instances to serve as the test set. We evaluate various large language models (LLMs) on INSTRUCTIE under multiple settings, such as zero-shot learning, in-context learning, and fine-tuning. Empirically, LLMs fine-tuned with INSTRUCTIE can not only enhance their performance in instruction-based IE

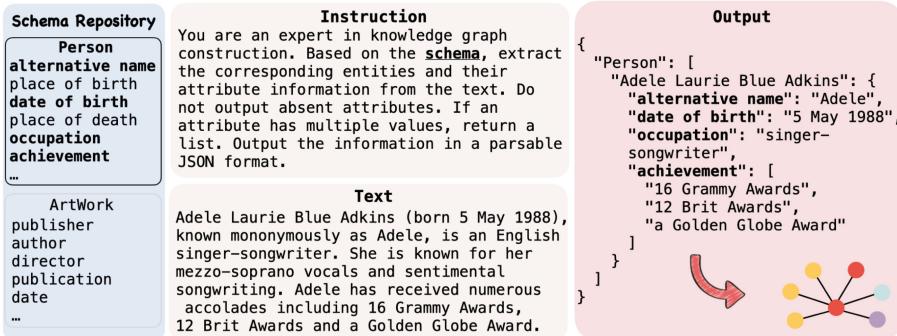


Fig. 2. Examples of instructions and their outputs for knowledge graph construction, with the Schema Repository containing labels under various domains.

tasks but also show certain advantages in generalizing to other domains. INSTRUCTIE is accessible as Linked Data at <https://w3id.org/instructie/> and available on Zenodo¹ under CC BY-SA 4.0 license.

The **main contributions** of this research can be summarized as:

1. We introduce a framework KG2Instruction, specifically designed for the automatic generation of IE instruction datasets across various domains.
2. Based on KG2Instruction, we successfully construct an bilingual IE instruction dataset named INSTRUCTIE, which encompasses 12 distinct domains. Furthermore, we manually annotate the test set.
3. We conduct a comprehensive evaluation of Instruction-based IE task on INSTRUCTIE, emphasizing the strengths and weaknesses of baseline, and the models' ability to generalize to unseen domains

2 Instruction-Based IE

We frame Instruction-based Information Extraction (IE) as an instruction-following auto-regressive generation task. The model first needs to understand the instructions to identify its intent, and then, based on the content of the instructions, extracts relevant information from the input text and outputs them in a specified format. Specifically, the instructions consist of two main parts: 1) **Task Description**: It specifies the task that the model is expected to perform, such as Named Entity Recognition, Relation Extraction, and Event Extraction; 2) **Schema**: A list of labels (entity types, relations, etc.) to be extracted, reflecting the user's requirements, which is dynamic and changeable.

To better adapt to the task of KG construction, we design specialized instruction templates. As shown in Fig. 2, the model's input includes two parts: the instructions and the text input. The instructions clarify the task to be performed by the model, namely the extraction of entities and their attributes, and also indicate the required schema to

¹ <https://zenodo.org/records/10970777>.

Table 1. A categorization of 12 textual domains, meticulously curated to ensure expansive coverage of real extraction requirements across diverse fields.

 Astronomy	 Transportation	 Building	 Creature	 Science	 Event
 Medicine	 Organization	 Person	 Artworks	 Product	 GPE

be extracted, as demonstrated in the Schema Repository. This may involve the entity type “Person” and related properties such as “alternative name”, “place of birth”, and “occupation”. The output displays the results extracted by the model, arranged in the order of entity type, entity, and attributes.

3 Construction of INSTRUCTIE

The traditional process of constructing IE datasets typically involves domain experts selecting relevant corpus and guiding data engineers in data collection and manual annotation. This process is not only costly and time-consuming but also inefficient. To address this issue, we introduce KG2Instruction, a framework aimed at automating the generation of such datasets. Through KG2Instruction, we construct a bilingual IE instruction dataset named INSTRUCTIE. In this section, we will detail the framework of the KG2Instruction as well as the construction of the INSTRUCTIE dataset.

3.1 Data Source and Preparation

Our data primarily originates from two platforms: Wikidata² and Wikipedia³. Initially, we examine both Chinese and English Wikipedia documents, selecting paragraphs with a token length between 50 and 512. Subsequently, we manually annotate 5,000 Chinese and English text paragraphs for classification. Based on these datasets, we train Chinese and English text classifiers on the chinese-roberta-wwm-ext-large⁴ and roberta⁵ models, respectively. In a manual evaluation of 1,000 data points, these text classifier achieves a 92% F1 score. In total, we define 12 text domains and conceive a specialized schema template for each domain. Table 1 enumerates our classification outcomes.

3.2 KG2Instruction

KG2Instruction automatically generates relational triples through a three-step process: 1) aligning KG with existing corpora, 2) supplementing missing triples with a trained IE model, and 3) filtering out unreal triples using a natural language inference model. Figure 3 provides an exhaustive visualization of the entire procedure.

² <https://www.wikidata.org>.

³ <https://www.wikipedia.org/>.

⁴ <https://huggingface.co/hfl/chinese-roberta-wwm-ext-large>.

⁵ <https://huggingface.co/FacebookAI/roberta-large>.

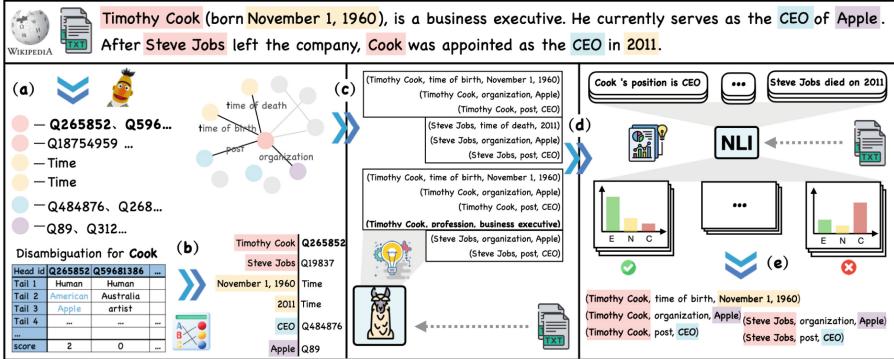


Fig. 3. Overview of INSTRUCTIE dataset construction. (a) Identify Entity Mentions. (b) Disambiguation. (c) Schema Constraint Matching. (d) Missing Triplets Supplement with LLM. (e) Hallucinatory Triplets Filtering with NLI.

Identify Entity Mentions. Firstly, we identify all the human-provided links between Wikipedia articles to create an initial entity mentions set. Although these links provide gold entity annotations, they only annotate the first appearance of entities, leaving subsequent mentions unannotated. To bridge this gap, we employ a NER model [9] to capture as many of the remaining entity mentions as possible. Next, we utilize the entity mention to query Wikidata and retrieve all associated IDs (unique identifiers in Wikidata) associated with the entity mentions.

Disambiguation. Within Wikidata, an identical entity mention can map to different IDs. For instance, “Apple” might refer to either a corporate entity (Q312) or a fruit (Q89). To mitigate such ambiguities, we instate an intuitive disambiguation strategy. Assume that all entity mentions in a text paragraph are defined as the set \mathcal{M} . For each entity mention m in \mathcal{M} , we calculate a score for all possible corresponding IDs. The score is based on the frequency of mentions of the tail entity corresponding to that ID appearing in the \mathcal{M} . For each m , we select the ID with the highest score as its unique corresponding Wikidata ID. Additionally, we iteratively query the “**instance of**” property of each entity to assign a type to every entity. As illustrated in Fig. 4, we identify 14 categories of entity types.

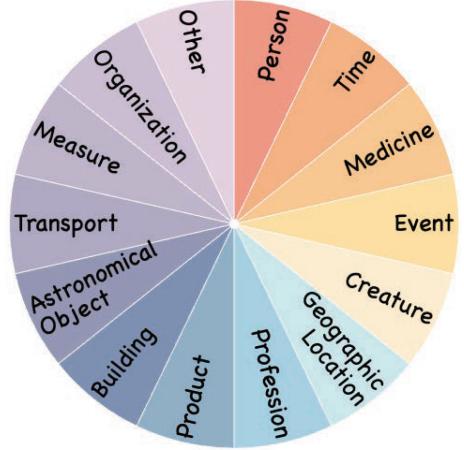


Fig. 4. Classification of 14 entity types, aiming at covering a diverse array of entities with distinct boundaries.

Schema Constraint Matching. The assumption held by traditional methods [29] is that “If two entities participate in a relation, any sentence that contains those two entities might express that relation.”. However, under the domain schema setting, such an assumption may lead to the generation of a large number of unnecessary triples. For example, in the sentence “Qiqi Technology has strongholds in both China and Japan.”, although China and Japan have the diplomatic relation in Wikidata, this relation is not the main topic of the sentence. To address this issue, we introduce schema constraints as a filtering mechanism to optimize the matching process. Specifically, we define a series of **schema mappers** for the 12 domains. Each mapper under a domain contains some relations that are most relevant to that domain, and specifies the entity type constraints that the head and tail entities of each relation must satisfy. For example, the domain of “Person” includes relation such as “date of birth”, where the head entity must be a person, and the tail must be time. When iterating over all potential pairs of entities in the entity set \mathcal{M} and all their corresponding relations, we only include those triples in the final results whose relations, as well as their head and tail entities, meet the constraints.

Missing Triplets Supplement with LLM. However, we notice that triplets generated solely from KG often suffer from missing issues, primarily due to the inherent incompleteness of the KG [38]. To address this issue, we propose the utilization of a LLM to complete those missing triplets. Specifically, we select 50 samples from each domain for manual annotation, then use these annotated samples for incremental training of an existing information extraction LLM⁶ proposed by [6] and following its training and prediction methods. This model has already undergone extensive fine-tuning with general IE instructions, so we expect that after incremental training on our limited domain-specific data, it will effectively handle domain-specific texts. We provide the model with instructions, the domain-specific schema mapper, and the input text. The model returns the relevant triples present in the text. These predicted triples can effectively supplement those missing from the KG generation, such as (Timothy Cook, profession, business executive) shown in Fig. 3. Finally, the triples predicted by the LLMs are merged with and deduplicated against those generated by the KG.

Hallucinatory Triplets Filtering with NLI. Another issue is that triples generated by KG or LLMs may either correspond to the original text or not. For example, in Fig. 3, the KG-generated triple (Steve Jobs, time of death, 2011) does not match the sentence, as it does not confirm that “Steve Jobs died in 2011”, and therefore should be removed. Following previous work [35, 42], we apply a Natural Language Inference model⁷, to filter out triples that are not entailed by the sentences. The specific steps are as follows: First, we utilize ChatGPT to generate 3 templates for each relation, for example, a template for “date of death” could be “[X] died on [Y]”. Then, with each source sentence as the premise, we transform the triplet into 3 hypotheses with the templates and compute the entailment probability scores for the corresponding premise-hypothesis pairs. We select the highest score as the final entailment score for each triplet. We set a threshold

⁶ <https://huggingface.co/zjunlp/baichuan2-13b-iepile-lora>.

⁷ <https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7>.

Table 2. Distribution of instances in INSTRUCTIE by domains. The label **ZH** denotes Chinese entries, and **EN** indicates English entries. The term $\mathbb{E}[Triples]$ refers to the average number of triples associated with each instance. Conversely, $\mathbb{E}[Tokens]$ signifies the average token count per instance.

Domain	ZH			EN		
	#Instance	$\mathbb{E}[Tokens]$	$\mathbb{E}[Triples]$	#Instance	$\mathbb{E}[Tokens]$	$\mathbb{E}[Triples]$
GPE	20,200	131.64	5.6	20,176	81.24	4.12
Event	19,201	194.8	4.97	20,185	117.23	5.33
Person	20,200	267.37	11.96	20,201	119.78	9.23
Science	4,508	192.98	2.47	8,765	98.19	1.78
Product	10,000	222.47	2.26	9,969	109.43	2.21
Creature	10,200	113.15	7.5	10,103	97.95	5.76
Building	16,727	173.17	5.51	20,181	102.56	4.56
Artworks	20,200	201.99	7.52	20,100	128.72	7.07
Medicine	3,444	258.29	4.59	6,676	153.58	3.98
Transport	20,200	106.34	5.53	20,165	80.67	5.01
Astronomy	10,200	107.58	3.82	11,846	107.65	3.17
Organization	18,590	176.43	4.5	20,039	113.78	4.71

of 0.5, and only triplets with scores above this threshold are retained. Through this filtering mechanism, we exclude approximately 15% of triplets from the final annotated dataset, thereby enhancing the quality and reliability of the dataset.

3.3 Data Sampling

While the KG2Instruction framework has yielded a rich set of annotated data, to foster diversity in the dataset and further optimize data balance, we employ a schema-centric sampling strategy to select a subset from the full data. Specifically, we conduct a statistical analysis of the schema combinations for each sample, wherein the likelihood of a sample being chosen decreased as the frequency of its schema combination increased within the chosen samples. Finally, we construct the INSTRUCTIE dataset, which comprises 173,670 Chinese instances and 188,406 English instances. Details on the data distribution and composition are presented in Table 2. Additionally, we also automatically construct entity data by KG2Instruction, but our evaluation is more focused on the relationships between entities. Therefore, the entity data is merely a byproduct.

3.4 Crowdsourcing Annotation

We carefully select around 1,000 Chinese instances for annotation. A dedicated team of 20 specialists is tasked to annotate these instances in two iterative rounds, striving to adhere to predefined guidelines. Before embarking on the actual annotation, each annotator partakes in thorough training to ascertain uniformity in their efforts. To ensure reliability, every instance undergoes independent evaluation by two separate annotators. When discrepancies arise between their annotations, administrators seek to achieve unanimity. Once annotated, these instances are translated into English using the GPT-4

model [30], and subsequent refinements are made to ensure the precision of translations. Collectively, these 2,000 instances constitute the **test sets** for INSTRUCTIE.

3.5 Quality Control

To evaluate the quality of the INSTRUCTIE dataset, we randomly select 500 samples from each domain for meticulous manual review. During this evaluation process, accuracy is adopted as the primary criterion for assessment. The results indicate that the average accuracy for INSTRUCTIE-ZH reaches 82%, while for INSTRUCTIE-EN, the average accuracy is 75%. We also observe that the data quality is relatively lower in certain specific domains, such as “Event”, “Medical”, and “Science”. This phenomenon can be attributed to the complexity and specialization inherent in these fields.

4 Experiments

4.1 Experimental Setup

We evaluate various LLMs and strategies within the INSTRUCTIE dataset to explore the performance of different approaches in instruction-based IE tasks.

Base Model. We compare the performance of various LLMs, including ChatGPT [31] accessible via the OpenAI API, the earlier and smaller-scale MT5-base [54], as well as the more recently released and more powerful LLaMA2 (7B/13B) [41] and Baichuan2 (7B/13B) [1]. Note that all models used in the following experiments are **Chat** versions.

Settings. Our experimental design seeks to systematically investigate the efficacy and applicability of diverse methodologies within the realm of instruction-based IE. Central to this inquiry are several strategies:

1. **Zero-shot learning:** which gauges a model’s intrinsic capability in the absence of specific IE instruction training.
2. **In-context learning:** which assesses the model’s capability to extract information by learning from contextual examples.
3. **Fine-tuning (including LoRA):** which examines the model’s actual extraction performance after thorough instruction tuning.

For fine-tuning, we employ the parameter-efficient Low-Rank Adaptation (LoRA) fine-tuning strategy. Specifically, we train the model for 3 epochs in accordance with the recommended hyperparameters from⁸. For in-context learning, we randomly select 5 samples from the training set for each domain. In the zero-shot and in-context learning settings, we exclusively compare the 13B Chat models and do not consider smaller models because smaller models typically have weaker capabilities in instruction following and context learning. **Metric:** In our assessment, we adopt span-based micro-F1,

⁸ <https://github.com/tloen/alpaca-lora>.

Table 3. Evaluation results of different models and strategies on INSTRUCTIE-ZH. We employ abbreviations to simplify the display of each domain. The specific correspondences are as follows: **PRO** (Product), **PER** (Person), **GPE** (Geopolitical Entity), **ORG** (Organization), **EVE** (Event), **BUD** (Building), **ART** (Artwork), **CRE** (Creature), **AST** (Astronomy), **MED** (Medicine), **SCI** (Science), **TRA** (Transport). **Overall** denotes the total micro F1 results, while **bold** indicates the best result. Entries marked with an † denote parameter-efficient fine-tuning using the **LoRA** approach.

	Evaluator	PRO	PER	GPE	ORG	EVE	BUD	ART	CRE	AST	MED	SCI	TRA	Overall
0-shot	ChatGPT	26.13	32.93	18.28	27.55	17.09	20.04	8.84	36.10	51.39	25.72	15.98	15.82	24.01
	Baichuan2-13B	14.59	17.79	19.53	14.52	10.45	8.13	1.57	15.45	14.57	1.30	12.54	8.72	12.11
	LLaMA2-13B	0.00	0.91	4.47	2.11	0.00	0.85	0.00	1.61	2.84	2.11	0.00	3.97	1.80
ICL	ChatGPT	31.82	25.41	12.06	31.24	21.42	36.88	33.51	67.05	21.85	11.11	16.00	56.07	37.62
	Baichuan2-13B	14.63	11.72	11.06	30.79	13.71	38.21	18.06	30.43	18.13	31.97	19.00	31.94	23.38
	LLaMA2-13B	19.63	19.06	27.11	27.97	22.84	31.79	31.39	48.03	32.86	27.62	14.04	34.87	29.63
FT	MT5-Base	64.20	63.49	57.98	72.78	53.69	79.21	66.20	75.67	82.13	41.03	38.52	79.14	68.02
	LLaMA2-7B †	63.96	57.37	63.14	71.88	54.10	77.76	58.19	83.82	86.42	48.20	46.74	81.32	68.61
	LLaMA2-13B †	70.96	58.91	64.86	73.70	53.00	79.20	57.89	85.97	87.01	48.80	46.16	81.71	69.78
	Baichuan2-7B †	57.86	63.41	62.99	71.48	50.00	75.06	71.40	89.87	86.73	56.22	39.39	77.61	70.53
	Baichuan2-13B †	63.58	68.37	61.85	74.30	56.33	76.66	63.99	91.84	87.79	59.32	45.93	78.37	72.18

Table 4. Result of different models and strategies on INSTRUCTIE-EN.

	Evaluator	PRO	PER	GPE	ORG	EVE	BUD	ART	CRE	AST	MED	SCI	TRA	Overall
0-shot	ChatGPT	14.65	35.58	16.72	20.29	4.07	16.05	10.25	36.23	54.81	15.62	14.40	18.35	21.01
	Baichuan2-13B	11.30	20.53	14.84	14.69	1.20	14.12	4.41	12.30	6.53	11.74	10.55	12.41	11.44
	LLaMA2-13B	8.38	16.51	15.49	9.35	1.34	11.30	6.35	16.31	9.34	8.86	5.65	10.45	11.06
ICL	ChatGPT	20.42	46.61	32.11	30.05	17.38	37.29	46.19	31.08	68.16	21.33	17.19	57.17	36.68
	Baichuan2-13B	17.91	26.30	31.72	9.93	5.76	44.23	20.21	13.26	24.42	14.41	18.56	36.25	24.50
	LLaMA2-13B	24.73	42.27	37.20	20.40	21.47	38.32	32.91	32.75	44.17	19.13	18.05	52.51	34.32
FT	MT5-Base	50.31	49.37	59.07	58.92	26.94	68.09	69.79	50.20	59.22	33.65	32.26	69.42	55.21
	LLaMA2-7B †	53.25	58.65	61.88	63.70	40.65	70.73	74.41	43.63	81.14	45.92	30.51	71.77	60.31
	LLaMA2-13B †	54.54	66.80	67.05	67.22	49.19	70.44	72.95	58.59	81.71	45.00	28.81	72.75	64.97
	Baichuan2-7B †	53.11	59.16	61.44	63.94	40.17	70.13	73.92	50.30	80.37	46.91	31.01	73.62	62.49
	Baichuan2-13B †	54.67	64.38	63.32	66.38	45.95	70.28	70.57	63.70	80.58	47.62	33.06	75.84	64.75

which considers a relation triple accurate only when the head entity, tail entity, and relation strings are precisely predicted. We report not only the F1 scores for each domain but also the **overall micro F1** score for the entire test set. Please note that the micro F1 score is not the average across all domains (**macro F1**).

4.2 Main Results

The empirical evaluation results for INSTRUCTIE-ZH and INSTRUCTIE-EN are respectively presented in Tables 3 and 4. These findings reveal that while existing LLMs demonstrate limited proficiency in instruction-based IE tasks, their performance can be significantly enhanced through instruction tuning on the INSTRUCTIE dataset.

Zero-Shot Learning Performance. Tables 3 and 4 (0-shot) respectively provide an assessment of current LLM under a zero-shot learning setting on INSTRUCTIE-ZH and INSTRUCTIE-EN. The results indicate that even for advanced LLM like ChatGPT, significant challenges persist in zero-shot learning, with all F1 metrics being quite low and none exceeding 60. When compared to the 13B open-source model, these challenges appear even more pronounced. We observe that a non-negligible proportion of the output produced by the models could not be parsed into structured data, rendering evaluation impossible. Therefore, we posit that the challenges of zero-shot learning lie in the models’ inadequacies in adhering to specific instructions for information extraction and in presenting the outputs in a designated format.

In-Context Learning Performance. As shown in Tables 3 and 4 (ICL), all models exhibit significant improvements on the INSTRUCTIE dataset as measured by the Overall metric under the in-context learning setting. Notably, ChatGPT achieves absolute increases of $\uparrow 13.61$ and $\uparrow 15.67$ in the Chinese and English evaluations, respectively. This reflects the model’s ability to discern the intention of instructions from contextual examples provided and to format its output accordingly. Strikingly, when introduced to contextual examples, LLaMA2-13B demonstrates a pronounced surge in extraction capabilities across both Chinese and English evaluation, with improvements of $\uparrow 27.83$ and $\uparrow 23.26$, respectively. A salient observation from our experimental endeavors is the model’s acute sensitivity to prompt templates, implying that the template we choose might be particularly compatible with LLaMA2-13B.

Fine-Tuning Performance. Although in-context learning leads to marked improvements in model performance, the overall metrics remain suboptimal. In this study, we fine-tune open-source models on INSTRUCTIE dataset to enhance their capabilities in relationship cognition and entity boundary recognition. As shown in Tables 3 and 4 (FT), the experimental results clearly reveal a significant enhancement in the performance of all open-source models across all domains. Particularly, Baichuan2-13B in Chinese tasks and LLaMA2-13B in English tasks stand out, achieving the highest scores of 72.18 and 64.97, respectively. We attribute the discrepancy between the results of the fine-tuning in Chinese and English to our English test set being translated from Chinese, thereby presenting syntactic differences from the training set. These results highlight, on the one hand, the importance of task-specific training data in instruction-based IE tasks, and on the other hand, they reflect the significant contribution that INSTRUCTIE makes in enhancing model performance.

Model Scaling. From the comparisons between different scale versions of Baichuan2 on INSTRUCTIE-ZH and LLaMA2 on INSTRUCTIE-EN, as shown in Tables 3 and 4 (FT), we observe that model size plays a positive role in enhancing performance on instruction-based IE tasks. Moreover, even among models of the same scale, the initial capacity of the base model significantly impacts the performance. Baichuan2 exhibits superior performance in Chinese tasks, while LLaMA2 demonstrates better results in processing English content. Simultaneously, we compare MT5-Base with other models

Table 5. The zero-shot generalization performance of the Baichuan2-13B and LLaMA2-13B models, each fine-tuned using LoRA on the DuIE2.0, INSTRUCTIE-ZH, NYT10, and INSTRUCTIE-EN datasets separately. This assessment is conducted across five distinct relation extraction datasets.

Model	Training	Evaluation (EN)			Evaluation (ZH)			
		FewRel	Wiki-ZSL	Avg	COAE2016	IPRE	CMeIE	Avg
Baichuan2-13B	DuIE2.0	17.24	21.58	19.41	51.21	22.41	10.09	27.90
Baichuan2-13B	INSTRUCTIE-ZH	23.67	21.66	22.67	43.75	23.02	16.72	27.83
LLaMA2-13B	NYT10	19.34	12.48	15.91	2.99	0.18	0.22	1.13
LLaMA2-13B	INSTRUCTIE-EN	22.51	24.97	23.74	16.51	0.54	0.43	5.71

and find that fine-tuning only a small fraction of the parameters in LLMs often yields better outcomes than fine-tuning all the parameters of smaller models. We speculate that this phenomenon may be due to the LoRA technology, which allows the model to learn more about the format sub-distribution patterns during user interactions rather than factual knowledge.

5 Analysis

5.1 Generalization to Unseen Schemas

In this section, we evaluate the model’s ability to handle unseen schemas after being trained on the INSTRUCTIE dataset. To this end, we separately train the Baichuan2-13B-Chat model using the INSTRUCTIE-ZH and the DuIE2.0 [19] datasets. Similarly, we also train the LLaMA2-13B-Chat on the INSTRUCTIE-EN and the NYT10 [33] datasets separately. Subsequently, we conduct zero-shot evaluations on the FewRel [8] and Wiki-ZSL [4] English relation extraction datasets, as well as the COAE2016⁹, IPRE [45], and CMeIE [26] Chinese relation extraction datasets. As shown in Table 5, in terms of average performance, the model train with INSTRUCTIE-EN outperforms the model train with NYT10 on both Chinese and English evaluations, achieving improvements of $\uparrow 7.83$ and $\uparrow 4.58$ respectively. The model train with INSTRUCTIE-ZH achieves comparable results to DuIE2.0 in Chinese evaluation, and a $\uparrow 3.26$ improvement in English evaluation. These outcomes suggest that the trained models acquire the enhanced ability to extend to unseen schemas, and the dataset created by the KG2Instruction framework possesses a higher quality than those purely generated through distant supervision, like the NYT10 dataset. In comparison with manually crowdsourced datasets like DuIE2.0, they also demonstrate a comparable level of performance.

5.2 Ablations on KG2Instruction

For our analysis, we randomly sample 1,000 examples from each domain within INSTRUCTIE-ZH and INSTRUCTIE-EN, conducting a series of analyses on two models: Baichuan2-13B-Chat and LLaMA2-13B-Chat. Our investigation focuses on evaluating the effectiveness of two key steps within KG2Instruction: (a) **Missing Triplets**

⁹ <https://github.com/Sewens/COAE2016>.

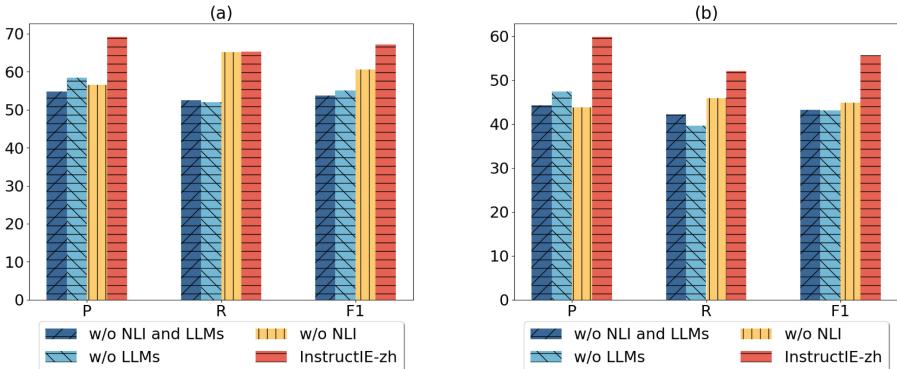


Fig. 5. (a) The results of Baichuan2-13B-Chat (LoRA tuning) on the INSTRUCTIE-ZH subset, (b) The results of LLaMA2-13B-Chat on the INSTRUCTIE-EN subset. The label **w/o LLMs** denotes the removal of the step “Missing Triplets Supplement with LLM”, **w/o NLI** indicates the removal of the step “Hallucinatory Triplets Filtering with NLI”, and **w/o NLI and LLMs** signifies the removal of both steps.

Supplement with LLM and **(b) Hallucinatory Triplets Filtering with NLI**. The results, as illustrated in Fig. 5, indicate that incorporating step (a) generally led to an increase in the recall values. We infer this improvement stems from the role of LLMs in supplementing triplet absences caused by the incompleteness of KGs. Concurrently, the introduction of step (b) results in a consistent increase in the precision values. By identifying and eliminating triplets generated from incorrect or false information, this mechanism enhances the accuracy of model predictions. In summary, the experimental results indicate that by integrating triplet augmentation with LLMs and implementing an NLI-based filtering mechanism for hallucinatory triplets, the KG2Instruction framework can significantly enhance the quality of the generated triplets.

5.3 Error Analysis

We conduct a detailed analysis of the predictions generated by the models and identify that the errors predominantly occur in the following four types. Additionally, examples of each error type are provided in Fig. 6.

1. **Entity Mismatch:** Within a triplet, either only the head entity or only the tail entity fails to align with the gold standard, while all other components remain accurate.
2. **Spurious Relation:** The model produces relations not reflected in the gold standard, indicating potential over-generation or hallucination.
3. **Boundary Mismatch:** The predicted head or tail entity depiction partially overlaps with the gold standard but fails to capture it entirely.
4. **Incongruent Predictions:** Several components of the prediction fail to align, rendering the output akin to arbitrary generation.

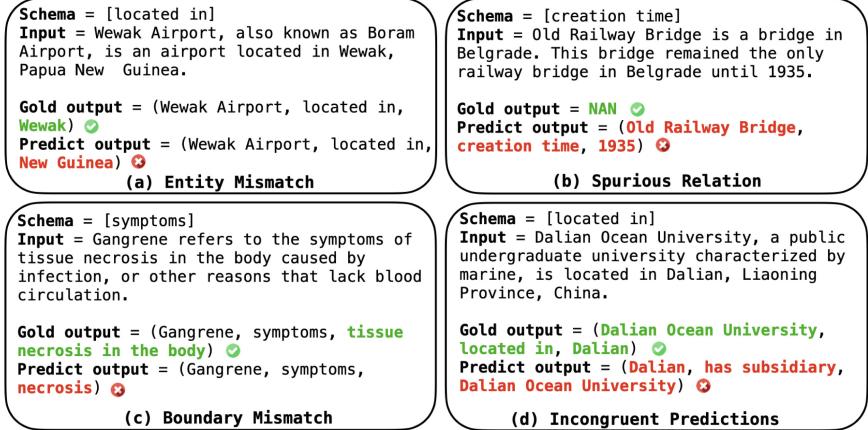


Fig. 6. Illustrations of four types of model prediction errors.

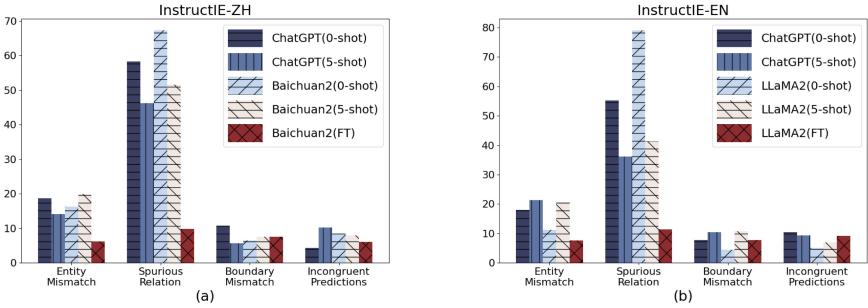


Fig. 7. The percentage of the 4 error predictions relative to all predictions on INSTRUCTIE-ZH and INSTRUCTIE-EN, where a smaller value indicates better performance.

In Fig. 7, we present the distribution of various error types across different strategies. When contextual examples are introduced, the model exhibits a significant reduction in the error rate associated with “Spurious Relation”. However, at the same time, the error rates in three entity-related categories: “Entity Mismatch”, “Boundary Mismatch”, and “Incongruent Predictions” show an increasing trend. We speculate that the contextual examples enhance the model’s ability to interpret instructions, making it more focused on extracting and outputting relationships according to the instructions, subsequently reducing the generation of inaccurate relations. Regrettably, despite the model’s improved capability to capture explicit relationships from the instructions, its ability to precisely determine entity boundaries remains suboptimal, leading to an increased incidence of entity-related errors. Notably, the model, after instruction tuning with INSTRUCTIE, exhibits lower error rates across all four mentioned error categories. This result suggests that targeted instruction tuning not only further enhances

the model’s ability to interpret instructions but also effectively reduces errors related to entity recognition, thereby improving the model’s overall accuracy.

6 Related Work

6.1 Information Extraction

Information Extraction (IE) aims to extract structured information from unstructured data sources automatically. Although traditional methods [15, 20, 58] that design specific architectures for different tasks and generative IE [23, 25, 32, 55, 56], which unifies various IE tasks into a sequence-to-sequence text generation task, have proven their efficacy in the past, their inherent limitations on pre-defined classes and a static training paradigm considerably hamper their adaptability, especially in the dynamic world. In contrast, instruction-based IE formulates IE as an instruction-driven generation task, capable of responding promptly to changes in instructions, and offers a more scalable solution. Some research such as [24, 34, 48, 51, 59] employ IE instruction data to fine-tune LLMs, aiming to enhance the model’s capability to understand and follow instructions to achieve instruction-based IE. However, many studies [5, 11, 13, 18, 27, 43, 44, 46, 53] indicate that instruction-based IE still faces performance issues, partly due to the limited availability of datasets annotated for IE and the lack of comprehensive coverage in domain labels. To address this problem, we introduce a new bilingual (Chinese and English) instruction-based IE dataset, encompassing 12 distinct domains and 123 types of schemas. This dataset aims to expand the corpus used in instruction tuning, with the hope of further advancing the development of LLMs in the field of IE.

6.2 Information Extraction Datasets

As depicted in Table 6, existing information extraction (IE) datasets [3, 7, 12, 14, 36, 39] currently face challenges such as small size, a narrow scope of covered domains, and a lack of richness in labels. These issues significantly constrain the applicability and development of IE technologies across wider domains. Moreover, the creation of high-quality IE datasets [10, 26, 36, 37, 39] typically relies on domain experts to select relevant corpus and guide data engineers in collecting and manually annotating data. This process is costly, time-consuming, and inefficient. Although some datasets [24, 33, 50] are automatically generated through distant supervision via KGs, these methods often result in incomplete labeling or inconsistencies between the labels and the original text. In light of this, we propose an automated IE data generation framework, KG2Instruction. This framework aims to address the shortcomings of distant supervision by leveraging a trained IE model to automatically complete missing triples in KGs and employing a natural language inference model to filter out unreliable triples. Through this methodology, we can generate IE datasets that are larger in scale, higher in quality, and more extensive in domain coverage, all at a lower cost and with greater efficiency.

Table 6. Descriptive metrics of some open-source IE datasets. **Domain-based** signifies the categorization of the label set based on different domains. **Annotation** refers to the processes involved in the data annotation of the training set, where **KG** represents distant supervision, **LLM** denotes model labeling, and **Human** indicates crowdsourced annotation. It is important to note that while INSTRUCTIE also contains a certain amount of general annotation, this is minimal compared to the total volume of the final dataset.

Dataset	Language	Task	#Class	#Size	Domain-based	Annotation	Human Engagement
CoNLL2003 [36]	en	NER	4	18,867	✗	Human	high
MSRA [16]	zh	NER	3	48,437	✗	Human	high
NYT10 [33]	en	RE	24	56,196	✗	KG	low
FewRel [8]	en	RE	100	70,000	✗	KG+Human	high
DuIE2.0 [19]	zh	RE+Entity	49	173,108	✗	KG+Human	high
INSTRUCTIE	en&zh	RE+Entity	123	364,076	✓	KG+LLMs	middle

7 Conclusion

In this paper, we introduce INSTRUCTIE, a Chinese-English bilingual instruction dataset for IE, designed to enhance the capability of LLMs to extract structured knowledge from text. We design and implement the KG2Instruction framework, which allows for the automated creation of the dataset.

INSTRUCTIE. The KG2Instruction framework not only generates a relationship extraction dataset but also automatically produces entity annotation data. Since our evaluation primarily focuses on the extraction of relationships between entities, i.e., triples, the entity data is considered a byproduct. Furthermore, we think that event extraction is also a form of triple extraction, and INSTRUCTIE includes data in the “events” domain, thus we consider it a comprehensive information extraction dataset.

Limitations. While the INSTRUCTIE dataset offers new opportunities for research in the construction of knowledge graphs from text, we also acknowledge several limitations of our study: Firstly, we only consider versions in two languages, Chinese and English, excluding other languages. Secondly, the dataset covers only 12 domains such as people, organizations, and works, without extending to more domains. Finally, although we have integrated LLMs to supplement the missing triples and natural language inference models to filter out hallucinatory triples, aiming to improve the quality of the distantly supervised labeled data, we still identify a certain amount of noise within the training set of INSTRUCTIE. Notably, we also found significant noise in crowdsourced datasets like DuIE2.0 and FewRel. Furthermore, as highlighted in [50], LLMs even when trained on noisy datasets, are capable of leveraging their robust capabilities to discover new triples that do not exist in INSTRUCTIE.

Future Work. In response to these limitations, we plan to expand the language dimension of the dataset to include more languages. Furthermore, we aim to enrich the domain covered by the dataset, specifically targeting domains with a higher degree of specialization, such as law, science, and finance. Concurrent with these efforts, we intend to conduct in-depth research into more precise and efficient data cleaning and augmentation techniques. This will allow us to optimize the framework for the automated construction of annotated data, thereby improving the quality of the dataset and its applicability to real-world applications.

Impact. With the rising popularity of LLMs, there is a surge of enthusiasm in research on knowledge graphs (KGs) construction and automatic data annotation using these models. The release of INSTRUCTIE aims at fostering research on LLMs for extracting knowledge from text to build KGs and automatic annotation, with the aspiration of eliciting substantial impact within the Semantic Web community. Notably, several works have already used InstructIE as a training set, including: YAYI-UIE [51], KnowCoder [21] and OneKE¹⁰.

Resource Availability Statement: The construction of our dataset utilizes a multitude of resources, including original corpus resources such as Wikidata¹¹ and Wikipedia¹², as well as several models: the chinese-roberta-wwm-ext-large¹³ and roberta-large¹⁴ models are employed for text classification; the Baichuan2-IEPile¹⁵ serves as the pre-training foundation for our IE model, and the mDeBERTa-v3-base-xnli-multilingual-nli-2mil7¹⁶ serves as our natural language inference model.

Acknowledgements. We would like to express our sincere gratitude to the anonymous reviewers for their thoughtful and constructive feedback. This work was supported by the National Natural Science Foundation of China (No. 62206246, No. NSFCU23B2055, No. NSFCU19B2027), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. This work was supported by Ant Group and Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph.

¹⁰ <https://huggingface.co/zjunlp/OneKE>.

¹¹ <https://www.wikidata.org/>.

¹² <https://www.wikipedia.org/>.

¹³ <https://huggingface.co/hfl/chinese-roberta-wwm-ext-large>.

¹⁴ <https://huggingface.co/FacebookAI/roberta-large>.

¹⁵ <https://huggingface.co/zjunlp/baichuan2-13b-iepile-lora>.

¹⁶ <https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7>.

References

1. Baichuan: Baichuan 2: open large-scale language models. arXiv preprint [arXiv:2309.10305](https://arxiv.org/abs/2309.10305) (2023). <https://arxiv.org/abs/2309.10305>
2. Brown, T.B., Mann, B., Ryder, N., et al.: Language models are few-shot learners. In: NeurIPS 2020 (2020)
3. Carreras, X., Màrquez, L.: Introduction to the CONLL-2004 shared task: semantic role labeling. In: Ng, H.T., Riloff, E. (eds.) 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, pp. 89–97. ACL (2004), <https://aclanthology.org/W04-2412/>
4. Chen, C., Li, C.: ZS-BERT: towards zero-shot relation extraction with attribute representation learning. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT2021, Online, June 6-11, 2021, pp. 3470–3479. Association for Computational Linguistics (2021). <https://doi.org/10.18653/V1/2021.NAACL-MAIN.272>
5. Gao, J., Zhao, H., Zhang, Y., Wang, W., Yu, C., Xu, R.: Benchmarking large language models with augmented instructions for fine-grained information extraction. CoRR [abs/2310.05092](https://arxiv.org/abs/2310.05092) (2023). <https://doi.org/10.48550/ARXIV.2310.05092>
6. Gui, H., et al.: IEPile: unearthing large-scale schema-based information extraction corpus. CoRR [abs/2402.14710](https://arxiv.org/abs/2402.14710) (2024). <https://doi.org/10.48550/ARXIV.2402.14710>
7. Gurulingappa, H., Rajput, A.M., Toldo, L.: Extraction of adverse drug effects from medical case reports. J. Biomed. Semant. **3**, 15 (2012). <https://doi.org/10.1186/2041-1480-3-15>
8. Han, X., et al.: FewRel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp. 4803–4809. Association for Computational Linguistics (2018). <https://doi.org/10.18653/V1/D18-1514>
9. He, H., Choi, J.D.: The stem cell hypothesis: dilemma behind multi-task learning with transformer encoders. In: EMNLP 2021, pp. 5555–5577. Association for Computational Linguistics (2021)
10. Hendrickx, I., et al.: SemEval-2010 Task 8: multi-way classification of semantic relations between pairs of nominals. In: Erk, K., Strapparava, C. (eds.) Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010, pp. 33–38. The Association for Computer Linguistics (2010). <https://aclanthology.org/S10-1006/>
11. Huang, K., et al.: A reevaluation of event extraction: Past, present, and future challenges. CoRR [abs/2311.09562](https://arxiv.org/abs/2311.09562) (2023). <https://doi.org/10.48550/ARXIV.2311.09562>
12. Jat, S., Khandelwal, S., Talukdar, P.P.: Improving distantly supervised relation extraction using word and entity based attention. In: 6th Workshop on Automated Knowledge Base Construction, AKBC@NIPS 2017, Long Beach, California, USA, December 8, 2017. Open-Review.net (2017)
13. Jiao, Y., et al.: Instruct and Extract: instruction tuning for on-demand information extraction. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pp. 10030–10051. Association for Computational Linguistics (2023). <https://aclanthology.org/2023.emnlp-main.620>

14. Kocaman, V., Talby, D.: Biomedical named entity recognition at scale. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I, pp. 635–646. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-68763-2_48
15. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: HLT-NAACL 2016, pp. 260–270. The Association for Computational Linguistics (2016)
16. Levow, G.: The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In: Ng, H.T., Kwong, O.O.Y. (eds.) Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22–23, 2006, pp. 108–117. Association for Computational Linguistics (2006). <https://aclanthology.org/W06-0115/>
17. Li, B., et al.: Evaluating ChatGPT’s information extraction capabilities: an assessment of performance, explainability, calibration, and faithfulness. CoRR **abs/2304.11633** (2023). <https://doi.org/10.48550/ARXIV.2304.11633>
18. Li, P., et al.: CodeIE: large code generation models are better few-shot information extractors. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023, pp. 15339–15353. Association for Computational Linguistics (2023). <https://doi.org/10.18653/V1/2023.ACL-LONG.855>
19. Li, S., et al.: DuIE: a large-scale Chinese dataset for information extraction. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11839, pp. 791–800. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_72
20. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A unified MRC framework for named entity recognition. In: ACL 2020, pp. 5849–5859. Association for Computational Linguistics (2020)
21. Li, Z., et al.: KnowCoder: coding structured knowledge into LLMs for universal information extraction. CoRR **abs/2403.07969** (2024). <https://doi.org/10.48550/ARXIV.2403.07969>
22. Liu, X., Huang, H., Shi, G., Wang, B.: Dynamic prefix-tuning for generative template-based event extraction. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22–27, 2022, pp. 5216–5228. Association for Computational Linguistics (2022). <https://doi.org/10.18653/V1/2022.ACL-LONG.358>
23. Lou, J., et al.: Universal information extraction as unified semantic matching. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7–14, 2023, pp. 13318–13326. AAAI Press (2023). <https://doi.org/10.1609/aaai.v37i11.26563>
24. Lu, K., Pan, X., Song, K., Zhang, H., Yu, D., Chen, J.: PIVOINE: instruction tuning for open-world information extraction. CoRR **abs/2305.14898** (2023). <https://doi.org/10.48550/arXiv.2305.14898>
25. Lu, Y., et al.: Unified structure generation for universal information extraction. In: ACL 2022, pp. 5755–5772. Association for Computational Linguistics (2022)
26. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp. 3219–3232. Association for Computational Linguistics (2018). <https://doi.org/10.18653/V1/D18-1360>

27. Ma, Y., Cao, Y., Hong, Y., Sun, A.: Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023, pp. 10572–10601. Association for Computational Linguistics (2023). <https://aclanthology.org/2023.findings-emnlp.710>
28. Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K.: Text2KG Bench: a benchmark for ontology-driven knowledge graph generation from text. In: Payne, T.R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., Li, J. (eds.) The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part II, pp. 247–265. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-47243-5_14
29. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Su, K.Y., Su, J., Wiebe, J., Li, H. (eds.) Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011. Association for Computational Linguistics, Suntec, Singapore (Aug 2009). <https://aclanthology.org/P09-1113>
30. OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023). <https://doi.org/10.48550/ARXIV.2303.08774>
31. Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. In: NeurIPS 2022 (2022)
32. Paolini, G., et al.: Structured prediction as translation between augmented natural languages. In: ICLR 2021. OpenReview.net (2021)
33. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
34. Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O.L., Rigau, G., Agirre, E.: GoLLIE: annotation guidelines improve zero-shot information-extraction. CoRR **abs/2310.03668** (2023). <https://doi.org/10.48550/ARXIV.2310.03668>
35. Sainz, O., de Lacalle, O.L., Labaka, G., Barrena, A., Agirre, E.: Label verbalization and entailment for effective zero and few-shot relation extraction. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7–11 November, 2021, pp. 1199–1212. Association for Computational Linguistics (2021). <https://doi.org/10.18653/V1/2021.EMNLP-MAIN.92>
36. Sang, E.F.T.K., Meulder, F.D.: Introduction to the CONLL-2003 shared task: language-independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003, pp. 142–147. ACL (2003). <https://aclanthology.org/W03-0419/>
37. Satyapanich, T., Ferraro, F., Finin, T.: CASIE: extracting cybersecurity event information from text. 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, pp. 8749–8757. AAAI Press (2020). <https://doi.org/10.1609/AAAI.V34I05.6401>
38. Smirnova, A., Cudré-Mauroux, P.: Relation extraction using distant supervision: a survey. ACM Comput. Surv. **51**(5), 1–35 (2019). <https://doi.org/10.1145/3241741>

39. Sun, Z., et al.: PHEE: a dataset for pharmacovigilance event extraction from text. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pp. 5571–5587. Association for Computational Linguistics (2022). <https://doi.org/10.18653/V1/2022.EMNLP-MAIN.376>
40. Touvron, H., Lavril, T., Izacard, G., et al.: LLaMA: open and efficient foundation language models. CoRR [abs/2302.13971](https://arxiv.org/abs/2302.13971) (2023)
41. Touvron, H., Martin, L., Stone, K., et al.: LLaMA 2: open foundation and fine-tuned chat models. CoRR [abs/2307.09288](https://arxiv.org/abs/2307.09288) (2023). <https://doi.org/10.48550/arXiv.2307.09288>
42. Vania, C., Lee, G., Pierleoni, A.: Improving distantly supervised document-level relation extraction through natural language inference. In: Cherry, et al: (eds.) Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pp. 14–20. Association for Computational Linguistics, Hybrid (Jul 2022). <https://doi.org/10.18653/v1/2022.deeplo-1.2>
43. Wadhwa, S., Amir, S., Wallace, B.C.: Revisiting relation extraction in the era of large language models. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 15566–15589. Association for Computational Linguistics (2023). <https://doi.org/10.18653/V1/2023.ACL-LONG.868>
44. Wan, Z., et al.: GPT-RE: in-context learning for relation extraction using large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pp. 3534–3547. Association for Computational Linguistics (2023). <https://aclanthology.org/2023.emnlp-main.214>
45. Wang, H., He, Z., Ma, J., Chen, W., Zhang, M.: IPRE: a dataset for inter-personal relationship extraction. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11839, pp. 103–115. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_9
46. Wang, J., et al.: TechGPT-2.0: a large language model project to solve the task of knowledge graph construction (2024)
47. Wang, S., et al.: GPT-NER: named entity recognition via large language models. CoRR [abs/2304.10428](https://arxiv.org/abs/2304.10428) (2023). <https://doi.org/10.48550/arXiv.2304.10428>
48. Wang, X., et al.: InstructUIE: multi-task instruction tuning for unified information extraction. CoRR [abs/2304.08085](https://arxiv.org/abs/2304.08085) (2023). <https://doi.org/10.48550/ARXIV.2304.08085>
49. Wei, X., et al.: Zero-shot information extraction via chatting with ChatGPT. CoRR [abs/2302.10205](https://arxiv.org/abs/2302.10205) (2023). <https://doi.org/10.48550/arXiv.2302.10205>
50. Whitehouse, C., Vania, C., Aji, A.F., Christodoulopoulos, C., Pierleoni, A.: WebIE: faithful and robust information extraction on the web. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 7734–7755. Association for Computational Linguistics (2023). <https://doi.org/10.18653/v1/2023.acl-long.428>
51. Xiao, X., et al.: YAYI-UIE: a chat-enhanced instruction tuning framework for universal information extraction. CoRR [abs/2312.15548](https://arxiv.org/abs/2312.15548) (2023). <https://doi.org/10.48550/ARXIV.2312.15548>
52. Xie, T., Li, Q., Zhang, Y., Liu, Z., Wang, H.: Self-improving for zero-shot named entity recognition with large language models. CoRR [abs/2311.08921](https://arxiv.org/abs/2311.08921) (2023). <https://doi.org/10.48550/ARXIV.2311.08921>
53. Xu, D., et al.: Large language models for generative information extraction: a survey. CoRR [abs/2312.17617](https://arxiv.org/abs/2312.17617) (2023). <https://doi.org/10.48550/ARXIV.2312.17617>

54. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. In: HLT-NAACL 2021, pp. 483–498. Association for Computational Linguistics (2021)
55. Ye, H., Gui, H., Xu, X., Chen, H., Zhang, N.: Schema-adaptable knowledge graph construction (2023)
56. Ye, H., Zhang, N., Chen, H., Chen, H.: Generative knowledge graph construction: a review. In: EMNLP 2022, pp. 1–17. Association for Computational Linguistics (2022)
57. Zeng, D., Zhang, H., Liu, Q.: CopyMTL: copy mechanism for joint extraction of entities and relations with multi-task learning. 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, pp. 9507–9514. AAAI Press (2020). <https://doi.org/10.1609/AAAI.V34I05.6495>
58. Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., Xu, B.: Joint extraction of entities and relations based on a novel tagging scheme. In: ACL 2017, pp. 1227–1236. Association for Computational Linguistics (2017)
59. Zhou, W., Zhang, S., Gu, Y., Chen, M., Poon, H.: UniversaLNER: targeted distillation from large language models for open named entity recognition. CoRR **abs/2308.03279** (2023). <https://doi.org/10.48550/ARXIV.2308.03279>